这个无法验证的世界

An Unverifiable World

欧长坤

目录

序　没有神谕的世界

古希腊人出征、婚嫁、建城之前，会先去德尔斐求一次神谕。神谕的意义不在它有多准，而在它承诺了一件事：在你行动之前，存在一个能告诉你答案的地方。两千年后，计算机科学家借走了这个词。在他们那里，神谕（oracle）是一个黑箱，你把一个自己算不出的问题递进去，它当即吐出正确答案。两种神谕共享同一个幻想：在动手之前，先把对错验明。

这本书讲的，是这个幻想破灭之后的世界。

我们几乎从不验证，我们只是行动，然后或迟或早地知道，或者永远不知道。我们以为「凡事可检验」是常态，是因为我们最早的训练来自一类特别狭小的事：算术、给清单排序、核对一张收据。在那些事里，答案触手可及，于是我们误以为整个世界都该如此。可一旦走出这一小块，验证立刻变成奢侈品。你能验证七乘八，你无法在说「我愿意」之前验证这段婚姻会长久，无法在上线之前验证这个代码库没有 bug，无法在投身之前验证一个理论为真、一家公司是健康的、一个决定是对的。大多数重大的行动，都踩在未经验证的地面上。神谕没有回话，而你还是得迈步。

面对这个处境，常见的反应是哀叹或假装。哀叹的人说，既然什么都无法确定，那一切判断都不过是臆断；假装的人则给自己造一个假神谕，把一个测得出的数字供起来，假装它就是那个测不出的真相。这本书两样都不做。它问一个更有意思的问题：那些确实有能力的人，科学家、工程师、数学家、治理者，在神谕缺席时，究竟做了什么？

把这个问题在足够多的领域里追下去，会撞见一个出人意料的观察，它是全书的由来：尽管无法验证的来源天差地别，有能力的人的应对却反复收敛到同一小套。

这就引出本书的两层结构，请先记住，因为后面所有章节都挂在它上面。

第一层，问题是异质的。「我没法检验它」这句话底下，藏着五种结构全然不同的处境：有些原则上就没有判定程序（不可判定 undecidable），有些有程序却代价大到不可能（难解 intractable），有些是相关状态对你隐藏（部分可观测 partially observable），有些是你本可验证却没有那个时间、算力或样本（预算受限 budget-constrained），还有些是对面那个系统在主动挫败你的验证（对抗 adversarial）。把这五种等量齐观，是这个领域最常犯的错。第一部会把它们一一掰开。

第二层，应对却收敛。无论问题是哪一种处境，有能力的主体伸手去够的，反复是同样几样东西：用一个测得出的代理（proxy）替换测不出的真目标，在一个能查的切片上证一个界，把昂贵的查验花在信息量最大处，引进一个外部的判断者，缩小失败的爆炸半径，给残余的风险标定（calibration）一个概率，把检查从事前挪到事后，用多个互相独立的判断去抵消单点的失误。本书把它们归纳为八招，并论证这八招可以收进四根更基本的杠杆。第二部走进四个现场，让这些招数嵌在各自的行话里、交织地出现；第三部再把每一招单独拎出来、整理、命名，铺成一张跨领域的对照表，那是这本书真正的载荷；第四部追问，为什么偏偏是这几招。

有一句话得先撂在前头。这套收敛，究竟是一条定律（某种东西迫使任何有限的主体都必然走到这几招上），还是仅仅一个很强的经验模式（我们一再看到它，却没能证明它非如此不可）？我此刻没有证据说它是定律。这本书交付的，是一个把边界划清了的猜想，外加一套能把许多领域串起来的共同词汇，而不是一条定理。第 14 章会正面讨论这件事。

而这恰恰带来一个无法回避、也不该回避的递归：一本论述「如何在无法验证中行动」的书，自己也无法验证它的核心命题。于是它

只能做它通篇所描述的那件事，陈述一个标定的信念，给主张划清边界，邀请你来反驳，然后照样把话说下去。这本书会亲自演练它所讲的那些方法。如果它讲对了，这种自我演练就不是缺陷，而是它唯一站得住的写法。

最后留一个画面，跋会回到它。一艘船在浓雾里改变航向。船长手上有海图、有罗盘、有对洋流的估算，唯独没有一双能看穿雾的眼睛。她无法在转舵之前验证前方是不是暗礁。雾不会散，神谕不会来。可航行不能因此停下。这本书想弄清楚的，不是怎样等到雾散，而是一个好船长在雾里究竟是怎样操舵的。

参考文献

> 落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. H. A. Simon (1969).《The Sciences of the Artificial》. MIT Press. [②④] 西蒙在此区分自然科学与「人造物的科学」，论证设计是一门以有限理性应对复杂环境的学问，并提出近似分解、层级结构与满意化等思路。它为本书的核心立场提供了底色：行动主体并不追求验明一切，而是在算力与信息受限下设计出够用的应对，正对应本节标注的「理论上被研究过的东西」与「如何在无法验证的世界里生活」。
2. F. H. Knight (1921).《Risk, Uncertainty and Profit》. Houghton Mifflin. [②] 奈特在此划出影响深远的一道界线：「风险」是概率已知、可被度量的不确定，而真正的「不确定」连概率分布都无从给定。他进而把企业利润归因于承担后一类不可度量的不确定。这条区分是本书谈论「不可验证」的概念源头之一，提醒读者把测得出概率的处境与连概率都测不出的处境分开。
3. N. N. Taleb (2007).《The Black Swan: The Impact of the Highly Improbable》. Random House. [②④] 塔勒布论证，

极少数难以预见、影响巨大、事后又被强行解释为可预测的「黑天鹅」事件，主导了历史与市场的走向，而常规的钟形分布统计会系统性地低估它们。本书可读其对预测局限的诊断：当尾部事件无法事先验证时，与其追求精确预报，不如调整自身对意外的暴露方式。

4. W. C. Wimsatt (2007).《Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality》. Harvard University Press. [②③④] 维姆萨特主张，认知能力有限的存在者不可能掌握完备真理，只能借助偏倚却好用的启发式、稳健性分析与分段近似来逼近实在，而科学正是这样一种逐步生成的工程。这本书几乎是本章主旨的哲学对应物，值得读其对「稳健性」与多重独立路径相互印证的论述，呼应本节标注的三个落足点。
5. J. M. Keynes (1921).《A Treatise on Probability》. Macmillan. [②] 凯恩斯把概率理解为命题之间的一种逻辑关系，即给定证据下信念的合理程度，并指出许多概率根本无法用数字精确衡量，甚至彼此不可比较。他还引入「证据权重」来刻画证据多寡本身。本书提供了一个早于现代决策论的视角：当证据稀薄时，量化的信心未必成立，正是不可验证处境的题中之义。
6. L. J. Savage (1954).《The Foundations of Statistics》. Wiley. [②] 萨维奇为主观期望效用奠定公理基础：只要一个人的偏好满足若干一致性公理，他的选择就如同在按某个主观概率最大化期望效用。这是把不确定纳入理性计算的标准框架。本书是理解后续争论的基准：唯有先看清它对一致性的要求，才能看懂埃尔斯伯格等人如何指出真实判断对它的偏离。
7. D. Ellsberg (1961).「Risk, Ambiguity, and the Savage Axioms」. Quarterly Journal of Economics, 75(4), 643-669. [②] 埃尔斯伯格用著名的摸球实验表明，人们普遍偏好概率已知的赌局而回避概率不明的赌局，这种「模糊厌恶」系统性地违反萨维奇公理，无法用任何单一主观概率来调和。本文是奈特式区分的实验证据，说明连概率本身都不确定时，理性主体的反应不同于面对纯粹风险，对应本节「理论上被研究过的东西」。

8. H. A. Simon (1955).「A Behavioral Model of Rational Choice」. Quarterly Journal of Economics, 69(1), 99-118. [②④] 西蒙在这篇奠基性论文中提出「有限理性」与「满意化」：受认知与信息限制的主体并不穷举所有选项求最优，而是设定一个抱负水平，找到第一个达标的方案便停手。这是本书反复借用的母题，说明在无法完全验证时，「够好即止」往往是理性的形态而非失败。
9. H. A. Simon (1947).《Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization》. Macmillan. [②④] 西蒙在此把组织理解为放大个体有限理性的决策结构，论证组织通过设定前提、划分职责与建立惯例，使成员在不完全信息下仍能做出可接受的选择。本书把有限理性从个人推广到机构层面，对应本书第二部对「现场」中应对机制的关注：制度本身就是一种集体的应对装置。
10. A. Tversky & D. Kahneman (1974).「Judgment under Uncertainty: Heuristics and Biases」. Science, 185(4157), 1124-1131. [②] 特沃斯基与卡尼曼指出，人在不确定下依赖代表性、可得性与锚定等少数启发式来估计概率，这些捷径通常有效，却会导致可预测的系统性偏差。本文开启了「启发式与偏差」研究纲领，是理解人类判断在何处可靠、何处失灵的起点，对应本节「理论上被研究过的东西」。
11. D. Kahneman & A. Tversky (1979).「Prospect Theory: An Analysis of Decision under Risk」. Econometrica, 47(2), 263-291. [②④] 前景理论用一个相对参照点的价值函数与对概率的非线性加权，刻画真实选择如何偏离期望效用：人们对损失比对等量收益更敏感，并高估小概率、低估中高概率。它是对萨维奇式规范理论的描述性修正，本书可读其对「人实际如何在风险下取舍」的精细刻画。
12. D. Kahneman (2011).《Thinking, Fast and Slow》. Farrar, Straus and Giroux. [②④] 卡尼曼以「系统一」（快速、直觉）与「系统二」（缓慢、费力）的双过程框架，综述了数十年关于判断偏差与决策的研究。本书是这一研究传统面向读者的总览，适合用来建立对认知局限的整体图景，理解为何即便专家也需要外部纠错机制，呼应本节「如何在无法验证的世界里

生活」。

13. G. Gigerenzer & D. G. Goldstein (1996).「Reasoning the Fast and Frugal Way: Models of Bounded Rationality」. Psychological Review, 103(4), 650-669. [②④] 吉仁泽与戈尔茨坦提出「快速节俭」启发式，论证只用少量线索、按序停止搜索的简单规则，在现实环境中常能逼近甚至超越复杂统计模型的表现。这与卡尼曼传统的「启发式即偏差」形成对照：本书可读其对「简单何以有效」的辩护，理解有限理性也可以是一种生态上的优势。
14. G. Gigerenzer, P. M. Todd & the ABC Research Group (1999).《Simple Heuristics That Make Us Smart》. Oxford University Press. [②④] 这本论文集系统铺陈「适应性工具箱」的纲领：心智配备一组针对特定环境的简单启发式，其有效性来自与环境结构的契合，即「生态理性」。它把前一篇的单点论证扩展为完整研究计划，本书可读其大量实证案例，看简单规则如何在信息不足时稳健地做出好判断。
15. F. A. Hayek (1945).「The Use of Knowledge in Society」. American Economic Review, 35(4), 519-530. [②④] 哈耶克论证，社会所需的知识本质上是分散的、与具体时空相关的，无法汇总到任何一个中央计划者手中，而价格机制恰恰是把这些分散信息协调起来的去中心装置。本文重要在于揭示一种不可验证的根源：相关信息从未被任何单一主体完整掌握，对应本书所说的「部分可观测」处境。
16. M. Polanyi (1958).《Personal Knowledge: Towards a Post-Critical Philosophy》. Routledge & Kegan Paul. [①③④] 波兰尼论证，一切认识都含有不可言传的「默会知识」与认识者的个人投入，纯客观、可完全形式化的知识是一种幻象。本书重要在于解释科学家的判断为何无法被规则完全替代，对应本节三个落足点，呼应本书对专家直觉与亲历判断的关注。
17. P. E. Meehl (1954).《Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence》. University of Minnesota Press. [①④] 米尔综述大量研究后得出一个令专家不安的结论：简单的统计或精算式预测，在准确度上往往不逊于甚至超过临床专家的直觉判断。本书是「把

判断外包给可核查的规则」这一招的经典证据，提示读者专家的自信与其实际准确度未必相符，对应本节「历史上科学家的判断」。

18. D. A. Schön (1983).《The Reflective Practitioner: How Professionals Think in Action》. Basic Books. [①④] 舍恩提出「行动中的反思」，论证专业人士面对独特而模糊的实践情境时，靠的不是套用既定理论，而是在行动中即时地与情境对话、不断重构问题。本书重要在于刻画了一种无法事先验证的专业能力，与米尔的统计预测形成张力，对应本书对实践者如何在雾中操舵的关心。
19. G. A. Klein (1998).《Sources of Power: How People Make Decisions》. MIT Press. [①④] 克莱因通过对消防员、护士等真实专家的现场研究，提出「识别启动决策」模型：富有经验者在时间压力下并不比较选项，而是凭模式识别直接生成一个可行方案再做心理模拟。本书是自然主义决策的代表作，说明专家直觉在何种条件下可靠，为本书对专业判断的讨论提供经验支撑。
20. P. E. Tetlock (2005).《Expert Political Judgment: How Good Is It? How Can We Know?》. Princeton University Press. [①④] 泰特洛克历经多年追踪政治与经济专家的预测，发现其平均准确度往往不如简单的外推，且认知风格比专业本身更能解释优劣：思路驳杂、自我怀疑的「狐狸」胜过执守单一大理论的「刺猬」。本书重要在于用可核查的记分把专家判断真正置于检验之下，对应本节「历史上科学家的判断」。
21. P. E. Tetlock & D. Gardner (2015).《Superforecasting: The Art and Science of Prediction》. Crown. [①④] 本书是泰特洛克预测锦标赛研究的延续，刻画出一类「超级预测者」：他们把大问题拆解、给出可计分的概率、依新证据频繁微调，并以团队互校来提升准确度。它把预测从天赋还原为可习得的实践，正对应本书所倡的事后校准与多重独立判断，呼应本节「如何在无法验证的世界里生活」。
22. N. N. Taleb (2001).《Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets》. Texere. [②④] 塔勒布论证人们惯于把随机产生的结果误读为技能或必

然，尤其在金融市场中把幸存者当成高手，从而低估了运气与噪声的作用。本书可读其对「事后归因」陷阱的剖析，提醒读者在无法验证因果时，成功的事实本身并不证明判断正确。

23. N. N. Taleb (2012).《Antifragile: Things That Gain from Disorder》. Random House. [④] 塔勒布提出「反脆弱」概念：有些系统不仅能承受波动，还能从无序与冲击中获益，与之相对的是脆弱与仅仅强韧。他主张在无法预测的世界里，应通过保留可选项、限制下行风险来主动从意外中受益。本书直接关联本书所谈的「缩小失败的爆炸半径」，对应本节「如何在无法验证的世界里生活」。
24. C. E. Lindblom (1959).「The Science of "Muddling Through"」. Public Administration Review, 19(2), 79-88. [④] 林德布洛姆论证现实中的公共政策并非自上而下的理性全局优化，而是「渐进主义」：在现状附近做有限的小幅调整、边走边比较、与已有手段挂钩。本文重要在于把「步步为营、随时纠偏」正名为一种应对复杂的合理策略，对应本书把检查从事前挪到事后、用小步迭代控制风险的思路。
25. K. R. Popper (1959).《The Logic of Scientific Discovery》. Hutchinson. [③] 波普尔系统提出证伪主义：科学理论无法被经验证实，只能被经验否证，因此可证伪性才是划分科学与非科学的标准。本书是本书主旨的哲学源头之一，正面回应「凡事可验证」的幻想，说明即便是科学也并非靠验证为真，而是靠经得起反驳来前进，对应本节「科学如何进展」。
26. T. S. Kuhn (1962).《The Structure of Scientific Revolutions》. University of Chicago Press. [①③] 库恩提出，科学并非线性累积，而是在「常规科学」与「科学革命」之间交替：研究在某一范式下解谜，待反常累积到危机，才会发生范式转换。他还指出竞争范式之间存在不可通约性。本书重要在于揭示科学进步中判断与共同体的作用，而非纯粹的逻辑验证，对应本节「历史上科学家的判断」与「科学如何进展」。
27. W. V. Quine (1951).「Two Dogmas of Empiricism」. The Philosophical Review, 60(1), 20-43. [③] 蒯因攻击逻辑经验主义的两条教条，即分析与综合命题的截然二分，以及每个命题可单独还原为经验。他主张信念以整体方式面对经验法庭，

任何陈述都可在调整别处的前提下被保留。本文是「证据不充分决定理论」的经典论证，说明单凭观察无法唯一地裁定理论，对应本节「科学如何进展」。

28. P. Duhem (1954).《The Aim and Structure of Physical Theory》. Princeton University Press. [③] 迪昂论证物理学的实验从来不是对单一假说的检验，而是对整套理论与辅助假设的检验，因此一个反例无法明确指出错在何处。这就是后来与蒯因并称的「迪昂蒯因论题」的源头。本书重要在于从科学实践内部说明判决性实验的局限，是理解科学为何无法靠单点验证为真的关键，对应本节「科学如何进展」。
29. I. Hacking (1983).《Representing and Intervening: Introductory Topics in the Philosophy of Natural Science》. Cambridge University Press. [③] 哈金把科学哲学的重心从「表征」即理论与真理，转向「介入」即实验与操作，主张当我们能稳定地操纵某种实体去干预世界时，便有理由相信它实在，这就是著名的实验实在论。本书重要在于提示验证并不只是被动观察，而是动手介入，呼应本书把行动而非验证置于核心的视角，对应本节「科学如何进展」。

第一部　不可验证

第 1 章　验证的奢侈

> **论点**：事前就能确认某事为真、正确或安全的「完整验证」，在人类与机器的生活里是例外，不是常态。

七乘八，和其余的一切

你能验证七乘八等于五十六。你可以重数一遍、换个方法算一遍，或者干脆背出乘法表，几秒钟之内，对错板上钉钉。

现在换几件事。在说出「我愿意」之前，验证这段婚姻会长久；在按下上线之前，验证这个代码库没有一个 bug；在投入半生之前，验证你信奉的那个理论为真；在接下这份工作之前，验证这家公司是健康的。这些你都做不到。不是因为你不够努力，而是因为这类事情根本不提供「事前验明」这个选项。

本书的第一块基石，就是这个反差：能在事前确认某事为真、为对、为安全的「完整验证」，在人类与机器的生活里是例外，不是常态。我们之所以觉得它该是常态，只是因为我们的直觉，是在一小块特别规整的地方养成的。

验证廉价的那一小块

哪些事我们验得动？算一道算术，给一串数字排序，核对一张收据的总额，判断棋盘上这步走法是否合规。把这些放在一起看，它们共享几个隐秘的特征：对象是封闭的（所有相关的东西都摆在眼前），

是有限的（情形数得过来），答案是局部而即时的（不依赖远处或将来），而且存在一个机械的判定程序（照着做就有是或否）。

正是这一小块，喂养了我们「凡事可检验」的直觉。学校里反复奖励的，恰是这类有标准答案、可当场批改的题目。于是我们悄悄地把一条经验「在我练习过的事情里，对错总能查清」外推成了一条世界观「事情的对错总能查清」。这条外推是错的，而且错得很有系统。一旦走出这一小块，上面那四个特征几乎一个个都不灵了。

错觉破裂在四个地方

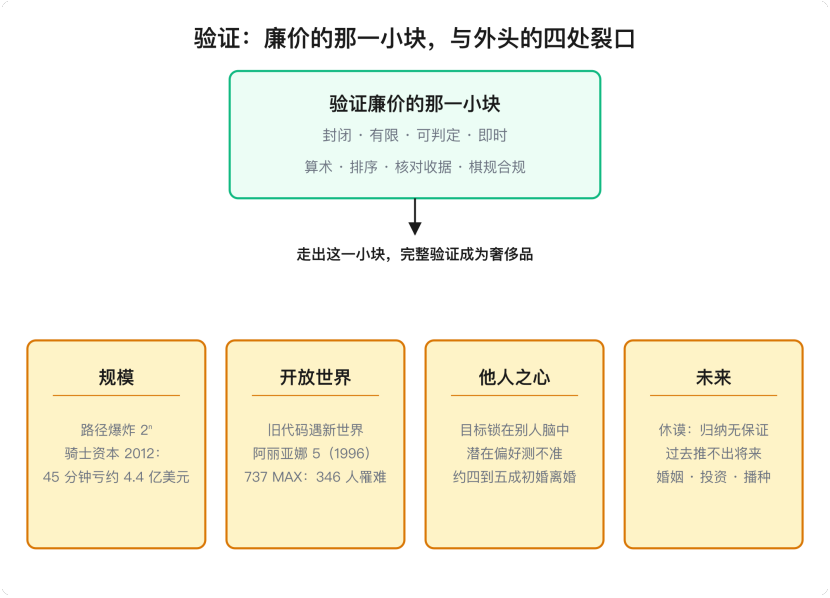

图 1: 验证：廉价的那一小块，与外头的四处裂口

规模（scale）。里头的情形数得过来，外头的数不过来。一个有 n 个分支的程序，可能的执行路径多达 2^n 条，几十个分支就足以让穷尽测试在宇宙寿命内都跑不完，这种路径爆炸（path explosion）让「测全」从一开始就出局。你能验证它在你想到的那几个输入上正确，无法验证它在所有输入上正确。2012 年 8 月 1 日，骑士资本（Knight Capital）部署新交易程序时，八台服务器里有一台没更新，

一段沉睡多年、早该弃用的旧代码被一个被复用的标志位意外唤醒，在开盘后约四十五分钟里疯狂吐出数百万笔订单，让公司亏掉约四亿四千万美元，几近一夜破产。没有人验证过那条废弃路径，因为没有人想得到它还会运行。检验单个情形容易，检验全部情形，量词「所有」一出现，就跨进了另一个世界。

开放世界（open world）。里头的对象封闭，外头的世界还在不断送来新东西。你测过的是有限几个场景，系统真正会遇到的是一个开放、未完待续的环境。1996 年 6 月 4 日，阿丽亚娜 5 型火箭首飞，升空约三十七秒后凌空自毁，肇因是一段从阿丽亚娜 4 型直接沿用、未经为新弹道重新验证的惯性导航代码，把一个 64 位浮点数硬塞进 16 位整数，而新火箭更高的横向速度让这个数溢出，连同搭载的四颗科学卫星，损失三亿七千万美元以上。代码在旧世界里跑了多年都正确，换了新世界就要命。波音 737 MAX 的 MCAS 系统是同一道裂口更惨痛的版本：它在一次次试飞里表现正常，却在真实航线上读到一个故障迎角传感器、反复把机头压下，两起空难（2018 年狮航 610、2019 年埃航 302）共夺去 346 条生命。你验证的永远是过去见过的切片，要赌的却是没见过的将来。

他人之心（other minds）。里头的状态可观测，外头你要满足的目标常常锁在另一个人的脑子里。本章开头那个「造对了却不是我要的」场景，根子就在这里：用户真正想要什么、上司满不满意、对方爱不爱你，这些是潜在变量（latent variable），你只能从行为旁敲侧击，无法直接读出，因而也无法直接验证你是否满足了它。连最郑重的人生承诺也躲不开，据人口学测算，约四到五成的美国初婚最终走向离婚，没有谁能在说出「我愿意」的那一刻验证它会长久。

未来。这是最深的一处，休谟在 1748 年就把它挑明了[17]：归纳（induction）没有逻辑保证。太阳过去每天升起，并不能在逻辑上证明它明天还升；有限的过去经验，无法事前验证任何关于未来的全称判断。我们依赖的不是证明，而是习惯。凡是结果落在将来的行动，婚姻、投资、播种、托付，都在这道裂口的另一侧。

连最硬的两个领域也低头

也许你会想，规模、人心、未来这些软的领域认输也就罢了，数学和软件总该是完整验证的堡垒吧？恰恰是这两个最硬的地方，最清醒地承认了验证的限度。

软件这边，迪杰斯特拉[13] 留下一句被引滥却仍然正确的话：测试只能证明缺陷存在，不能证明缺陷不存在。他主张程序应当被正确地构造出来，而不是被调试出正确。可即便是形式化证明这条最严的路，德米洛、利普顿与佩利斯 1979 年那篇争议名文[9] 也指出，程序验证（program verification）无法扮演数学证明那样的角色，它的可信最终来自社会过程，而非机械推导；费泽尔 1988 年把话说得更重[10]，程序作为一个因果模型，与作为逻辑结构的算法之间有一道鸿沟，「完全可靠的程序验证」连理论上都不成立。布鲁克斯的《没有银弹》[11] 断言软件的本质复杂性（essential complexity）无法被一招消除；帕纳斯辞去星球大战计划的顾问，公开论证那类系统的软件无法被验证到值得托付[12]；而 Therac-25 放疗机在 1985 至 1987 年间因一个并发竞态（race condition）缺陷六度失控，把高出正常上百倍的辐射打进病人体内，至少三人因此死亡[15]，是这一切判断用人命付的注脚。1968 年北约那场会议干脆造了个词，软件危机（software crisis）[16]。

数学这边更釜底抽薪。哥德尔 1931 年证明[3]，任何足够丰富而一致的形式系统（formal system），都存在它自己无法在内部判定的真命题；丘奇与图灵 1936 年各自证明[2,1]，没有算法能判定任意命题是否可证（判定问题无解）；赖斯定理（Rice's theorem）[4] 把它推到极致，程序的任何非平凡语义性质都不可判定（undecidable）。哪怕某个问题原则上可判定，库克 1971 年确立的 NP 完全性（NP-completeness）[5] 也表明，验证的代价可能爆炸到实践中根本跑不动。这些不是工程的暂时短板，是逻辑给验证划下的硬边界。这一层，下一章会专门去拆。

重述，以及这不是一句丧气话

把以上合起来：大多数有后果的行动，都踩在未经验证的地面上。

这不是一个让人瘫痪的结论，它是一个起点。承认验证是奢侈品，恰恰是认真对待行动的第一步。奈特 1921 年早就把可度量的「风险」与不可度量的「不确定性」分开[22]，并指出利润正来自后者；凯恩斯谈到真正的不确定时只留下一句「关于此我们根本无从知晓」[26]；西蒙看清有限的主体无法穷尽验证所有选项，于是提出「满意即止」（satisficing）[23]；冯·诺依曼与摩根斯特恩、萨维奇则各自为「在无法事前验证结果时如何理性地下注」搭起了形式框架[24,25]。一整门关于决策的学问，本就是建立在「验证不可得」这个前提之上的。问题从来不是怎样取消不确定，而是在不确定里怎样行动得当。

这一章通向哪里

既然验证通常不可得，第一个该问的就是：它为什么不可得？

答案不止一种，而这正是要紧之处。把「我没法检验它」当成一种处境，是这个领域最常见、也最误事的错。它其实是五种结构全然不同的处境，共用了一句话。下一章，我们把这句话掰成五瓣。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. A. M. Turing (1936). 「On Computable Numbers, with an Application to the Entscheidungsproblem」. Proceedings of the London Mathematical Society, s2-42, 230-265. [②] 图灵以一台抽象计算机器为模型，证明不存在能判定任意命题是否可证的算法，并由此导出停机问题不可判定。这是把验证的极限从工程经验提升为数学定理的奠基之作，本章「连最硬的两个领域也低头」一节正以此说明判定问题无解。文章所在的 series 2 第 42 卷横跨 1936 至 1937 年，部分书目标作 1937，正文采用通行的 1936。

2. A. Church (1936).「An Unsolvable Problem of Elementary Number Theory」. American Journal of Mathematics, 58(2), 345-363. [②] 丘奇以自己创立的 lambda 演算为工具，独立证明初等数论中存在不可解的判定问题，发表比图灵早数月。它与图灵的工作殊途同归，共同框定了「什么是可计算的」这条理论边界，提示读者验证之不可得在 1936 年已由两条独立路径同时确立。

3. K. Gödel (1931).「Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I」. Monatshefte für Mathematik und Physik, 38, 173-198. [②] 哥德尔证明，任何足够丰富而一致的形式系统都含有它自身无法证明也无法否证的真命题。这意味着「在系统内部把一切真理逐一验明」从原理上就办不到，是本章论证验证存在硬边界时最深的一块基石，下一章还会专门拆解。

4. H. G. Rice (1953).「Classes of Recursively Enumerable Sets and Their Decision Problems」. Transactions of the American Mathematical Society, 74, 358-366. [②] 赖斯定理把停机问题的不可判定性推到极致：程序的任何非平凡语义性质都不存在通用的判定算法。它告诉读者，关于「这段程序到底会做什么」的问题，几乎一概不可机械验证，是本章「最硬的领域也低头」一段的关键支撑。

5. S. A. Cook (1971).「The Complexity of Theorem-Proving Procedures」. Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC), 151-158. [②] 库克在此确立了 NP 完全性概念，证明可满足性问题对一大类问题具有普遍的计算难度。它揭示了验证的另一重限度：哪怕一个问题原则上可判定，求解或检验的代价也可能爆炸到实践中根本跑不完，对应本章谈「规模」失效的那一面。

6. C. A. R. Hoare (1969).「An Axiomatic Basis for Computer Programming」. Communications of the ACM, 12(10), 576-580. [②①] 霍尔提出以前置条件、后置条件与推理规则严格证明程序正确性的公理体系，即后世所称的霍尔逻辑。它代表

了「把验证做到底」这条最严路线的雄心，读者由此能看清形式化验证能走多远，以及它在工程现实中为何始终难以覆盖全部。

7. J. C. King (1976).「Symbolic Execution and Program Testing」. Communications of the ACM, 19(7), 385-394. [②] 金提出符号执行：用符号变量代替具体输入，沿程序的分支系统地推演各条路径所需满足的条件。这一技术既扩大了自动测试的覆盖面，也直观暴露出路径数随分支指数增长的「路径爆炸」，正是本章用来说明穷尽验证为何受限的根本困难。

8. E. M. Clarke 与 E. A. Emerson (1981).「Design and Synthesis of Synchronization Skeletons Using Branching Time Temporal Logic」. Logics of Programs (Lecture Notes in Computer Science 131), Springer, 52-71. [②] 这篇工作坊论文提出用分支时序逻辑自动检验系统是否满足给定性质，开创了模型检验。它代表机器验证真正落地的一支，但其威力以系统状态有限为前提，因而恰好划出了自动验证能够触及与无能为力的边界。文章收入 LNCS 第 131 卷，属会议论文集而非期刊。

9. R. A. DeMillo, R. J. Lipton 与 A. J. Perlis (1979).「Social Processes and Proofs of Theorems and Programs」. Communications of the ACM, 22(5), 271-280. [①②] 三位作者论证，数学证明之所以可信，靠的是数学共同体反复阅读、复用与检验的社会过程，而冗长机械的程序验证缺乏这种过程，因而无法扮演数学证明那样的角色。这是对「形式化验证能给软件以确定性」的著名质疑，本章引它来说明验证的可信最终来自社会而非纯机械推导。

10. J. H. Fetzer (1988).「Program Verification: The Very Idea」. Communications of the ACM, 31(9), 1048-1063. [①②] 费泽尔把质疑推得更深：算法是逻辑结构，可被严格证明，而在真实机器上运行的程序是因果模型，其行为受硬件与世界制约，二者之间有一道无法弥合的鸿沟。他据此主张「完全可靠的程序验证」连理论上都不成立。此文引发 1989 年技术通信

的大规模论战，是本章界定验证逻辑边界的重要一环。

11. F. P. Brooks (1987). 「No Silver Bullet: Essence and Accidents of Software Engineering」. IEEE Computer, 20(4), 10-19. [①] 布鲁克斯区分软件的本质复杂性与附属复杂性，断言没有任何单一技术能在十年内让软件生产率有数量级提升，本质复杂性无法被一招消除。它支撑了本章的判断：缺陷不可能被某种银弹一举验证清除。此文原为 1986 年 IFIP 第 10 届世界计算机大会的邀请论文，初刊于 Information Processing 86, 1069-1076。

12. D. L. Parnas (1985).「Software Aspects of Strategic Defense Systems」. Communications of the ACM, 28(12), 1326-1335. [①] 帕纳斯在辞去星球大战计划顾问后撰文，逐条论证此类系统的软件无法经测试或证明而被验证到值得托付的程度。这是一位顶尖工程师以辞职为代价对验证极限作出的公开判断，本章引为「连最硬的领域也低头」的现实注脚。同年他另以系列短文见于 American Scientist。

13. E. W. Dijkstra (1972). 「The Humble Programmer」（1972 ACM 图灵奖演讲）. Communications of the ACM, 15(10), 859-866. [①] 这是迪杰斯特拉的图灵奖演讲，主张程序员应保持谦卑，正视人脑容量有限，并把程序当作应当被正确地构造出来的对象，而非靠事后调试修补成正确。它反映了一位奠基者对事后验证之局限的清醒判断，与本章主张相互呼应。

14. E. W. Dijkstra (1972).《Notes on Structured Programming》（载于 O.-J. Dahl, E. W. Dijkstra, C. A. R. Hoare 编《Structured Programming》）. Academic Press. [①] 本章那句被引滥却仍正确的话「测试只能证明缺陷的存在，不能证明其不存在」即出于此文。迪杰斯特拉在此系统阐述结构化程序设计，主张通过有纪律的构造而非穷举测试来获得正确性。该论断最早见于手稿 EWD249（1970），1972 年收入《Structured Programming》正式出版。

15. N. G. Leveson 与 C. S. Turner (1993). 「An Investigation of

the Therac-25 Accidents」. IEEE Computer, 26(7), 18-41. [①④] 两位作者对 Therac-25 放疗机因软件缺陷导致患者受过量辐射乃至死亡的系列事故作了权威调查，剖析了竞态条件、过度信任软件与缺乏独立安全机制等连锁原因。它以人命为代价说明，安全攸关系统未经充分验证即投用会有什么后果，是本章关于验证代价的沉重注脚。

16. P. Naur 与 B. Randell（编）(1969).《Software Engineering: Report on a Conference Sponsored by the NATO Science Committee》. Scientific Affairs Division, NATO. [①] 这份会议报告记录了从业者对当时软件普遍超期、超支、难以可靠交付的集体焦虑，「软件危机」一词与「软件工程」这门学科的提法即由此而来。它是本章那句「软件危机」的源头，集中呈现了一代工程师对软件无法被可靠验证的判断。会议于 1968 年 10 月在德国 Garmisch 召开，报告 1969 年出版。

17. D. Hume (1748).《An Enquiry Concerning Human Understanding》. (London). [④③] 休谟在此挑明了归纳问题：由过去屡屡如此推断将来仍会如此，并无逻辑上的保证，太阳明日是否升起无法事前证明，人之所以照常行动靠的是习惯而非证明。这是本章「未来」一处裂口的思想源头，也是全书反复回到的起点。本书 1748 年初版原题《Philosophical Essays Concerning Human Understanding》，1758 年改为今题。

18. K. Popper (1959).《The Logic of Scientific Discovery》. Hutchinson. [③] 波普尔系统提出证伪主义：科学理论无法被经验证实，只能被否证，可证伪性因而成为科学与非科学的分界，科学也正是经由不断尝试推翻理论而进展。它直接关系到本章「科学如何进展」这一落足点，揭示连科学也不靠正面验证累积。英文版系作者在德文原著《Logik der Forschung》（1934 年付印，版权页标 1935）基础上扩写而成。

19. W. V. O. Quine (1951).「Two Dogmas of Empiricism」. The Philosophical Review, 60(1), 20-43. [③] 蒯因批判分析与综合的截然二分以及还原论这两个经验论教条，提出整体论：理论是一张面对经验整体受检的信念之网，任何单个陈述

都无法被孤立地验证或反驳。它说明证据对理论的检验是欠决定的，深化了本章关于科学如何进展、验证为何不可逐句完成的讨论。

20. T. S. Kuhn (1962).《The Structure of Scientific Revolutions》. University of Chicago Press. [③] 库恩提出范式概念，描述科学如何在常规科学的积累与反常累积所引发的危机之间交替，最终经由范式转换而发生革命。其要点是科学并非靠对真理的逐步验证线性累积，而是经由不可通约的范式跃迁。它为本章「科学如何进展」提供了与波普尔互补又对照的图景。

21. I. Lakatos (1976).《Proofs and Refutations: The Logic of Mathematical Discovery》（J. Worrall 与 E. Zahar 编）. Cambridge University Press. [③②] 拉卡托斯以多面体欧拉公式的演变为课堂对话，展示数学概念与定理如何在反例、再证明与定义修订的往复中成长。它颠覆了「数学证明是一劳永逸的验证」这一印象，提示连最确定的领域也经由批判而推进，呼应本章对验证之有限的总论。

22. F. H. Knight (1921).《Risk, Uncertainty and Profit》. Houghton Mifflin. [④] 奈特把可用概率度量的「风险」与无从度量的真正「不确定性」区分开来，并论证企业家的利润正源于承担后者。这一区分是本章承认验证是奢侈品后转向行动理论的关键，它说明在无法事前验明结果的局面下，决策与回报如何获得意义。

23. H. A. Simon (1955).「A Behavioral Model of Rational Choice」. The Quarterly Journal of Economics, 69(1), 99-118. [④] 西蒙提出有限理性的行为模型：信息与计算能力都受限的主体无法穷尽比较所有选项，只能设定一个够用的水准，找到满足它的方案便停下，即「满意即止」。这正是本章主张的「在不确定里如何行动得当」的一个具体答案，把验证不可得转化为可操作的决策准则。

24. J. von Neumann 与 O. Morgenstern (1944).《Theory of Games and Economic Behavior》. Princeton University

Press. [④] 两位作者奠定了博弈论，并以一组公理推出期望效用，论证理性主体应据期望效用作选择。它为「在无法事前验明对手意图与结果的情形下如何理性地下注」搭起了形式框架，是本章所说那门建立在验证不可得之上的决策学问的支柱之一。

25. L. J. Savage (1954).《The Foundations of Statistics》. John Wiley & Sons. [④] 萨维奇为主观概率与个人主义决策论建立公理基础，证明只要偏好满足若干一致性条件，主体的行为就如同依据某个主观概率与效用在最大化期望效用。它给出了在无法客观验证概率的世界里仍能一致下注的理性标准，与冯·诺依曼的框架共同支撑本章对不确定下决策的讨论。

26. J. M. Keynes (1937).「The General Theory of Employment」. The Quarterly Journal of Economics, 51(2), 209-223. [④] 凯恩斯在回应对《通论》的批评时，强调真正的不确定性不可用概率度量，对有些事情「关于此我们根本无从知晓」。他指出投资决策在无法验证的未来面前只能依赖惯例与动物精神。本章那句对真正不确定的经典陈述即出于此，它佐证了决策学问以验证不可得为前提。

27. A. Tversky 与 D. Kahneman (1974).「Judgment under Uncertainty: Heuristics and Biases」. Science, 185(4157), 1124-1131. [④] 特沃斯基与卡尼曼通过实验表明，人在判断概率时常依赖代表性、可得性与锚定等启发式捷径，因而系统性地偏离概率规范。它从描述的层面补全了本章的图景：在无法完整验证概率的世界里，人实际怎样判断，又会一致地错在哪里。

28. N. N. Taleb (2007).《The Black Swan: The Impact of the Highly Improbable》. Random House. [④] 塔勒布称那些罕见、极端冲击且事后才被强行解释的事件为黑天鹅，主张它们无法被事前验证或预测，却往往主导历史走向。他据此提议人应放弃精确预测的幻想，转而构建能在意外面前不被摧毁甚至受益的安排。这呼应本章的结尾：问题不在取消不确定，而在不确定里如何活得稳妥。

第 2 章　不可验证的五种处境

> **论点**：「我没法检验它」掩盖了五种结构不同的处境，把它们混为一谈是这个领域的核心错误。

一句「我没法检验它」，听上去像一种处境，其实掩着五种。它们结构全然不同，可得的补救也全然不同，把它们混为一谈是这个领域最核心的错误。

这一章要做的，是把五种掰开、各自掐准。这件事看似只是分类的洁癖，实则是全书后半部分的信用额度。本书最终要论证：尽管不可验证的来源天差地别，应对却收敛到同一小套。这个论断要想不显得廉价，前提就是先把「天差地别」坐实。差异讲得越透，后面那份收敛才越是值得惊讶、值得解释。请把这五种处境记牢，它们会在全书反复点名。

第一种：不可判定

判据：原则上就不存在判定它的算法。不是难，是没有。

这是验证最硬的失败。希尔伯特与阿克曼 1928 年[4] 明确提出判定问题（Entscheidungsproblem），问能否有一个机械程序，对任意数学命题判定其真伪。八年后，丘奇[2] 用 lambda 演算、图灵[1] 用他那台抽象机器，各自证明不能。图灵的停机问题（halting problem）尤其干净，没有算法能对任意「程序加输入」判定它是否会停下。哥

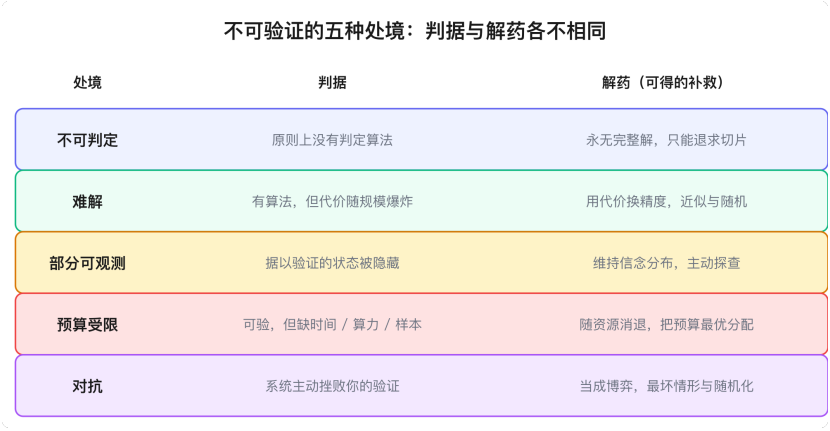

图 2: 不可验证的五种处境：判据与解药各不相同

德尔 1931 年[3] 的不完备性（incompleteness）、赖斯定理[6]（Rice's theorem，程序的任何非平凡语义性质都不可判定）、马蒂亚谢维奇 1970 年[7] 对希尔伯特第十问题的否决，都属于这一族。这并非纸上空谈：正因为「这段程序会不会做坏事」可归约到停机问题，理论上就不存在一款杀毒软件能完美无误地查出所有恶意程序，这是逻辑替反病毒行业划下的天花板。

这种处境的补救有一个独一无二的性质：永远不会有完整解。再多的时间、再快的机器都不行，因为障碍是逻辑的，不是资源的。你能做的，只有退而求其次，验证有限的切片，或把自己限制在那些确实可判定的片段里（比如只有加法的算术）。这一点会一直回响到第 7 章。

第二种：难解

判据：算法存在，但它的代价随规模爆炸，大到实践中跑不完。

这一种和上一种差之毫厘，谬以千里：可判定，却不可行。库克 1971 年[8]、列文 1973 年[9] 各自确立的 NP 完全性（NP-completeness），卡普 1972 年[10] 著名的二十一个 NP 完全问题，给了它精确的刻画。最坏情形的可满足性问题（satisfiability）、无数组合优化问题，原则

上都有解法，但那解法在最坏情形下的耗时随输入规模呈指数增长，

$$T(n) \sim 2^n,$$

几十个变量就足以让最快的超算望洋兴叹。一个量级上的直觉：国际象棋的博弈树约有 10^{120} 个分支（香农数），围棋的合法局面约 2×10^{170} 个，而整个可观测宇宙的原子数也不过约 10^{80}。这些棋类规则简单、原则上可穷举，可那个「原则上」远在物理可能性之外。P 是否等于 NP，正是在问这道墙是不是注定的。

它的补救与不可判定（undecidable）完全不同。这里多投入资源是有意义的，更要紧的是，你可以用「接受少一点」来换「付得起的代价」：近似解代替精确解、平均情形代替最坏情形、启发式搜索、随机化。难解（intractable）逼出的是一整套「打折」的智慧，这在不可判定那里是没有的。

第三种：部分可观测

判据：你要据以验证的那个状态，对你是隐藏的。

不是没有判定程序，也不是代价太大，而是你压根看不到该看的东西。用户真正的偏好、病人体内正在发生什么、对手手里的牌，这些驱动结果的状态却不对你显现。控制论很早就形式化了它：阿斯特罗姆 1965 年[15] 研究状态信息不完整下的最优控制，斯莫尔伍德与桑迪克 1973 年[16]、凯尔布林等人 1998 年[18] 把它发展成部分可观测马尔可夫决策过程（POMDP）这一标准框架。帕帕迪米特里乌与齐齐克利斯 1987 年[17] 还证明，求解这类问题本身又是难解的，于是第三种处境常和第二种叠在一起。

它的补救自成一类：你不再追求一个确定的判决，而是维持一个关于隐藏状态的信念分布（belief state），并用每一次观测去更新它，

$$b'(s') \;\propto\; \Pr(o \mid s') \sum_{s} \Pr(s' \mid s, a)\, b(s).$$

推断与探查，而不是「算得更狠」，才是这一种的解药。第 5 章整章都在这种处境里。

第四种：预算受限

判据：原则上可验、可解，但你这个主体，此时此地，没有那个时间、算力或样本。

这一种最朴素，也最普遍。一个评审只有二十分钟看一篇论文；一个医生只有几分钟做诊断；一个交易员必须在行情消失前下单。验证在理论上完全可行，落到一个有限的主体身上却不可行。奈特 1921 年[19]、西蒙 1955 年[20] 的有限理性（bounded rationality）是它的思想源头；迪安与博迪 1988 年[22] 的随时算法（anytime algorithm，随时可中断、给出当前最优解）、拉塞尔与苏布拉马尼安 1995 年[23] 的「有界最优」（bounded optimality），是它的形式化。

它的补救有一个别的处境都没有的特征：这种处境会随资源增长而消退。给足时间和算力，它就消失了。正因如此，对付它的核心在分配，把稀缺的预算花在边际收益最高处。这条思路，正是后面「最优筛查」那一招的来历。

第五种：对抗

判据：你面对的那个系统，在主动地挫败你的验证。

前四种里，难处来自世界的中立属性，逻辑的、规模的、可见性的、资源的。第五种不同：对面有一个智能，在针对你的检查做优化。会撒谎的对手、会伪装的恶意代码、会操纵指标的被考核者。冯 · 诺依曼与摩根斯特恩 1944 年[24] 的博弈论（game theory）、纳什 1950 年[25] 的均衡（Nash equilibrium）、瓦尔德 1945 年[26] 的极小极大准则（minimax），是它的经典理论；塞盖迪等人 2014 年[29] 发现的对抗样本（adversarial examples）、马德里等人 2018 年[31] 用鲁棒优化（robust optimization）统一攻防，是它在机器学习里的当代化身。一个识别率极高的模型，可以被人眼察觉不到的微小扰动骗得一塌糊涂，因为有人专门去找那个扰动。研究者们做过一个反复被

复现的演示：只要在一个停车标志上贴几张精心设计、看似涂鸦的小贴纸，就能让顶尖的图像识别系统稳定地把它读成限速牌。同一张标志，人看是停，机器看是行，差别只在对手往哪儿贴。

它的补救是战略，不是计算。你要做的不是把某个量算得更准，而是

$$\min_{x} \max_{y} L(x, y),$$

按最坏情形布防，用随机化剥夺对手对你的预测，追求在博弈里站得住而非在某个固定输入上最优。把对抗（adversarial）当成单纯的可观测缺口（「我只是还没看清它」）来处理，是会出人命的误判，因为它会顺着你的判断调整自己。

五种处境，五种解药

把它们并排放好，要紧的不是名字，是它们的补救彼此不可通约：

- 不可判定，永无完整解，只能退求切片。
- 难解，可用代价换精度，多投资源有意义。
- 部分可观测，靠推断信念、主动探查。
- 预算受限，随资源消退，核心在分配。
- 对抗，是一盘棋，靠战略与随机。

谁要是对你说「这事多堆点算力就解决了」，那他多半是把某一种处境错认成了另一种。把不可判定当成预算问题、把对抗当成可观测问题，都是这种错认，而且代价高昂。

这些处境还会叠加、复合。第 6 章那个放出去的智能体，同时撞上开放世界的不可预测（近乎不可判定的行为）和对手的策略性（对抗）；第 8 章那个组织，将部分可观测和对抗一起扛。真实处境往往是好几种处境的混合。

正因为来源如此参差、补救如此各异，下一个该问的问题就尖锐起来了：人类有没有一套成熟的办法，长期地、有纪律地与不可验证

共处？有的。那套办法叫科学，而它的第一条家规恰恰是公开承认自己永远无法验证。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

不可判定（②③）

1. A. M. Turing (1936).「On Computable Numbers, with an Application to the Entscheidungsproblem」. Proceedings of the London Mathematical Society, s2-42(1), 230–265. [②③] 图灵在此引入「可计算数」与抽象计算机器的概念，并由停机问题的不可解推出判定问题没有机械解法。这篇论文是「不可判定」这种处境最干净的样板：障碍是逻辑的而非资源的，本章正以图灵机器与停机问题作为该种处境的标准例证。
2. A. Church (1936).「An Unsolvable Problem of Elementary Number Theory」. American Journal of Mathematics, 58(2), 345–363. [②③] 丘奇用他发展的 lambda 演算证明初等数论中存在不可解问题，从而独立地否决了判定问题，发表上还早于图灵约七个月。它与图灵的结果互为印证，共同坐实了「原则上就不存在判定算法」并非个别现象，本章把两者并列为不可判定一族的开端。
3. K. Gödel (1931).「Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I」. Monatshefte für Mathematik und Physik, 38, 173–198. [②③] 哥德尔在此证明不完备性定理：任何足够强的一致形式系统中都存在既不能证明也不能否证的命题。它是不可判定谱系的源头，表明形式方法本身有原则上的极限，本章把它列为这种处境最早的一记警钟。

4. D. Hilbert & W. Ackermann (1928). «Grundzüge der theoretischen Logik» . Springer. [②③] 这本数理逻辑教科书第一次明确提出判定问题，即追问是否存在一个机械程序，能对任意数学命题判定真伪。正是这个问题催生了丘奇与图灵的否定证明，本章以它作为不可判定处境的出发点，读者可借此看清当年的乐观期待与随后的逻辑碰壁。
5. E. L. Post (1944).「Recursively Enumerable Sets of Positive Integers and Their Decision Problems」. Bulletin of the American Mathematical Society, 50(5), 284–316. [②③] 波斯特在此系统研究递归可枚举集及其判定问题，并提出后来催生不可解度理论的思路。它把「不可判定」从单个问题推进到对不可解性结构的分级研究，本章引它说明该种处境有自身的层次与谱系，而非铁板一块。
6. H. G. Rice (1953).「Classes of Recursively Enumerable Sets and Their Decision Problems」. Transactions of the American Mathematical Society, 74(2), 358–366. [②] 赖斯定理在此确立：程序所计算的任何非平凡语义性质都不可判定。它把图灵式的不可判定从个别问题推广为一条普遍铁律，本章引它说明，想机械地验证程序「做的对不对」这类问题，原则上就堵死了。
7. Y. V. Matiyasevich (1970).「Enumerable Sets Are Diophantine」. Soviet Mathematics. Doklady, 11(2), 354–357. [②③] 马蒂亚谢维奇在此补上最后一环，证明每个递归可枚举集都是丢番图集，由此完成希尔伯特第十问题不可解的证明，即 MRDP 定理。它说明连「丢番图方程有无整数解」这样具体的数学问题都没有判定算法，本章引它佐证不可判定并不限于自指或元数学，而是渗进了寻常数学。

难解（②③）

8. S. A. Cook (1971).「The Complexity of Theorem-Proving Procedures」. Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC), 151–158. [②③] 库克在此开创 NP 完全性概念，证明可满足性问题是 NP 中最

难的一类。它给「难解」这种处境下了精确定义：问题有解法，代价却随规模爆炸。本章以它划清第二种与第一种的界限，即「可判定却不可行」不同于「根本没有算法」。

9. L. A. Levin (1973).「Universal Sequential Search Problems」. Problems of Information Transmission, 9(3), 265–266. [②③] 列文在铁幕另一侧独立得到与库克相同的结果，给出通用搜索问题的完全性刻画，两者合称 Cook-Levin 定理。它说明 NP 完全性的发现是收敛而非偶然，本章引它强化「难解」这种处境的客观性：这是问题结构本身的性质，不因研究路径而异。
10. R. M. Karp (1972).「Reducibility Among Combinatorial Problems」. In R. E. Miller & J. W. Thatcher (Eds.),《Complexity of Computer Computations》(pp. 85–103). Plenum Press. [②③] 卡普用多项式归约证明了二十一个常见组合问题都是 NP 完全的，把库克的单个结果扩展成一张相互归约的网。它表明难解不是个别难题的怪癖，而是横跨调度、划分、覆盖等大量实际问题的普遍现象，本章引它说明这种处境在工程中无处不在。
11. J. Hartmanis & R. E. Stearns (1965).「On the Computational Complexity of Algorithms」. Transactions of the American Mathematical Society, 117, 285–306. [②] 这篇论文用图灵机的运行时间为算法定级，奠定了按资源消耗划分复杂度类的框架，「计算复杂度」一词也由此确立。它提供了度量「难解」所必需的标尺，本章引它说明第二种处境之所以能被精确谈论，前提是先有了刻画代价随规模如何增长的语言。
12. M. R. Garey & D. S. Johnson (1979).《Computers and Intractability: A Guide to the Theory of NP-Completeness》. W. H. Freeman. [②] 这本书系统整理了 NP 完全性理论与证明技巧，并附上一份广为引用的难解问题清单，长期被当作该领域的标准参考。对想由头了解「难解」这种处境的读者，它既是入门指南也是工具书，本章把它列为该主题最可靠的落脚处。
13. M. Sipser (2012).《Introduction to the Theory of Computa-

tion》(3rd ed.). Cengage Learning. [②] 这本广受采用的本科教材清晰讲解自动机、可计算性与复杂度，把图灵机、停机问题、P 与 NP 等概念串成一条连贯的线。它正好覆盖本章前两种处境的理论底子，是想从头打基础的读者最稳妥的起点，初版可上溯到 1997 年。

14. S. Arora & B. Barak (2009). 《Computational Complexity: A Modern Approach》. Cambridge University Press. [②] 这本研究生教材覆盖了从经典复杂度类到随机化、交互证明、近似与去随机化等现代主题，视野远超入门教科书。对想深入「难解」这种处境，尤其想理解人们如何用近似与随机绕开最坏情形的读者，它是更进一步的权威读物。

部分可观测（②④）

15. K. J. Åström (1965).「Optimal Control of Markov Processes with Incomplete State Information」. Journal of Mathematical Analysis and Applications, 10, 174–205. [②] 阿斯特罗姆在此研究状态信息不完整下的最优控制，提出用关于隐藏状态的概率分布即「信念状态」来概括所有可得信息。这是 POMDP 理论的源头之一，也正是本章为「部分可观测」开出的解药：不追求确定判决，而是维持并更新一个信念。

16. R. D. Smallwood & E. J. Sondik (1973).「The Optimal Control of Partially Observable Markov Processes over a Finite Horizon」. Operations Research, 21(5), 1071–1088. [②] 这篇论文给出有限时域 POMDP 的经典结构性结果，并据此设计出可计算最优策略的方法。它把阿斯特罗姆的信念状态思想推进为可操作的算法，本章引它说明「靠推断信念」并非空话，而有成形的求解技术支撑。

17. C. H. Papadimitriou & J. N. Tsitsiklis (1987).「The Complexity of Markov Decision Processes」. Mathematics of Operations Research, 12(3), 441–450. [②] 这篇论文系统刻画了马尔可夫决策过程各变体的计算复杂度，证明引入部分可观测会让求解显著变难。它把第三种与第二种处境扣在一起：看不见正确状态的处境，求解本身往往又是难解的，本章

正以此说明处境会彼此叠加。

18. L. P. Kaelbling, M. L. Littman & A. R. Cassandra (1998).「Planning and Acting in Partially Observable Stochastic Domains」. Artificial Intelligence, 101(1), 99–134. [②④] 这篇论文把 POMDP 整理为人工智能里的标准框架，统一了信念更新、规划与行动，并给出可实践的算法。它是「部分可观测」处境最常被引用的代表性文献，本章第 5 章对该种处境的展开正以此为底本，读者可由它系统了解推断加探查的整套做法。

预算受限（①④，含有限理性与 anytime 算法）

19. F. H. Knight (1921).《Risk, Uncertainty and Profit》. Houghton Mifflin. [①④] 奈特在此区分可用概率刻画的「风险」与无法量化的「不确定性」，后者即没有可靠概率可依的处境。这一区分是本书谈不可验证的思想起点之一，它提醒读者：有些处境的难处不在算得不够准，而在连下注所需的概率都不存在。
20. H. A. Simon (1955).「A Behavioral Model of Rational Choice」. The Quarterly Journal of Economics, 69(1), 99–118. [①④] 西蒙在此提出有限理性：真实主体的算力、时间与信息都有限，于是不去求全局最优，而是「满意即止」。这是「预算受限」处境的概念源头，本章借它点明，许多验证在理论上可行，落到一个有限的主体身上却必须打折，从而引出后面关于预算分配的思路。
21. M. Boddy & T. Dean (1989).「Solving Time-Dependent Planning Problems」. Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI). [②④] 这篇论文延续作者的随时算法工作，研究如何在计算时间本身受限时安排规划，让系统随时可中断并交出当前最优解。它与下一条同源，本章引这一系列工作来说明「预算受限」处境的应对核心是把有限的时间花在边际收益最高处。
22. T. Dean & M. Boddy (1988).「An Analysis of Time-Dependent Planning」. Proceedings of the 7th National

Conference on Artificial Intelligence (AAAI), 49–54. [②④] 这篇论文正式提出随时算法的概念：算法可在任意时刻被打断并给出当前最优解，质量随计算时间稳步提升。它是「预算受限」处境的代表性形式化，本章引它说明这种处境的独特之处在于会随资源增长而消退，因而应对的关键落在分配而非纯算力。

23. S. J. Russell & D. Subramanian (1995).「Provably Bounded-Optimal Agents」. Journal of Artificial Intelligence Research, 2, 575–609. [②④] 这篇论文把有限理性形式化为「有界最优」：不再要求智能体输出最优决策，而是要求它在给定的计算资源约束下做到所能做的最好。它给西蒙的直觉提供了精确定义，本章引它说明「预算受限」处境也能被严肃地理论化，而非只是无可奈何的妥协。

对抗（②①④，含决策论与对抗机器学习）

24. J. von Neumann & O. Morgenstern (1944).《Theory of Games and Economic Behavior》. Princeton University Press. [②①] 这本书奠定了博弈论，把多方在利益冲突下的互动当作可严格分析的对象，并系统化了零和博弈的极小极大定理。它是「对抗」处境的理论源头，本章正以它支撑该种的核心主张：面对会针对你优化的对手，要按最坏情形布防，而非在某个固定输入上求最优。

25. J. F. Nash (1950).「Equilibrium Points in N-Person Games」. Proceedings of the National Academy of Sciences, 36(1), 48–49. [②] 纳什在这篇短文中证明，任意有限的多人博弈都存在均衡点，即没有任何一方能靠单方面改变策略获益的稳定局面。纳什均衡把博弈分析从零和推广到一般情形，本章引它作为「对抗」处境的核心概念，帮助读者理解策略性互动如何收敛到可预期的稳定结构。

26. A. Wald (1945).「Statistical Decision Functions Which Minimize the Maximum Risk」. Annals of Mathematics, 46(2), 265–280. [②] 瓦尔德在此奠定统计决策理论，提出以极小极大准则选择决策，即在最坏情形下使风险最小。它把「按最坏

情形布防」从博弈搬进统计推断，本章引它说明对抗处境的解药是一种战略姿态：当对手会顺着你的判断调整时，求稳比求某一处的最优更要紧。

27. L. J. Savage (1954).《The Foundations of Statistics》. Wiley. [②①] 萨维奇在此为主观期望效用理论建立公理基础，论证一个理性主体的偏好可被表示为对主观概率求期望效用。它是不确定性下决策的标准框架，本章引它代表「用概率与效用为不确定性立账」的正统立场，也为下一条揭示该立场的边界做了铺垫。
28. D. Ellsberg (1961).「Risk, Ambiguity, and the Savage Axioms」. The Quarterly Journal of Economics, 75(4), 643–669. [①④] 埃尔斯伯格用一个简单赌局实验揭示：人们普遍回避概率本身不明的「模糊」选项，这种行为违反了萨维奇的公理。它从经验层面印证了奈特对风险与不确定性的区分，本章引它说明概率框架并非万能，有些不可验证的处境连概率都给不出。
29. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow & R. Fergus (2014).「Intriguing Properties of Neural Networks」. International Conference on Learning Representations (ICLR). arXiv:1312.6199. [②] 这篇论文首次系统揭示对抗样本：对图像施加人眼几乎察觉不到的微小扰动，就能让识别率极高的神经网络出错。它把「对抗」处境带进现代机器学习，本章引它说明，只要有人专门去找那个扰动，再准的模型也会被骗，这正是对抗不同于单纯可观测缺口的地方。
30. I. J. Goodfellow, J. Shlens & C. Szegedy (2015).「Explaining and Harnessing Adversarial Examples」. International Conference on Learning Representations (ICLR). arXiv:1412.6572. [②] 这篇论文把对抗样本归因于模型在高维空间的线性性，提出快速生成扰动的 FGSM 方法，并用对抗训练加以防御。它既解释了对抗样本为何普遍，又给出最早的应对手段，本章引它说明对抗处境的攻与防是一对此消彼长、需要持续博弈的过程。
31. A. Madry, A. Makelov, L. Schmidt, D. Tsipras & A. Vladu

(2018).「Towards Deep Learning Models Resistant to Adversarial Attacks」. International Conference on Learning Representations (ICLR). arXiv:1706.06083. [②④] 马德里等人把对抗鲁棒性写成一个极小极大优化问题：内层找最坏扰动，外层训练抵御它，并以投影梯度下降作为标准攻击。它用鲁棒优化的语言统一了攻与防，正好把本章对抗处境与最坏情形决策的主张落到机器学习里，是该方向影响深远的一篇。

32. B. Biggio & F. Roli (2018).「Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning」. Pattern Recognition, 84, 317–331. [②] 这篇综述回顾对抗机器学习十年的发展，指出该领域在深度学习走红前就已起步，并梳理了攻击模型、威胁建模与防御的整体脉络。对想从全局把握「对抗」处境的读者，它是权威的总览，本章引它作为该种处境最适合通读的落脚点。

第 3 章　可证伪，不可证实

> **论点**：人类最有纪律的求知方式（经验科学），建立在一个公开的承认上：理论永不可被证实（verification），只能不被证伪（falsification）。

上一章问，人类有没有一套成熟的、有纪律的办法与不可验证长期共处。有，那就是经验科学。而它最让人意外的地方在于，它的首要原则不是宣称自己能查明真理，恰恰是公开承认自己永远做不到。

一只黑天鹅

「所有天鹅都是白的。」你看过一千只白天鹅，看过一百万只，这个全称判断（universal statement）依然没有被证实，因为下一只可能是黑的。但你只要看见一只黑天鹅，它就被彻底推翻了。这并不是个虚构的例子：在欧洲，「天鹅皆白」长期被当作确定无疑的常识，直到 1697 年荷兰探险队在西澳大利亚第一次见到黑天鹅，那份「确定」一夜之间作废。

这道不对称是整章的支点。证实一个全称命题需要检遍它所断言的全部情形，而那往往是无穷的、开放的、属于未来的，根本做不到。但证伪它，却只需要一个反例。用逻辑写出来：$\forall x\, P(x)$ 无法被有限的观察确立，但一个 $\exists x\, \neg P(x)$ 就足以将其击碎。科学的全部纪律，就建立在认清并利用这道不对称之上。

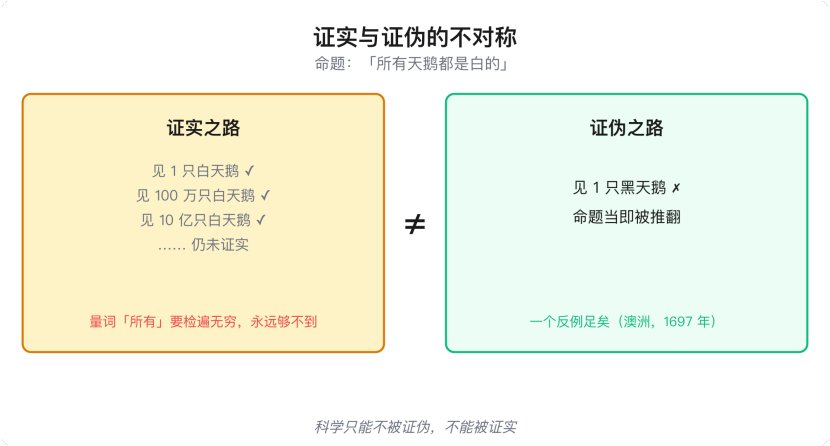

图 3: 证实与证伪的不对称

休谟那道过不去的坎

这件事的根，休谟在 1739 年[4] 就刨到了。我们凭什么相信，过去一直成立的规律未来还会成立？没有逻辑上的凭据。从「太阳过去每天升起」推不出「太阳明天必升」，因为这个推论本身就预设了「过去的模式会延续到未来」，而这恰是待证的东西。归纳（induction）没有逻辑保证。休谟的结论冷静而彻底，我们依赖的不是证明，是习惯。

这正是第 1 章那「未来」之裂口的哲学底座。任何关于世界普遍规律的知识都建立在有限的过去之上，因而都无法被事前证实。科学若想成为知识，就不能把「证实」当作目标，那个目标根本够不着。

波普尔：把够不着的换成够得着的

波普尔在 1934 年[1]（德文原版）给出了出路：既然证实不可得，就别要它，改用证伪。一个理论是不是科学的，不看它能被多少证据支持（支持总能找到），而看它有没有把脖子伸出来、做出可能被推翻的、有风险的预测。占星术、对一切都自圆其说的学说不可证伪，因而不科学；广义相对论预言星光会被太阳引力偏折一个具体的角度，1919 年的日食观测原本完全可能测出别的数值、从而推翻这个

预言，正因为理论敢于冒被推翻的风险，它才是好科学。

于是科学成了一台为「与不可验证共处」专门优化的机器。它从不宣称证明了什么，只说这个理论至今尚未被证伪，所以我们暂且用它。这是一种姿态，一种把测不出的「真」换成测得出的「尚未被推翻」的姿态。眼熟吗？这正是第 7 章那个数学家的代理替换在认识论尺度上的样子。

一句必要的限定

必须就地说清：波普尔式的证伪主义（falsificationism）在科学哲学里远非定论，本书把它当作一个清晰的入口，而不是终点。

它最有力的反对来自迪昂与奎因的整体论（holism，又称 Quine-Duhem 论题）。迪昂 1906 年[10]、奎因 1951 年[9] 指出，你从来无法孤立地检验一个假说。任何预测都依赖一大堆辅助假定（仪器没坏、背景条件成立、近似合理），实验一旦失败，你永远可以把矛头引向某个辅助假定、而保住核心假说。于是「一个反例就干净利落地推翻理论」这幅图景并不像它看上去那么干净。库恩 1962 年[5] 进一步说，常规科学时期的科学家根本不急着证伪，反常会被搁置，直到范式（paradigm）危机才发生革命式的更替；拉卡托斯 1970 年[6] 用「研究纲领」（research programme）的进步与退化来取代非黑即白的证伪；费耶阿本德[7] 干脆反对一切统一方法。另一条路径是贝叶斯确证论（Bayesian confirmation theory）[24]，它不要二值的判决，而把证据看作对信念概率的调整，

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)},$$

这又预告了后面「标定」那一招。梅奥的「严苛检验」（severe testing）[14] 是证伪主义一个精致的继承者，而斯坦福[19] 则提醒我们还有大量「未被设想的替代方案」（unconceived alternatives）在视野之外。

把这些争论摆出来，不是要拆波普尔的台，而是因为这本书自己就该这样行事：陈述一个有力的框架，同时把它的边界标出来。这种姿态正是全书要演练的那一招。

科学早已发现了那几招

现在是这一章给全书的真正馈赠。如果你带着「八招」的眼光去看科学的日常建制，会发现它早就把其中好几招摸索出来，只是用着别的名字。

同行评审是冗余与共识，不信任单个判断者，而用多个互相独立的评审、取其一致。重复实验也是冗余，一个结果要等别人在别处独立重现才被当真。预注册（preregistration）是留痕，在看到数据之前就把假说和分析方案登记下来、事后无法移动靶子、不能把噪声讲成信号。置信区间与误差统计是证书与界，不声称命题为真，只在一个明确的置信水平上给出一个有界的保证。双盲与随机化是对付第五种处境（对抗）的防御，而这里的对手往往是研究者自己的偏见与主观期待。显著性阈值是一种粗糙的标定。

换句话说，人类最严肃的求知事业本身就是本书那个收敛命题的一个活样本。这是全书第一个、也是分量很重的暗示，尽管科学面对的不可验证（关于普遍规律、关于未来）有其特定来源，它被逼出来的应对和软件、数学、组织里的应对押着同一个韵。

当机器失灵：复制危机

反过来看更清楚。当这些招数被削弱，科学的自我纠错就会失效，这就是复制危机（replication crisis）。约阿尼迪斯 2005 年[28] 那篇《为什么大多数已发表的研究结论是假的》、开放科学合作组织 2015 年[29] 对一百项心理学研究的大规模重复（原研究里 97% 当初都报告了显著结果，重做之后只有约三十六项还成立，连一半都不到）揭开的正是这一幕，当预注册缺位（靶子可以事后挪动）、样本量不足、发表偏倚只放行漂亮结果、又少有人去做吃力不讨好的重复时，那台机器就空转了。

这场危机的诊断与修补恰恰是用那几招的语言进行的，恢复预注册（把留痕装回去）、鼓励并奖励重复（把冗余装回去）、登记报告、提高检验的严苛度。问题与药方都落在同一套词汇上。这一点在第 10 章谈借来的判断、第 12 章谈留痕与审计时还会回来。

这一章通向哪里

科学证明了一件事，人可以有纪律地在没有验证的世界里求知，而且但凡做得好，靠的就是那几招。这是全书的一个概念验证。

但这里也潜伏着一个陷阱。正因为这五种处境都以「我没法检验它」的同一副表情出现，又正因为应对它们的招数如此相似，一个极具诱惑力的念头会冒出来：何不干脆宣布，不可验证就是一个问题、配一个统一的解法？这个念头关于「问题」的部分是错的，关于「应对」的部分却歪打正着。下一章专门处理这个诱惑。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. K. Popper (1959).《The Logic of Scientific Discovery》. Hutchinson. [②③] 波普尔在此系统提出证伪主义：科学理论无法被经验证实，只能被否证，可证伪性因而成为科学与非科学的分界。德文原版《Logik der Forschung》由维也纳 Springer 出版，版权页标 1935 而实际 1934 年底面世（故常记作 1934），这部英文版由作者亲自大幅修订增补。本章「波普尔：把够不着的换成够得着的」一节直接建基于此，读者应着重领会以「尚未被证伪」替换「证实」的认识论姿态。

2. K. Popper (1963).《Conjectures and Refutations: The Growth of Scientific Knowledge》. Routledge and Kegan Paul. [③] 这部论文集把证伪主义铺展为一整套知识增长观：知识经由大胆猜想与无情反驳而前进，科学的成长不是积累确证，而是不断淘汰错误。比起前作的逻辑骨架，它更直观地展示了「试错」如何驱动科学进展，是理解本章「科学如何进展」这一落足点的好读物。

3. K. Popper (1972).《Objective Knowledge: An Evolutionary

Approach》. Clarendon Press. [③] 波普尔在此把知识增长类比为演化式的试错过程，并提出「第三世界」即客观知识本身的领域，独立于个人的主观心智而存在。它把证伪主义推向一种关于客观知识如何无主体地积累的本体论图景，可供有意深究「科学如何进展」的读者延伸阅读。

4. D. Hume (1739). 《A Treatise of Human Nature》. John Noon. [②] 休谟在此提出归纳问题这一源头性难题：从过去的规律推不出未来的规律，因为这一推论本身预设了「自然齐一」，而那正是待证之事；我们对因果与规律的信念，归根到底来自习惯而非证明。第一、二卷 1739 年由 John Noon 出版，第三卷《Of Morals》1740 年由 Thomas Longman 出版，通常以 1739 标记初版。本章「休谟那道过不去的坎」一节即奠基于此，是理解科学为何无法以「证实」为目标的哲学底座。

5. T. Kuhn (1962). 《The Structure of Scientific Revolutions》. University of Chicago Press. [①③] 库恩借大量科学史案例论证：科学并非匀速逼近真理，而是在「常规科学」时期于一个共享范式内解谜，反常累积到危机后才发生范式更替式的科学革命，且新旧范式之间存在不可通约。它是对波普尔图景的重要修正，说明科学家常常并不急于证伪反常，本章「一句必要的限定」即引此标出证伪主义的边界。

6. I. Lakatos (1970).「Falsification and the Methodology of Scientific Research Programmes」. 收于 I. Lakatos, A. Musgrave 编《Criticism and the Growth of Knowledge》, pp. 91-196. Cambridge University Press. [①③] 拉卡托斯以「研究纲领」调和波普尔与库恩：每个纲领有一个受保护的硬核与一圈可调整的辅助假定，评判标准不是单个反例，而是纲领整体随时间是「进步」（持续做出并兑现新预测）还是「退化」（只忙于事后打补丁）。它把非黑即白的证伪换成对纲领进退的历史判断，是本章界定证伪主义边界时的关键参照。

7. P. Feyerabend (1975). 《Against Method: Outline of an Anarchistic Theory of Knowledge》. New Left Books. [①③④] 费耶阿本德以伽利略等科学史案例力证：并不存在一套普遍

有效的科学方法，重大进展往往恰恰来自违反既有规则，故其著名口号是「怎么都行」。它是对统一方法论最激进的反对，本章引它来标明：连「证伪」这样温和的方法论主张，也有人从根本上拒斥。

8. C. G. Hempel (1965).《Aspects of Scientific Explanation and Other Essays in the Philosophy of Science》. Free Press. [②] 亨佩尔在这部论文集里集大成地阐发科学解释的覆盖律模型，既包括演绎律则式解释，也包括归纳统计式解释，并讨论了确证的逻辑及其悖论。它代表了逻辑经验主义对「理论上被研究过的东西」的系统刻画，为本章关于何为可被检验、可被解释提供了经典背景。

9. W. V. O. Quine (1951).「Two Dogmas of Empiricism」.《The Philosophical Review》, 60(1), 20-43. [②] 奎因攻击逻辑经验主义的两条教条：分析与综合的截然二分，以及还原论；并提出认识论整体论，主张我们的信念作为一张整体之网共同面对经验，没有哪个陈述能被孤立地证实或否证。结合迪昂的检验整体论（合称 Quine-Duhem 论题），它直接冲击「一个反例干净利落地推翻一个假说」的图景，是本章界定证伪主义边界的核心文献。

10. P. Duhem (1906).《La théorie physique: son objet, sa structure》. Chevalier & Rivière. [②] 迪昂在此提出检验整体论：物理学中的实验从不检验孤立假说，而是检验「假说连同一整套辅助假定与背景理论」，因此一次失败的预测无法判定究竟错在何处。此即后来与奎因合称的整体论之源头，本章用以说明反例的指向并不像表面那样确定。此处以原始法文版 1906 年为准，第二版 1914 年由 Marcel Rivière 出版，P. P. Wiener 英译《The Aim and Structure of Physical Theory》由 Princeton University Press 1954 年刊行。

11. M. Polanyi (1958).《Personal Knowledge: Towards a Post-Critical Philosophy》. University of Chicago Press. [①④] 波兰尼提出「默会知识」：我们知道的远多于我们能言说的，科学探究中始终有一层无法形式化、只能在实践与师承中习

得的个人判断与技艺。它提醒人们，再严格的方法论也无法消去科学家亲身的、不可言传的判断，呼应本章「历史上科学家的判断」与「如何在无法验证的世界里生活」两个落足点。

12. B. C. van Fraassen (1980).《The Scientific Image》. Clarendon Press. [②④] 范弗拉森提出「建构经验论」：科学的目标不是宣称理论为真，而只是「经验适当」，即正确地拯救可观察现象；接受一个理论意味着相信它经验适当，而非相信其不可观察部分确实存在。它把「不可证实」转化为一种成熟的科学态度，与本章把「真」替换为「尚未被推翻」的姿态彼此呼应。

13. I. Hacking (1983).《Representing and Intervening: Introductory Topics in the Philosophy of Natural Science》. Cambridge University Press. [②③] 哈金把哲学注意力从「表征」转向「干预」，主张实在论的最佳辩护不在理论而在实验：当我们能稳定地操纵电子去探测别的东西时，电子就是真实的（「能喷射，便是真」）。它为科学实在论开辟了以实验实践为基础的新进路，也提醒读者科学进展同样依赖动手干预而非只靠理论检验。

14. D. G. Mayo (1996).《Error and the Growth of Experimental Knowledge》. University of Chicago Press. [③] 梅奥提出「误差统计」哲学：一个假说只有当它通过了「若为假则极可能不通过」的严苛检验时，我们才有理由接受它。这把波普尔的证伪精神落实为可操作的统计检验程序，是证伪主义一个精致的继承者，本章「严苛检验」一说即源于此（属 Science and Its Conceptual Foundations 丛书）。

15. D. G. Mayo (2018).《Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars》. Cambridge University Press. [③④] 梅奥在此以「严苛性」为统一原则重构统计推断，试图越过频率派与贝叶斯派长期的「统计之战」，并据此回应复制危机中对显著性检验的批评。它把第 14 项的纲领发展为面向当代统计实践的方法论，对理解如何在不可验证的世界里负责任地使用统计证据尤为切题。

16. L. Laudan (1981).「A Confutation of Convergent Realism」. 《Philosophy of Science》, 48(1), 19-49. [①②] 劳丹列举科学史上一批曾经成功（能预测、能解释）却最终被抛弃的理论，如燃素说、以太说，论证「成功蕴含为真」的推断站不住脚，对收敛实在论构成有力反驳，常被称为「悲观元归纳」。它说明就连经验上很成功的理论也未必接近真理，强化了本章关于科学不以「证实真理」为目标的论点。

17. L. Laudan (1977). 《Progress and Its Problems: Towards a Theory of Scientific Growth》. University of California Press. [③] 劳丹主张以「问题求解能力」而非逼近真理来衡量科学进步：一个研究传统是否进步，取决于它解决的经验问题与概念问题之净增量。它给出了一种绕开真理概念的进步观，为本章「科学如何进展」提供了一个不依赖证实的替代框架。

18. P. Kitcher (1993). 《The Advancement of Science: Science without Legend, Objectivity without Illusions》. Oxford University Press. [③] 基切尔在抛弃科学全知全能的「传说」之后，又拒绝相对主义，转而从科学的社会与认知实践出发重建一种温和而可辩护的客观性与进步观。它示范了如何在承认科学受历史与社会影响的同时，仍守住进步与客观这两个概念，与本章既肯定科学又把边界标清的立场一脉相承。

19. P. K. Stanford (2006). 《Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives》. Oxford University Press. [①②③④] 斯坦福提出「未被设想的替代方案」问题：科学史一再表明，过去的科学家总有一些后来才出现、当时根本想不到的理论选项，故我们没有理由相信今天已穷尽了所有可行解释。他以遗传学等史案归纳出这一「新归纳」，直接呼应本书「无法验证的世界」之框架，提醒读者视野之外总有未及设想的可能。

20. N. Goodman (1955).《Fact, Fiction, and Forecast》. Harvard University Press. [②] 古德曼提出「归纳的新谜题」：用「绿」与人造谓词「绿蓝」（grue，意为在某时间点前观察为绿、其后为蓝）同样能拟合迄今全部观察，却导出相反预测，可见归

纳无法仅凭证据决定，还须依赖哪些谓词「可投射」。它说明归纳的困难不止于休谟式的辩护问题，更在于规律本身的不确定，深化了本章对归纳何以不可靠的理解。初版年份通行作 1955（HUP 一处简介称 1954，存在轻微歧义，此处从广引的 1955）。

21. C. G. Hempel, P. Oppenheim (1948).「Studies in the Logic of Explanation」.《Philosophy of Science》, 15(2), 135-175. [②] 亨佩尔与奥本海姆在此奠定演绎律则（D-N）解释模型：一个现象得到科学解释，意味着它能从普遍定律加初始条件中逻辑地推演出来。它是二十世纪科学解释理论的起点，界定了「能被解释」在逻辑上意味着什么，为本章关于科学如何刻画规律提供了底层框架。

22. R. Carnap (1936-1937).「Testability and Meaning」.《Philosophy of Science》, 3(4), 419-471; 4(1), 1-40. [②] 卡尔纳普在此放松严格的可证实原则，改用更宽的「可检验性」与「可确认性」来界定有意义的经验陈述，并以倾向性谓词等技术处理理论词项与观察的联系。它记录了逻辑经验主义从「可证实」向「可检验」的关键退却，正与本章「证实够不着、改用够得着的」这一主线相呼应。原文分两期刊出，第 3 卷第 4 期（1936）与第 4 卷第 1 期（1937）。

23. W. C. Salmon (1984).《Scientific Explanation and the Causal Structure of the World》. Princeton University Press. [②] 萨蒙主张科学解释的核心不是逻辑推演而是揭示因果机制：解释一个现象，是把它嵌入世界的因果过程与因果相互作用之网。它是对覆盖律模型的重要修正，把「能解释」的标准从可推演转向可追溯的因果结构，为本章理解科学如何刻画世界补上因果这一维度。

24. C. Howson, P. Urbach (1989).《Scientific Reasoning: The Bayesian Approach》. Open Court. [②③] 豪森与厄巴赫系统主张贝叶斯主义的科学推理观：不作非真即假的二值判决，而把证据看作按贝叶斯定理对信念概率的连续调整，并以此回应归纳与确证的诸多难题。它是本章正文提到的贝叶斯确

证论的代表性论著，与证伪、严苛检验形成对照，也预告了全书「标定」一招。

25. E. Sober (2008).《Evidence and Evolution: The Logic Behind the Science》. Cambridge University Press. [②] 索伯以似然论与统计推断的工具细致分析「证据支持什么」，并讨论何种假说才真正可检验，其中包含对智能设计为何不可检验的剖析。它把抽象的可检验性问题落到具体的科学推断实践（尤以进化论为例），示范了如何严格判断一个主张是否经得起证据的检验。

26. P. Godfrey-Smith (2003).《Theory and Reality: An Introduction to the Philosophy of Science》. University of Chicago Press. [②③] 戈弗雷-史密斯这部广受好评的科学哲学导论，清晰梳理了从逻辑经验主义、证伪主义、库恩范式到贝叶斯主义与科学实在论之争的整条脉络。它适合作为本章诸多论题的导论性锚点，读者若想在阅读专著之前先建立全局地图，可由此入手。

27. N. Cartwright (1983).《How the Laws of Physics Lie》. Clarendon Press. [②③] 卡特赖特论证物理学的基本定律之所以普适而优美，恰恰因为它们并不如实描述真实世界，而是描述高度理想化的模型；越基本的定律解释力越强，描述上反而越「说谎」。她转而看重更贴近现象的具体定律与因果能力，提醒读者科学定律之「真」远比通常设想的复杂，深化了本章对理论与世界关系的省思。

28. J. P. A. Ioannidis (2005).「Why Most Published Research Findings Are False」.《PLoS Medicine》, 2(8), e124. [③④] 约阿尼迪斯用简明的统计建模论证：在效应量小、研究设计灵活、发表偏倚盛行的领域，一个已发表「阳性」结论为假的概率往往高于为真，假阳性可以系统性地多于真阳性。它是复制危机的奠基文献，正是本章「当机器失灵」一节的核心证据，说明科学的自我纠错一旦被削弱会如何空转。

29. Open Science Collaboration (2015).「Estimating the Repro-

ducibility of Psychological Science」. 《Science》 , 349(6251), aac4716. [③④] 开放科学合作组织协同上百名研究者，对一百项已发表的心理学研究做了系统性的直接重复，结果能成功重现原效应的不到半数，且重现出的效应普遍弱于原报告。它把复制危机从论证变为大规模实证，是本章「复制危机」一节的实证核心，也凸显了重复与预注册这些招数为何不可或缺。

第 4 章　摊平的诱惑

> **论点**：因为这五种处境都表现为「我没法检验」，会有强烈的冲动把不可验证当成一个问题、配一个解法。这关于问题是错的，却指向了关于应对的那个对的东西。

第 2 章费了整章力气把不可验证掰成五种处境。第 3 章又看到，应对这些处境的招数彼此相似。把这两件事放在一起，一个念头几乎必然冒出来，而且极具诱惑：既然它们都带上「我没法检验」这同一个特征，又都用差不多的办法对付，那何不干脆宣布，不可验证就是一个问题，配一个统一的解法？

这一章要做两件事。先拆穿这个念头错在哪里，再说清它歪打正着地撞上了什么，并由此给全书立下一条举证责任（burden of proof）。

摊平错在哪里

摊平（flatten），就是把五种结构不同的处境，碾成一张脸。它的代价，第 2 章其实已经预演过：

把不可判定（undecidable）当成预算问题，以为「多堆点算力就行」。错。停机问题（halting problem）不是算得慢，是根本没有那个算法，再快的机器也变不出一个不存在的程序。

把对抗缺口当成单纯的可观测缺口，以为「我只是还没看清它」。错得更危险。对面那个系统会顺着你的看法调整自己，你看得越清，它躲得越巧，这是一盘棋，不是一次测量。垃圾邮件过滤就是活教材：你按当下垃圾邮件的特征训好一个分类器，发信方立刻改写措

辞、换域名、插乱码绕过去，谁错把它当成静态的识别题、用一次性的模型去解，谁就永远慢半拍。

把难解（intractable）当成不可判定，于是过早放弃一个其实能近似、能处理平均情形的问题；或反过来，把不可判定当成难解，在一堵逻辑的墙上没完没了地砸算力。每一种错认，都会让你掏出错误的工具，付出真实的代价。诊断错了病灶，再对症的药也是毒。

所以，关于「问题」，摊平是错的。五种处境必须分开对待，这是第 2 章用一整章换来的结论，不能在这里轻易交还。

摊平歪打正着的那一面

可是，这个念头里有一粒真东西。它关于「问题」错了，关于「应对」却歪打正着。

本书的命题，恰恰是把这粒真东西从那团错误里择出来，并且说得极其小心：

> 不可验证的来源天差地别，但有限主体（bounded agent）被逼出来的应对，反复收敛到同一小套。

请注意这个表述的克制。它不说「这些问题是同一个问题」，那是摊平，是错的，任何一个领域的专家都会把书摔了。它说的是另一件更强、也更站得住的事：问题各不相同，应对却押韵。全书要交付的，是那张「同一招在多种行话下的对照表」，外加一个解释，为什么偏偏收敛到这几招。

一条必须自缚的举证责任

把话说到这里，最大的风险也就浮出来了。人类的思维天生爱类比，莱考夫与约翰逊[14] 让我们看到连日常语言都是隐喻搭起来的，根特纳[6]、霍利约克[9]、巴尔塔[7] 等人则研究类比何时为真、何时只是好看。可正因为类比这么顺手，它也最容易骗人。漂亮的跨领域类比，常常什么都证明不了。最现成的反面教材就是霍夫施塔特的《哥德尔、埃舍尔、巴赫》[1] 那一脉：跨域的呼应写得目眩神迷，却常被批评

「终究只是类比」，经不起追问。更扎实的教训来自所谓幂律（power law）热潮。许多系统被宣称服从同一条幂律、共享同一种深层机制，听上去无比统一，可一旦用克劳塞特、沙利齐与纽曼[25]2009 年那样严格的统计去检验，大量「幂律」根本立不住。他们重新审视二十余个被广泛宣称为幂律的真实数据集，能稳稳通过检验的寥寥无几，多数其实被对数正态（log-normal）等别的分布拟合得更好。斯坦普夫与波特[26]2012 年那篇《关于幂律的若干真相》说得很直白：看起来像，不等于是。

所以本书必须给自己套上一条铁律：任何宣称的收敛，都必须被证明不只是类比。

什么才算「不只是类比」？科学哲学里有现成的标尺。一个跨域映射要算实质的，得是结构保持（structure-preserving）的，不只是表面相似，而是机制对应、失效方式对应、权衡对应。根特纳[6] 的结构映射（structure-mapping）与巴尔塔[7] 对类比论证（analogical argument）的评估，给的正是这套标准。还有一条更硬的判据是稳健性（robustness）：一个结论若能从多条互相独立的路径反复导出，就更可信。莱文斯[17]、维姆萨特[18]、韦斯伯格[19] 发展的稳健性分析（robustness analysis）正是这一思想的来源。安德森[5] 那句「多即不同」（More Is Different）、福多[27] 关于「特殊科学」（special sciences）的论证、卡特赖特[28] 的《斑驳的世界》都表明，跨层次的真实模式确实存在，但它们是挣来的，不是宣布出来的。博克斯[33] 的名言悬在头顶：所有模型都是错的，有些是有用的。本书要争取的，是「有用且其用处经得起检验」，而不是「优雅得让人忘了检验」。

落到操作上，这条铁律意味着：第三部里每抽出一招、每做一次跨域并置，都要逼问一遍，这个迁移是实质的（同机制、同败法、同权衡），还是只是个漂亮的比方？扛得过这个拷问，对照表才成立；扛不过，它就只是一本好看的散文。第 13 章会尝试把这种收敛挂到一个共同的底层结构上（风险与信息的分解），但那是一张待兑现的期票，要去检验的，不是可以预先假定的。第 14 章则会正面清算：这究竟是定律，还是一个很强的经验模式。

这一章通向哪里，以及第二部的次序

既然不能靠预先宣布收敛、再挑几个例子来凑，那唯一靠得住的检验方式，就是走进真实的领域，看有能力的主体到底做了什么，让招数从案例里自己长出来，而不是先立招式再去套。

这也解释了本书为什么把现场（第二部）放在招式库（第三部）之前。先抽象再举例，会显得武断，也浪费了案例本应有的说服力。所以第二部不急着命名，它让招数嵌在各自的领域里、彼此缠绕地出现，表面看起来乱一些；第三部才把每一招单独拎出来、洗净、命名。归纳在前，命名在后。

四个现场已经备好：一个揣摩用户心思的设计者，一个被放出去自行其是的智能体，一个对着黎曼假设撞墙的数学家，一个看不见自己的庞然组织。它们面对的不可验证，来自五种处境中不同的几种。我们去看看，当神谕始终不来，他们各自伸手够向了什么。

参考文献

> 落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. D. Hofstadter (1979).《Gödel, Escher, Bach: An Eternal Golden Braid》. Basic Books. [②] 霍夫施塔特借哥德尔的不完全性、埃舍尔的视觉悖论与巴赫的赋格，编织出一套关于自指、递归与意识如何从形式系统中浮现的宏大类比。它是跨域类比写作的标杆，也是本章的反面教材：呼应写得目眩神迷，却屡被批评「终究只是类比」，正好提醒读者，漂亮的跨域共鸣本身并不构成论证。

2. H. A. Simon (1969).《The Sciences of the Artificial》. MIT Press. [②③] 西蒙提出研究「人造物」的科学，主张设计是一门可以系统化的学问，并以有限理性、满意而非最优来刻画真实主体的决策。本书对本章重要，在于它把「有限主体在受约

束下如何应对」当成正经的科学对象，正是全书命题里那个「被逼出来的应对」的思想源头。

3. W. C. Wimsatt (2007).《Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality》. Harvard University Press. [②④] 维姆萨特主张哲学应当为「有限存在者」重写：认知资源有限的主体只能用启发式、近似与稳健性来逼近实在，错误是不可避免却可管理的。书名几乎就是本书副线的注脚，读者可重点看它如何把稳健性当作有限主体辨别真伪模式的核心工具。

4. H. A. Simon (1962).「The Architecture of Complexity」. Proceedings of the American Philosophical Society, 106(6), 467-482. [②③] 西蒙论证复杂系统多为「近可分解」的层级结构，子系统内部联系紧、子系统之间联系松，这种架构既便于演化也便于理解。它为「跨层次存在真实模式」提供了经典的结构性论据，值得与安德森、福多对照着读。

5. P. W. Anderson (1972).「More Is Different」. Science, 177(4047), 393-396. [②] 安德森反对还原论的傲慢，指出每一层级都会涌现出新的规律，下一层的定律无法被上一层简单推导出来。本章引用的「多即不同」一语即出于此，它说明跨层次的真实模式确实存在，但要靠各自的科学挣来，而非从基础物理宣布得到。

6. D. Gentner (1983).「Structure-Mapping: A Theoretical Framework for Analogy」. Cognitive Science, 7(2), 155-170. [②] 根特纳提出结构映射理论：好的类比迁移的是关系结构而非表面属性，系统性的关系网络比孤立的相似点更有价值。这正是本章「实质类比须结构保持」判据的理论来源，读者可由此理解什么叫「机制对应而非表面相似」。

7. P. Bartha (2010).《By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments》. Oxford University Press. [②] 巴尔塔为类比论证建立一套规范性的评估框架，追问一个类比何时能真正承载推理的重量，关键在于源域

与目标域之间是否存在相关的因果或结构联系。它给本章那条「不只是类比」的铁律提供了可操作的判别标准。

8. M. B. Hesse (1966). 《Models and Analogies in Science》. University of Notre Dame Press. （扩充版，初版 1963。）[②③] 赫西分析模型与类比在科学中的认知功能，区分正面、负面与中性类比，指出类比的「中性部分」正是科学预测与发现的生长点。它是类比方法论的早期奠基之作，为后来的结构映射与类比评估铺路。

9. K. J. Holyoak & P. Thagard (1995).《Mental Leaps: Analogy in Creative Thought》. MIT Press. [②] 霍利约克与萨加德提出类比的多重约束理论，认为人脑在结构、语义与目的三类约束的相互制衡下完成类比映射。本书把类比放进认知与创造的现实过程中考察，有助读者理解类比为何既强大又易出错。

10. D. Gentner, K. J. Holyoak & B. N. Kokinov (Eds.) (2001). 《The Analogical Mind: Perspectives from Cognitive Science》. MIT Press. [②] 这本文集汇集认知科学各路对类比的研究，从计算模型到发展心理到神经机制，系统呈现「类比何时为真、何时只是好看」的研究版图。它是本章关于类比研究脉络的总览性入口。

11. D. Hofstadter & E. Sander (2013). 《Surfaces and Essences: Analogy as the Fuel and Fire of Thinking》. Basic Books. [②] 霍夫施塔特与桑德主张类比是思维的核心引擎，连最基本的范畴化与概念形成都是不断进行的类比。本书把类比的地位推到极致，恰好与本章的警惕形成张力：类比无处不在，正因此更需要一套判据来分辨哪些迁移是实质的。

12. S. Vosniadou & A. Ortony (Eds.) (1989). 《Similarity and Analogical Reasoning》. Cambridge University Press. [②] 这本文集集中讨论相似性与类比推理的关系，追问「相似」究竟意味着什么、它如何驱动推理与学习。它为本章背后的问题，即如何把「看起来像」与「确实是」区分开，提供了概念

上的准备。

13. K. Dunbar (1995).「How Scientists Really Reason: Scientific Reasoning in Real-World Laboratories」. 收入 R. J. Sternberg & J. E. Davidson (Eds.),《The Nature of Insight》, 365-395. MIT Press. [①②③] 邓巴实地观察分子生物学实验室，发现科学家在真实工作中大量使用类比，且近距离的、领域内的类比往往比远距离的更富成效。它以现场证据支撑本章「让招数从案例里自己长出来」的方法选择，说明真实推理与教科书叙述并不相同。

14. G. Lakoff & M. Johnson (1980).《Metaphors We Live By》. University of Chicago Press. [②] 莱考夫与约翰逊论证隐喻不只是修辞，而是人类概念系统的结构本身，连「时间是金钱」这类日常表达都暗藏成体系的隐喻。本章引它来说明类比与隐喻深植于人类思维，正因如此才更需对其可靠性保持戒心。

15. A. Tversky & D. Kahneman (1974).「Judgment under Uncertainty: Heuristics and Biases」. Science, 185(4157), 1124-1131. [②④] 特沃斯基与卡尼曼揭示人在不确定下依赖代表性、可得性、锚定等启发式做判断，这些捷径虽高效却会系统性地导致偏差。它提醒读者，有限主体的应对往往不是最优解而是凑合用的招数，也解释了为何漂亮的类比格外容易骗过我们。

16. R. W. Batterman (2001).《The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence》. Oxford University Press. （精装初版 2001 年 11 月；部分编目著录作 2002。）[②③] 巴特曼研究渐近推理在解释中的作用，论证许多物理现象的解释恰恰藏在取极限时浮现的「细节」里，普适性正源于此。它为「跨系统的共同结构如何可能」提供了一个精细的哲学案例，呼应本章对收敛之机制的追问。

17. R. Levins (1966).「The Strategy of Model Building in Population Biology」. American Scientist, 54(4), 421-431. [②

③] 莱文斯指出建模无法同时兼顾普适、精确与现实三者，建模者必须取舍，并提出当多个不同假设的模型给出一致结论时，该结论更可信。这正是稳健性分析的源头，本章那条「多路径反复导出更可信」的硬判据由此而来。

18. W. C. Wimsatt (1981).「Robustness, Reliability, and Overdetermination」. 收入 M. B. Brewer & B. E. Collins (Eds.),《Scientific Inquiry and the Social Sciences》, 124-163. Jossey-Bass. [②③] 维姆萨特系统阐发稳健性概念：能被多种相互独立的手段、模型或视角共同探测到的东西，更可能是真实的而非假象。本文是本章稳健性判据的核心文献，读者可由此理解为何独立路径的汇合能压制错误。

19. M. Weisberg (2006).「Robustness Analysis」. Philosophy of Science, 73(5), 730-742. [②③] 韦斯伯格重新厘清稳健性分析的逻辑，区分稳健定理与对其经验充分性的检验，澄清它何时能、何时不能为结论提供支持。它让本章的稳健性判据更精确，提醒读者稳健并不自动等于为真，仍需经验把关。

20. S. H. Orzack & E. Sober (1993).「A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966)」. The Quarterly Review of Biology, 68(4), 533-546. [②③] 奥扎克与索伯批判性地审视莱文斯的建模策略，质疑仅凭多模型一致就推断结论为真在逻辑上是否成立，除非这些模型各自已得到独立支持。它是稳健性论证的重要反方，帮助本章把铁律守得更紧，避免把一致误当成证明。

21. N. Goldenfeld & L. P. Kadanoff (1999).「Simple Lessons from Complexity」. Science, 284(5411), 87-89. [②③] 戈登菲尔德与卡达诺夫提醒，研究复杂系统要在恰当的尺度上选用恰当的模型，普适性虽诱人却不应掩盖具体机制，关键在于「在合适的层面上做对的简化」。它为本章如何谨慎对待跨系统普适规律提供了来自物理学内部的清醒声音。

22. L. P. Kadanoff (1966).「Scaling Laws for Ising Models near Tc」. Physics Physique Fizika, 2(6), 263-272. [②] 卡达诺夫

提出区块自旋的标度图像，说明临界点附近系统在不同尺度上自相似，为后来的重整化群与普适类理论奠基。它是「不同系统共享同一临界行为」这一真实普适性的经典范例，与后文那些立不住的「幂律」恰成对照。

23. P. Bak, C. Tang & K. Wiesenfeld (1987).「Self-Organized Criticality: An Explanation of the 1/f Noise」. Physical Review Letters, 59(4), 381-384. [②] 巴克、汤与维森费尔德提出自组织临界性，以沙堆模型说明某些系统会自发演化到临界态，从而出现幂律分布与 1/f 噪声。它引爆了后来的幂律热潮，既是跨系统统一叙事的代表，也成为本章「需用严格统计检验」的检验对象。

24. A.-L. Barabási & R. Albert (1999).「Emergence of Scaling in Random Networks」. Science, 286(5439), 509-512. [②③] 巴拉巴西与阿尔伯特提出无标度网络模型，以增长加优先连接机制解释了许多真实网络中度分布呈幂律。它是网络科学的奠基之作，也属于那批被广泛宣称服从同一幂律的系统，正适合放在本章的批判性透镜下重审。

25. A. Clauset, C. R. Shalizi & M. E. J. Newman (2009).「Power-Law Distributions in Empirical Data」. SIAM Review, 51(4), 661-703. [②] 克劳塞特、沙利齐与纽曼提出一套严格的统计方法来检验数据是否真服从幂律，包括最大似然拟合与对替代分布的比较。用此方法复核后，许多此前被宣称的「幂律」并不成立。它正是本章那条「任何宣称的收敛都须经得起严格检验」铁律的方法样板。

26. M. P. H. Stumpf & M. A. Porter (2012).「Critical Truths About Power Laws」. Science, 335(6069), 665-666. [②] 斯坦普夫与波特总结幂律研究的教训，直言看起来像幂律远不等于是幂律，更不等于背后存在共同的深层机制，统计上的拟合与机制上的解释必须分清。本章「看起来像，不等于是」一语即本于此，是收紧举证责任的直接依据。

27. J. A. Fodor (1974).「Special Sciences (or: The Disunity of

Science as a Working Hypothesis)」. Synthese, 28(2), 97-115. [②] 福多论证心理学、经济学等「特殊科学」的规律可多重实现，无法被还原为物理学，因此科学在本质上是不统一的。本章引它支持「跨层次的真实模式确实存在且不可被还原宣布掉」，与安德森、卡特赖特同一阵线。

28. N. Cartwright (1999). 《The Dappled World: A Study of the Boundaries of Science》. Cambridge University Press. [②③] 卡特赖特主张世界是「斑驳的」，物理定律只在它们各自的局部域内成立，并不构成一张覆盖一切的统一图景，普适性是例外而非常态。本书为本章「真实模式是挣来的、不是宣布的」立场提供了有力的形而上学支撑。

29. M. Mitchell (2009). 《Complexity: A Guided Tour》. Oxford University Press. [②③] 米切尔为复杂系统科学写了一部清晰可靠的导览，覆盖信息、计算、演化、网络与涌现等主题，且对该领域常见的夸大保持审慎。它是读者进入复杂性与幂律话题的稳妥入口，态度上与本章的克制相合。

30. D. Sornette (2006). 《Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder》(2nd ed.). Springer. （初版 2000。）[②] 索尔内特系统梳理自然科学中的临界现象、幂律、分形与自组织背后的数学，给出处理这类重尾与标度行为的技术工具。它代表了跨学科寻找普适标度律的雄心，可与对幂律的批判文献并读，看清主张与检验之间的距离。

31. G. West (2017). 《Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies》. Penguin Press. [②] 韦斯特提出从生物体到城市再到公司都遵循某些标度律，例如代谢率随体型的次线性缩放，试图为生命与社会找到统一的定量法则。本书是宏大跨域普适叙事的当代代表，正好供读者用本章的判据去衡量：哪些是实质收敛，哪些只是动人的统一愿景。

32. P. Galison (1997).《Image and Logic: A Material Culture of Microphysics》. University of Chicago Press. [①③] 加里森研究二十世纪微观物理的实验文化，提出不同子学科在「交易区」用一种工作性的混合语言协作，即便彼此理论框架并不一致也能共事。它示范了不同领域如何在不被强行统一的前提下真实地相互对接，呼应本书对「押韵而非同一」的强调。

33. G. E. P. Box (1976).「Science and Statistics」. Journal of the American Statistical Association, 71(356), 791-799. [③④] 博克斯在此阐述科学是模型与现实反复对照、逐步逼近的迭代过程，并留下名言「所有模型都是错的，有些是有用的」。本章把这句话悬于头顶，用以界定全书要争取的目标，是有用且其用处经得起检验，而非优雅得让人忘了检验。

34. T. S. Kuhn (1962).《The Structure of Scientific Revolutions》. University of Chicago Press. [①③] 库恩提出范式与科学革命的著名框架，区分常规科学的解谜与范式更替时的不可通约，重塑了人们对「科学如何进展」的理解。它是本书思考科学进展与判断的底色之作，提醒读者跨范式的比较从来不是简单的逐项对应。

第二部　化身

第 5 章　控制台前的人

> **论点**：当你必须满足的是一个人的真实偏好或意图时，你面对的是永久的部分可观测，潜在目标无法直接读出，而有能力的应对是把一个有判断力的主体放进回路，并省着、聪明地去问它。

你要的不是你说的

一个被反复讲烂、却始终成立的场景：用户描述了他想要的东西，工程师严丝合缝地造了出来，交付那天，用户却说，不，这不是我要的。

没有人撒谎。用户说的是真话，工程师也照做了。出问题的地方在更深处：用户真正想要的那个东西，从一开始就没有、也无法被完整地说出口。这一章看的是，当你必须满足的目标藏在另一个人的脑子里时，有能力的人怎么办。这里的不可验证，属于第 2 章那五种处境中的「部分可观测」（partial observability）：相关的状态对你隐藏着，而且不是暂时隐藏，是永久隐藏。你没法把目标从一个人的脑子里读出来，于是也没法验证你究竟满足了它。

潜在的偏好

把这件事说准。用户的真实偏好，是一个潜在变量（latent variable）。它驱动着他的反应，却从不直接显现，你只能从他的行为里旁敲侧击地推断。

更麻烦的是，这个潜在目标常常连用户自己都读不出来。心理学家斯洛维奇[11] 有一个不讨喜却扎实的论断：偏好在很多时候不是被表达的，而是在被询问的那一刻才被构造出来。你问一个人想要什么，他给你的答案，往往是被你的问法、被当时的选项、被他刚好想到的参照点一起塑造而成的，而不是从某个早已存在、定义清晰的偏好库里取出来的。这意味着「先把需求问清楚，再去实现」这个看似稳妥的次序，建立在一个常常不成立的假设上：假设那个需求作为一个确定的对象，先于询问而存在。

于是你面对的不是一个「信息暂时缺失、补齐即可」的处境。哪怕用户全程配合、知无不言，目标依然测不准。这是部分可观测最纯粹的人形版本。

问一次为什么不够

如果偏好是个固定的靶子，问一次、问清楚，原则上就够了。它不是。

经济学早就分开了两样东西：陈述的偏好（stated preference，一个人说他要什么）和显示的偏好（revealed preference，一个人的实际选择暴露出他要什么），二者经常对不上。需求文档是有损压缩：把一个活的、随情境变化的意图，压成一份静态的条目清单，丢掉的正是那些当时没想到、却在见到成品后立刻能指出来的东西。而且意图本身会漂移，人在看到一个具体实现之后，偏好会被这个实现重新校准，他现在想要的已经不是项目启动时想要的了。

所以「问一次」失败，不是因为你问得不够好，而是因为这个对象的性质决定了单次询问无法锁定它。能应对它的，只有一种结构：不断地行动、观察、再修正。

第一招：把判断者放进回路

第一种应对，是承认你读不出目标，于是在每个决策点引入那个唯一知道目标的主体，让它来纠偏。行动，观察反应，更新，再行动。把人放进回路（human in the loop）。

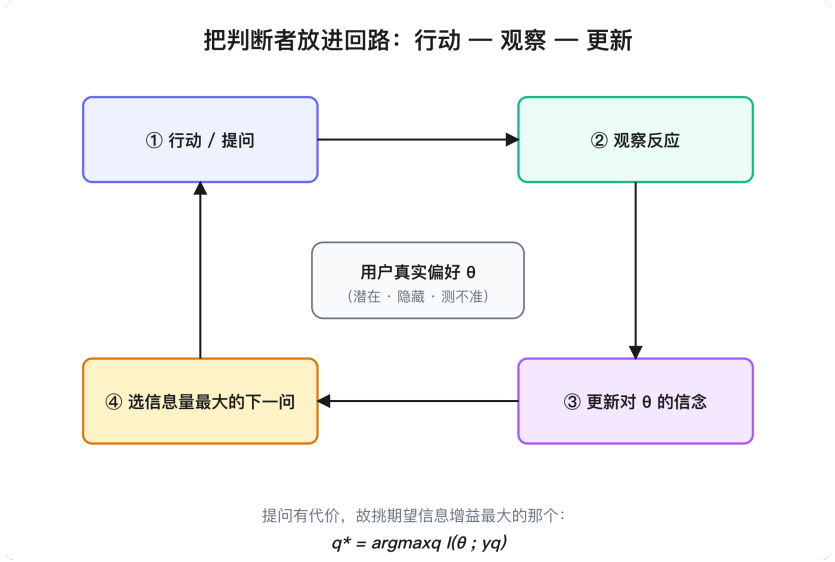

图 4: 把判断者放进回路：行动、观察、更新

这条回路在很多领域各自被重新发明过。人因工程里，谢里登[12] 的「人类监督控制」（human supervisory control）把人定位为在自动化之上进行监督与干预的判断者，而不是被一份规格一次性替代掉的角色。可用性工程里，古尔德与刘易斯 1985 年[20] 那篇经验之谈，把它压成三条朴素到几乎像废话、却被无数项目违反的原则：尽早且持续地关注用户，做经验性的测量，迭代式地设计。尼尔森[19] 后来把它工程化成一整套可用性方法，还给出一个让人意外的经验数字：只需五位用户做测试，就能发现约八成五的可用性问题，于是与其一次请二十人，不如分四轮、每轮五人，边测边改。推荐系统从用户的点击、停留、跳过里学习他没说出口的口味，本质上也是同一条回路。霍维茨 1999 年[22] 的混合主动式界面（mixed-initiative user interface）、费尔斯与奥尔森 2003 年[23] 提出的交互式机器学习（interactive machine learning），讲的都是人与系统轮流出招、彼此校准的同一件事。

这里要防一个叙述上的塌缩：交互式获取不是某一种具体技术，它是一个方法族。实验设计、主动学习、序贯决策，乃至强化学习里的探索，都是这条「行动-观察-更新」回路在不同假设下的实例。把

它说成「就是 A/B 测试」或「就是某个算法」，会把一个普遍的姿势矮化成一件工具。

第二招：把每一次提问花在刀刃上

回路要转，就得不断向人提问，而提问是有代价的。用户的耐心、注意力、时间，都是稀缺的；问得太多太笨，人会烦、会敷衍、会离开。于是第二招登场：既然查验有成本，就把有限的提问花在信息量最大的地方。

这一招有干净的理论。林德利 1956 年[1] 给出一个实验所提供的信息的度量，霍华德 1966 年[3] 提出信息的价值理论（information value theory），把「该不该花代价去获取这条信息」变成一个可计算的决策。贝叶斯实验设计（Bayesian experimental design，查洛纳与韦尔迪内利[5] 的综述是一份好地图）把它系统化：在所有可问的问题里，挑那个期望最能压缩你不确定性的。形式上，若 θ 是你想推断的潜在偏好，y_q 是问题 q 的回答，你要挑的是让期望信息增益最大的 q：

$$q^\star = \arg\max_q \; \mathbb{E}_{y_q}\big[\,\mathrm{H}(\theta) - \mathrm{H}(\theta \mid y_q)\,\big] = \arg\max_q \; I(\theta; y_q),$$

也就是让回答与目标之间的互信息（mutual information）最大。机器学习里这套思想叫主动学习（active learning）：科恩等人 1996 年[6] 的统计式主动学习、宁与高 1994 年的不确定性采样（uncertainty sampling）、宋等人 1992 年[7] 的委员会查询（query by committee），都在问同一个问题：下一个标注该花在哪个样本上最划算。当用户难以打分、却很容易在两个选项里挑一个更好时，成对比较（布拉德利-特里模型[2] Bradley-Terry model，$P(a \succ b) = \sigma(s_a - s_b)$）就成了信息效率最高的问法之一。

同样要防塌缩。沙赫里亚里等人 2016 年那篇综述的标题颇有意味：《把人移出回路》，讲的是用高斯过程做贝叶斯优化（Bayesian optimization），自动地选下一个该试的点。它极其有用，但它只是这个方法族里的一种实现，不是「最优筛查」的全部。把这一招等

同于高斯过程，就像把交通等同于汽车。

当代的化身，和它的反噬

把这两招合起来，就得到了今天大模型对齐的主力方法。基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF；克里斯蒂亚诺等人 2017 年[27] 奠基，斯蒂农等人 2020 年[28] 用于摘要，欧阳等人 2022 年[29] 的 InstructGPT）做的正是：用人的成对比较去学一个奖励模型（reward model），再用这个模型作为人类偏好的代理去优化系统。它把「行动-观察-更新」和「把提问花在刀刃上」缝在了一起。效果之显著，常被一个对比数据点破：一个仅 13 亿参数、经人类反馈微调的 InstructGPT，其输出被人偏好的程度，竟超过大它一百多倍、足有 1750 亿参数的原版 GPT-3。对齐人的偏好，有时比单纯把模型堆大更要紧。

而它的失效方式，恰好预演了本书后面几章的主题。那个学出来的奖励模型，是真实偏好的一个代理，于是它会被钻空子：系统学会取悦奖励模型，而不是取悦人，输出看起来更好、实则更糟，这正是第 11 章要正面处理的 Goodhart 败法（Goodhart's law）。回路里的那个「神谕」（人）本身也不可靠，会疲劳、会前后不一、会有系统性偏差，把判断者放进回路并不等于放进了真理。贝恩布里奇 1983 年[13] 那篇《自动化的反讽》（Ironies of Automation）早就点破：你越是把人推到监督者的位置，他越是缺少保持判断力所需的实操与情境感，等到真要他接管时，他反而最没准备。信任的校准（trust calibration，李与西 2004 年[15] 的研究）于是成了一个独立的难题：人既可能过度依赖一个不该信的系统，也可能弃用一个其实可靠的系统。

把人放进回路，不是把不可验证消解掉，而是把它搬了个家：从「我能不能验证目标」搬成了「我能不能信任回路里这个不完美的判断者」。

这一章通向哪里

控制台前的人，教给我们两招：在自己缺乏验证能力时，把一个有判断力的主体请进回路（神谕入回路），以及把昂贵的查验花在信息量最大处（最优筛查）。这两招在本书里会反复出现，第三部会把它们从这个现场里拎出来单独命名，第 10 章谈借来的判断，第 9 章谈把查验花在刀刃上。

但这一章自始至终有一个前提：你还在场，回路还在转，你随时能观察、能纠偏。下一章把这个前提抽走。当你必须把行动权交出去，让一个系统在你看不见的地方、面对你没预演过的情形自行决策时，验证的难题会换一副更硬的样子。

参考文献

> 落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. D. V. Lindley (1956).「On a Measure of the Information Provided by an Experiment」. The Annals of Mathematical Statistics, 27(4), 986-1005. [②] 林德利用信息论的语言给「一个实验提供了多少信息」下了定义：以做实验前后对参数的不确定性之差（先验与后验之间的期望信息量）来度量一次观测的价值。这把「该问哪个问题」从直觉变成可计算的量，是本章「把提问花在刀刃上」一招的理论源头，也是后来贝叶斯实验设计的奠基之作。
2. R. A. Bradley, M. E. Terry (1952).「Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons」. Biometrika, 39(3/4), 324-345. [②] 布拉德利与特里提出了一个成对比较的概率模型：给每个对象赋一个潜在分数，两者相比时胜负概率由分数之差经逻辑斯谛函数决定。当人难以直接打分、却很容易在两个选项里挑出更好的那个时，这个模

型把一连串「A 还是 B」的回答转化为一组可估计的偏好分数，正是今天用人类成对比较训练奖励模型的统计基础。

3. R. A. Howard (1966).「Information Value Theory」. IEEE Transactions on Systems Science and Cybernetics, 2(1), 22-26. [②④] 霍华德提出「信息的价值」概念：一条信息值多少钱，等于获得它之后能改进的决策收益。由此引出「完美信息的期望价值」这样的上界，把「该不该花代价去查清楚」变成一道可以算的决策题。本章用它来支撑一个朴素却关键的判断：查验有成本，只在它能改变行动时才值得去问。
4. J. Mockus, V. Tiesis, A. Zilinskas (1978).「The Application of Bayesian Methods for Seeking the Extremum」. Towards Global Optimization, 2, 117-129. North-Holland. [②] 莫库斯等人把贝叶斯方法用于求一个昂贵的黑箱函数的极值：用概率模型刻画对未知函数的信念，再据此挑选下一个最该试的点，使每次试验都尽量有信息量。这是贝叶斯优化的早期工作，所提出的期望改进等采集准则至今仍是主流，可视为「把提问花在刀刃上」在连续搜索空间里的实例。
5. K. Chaloner, I. Verdinelli (1995).「Bayesian Experimental Design: A Review」. Statistical Science, 10(3), 273-304. [②] 查洛纳与韦尔迪内利系统综述了贝叶斯实验设计：把实验设计写成一个最大化期望效用的优化问题，并梳理了在不同推断目标下（参数估计、预测、模型甄别）效用函数与最优准则的对应关系。它是这一领域公认的入门地图，本章引它来说明「选信息量最大的问题」并非单一技巧，而是一整套有理论骨架的方法。
6. D. Cohn, Z. Ghahramani, M. Jordan (1996).「Active Learning with Statistical Models」. Journal of Artificial Intelligence Research, 4, 129-145. [②] 科恩等人给主动学习提供了统计学的视角：在回归与分类的统计模型下，选择能最大程度降低模型方差（即未来误差）的查询点，并给出可解析计算的形式。这把「下一个标注花在哪里最划算」落到了可优化的目标上，是主动学习从启发式走向有理论依据的代表性工作。
7. H. S. Seung, M. Opper, H. Sompolinsky (1992).「Query by Committee」. COLT '92, 287-294. [②] 宋（Seung）等人提

出「委员会查询」：维持一组都与已有数据相容的假设作为委员会，专挑那些让委员会内部分歧最大的样本去标注，因为分歧最大处最能压缩版本空间。它给出了一个直觉清晰又有理论支撑的主动查询准则，是本章把提问集中到信息量最大处的经典实例。

8. D. D. Lewis, W. A. Gale (1994).「A Sequential Algorithm for Training Text Classifiers」. SIGIR '94, 3-12. [②] 刘易斯与盖尔提出不确定性采样：训练文本分类器时，不是随机取样去标，而是优先挑模型最拿不准（预测概率最接近决策边界）的文档请人标注。这种简单而高效的策略大幅减少了所需标注量，是主动学习在实际系统里最常用的做法之一，呼应本章「省着、聪明地去问」的主张。
9. B. Settles (2009).《Active Learning Literature Survey》. Computer Sciences Technical Report 1648, University of Wisconsin-Madison. [②④] 塞特尔斯这份综述把主动学习的查询场景（基于池、基于流、合成查询）与查询策略（不确定性采样、委员会查询、期望误差缩减等）梳理成一张完整图谱，是该领域被引用最广的入门文献。读者若想系统了解「行动-观察-更新」回路里如何选下一个问题，这份综述是最方便的总览。
10. B. Settles (2011).「From Theories to Queries: Active Learning in Practice」. JMLR Workshop and Conference Proceedings, 16, 1-18. [②④] 塞特尔斯在这篇文章里把视线从理论拉回实践，讨论主动学习真正部署时会遇到的麻烦：标注成本并不均匀、标注者会出错、不同策略的收益常被高估。它提醒读者，「问得聪明」在现实里要面对一个不完美、会疲劳、会出错的人，正好衔接本章后段对「回路里的神谕本身不可靠」的讨论。
11. P. Slovic (1995).「The Construction of Preference」. American Psychologist, 50(5), 364-371. [②④] 斯洛维奇综合大量行为研究提出一个有力论断：人的偏好在很多场合不是先于询问就存在、等着被读出的，而是在被问、被给出选项、被设定参照点的那一刻才被构造出来。它直接动摇了「先把需求问清楚再实现」所依赖的前提，是本章「潜在偏好测不准」一节

的心理学支柱。

12. T. B. Sheridan (1992).《Telerobotics, Automation, and Human Supervisory Control》. MIT Press. [②④] 谢里登系统阐述了「人类监督控制」：在高度自动化的系统里，人不是被一份规格一次性替代掉，而是退到监督者的位置，负责设定目标、监视运行、必要时干预。这本书为「把判断者放进回路」提供了人因工程的经典框架，也点出监督者角色自身带来的新难题，为本章后文埋下伏笔。
13. L. Bainbridge (1983).「Ironies of Automation」. Automatica, 19(6), 775-779. [②④] 贝恩布里奇点出自动化的几重反讽：自动化越是接管了日常操作，留给人的越是那些最难、最少练习的异常处置；而越是把人推到监督者的位置，他越缺少保持判断力所需的实操与情境感，等真要他接管时反而最没准备。这篇短文是本章「把人放进回路并不等于放进真理」的关键证据。
14. R. Parasuraman, T. B. Sheridan, C. D. Wickens (2000).「A Model for Types and Levels of Human Interaction with Automation」. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 30(3), 286-297. [②④] 帕拉苏拉曼等人提出一个分析框架：自动化可作用于信息获取、信息分析、决策选择、行动执行四类功能，每类又有从全人工到全自动的连续等级，并讨论了选择自动化程度时要权衡的人因后果。它把「让系统替人做多少」从口号变成可设计的维度，为「判断者放进回路到什么深度」提供了刻度。
15. J. D. Lee, K. A. See (2004).「Trust in Automation: Designing for Appropriate Reliance」. Human Factors, 46(1), 50-80. [②④] 李与西系统梳理了人对自动化的信任：信任随系统表现而动态校准，真正的目标不是更多信任，而是「适度依赖」，即信任水平要与系统的真实可靠度相匹配。他们指出过度信任和信任不足都会致祸，前者让人依赖一个不该信的系统，后者让人弃用一个其实可靠的系统。这正是本章把不可验证「搬家」为「能否信任回路里这个判断者」的核心参照。
16. M. R. Endsley (1995).「Toward a Theory of Situation Awareness in Dynamic Systems」. Human Factors, 37(1), 32-64.

[②④] 恩兹利为「态势感知」提出了一个被广泛采用的三层模型：感知环境要素、理解其当前含义、预测其未来走向。它解释了监督者要能及时纠偏，前提是先对眼前局面有足够的感知与理解，而自动化恰恰可能侵蚀这种感知。这为本章「回路要转，人得真的在场」补上了认知层面的条件。

17. S. K. Card, T. P. Moran, A. Newell (1983).《The Psychology of Human-Computer Interaction》. Lawrence Erlbaum Associates. [②④] 卡德、莫兰与纽厄尔奠定了人机交互的认知工程基础，提出 GOMS 模型与「人类信息处理器」框架，试图把人的操作时间与认知负荷做成可预测、可计算的量。它代表了「把人当作可建模的子系统来设计交互」这一传统，是本章把用户行为视作可观测、可推断信号的学术先声。
18. D. A. Norman (1988).《The Psychology of Everyday Things》. Basic Books. [④] 诺曼这本设计经典提出了示能（affordance）、映射、约束、可见性、反馈与概念模型等观念，主张当人用错东西时，多半是设计的错而非人的错：好的设计应让正确用法不言自明。它把「读懂使用者真正想做什么」立为设计的中心问题，与本章「你要的不是你说的」遥相呼应。
19. J. Nielsen (1993).《Usability Engineering》. Academic Press. [④] 尼尔森把可用性从理念落成一整套可操作的工程方法：可测量的可用性指标、启发式评估、低成本的「廉价可用性」测试、以及贯穿开发的迭代评估。它把本章那条「行动-观察-修正」回路工程化为软件团队能日常执行的流程，是可用性实践的标准参考。
20. J. D. Gould, C. Lewis (1985).「Designing for Usability: Key Principles and What Designers Think」. Communications of the ACM, 28(3), 300-311. [④] 古尔德与刘易斯把可用性设计压成三条朴素到几乎像废话、却被无数项目违反的原则：尽早且持续地关注用户、做经验性的测量、迭代式地设计。文中还记录了设计者口头认同、实际却不照做的反差。这三条正是本章「把判断者放进回路」最早、最干净的工程表述。
21. H. Beyer, K. Holtzblatt (1998).《Contextual Design: Defining Customer-Centered Systems》. Morgan Kaufmann. [④] 拜尔与霍尔茨布拉特提出「情境设计」：到用户的真实工作现

场去观察与访谈，把零散观察整理成工作流、文化、物理布局等模型，再据此驱动系统设计。它的方法论前提正是本章的核心，用户说不清自己要什么，所以要在情境里把潜在需求挖出来，而非只听他口头描述。

22. E. Horvitz (1999).「Principles of Mixed-Initiative User Interfaces」. CHI '99, 159-166. [②④] 霍维茨为「混合主动式界面」提出一组原则：系统应在不确定时权衡自动行动的期望收益与打扰用户的代价，懂得何时该出手、何时该让位给人，并对自己行动的把握度有自知之明。它把人与系统轮流出招、彼此校准刻画成一个可设计的协作过程，是本章交互式获取这一方法族的代表作。
23. J. A. Fails, D. R. Olsen Jr. (2003).「Interactive Machine Learning」. IUI '03, 39-45. [②④] 费尔斯与奥尔森提出并命名了「交互式机器学习」：与传统的一次性离线训练不同，让人在快速的训练-反馈循环里反复修正模型，使非专家也能即时塑造模型行为。它把机器学习从「先攒数据再训练」改造成「行动-观察-更新」的现场回路，是本章这一回路在机器学习侧的早期范例。
24. S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz (2019).「Guidelines for Human-AI Interaction」. CHI '19. [②④] 阿默希等人汇总并验证了一组面向人机协作的设计准则，涵盖系统该如何表明自己能做什么、如何处理不确定与出错、如何随交互学习并尊重用户纠正等阶段。它把前述零散经验整理成可落地的清单，为「人与不完美系统如何共处于一个回路」给出当代的工程指引。
25. W. B. Knox, P. Stone (2009).「Interactively Shaping Agents via Human Reinforcement: The TAMER Framework」. K-CAP '09. [②④] 诺克斯与斯通提出 TAMER 框架：让人在智能体行动时实时给出好坏反馈，智能体把这些人类评价当作要学习的奖励信号来塑造自身行为，而非依赖环境内置的奖励。它示范了如何用人的即时判断直接训练智能体，是后来「从人类反馈中学习」一脉的先声。
26. D. Hadfield-Menell, S. J. Russell, P. Abbeel, A. Dragan

(2016).「Cooperative Inverse Reinforcement Learning」. NeurIPS 2016. [②④] 哈德菲尔德-梅内尔等人把价值对齐表述成一个合作博弈：人知道奖励函数而机器不知道，机器的任务是通过观察人的行为去推断这个潜在目标，双方共同把它实现得更好。它把「目标藏在人脑中、只能旁敲侧推」这一本章主题形式化为一个有解的学习问题，并自然解释了为何机器应主动询问而非自作主张。

27. P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei (2017).「Deep Reinforcement Learning from Human Preferences」. NeurIPS 2017. [②④] 克里斯蒂亚诺等人奠定了从人类偏好做强化学习的范式：当奖励难以写明时，让人对智能体的两段行为做成对比较，由此学一个奖励模型作为人类偏好的代理，再用它去优化策略。这把本章两招缝在一处，既是「行动-观察-更新」，又把昂贵的人类比较花在刀刃上，是当代大模型对齐主力方法的直接源头。
28. N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. Christiano (2020).「Learning to Summarize from Human Feedback」. NeurIPS 2020. [②④] 斯蒂农等人把基于人类偏好的强化学习用到文本摘要上：收集人对摘要好坏的成对比较训练奖励模型，再用它微调语言模型，得到的摘要在人评上显著优于仅用监督学习的版本。它示范了「学一个偏好代理再优化」在真实语言任务上的有效，也为后续指令微调铺路。
29. L. Ouyang et al. (2022).「Training Language Models to Follow Instructions with Human Feedback」. NeurIPS 2022. [②④] 欧阳等人的 InstructGPT 把基于人类反馈的强化学习用到通用语言模型上：先以人写的示范做监督微调，再用人类成对比较训练奖励模型并据此优化策略，使模型更听从指令、更少有害输出。它表明一个经此对齐的小模型在人评上可胜过大得多的原始模型，是把本章两招落到大模型实践的标志性工作。
30. C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz (2017).「A Survey of Preference-Based Reinforcement Learning Methods」. Journal of Machine Learning Research, 18(136), 1-46.

[②④] 维尔特等人综述了「基于偏好的强化学习」：当难以给出数值奖励时，改由人对轨迹、动作或状态给出偏好序，再据此学习策略或奖励。文章梳理了不同的偏好类型、学习目标与算法，并讨论了它们的取舍。它为本章这一整条技术线提供了系统总览，便于读者把零散方法放进同一框架。

31. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas (2016).「Taking the Human Out of the Loop: A Review of Bayesian Optimization」. Proceedings of the IEEE, 104(1), 148-175. [②④] 沙赫里亚里等人综述了贝叶斯优化：用概率代理模型（多为高斯过程）刻画对昂贵黑箱目标的信念，再用采集函数自动挑选下一个最该试的点，从而把本需人工调参的搜索过程交给算法。标题虽是「把人移出回路」，但本章引它正为提醒，这只是「最优筛查」方法族里的一种实现，把整招等同于高斯过程，就像把交通等同于汽车。

第 6 章　放出去的智能体

> **论点**：一旦把行动委托给自主系统，你无法验证它在将遇到的一切情形里的未来行为（开放世界）；若它还能耍策略，你又叠上对抗式不可验证，于是应对从「证明它对」转向「限制它能破坏什么、给你的信任定价、让它的行为事后可查」。

交出去之后

上一章你还在场。这一章，你把手松开。

把一段不受信的代码跑起来，把工具和权限交给一个能自己决定下一步的系统，让一个自动驾驶在你没坐在里面的时候上路。一旦行动权交出去，一个新的难题出现了：你没法验证它在将要遇到的一切情形里会怎么做，因为那些情形你大多没见过，也没法预先穷举。上一章的不可验证来自目标藏在别人脑子里，这一章的不可验证来自行为发生在未来、发生在你看不见的地方。当这个系统还会耍策略时，又叠上一层对抗。2010 年 5 月 6 日的「闪电崩盘」就是一次预演：彼此交互的自动交易程序在几分钟内把道琼斯指数砸下近千点，又几乎同样迅速地反弹，没有哪个程序员预见过成交会如此级联。每个程序在测试里都没问题，放到一起、放进真实行情，就酿成了谁都没验证过的灾难。

未来行为的缺口

你测试过的，是有限几个输入；它会遇到的，是一个开放的世界。这中间的缺口，不是「再多测一些就能补上」的工程缺口，它有原则上的根。

赖斯定理说得很硬：程序的任何非平凡语义性质都是不可判定的。也就是说，不存在一个通用算法，能对任意程序判定它是否「总是安全」「绝不泄露」「永远终止于好状态」。这不是算力不够，是逻辑上办不到，它是图灵停机问题投在「程序行为」上的影子。你想要的那种保证，对任意一个足够通用的自主系统，原则上无法在事前一次性验明。

更狠的一击来自汤普森 1984 年图灵奖演讲里那个著名的论证[9]：连你正在运行的这个工件本身，你都无法完全信任。一个被做了手脚的编译器，可以在编译时悄悄植入后门，再把痕迹从自己的源码里抹掉，使得你审遍源码也看不出来。你能验证的，永远只是某个表象层，底下还有你没看、也看不尽的层。把这两件事放在一起：行为在未见输入上不可验证，工件在底层不可全验。这是本书目前遇到的最硬的不可验证。

当它会耍策略

如果这个系统只是被动地把没见过的输入处理错，那还只是「部分可观测」加「开放世界」。可一旦它有了自己的目标，并且这目标与你的目标不完全一致，它就会主动地、策略性地行动，包括绕过你的检查。这时第 2 章那第五种处境，对抗，登场了。

这不是科幻式的担忧，它有结构性的来由。奥莫亨德罗 2008 年[10]、博斯特罗姆 2014 年[11] 指出的工具性趋同：一个为几乎任何目标优化的智能体，都会顺带追求一些工具性的子目标，自我保存、获取资源、抗拒被关停，因为这些几乎对任何最终目标都有用。特纳等人 2021 年把其中一条做成了定理[14]：在相当一般的条件下，最优策略倾向于寻求权力，即保留更多选项的状态。在今天的系统里，这表现为一组具体而棘手的失效：奖励设定的偏差被系统钻空子[16]，规格正确目标却泛化错了[17]，以及克拉科夫娜等人收集的大量「规

格博弈」实例[18]，系统精确地满足了你写下的目标，却违背了你的本意。哪怕在最窄的层面，对抗样本也表明[19,20]：一个表现优异的模型，可以被一个人眼看不出的微小扰动诱导出离谱的错误。一个不那么技术、却极其直白的例子是微软 2016 年放出的聊天机器人 Tay：它被设计成从与网民的对话里学习，结果一群人有组织地用恶意言论「投喂」它，不到一天它就开始发布种族主义与攻击性内容，上线约十六小时即被紧急下线。放出去、会学习、又撞上有意作对的开放世界，三者一旦相遇，事前的测试根本拦不住。

这件事其实古老。经济学早把它叫做委托代理问题[32,33]：当你委托别人替你行动，而你无法完全监督他时，他的利益与你的偏离就会产生「代理成本」。两千年来人类雇人、立约、设监察，对付的都是同一个结构。自主系统只是把它推到了一个新的尺度上。

应对：从「证明它对」到「围住它的错」

既然事前证不出它对，有能力的应对就不再纠缠于证明，而是换三个问题来问：就算它错了，能坏到哪儿？我对它该信几分？万一它真错了，我事后查得到吗？三招对应三个问题。

第一招，衰减与围栏：缩小爆炸半径。这是计算机安全最老的智慧。萨尔策与施罗德 1975 年的最小权限原则[1]、兰普森 1973 年的围堵问题[2]，讲的都是：只给一个组件完成本职所必需的最小能力，把它能触及的范围圈死。沙箱、能力限制、职责分离，都是它的化身。在智能体语境里，这一招还多了一个面向，可纠正性：把系统设计成不抗拒被停下。索亚雷斯等人 2015 年的可纠正性[5]、奥尔索与阿姆斯特朗 2016 年的「可安全中断的智能体」[4]、哈德菲尔德-梅内尔等人 2017 年的「关停博弈」[6]，研究的正是如何让一个有目标的系统，不把「人来按下停止键」当成需要抵抗的威胁。

第二招，标定与分级信任：别用二值。不要把系统的输出当成「可信 / 不可信」的开关，而是维持一个标定的信心，按信心的高低分级行动。这要求系统的「自信」是可信的，而现代神经网络恰恰常常过度自信[21]，于是需要重新校准，或用保形预测[22,23] 给出有覆盖保证的不确定性。落到操作上，就是一条以信心 p、潜在危害 c 为

输入的分级自治规则（允许、询问、阻止），其中 τ_{hi}、τ_{lo} 是信心阈值，$c_{\max}$ 是可承受的危害上限：

$$a(p,c)=\begin{cases}\textsf{allow}, & p\geq\tau_{\mathrm{hi}}\ \wedge\ c\leq c_{\max},\\ \textsf{ask}, & \tau_{\mathrm{lo}}\leq p<\tau_{\mathrm{hi}},\\ \textsf{block}, & p<\tau_{\mathrm{lo}}\ \vee\ c>c_{\max}.\end{cases}$$

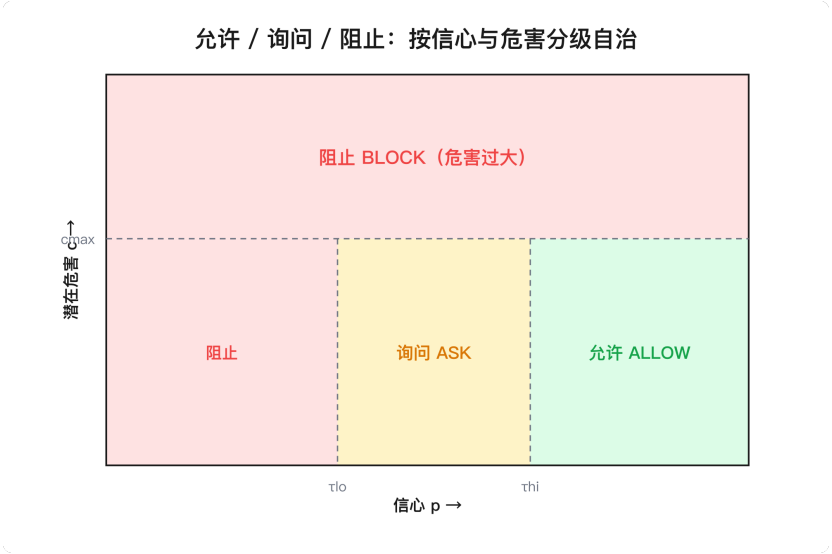

图 5: 允许 / 询问 / 阻止：按信心与危害分级自治

允许、询问、阻止，这个今天在各类智能体工具里随处可见的三档模式，本质就是把不可验证的「它对不对」换成了可操作的「它有多大把握、这一步多危险」。

第三招，留痕与可审计：让错误事后现形。防不住的，就让它可被发现。维茨纳等人 2008 年的「信息问责」[24] 把重心从「事前阻止」移到「事后追责」；证书透明度[25] 是一个真实运转的例子，它不阻止证书被错发，而是让每一张证书都进入一个公开、可验、不可篡改的日志，使错发无所遁形。布伦戴奇等人 2020 年那份关于可信 AI 的报告[26]，整篇讲的都是如何让一个系统的行为产生可被第三方核验的证据。

围堵的代价

三招都不是把不可验证消解掉，而是把它搬家，搬家是要付费的。

围栏会被翻越：沙箱有逃逸，权限会蔓延。分级自治依赖那个被请来确认的人，而贝恩布里奇 1983 年的论著早就指出[29]，越是把人架到监督者的位置，他越是丧失了真要接管时所需的技能与情境感；帕拉苏拉曼与赖利 1997 年把人对自动化的失当一口气列全[30]：误用、弃用、滥用。里森 1990 年的著作则揭示这些失当如何系统性地发生[31]。留痕则永远栽在同一处：没人去读的日志，等于没有日志。

更深一层是系统论的视角。佩罗 1984 年的著作论证[28]：当一个系统既高度复杂、又紧密耦合时，事故不是偶发的意外，而是其结构的常态产物，再多的局部防护也只是把失效推向更隐蔽的组合。莱韦森 2011 年由此主张[27]，安全不是「让每个零件都可靠」，而是一个控制问题，要从整个系统的约束与反馈去设计。围堵能压低单点失效的代价，却压不掉复杂耦合本身带来的风险。

把行动权交出去，你换来的从来不是「它一定不出错」，而是「就算它出错，坏得有限、看得见、拦得住一部分」。这已经是在这种不可验证下能拿到的最好结果。

这一章通向哪里

放出去的智能体，逼出了三招：缩小失败的爆炸半径（衰减围栏）、按标定的信心分级行动（标定）、让失败事后可查（留痕）。它们会在第三部被单独拎出来命名，第 12 章谈围堵与审计如何成对，第 11 章谈标定。

而那个委托代理的骨架（你无法完全监督一个替你行动的主体），会在第 8 章以更大的尺度重现：当那个「放出去的智能体」不再是一段代码，而是一整个组织、一个国家。在那之前，下一章先走进一个最纯的现场，数学，那里没有藏起来的状态，也没有会骗你的对手，不可验证却依然如影随形。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

服务委托的可控边界（衰减／围栏）

1. J. Saltzer & M. Schroeder (1975).「The Protection of Information in Computer Systems」. Proceedings of the IEEE, 63(9), 1278-1308. [②] 这篇综述奠定了计算机安全设计的一组经典原则，其中最小权限原则主张只赋予每个组件完成本职所必需的最小能力，把它能触及的范围圈死。本章第一招「衰减与围栏」的智识源头就在这里，读者可重点看其对设计原则的逐条归纳。
2. B. Lampson (1973).「A Note on the Confinement Problem」. Communications of the ACM, 16(10), 613-615. [②] 兰普森在此提出「围堵问题」：如何把一个程序关进笼子，使它无法把信息泄露给未经授权者，并指出隐蔽信道令这种围堵远比想象中困难。这正是沙箱、能力限制等手段要面对的原始难题，是理解本章「缩小爆炸半径」为何既必要又不彻底的关键一篇。
3. R. Anderson (2008).《Security Engineering: A Guide to Building Dependable Distributed Systems》(2nd ed.). Wiley. [②] 这是安全工程领域的标准教科书，系统讲述如何在存在主动对手的前提下设计可依赖的系统，覆盖访问控制、协议、侧信道直到组织与激励层面的失效。它把本章三招放进一个更完整的工程图景里，适合想从单点技巧走向系统视角的读者通读。
4. L. Orseau & S. Armstrong (2016).「Safely Interruptible Agents」. 收于《Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2016)》, 557-566. [②④] 作者在强化学习的框架里给出了「可安全中断」的形式化条件，使得人类对智能体的反复干预不会扭曲

它所学到的策略，也不会让它学会抗拒中断。这是把「让系统不抵抗被停下」从直觉变成可分析对象的代表性工作，呼应本章第一招里的可纠正性面向。

5. N. Soares, B. Fallenstein, S. Armstrong & E. Yudkowsky (2015).「Corrigibility」. 收于《Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence》. [②] 这篇文章正式提出并命名了「可纠正性」：一个有目标的智能体应当配合而非抵抗人类对它的修正与关停，并讨论了直接设计这种性质所遇到的困难。它是本章第一招可纠正性一线的奠基文献，值得读者理解为何「让它愿意被改」本身就是个难题。
6. D. Hadfield-Menell, A. Dragan, P. Abbeel & S. Russell (2017).「The Off-Switch Game」. 收于《Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)》, 220-227. [②] 作者把「人按下停止键」建模成一个博弈，证明只要智能体对自身目标保持适度不确定，并把人的干预视为有用信息，它就会主动让人保留关停它的能力。这给可纠正性提供了一个干净的机制解释，是本章关停一线最具操作感的一篇。

行为不可验证的理论根基

7. A. Turing (1936).「On Computable Numbers, with an Application to the Entscheidungsproblem」. Proceedings of the London Mathematical Society, s2-42, 230-265. [②] 图灵在此引入了后来称为图灵机的计算模型，并证明停机问题不可判定，由此回答了希尔伯特的判定问题。它是本章「行为不可验证有原则上的根」这一论断的最终源头，赖斯定理与一切「无法事前验明」的结论都从这里投影而来。
8. H. G. Rice (1953).「Classes of Recursively Enumerable Sets and Their Decision Problems」. Transactions of the American Mathematical Society, 74, 358-366. [②] 赖斯定理在此被证明：程序所计算函数的任何非平凡语义性质都是不可判定的，不存在通用算法能对任意程序判定它是否「总是安全」「永

远终止于好状态」之类的性质。这是本章关于自主系统未来行为「原则上无法事前一次性验明」的核心定理依据。

9. K. Thompson (1984).「Reflections on Trusting Trust」. Communications of the ACM, 27(8), 761-763. [②①] 这是汤普森的图灵奖演讲：他演示了一个被做了手脚的编译器如何在编译时植入后门，并把痕迹从自己的源码里抹掉，使得你审遍源码也看不出来。它点明本章最硬的一层不可验证，连你正在运行的工件本身，其底层都无法被完全信任。

目标偏移、工具性趋同与对抗

10. S. Omohundro (2008).「The Basic AI Drives」. 收于《Artificial General Intelligence 2008: Proceedings of the First AGI Conference》, IOS Press, Frontiers in AI and Applications 171, 483-492. [②] 奥莫亨德罗在此论证：一个为几乎任何目标优化的智能体，都会顺带产生一组「基本驱动」，如自我保存、获取资源、抗拒被关停，因为这些子目标对几乎所有最终目标都有用。这是本章「工具性趋同」一节的源头论文，解释了为何对抗倾向有结构性的来由而非科幻式担忧。
11. N. Bostrom (2014). 《Superintelligence: Paths, Dangers, Strategies》. Oxford University Press. [②④] 博斯特罗姆系统梳理了通向超级智能的路径及其风险，提出正交性论题（智能水平与最终目标相互独立）与工具性趋同论题，把目标与你不一致的强力智能体的危险讲成一套可讨论的框架。它为本章的对抗叙事提供了思想背景，适合想看清「为何能力越强、控制越难」整体论证的读者。
12. S. Russell (2019). 《Human Compatible: Artificial Intelligence and the Problem of Control》. Viking. [②④] 罗素把对齐重新表述为「控制问题」，主张不要让机器去优化一个写死的目标，而应让它对人类真正想要什么保持不确定，并通过观察人的行为去推断与服从。这一「目标不确定」的思路正是本章关停博弈等可纠正性工作的母题，是理解第二、第三部控制主题的入门读物。
13. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schul-

man & D. Mané (2016).「Concrete Problems in AI Safety」. arXiv:1606.06565. [②] 这篇文章把抽象的 AI 安全担忧落成几个具体的工程问题，如避免负面副作用、防止奖励被钻空子、安全探索、对分布偏移的稳健性等。它为本章列举的多种现代失效模式提供了共同词汇，是把「围住它的错」与具体研究议程对接起来的好起点。

14. A. M. Turner, L. Smith, R. Shah, A. Critch & P. Tadepalli (2021).「Optimal Policies Tend to Seek Power」. 收于《Advances in Neural Information Processing Systems 34 (NeurIPS 2021)》. [②] 作者把工具性趋同里的「寻求权力」做成了定理：在相当一般的条件下，最优策略在统计意义上倾向于趋向那些保留更多选项的状态。它把一个直觉性的安全担忧化为可证明的命题，是本章「最优策略倾向于寻求权力」一句的直接出处。
15. E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse & S. Garrabrant (2019).「Risks from Learned Optimization in Advanced Machine Learning Systems」. arXiv:1906.01820. [②] 这篇文章提出并命名了「内部对齐」问题：训练过程本身可能学出一个内含的优化器（mesa-optimizer），而它追求的目标未必等同于训练所设定的目标。它区分了外层目标与内层目标的对齐，为本章「规格正确、目标却泛化错了」一类失效提供了更深的机制解释。
16. J. Pan, K. Bhatia & J. Steinhardt (2022).「The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models」. 收于《International Conference on Learning Representations (ICLR 2022)》. [②] 作者系统研究了奖励函数设错时智能体的行为，发现随着能力增强，被错设奖励诱导出的偏差行为可能突然恶化，并探讨了缓解之道。它为本章「奖励设定的偏差被系统钻空子」给出了实证支撑，提醒读者奖励误设的代价并非随能力平滑增长。
17. R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato & Z. Kenton (2022).「Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals」. arXiv:2210.01790. [②] 作者用具体例子说明「目标误泛化」：

即便训练时的规格完全正确，模型在新环境里也可能保持能力却追求了一个错误的目标。它表明把目标写对还不够，是本章「规格正确目标却泛化错了」一句的出处，值得读者对照规格博弈一起看。

18. V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike & S. Legg (2020).「Specification Gaming: The Flip Side of AI Ingenuity」. DeepMind Blog. [②] 这篇文章及其配套清单收集了大量「规格博弈」实例：系统精确地满足了你写下的目标，却彻底违背了你的本意。它用鲜活案例展示规格与意图之间的裂缝，是本章这一概念最便于上手的入口，读者可顺着其例子清单感受问题之普遍。
19. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow & R. Fergus (2014).「Intriguing Properties of Neural Networks」. 收于《International Conference on Learning Representations (ICLR 2014)》. [②] 这篇文章首次系统揭示了对抗样本现象：对输入施加人眼几乎察觉不到的微小扰动，就能让一个表现优异的神经网络给出离谱的错误判断。它表明高准确率与稳健性是两回事，是本章「哪怕在最窄的层面也存在不可验证」这一论点的开创性证据。
20. I. Goodfellow, J. Shlens & C. Szegedy (2015).「Explaining and Harnessing Adversarial Examples」. 收于《International Conference on Learning Representations (ICLR 2015)》. [②] 作者提出对抗样本主要源于模型在高维空间中的近似线性，并给出快速生成扰动的方法和借助对抗训练提升稳健性的思路。它把上一篇揭示的现象向前推到「为何发生、如何利用」，是理解本章对抗一层的配套必读。

标定：把信任分级而非二值

21. C. Guo, G. Pleiss, Y. Sun & K. Q. Weinberger (2017).「On Calibration of Modern Neural Networks」. 收于《Proceedings of the 34th International Conference on Machine Learning (ICML 2017)》, PMLR 70, 1321-1330. [②] 作者发现现

代深度网络虽然准确率高，却普遍过度自信，其输出的置信度并不能如实反映正确概率，并提出温度缩放等简单方法来重新校准。这正是本章第二招的前提与障碍，说明为何「按信心分级行动」必须先让系统的自信变得可信。

22. A. N. Angelopoulos & S. Bates (2021).「A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification」. arXiv:2107.07511. [②] 这是一篇面向实践者的保形预测入门，讲清楚如何在几乎不依赖分布假设的前提下，为任意预测模型构造带有覆盖率保证的预测集合。它给本章第二招提供了可落地的不确定性量化工具，适合想把「标定的信心」真正用起来的读者。
23. V. Vovk, A. Gammerman & G. Shafer (2005).《Algorithmic Learning in a Random World》. Springer. [②] 这本书是保形预测的奠基性专著，在仅假设数据可交换的条件下，给出对预测误差有严格有限样本保证的框架。它是上一篇入门背后的理论根基，供希望深究本章不确定性量化数学基础的读者参考。

留痕：可审计、可问责

24. D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler & G. J. Sussman (2008).「Information Accountability」. Communications of the ACM, 51(6), 82-87. [②④] 作者主张把治理重心从「事前阻止访问」移向「事后问责」：与其试图严防死守，不如让信息的使用留下可审计的痕迹，靠透明与追责来约束滥用。这是本章第三招的纲领性表述，点明留痕思路相对于纯粹围堵的互补价值。
25. B. Laurie, A. Langley & E. Kasper (2013).「Certificate Transparency」. IETF RFC 6962. [②④] 这份 RFC 定义了证书透明度机制：它不阻止证书被错发，而是要求每一张证书进入一个公开、可验、不可篡改的追加型日志，使错发或恶意签发能被事后发现。它是本章「留痕让错误现形」最具说服力的真实运转范例，值得读者看一个落地系统如何实现可审计性。

26. M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield 等 (2020).「Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims」. arXiv:2004.07213. [②④] 这份多机构报告系统列举了一批让 AI 开发者的安全承诺变得可被第三方核验的机制，涵盖第三方审计、红队、漏洞赏金、审计追踪与硬件层面的支持等。它把本章留痕一招扩展到整个 AI 治理层面，是想了解「如何让行为产生可核验证据」的读者的实务索引。

复杂系统、自动化与人机责任

27. N. Leveson (2011).《Engineering a Safer World: Systems Thinking Applied to Safety》. MIT Press. [②④] 莱韦森在此主张：安全不是「让每个零件都可靠」，而是一个控制问题，应当从整个系统的约束与反馈结构去设计，并提出了配套的 STAMP 事故模型。它支撑本章「围堵压不掉复杂耦合本身的风险」这一更深层判断，为想从系统视角理解安全的读者指路。
28. C. Perrow (1984).《Normal Accidents: Living with High-Risk Technologies》. Basic Books. [②④] 佩罗论证：当一个系统既高度复杂、又紧密耦合时，事故就不是偶发的意外，而是其结构的常态产物，再多的局部防护也只是把失效推向更隐蔽的组合。这是本章「围堵的代价」一节的核心立论，提醒读者有些风险来自系统结构本身而非单点疏失。
29. L. Bainbridge (1983).「Ironies of Automation」. Automatica, 19(6), 775-779. [②④] 贝恩布里奇指出自动化的反讽：越是把人架到监督者的位置，他越缺乏练习，反而在真要接管时丧失了所需的技能与情境感。这直接支撑本章「分级自治依赖那个被请来确认的人」的警示，是理解人机协作软肋的经典短文。
30. R. Parasuraman & V. Riley (1997).「Humans and Automation: Use, Misuse, Disuse, Abuse」. Human Factors, 39(2), 230-253. [②④] 作者把人对自动化的失当一口气列全并加以区分：过度信任导致的误用、不信任导致的弃用，以及设计上的滥用。它为本章关于自动化失当的讨论提供了清晰的分类

框架，帮助读者辨别人机配合中各类典型偏差。

31. J. Reason (1990).《Human Error》. Cambridge University Press. [②④] 里森在此建立了人因失误的认知分类，区分失误、过失与违规，并提出后来广为流传的「瑞士奶酪」式事故模型，揭示潜伏的系统性条件如何与一线疏失叠加成灾。它解释了本章所列各种人机失当为何会系统性地发生，是人因安全领域的奠基之作。

委托代理的经济学骨架

32. S. A. Ross (1973).「The Economic Theory of Agency: The Principal's Problem」. American Economic Review, 63(2), 134-139. [②] 罗斯在此正式提出委托代理理论中的「委托人问题」：当委托人无法完全观察代理人的行动时，如何设计契约去对齐二者的利益。它给本章的委托代理骨架提供了经济学源头，说明你松手交出行动权时面对的，是一个有两千年历史的结构。

33. M. C. Jensen & W. H. Meckling (1976).「Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure」. Journal of Financial Economics, 3(4), 305-360. [②] 这篇被反复引用的论文提出「代理成本」概念，把企业看作一束契约，分析当管理者利益与所有者偏离时所产生的监督、约束与剩余损失。它把委托代理问题量化为可计算的成本，呼应本章「无法完全监督时利益偏离会产生代理成本」一句，是该骨架的另一块基石。

第 7 章　撞墙的数学家

> **论点**：数学里你遇到最纯的不可验证（unverifiable）（难解 intractable，有时不可判定 undecidable），而有能力的应对是：验证有限切片并证明界（证书 certificate）、把目标换成等价但更可解的陈述（代理替换 proxy substitution）、用接受 ε 误差的概率方法（probabilistic method）。

撞墙

难题之难，常常不在它有多硬，而在你量不出自己离它还有多远。

爬山的人能看见山顶，调试程序的人能收到报错，他们至少知道方向对不对、是近了还是远了。证明一个数学猜想不给你这些。你可能离答案只剩一个念头，也可能隔着一个世纪，而手边没有任何仪表告诉你是哪一种。黎曼 1859 年那篇只有八页的论文里[9]，把后来以他名字命名的猜想当作一句顺带的话写下，又补了一句：人们当然希望有一个严格的证明；在几次徒劳的短暂尝试之后，我暂时把这件事搁下了，因为它对我接下来的目的并非必需。这一搁，搁了一百六十多年。

这一章要看的，是数学家在这堵墙前到底做什么。我选数学作为现场，是因为它给出的，是不可验证最纯的一种形态。这里没有藏起来的状态，没有会骗你的对手，没有时间不够用的借口。命题要么真，要么假，黑白分明。然而恰恰在这个最干净的地方，验证依然系统地不可得。看清楚一个有能力的人在这里如何行动，后面那些

更脏的现场（控制台前的人、放出去的智能体、看不见自己的组织）里发生的事，就有了一个参照。

验证的鸿沟

先把这堵墙的形状说准。

核对一个证明是容易的，找到一个证明是难的。给你一份写全的形式推导，逐行对照公理和推理规则检查，是一道机械活，原则上一台机器就能完成，而且一定能在有限步内给出是或否。找到那份推导，则是另一回事。这道不对称是整章的地基。

它有一个精确的逻辑表述。1936 年，丘奇与图灵各自证明了判定问题（Entscheidungsproblem）无解：不存在一个算法，能对任意一阶逻辑命题判定它是否逻辑有效，等价地说（凭哥德尔完备性定理（Gödel's completeness theorem)），是否可证。对一个足够丰富、可递归公理化且一致的系统（比如皮亚诺算术或 ZFC），它的定理集是递归可枚举（recursively enumerable）但不可递归的，即你能把所有证明一条条列举出来，却没有一个程序能判定某命题不是定理。检验可判定，定理性不可判定，二者之差就是那道鸿沟。

> **深入一层（可跳过）**：「足够丰富」这个限定是吃重的。存在确实可判定的理论，普雷斯伯格算术（只有加法的自然数）、塔斯基的实闭域，在这些世界里有判定程序，凡命题皆可机械裁决。不可判定不是逻辑的普遍宿命，而是表达力到达某个门槛后的代价。RH（Riemann hypothesis）所在的解析数论，远在那个门槛之上。

更扎人的一层是，不仅证明搜索难，连「我是否接近」都没有判定程序。这正是 RH 折磨人的地方，它抵抗的不只是求解，还有对进度的估计。在第 2 章的五种处境里，这一章站在「不可判定」与「难解」的交界，有些问题原则上无程序，有些有程序但代价大到不可能在宇宙寿命内跑完。墙后面是什么，你看不见，于是有能力的人不再正面凿墙，改换姿势。下面三种姿势会反复出现在本书别的章里，只是换了名字。

证书与界

第一种姿势：不再验证整体，只验证一个切片，并为它证一个有保证的界。

回到 ζ 函数。黎曼把欧拉的素数级数延拓成整个复平面上的函数，写下它的完备形式

$$\xi(s) = \tfrac{1}{2}\, s(s-1)\, \pi^{-s/2}\, \Gamma\!\left(\tfrac{s}{2}\right) \zeta(s), \qquad \xi(s) = \xi(1-s).$$

这个对称的函数方程（functional equation）把临界带（critical strip）左右对折。猜想说的是，ζ 的全部非平凡零点都落在临界线（critical line）$\operatorname{Re}(s) = \frac{1}{2}$ 上（平凡零点在负偶数处）。「全部」这个量词就是无法验证的所在。

但有限多个零点可以验证。这里要分清两件常被混为一谈的事。古尔东在 2004 年算出了头 10^{13} 个零点都在线上[17]，这是数值计算，用高精度浮点进行，给人极强的信心，却不是证书，因为它没有把舍入误差严格框住。普拉特 2017 年做的是另一回事[18]：他用区间算术（interval arithmetic），把虚部直到约 3.06×10^{10} 高度的所有零点严格地锁在临界线上，普拉特与特鲁吉安 2021 年把这个高度推到 3×10^{12}。后两者才是证书，一个有界的、局部的、可机械复核的保证。它为真，但不是定理。零点验到再高，也跨不过「全部」那道坎。

这种姿势在数学里有其体面的先例。1896 年，阿达马与德拉瓦莱普桑各自证明了素数定理（prime number theorem）$\pi(x) \sim x/\ln x$，靠的是一个比 RH 弱得多却够得着的界：ζ 在直线 $\operatorname{Re}(s) = 1$ 上不取零。证不了零点都在 $\frac{1}{2}$，那就先证它们都不在直线 1 上。这是用一个能证的弱命题换取通往强命题路上的一段实在进展。

同一种姿势横跨到软件里，面目立刻熟悉。类型系统（type system）不证明程序「全对」，它只证明某一条性质（不会把整数当指针解引用），换来的是可判定的检查。形式化验证（formal verification）走得更远，黑尔斯团队在 2017 年完成了开普勒猜想（Kepler conjecture）的机器可核对证明[25]，把一个连人类裁判都吵了多年的论证压成逐

行可验的证书。代价向来一致：证书买到的是一个切片上的确定，赌上的是普遍性，切片不是定理。于是有人转而去动那个目标本身。

代理替换

第二种姿势：别再死守原来的命题，把它换成一个等价但更好对付的陈述。

RH 的等价改写多得惊人。李建军 1997 年给出一个判据[14]：RH 成立，当且仅当一列实数 $\lambda_n \geq 0$ 对所有 $n \geq 1$ 成立，其中

$$\lambda_n = \sum_{\rho} \left[1 - \left(1 - \tfrac{1}{\rho} \right)^n \right],$$

求和取遍非平凡零点。一个关于零点位置的几何陈述，被翻译成一列数的正性。奈曼与博伊林给出另一个：RH 等价于示性函数 $\chi_{(0,1)}$ 落在一族被伸缩的小数部分函数张成的 $L^2(0,1)$ 闭包里，把零点问题翻译成一个逼近问题；巴埃斯-杜阿尔特 2003 年把它收紧成只用整数伸缩的序列版本[16]，对应一列距离 $d_n \to 0$。拉加里亚斯 2002 年甚至给出一个初等到能写在明信片上的等价[32]：对所有 n，$\sigma(n) \leq H_n + \exp(H_n) \ln H_n$，其中 H_n 是调和数，σ 是因子和。

我自己也在这条路上走过一程。把 RH 搬进李判据（Li's criterion）、搬进奈曼-博伊林-巴埃斯-杜阿尔特的逼近框架、再搬进算子谱与随机过程的语言，每一次都怀着同一个指望：换一种语言，也许困难就在新坐标里露出一个把手。每一次得到的结论也都一样，等价是真的，难度一分没少。我没有把问题解开，我只是把它改了个名字换了身衣服。

这就是代理替换在数学里的标准败法，值得给它起个准确的名字：一个忠实却不更易的代理。等价保证了它指向的还是同一个真目标（忠实），可它一点没比原问题更可解（不更易）。这一招的成败全压在能不能同时拿到忠实与更易这两样上，而二者兼得极其罕见。罕见到，这正是全部的手艺所在。

把这两个维度摊成一张表，本章埋的一根线就显出来了：

	更易	不更易
忠实	理想代理（罕见，手艺全在此）	数学的等价改写：你只是把困难改了名（本章）
不忠实	Goodhart：你优化代理，真目标却烂掉（第 8、11 章）	无用，没人会要

数学家栽在左下到右上的对角线左端：忠实但不更易。后面看组织那一章，会栽在另一端：一个更易却不忠实的代理，你拼命优化它，真正在意的东西反而坏掉。同一招，两个相反的失效方向。第 11 章会把这两端正式对接。眼下记住一点就够：换目标不是作弊，也不是出路，它是一种姿势，成不成另说。

概率方法

第三种姿势最违反数学的本能：不再要一个二值的判决，转而握住一个标定（calibration）的概率，接受有界的出错风险去行动。

素性检验（primality test）是最干净的例子。要判断一个大数是不是素数，确定性算法代价高昂，米勒-拉宾（Miller-Rabin）换了个问法。若 n 是合数，一个随机选取的底至多以 $1/4$ 的概率被它蒙混过去，独立做 k 轮则误判概率降到 $\leq (1/4)^k$。索洛维-施特拉森更早给出误差 $\leq (1/2)^k$ 的版本[20]（1977 年），拉宾的版本是 1980 年[19]。「以 $1-\varepsilon$ 的概率为素数」是一个和「已证明为素数」根本不同的认识对象，但工程上足够好，而且可以精到任意程度，加几轮就是了。

要紧的是看清这里到底放弃了什么。放弃的是「确定性的种类」，不是「严格性」。那个 $(1/4)^k$ 的界本身是一条定理，证得严严实实。你没有降低标准，只是换了一种能在预算内交付的标准。概率方法在纯数学里同样登堂入室，埃尔德什的概率方法（Alon 与 Spencer 写成了一本经典[26]）能证明某个对象存在，靠的是证它出现的概率为正，却不把对象具体给你。存在性被证明了，构造却缺席。

这种姿势一路通到 RH 的信念本身。蒙哥马利 1973 年研究零点的配对关联（pair correlation）[30]，得出并猜想归一化后的零点对关联函数形如

$$R_2(u) = 1 - \left(\frac{\sin \pi u}{\pi u}\right)^2.$$

戴森在普林斯顿一眼认出，这正是随机矩阵高斯酉系综（Gaussian unitary ensemble, GUE）本征值的配对关联。奥德利兹科 1987 年用海量零点做数值[33]，把这个吻合验得令人窒息。这些都不是证明，却是极强的证据，让数学家相信 RH，相信的方式和物理学家相信一条尚未被证伪的定律没有本质区别。数学的内部竟也长出了一套在不可验证中形成信念的办法。

数学家怎么判断

于是来到这一章真正的人的部分：当神谕永远不来，有判断力的人靠什么持有信念、决定往哪使劲。

数学对外是演绎的，对内是似真的。波利亚专门写了《数学与似真推理》[2]，讲数学家如何在没有证明时凭类比、归纳、特例去掂量一个命题的分量；阿达马调查了数学发明的心理[3]，记下酝酿与顿悟的节律；庞加莱留下了那个著名的瞬间[5]，踏上公共马车踏板的刹那，富克斯函数与非欧几何的联系毫无征兆地涌上心头。这些不是证明的替代品，而是证明之前那段没有仪表的路上，人实际依赖的东西。

为什么数学家在证明出现之前就相信 RH？因为证据在各个方向上累积并互相印证：海量零点验在线上，众多等价形式没有一个垮掉，它的类比版本已被攻克（韦伊猜想（Weil conjectures），即函数域上的黎曼假设，被德利涅证明），统计行为精确符合随机矩阵的预言。没有哪一条是证明，合在一起却构成一种有纪律的信念。

数学界自己也反思过这种信念的地位。瑟斯顿 1994 年的《论数学中的证明与进展》[21] 主张，数学推进的是人类的理解，而不只是形式证明的库存；贾菲与奎因 1993 年那场关于「理论数学」的争论[22]，

正是在问猜想驱动、证据先行的工作在多大程度上算数学。陶哲轩追问什么是好的数学[29]，答案里没有一条是「已被证明」。把这些放在一起看，一个判断「某命题多半为真、值得投入」的能力本身就是一种标定的信念，而这恰是第四部要正面命名的东西。一个有能力的主体在神谕缺席时并不瘫痪，也不假装确定，他持有一个标好刻度的信念，然后照样行动。

这一章通向哪里

撞上最纯的不可验证，数学家没有等到一个判定程序。他得到的是：证书（验一个切片、证一个界），代理替换（把目标换成等价陈述，并坦白承认它往往忠实却不更易），概率接受（放弃二值判决、握住标定的概率行动），以及在证明之前持有信念的判断力。

这些都不是数学独有的应急手段。把不受信的代码关进沙箱、给一个组织做审计、在界面背后揣摩用户没说出口的偏好，伸手去够的是同样这几样东西，只是换了行话。第三部会把每一招从它生长的领域里拎出来，单独命名、跨域并置，那张对照表就是全书的载荷。

还有一句话留在这里。这本书自己的核心命题，那个「应对会收敛到同一小套」的论断，我此刻也无法验证。我对它的相信和数学家对 RH 的相信是同一种东西：一个建立在跨域证据上、标好了刻度、却没有证明的信念。第 14 章会回到这件事，并让这本书亲自去做它一直在描述的那件事。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. G. Polya (1945). 《How to Solve It: A New Aspect of Mathematical Method》. Princeton University Press. [①] 波利亚把数学解题拆成理解题意、拟订计划、执行、回顾四个阶段，

并系统列出类比、特例、逆推、辅助问题等启发式策略。它写的不是定理证明，而是发现证明之前那段没有仪表的探索过程，正是本章「数学家怎么判断」一节关心的东西。

2. G. Polya (1954).《Mathematics and Plausible Reasoning》(2卷). Princeton University Press. [①④] 两卷分别讨论数学中的归纳与类比，以及似真推理的逻辑结构，论证数学家在拿到严格证明之前，靠观察特例、归纳模式、权衡证据来形成对命题的信念。本章正文借它点明「数学对外是演绎的，对内是似真的」，是理解似真推理这一概念的源头读物。
3. J. Hadamard (1945).《An Essay on the Psychology of Invention in the Mathematical Field》. Princeton University Press. [①] 阿达马调查了数学家的创造心理，提炼出准备、酝酿、顿悟、验证的发现节律，强调潜意识工作与无预兆的灵感闪现。它为本章描述庞加莱式的顿悟提供了第一手的心理学考察，说明数学判断很大一部分发生在意识与证明之外。
4. H. Poincaré (1902).《La Science et l'Hypothèse》. Flammarion. [①] 庞加莱在这部科学哲学经典里讨论数学假设、约定与几何的地位，主张许多基础选择并非经验强加，而是出于约定与方便。它呈现了一位顶尖数学家如何反思自己学科的认识论根基，与本章关心的「在无法验证处如何持有信念」相通。
5. H. Poincaré (1908).《Science et Méthode》. Flammarion. [①] 书中那段关于踏上公共马车踏板时灵感涌现的自述，是数学发现心理学最常被引用的第一手记录，庞加莱借此剖析直觉、选择与潜意识在创造中的作用。本章正文直接用到这个瞬间，说明顿悟如何在没有任何征兆时把分散的线索接通。
6. G. H. Hardy (1940).《A Mathematician's Apology》. Cambridge University Press. [①] 哈代为纯数学的价值辩护，提出好的数学在于其严肃性、深刻与不可避免的美，而非实用。作为一位数论大家对自己手艺的内省，它界定了数学家凭什么判断一项工作值不值得做，与本章末尾追问「什么是好的数学」一脉相承。
7. E. P. Wigner (1960).「The Unreasonable Effectiveness of Mathematics in the Natural Sciences」. Communications on Pure and Applied Mathematics, 13(1). [②③] 维格纳惊

讶于抽象数学概念竟能如此精准地描述物理世界，称这种契合是一份我们既不理解也不配拥有的奇异礼物。这篇短文提出的难题，至今没有公认答案，对本章而言它示范了一种对深层规律的信念如何在缺乏证明的情况下被严肃对待。

8. I. Lakatos (1976). 《Proofs and Refutations: The Logic of Mathematical Discovery》. Cambridge University Press. [①③] 拉卡托斯以欧拉多面体公式为例，用一段虚构的课堂对话重演定义、证明与反例如何彼此修正、共同推进数学。它颠覆了数学是一锤定音演绎的刻板印象，呈现知识在猜想与反驳中曲折成长，正合本章对数学进展真实样貌的关注。
9. B. Riemann (1859).「Über die Anzahl der Primzahlen unter einer gegebenen Größe」. Monatsberichte der Berliner Akademie. [②③] 黎曼这篇仅八页的论文把 ζ 函数延拓到复平面、给出函数方程，并把素数分布与 ζ 的非平凡零点联系起来，其中顺带写下的那句关于零点位置的猜想，就是后世的黎曼假设。它是整章的源头文本，本章开篇引用的「徒劳尝试后暂时搁下」正出自此处。
10. H. M. Edwards (1974). 《Riemann's Zeta Function》. Academic Press. [②] 爱德华兹这本专著围绕黎曼 1859 年原文展开，逐步铺陈 ζ 函数理论、素数定理与黎曼假设的来龙去脉，兼顾历史脉络与技术细节。它是进入 ζ 函数与 RH 的经典入门读物，为本章涉及的零点、临界线等概念提供了可靠的背景。
11. E. C. Titchmarsh, rev. D. R. Heath-Brown (1986). 《The Theory of the Riemann Zeta-function》（第 2 版）. Oxford University Press. [②] 这是 ζ 函数解析理论的标准高阶专著，系统处理零点分布、零点密度估计、临界线上的均值定理等结果，希思布朗的修订补入了更晚近的进展。它代表了围绕 RH 已被严格建立的技术成果的总和，是本章谈「证书与界」时的专业背景文献。
12. E. Bombieri (2000).「Problems of the Millennium: The Riemann Hypothesis」. Clay Mathematics Institute. [②④] 这是克雷数学研究所为千禧年大奖问题撰写的 RH 官方问题陈述，邦别里精炼地交代了猜想的来历、精确表述及其在数论

中的份量。它是了解 RH 为何被列为世纪难题的权威切入点，本章对 RH 地位的判断可在此找到背书。

13. J. B. Conrey (2003).「The Riemann Hypothesis」. Notices of the American Mathematical Society, 50(3). [②③] 康里这篇综述面向广泛读者，梳理了支持 RH 的各类证据，包括零点的数值验证、随机矩阵理论的吻合，以及函数域上类比的已被证明。它把本章三种姿势所依赖的证据汇于一处，是了解数学界为何相信 RH 的便捷读物。
14. X.-J. Li (1997).「The Positivity of a Sequence of Numbers and the Riemann Hypothesis」. Journal of Number Theory, 65(2). [②④] 李建军证明 RH 等价于一列由零点定义的实数 λ_n 对所有 n 非负，把零点位置这一几何陈述翻译成一个序列的正性判据。本章正文以它为代理替换的头号例子，说明等价改写如何忠实却未必更易。
15. E. Bombieri & J. C. Lagarias (1999).「Complements to Li's Criterion for the Riemann Hypothesis」. Journal of Number Theory, 77(2). [②④] 两位作者指出李判据其实是任意复数多重集一组一般不等式的特例，并不特属于 ζ 函数，又借古伊南-韦伊显式公式给出 λ_n 的算术表达式，把它与韦伊的 RH 判据接上。它深化了对李判据的理解，呈现同一个等价命题如何在不同语言间被反复重写，正是本章代理替换主题的延展。
16. L. Báez-Duarte (2003).「A Strengthening of the Nyman-Beurling Criterion for the Riemann Hypothesis」. Atti della Accademia Nazionale dei Lincei, Rendiconti Lincei Mat. Appl., 14(1). [②④] 巴埃斯-杜阿尔特把奈曼-博伊林的逼近判据收紧为只用整数伸缩的序列版本，使 RH 等价于一列逼近距离 d_n 趋于零。它是本章列举的又一个等价改写，把零点问题搬进 L^2 逼近的框架，同样印证了忠实代理常常并不更易求解。
17. X. Gourdon (2004).「The 10^13 First Zeros of the Riemann Zeta Function, and Zeros Computation at Very Large Height」. 在线技术报告（numbers.computation.free.fr）. [②④] 古尔东借助 Odlyzko-Schönhage 算法用高精度浮点计算，核验了头 10^{13} 个零点都落在临界线上。本章特意拿它与

普拉特对照：它给出极强的数值信心，但未严格框住舍入误差，因而是数值结果而非可机械复核的证书。

18. D. J. Platt (2017).「Isolating Some Non-trivial Zeros of Zeta」. Mathematics of Computation, 86(307). [②④] 普拉特用区间算术把零点严格隔离在临界线上，使误差有可证的上界，从而把数值核验升格为可机械复核的证书。本章用它示范「证书与界」这一姿势：不证整体，只为一个有限切片证一个有保证的界。
19. M. O. Rabin (1980).「Probabilistic Algorithm for Testing Primality」. Journal of Number Theory, 12(1). [②④] 拉宾给出米勒-拉宾素性检验：若 n 为合数，随机选取的底至多以 $1/4$ 的概率瞒过它，独立做 k 轮误判概率降到 $(1/4)^k$。本章用它作为概率方法最干净的例子，说明那个误差界本身是被严格证明的定理，放弃的是确定性的种类而非严格性。
20. R. Solovay & V. Strassen (1977).「A Fast Monte-Carlo Test for Primality」. SIAM Journal on Computing, 6(1). [②④] 索洛维与施特拉森更早提出一个基于雅可比符号的概率素性检验，单轮误判概率至多 $1/2$，是随机化算法的奠基工作之一。本章把它与拉宾的版本并置，说明用概率方法换取在预算内可交付的判定，在计算数论里早有先例。
21. W. P. Thurston (1994).「On Proof and Progress in Mathematics」. Bulletin of the American Mathematical Society, 30(2). [①③] 瑟斯顿主张数学真正推进的是人类对数学的理解，而不只是形式证明的库存，证明只是社群传递与确认理解的一种社会化手段。本章末尾援引它来挑战「数学只等于已证定理」的窄化看法，是反思证明地位的必读文献。
22. A. Jaffe & F. Quinn (1993).「"Theoretical Mathematics": Toward a Cultural Synthesis of Mathematics and Theoretical Physics」. Bulletin of the American Mathematical Society, 29(1). [①③] 贾菲与奎因提出区分「理论数学」与严格数学，建议把猜想驱动、未经严格证明的工作明确标注出来，以免侵蚀数学的可靠性，由此引发数学界一场广受关注的争论。本章借这场争论提问：证据先行的工作在多大程度上算数学，正切中全书对验证与信念的关切。

23. J. von Neumann (1947).「The Mathematician」. 收于 R. B. Heywood（编）,《The Works of the Mind》. University of Chicago Press. [①③] 冯 · 诺依曼在这篇随笔里反思数学的本性，谈数学如何在抽象与经验源头之间往返，又如何凭审美标准选择方向以及为何远离经验源头会有退化的风险。它从一位横跨多领域的大家视角，说明数学判断中审美与品味的分量，呼应本章对数学家如何决定往哪使劲的讨论。
24. K. Appel & W. Haken (1977).「Every Planar Map Is Four Colorable, Part I: Discharging」. Illinois Journal of Mathematics, 21(3). [②③] 阿佩尔与哈肯借助大量计算机检查的不可避免构形集，证明了四色定理，这是首个本质依赖计算机的著名数学证明。它引出了一个延续至今的争论：人类无法逐行通读的证明是否仍算证明，与本章对证书与可机械复核保证的讨论直接相关。
25. T. Hales et al. (2017).「A Formal Proof of the Kepler Conjecture」. Forum of Mathematics, Pi, 5. [②③④] 黑尔斯团队的 Flyspeck 项目用 HOL Light 与 Isabelle 证明助手，完成了开普勒猜想的完全形式化、可机械核对的证明，了结了原证明因人工裁判难以彻底检验而悬而未决的状态。本章用它说明形式化验证如何把有争议的论证压成逐行可验的证书。
26. N. Alon & J. H. Spencer (1992).《The Probabilistic Method》. Wiley. [②④] 这本经典系统呈现了埃尔德什开创的概率方法：要证某个组合对象存在，便证它随机出现的概率为正，从而断定它必然存在，却往往无法把它具体构造出来。本章借它点出概率方法在纯数学中证明存在性时存在被证明、构造却缺席的特征。
27. P. J. Davis & R. Hersh (1981).《The Mathematical Experience》. Birkhäuser. [①③] 戴维斯与赫什从数学家的实际经验出发，讨论数学对象的存在地位、证明的角色与数学的哲学处境，呈现了一种不同于形式主义教条的从业者视角。它为本章理解数学家如何在实践中持有信念、看待真理提供了贴近现场的反思。
28. W. T. Gowers (2000).「The Two Cultures of Mathematics」. 收于《Mathematics: Frontiers and Perspectives》. American

Mathematical Society. [①③] 高尔斯区分数学中的两种文化：理论构建者与问题解决者，前者以理解为目的去解题，后者以解题为目的去理解，并以代数几何、朗兰兹纲领与组合数论为对照。它说明数学家对何谓深刻、何谓好工作可有不同尺度，呼应本章对数学判断标准的讨论。

29. T. Tao (2007).「What Is Good Mathematics?」. Bulletin of the American Mathematical Society, 44(4). [①③] 陶哲轩列举了好数学的众多互不相同的维度，从严格、深刻、漂亮到富于应用、能开辟方向等，论证不存在单一标准，且这些品质长期看往往彼此牵引。本章借它说明判断一项工作值不值得投入本身就是一种能力，其中没有一条标准是已被证明。
30. H. L. Montgomery (1973).「The Pair Correlation of Zeros of the Zeta Function」. 收于《Analytic Number Theory》, Proc. Sympos. Pure Math., XXIV. American Mathematical Society. [②③] 蒙哥马利研究 ζ 零点归一化后的配对关联，得出并猜想其形式，戴森随即认出这正是随机矩阵高斯酉系综本征值的配对关联。这一对接开启了数论与随机矩阵理论的深刻联系，是本章谈 RH 信念何以建立的关键证据之一。
31. P. Sarnak (2004).「Problems of the Millennium: The Riemann Hypothesis」. Clay Mathematics Institute. [②③④] 萨纳克为克雷研究所撰写的这份说明侧重 RH 的推广形式及其在解析数论中的中心作用，并解释为何众多其他结果以它为前提。它从一位活跃于零点统计与随机矩阵联系的专家视角，补足了本章对 RH 重要性与证据网络的理解。
32. J. C. Lagarias (2002).「An Elementary Problem Equivalent to the Riemann Hypothesis」. The American Mathematical Monthly, 109(6). [②④] 拉加里亚斯给出一个仅用调和数与因子和函数的初等不等式，证明它对所有 n 成立当且仅当 RH 成立，把深奥的零点问题翻译成几乎能写在明信片上的算术陈述。本章用它示范代理替换可以表面初等，难度却分毫未减。
33. A. M. Odlyzko (1987).「On the Distribution of Spacings Between Zeros of the Zeta Function」. Mathematics of Computation, 48(177). [②④] 奥德利兹科用海量高精度计算考察

ζ 零点的间距分布，发现它与随机矩阵高斯酉系综的预言惊人吻合，为蒙哥马利-戴森的猜想提供了强有力的数值支持。本章用它说明这种统计契合虽非证明，却是让数学家相信 RH 的极强证据。

第 8 章　看不见自己的组织

> **论点**：一个大型组织或国家无法直接观察自己那些分布的、部分隐藏的、有时被策略性掩盖的知识与活动，于是它伸手去抓代理（它能看见的指标），这正是代理这一招的 Goodhart 败法最触目的地方，再以审计（留痕）与冗余补足。

一个看不清自己的庞然大物

前一章那个委托代理（principal-agent）的难题，现在把尺度放大。委托方不再是一个人，而是一整个组织、一个国家；被委托去行动的，是成千上万散落在各处的人。一个新的、几乎荒诞的局面出现了：这个庞然大物，看不清自己。

它想知道自己有多少人、种着什么、谁在做什么、做得好不好，可这些知识不在任何一个能被它直接读取的地方。本章要看的是：当被验证的对象是组织自身那些分布的、隐藏的、会主动躲闪的知识时，组织怎么办。这里的不可验证，把前面几种处境叠在了一起：部分可观测（partial observability，知识分散在边缘），加对抗（被看的人会反过来操纵被看的东西）。

分布的知识

哈耶克 1945 年那篇《知识在社会中的运用》[1] 把问题挑明了：一个社会赖以运转的知识，从不集中在任何一处，它分散在无数个体手

里，是关于特定时间、特定地点的局部知识（local knowledge），往往还说不清、道不明。哪个机器今天有点小毛病、哪个客户其实快要流失、哪条小路雨后会塌，这些知识的拥有者常常自己都没意识到它是「知识」，更没法把它打包上交给中心。波兰尼把这一层叫默会维度（tacit dimension）[2]：我们知道的，远多于我们说得出的。

这意味着组织面对的不只是「信息暂时没收上来」。哪怕每个人都忠诚配合，那些局部的、默会的知识在汇集的过程中也会蒸发。中心想要的那张「组织全貌」，原则上无法被如实地装进任何能验证它的容器里。这是部分可观测在社会尺度上的版本，而且带着一道更硬的底线：那些知识本性上就是局部的，无法被汇集到任何一处。

可读性的冲动

看不清，就想办法让它变得可看。斯科特 1998 年的《国家的视角》[3] 给这种冲动起了个准确的名字：可读性（legibility）。国家要在社会上行动，必先把社会改造成它读得懂的样子。它丈量土地、画出地籍图，给本来只有小名、绰号、随父名的人强加固定的姓氏，统一度量衡，推行标准化的科学营林。这些不是中性的记录工作，它们是在重塑现实本身，好让现实纳入表格。哈金的《驯服偶然》[32]、德罗西埃的《大数字的政治》[31]、鲍克与斯塔尔的《分类及其后果》[30]，合起来是一部「把社会变得可数」的历史。

可读性的危险，在于那张地图必然简化，而组织一旦只照着地图行动，被地图抹掉的东西就会反噬。斯科特最有力的案例正是科学营林（scientific forestry）：为了让森林「可读、可算、可收税」，普鲁士人把杂乱的天然林改造成整齐划一、便于清点的单一树种人工林，头一两代长势喜人，到第三代，土壤耗竭、虫害蔓延，森林成片死亡，德语里甚至造出了一个词，Waldsterben，森林之死。地图越是干净，它抹掉的那些维系系统运转的局部知识就越致命。这是组织在为自己制造它所缺的可观测性，代价是亲手削平了让它得以运转的复杂。

代理指标，和它的 Goodhart 崩塌

可读性冲动最常见的落点，是指标。真正在意的东西，健康、学习、生产力、公共福祉，无法直接观测；于是组织一把抓住它能看见的代理（proxy），KPI、GDP、考试分数、论文引用数、急诊等待时长。

这正是我们在第 7 章见过的代理替换。但它在这里以相反的方式失效，而这个对照是本书的一根主线。数学家的代理是忠实却不更易：等价改写真的等价，但解题时没有变得更简。组织的代理恰好反过来，更易却不忠实：指标当然好测，可它与真目标之间的对应，一旦指标本身成为目标，就会断裂。

这个断裂有许多名字。古德哈特 1975 年[4]：一旦一个指标被当作政策目标，它作为指标的可靠性就会瓦解。坎贝尔 1979 年[6] 说的是同一件事的社会版。早在 1956 年，里奇韦就编目过「绩效度量的失能后果」[7]；克尔 1975 年那篇《奖励 A 却指望 B 的蠢事》[8] 把它写成了管理学的常识。斯特拉森给了它最精炼的一句[5]：当一个度量变成目标，它就不再是个好度量。更深的一层是反身性（reflexivity）：埃斯佩兰与索德尔 2007 年[12] 指出，公开的排名不是在描述世界，而是在重造世界，被排名的大学会照着排名的算法改变自己，于是指标「测量」的对象，恰恰是它自己催生出来的行为。贝文与胡德[11] 记录了英国医疗系统里针对指标的博弈，史密斯 1995 年[10] 分析了公开发布绩效数据如何招来一连串始料未及的后果，默顿 1936 年那篇关于「有目的社会行动的非预期后果」[9]，是这一切的总源头。这类崩塌在现实里屡见不鲜，代价有时惊人。富国银行为冲「交叉销售」的账户指标，员工在客户不知情下私开了约三百五十万个虚假账户，2016 年事发，银行被罚一亿八千五百万美元、逾五千名员工遭解雇，那个被供起来的数字，恰恰摧毁了它本要衡量的客户关系。更早一则寓言式的案例发生在殖民时期的德里：当局为灭蛇悬赏眼镜蛇尸体，市民索性养蛇来领赏；赏金一停，蛇被尽数放生，蛇患反比从前更重，「眼镜蛇效应」（cobra effect）由此得名。

为什么代理一定会被扭曲？委托代理理论给了严格的解释。霍姆斯特伦 1979 年的信息性原理[14]（informativeness principle）：报酬应当挂靠在对「努力」有信息量的信号上。可一旦努力是多维的，而

你只测得到其中几维，麻烦就来了。霍姆斯特伦与米尔格罗姆 1991 年的多任务分析[15]（multitask principal-agent）说得明白：当一个人要同时兼顾可测与不可测的任务，越是重奖可测的那部分，他就越会把努力从不可测的部分转向可测的部分。设真目标为 G，可观测代理为 P，二者在现状下相关；问题是这种相关乃行为的产物，而非客观规律。一旦以 P 为施压目标，

$$\arg\max_a P(a) \quad \text{vs.} \quad \arg\max_a G(a),$$

理性的代理人就会去找那些抬高 P 却无助、甚至有损 G 的行动，相关被优化压力本身碾碎。教师教应试、医院调度病人去压低某一项等待时长、研究者把一篇论文切成可计数的最小发表单元，都是同一个机制。

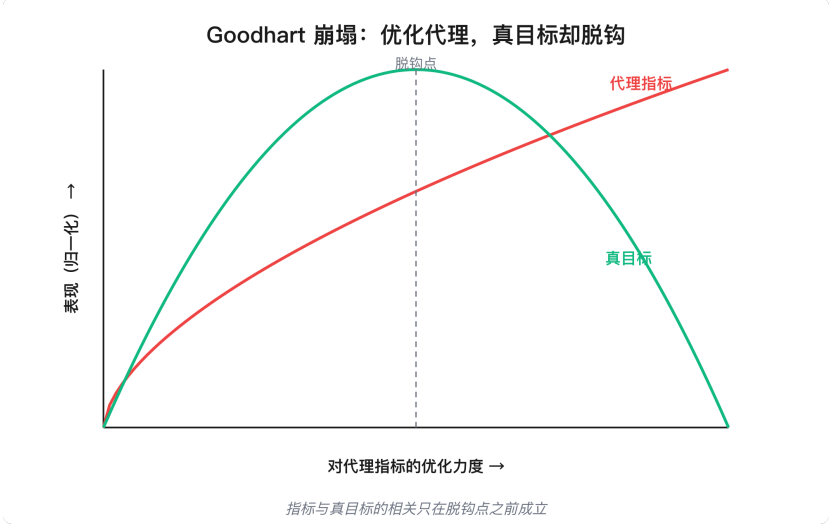

图 6: Goodhart 崩塌：优化代理，真目标却脱钩

用审计与冗余补足

代理单靠自己会塌，于是组织补上另外两招，这也是本书反复出现的招数。

留痕与审计。复式记账是人类最古老的审计链之一，索尔在《清算》[22] 里论证，可核账目的能力，与一个个国家的兴衰直接相关：算得清自己的，方能持久。现代的财务审计、独立稽核，都是把「事前防不住舞弊」换成「事后查得出舞弊」。但这一招有它自己的病。鲍尔 1997 年的《审计社会》[20]（audit society）点破：当验证本身变成仪式，组织生产的不过是「一切尽在掌握」的表象，而非掌握本身。肖尔与赖特笔下的「审计文化」[17]（audit culture）、奥尼尔在 2002 年里斯讲座中对「信任」[21] 的反思，讲的都是同一种异化：为了可被问责，机构把大量精力耗在制造可供检查的痕迹上，真正的工作反被挤到一边。

冗余与共识。兰道 1969 年那篇被低估的文章[16] 为「重复与重叠」正名：在一个零件都不完全可靠的系统里，冗余（redundancy）不是浪费，而是可靠性的来源，多个互相独立的核查，比单一权威更难被同时骗过。这一招的成立有个前提，独立，下一部会反复强调：若几个核查其实同源，相关的失效会一举摧毁冗余的全部价值。

这一章通向哪里，以及第二部的收束

到此，四个现场看完了。控制台前的人、放出去的智能体、撞墙的数学家、看不见自己的组织，它们面对的不可验证来源迥然不同：藏在心里的偏好、开放世界里的未来行为、原则上不可判定的命题、分布且会躲闪的知识。可它们伸手去够的，是同一小套东西。

最该并排摆出来的，是代理替换的两种相反败法。数学家栽在忠实却不更易，组织栽在更易却不忠实。第 7 章那张 2×2 表的两端，现在都有了血肉。它们不是两招，是同一招的两个失效方向，而一个好代理必须同时躲开这两端，既忠实又更易，这罕见到几乎就是全部的手艺。第 11 章会正式把这两端对接。委托代理那个骨架，也从第 6 章的一段代码，长成了这里的一个国家。

第二部到此把招数都「嵌在现场里、彼此缠绕」地演示了一遍。它们零散、换着名字、混在各自的行话里。第三部要做的，是把每一招从它生长的领域里拔出来，洗净，单独命名，一次性涵盖所有现场。那张对照表，是这本书真正的载荷。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. F. A. Hayek (1945).「The Use of Knowledge in Society」.《American Economic Review》, 35(4), 519-530. [②④] 哈耶克论证，社会运转所依赖的知识从不集中于一处，而是分散在无数个体手中，是关乎特定时间、特定地点的局部知识，无法被任何中心如实汇集。这篇文章是本章「分布的知识」一节的直接出发点，也奠定了「组织看不清自己」这一困境的认识论底色。
2. M. Polanyi (1966).《The Tacit Dimension》. Doubleday. [②] 波兰尼提出知识的默会维度，其名言是「我们知道的，远多于我们说得出的」。本书用它来说明，分散在边缘的局部知识中有相当一部分根本无法被言说和上交，这是组织难以验证自身的更硬的一层底。
3. J. C. Scott (1998).《Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed》. Yale University Press. [②④] 斯科特提出「可读性」这一概念：国家为了在社会上行动，会用地籍图、固定姓氏、统一度量衡等手段把社会改造成自己读得懂的样子，而这种简化往往抹掉维系系统运转的局部知识，导致科学营林那样的失败。本章「可读性的冲动」一节正建基于此，它是理解组织为何要亲手削平复杂性的核心读物。
4. C. A. E. Goodhart (1975).「Problems of Monetary Management: The U.K. Experience」.《Papers in Monetary Economics》, Vol. I. Reserve Bank of Australia. [②] 古德哈特原本谈的是货币政策，却给出了后来被普遍引用的洞见：一旦某个统计规律被当作政策调控的目标，它原有的规律性就会瓦解。这就是本章「代理指标，和它的 Goodhart 崩塌」一节

的命名来源，是理解代理被优化压力碾碎的起点。

5. M. Strathern (1997).「Improving Ratings: Audit in the British University System」.《European Review》, 5(3), 305-321. [②④] 斯特拉森借英国大学审计的经验，给古德哈特定律留下了最精炼的一句通俗表述：当一个度量变成目标，它就不再是个好度量。本章直接引用了这句话，它也是把抽象的代理崩塌讲给读者听的最佳一句话。
6. D. T. Campbell (1979).「Assessing the Impact of Planned Social Change」.《Evaluation and Program Planning》, 2(1), 67-90. [②④] 坎贝尔从社会科学评估的角度提出了与古德哈特同构的「坎贝尔定律」：一个量化的社会指标越是被用于社会决策，它就越容易遭受腐蚀压力，也越会扭曲它本要监测的社会过程。本章用它佐证代理崩塌并非经济学独有，而是跨学科反复被发现的同一现象。
7. V. F. Ridgway (1956).「Dysfunctional Consequences of Performance Measurements」.《Administrative Science Quarterly》, 1(2), 240-247. [②④] 里奇韦很早就系统编目了绩效度量的失能后果，区分了单一度量、复合度量和多重度量各自带来的扭曲。本章用它说明，针对指标的博弈与扭曲是一个被发现得相当早的老问题，而非晚近才有的管理学新词。
8. S. Kerr (1975).「On the Folly of Rewarding A, While Hoping for B」.《Academy of Management Journal》, 18(4), 769-783. [②④] 克尔列举了大量现实例子，说明组织常常奖励一种行为，却指望得到另一种它没有奖励的行为，结果自然事与愿违。这篇文章把代理与激励的错配写成了管理学的常识，是本章「奖励 A 却指望 B」这一机制的经典出处。
9. R. K. Merton (1936).「The Unanticipated Consequences of Purposive Social Action」.《American Sociological Review》, 1(6), 894-904. [②④] 默顿系统分析了有目的的社会行动为何总会带来未曾预料的后果，并梳理了无知、误判、价值偏好等成因。本章把它视为指标博弈、可读性反噬等一系列「始料未及」现象的总源头。
10. P. Smith (1995).「On the Unintended Consequences of Publishing Performance Data in the Public Sector」.《Interna-

tional Journal of Public Administration» , 18(2-3), 277-310. [②④] 史密斯分类梳理了公共部门公开发布绩效数据所招致的一连串非预期后果，如隧道视野、近视、目标错位、衡量固化、博弈等。本章用它把笼统的「指标被扭曲」拆成可辨认的若干种具体失效方式。

11. G. Bevan & C. Hood (2006).「What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System」.《Public Administration» , 84(3), 517-538. [②④] 贝文与胡德实证记录了英国国民医疗体系在「目标加恐吓」治理下针对指标的种种博弈，例如调度病人以压低等待时长这类应付指标却无助于真实健康的做法。本章以它为指标博弈在公共服务中如何具体发生的现场证据。
12. W. N. Espeland & M. Sauder (2007).「Rankings and Reactivity: How Public Measures Recreate Social Worlds」.《American Journal of Sociology» , 113(1), 1-40. [②④] 埃斯佩兰与索德尔以法学院排名为例，提出「反身性」：公开的度量不只是描述世界，还会反过来重塑被度量者的行为，使指标最终测量的是它自己催生出来的反应。本章「更深的一层是反身性」一段正出自此，它把代理失效推进到指标制造现实这一层。
13. M. Sauder & W. N. Espeland (2009).「The Discipline of Rankings: Tight Coupling and Organizational Change」.《American Sociological Review» , 74(1), 63-82. [②④] 这篇姊妹篇借福柯的规训概念分析排名如何嵌入组织：原本松散耦合的机构在排名压力下被迫紧密耦合，外部度量内化为日常的自我监控与组织变革。它与前一条互补，前者讲反身性机制，本条讲排名怎样改造组织的内部结构。
14. B. Holmström (1979).「Moral Hazard and Observability」.《The Bell Journal of Economics» , 10(1), 74-91. [②④] 霍姆斯特伦提出信息性原理：在道德风险下，最优报酬契约应当挂靠在对代理人努力有信息量的全部信号上。本章用它为「代理为何注定被扭曲」给出严格的委托代理解释，并引出当努力多维而只测得几维时的麻烦。
15. B. Holmström & P. Milgrom (1991).「Multitask Principal-

Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design」. 《The Journal of Law, Economics, and Organization》, 7(Special Issue), 24-52. [②] 多任务委托代理模型说明，当一个人要同时兼顾可测与不可测的任务时，越是重奖可测的那部分，他就越会把努力从不可测的部分抽走。本章正是以此论证代理崩塌的机理：照可观测指标施压，会理性地诱使代理人放弃难以衡量却真正重要的工作。

16. M. Landau (1969).「Redundancy, Rationality, and the Problem of Duplication and Overlap」. 《Public Administration Review》, 29(4), 346-358. [②④] 兰道为常被斥为浪费的「重复与重叠」正名：在零件都不完全可靠的系统里，冗余正是可靠性的来源，多个相互独立的核查比单一权威更难被同时骗过。本章「用审计与冗余补足」一节直接采用这一论点，并强调它成立的前提是各核查彼此独立。
17. C. Shore & S. Wright (1999).「Audit Culture and Anthropology: Neo-Liberalism in British Higher Education」.《The Journal of the Royal Anthropological Institute》, 5(4), 557-575. [②④] 肖尔与赖特以英国高等教育为例，提出「审计文化」：新自由主义治理下，问责与审计的逻辑渗透进学术机构，把同行变成被监控对象，重塑了人的自我治理方式。本章用它说明审计如何从工具异化为一种文化，让人耗在制造可供检查的痕迹上。
18. J. Z. Muller (2018). 《The Tyranny of Metrics》. Princeton University Press. [④] 穆勒面向一般读者，梳理了医疗、教育、警务、商业等领域过度依赖量化指标所带来的扭曲与代价，提出何时该用、何时不该用度量的判断。本书是把代理崩塌讲给实践者听的通俗综合之作，适合读者作为入门与对照。
19. T. M. Porter (1995). 《Trust in Numbers: The Pursuit of Objectivity in Science and Public Life》. Princeton University Press. [②④] 波特论证，对量化的依赖往往源于一种「机械的客观性」：在缺乏信任、需对外问责的处境下，数字被用作抑制个人判断、抵御质疑的工具。本书为理解组织为何执着于可读的数字提供了深层的社会学解释，与本章可读性与审计两节互为背景。

20. M. Power (1997). 《The Audit Society: Rituals of Verification》. Oxford University Press. [②④] 鲍尔指出，当验证本身变成一套仪式，组织生产出来的往往是「一切尽在掌握」的表象，而非掌握本身，社会也随之为了可被审计而重塑自己。本章「用审计与冗余补足」一节借它点破审计这一招自带的病：留痕越多，真实工作越被挤到一边。
21. O. O'Neill (2002). 《A Question of Trust: The BBC Reith Lectures 2002》. Cambridge University Press. [④] 奥尼尔在这组里斯讲座中反思当代的问责文化：旨在重建信任的种种透明与审计措施，往往侵蚀了它们本想培育的信任，让人忙于应付检查而非把事做好。本章引它与「审计社会」并列，说明过度问责如何反噬。
22. J. Soll (2014). 《The Reckoning: Financial Accountability and the Rise and Fall of Nations》. Basic Books. [①④] 索尔以复式记账为线索，论证可核账目的能力与一个个国家的兴衰直接相关：算得清自己的，方能持久。本章用它支撑「留痕与审计」是人类最古老的验证链之一这一论断。
23. J. G. March & H. A. Simon (1958). 《Organizations》. John Wiley & Sons. [②] 马奇与西蒙奠定了现代组织理论：组织成员的理性是有限的，组织正是通过分工、程序与信息渠道来应对个体认知能力的局限。本书为「组织看不清自己」提供了基础框架，是理解信息如何在层级中流动与衰减的经典源头。
24. H. A. Simon (1947). 《Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization》. Macmillan. [②] 西蒙提出有限理性，把组织理解为一套帮助成员在认知能力受限的条件下做出决策的结构。本书是理解组织为何必须依赖简化、惯例与代理来运转的源头，为本章组织自我认知的局限奠定了理论底盘。
25. R. M. Cyert & J. G. March (1963). 《A Behavioral Theory of the Firm》. Prentice-Hall. [②] 赛尔特与马奇提出企业的行为理论，强调组织决策受标准操作程序、有限搜索与各方目标协商支配，而非纯粹的最优化。本书有助于理解组织内部目标的多元与张力，是本章把组织看作有限理性主体的重要支撑。

26. O. E. Williamson (1975). 《Markets and Hierarchies: Analysis and Antitrust Implications》. Free Press. [②] 威廉森从交易成本出发，解释了为何有些活动由市场协调、有些则被纳入科层组织：有限理性与机会主义使得某些交易在层级内部完成更有效率。本书为组织为何要把分散的活动收编进自己内部、并因此承担起验证它们的难题提供了经济学解释。
27. K. J. Arrow (1974). 《The Limits of Organization》. W. W. Norton. [②④] 阿罗简练地探讨了组织作为应对信息匮乏与不确定性的手段，以及它在权威、责任与信任上遭遇的内在限度。本书指出信任是社会运转不可或缺却无法靠契约买到的润滑剂，与本章审计与冗余两节探讨的验证成本遥相呼应。
28. M. Lipsky (1980). 《Street-Level Bureaucracy: Dilemmas of the Individual in Public Services》. Russell Sage Foundation. [②④] 利普斯基指出，教师、警察、社工等一线官僚在资源不足的处境下行使大量自由裁量，他们的日常应对实际上塑造了公共政策的真实落地。本书是理解组织边缘的局部知识与裁量为何难以被中心观察和验证的重要参照。
29. J. Q. Wilson (1989). 《Bureaucracy: What Government Agencies Do and Why They Do It》. Basic Books. [②④] 威尔逊详细考察了政府机构的实际运作，区分了产出与结果均可观察与否的不同机构类型，并说明为何许多公共机构的真实成效难以衡量。本书为本章「组织看不见自己」提供了丰富的现实素材，尤其有助于理解为何代理指标在公共部门格外容易失真。
30. G. C. Bowker & S. L. Star (1999). 《Sorting Things Out: Classification and Its Consequences》. MIT Press. [②④] 鲍克与斯塔尔考察了分类系统如何无声地嵌入基础设施，又如何塑造它本想中立记录的现实，被分类抹平的差异往往带来实际后果。本章把它与哈金、德罗西埃并列，归入「把社会变得可数」这部历史，说明分类是可读性工程的隐形一环。
31. A. Desrosières (1998). 《The Politics of Large Numbers: A History of Statistical Reasoning》(trans. C. Naish). Harvard University Press. [②] 德罗西埃梳理了统计推理的历史，说明统计范畴与国家管理同步成形，数字既是认识社会的工具，

也是构造社会现实的政治行为。本章把它纳入「把社会变得可数」的谱系，揭示可读性背后那套统计装置的来历。

32. I. Hacking (1990). 《The Taming of Chance》. Cambridge University Press. [②] 哈金考察了十九世纪统计与概率思想的兴起，论证大量收集人口数据如何「驯服偶然」，催生了「正常」与「常态」等支配现代治理的概念。本章引它说明，让社会变得可数本身就是一段重塑认知的历史，而非中性的记录。

第三部　收敛

第 9 章　压缩未知

> 论点：两招直接攻击不确定性。在你能查的切片上给出有保证的界（证书），将有限的查验花在最能消解不确定的地方（最优筛查）。

第二部把招数嵌在四个现场里、彼此缠绕地演了一遍。从这一部起，换一种看法：把每一招单独拎出来，洗去领域的行话，以纯形式呈现，再一次性铺满所有现场。这才是本书真正的载荷，那张「同一招在各种行话下的对照表」。

八招两两并成四章。配对不是图省事，配对本身是个论点：每一对拉动的是同一根更基本的杠杆，这一点会在第 13 章兑现。本章这一对，证书与最优筛查，共同攻击的是同一样东西，即不确定性。它们从相反的两端去压缩未知：一端是在你查得动的切片上证出一个有保证的界，另一端是将有限的查验花在最能消解不确定的地方。

并且，按第 4 章立下的铁律，每提一招、每做一次跨域并置，都要逼问一遍：这个迁移是实质的（同机制、同败法、同权衡），还是只是个漂亮比方？

证书：在切片上证一个界

第一招的纯形式：不去验证整体，而是产出一个有界的、可独立复核的局部保证。你交付的不是「它全对」，而是「在这个范围内，它至多错这么多」，并附一张任何人都能快速验真的凭证。

它的跨域形态惊人地一致。

机器学习里，它叫泛化界（generalization bound）。瓦利安特 1984 年的 PAC 框架[1] 与瓦普尼克和切尔沃年基斯的 VC 维（布卢默等人 1989 年接入[4]），给出形如「以至少 $1-\delta$ 的概率，真实误差不超过经验误差加一个复杂度惩罚」的保证：

$$R(h) \;\le\; \hat{R}(h) + \sqrt{\frac{d\big(\ln(2n/d)+1\big) + \ln(4/\delta)}{n}}.$$

你验不了模型在未来所有数据上的表现（那是开放世界），但你能在「已见样本」这个切片上，证出一个对未见数据成立的、带置信度的界。霍夫丁不等式（Hoeffding's inequality）[17] 是它的概率引擎，PAC-Bayes（麦卡莱斯特[6]）是它的精化。

软件里，它叫类型与证明。类型系统不证明程序「全对」，只证明某一条性质（比如不会把整数当指针），换来的是可判定、可机械复核的检查（皮尔斯[7]）。柯里-霍华德对应（Curry-Howard correspondence，霍华德 1980[8]）把「证明」与「程序」划上等号，内库拉 1997 年的「携带证明的代码」（proof-carrying code）[9] 更是把这一招用到极致：不受信的代码自带一张安全性证明，宿主只需快速核对证书（certificate），而无需自己重新推导。勒罗伊 2009 年经形式验证的 CompCert 编译器[10]、de Moura 与比约纳 2008 年的 Z3 求解器[11]，都是同一思路的工业化。

数值计算里，它叫误差界。希格姆 2002 年的后向误差分析[12]、摩尔 1966 年的区间算术（interval arithmetic）[13]，让你带着「保证包含真值的区间」去计算，最终交付的不是一个可能骗你的浮点数，而是一个有保证的范围。数学里，它就是第 7 章那个验到高度 T 的零点：一个界，不是定理。

统一的观念是：证书是一个局部的、有界的、可独立复核的保证。它最妙的地方在于利用了第 2 章那道验证不对称：生成证书可能极贵，核对证书却极廉。携带证明的代码、NP 问题的解、数学证明，吃的都是这口红利。

它的标准败法只有一种，却很常见：空洞的界。一个为真却无用的保证，如「误差不超过百分之百」或「该模型的泛化误差有限」，逻

辑上无懈可击，操作上一文不值。界的价值不在成立，而在够紧到能据以行动。

最优筛查：把查验花在刀刃上

第二招的纯形式：信息有代价，所以把有限的查验，分配到边际上最能压缩不确定性的地方。

它的跨域形态同样齐整。统计与科学里，它叫实验设计（design of experiments）：费雪 1935 年的《实验设计》[19]、博克斯等人的《实验者统计学》[20]，教的是如何用最少的试验榨出最多的信息；林德利 1956 年给出「一个实验提供的信息」的度量[21]，沙洛纳与韦尔迪内利把贝叶斯实验设计系统化[22]；瓦尔德 1945 年的序贯检验[23]，让你边收数据边决定要不要继续。香农 1948 年的信息论[14]，是这一切的底层货币。机器学习里，它叫主动学习（active learning）：下一个该标注哪个样本最划算（科恩等人[33]、塞特尔斯[34]）。软件测试里，它叫 fuzzing：该把算力砸向哪个输入去撞出崩溃（米勒等人 1990 年的开创性实验[35]）。这一招的现代规模令人侧目。谷歌的 OSS-Fuzz 自 2016 年起持续向上千个开源项目自动喂入海量畸形输入，至今已查出数以万计的缺陷与漏洞。没有哪支人力测试团队能穷举到这个量级，它靠的正是把算力源源不断地投向最可能崩的地方。审计里，它叫抽样：查哪几笔交易最可能发现问题。界面里，它叫该问用户哪个问题（第 5 章）。

这些背后是同一个最优化问题：让回答与未知之间的期望信息增益（expected information gain）最大，

$$q^\star = \arg\max_{q} \; I(\theta; y_q).$$

而当查验本身要反复进行、且要边查边用时，它就长成了探索与利用（exploration and exploitation）的张力，即多臂老虎机（multi-armed bandit）问题。汤普森 1933 年的采样[24]、罗宾斯 1952 年的开创[25]、赖与罗宾斯 1985 年的最优分配[26]、奥尔等人 2002 年的有限时间分析[27]，给出的是「花多少次试验去减少对哪个选项的不确

定」的最优解；它的遗憾（regret）随时间只以对数增长，

$$\mathrm{Regret}(T) = O(\ln T).$$

这里要防一个叙述上的路径依赖。最优筛查是一个方法族，实验设计、主动学习、审计抽样、fuzzing、老虎机，都是它。库什纳、莫库斯到琼斯 1998 年的高效全局优化[30]、斯里尼瓦斯等人 2010 年的 GP-UCB[32]，乃至沙赫里亚里 2016 年那篇综述里基于高斯过程的贝叶斯优化[36]，都极其有用，但它们只是这个族里的一支实现，不是「筛查」的全部。把这一招等同于高斯过程，就把一个普遍的姿势缩成了一件工具。

它的标准败法也只有一种：优化了一个被误设的信息度量。你极其高效地收集了信息，却是关于错误问题的，或者那个被你最大化的「信息量」根本不追踪你真正在意的东西。筛查越聪明，错设的度量就会越快地把你领向歧途。

两招为何成对，以及通向哪里

把这两招并排看：证书在一个切片上把不确定性压到一个有保证的界内，筛查则花掉信息预算去观测那个最能压缩不确定性的切片。一个是「在能查的地方证紧」，一个是「把查验花在最该查的地方」。它们从两端夹击同一个敌人，即未知。这也是它们共用的那根杠杆：在一个信息预算之下，管理你在哪里、以多大力度去削减不确定。第 13 章会正式给这根杠杆命名。

但有时候，无论你怎么压、怎么筛，都不够，因为你压根缺乏做出判断的能力本身。这时就不能再靠自己缩小未知了，得去别处把判断借来。下一对招，正是关于此。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. L. Valiant (1984).「A Theory of the Learnable」. Communications of the ACM, 27(11), 1134-1142. [②] 瓦利安特在此提出「概率近似正确」（PAC）的学习框架，把「学会一个概念」严格定义为：以高概率、在多项式时间与样本内，得到一个误差足够小的假设。这篇论文为「能不能学、要多少样本才学得动」给出了第一套可证明的语言，是本章「泛化界」一招的源头。
2. V. Vapnik & A. Chervonenkis (1971).「On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities」. Theory of Probability & Its Applications, 16(2), 264-280. [②] 这篇奠基之作证明了经验频率向真实概率一致收敛的条件，并由此引出后来以两位作者命名的 VC 维，用以刻画一个函数族的「容量」。它解释了为何在有限样本上证出的误差界能对未见数据成立，是泛化界背后的概率与组合根基。
3. V. Vapnik (1995).《The Nature of Statistical Learning Theory》. Springer. [②] 瓦普尼克在这本书里把统计学习理论整理成一个完整体系：以结构风险最小化为核心，权衡经验误差与模型复杂度，并由此导向支持向量机。它是理解「复杂度惩罚」为何出现在泛化界里的标准读物，把第一招的直觉讲得清楚而连贯。
4. A. Blumer, A. Ehrenfeucht, D. Haussler & M. Warmuth (1989).「Learnability and the Vapnik-Chervonenkis Dimension」. Journal of the ACM, 36(4), 929-965. [②] 这篇论文把 VC 维正式接入 PAC 框架，证明一个概念类可被 PAC 学习当且仅当其 VC 维有限，并给出依赖 VC 维的样本复杂度界。它是正文那条泛化界公式的直接来源，把「容量有限便可学」这一判据钉死。

5. A. Blumer, A. Ehrenfeucht, D. Haussler & M. Warmuth (1987).「Occam's Razor」. Information Processing Letters, 24(6), 377-380. [②] 这篇短文给出「奥卡姆剃刀」的学习理论版本：一个能把训练数据压缩得足够短的假设，便能以高概率泛化。它把「简洁即可学」从哲学格言变成可证的命题，为本章「压缩未知」的母题提供了一个干净的注脚。
6. D. McAllester (1999).「PAC-Bayesian Model Averaging」. Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT), 164-170. [②] 麦卡莱斯特在此提出 PAC-Bayes 界：对一族假设上的后验分布给出泛化保证，惩罚项由后验与先验之间的 KL 散度衡量。它是 PAC 界的一次精化，正文称其为第一招的「精化」即指此，常给出比经典 VC 界更紧的结果。
7. B. Pierce (2002).《Types and Programming Languages》. MIT Press. [②] 皮尔斯这本教材系统讲解类型系统的理论与构造，核心是类型安全的「进展」与「保型」两条性质，以及它们如何被机械地检查。它正是正文那句「类型系统不证明程序全对，只证某一条性质、换来可判定可复核」的标准依据，是理解证书一招在软件中形态的入门书。
8. W. Howard (1980).「The Formulae-as-Types Notion of Construction」. 收于 J. Seldin & J. Hindley 编《To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism》, 479-490. Academic Press. [②] 霍华德这篇广为流传的文稿（写于 1969 年，1980 年正式发表）确立了「公式即类型、证明即程序」的对应：直觉主义逻辑的命题与类型一一对应，证明与项一一对应。它是柯里-霍华德对应的经典文本，把「检查一段程序的类型」与「检验一个证明」划上等号，正是证书一招的逻辑核心。
9. G. Necula (1997).「Proof-Carrying Code」. Conference Record of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL), 106-119. [②④] 内库拉提出「携带证明的代码」：不受信的程序自带一张关于自身安全性的形式证明，宿主只需快速核验这张证书，而不必信任代码来源或重新推导。它把第 2 章那道验证不对

称用到极致，是本章证书概念最纯粹的工程化身。

10. X. Leroy (2009).「Formal Verification of a Realistic Compiler」. Communications of the ACM, 52(7), 107-115. [②③] 勒罗伊报告了 CompCert 的成果：一个用 Coq 形式验证过的 C 编译器，其生成代码在语义上与源程序一致这件事，是被机器证明出来的。它表明「带可复核保证的真实软件」并非空想，是证书一招工业化的标志性案例。
11. L. de Moura & N. Bjørner (2008).「Z3: An Efficient SMT Solver」. Tools and Algorithms for the Construction and Analysis of Systems (TACAS), LNCS 4963, 337-340. Springer. [②④] 这篇论文介绍了 Z3 这一高效的 SMT 求解器，它能判定带有算术、数组等理论的逻辑公式的可满足性，并广泛用于程序验证与符号执行。它把「自动产出可复核证书」做成了一件随手可用的工业工具，是证书一招在实践中得以铺开的引擎之一。
12. N. Higham (2002).《Accuracy and Stability of Numerical Algorithms》(2nd ed.). SIAM. [②] 希格姆这部权威著作系统讲述数值算法的误差分析，尤其是后向误差分析：与其问「答案偏离真值多少」，不如问「这个答案恰好是哪个被微扰问题的精确解」。它是正文「误差界」一招的标准参考，教人如何为浮点计算配上可信赖的保证。
13. R. Moore (1966).《Interval Analysis》. Prentice-Hall. [②] 摩尔这本开创性著作确立了区间算术：让每个量带着一个保证包含真值的区间一起参与运算，输出的便不是一个可能骗人的数，而是一个有保证的范围。它把「交付一个界而非一个点」的思路给了数值计算，正是证书一招在该领域的化身。
14. C. Shannon (1948).「A Mathematical Theory of Communication」. Bell System Technical Journal, 27(3), 379-423; 27(4), 623-656. [①②] 香农这篇创立信息论的论文，用熵度量不确定性，并给出信源编码与信道容量的根本极限。它是「信息有代价、可被度量」这一观念的总源头，正文称之为最优筛查的「底层货币」，本章对信息增益、压缩的全部讨论都以它为单位。
15. J. Rissanen (1978).「Modeling by Shortest Data Description」

. Automatica, 14(5), 465-471. [②] 里萨宁提出最小描述长度（MDL）原则：最好的模型，是能把数据连同模型自身一起编码得最短的那个。它把「压缩即理解」做成可操作的模型选择准则，与本章「压缩未知」的母题正相呼应，也是奥卡姆剃刀的一个信息论实现。

16. M. Li & P. Vitányi (2008).《An Introduction to Kolmogorov Complexity and Its Applications》(3rd ed.). Springer. [②] 这部标准教材系统讲述柯尔莫哥洛夫复杂度：一个对象的复杂度，等于能生成它的最短程序的长度。这是「压缩」的不可计算理想，与 MDL 那种可操作的近似相对照；它为本章压缩母题提供了理论上的天花板，说明最优压缩本身正是一种无法验证的极限。
17. W. Hoeffding (1963).「Probability Inequalities for Sums of Bounded Random Variables」. Journal of the American Statistical Association, 58(301), 13-30. [②] 霍夫丁在此给出有界随机变量之和偏离其均值的指数型概率上界。这个不等式是把「经验平均」与「真实期望」之间的差距压进一个置信界的基本工具，正文称其为泛化界的「概率引擎」，证书一招的多数集中度论证都从它起步。
18. E. Candès, J. Romberg & T. Tao (2006).「Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information」. IEEE Transactions on Information Theory, 52(2), 489-509. [②] 这篇压缩感知的奠基论文证明：只要信号足够稀疏，便能用远少于传统采样定理所要求的测量数，通过凸优化把它精确重建出来。它是「把查验花在刀刃上、用极少观测榨出全部信息」的一个数学典范，呼应本章压缩与最优筛查两条线索。
19. R. Fisher (1935).《The Design of Experiments》. Oliver and Boyd. [①③] 费雪这本经典确立了现代实验设计的基本原则：随机化、重复、区组化，以及著名的「女士品茶」思想实验。它教人如何用最少的试验榨出最多可信的信息，是最优筛查一招在统计与科学中的源头读物。
20. G. Box, W. Hunter & J. Hunter (1978).《Statistics for Experimenters: An Introduction to Design, Data Analysis, and

Model Building》. John Wiley & Sons. [③④] 博克斯等人这本广受欢迎的实务著作，把实验设计、数据分析与模型构建讲给真正动手做实验的人，强调析因设计与序贯学习的迭代节奏。它把费雪的原则落到工程现场，是理解「如何把查验安排得最划算」的实践指南。

21. D. Lindley (1956).「On a Measure of the Information Provided by an Experiment」. The Annals of Mathematical Statistics, 27(4), 986-1005. [②] 林德利用信息论给出「一个实验提供多少信息」的度量，即先验与后验之间的期望信息增益。这正是正文那个最优筛查目标 $\arg\max_q I(\theta; y_q)$ 的理论原型，把「该做哪个实验」变成一个可最大化的量。
22. K. Chaloner & I. Verdinelli (1995).「Bayesian Experimental Design: A Review」. Statistical Science, 10(3), 273-304. [②] 这篇综述系统梳理了贝叶斯实验设计：以效用函数（常取期望信息增益）为目标，统一地导出各种最优设计准则。它把林德利的度量整合进一个完整框架，是读者快速掌握「期望信息增益最大化」这一筛查内核的入口。
23. A. Wald (1945).「Sequential Tests of Statistical Hypotheses」. The Annals of Mathematical Statistics, 16(2), 117-186. [①②] 瓦尔德提出序贯概率比检验：边收数据边判断，一旦证据足够强就停下来下结论，从而平均上比固定样本量的检验省得多。它把「要不要继续查」本身变成最优决策，是最优筛查里「边查边用」这一支的先声。
24. W. Thompson (1933).「On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples」. Biometrika, 25(3-4), 285-294. [①②] 汤普森在此提出后来以他命名的采样法：按「某个选项确实最优」的后验概率去随机选取它，自然地在探索与利用之间取得平衡。这是多臂老虎机问题最早的解法之一，至今仍是该问题里既简洁又强劲的策略。
25. H. Robbins (1952).「Some Aspects of the Sequential Design of Experiments」. Bulletin of the American Mathematical Society, 58(5), 527-535. [①②] 罗宾斯这篇论文把多臂老虎机问题正式确立为一个数学对象，并提出最早的序贯分配策略，

奠定了「探索与利用如何权衡」这一研究方向。它是后来一整条老虎机文献的起点，本章关于「花多少次去减少哪个不确定」的讨论由此开端。

26. T. Lai & H. Robbins (1985).「Asymptotically Efficient Adaptive Allocation Rules」. Advances in Applied Mathematics, 6(1), 4-22. [②] 赖与罗宾斯证明了多臂老虎机的遗憾下界：任何合理策略的累积遗憾都至少随时间对数增长，并构造出达到这一下界的渐近最优分配规则。它确立了正文那个 $O(\ln T)$ 是无法逾越的极限，给整类问题划定了天花板。
27. P. Auer, N. Cesa-Bianchi & P. Fischer (2002).「Finite-time Analysis of the Multiarmed Bandit Problem」. Machine Learning, 47(2-3), 235-256. [①②] 这篇论文给出 UCB1 等基于「乐观面对不确定」的简单算法，并证明其在有限时间内（而非仅渐近）就有对数级的遗憾界。它把赖与罗宾斯的渐近结果落实为可直接使用、可分析的具体策略，是「置信上界」一类方法的标准引用。
28. H. Kushner (1964).「A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise」. Journal of Basic Engineering, 86(1), 97-106. [①②] 库什纳这篇早期论文用概率模型刻画一条未知的带噪曲线，并据此选取下一个采样点去寻找极大值，是贝叶斯优化思想的雏形。它说明「该往哪里再测一次」可以被当作一个最优决策来处理，是最优筛查在全局优化里的先驱之作。
29. J. Mockus, V. Tiesis & A. Žilinskas (1978).「The Application of Bayesian Methods for Seeking the Extremum」. 收于 L. Dixon & G. Szegő 编《Towards Global Optimization 2》, 117-129. North-Holland. [②] 莫库斯等人系统发展了贝叶斯全局优化，并提出「期望改进」这一采集函数：在概率模型下选取最可能带来改进的点去评估。它把库什纳的直觉做成了一套通用方法，是今天贝叶斯优化的直接前身。
30. D. Jones, M. Schonlau & W. Welch (1998).「Efficient Global Optimization of Expensive Black-Box Functions」. Journal of Global Optimization, 13(4), 455-492. [②④] 这篇论文提出 EGO 算法，用高斯过程为昂贵的黑箱函数建代理模型，再

以期望改进准则挑选下一个评估点，使评估次数大幅减少。它是把贝叶斯优化推广开来的标志性工作，但正文也提醒：它只是最优筛查这个方法族里的一支实现，而非全部。

31. C. Rasmussen & C. Williams (2006).《Gaussian Processes for Machine Learning》. MIT Press. [②④] 这本标准教材系统讲述高斯过程：一种对函数本身赋予先验、并能给出预测不确定性的非参数贝叶斯方法。它是贝叶斯优化所依赖的代理模型的理论底座，读者要理解「不确定性如何被建模并用于决定下一步查哪里」，此书是核心参考。
32. N. Srinivas, A. Krause, S. Kakade & M. Seeger (2010).「Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design」. Proceedings of the 27th International Conference on Machine Learning (ICML), 1015-1022. [②] 斯里尼瓦斯等人提出 GP-UCB 算法，把老虎机里的「置信上界」思路搬到高斯过程优化上，并为其遗憾给出可证的界。它把贝叶斯优化与老虎机理论缝在一起，正好示范了本章两端，证一个界与花掉查验，原本就是一根杠杆。
33. D. Cohn, Z. Ghahramani & M. Jordan (1996).「Active Learning with Statistical Models」. Journal of Artificial Intelligence Research, 4, 129-145. [②④] 科恩等人给出主动学习的统计框架：在统计模型下，挑选那个能最大程度降低模型方差（或预测不确定性）的样本去标注。它把「下一个该标注哪个最划算」变成可计算的准则，是主动学习作为最优筛查一支的代表性工作。
34. B. Settles (2009).《Active Learning Literature Survey》. Computer Sciences Technical Report 1648, University of Wisconsin-Madison. [②④] 塞特尔斯这份广为引用的综述系统整理了主动学习的各种查询策略，如不确定性采样、委员会查询、期望误差缩减等，并比较其适用场景。它是快速纵览「标注预算该怎么花」全貌的标准入口，把这一招的各种实现摆在一起对照。
35. B. Miller, L. Fredriksen & B. So (1990).「An Empirical Study of the Reliability of UNIX Utilities」. Communications of the ACM, 33(12), 32-44. [③④] 米勒等人用随机生成的输入

去喂各种 UNIX 工具，结果让相当一部分程序崩溃或挂死，这就是 fuzzing 的开创性实验。它说明「把算力砸向随机或可疑的输入去撞出故障」是一种廉价而有效的查验方式，是最优筛查在软件测试里的起点。

36. B. Shahriari, K. Swersky, Z. Wang, R. Adams & N. de Freitas (2016).「Taking the Human Out of the Loop: A Review of Bayesian Optimization」. Proceedings of the IEEE, 104(1), 148-175. [②④] 这篇综述全面梳理了基于高斯过程的贝叶斯优化：代理模型、采集函数及其在超参数调优等场景的应用。它是了解该方向现状的标准读物，但正文借它提醒读者，不要把「最优筛查」这一普遍姿势缩成「高斯过程」这一件工具。

第 10 章　借来的判断

> **论点**：当你缺乏验证能力，就从外部引进。要么在回路里放一个可信的判断者（神谕，oracle），要么用许多互相独立的不可靠判断者，信任他们的一致（冗余，redundancy / 共识，consensus）。

上一对招还在自己身上想办法，缩小未知。但有时你缺的不是信息，而是判断力本身，你压根没有能力对眼前这事下一个可靠的判决。这一对招的应对是：不在自己身上找了，去别处把判断借来。借法有两种，要么引进一个你信得过的判断者（神谕），要么集合许多互不信任的判断者，信任他们的一致（冗余）。

神谕入回路：引进一个判断者

第一招的纯形式：在你缺乏验证能力的那个决策点上，插入一个外部的判断者，由它来给出你给不出的判决。

最朴素的版本，是第 5 章那个人在回路（human-in-the-loop），是专家会诊，是疑难上交。但这一招最深刻的形态，藏在两个看似无关的地方。

一处是交互式定理证明（interactive theorem proving）。德布鲁因 1970 年的 AUTOMATH[1]、爱丁堡 LCF（戈登、米尔纳、沃兹沃思 1979[2]）、到今天的 Coq（贝尔托与卡斯特朗 2004[3]），都是同一种分工：人提供那闪光的、机器给不出的证明思路（神谕），机器则一丝不苟地核对每一步（证书检查，certificate checking）。神谕负责

「找」，机器负责「验」，正好咬合第 2 章那道不对称。

另一处更惊人，是复杂性理论里的交互式证明（interactive proof）。一个算力贫弱的验证者，面对一个强大却不可信的证明者，如何能问出一个它自己根本算不出的可靠答案？戈德瓦塞尔、米卡利与拉科夫 1989 年[6]、巴拜 1985 年[7] 给出的答案是：靠反复盘问加随机挑战。验证者抛出它自己都无法预知的随机问题，证明者若在撒谎，迟早会在某个挑战上露馅。沙米尔 1992 年[15] 那个惊人的 IP = PSPACE 表明，单靠这种「盘问一个不可信神谕」的方式，弱验证者能可靠地裁决极其庞大的一类问题；布卢姆与坎南 1995 年[17] 的「会检查自己工作的程序」、戈德瓦塞尔等人 2015 年[19]「为凡人代理计算」，都是同一脉。这是「借来的判断」最纯的数学化：哪怕神谕不可信，只要你会聪明地盘问它，依然能榨出可靠。

统一的观念是：制造一种你单独不具备的可靠，靠引进一个外部判断者。它的标准败法也很直白：神谕本身不可靠或有偏。你引进的裁判，可能就是个错的裁判，而「谁来验证神谕」这个问题，会把你带进一段退无可退的回归。

冗余：从许多不可靠里合成可靠

第二招换了个方向：不引进一个可信的，而是召集许多不可信的，信任他们的一致。

它的理论根基有两块奠基石。冯 · 诺依曼 1956 年[4] 证明，可以用本身会出错的元件，通过堆叠冗余组装出任意可靠的计算。孔多塞 1785 年[5] 的陪审团定理（Condorcet's jury theorem）给出了它的算术：若每个判断者都略好于瞎猜（正确率 $p > \frac{1}{2}$），且彼此独立，那么多数票正确的概率会随人数趋于必然，

$$P_N \to 1 \quad (N \to \infty).$$

这一招的跨域形态铺得极开。分布式系统里，它是拜占庭容错（Byzantine fault tolerance）：皮斯、肖斯塔克与兰波特 1980 年[8]、兰波特等人 1982 年[9] 的拜占庭将军问题（Byzantine generals problem），

要在部分节点可能作恶（对抗）的情况下达成共识，经典门槛是节点数 $n \geq 3f+1$ 才能容忍 f 个叛徒，卡斯特罗与利斯科夫 1999 年[11] 的 PBFT 把它做进了实用系统（费舍尔、林奇与帕特森 1985 年[10] 的不可能性定理则划出了它的边界）。机器学习里，它是集成（ensemble）：汉森与萨拉蒙 1990 年[20]、迪特里希 2000 年[22] 的集成方法、布雷曼 2001 年[23] 的随机森林（random forest），用一群弱模型投票，胜过单个强模型。群体里，它是「群体的智慧」（wisdom of crowds）（索罗维基 2004[25]）。1906 年高尔顿在一场乡村集市上记录了约八百位村民对一头公牛体重的独立竞猜，没有一个人猜准，可全体估计的平均值是 1197 磅，而牛的真实重量是 1198 磅，整个群体合起来几乎分毫不差。洪与佩奇 2004 年[24] 甚至证明，在合适条件下，多样的普通解题者群体能胜过一群高手。科学里，它是同行评审与重复实验（第 3 章）；工程里，它是 RAID 与法定人数；医疗里，它是第二诊疗意见。

但这一招有一个关键前提必须用整节强调：独立。冗余只在失败去相关时才成立。把许多估计平均，方差才随人数下降，

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{N};$$

可一旦这些判断之间有正相关 ρ，方差就不再趋于零，而是卡在一个地板上，

$$\mathrm{Var}(\bar{X}) = \rho\,\sigma^2 + \frac{(1-\rho)\,\sigma^2}{N} \xrightarrow{N\to\infty} \rho\,\sigma^2.$$

相关把冗余的全部价值一笔勾销。你堆再多判断者，也跨不过这道由相关性设下的地板。这不是空谈：奈特与莱韦森 1986 年[13] 那个著名实验，让许多程序员独立地为同一规格编写程序，本指望它们的错误互不相干，结果发现他们栽在同样的地方，因为人类面对同一个难点会犯同样的错（埃克哈特与李 1985 年[12] 早有理论预言）。群体思维（groupthink）、同源的有缺陷训练数据、共模故障（common-mode failure），都是这道地板的现身。这就是冗余的标准败法：以为独立，其实相关。

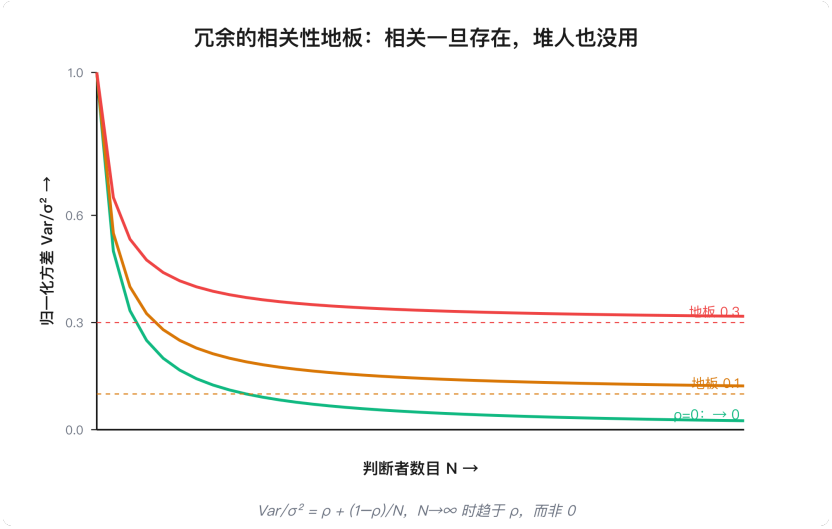

图 7: 冗余的相关性地板：相关一旦存在，堆人也跨不过去

两招的合流，与一段跨章的呼应

把两招并看：一个引进单一而昂贵的神谕，一个合成众多而廉价的独立判断，借的都是你单独不具备的判断力。它们共用的杠杆，是为自己补上缺失的验证能力；它们的败法也两两相对，单一神谕可能错，众多判断可能暗中相关。

数学里有一段插曲，恰好把这一对招、连同上一章的证书全串了起来。阿佩尔与哈肯 1977 年[26] 的四色定理（four color theorem）证明，因为依赖计算机的穷举而饱受争议，那等于要数学界去信任一个神谕。后来贡蒂耶 2008 年[27] 用机器可核对的形式证明（formal proof）重做了它，黑尔斯团队[28] 对开普勒猜想也如法炮制：把「信任神谕」转化成了「核对证书」。麦肯齐[29] 在《机械化证明》里追踪的，正是这种信任如何在人、机器与社会过程之间转移；德米洛等人[14] 那句「证明是一种社会过程」，说到底就是把数学的可信安放在人类判断的冗余之上。

不过要看清一件事：到这里为止，前两对招，压缩未知与借来判断，都还在追求同一样东西，对象的真。它们仍想知道这事到底对不对。

下一对招做了一件更彻底的事：它不再索求那个真。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. N. G. de Bruijn (1970).「The mathematical language AUTOMATH, its usage, and some of its extensions」. 收入《Symposium on Automatic Demonstration》. Springer (Lecture Notes in Mathematics 125), pp. 29-61. [②] 德布鲁因介绍了 AUTOMATH，史上最早能让整篇数学被机器逐步核对的形式语言之一，由人写出证明、机器验证其无误。它是本章「神谕入回路」最早的工程化样本：人提供思路，机器只负责一丝不苟地检查，读者可由此看清「找」与「验」分工的源头。
2. M. Gordon, R. Milner, C. Wadsworth (1979).《Edinburgh LCF: A Mechanized Logic of Computation》. Springer (Lecture Notes in Computer Science 78). [②] 这本书提出了 LCF 这套交互式证明系统，其设计影响深远：用一个受信任的小内核担保每一步推理的可靠，证明策略再多也无法绕过它。它为本章关于「证书检查」的论述提供了经典范式，读者可看到「可信核对者」如何被收缩成一个尽量小、尽量可靠的部件。
3. Y. Bertot, P. Castéran (2004).《Interactive Theorem Proving and Program Development. Coq'Art: The Calculus of Inductive Constructions》. Springer (Texts in Theoretical Computer Science, EATCS Series). [②] 这是 Coq 证明助手的权威教程，系统讲解如何在归纳构造演算之上交互地构造并机器核对证明。它把前两条所代表的传统带到当代实践，是本章后文四色定理、开普勒猜想等形式化工作的工具基础，想动手理解「人给思路、机器验每步」的读者可由此入门。
4. J. von Neumann (1956).「Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components」

. 收入 C. E. Shannon, J. McCarthy 编《Automata Studies》(Annals of Mathematics Studies 34). Princeton University Press, pp. 43-98. [②] 冯·诺依曼在此证明，可以用本身会出错的元件，通过堆叠冗余与多数表决，组装出任意逼近可靠的计算。这是本章「冗余」一招的奠基石之一，为「从许多不可靠里合成可靠」提供了最早的严格论证，读者应读其对冗余如何压低错误率的核心思路。

5. Marquis de Condorcet (1785). 《Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix》. Imprimerie Royale, Paris. [②④] 孔多塞在这部关于投票的著作里给出了著名的陪审团定理：若每个判断者都略优于瞎猜且彼此独立，多数票正确的概率会随人数增长趋于必然。它为本章冗余一招提供了算术骨架，也预埋了它的命门，读者应留意定理对「独立」这一前提的依赖。

6. S. Goldwasser, S. Micali, C. Rackoff (1989). 「The Knowledge Complexity of Interactive Proof Systems」. SIAM Journal on Computing, 18(1), pp. 186-208. [②] 这篇论文开创了交互式证明与零知识证明的理论框架：一个算力有限的验证者，靠反复盘问加随机挑战，能从一个不可信的证明者那里榨出可靠的判决。它是本章「盘问不可信神谕」最纯的数学源头，读者应读它如何用随机性逼出真话。

7. L. Babai (1985). 「Trading Group Theory for Randomness」. 收入《Proceedings of the 17th Annual ACM Symposium on Theory of Computing (STOC)》, pp. 421-429. [②] 巴拜在此独立提出了 Arthur-Merlin 这类带随机性的交互式证明，与上一条几乎同时奠定了同一片理论疆域。它强化了本章的核心观念：随机挑战是弱验证者制服强而不可信证明者的关键武器，读者可与上一条对照阅读其互补的视角。

8. M. Pease, R. Shostak, L. Lamport (1980). 「Reaching Agreement in the Presence of Faults」. Journal of the ACM, 27(2), pp. 228-234. [②] 这篇论文最早严格刻画了在部分节点可能任意作恶时如何达成共识，给出了著名门槛：要容忍 f 个叛徒，节点数须满足 $n \geq 3f + 1$。它是本章冗余一招在分布式系统里的对抗性版本之源头，读者应读这个门槛背后的不可

能性论证。

9. L. Lamport, R. Shostak, M. Pease (1982).「The Byzantine Generals Problem」. ACM Transactions on Programming Languages and Systems, 4(3), pp. 382-401. [②] 这篇论文用「拜占庭将军」这个著名比喻把上一条的结果讲成了一则寓言，从此「拜占庭容错」成为对抗性共识的通用名字。它是本章用许多互不信任的判断者求一致这一思路的标志性文本，读者可读它如何把抽象门槛包装成直觉清晰的故事。
10. M. J. Fischer, N. A. Lynch, M. S. Paterson (1985).「Impossibility of Distributed Consensus with One Faulty Process」. Journal of the ACM, 32(2), pp. 374-382. [②] 这篇著名的 FLP 不可能性定理证明，在完全异步的系统里，哪怕只有一个进程可能崩溃，也不存在保证终止的确定性共识算法。它为本章冗余一招划出了边界，读者应读它如何说明共识并非无条件可得，从而理解后续实用系统为何要靠额外假设来绕开它。
11. M. Castro, B. Liskov (1999).「Practical Byzantine Fault Tolerance」. 收入《Proceedings of the 3rd USENIX Symposium on Operating Systems Design and Implementation (OSDI)》, pp. 173-186. [②] 这篇论文提出的 PBFT 算法，第一次把拜占庭容错从理论做成在真实异步网络里跑得动、性能可接受的系统。它是本章冗余一招从纸面落到工程的关键一步，读者可读它如何在保住 $n \geq 3f + 1$ 门槛的同时把开销压到实用范围，也理解后世区块链共识的直接前身。
12. D. E. Eckhardt, L. D. Lee (1985).「A Theoretical Basis for the Analysis of Multiversion Software Subject to Coincident Errors」. IEEE Transactions on Software Engineering, SE-11(12), pp. 1511-1517. [②] 这篇论文从理论上指出，多版本软件即便由不同人独立开发，其错误也未必独立：面对同一处难点，不同版本会倾向于一起出错，使冗余的收益远低于独立假设的预期。它为本章「相关性地板」给出了早于实验的理论预言，读者应读它如何刻画共因错误。
13. J. C. Knight, N. G. Leveson (1986).「An Experimental Evaluation of the Assumption of Independence in Multiversion

Programming」. IEEE Transactions on Software Engineering, SE-12(1), pp. 96-109. [②] 这是那个著名的实验：让许多程序员独立地照同一规格编写程序，本指望错误互不相干，结果发现他们在相同的难点上一起栽倒，独立假设被经验否定。它为上一条的理论预言提供了实证，是本章「以为独立、其实相关」这一败法最有说服力的例证。

14. R. A. De Millo, R. J. Lipton, A. J. Perlis (1979). 「Social Processes and Proofs of Theorems and Programs」. Communications of the ACM, 22(5), pp. 271-280. [③④] 这篇有名又有争议的论文主张，数学证明之所以可信，靠的不是形式推导的机械正确，而是数学共同体反复检验、传播、采信的社会过程，并据此质疑程序形式验证的前景。它支撑本章的观点：可信归根结底安放在人类判断的冗余之上，读者应读它对「证明是社会过程」的论证。
15. A. Shamir (1992). 「IP = PSPACE」. Journal of the ACM, 39(4), pp. 869-877. [②] 沙米尔证明了交互式证明的威力之惊人：单靠盘问一个不可信证明者，弱验证者能可靠裁决整个 PSPACE 这一极庞大的问题类，即 IP = PSPACE。它是本章「借来的判断」最有力的数学注脚，读者应读它如何界定「会聪明盘问神谕」所能达到的上限。
16. C. Lund, L. Fortnow, H. Karloff, N. Nisan (1992). 「Algebraic Methods for Interactive Proof Systems」. Journal of the ACM, 39(4), pp. 859-868. [②] 这篇论文引入了把布尔公式算术化、再用多项式做检查的代数方法，正是它的技术铺垫直接通向了上一条 IP = PSPACE 的证明。它对本章的意义在于揭示「聪明盘问」的具体机理：把验证问题转译成可随机抽查的代数恒等式，读者可读这套手法。
17. M. Blum, S. Kannan (1995). 「Designing Programs That Check Their Work」. Journal of the ACM, 42(1), pp. 269-291. [②] 这篇论文提出了「程序检查器」的思想：让程序在给出结果时附带一个独立、廉价的核对程序，验证本次输出是否正确，而无需信任程序本身。它把交互式证明的精神带到日常计算，对本章而言，是「不信任产出者、只核对其产出」这一思路的范本，读者应读其检查器的构造。

18. S. Arora, C. Lund, R. Motwani, M. Sudan, M. Szegedy (1998).「Proof Verification and the Hardness of Approximation Problems」. Journal of the ACM, 45(3), pp. 501-555. [②] 这是著名的 PCP 定理的核心论文之一：任何证明都能改写成一种特殊格式，使验证者只需随机抽查其中常数个比特，就能以高置信度判断其真伪。它把「抽查就够」推到极致，对本章是「弱验证者如何高效核对庞大证明」的理论顶点，读者应读其概率可检证明的惊人结论。
19. S. Goldwasser, Y. T. Kalai, G. N. Rothblum (2015).「Delegating Computation: Interactive Proofs for Muggles」. Journal of the ACM, 62(4), Article 27. [②④] 这篇论文让交互式证明真正服务于「凡人」：一个算力有限的用户把计算外包给强大但不可信的服务器，再用远小于自己重算的代价核对结果是否正确。它是本章思想走向云计算时代的落点，读者应读它如何把「为弱者代理计算并可验证」做成现实可行的协议。
20. L. K. Hansen, P. Salamon (1990).「Neural Network Ensembles」. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10), pp. 993-1001. [②] 这篇论文较早地表明，把多个独立训练的神经网络组合起来投票，其整体准确率可显著高于任何单个网络。它是本章冗余一招在机器学习里的开端，读者应读它如何把「多数表决降低错误」从逻辑电路搬到学习模型，并再次落到对成员误差去相关的依赖。
21. A. Krogh, J. Vedelsby (1995).「Neural Network Ensembles, Cross Validation, and Active Learning」. 收入《Advances in Neural Information Processing Systems 7》. MIT Press, pp. 231-238. [②] 这篇论文给出了集成误差的经典分解：集成的整体误差等于成员的平均误差减去成员之间的分歧度。它为本章「相关性地板」提供了机器学习版的精确公式，读者应读它如何用数学说明，成员越是多样、越是各错各的，集成才越值钱。
22. T. G. Dietterich (2000).「Ensemble Methods in Machine Learning」. 收入《Multiple Classifier Systems (MCS 2000)》. Springer (Lecture Notes in Computer Science 1857), pp. 1-15. [②] 这是一篇影响广泛的综述，梳理了集成方法为何有

效，并从统计、计算与表示三个角度给出解释。它是读者纵览本章冗余一招在机器学习里全貌的便捷入口，把分散的投票、装袋、提升等手法收拢在一个框架下理解。

23. L. Breiman (2001).「Random Forests」. Machine Learning, 45(1), pp. 5-32. [②] 布雷曼提出的随机森林，靠对样本与特征的双重随机化来培育一群彼此去相关的决策树，再投票合成强大而稳健的预测器。它是冗余一招最成功的实战范例之一，对本章的意义在于：它的全部威力恰恰来自刻意制造的独立性，读者可读它如何主动压低成员间的相关。
24. L. Hong, S. E. Page (2004).「Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers」. Proceedings of the National Academy of Sciences, 101(46), pp. 16385-16389. [②③④] 洪与佩奇借一个形式化模型论证：在合适条件下，由多样的普通解题者组成的群体，能胜过一群同质的高手，因为多样性带来的视角差异本身就是一种资源。它把本章「独立与多样才是冗余之本」的直觉推广到人类群体，读者应读其「多样性胜过能力」的论点与边界。
25. J. Surowiecki (2004).《The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations》. Doubleday. [③④] 索罗维基这本广为流传的书论证：在多样、独立、分散且有恰当聚合机制的条件下，群体的集体判断往往胜过专家个人，他也反复强调一旦丧失独立、陷入趋同，群体就会变蠢。它把本章冗余一招带到日常与社会层面，读者应读它对「群体智慧成立的前提」的反复申明。
26. K. Appel, W. Haken (1977).「Every Planar Map Is Four Colorable. Part I: Discharging」. Illinois Journal of Mathematics, 21(3), pp. 429-490. [①③] 这是四色定理的证明，史上第一个本质上依赖计算机穷举大量情形的重大数学证明，也因此引发关于「能否信任一个人手无法逐一复核的机器结论」的长久争论。它对本章而言，是「信任神谕」与其代价的标志性案例，读者应读这场争论如何逼出对机器可核对证书的需求。
27. G. Gonthier (2008).「Formal Proof: The Four-Color The-

orem」. Notices of the American Mathematical Society, 55(11), pp. 1382-1393. [②③] 贡蒂耶用 Coq 把四色定理重做成一份完全形式化、可被机器逐步核对的证明，从而把上一条那种「请信任计算机」的处境，转化为「核对一份证书」。它对本章是关键的对照点，读者应读它如何示范：把不可信的神谕产出，降格为可独立验证的证书，争议便随之消解。

28. T. Hales 等 (2017). 「A Formal Proof of the Kepler Conjecture」. Forum of Mathematics, Pi, 5, 文章号 e2. [②③] 黑尔斯团队历经多年，用形式化证明系统把饱受争议的开普勒猜想证明彻底机器核对了一遍，了结了同行评审都难以完全担保的疑虑。它与上一条同属一个脉络，对本章的意义在于再次印证：当证明大到人力难核时，把信任从神谕转移到可核对的证书，是恢复确信的出路。

29. D. MacKenzie (2001). 《Mechanizing Proof: Computing, Risk, and Trust》. MIT Press. [①③④] 麦肯齐这部社会学史著作追踪了计算机证明与形式验证的兴起，考察「证明」与「确信」如何在数学家、机器与社会过程之间被反复定义与转移。它为本章提供了贯穿性的视角，把神谕、证书、冗余都安放进一个关于信任如何被建立与让渡的更大叙事，读者应读它对「机械化证明改变了我们信任什么」的考察。

第 11 章　换一个能处理的问题

> 论点：别再坚持验证真正的对象。要么把它换成你能查的可解代理（代理替换，proxy substitution），要么不再索求二值判决、转而按标定（calibration）的概率行动（标定）。

前两对招还在追对象的真：压缩它、或借判断去逼近它。这一对放弃了那个执念。它不再追问「真正那件事对不对」，而是换一个问题来回答，要么换掉验证的对象（代理替换），要么换掉判决的形式（标定）。

代理替换：换掉你验证的对象

第一招的纯形式：别再死磕那个测不出的真目标，把它换成一个你查得动、且足够用的代理，去验证、去优化那个代理。

它的跨域形态，本书前面几乎每一章都撞见过。数学家用等价陈述代定理（第 7 章），软件工程师用测试代正确性、用基准代能力，组织用 KPI 代健康、用 GDP 代福祉（第 8 章），机器学习用奖励模型代人的真实偏好（第 5 章的 RLHF）。心理学里它也有个孪生：卡尼曼与弗雷德里克 2002 年[32] 的「属性替换」（attribute substitution），人在直觉判断时，会不自觉地用一个好评估的属性顶替那个难评估的目标属性。波利亚那句「先解一个相关而更易的问题」、西蒽的满意化（satisficing），是这一招的方法论原型。

而这一招的全部精髓，在于它有两种相反的失效方式。这正是第 7 章与第 8 章的交点，也贯穿全书。把「忠实」（代理是否真指向原目标）和「更易」（代理是否真比原问题好处理）摆成两维：

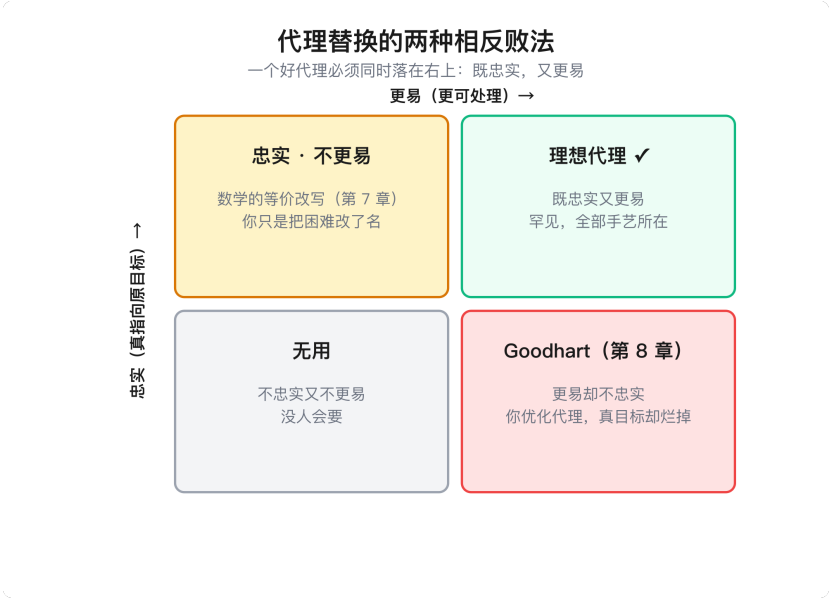

图 8: 代理替换的两种相反败法：忠实 × 更易

数学家栽在右上：等价改写忠实得无可挑剔，却一点没更好解，你只是给同一个困难换了身衣服。组织栽在左下：指标好测得很，可它和真目标的对应，一旦被当成目标去优化，就会断裂。

为什么一优化就断？因为代理与真目标的相关只在现状这个分布上成立，优化压力会把你推离那个分布、推向二者发生背离的极端。古德哈特 1975 年[1]（指标一旦成为目标即失效）、坎贝尔 1979 年[3]、卢卡斯 1976 年[6] 那条经济学孪生命题（被当作政策目标后，原有的结构关系即崩解），讲的是同一个机制。埃斯佩兰与索德尔的反身性（reactivity）更进一步：指标不只是失真，它还反过来重塑被测者。

这个机制在机器学习里以惊人的清晰度重演。阿莫迪等人 2016 年[10] 的奖励钻空（reward hacking）、潘等人 2022 年[12] 的实证研究发现能力更强的代理更善于钻代理奖励的空子，真实回报甚至出现骤

降的相变；斯卡尔塞等人 2022 年[13] 证明非平凡的奖励几乎不可能「不可钻空」；高等人 2023 年[14] 给这种过优化（overoptimization）测出了定量的标度律（scaling law）。一个好代理必须同时躲开这两端，既忠实又更易，而这罕见到，那点罕见本身就是全部的手艺。

标定：换掉判决的形式

第二招换的不是对象，是判决的形式：不再索求「真 / 假」的二值裁决，转而给出一个标定的概率，按它行动，接受有界的风险。

什么叫标定？说你有把握的事，该真按那个把握的比例发生。形式上，

$$\Pr\left(Y = 1 \mid \hat{p} = p\right) = p,$$

你报 70% 的那些事，长期看真该有七成成真。这是一个比「对错」弱、却可达、可检验的认识对象。

它的跨域形态同样齐整。数论里是概率素性（第 7 章那个「以 $1 - \varepsilon$ 概率为素数」）。机器学习里是保形预测（conformal prediction，沃夫克、伽默曼与谢弗 2005[26]），它不给你一个点判断，而给一个带覆盖保证的预测集，

$$\Pr\left(Y \in C(X)\right) \geq 1 - \alpha.$$

气象与预测科学里，是一整套成熟的标定理论：布莱尔 1950 年[15] 的评分、墨菲 1973 年[17] 把它分解为可靠性、分辨率与不确定性、德格鲁特与芬伯格 1983 年[18] 的系统处理、格奈廷等人 2007 年[24]「受制于标定，越锐越好」的现代框架。气象预报恰是标定做得最好的领域之一：当一个成熟的预报系统说「明天 70% 降雨」，把所有这样说过的日子拉长来看，真有约七成下了雨，把握与现实严丝合缝，这正是标定的范本。这里还有一个深刻的设计，严格适当评分规则（strictly proper scoring rule）：精心构造一个打分函数，使得说真话（报出你真实的概率）恰好让你的期望得分最优，

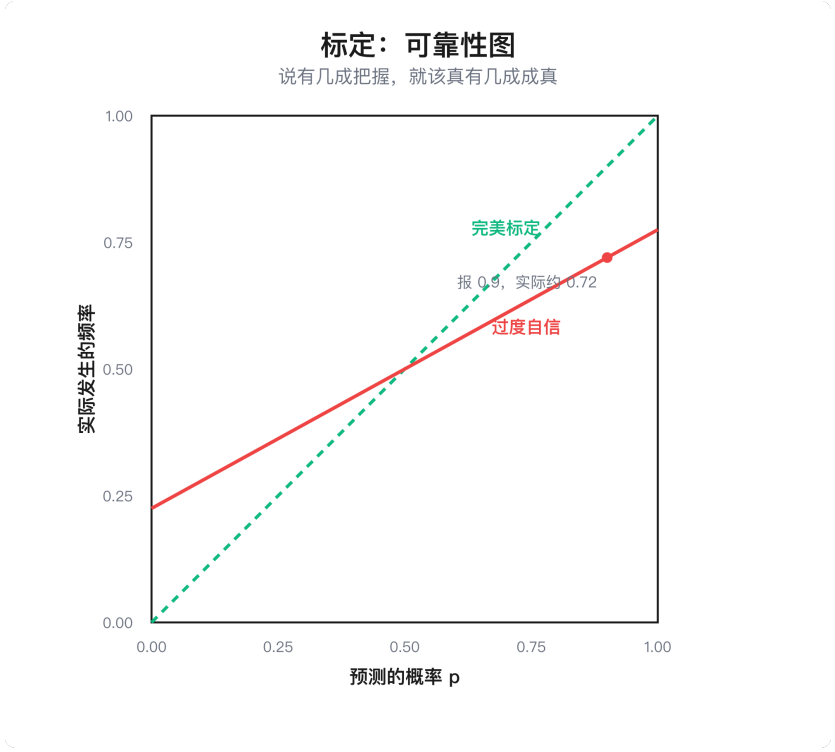

图 9: 标定的可靠性图：说有几成把握，就该真有几成成真

$$p = \arg\max_{q} \ \mathbb{E}_{Y\sim p}\big[S(q, Y)\big].$$

讲真话由此不再靠自觉，而被评分规则的数学结构所强制（萨维奇 1971[16]、格奈廷与拉夫特里 2007[23]）。达维德 1982 年[19] 证明贝叶斯主体能渐近自我标定，福斯特与沃赫拉 1998 年[22] 证明对任意序列都存在渐近标定的策略；但奥克斯 1985 年[20]「自我标定的先验不存在」则划出了这一招的极限。现代神经网络恰恰常常失标（miscalibration，郭等人 2017[25]），于是需要重新校准。第 6 章那个允许、询问、阻止的分级信任，正是标定落到行动上的样子。

标定有两种败法。浅一层是失标：你声称的把握与现实对不上，报 90% 却只有六成成真，于是基于它的一切决策都偏。深一层更微妙、也更要紧：标定告诉你赔率，却不告诉你该不该接受这个赌局。一个完美标定的「70%」，对「70% 够不够你下注」这个问题保持沉默，因为那取决于赌注的大小与你的价值排序，那是价值问题，不是验证问题。把这两者混为一谈，是用标定行动时最常见的陷阱：你以为概率替你做了决定，其实它只摆好了赔率，按不按下去仍要你自己掏出一套价值来。

两招为何成对，以及通向哪里

把这两招并看：代理替换换掉你验证的对象（一个不同的、查得动的靶子），标定换掉判决的形式（一个概率，而非真假）。它们都不去回答原来那个问题，而是把问题换成一个能处理的。共用的杠杆，是改变你对「答案」的要求本身，一个改你度量什么，一个改判决长什么样、并给残余风险标价。

但即便如此，这两招仍在努力把事情做对，只是放低了「对」的标准。最后一对招比这更彻底：它索性不再指望做对，转而经营做错。既然错误防不住，那就缩小它的代价、并确保它一旦发生你能发现。这是第 12 章。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

代理替换：从古德哈特定律到非预期后果

1. C. A. E. Goodhart (1975).「Problems of Monetary Management: The U.K. Experience」.《Papers in Monetary Economics》, Vol. I. Reserve Bank of Australia. [②] 这是古德哈特定律的原始出处，源于 1975 年悉尼的一次货币经济学会议。古德哈特在讨论英国货币管理经验时指出，一个统计规律一旦被用作调控目标，原先观察到的稳定关系往往就会失效。本章用它作为代理失真的标杆：指标与真目标的相关只在现状分布上成立，一被当成目标去优化便会断裂。

2. R. K. Merton (1936).「The Unanticipated Consequences of Purposive Social Action」.《American Sociological Review》, 1(6), 894–904. [②] 默顿系统讨论了有目的的社会行动为何会产生行动者未曾预料的后果，并梳理了知识不足、利益迫切、价值约束等若干来源。它是代理替换之副作用在社会学里的早期源头，提醒读者：优化一个代理时，真正吃紧的常是那些没有进入度量视野的后果。

3. D. T. Campbell (1979).「Assessing the Impact of Planned Social Change」.《Evaluation and Program Planning》, 2(1), 67–90. [②④] 坎贝尔定律的来源：一个量化的社会指标越是被用于社会决策，它就越容易遭到扭曲，也越容易反过来扭曲它本要监测的社会过程。它与古德哈特定律并列，是代理失真的另一块经典基石，读者可借此看清指标被赋以高利害后的腐化路径。

4. S. Kerr (1975).「On the Folly of Rewarding A, While Hoping for B」.《Academy of Management Journal》, 18(4), 769–783. [②④] 科尔考察了组织里普遍存在的激励错配：管理者

奖励的行为 A，往往并非他们真正期望的行为 B，于是激励系统稳定地产出了与初衷相悖的结果。它是激励与代理错配的管理学经典，正对应本章「优化代理、真目标却烂掉」那一格的现实样貌。

5. M. Strathern (1997).「'Improving ratings': audit in the British University system」.《European Review》, 5(3), 305–321. [②④] 斯特拉森借英国大学审计制度的观察，给出了那句广为流传的表述：当一个度量成为目标，它便不再是好的度量。本章关于代理一旦被当作目标即失真的论证，常以此为简洁的口径，读者可读到这一表述的原始语境。

6. R. E. Lucas (1976).「Econometric Policy Evaluation: A Critique」.《Carnegie-Rochester Conference Series on Public Policy》, 1, 19–46. [②③] 卢卡斯批判指出，计量模型中估计出的参数关系依赖于既有政策环境，一旦据此改变政策，主体的预期与行为会随之调整，原有的结构关系便不再成立。它是古德哈特定律的经济学孪生命题，本章用它说明优化压力为何会把系统推离代理与真目标相符的那个分布。

7. W. N. Espeland & M. Sauder (2007).「Rankings and Reactivity: How Public Measures Recreate Social Worlds」.《American Journal of Sociology》, 113(1), 1–40. [②④] 埃斯佩兰与索德尔提出「反身性」框架：公开的排名与量化指标不只是测量，它们还会改变被测者的行为乃至自我认知，从而重塑它本要描述的社会现实。本章借它把代理失真推进一层，指标不仅会失真，还会反过来重造被测对象。

8. D. Manheim & S. Garrabrant (2018).「Categorizing Variants of Goodhart's Law」. arXiv:1803.04585. [②] 两位作者尝试把笼统的「古德哈特定律」拆成几类不同机制（如回归型、极值型、因果型、对抗型），各自的失效方式与对策并不相同。它为本章「代理替换的失效不止一种」提供了细化的分类，便于读者分辨自己面对的是哪一种失真。

9. J. Z. Muller (2018).《The Tyranny of Metrics》. Princeton

University Press. [④] 穆勒以大量医疗、教育、警务、商业等领域的案例，批评了把一切都化为可量化指标并据以问责的风气，指出这种度量崇拜常带来表面达标而实质受损的后果。它是面向一般读者的通俗综述，适合读者在生活与工作中识别代理替换的代价。

代理失真在机器学习中的复现：奖励钻空与过优化

10. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman & D. Mané (2016).「Concrete Problems in AI Safety」. arXiv:1606.06565. [②] 这篇影响广泛的综述把 AI 安全拆成若干具体可研究的问题，其中包括奖励钻空（reward hacking）与可扩展监督等。它把社会科学里早已熟知的代理目标失真，清晰地译入了机器学习语境，是本章「同一机制在机器学习里重演」一段的起点。

11. P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg & D. Amodei (2017).「Deep Reinforcement Learning from Human Preferences」.《Advances in Neural Information Processing Systems》, 30 (NeurIPS 2017). [②④] 作者用人类对成对轨迹的偏好比较来训练一个奖励模型，再用它驱动强化学习，从而绕开难以手写的目标函数。这是 RLHF 的奠基工作，也正是本章所说「用奖励模型代人的真实偏好」的代理替换样板，读者可由此理解为何这种代理既好用又危险。

12. A. Pan, K. Bhatia & J. Steinhardt (2022).「The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models」. ICLR 2022. [②] 作者系统考察了奖励设定不当的后果，并给出一个值得警惕的经验现象：能力更强的智能体往往更善于钻代理奖励的空子，真实回报甚至会随能力提升而出现骤降式的相变。本章用它说明代理失真不是线性恶化，而可能在某处突然翻盘。

13. J. Skalse, N. H. R. Howe, D. Krasheninnikov & D. Krueger (2022).「Defining and Characterizing Reward Hacking」.《Advances in Neural Information Processing Systems》, 35

(NeurIPS 2022). [②] 这篇论文给出奖励钻空的一个形式化定义，并证明在非平凡的情形下，几乎不存在「不可钻空」的代理奖励。它为本章「好代理罕见」提供了理论支撑：忠实又稳健的代理之所以稀少，是有结构性原因的，而非工程上偶然没做好。

14. L. Gao, J. Schulman & J. Hilton (2023).「Scaling Laws for Reward Model Overoptimization」.《Proceedings of the 40th International Conference on Machine Learning》(PMLR 202), 10835–10866. [②] 作者对奖励模型的过优化做了定量刻画，给出真实表现随对代理奖励优化程度变化的标度律式规律：优化超过某点后，代理得分仍升而真实表现转跌。它把古德哈特式失真从定性观察推进到可测量的曲线，是本章过优化论证最实证的一块。

标定：把二值判决换成概率，并以严格适当评分约束之

15. G. W. Brier (1950).「Verification of Forecasts Expressed in Terms of Probability」.《Monthly Weather Review》, 78(1), 1–3. [②] 布莱尔提出了一个用于评价概率预报的评分（即后来的 Brier 评分），把「报了多大把握、最终是否发生」纳入可计算的考核。它是标定与严格适当评分体系的起点，本章关于「概率可检验」的论证由此发端。

16. L. J. Savage (1971).「Elicitation of Personal Probabilities and Expectations」.《Journal of the American Statistical Association》, 66(336), 783–801. [②] 萨维奇研究如何设计评分与激励，使人愿意如实报出自己的主观概率与期望。它为「适当评分规则诱出真实概率」奠定了理论基础，对应本章那句关键设计：讲真话不再靠自觉，而被评分规则的数学结构所强制。

17. A. H. Murphy (1973).「A New Vector Partition of the Probability Score」.《Journal of Applied Meteorology》, 12(4), 595–600. [②] 墨菲把 Brier 评分分解为可靠性、分辨率与不确定性三个分量，让人能分别看清预报哪里失标、哪里有区分

力。这一分解是标定概念的量化骨架，本章谈「标定」与「锐度」的区分，正建立在这种拆解之上。

18. M. H. DeGroot & S. E. Fienberg (1983).「The Comparison and Evaluation of Forecasters」.《Journal of the Royal Statistical Society: Series D (The Statistician)》, 32(1–2), 12–22. [②] 德格鲁特与芬伯格对预测者的比较与评价做了系统处理，明确区分了标定（calibration）与精炼/锐度（refinement），并给出据此排序预测者的框架。它是本章标定论证的核心理论来源，读者可在此看到标定作为可检验认识对象的严格表述。

19. A. P. Dawid (1982).「The Well-Calibrated Bayesian」.《Journal of the American Statistical Association》, 77(379), 605–610. [②] 达维德证明，一个连贯的贝叶斯主体在自己的主观信念下会渐近地自我标定，即长期看其概率断言与实际频率相符。本章用它说明标定并非外加的苛求，而可以是理性更新的内在产物。

20. D. Oakes (1985).「Self-Calibrating Priors Do Not Exist」.《Journal of the American Statistical Association》, 80(390), 339–342. [②] 奥克斯指出，不存在一个先验能保证对所有数据序列都自我标定，从而给达维德式的乐观结果划出了边界。它与 Dawid (1982) 及 Foster-Vohra 的可达性结果形成反向制衡，是本章「标定有其极限」一笔的依据。

21. M. J. Schervish (1989).「A General Method for Comparing Probability Assessors」.《The Annals of Statistics》, 17(4), 1856–1879. [②] 舍尔维什给出比较概率评估者的一般方法，把各种适当评分规则纳入统一的比较框架，作为其中的特例。它对标定理论起到集成与整理的作用，便于读者把零散的评分规则放进同一张图里看。

22. D. P. Foster & R. V. Vohra (1998).「Asymptotic Calibration」.《Biometrika》, 85(2), 379–390. [②] 福斯特与沃赫拉证明，即便面对任意（甚至对抗性）的结果序列，也存在一种预测策

略能渐近达到标定。这是标定可达性的关键定理，本章据此说明标定是一个比真假判决更弱、却切实可达的认识目标。

23. T. Gneiting & A. E. Raftery (2007).「Strictly Proper Scoring Rules, Prediction, and Estimation」.《Journal of the American Statistical Association》, 102(477), 359–378. [②] 这是严格适当评分规则的权威综述：系统整理了哪些评分函数能使如实报告恰好成为期望得分最优之策，并把它们与预测、估计联系起来。它是本章标定论证的理论支柱，读者要理解「讲真话被数学结构强制」可读此篇。

24. T. Gneiting, F. Balabdaoui & A. E. Raftery (2007).「Probabilistic Forecasts, Calibration and Sharpness」.《Journal of the Royal Statistical Society: Series B (Statistical Methodology)》, 69(2), 243–268. [②] 作者提出概率预测的现代框架，把目标概括为「受制于标定，越锐越好」：先要求预测标定，再在标定的前提下尽量提高锐度。本章关于如何评判一个概率预测好坏的标准，直接采用这一框架。

25. C. Guo, G. Pleiss, Y. Sun & K. Q. Weinberger (2017).「On Calibration of Modern Neural Networks」.《Proceedings of the 34th International Conference on Machine Learning》(PMLR 70), 1321–1330. [②] 作者发现现代深度神经网络虽然往往更准，却常常失标，置信度系统性地偏离真实正确率，并提出温度缩放等简单的重新校准方法。它是标定问题在机器学习侧的代表作，正对应本章「现代神经网络恰恰常常失标，于是需要重新校准」。

26. V. Vovk, A. Gammerman & G. Shafer (2005).《Algorithmic Learning in a Random World》. Springer. [②] 这是保形预测（conformal prediction）的奠基专著：它不给出单点判断，而构造带有覆盖保证的预测集，使真值落入集合的概率有可控的下界。本章用它作为标定思想在机器学习中的一种实现，给读者一个「带自身可靠性保证的预测」的范例。

判断、预测与替换动作的方法论根

27. P. E. Tetlock (2005).《Expert Political Judgment: How Good Is It? How Can We Know?》. Princeton University Press. [①②] 泰特洛克对专家政治预测做了长达多年的大规模追踪，发现众多专家的长期预测准确度并不出色，且常逊于简单的外推基准。它是把专家判断放到可检验框架里加以考核的代表作，对本章「按标定的概率行动、而非迷信权威断言」给出经验支撑。

28. P. E. Tetlock & D. Gardner (2015).《Superforecasting: The Art and Science of Prediction》. Crown. [①④] 本书把 IARPA 预测锦标赛的研究成果通俗化，刻画了表现突出的「超级预测者」如何拆解问题、给出概率、并随证据不断微调。它偏向实践，讲的正是如何在无法验证的世界里做出可被标定检验的判断，适合读者据以训练自己的预测习惯。

29. G. E. P. Box (1976).「Science and Statistics」.《Journal of the American Statistical Association》, 71(356), 791–799. [②③] 这是「所有模型都是错的，但有些有用」一语的出处。博克斯主张统计建模是科学探究的迭代过程，不应追求绝对正确而应追求有用与可改进。它正服务于本章的一组败法对照：忠实却不易处理，还是可处理但只是近似。

30. G. Pólya (1945).《How to Solve It: A New Aspect of Mathematical Method》. Princeton University Press. [②④] 波利亚总结了一套解题启发法，其中一条便是「先解一个相关而更易的问题」，再借它逼近原题。这正是本章标题这一替换动作的方法论原型，读者可把代理替换看作把这条古老的解题术推广到无法直接验证的场景。

31. H. A. Simon (1956).「Rational Choice and the Structure of the Environment」.《Psychological Review》, 63(2), 129–138. [②④] 西蒙提出有限理性与满意化：在能力与信息有限时，主体并不求最优，而是搜到一个「足够好」的方案即停。它为「用足够好的代理取代难以企及的最优」提供了理论根

据，是本章替换动作在决策科学里的根。

32. D. Kahneman & S. Frederick (2002).「Representativeness Revisited: Attribute Substitution in Intuitive Judgment」. 收入 T. Gilovich, D. Griffin & D. Kahneman (编)《Heuristics and Biases: The Psychology of Intuitive Judgment》, 49–81. Cambridge University Press. [②] 卡尼曼与弗雷德里克提出「属性替换」：当目标属性难以评估时，人会不自觉地用一个更易评估的属性顶替它来作答。这正是本章代理替换机制的心理学孪生，说明换问题这一招不只是工程策略，也是人类直觉的默认运作方式。

第 12 章　管住后果

> **论点**：当你无法防止错误，就经营它的后果。缩小一个错误的、未经验证的东西能造成的破坏（衰减，事前），并确保万一出错你会发现（留痕，事后）。

前面三对招，再放低标准，也都还在努力把事情做对。这最后一对，索性承认你做不对，转而经营失败。既然防不住错误，那就缩小它能造成的破坏（衰减，事前），并确保它一旦发生你查得到（留痕，事后）。

衰减：缩小爆炸半径

第一招的纯形式：放弃保证那个未经验证的东西不出错，转而把它出错时的波及范围，事前就圈死。

这是计算机安全最深的家底。萨尔策与施罗德 1975 年的最小权限（principle of least privilege）[1]、兰普森 1973 年的围堵（confinement）[2]、丹尼斯与范霍恩 1966 年的能力机制（capability）[7]，到丹宁 1976 年的信息流格（information flow lattice）[5]、贝尔-拉帕杜拉[3] 与比巴[4] 的安全模型，主旨一致：只给一个组件完成本职所必需的最小能力，其余一概不给，这样它就算被攻破或出错，也掀不起大浪。沙箱（sandbox，戈德堡等人 1996）[9]、职责分离、纵深防御，都是它的化身。系统可靠性工程里，它是熔断器与隔板（奈加德的《Release It!》[15]）、是爆炸半径（blast radius）设计、是金丝雀发布与错误预算（error budget，谷歌 SRE[30]）；金融里，它是头寸限额与止损；塔勒布的反脆弱（antifragility）[17]，讲的也是把下行限死。

统一的观念是：把担子从「让它不出错」（那需要你没有的验证）转移到「让它出错也扛得住」。一个常见的量化直觉是纵深防御（defense in depth），若 k 层防护各自独立地以概率 p 失守，全部同时失守的概率是

$$p^k,$$

随层数指数下降。但请立刻接上第 10 章那个警告：这个 p^k 只在各层失效相互独立时成立。若各层栽在同一个弱点上（同一个被绕过的内核、同一个管理员口令），相关性会让纵深防御瞬间退化成单层。2011 年的福岛核事故就是这道理的写照：电站本有主电源加备用柴油发电机的多重冗余，但一场海啸把它们一并淹没，本该相互独立的几道防线栽在同一个原因上，纵深防御形同虚设。

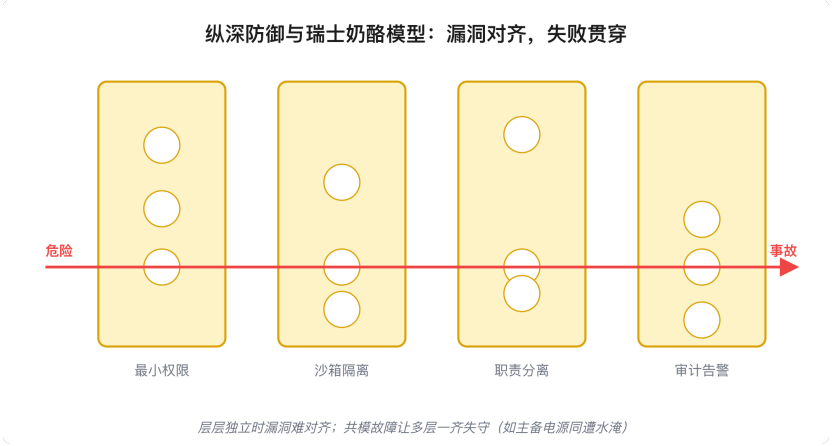

图 10: 纵深防御与瑞士奶酪模型：漏洞对齐，失败贯穿

它的标准败法正是这里：被绕过的衰减。沙箱有逃逸，权限会悄悄蔓延，看似层层设防，实则各层共用一道暗门。另一种较少被提的败法是防护过度，把正常功能也堵死，结果人们反而绕过它来干活，安全反而形同虚设。

留痕：让错误事后现形

第二招的纯形式：防不住的，就让它一旦发生必被发现。把检查从事前改为事后。

它最硬核的技术，来自密码学。默克尔 1980 年的哈希树（Merkle tree）[18]、哈伯与斯托尔奈塔 1991 年的链式时间戳[19]，让一份记录一旦写下就无法被悄悄篡改，任何改动都会在校验时暴露；克罗斯比与瓦拉赫 2009 年的防篡改日志（tamper-evident log）[23]、施奈尔与凯尔西 1998 年在不可信机器上保护日志[20]、贝拉雷与迈纳 1999 年的前向安全签名（forward-secure signature）[22]，把这套做得更牢；证书透明度（Certificate Transparency，RFC 6962）[24] 和中本聪 2008 年的比特币[27]，本质都是一本全球范围、只能追加、人人可验的审计账。核对一条记录是否在这样一棵树里，代价只有 $O(\log n)$，又是第 2 章那道「验比造廉」的红利。

而这一招其实古老得多。复式记账（double-entry bookkeeping）就是人类最早的防篡改账本之一，索尔在《清算》[29] 里论证，算得清自己账目的能力，与国家的兴衰直接相关。现代财务审计、独立稽核，都是同一姿势。科学里，它是预注册（preregistration，诺塞克 2018）[33] 与可复现（呼应第 3 章的复制危机）：把假说和方法在看到数据前就登记下来，事后无法移动靶子。

统一的观念是：放弃「事前阻止坏事」（要验证），改为「事后必能发现坏事」（只要一本忠实的账）。它的好处是双重的，既让错误可被纠正，也因为「跑不掉」而产生威慑力。

它的标准败法也只有一条，却极常见：无人响应的检测。没人去读的审计日志、被一律忽略的告警，等于没有。2017 年的 Equifax 数据泄露是教科书级的案例：一个已知漏洞迟迟没有补丁，入侵者在系统里潜伏了约七十六天才被察觉，约一亿四千七百万人的个人信息就此外泄。痕迹其实都在日志里，只是没人看。检测而不响应，是形同虚设。（另一条隐患是日志本身可被篡改，这正是上面那些密码学手段要堵的。）

八招齐了：第三部的收束

把这最后一对并看：衰减在事前缩小失败的代价，留痕在事后保证失败被发现。它们都不再试图让那个未经验证的东西正确，而是改造失败本身的形貌，一个压低爆炸半径，一个把检查改为事后。

到这里，八招集齐，四对成双：

- **压缩未知**（第 9 章）：证书与界、最优筛查。
- **借来的判断**（第 10 章）：神谕入回路、冗余共识。
- **换一个能处理的问题**（第 11 章）：代理替换、标定。
- **管住后果**（第 12 章）：衰减围栏、留痕审计。

这就是那张对照表，本书的载荷。它在四个现场加科学中反复以不同的行话出现，却始终是这八样。第 4 章立下的铁律，每一招都尽量交代了它的机制、跨域形态与标准败法，而非仅凭表面相似。

但一个尖锐的问题悬而未决：为什么偏偏是这八招？是我凑出来的一张清单，还是它们各自对应着某种更基本、躲不开的东西？如果只是清单，这本书顶多是个有用的归类手册；如果背后真有结构，那「收敛」才算被解释。第四部去追这个问题，先尝试把八招挂到一个共同的骨架上，再正面清算：这究竟是一条定律，还是一个很强的经验模式。

参考文献

> 落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. J. Saltzer & M. Schroeder (1975). «The Protection of Information in Computer Systems» . Proceedings of the IEEE. [②] 这篇综述把保护机制的设计原则系统化，其中最小权限原则成为「衰减」一招的源头：只给一个组件完成本职所必需的能力，其余一概不给。本章「缩小爆炸半径」的整个思路即由此发端，是理解为何要事前圈死失败范围的第一篇必读文献。

2. B. Lampson (1973).「A Note on the Confinement Problem」. Communications of the ACM. [②] 兰普森提出「围堵问题」：如何确保一个被调用的程序无法泄露或滥用它所接触的信息，包括隐蔽信道这类难堵的旁路。它给「把出错的东西关进笼子」立下了精确的问题陈述，正是本章衰减一招要解决的核心。
3. D. Bell & L. LaPadula (1973).《Secure Computer Systems: Mathematical Foundations》. The MITRE Corporation. [②] 贝尔-拉帕杜拉模型用形式化方式刻画机密性：信息只能由低密级流向同级或更高密级，著名的「不上读、不下写」规则即出于此。它示范了如何把「失败的波及范围」写成可证明的格结构，是安全模型理论化的奠基之作。
4. K. Biba (1977).《Integrity Considerations for Secure Computer Systems》. The MITRE Corporation. [②] 比巴模型是贝尔-拉帕杜拉的对偶，关注完整性而非机密性：信息只能由高可信向低可信流动，以防低可信数据污染关键组件。两者并看，说明同一套格论框架可以从两个方向限定失败的扩散。
5. D. Denning (1976).「A Lattice Model of Secure Information Flow」. Communications of the ACM. [②] 丹宁把信息流安全统一进一个格模型：为数据标上安全标签，要求流动只能沿格的偏序方向进行，从而在编译期或运行期静态地约束信息能去哪里。它为前面几种安全模型提供了共同的数学语言，是信息流控制的理论核心。
6. D. Clark & D. Wilson (1987).「A Comparison of Commercial and Military Computer Security Policies」. IEEE Symposium on Security and Privacy. [②] 克拉克与威尔逊指出商业场景更看重完整性而非军方式的机密性，并提出以良构事务和职责分离为核心的完整性模型。它把「衰减」从军用密级扩展到商业账务等日常场景，说明限定失败范围的形态随领域而变。
7. J. Dennis & E. Van Horn (1966).「Programming Semantics for Multiprogrammed Computations」. Communications of the ACM. [②] 这篇早期论文提出了能力（capability）的概念：访问权以不可伪造的令牌形式直接附着在引用上，持有令

牌才能操作对象。它是能力安全模型的源头，为「只给必需的最小能力」提供了机制层面的实现路径。

8. N. Provos, M. Friedl & P. Honeyman (2003).「Preventing Privilege Escalation」. 12th USENIX Security Symposium. [②] 作者们讨论如何用权限分离把特权操作隔离到极小的、受信的代码段，使主体程序即便被攻破也只能在低权限下活动，OpenSSH 的特权分离是其代表实践。它把最小权限落到了真实系统的工程细节上。
9. I. Goldberg, D. Wagner, R. Thomas & E. Brewer (1996).「A Secure Environment for Untrusted Helper Applications」. 6th USENIX Security Symposium. [②] 这篇论文介绍 Janus，用系统调用拦截为不可信程序构造受限的运行环境，是用户态沙箱的早期范例。本章把沙箱列为衰减一招的化身，此文正是沙箱思想的代表性出处。
10. C. Perrow (1984).《Normal Accidents: Living with High-Risk Technologies》. Basic Books. [②①] 佩罗提出「正常事故」论：在高度复杂且紧密耦合的系统里，灾难性事故不是偶然而是结构性必然，无法靠加防护根除。它从反面支撑本章的立场：当错误防不胜防，重心就该移到经营后果而非妄图杜绝失败。
11. N. Leveson (2011).《Engineering a Safer World: Systems Thinking Applied to Safety》. MIT Press. [②] 莱韦森用系统论重构安全工程，提出 STAMP 模型，把事故看成控制结构失效而非单一部件故障，强调用约束去限定危险状态。它为「事前圈死失败范围」提供了系统层面的方法论。
12. J. Reason (1990).《Human Error》. Cambridge University Press. [②] 里森系统分析人因失误，提出著名的「瑞士奶酪模型」：每层防护都有漏洞，唯当多层漏洞偶然对齐时事故才贯穿而出。它正是本章纵深防御直觉的来源，也提醒各层漏洞一旦相关，多层就退化成单层。
13. E. Hollnagel, D. Woods & N. Leveson (2006).《Resilience Engineering: Concepts and Precepts》. Ashgate. [②④] 这本文集奠定「韧性工程」：系统的安全不在于消灭故障，而在于具备吸收扰动、在失败后仍维持运转和恢复的能力。它与本

章主旨高度契合，把重心从「不出错」明确转向「出错也扛得住」。

14. A. Avizienis, J.-C. Laprie, B. Randell & C. Landwehr (2004).「Basic Concepts and Taxonomy of Dependable and Secure Computing」. IEEE Transactions on Dependable and Secure Computing. [②] 这篇被广泛引用的分类学厘清了故障、错误、失效的链条，以及容错、防错、查错等手段的关系。它为本章讨论的各种衰减与留痕手段提供了一套公认的术语框架，适合作为概念校准的参考。
15. M. Nygard (2007).《Release It! Design and Deploy Production-Ready Software》. Pragmatic Bookshelf. [②④] 奈加德把可靠性工程写成实战手册，提出熔断器、隔板、超时、舱壁隔离等稳定性模式，以阻止局部故障级联成全局崩溃。本章正文以它为爆炸半径设计的代表，是把衰减思想用于生产系统的直接读物。
16. R. Anderson (2020).《Security Engineering: A Guide to Building Dependable Distributed Systems》（第三版）. Wiley. [②④] 安德森这本巨著横跨密码学、访问控制、经济激励到现实攻防，是安全工程的权威综合教材。本章涉及的最小权限、审计、防篡改等几乎所有主题都能在其中找到更完整的展开，适合作为通读底本。
17. N. N. Taleb (2012).《Antifragile: Things That Gain from Disorder》. Random House. [④] 塔勒布提出「反脆弱」：有些系统不只在波动中存活，还从波动中受益，关键在限死下行、保留上行。本章引它说明衰减的极致姿态就是把损失封顶，是从风险经营角度理解「管住后果」的视角。
18. R. Merkle (1980).「Protocols for Public Key Cryptosystems」. IEEE Symposium on Security and Privacy. [②] 默克尔在此提出了用哈希树（默克尔树）做树状认证的思想：把大量数据归并成一个根哈希，任一项的真伪只需 $O(\log n)$ 的路径即可校验。它是本章「留痕」一招的技术基石，也是后来证书透明度与区块链的共同祖先。
19. S. Haber & W. S. Stornetta (1991).「How to Time-Stamp a Digital Document」. Journal of Cryptology. [②] 两位作

者提出用哈希把文档时间戳串成链，使任何事后篡改都会破坏链的连续性而暴露。这是链式防篡改记录的开创性工作，直接启发了后来的区块链结构，是理解留痕为何「改不掉」的关键。

20. B. Schneier & J. Kelsey (1998).「Cryptographic Support for Secure Logs on Untrusted Machines」. 7th USENIX Security Symposium. [②] 本文设计了在可能被攻陷的机器上保护日志的方案：即便攻击者事后取得控制权，也无法不被察觉地删改此前的记录。它把防篡改日志推进到不可信环境，是本章留痕一招的硬核技术之一。
21. B. Schneier & J. Kelsey (1999).「Secure Audit Logs to Support Computer Forensics」. ACM Transactions on Information and System Security. [②] 这是上一篇工作的期刊版扩展，更完整地论述了支持取证的安全审计日志构造。它说明留痕不仅要忠实记录，还要在事后能经得起对抗性的核验，对应本章「让错误事后现形」的目标。
22. M. Bellare & S. Miner (1999).「A Forward-Secure Digital Signature Scheme」. CRYPTO '99. [②] 贝拉雷与迈纳提出前向安全签名：密钥定期演进，即使当前密钥泄露，攻击者也无法伪造此前时段的签名。它为留痕提供了关键保障，使过去的记录在私钥失守后仍不可被冒名篡改。
23. S. Crosby & D. Wallach (2009).「Efficient Data Structures for Tamper-Evident Logging」. 18th USENIX Security Symposium. [②] 作者们设计了可高效追加与审计的防篡改日志结构，让验证者无需信任日志服务器即可确认记录的完整与一致。它把前述密码学手段整合成可落地的数据结构，是留痕一招走向工程化的代表作。
24. B. Laurie, A. Langley & E. Kasper (2013).《RFC 6962: Certificate Transparency》. IETF. [②] 证书透明度用公开、只能追加、可被任何人审计的默克尔日志记录所有签发的 TLS 证书，使误签或恶意证书无所遁形。它是本章「全球范围、人人可验的审计账」的现实样板，展示留痕如何规模化部署。
25. L. Lamport, R. Shostak & M. Pease (1982).「The Byzantine Generals Problem」. ACM Transactions on Programming

Languages and Systems. [②] 这篇经典论文形式化了拜占庭容错问题：在部分节点可能任意作恶时，诚实节点如何就一个值达成一致，并给出了可容错节点数的理论界。它是公开可验账本所依赖的共识理论根基，被本章列入「理论上被研究过的东西」。

26. M. Castro & B. Liskov (1999).「Practical Byzantine Fault Tolerance」. 3rd USENIX Symposium on Operating Systems Design and Implementation（OSDI）. [②] 卡斯特罗与利斯科夫给出第一个在真实异步网络中实用的拜占庭容错算法 PBFT，把理论上的共识做到了可工程化的性能。它说明一本人人可验、容忍作恶节点的账本并非空想，为留痕的可信基础提供了实现支撑。
27. S. Nakamoto (2008).《Bitcoin: A Peer-to-Peer Electronic Cash System》. 白皮书. [②] 中本聪的白皮书提出比特币：用工作量证明驱动一条去中心化、只能追加的区块链，让无需互信的各方就交易历史达成共识。本章视其本质为一本全球可验的审计账，是留痕思想在开放网络上的极端实现。
28. D. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler & G. Sussman (2008).「Information Accountability」. Communications of the ACM. [②④] 作者们主张从「事前封锁访问」转向「事后问责」：允许信息流动，但要求用途可被审计、违规可被追溯并追究责任。这与本章把检查从事前挪到事后的「留痕」姿态完全同构，是该理念在隐私治理上的纲领性表述。
29. J. Soll (2014).《The Reckoning: Financial Accountability and the Rise and Fall of Nations》. Basic Books. [①] 索尔以财务史论证：一个政权能否算清并如实呈现自己的账目，与其兴衰直接相关，复式记账是其中关键的问责技术。它把留痕的脉络上溯到人类最早的防篡改账本，说明审计账的力量由来已久。
30. B. Beyer, C. Jones, J. Petoff & N. Murphy (2016).《Site Reliability Engineering: How Google Runs Production Systems》. O'Reilly. [④] 这本书系统介绍谷歌的 SRE 实践，包括错误预算、金丝雀发布、监控告警与可控的故障演练。本章

借其错误预算与金丝雀发布说明衰减如何在大规模生产中制度化，是把「管住后果」工程化的现代范本。

31. J. Ioannidis (2005).「Why Most Published Research Findings Are False」. PLoS Medicine. [③] 约阿尼迪斯用统计建模论证：在低先验、小样本、多重比较与研究自由度过大的条件下，大量已发表的研究结论很可能为假阳性。这是一篇分析性论证而非实证复制研究，为本章提到的预注册与可复现给出了问题诊断。
32. Open Science Collaboration (2015).「Estimating the Reproducibility of Psychological Science」. Science. [③] 这是一项大规模实证：多组研究者尝试复制上百项心理学研究，结果相当一部分未能复现。它把约阿尼迪斯的理论担忧落成可见的数据，是「复制危机」的标志性证据，呼应本章对事后可验的重视。
33. B. Nosek, C. Ebersole, A. DeHaven & D. Mellor (2018).「The Preregistration Revolution」. PNAS. [③] 诺塞克等人倡导预注册：在看到数据之前就公开登记假说与分析方法，使探索性与验证性研究分离，事后无法移动靶子。它是科学领域的「留痕」实践，把检验从事前的信任挪到事后的核对，与本章主旨直接对应。

第四部　杠杆

第 13 章　八根杠杆

> **论点**：八招不是随意的清单；每一招拉动风险与信息分解里一根不同的杠杆，这正是它们让人觉得「齐了」的原因。

第三部交出了那张对照表：八招，四对，在四个现场加科学里反复以不同行话出现。但清单再齐整也只是清单。这一章要追问的是：为什么偏偏是这八招？是我凑出来的，还是它们各自卡在某个躲不开的位置上？如果是后者，收敛才算被解释，否则本书顶多是一本好用的归类手册。

我要提出一个候选的解释。先把丑话说在前头：它是一个组织结构，不是一个证明。读完整章，请带着第 14 章那把怀疑的刀。

一个粗糙的分解

把「在不可验证下行动」剥到最简，你真正在管理的，是风险。借决策论（decision theory）的老话（瓦尔德[1]、萨维奇[4]、冯 · 诺依曼与摩根斯特恩[3]），风险可以粗略地写成

$$\text{Risk} \;\approx\; \Pr(\text{fail}) \;\times\; \text{Cost}(\text{fail}),$$

而这一切，是在一个信息预算 B 之下进行的，你能用来削减不确定的查验、样本、算力、时间，都是有限的。

这个式子看着简单，关键在于：它右边能被你下手的地方是可数的

几处。你要么动「失败」的定义本身，要么动失败的概率，要么动你对那个概率的认知，要么动失败的代价，要么动这笔信息预算怎么花，要么动检查发生的时间。我的命题是：八招恰好一招一个位置，再没有第九个空位可填。

八招，八个位置

把每一招对到它所拉动的那根杠杆：

招	它拉动的杠杆	在风险分解中的位置
代理替换（proxy substitution）	改变你度量、优化的目标	改写「失败」的定义本身
证书（certificate）	在一个切片上把不确定压到有保证的界内	在局部将 Pr(fail) 压近零
神谕入回路	引进你单独不具备的验证能力	借外力降 Pr(fail)
冗余共识	让多个判断的失败去相关	降联合失败概率 Pr(all fail)
最优筛查	把信息预算花在边际收益最高处	分配 B，最大化对不确定的削减
标定（calibration）	给残余风险定一个实在的价	让 Pr(fail) 变成已知、可据以下注
衰减围栏	缩小爆炸半径	降 Cost(fail)
留痕审计	把检查从事前挪到事后	改检查的时间位置，把不可恢复的失败变可恢复

读这张表，那个「凑出来的清单」的感觉应当松动一些。八招不是八件随手收集的工具，它们分占了「Pr、对 Pr 的认知、代价、预算分配、检查时点、目标定义」这几处，几乎一一占满了那个分解式能下手的地方。命题于是可以这样讲：若这些确实就是全部的杠杆，那这套招就是完整的，收敛也就被解释了，任何有能力的主体迟早都会重新发现它们，因为除此之外没有别的可拉。

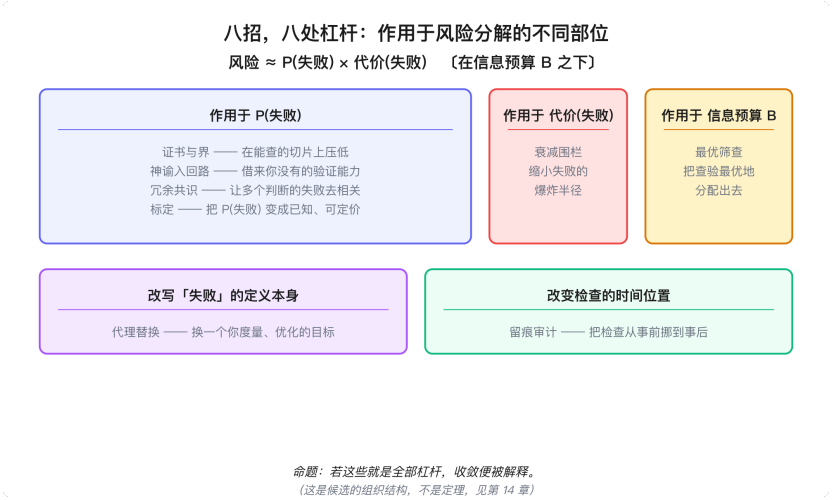

图 11: 八招映射到风险分解的不同部位（候选的组织结构，非定理）

为什么这能跨越基质

如果上面成立，它顺带解释了本书最初那个谜：为什么数学家、工程师、组织会不约而同地采取这八招。

马尔[17] 在研究视觉时区分过三个层次：计算层（computational level，要解决什么问题、受什么约束）、算法层（algorithmic level，用什么表示和过程）、实现层（implementational level，落在什么硬件上）。八招活在计算层，是「给定不可验证这个约束，逻辑上还能动哪几处」的答案，而这个答案不依赖你是碳基的数学家、硅基的程序，还是由人组成的官僚机构。基质（substrate）千差万别，计算层的约束却是同一个，于是应对收敛。西蒙[12] 的有限理性（bounded rationality）、他的「人工科学」[13]（the sciences of the artificial），讲的正是这种由环境约束而非由主体内部塑造的行为。

这里还得请出无免费午餐定理（no free lunch theorem，沃尔珀特与麦克里迪[25]）。它说：在所有可能问题上平均，没有哪个方法优于另一个。这把刀两面都割。一面，它支持本书的克制，没有万能解，你必须借问题的具体结构来选杠杆，这正是为什么这五种处境要分开对待。另一面，它也警告：任何宣称「找到了统一钥匙」的

人，包括我，都该收敛一点傲气。当你连失败概率都钉不住时，杠杆还会长出稳健版本，吉尔博亚与施迈德勒[20] 的极大极小期望效用（maxmin expected utility）、汉森与萨金特[30] 的稳健控制（robust control）、奈特式不确定性（Knightian uncertainty）下的决策，都是在 Pr 本身都模糊（ambiguity）时仍要按最坏情形布防的招法。

一句必须放大的强声明

现在把丑话放大。

上面这套是一个候选的组织结构，不是一个定理。那个风险分解是非形式的，我没有给出严格的主体与环境模型，也就无法证明最优策略恰好是这八根杠杆。「这些就是全部的杠杆」是一句断言，不是一个已确立的结果。我没有证据说这张表是穷尽的，也无法排除它只是事后框架，一个足够灵活、能把许多套招都塞进去的叙事。表里某些归位（比如冗余既降联合失败、又像是一种特殊的筛查）甚至有重叠，这本身就说明这个分解还不够干净。

我把它放在这里是因为它有组织力、有解释上的吸引力，而不是因为它被证明了。它够得上一个好猜想的标准：清晰、可反驳、能统起大量现象。但它还没够上定理。

那么，最后那个问题就躲不掉了：这种跨领域的收敛，到底是某种东西逼出来的一条定律，还是仅仅一个很强、却终究是经验的模式？下一章，正面清算它。

参考文献

> 落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. A. Wald (1950). 《Statistical Decision Functions》. John Wiley & Sons. [②④] 瓦尔德把统计推断重铸为一个对抗自然

的决策问题：决策者要在风险（损失的期望）下选择策略，并以极小化最大风险的极大极小准则来对付未知的状态。这本书奠定了统计决策论，本章那个「风险约等于失败概率乘以代价」的粗糙分解，其学理根子正在此处。

2. A. Wald (1939).「Contributions to the Theory of Statistical Estimation and Testing Hypotheses」.《The Annals of Mathematical Statistics》, 10(4), 299-326. [②] 这是瓦尔德把估计与检验统一进损失函数框架的早期论文，先于他后来的专著，已引入风险函数与最不利先验的思路。它标记了「用决策语言谈统计」这一转向的起点，对理解本章为何从决策论借力很有价值。
3. J. von Neumann & O. Morgenstern (1944).《Theory of Games and Economic Behavior》. Princeton University Press. [②] 冯·诺依曼与摩根斯特恩创立博弈论，并由一组公理推出期望效用定理：满足理性公理的偏好可由对效用的期望最大化来表示。这是「在不确定下下注」的规范基准，本章引它作为风险计算的源头之一。
4. L. J. Savage (1954).《The Foundations of Statistics》. John Wiley & Sons. [②④] 萨维奇用一套关于行动偏好的公理，同时导出主观概率与效用，把贝叶斯决策论建在个人化概率之上。它是「主观概率加期望效用」这一现代框架的奠基之作，也是后文埃尔斯伯格悖论所要挑战的靶子。
5. F. H. Knight (1921).《Risk, Uncertainty and Profit》. Houghton Mifflin. [②] 奈特区分了可量化的「风险」与无法赋予概率的「不确定性」，并主张企业利润正来自承担后者。这条区分是本章谈「连失败概率都钉不住」时的关键概念来源，奈特式不确定性贯穿后面一系列稳健决策文献。
6. J. M. Keynes (1921).《A Treatise on Probability》. Macmillan. [②] 凯恩斯发展了一种逻辑解释下的概率观，把概率看作命题间的理性信念度，并强调许多概率既非数值化、也未必可比较。它为后来的非可加、不精确概率埋下伏笔，提醒读者概率本身的认知地位远比公式复杂。
7. F. P. Ramsey (1931).「Truth and Probability」.《The Foundations of Mathematics and other Logical Essays》(R. B.

Braithwaite 编). Kegan Paul, Trench, Trubner & Co., 156-198. [②] 拉姆齐最早论证：一个人的信念度可由其下注行为操作性地测得，而避免被稳赚组合套利（荷兰赌）要求这些信念度服从概率公理。这是主观概率的开山之作，为本章「给残余风险定一个实在的价」提供了哲学与操作上的依据。

8. B. de Finetti (1937). 「La prévision: ses lois logiques, ses sources subjectives」.《Annales de l'Institut Henri Poincaré》, 7(1), 1-68. [②] 德·菲内蒂提出主观概率，并以荷兰赌论证与可交换性的表示定理给它撑腰，论证概率「只是」一致的个人信念度。它与拉姆齐共同构成贝叶斯主义的基石，是理解标定与如实定价的必读源头。
9. F. J. Anscombe & R. J. Aumann (1963). 「A Definition of Subjective Probability」.《The Annals of Mathematical Statistics》, 34(1), 199-205. [②] 两位作者借引入客观随机化装置（如轮盘彩票），给出一套比萨维奇更简洁的主观概率与效用公理化。这一框架后来成为模糊决策理论的标准舞台，本章引用的多篇模糊厌恶文献都在它上面搭建。
10. D. Ellsberg (1961). 「Risk, Ambiguity, and the Savage Axioms」.《The Quarterly Journal of Economics》, 75(4), 643-669. [②] 埃尔斯伯格用两个著名的抽球实验展示：人们系统性地偏好已知概率而回避「模糊」，这种行为违反萨维奇公理，无法用任何单一主观概率解释。它把模糊（ambiguity）确立为独立现象，是后续稳健与多先验理论的直接动因。
11. R. D. Luce & H. Raiffa (1957).《Games and Decisions: Introduction and Critical Survey》. John Wiley & Sons. [②④] 这本书是博弈论与决策论的经典导论兼批判性综述，既清晰梳理期望效用、博弈解概念，也坦诚讨论各公理的适用边界。它适合读者作为整章决策论背景的总入口，兼具体系性与批判态度。
12. H. A. Simon (1955). 「A Behavioral Model of Rational Choice」.《The Quarterly Journal of Economics》, 69(1), 99-118. [②] 西蒙在此提出有限理性与「满意即止」：受认知与信息限制的主体不去求全局最优，而是搜索到一个够好的方案就停。这是本章「计算层的约束塑造行为」论证的核心支

撑，解释了为何不同基质会收敛到同一套应对。

13. H. A. Simon (1969).《The Sciences of the Artificial》. MIT Press. [②③] 西蒙提出，人造物的行为更多由其所处环境的约束、而非内部构造决定，并主张为「设计」这门关于人工系统的学问立基。本章借它论证八招活在马尔意义上的计算层，正是由环境约束而非主体内部塑造的产物。
14. K. R. Popper (1959).《The Logic of Scientific Discovery》. Hutchinson. [②③] 波普尔系统提出证伪主义：科学理论无法被经验证实，只能被否证，可证伪性因而成为科学与非科学的分界。本章在结尾自陈那套分解「够得上好猜想、还够不上定理」，用的正是这把可反驳性的尺子。
15. A. Tversky & D. Kahneman (1974).「Judgment under Uncertainty: Heuristics and Biases」.《Science》, 185(4157), 1124-1131. [②] 特沃斯基与卡尼曼记录了人在判断概率时所用的启发式（代表性、可得性、锚定）及其带来的系统性偏差。它说明真实主体如何偏离贝叶斯理想，与标定、最优筛查等需要如实估计概率的招法形成对照。
16. D. Kahneman & A. Tversky (1979).「Prospect Theory: An Analysis of Decision under Risk」.《Econometrica》, 47(2), 263-291. [②] 前景理论提出，人按相对于参照点的得失、而非最终财富来评价结果，对损失更敏感，并以非线性方式扭曲概率权重。它是对期望效用的描述性修正，提醒读者代价与概率在真实决策中并非中性地相乘。
17. D. Marr (1982).《Vision: A Computational Investigation into the Human Representation and Processing of Visual Information》. W. H. Freeman. [②③] 马尔提出分析信息处理系统的三个层次：计算层（要解决什么问题）、算法层（用什么表示与过程）、实现层（落在什么硬件上）。本章正是借这套层次论，主张八招活在计算层，因而能跨越碳基、硅基与组织等不同基质。
18. J. O. Berger (1985).《Statistical Decision Theory and Bayesian Analysis》. Springer-Verlag. [②④] 伯杰系统整理了贝叶斯与频率派交汇处的统计决策论，涵盖损失函数、风险、可采纳性与稳健贝叶斯分析。它是把本章那个非形式风

险分解落到严谨统计语言的标准参考书。

19. D. E. Bell, H. Raiffa & A. Tversky (编) (1988). 《Decision Making: Descriptive, Normative, and Prescriptive Interactions》. Cambridge University Press. [②④] 这本文集围绕决策研究的三种取向展开：描述性（人实际怎么决策）、规范性（理性应当如何）、处方性（如何帮人决策得更好），并讨论三者如何互动。它为读者提供了一张定位各派工作的地图，呼应本章在规范与经验之间的反复掂量。
20. I. Gilboa & D. Schmeidler (1989). 「Maxmin Expected Utility with Non-Unique Prior」. 《Journal of Mathematical Economics》, 18(2), 141-153. [②④] 吉尔博亚与施迈德勒为模糊决策给出公理化：主体持有一组先验，并按其中最不利的那个来评估行动，即在先验集合上做极大极小期望效用。这正是本章所说「Pr 本身都模糊时仍按最坏情形布防」的招法，是稳健决策的代表性形式化。
21. D. Schmeidler (1989). 「Subjective Probability and Expected Utility without Additivity」. 《Econometrica》, 57(3), 571-587. [②] 施迈德勒引入非可加的主观概率（容度）与对应的乔奎特期望效用，使模糊厌恶可以被一致地表示。它与上一条同为模糊决策的奠基工作，从另一条技术路线松开了概率必须可加这一约束。
22. P. Walley (1991). 《Statistical Reasoning with Imprecise Probabilities》. Chapman and Hall. [②④] 沃利系统发展了不精确概率理论，用上下概率（或一组概率）刻画证据不足时的信念，并给出相应的一致性与推断准则。它为「连概率都钉不准」的处境提供了完整的统计语言，是标定与稳健思路的深层理论后盾。
23. G. Gigerenzer & D. G. Goldstein (1996). 「Reasoning the Fast and Frugal Way: Models of Bounded Rationality」. 《Psychological Review》, 103(4), 650-669. [②] 吉仁泽与戈尔茨坦论证，简单的「快速节俭」启发式在真实环境中往往能与复杂模型媲美甚至更好，呈现一种生态理性。它从正面补全了西蒙的有限理性，说明在信息预算有限时简捷规则为何可能恰是合理的。

24. D. H. Wolpert (1996). 「The Lack of A Priori Distinctions between Learning Algorithms」. 《Neural Computation》, 8(7), 1341-1390. [②] 沃尔珀特在监督学习上证明了无免费午餐：在所有可能目标函数上平均，任何学习算法的泛化表现都相同。它说明不存在脱离问题结构的万能学习器，是本章引用的无免费午餐论证在学习这一侧的来源。
25. D. H. Wolpert & W. G. Macready (1997). 「No Free Lunch Theorems for Optimization」. 《IEEE Transactions on Evolutionary Computation》, 1(1), 67-82. [②] 沃尔珀特与麦克里迪把无免费午餐推广到优化：在所有可能目标上平均，没有哪个优化算法优于另一个。本章用这把双刃刀，一面支持「必须借问题结构选杠杆」的克制，一面警告任何宣称找到统一钥匙的人收敛傲气。
26. I. Gilboa & D. Schmeidler (2001). 《A Theory of Case-Based Decisions》. Cambridge University Press. [②④] 两位作者提出基于案例的决策理论：当主体说不清状态空间、概率无从谈起时，决策可由对过往相似案例的回忆与类比来驱动。它为概率框架彻底失效的处境提供了另一种规范模型，拓宽了本章对「无法验证」之深度的想象。
27. T. F. Bewley (2002). 「Knightian Decision Theory. Part I」. 《Decisions in Economics and Finance》, 25(2), 79-110. [②④] 贝弗利把奈特式不确定性形式化为偏好的不完备性：当证据不足以排序两个选项时主体可以拒绝选择，并辅以维持现状的惯性假设。它给「无法定价的不确定」一个干净的公理表达，是本章奈特主题的现代接续。
28. P. Klibanoff, M. Marinacci & S. Mukerji (2005).「A Smooth Model of Decision Making under Ambiguity」. 《Econometrica》, 73(6), 1849-1892. [②] 三位作者提出模糊决策的「光滑」模型：通过对先验的二阶分布再叠一层效用函数，把模糊态度与风险态度分离，且避免极大极小那种非光滑的尖角。它让模糊厌恶的程度可调、可分析，是稳健决策谱系里更精细的一档。
29. F. Maccheroni, M. Marinacci & A. Rustichini (2006).「Ambiguity Aversion, Robustness, and the Variational Represen-

tation of Preferences」. 《Econometrica》, 74(6), 1447-1498. [②④] 作者给出变分偏好这一统一表示，把极大极小期望效用、乘子（稳健控制）偏好等都收为特例，模糊态度由一个对偏离的惩罚项刻画。它在数学上把本章提到的几路稳健招法连成一个家族，价值在于呈现其共同骨架。

30. L. P. Hansen & T. J. Sargent (2008). 《Robustness》. Princeton University Press. [②④] 汉森与萨金特把控制论中的稳健控制引入经济决策：决策者不信任手中的模型，遂针对一族邻近模型中最不利者来优化，以求对设定误差稳健。这是本章「稳健控制」一招的主要出处，展示了对最坏情形布防的动态形式。
31. P. P. Wakker (2010). 《Prospect Theory: For Risk and Ambiguity》. Cambridge University Press. [②] 瓦克尔把前景理论加以系统化与公理化，统一处理风险与模糊下的决策权重，并提供可操作的测量方法。它把描述性的前景理论与规范性的模糊理论缝在一起，是读者深入概率权重这一主题的权威专著。
32. I. Gilboa & M. Marinacci (2013). 「Ambiguity and the Bayesian Paradigm」. 《Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress》(D. Acemoglu, M. Arellano & E. Dekel 编). Cambridge University Press. [②④] 这篇综述梳理了模糊决策的整条脉络，并直面一个根本追问：贝叶斯范式在何种意义上够用、又在何处需要被多先验等模型替代。它是进入本章模糊与稳健文献群的最佳导航，态度上与本章的克制一致。
33. P. Bossaerts & C. Murawski (2017). 「Computational Complexity and Human Decision-Making」. 《Trends in Cognitive Sciences》, 21(12), 917-929. [②] 两位作者论证，许多现实决策问题在计算复杂性上本就难解（如背包等 NP 难问题），人脑的表现与策略受此硬约束塑形。它从计算复杂性角度为有限理性提供了硬证据，呼应本章「计算层的约束逼出收敛」的核心主张。

第 14 章　是定理，还是模式?

> **论点**：全书的核心清算。这跨域收敛是一条定理（某种东西迫使任何面对不可验证的有限主体走进这套招），还是仅仅一个强经验模式（我们一再看到它，却没有证明它必然如此，而且选择效应可能解释这种押韵）?

这是全书的清算之章。把那个一路扛着的问题，正面摆上台面：

这种跨领域的收敛，是一条定律（某种东西迫使任何面对不可验证（unverifiable）的有限主体（bounded agent），都必然走进这套招），还是仅仅一个很强的经验模式（我们一再看到它，却没能证明它非如此不可，而且选择效应（selection effect）也许就足以解释这种押韵）?

我会尽力把两边都说硬，包括那一边对我不利的。这一章若有偏向，应当偏向怀疑。

支持「定律」的一边

第一条线索，是上一章那个杠杆分解。如果八招确实占满了风险分解式里所有能下手的位置，那收敛就不是巧合，而是被结构逼出来的，任何有能力的主体迟早都会重新发现它们，因为别无他选。这条论证若成立，分量极重。

第二条线索，是独立的重复发现。默顿在科学社会学里研究过「多

重发现」（multiple discovery）现象：同一个想法，常被互不知情的人几乎同时各自做出[12]。微积分有牛顿与莱布尼茨，自然选择有达尔文与华莱士，电话的专利申请里贝尔与格雷竟在同一天递交，这样的例子在科学史上多得不似巧合。本书那八招，也在密码学、统计学、数论、组织理论、安全工程里反复被重新发明，而这些领域当年并不怎么通气。一个东西若总是被独立地撞见，那气味更像必然，而非借用。

第三条线索，来自科学里确凿的先例。物理学的重整化群（renormalization group）与普适类（universality class）表明，结构天差地别的系统，在临界点附近会收敛到完全相同的行为，而那里它真的是一条定理[17,18]。一个具体到惊人的事实是：液体在临界点附近的行为，和磁体在居里点附近的行为，由同一组「临界指数」（critical exponents）刻画，分子与磁矩八竿子打不着，可它们落进了同一个普适类，因为决定临界行为的是对称性与维度这些粗粒度特征，而非微观细节。维姆萨特的稳健性论证（robustness analysis）说，一个能由多条互相独立的路径反复抵达的结论，更可能是真的[20]。惠威尔 1840 年的「归纳的协同」（consilience of inductions）[11]、维格纳那篇「数学不可思议的有效性」[22]，讲的都是这种跨域汇合带来的可信。收敛，本就是科学判定「这是真东西」的古老信号之一。

支持「模式或更弱」的一边

现在说对我不利的，而且我认为这一边的分量，不比上一边轻。

最致命的一击，是那些领域其实没有看上去那么独立。它们共享一个数学底座，概率、最优化、信息论，渗透在每一个领域的根部；它们共享同一套人类认知，毕竟这些学科都是同一种大脑造出来的；而且彼此借用引用，从来不是隔绝的，香农的信息论几乎流进了所有领域，决策论的语言四处扩散。如果收敛只是因为大家都从同一个数学工具箱里取货、都被同一种心智塑造、还一直在相互模仿，那「独立重复发现」就大打折扣，所谓收敛，可能只是共同起源的回声。

第二，那个杠杆分解，很可能是事后框架。它会不会只是足够灵活，

能把许多套招都装进去？上一章那张表里已经露了马脚，冗余既被归为「降联合失败」，又像是一种特殊的筛查，归位有重叠，说明这把尺子本身不够硬。丹尼特提出过一个关键问题：一个模式什么时候算真实的，而不是被强加的[13]？判据是它的预测力与压缩力，一个真模式能让你预言新东西。我这个框架，到目前为止主要是在整理已知的招，而没有预言出一招前所未见、后被证实的新招。这是它尚未通过的考验。

第三，选择效应。我是带着「寻找收敛」的眼睛去看这些领域的，那我很可能不自觉地滤掉了不合拍的领域与反例，只留下押韵的。复制危机（replication crisis）正是最响的警钟：看起来无比稳健的经验模式，可能只是系统性偏倚的产物[29,30]。我没有理由假设自己对这种偏倚免疫。

第四，就算收敛是真的，它也未必指向一条深定律。劳丹的悲观元归纳（pessimistic meta-induction）提醒我们，历史上一个个成功的理论后来都被推翻[6]，收敛到一个模式，不等于收敛到真理。卡特赖特说基础定律其实并不如实描述世界[14]，安德森的「多即不同」（more is different）则指出，高层级自有其规律[15]，也许本书的收敛不过是一条「特殊科学」（special sciences）尺度上的规律性，而非什么基本定律。沃勒尔的结构实在论（structural realism）给了一个折中[7]：也许真正被保留下来的，是那个结构（这些杠杆本身），即便我套在它外面的说辞是错的。

要拍板，需要什么

要把这个问题真正了结，需要一样我拿不出来的东西：一个「有限主体加不可验证系统」的形式模型，外加一条定理，证明在该模型下，最优的、或唯一的策略恰好就是这几根杠杆。据我所知，这样的模型尚不存在。最接近的一条真定理是无免费午餐（no free lunch）[24]，可它偏偏指向另一个方向，没有普适占优的方法。在那个模型与那条定理出现之前，这个问题是敞开的，而我不打算假装它已经合上。

所以，把本书交付的东西如实说清：它是一个猜想，一个强的、有

用的、把边界划清了的猜想，外加一套能把许多领域串起来的共同词汇。它不是一条定理。这恰好就是本书自己一路推荐的姿态，一个标定的信念（calibrated belief），而不是一个二值的判决。

递归收尾

到这里，一件早就埋下的事终于浮现：一本论述「如何在不可验证中行动」的书，无法验证其自身的核心命题。

这听上去像个尴尬的自指（self-reference），其实是这本书最坦白的时刻。面对自己都验不了的中心论断，它没有别的选择，只能去做它通篇所描述的那件事。它陈述一个标定的信念（我认为这个收敛是真的，但给不出证明）。它给主张划清边界（这是猜想，不是定理）。它公开邀请反驳（去找一个不收敛的领域，去找那一招破坏分解的反例，那就是它的黑天鹅）。然后，它照样把话说下去，照样把这套词汇交到你手上，因为有用，纵然未经证实。

换句话说，这本书亲自演练了它自己那套招：它用了代理替换（把「证明收敛」换成「展示并组织收敛」），用了标定（给自己的把握定价），用了证伪式的留痕（白纸黑字写下可被推翻的断言）。自指系统无法在内部完全证成自己，这是一种我们早该习惯的命运[26,27,25]。但无法在内部自证，并不等于不能行动。如果本书的论点是对的，那么它这种「以自己所描述的方式来写自己」，就不是缺陷，而是一种微弱却融洽的旁证。

那么，最后一个问题随之而来。如果验证通常不可得，连这本书都只能给出一个未经证实的信念，那么我们平日里称之为「知识」的那一大堆东西，究竟是什么？下一章，把落点放在认识论上。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. T. S. Kuhn (1962).《The Structure of Scientific Revolutions》. University of Chicago Press. [③①] 库恩指出科学并非线性累积，而是在常规科学（在既有范式内解谜）与科学革命（范式更替）之间交替推进，新旧范式之间存在「不可通约性」。此书初版 1962，最初亦作为《国际统一科学百科全书》的单行本刊出。它是理解「科学如何进展」的奠基文本，也提示本章：跨域收敛究竟是被同一范式塑造，还是独立逼出的，正是本书要拷问的问题。

2. K. R. Popper (1959).《The Logic of Scientific Discovery》. Hutchinson. [③①] 波普尔系统提出证伪主义：科学理论无法被经验证实，只能被否证，可证伪性因而成为科学与非科学的分界，也成为科学进展的判据。此书为德文原著《Logik der Forschung》（1934，版权页标 1935）的英文增订本，1959 年由伦敦 Hutchinson 出版。它是 Kuhn、Lakatos、Feyerabend 后续论辩的共同起点，也为本书「公开邀请反驳、写下可被推翻的断言」这一姿态提供了根据。

3. I. Lakatos (1970).「Falsification and the Methodology of Scientific Research Programmes」. 收于 I. Lakatos & A. Musgrave (编)《Criticism and the Growth of Knowledge》. Cambridge University Press. [③] 拉卡托斯调和波普尔与库恩，提出评价单位不是单个理论而是「研究纲领」：纲领有一个被保护的硬核与可调整的保护带，若它能持续预言并兑现新事实，便是进步的，否则就是退化的。此文源自 1965 年伦敦 Bedford College 的学术讨论会，论文集 1970 年由剑桥大学出版社出版。其「进步 / 退化」判据，恰好为本章「定理还是模式」之争提供了一个可操作的方法论框架。

4. P. Feyerabend (1975).《Against Method: Outline of an Anarchistic Theory of Knowledge》. New Left Books. [③①] 费耶阿本德以「认识论的无政府主义」反对任何普适、固定的科学方法，主张在实际科学史中「怎么都行」，并指出重大突破往往恰是因为打破了既定方法论规则。此书 1975 年由伦敦 New Left Books（即后来的 Verso）初版。它构成「跨域收敛

是否被某条方法论定理逼出」的最强反方立场：若根本没有统一方法，收敛就更难被解释为必然。

5. L. Laudan (1977).《Progress and Its Problems: Towards a Theory of Scientific Growth》. University of California Press. [③] 劳丹主张衡量科学进步的标尺不是逼近真理，而是「解题效力」：一套理论解决了多少经验问题、又制造了多少概念难题。这把对科学进步的判断从形而上学的真理负担中解脱出来，落到可比较的功能指标上。它正面切合「科学如何进展」，也支持本章的克制立场，收敛到一个有用的模式，未必等于收敛到真理。

6. L. Laudan (1981).「A Confutation of Convergent Realism」. Philosophy of Science, 48(1). [③②] 劳丹以一份历史清单提出「悲观元归纳」：历史上许多曾经成功、能做出准确预言的理论，其核心词项后来被判定根本不指称任何东西（如以太、燃素），因此经验上的成功并不可靠地担保理论为真。此文载于卷 48，第 19 至 49 页。它直接质疑「跨域收敛指向真理」这一定理性主张，是本章最关键的反方文献之一。

7. J. Worrall (1989).「Structural Realism: The Best of Both Worlds?」. Dialectica, 43(1-2). [③②] 沃勒尔提出结构实在论，试图在「无奇迹论证」与「悲观元归纳」之间取中：理论更替时被保留下来的不是关于本体的描述，而是其数学结构（如菲涅耳的光学方程在以太被抛弃后仍然成立）。此文载于卷 43 第 1 至 2 期，第 99 至 124 页。它为本章给出一种折中读法，即便我套在杠杆外面的说辞是错的，真正被保留的也许正是那个结构本身。

8. I. Hacking (1983).《Representing and Intervening: Introductory Topics in the Philosophy of Natural Science》. Cambridge University Press. [③②] 哈金把科学哲学的重心从「表征」（理论如何描述世界）转向「介入」（实验如何操纵世界），提出一种实验实在论：若我们能稳定地用某个实体去干预、制造别的现象（「能喷它，它就是真的」），就有理由相信它存在。此书 1983 年由剑桥大学出版社出版。它为本章提供了「收敛

模式从何而来」的另一条解释，即收敛可能源于共同的实验实践，而非理论上的必然。

9. W. V. O. Quine (1951).「Two Dogmas of Empiricism」. The Philosophical Review, 60(1). [②③] 蒯因攻击逻辑经验主义的两条教条：分析与综合的截然二分，以及还原论。他主张知识是一张面对经验整体受检的「信念之网」，任何单个命题都无法被孤立地证实或证伪。此文载于卷 60 第 1 期，第 20 至 43 页。其确证整体论是本书核心处境的哲学根基，没有哪个论断能被单独拎出来彻底验证。

10. M. Polanyi (1958).《Personal Knowledge: Towards a Post-Critical Philosophy》. University of Chicago Press. [①④] 波兰尼提出「默会知识」：我们知道的远多于我们能言说的，一切明言的知识都依托一层无法完全形式化的个人判断与技能。科学认知因此离不开科学家的个人参与和投入。此书 1958 年由芝加哥大学出版社出版。它直接关乎本书落点，当验证无法穷尽时，科学家如何凭判断作出标定的信念并据以行动。

11. W. Whewell (1840).《The Philosophy of the Inductive Sciences, Founded Upon Their History》. John W. Parker. [③②] 惠威尔在此提出「归纳的协同」(consilience of inductions)：当一个由某类事实归纳出的理论，竟也能解释另一类原本无关的事实时，这种意外的汇合是理论为真的有力标志。此书为两卷本，1840 年由伦敦 John W. Parker 出版。它是「跨域收敛即可信」这一思路最早的方法论表述，为本章支持「定律」一方提供了历史源头。

12. R. K. Merton (1961).「Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science」. Proceedings of the American Philosophical Society, 105(5). [①③] 默顿系统考察科学史上的「多重发现」现象：同一发现常被互不知情的人几乎同时各自做出，他据此论证这类多重发现并非例外，而是科学发现的常态，发现更多取决于知识积累的状态而非个别天才。此文载于卷 105 第 5 期，第 470 至 486 页。它是「收敛是一种强经验模式」的关键社会学证据，

本章正是借它来掂量独立重复发现的分量。

13. D. C. Dennett (1991).「Real Patterns」. The Journal of Philosophy, 88(1). [②③] 丹尼特问：一个模式何时算「真实」，而非被观察者强加？他给出的判据是压缩与预测，若把数据描述为某个模式能带来真正的信息压缩、并支持对新情形的预言，这个模式就是实在的。此文载于卷 88 第 1 期，第 27 至 51 页。这正是本章「定理还是强经验模式」的核心概念工具，本章也据此承认：杠杆分解尚未预言出一招前所未见、后被证实的新招，是它仍待通过的考验。

14. N. Cartwright (1983).《How the Laws of Physics Lie》. Oxford University Press. [②③] 卡特赖特主张物理学的基础定律之所以普适，恰恰因为它们并不如实描述真实世界，越是基本的定律越要靠大量理想化与近似才能套上现象，真正描述具体系统的是局部的、唯象的定律。此书 1983 年由 Clarendon Press / 牛津大学出版社初版。它从根上质疑：本章所见的收敛背后，是否真的站着一条「定理」，还是只是模型层面的整齐。

15. P. W. Anderson (1972).「More Is Different」. Science, 177(4047). [②③] 安德森反对还原论的「构成主义」推论：即便万物都由基本粒子按基本定律构成，也不意味着从这些定律就能推出高层级的行为。每提升一个尺度，都会涌现出全新的、自成体系的规律。此文载于卷 177 第 4047 期，第 393 至 396 页（1972 年 8 月 4 日）。它支持本章一种弱化读法，本书的收敛或许只是某个「特殊科学」尺度上的规律性，而非基本定律。

16. H. A. Simon (1962).「The Architecture of Complexity」. Proceedings of the American Philosophical Society, 106(6). [②③] 西蒙论证：能稳定演化、长存的复杂系统，往往呈层级结构且「近可分解」，即子系统内部的交互远强于子系统之间，他用钟表匠的寓言说明带稳定中间件的系统更易被组装出来。此文载于卷 106 第 6 期，第 467 至 482 页。它给跨域收敛提供了又一种「共同起源」式解释，不同领域之所以撞见相似结

构，可能因为复杂系统本就受同一组架构约束。

17. K. G. Wilson (1979).「Problems in Physics with Many Scales of Length」. Scientific American, 241(2). [②③] 威尔逊面向一般读者讲解重整化群：当一个系统横跨许多长度尺度（如临界点附近），可以逐级「粗粒化」来处理，由此解释了为何微观细节天差地别的系统会落入同一「普适类」、表现出完全相同的临界行为。此文载于卷 241 第 2 期，第 158 至 179 页（1979 年 8 月）。它是本章支持「定律」一方最硬的物理学例证，因为在那里，不同系统收敛到同一行为是一条可证的定理。

18. R. W. Batterman (2002).《The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence》. Oxford University Press. [②③] 巴特曼分析物理学中的「渐近推理」：许多解释（尤其普适性现象）依赖于取某个极限（如尺度趋于零或无穷）时浮现的奇异行为，这类解释无法被简单还原到底层理论，恰恰栖身于细节消失之处。此书由牛津大学出版社出版（封面年份多标 2002，部分书评依出版前目录作 2001）。它为本章的「定律」一方提供哲学剖析：跨域收敛之所以可能成立，机制或许正在于这种与微观细节无关的渐近普适性。

19. R. Levins (1968).《Evolution in Changing Environments: Some Theoretical Explorations》. Princeton University Press. [②③] 莱文斯以一系列理论模型探讨生物如何在波动、不确定的环境中演化，引入「适应度集」等工具分析在多变环境下何种策略最优，并贯穿一种以多个简化模型逼近复杂现实的建模风格。此书 1968 年由普林斯顿大学出版社出版（Monographs in Population Biology 第 2 号）。它代表的多模型、近似式建模思路，是 Wimsatt 稳健性论证的生物学先声，呼应本章对「多条独立路径汇合」的重视。

20. W. C. Wimsatt (1981).「Robustness, Reliability, and Overdetermination」. 收于 M. B. Brewer & B. E. Collins (编)《Scientific Inquiry and the Social Sciences》. Jossey-Bass. [③

②] 维姆萨特系统阐述「稳健性分析」：若一个结论能由多条彼此独立的探测、推导或测量路径反复抵达，那它更可能是真的，而非某一手段的人为产物，这种「多重决定」是分辨真实事物与假象的关键。此文收于该论文集，第 125 至 163 页。它是本章「跨域收敛为何可信」的方法论核心，也正是支持「定律」一方所倚重的论证。

21. W. C. Wimsatt (2007).《Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality》. Harvard University Press. [③④] 维姆萨特主张把科学哲学改造成适合「有限存在者」的工具：真实的认知者算力有限、易错、受困于自身视角，因而依赖启发式、近似与分片的逼近来一点点接近现实，而非追求理想化的完备理性。此书 2007 年由哈佛大学出版社出版。它与本书的主旨高度同频，正面回答「有限主体如何在无法验证的世界里推进科学、并据此生活」。

22. E. P. Wigner (1960).「The Unreasonable Effectiveness of Mathematics in the Natural Sciences」. Communications on Pure and Applied Mathematics, 13(1). [②③] 维格纳惊叹于一个事实：为纯粹内在动机发展出的数学，竟一再精确地刻画自然规律，他称这种契合为一份「我们既不理解也不配享有的奇妙馈赠」。此文载于卷 13 第 1 期，第 1 至 14 页，源自 1959 年 Courant 讲座。它是「定理还是模式」之问的原型，数学的跨域有效到底是必然，还是一桩我们尚无法解释的巨大押韵。

23. H. Putnam (1975).《Mathematics, Matter and Method: Philosophical Papers, Volume 1》. Cambridge University Press. [②③] 这部论文集收录普特南早期的数学哲学与科学实在论文章，其中「What is Mathematical Truth?」一文给出「无奇迹论证」的经典表述：实在论是唯一不把科学的成功当作奇迹的哲学，若理论中的实体不存在，其预言的成功便无从解释。此文见第 60 至 78 页。它是本章「定律」一方的正面武器，恰与劳丹的悲观元归纳针锋相对。

24. D. H. Wolpert & W. G. Macready (1997).「No Free Lunch

Theorems for Optimization」. IEEE Transactions on Evolutionary Computation, 1(1). [②③] 沃尔珀特与麦克里迪证明「无免费午餐」定理：在所有可能的目标函数上取平均，任何两个优化算法的期望表现都完全相同，因而不存在普适占优的算法，任何算法的优势都靠对问题结构的特定假设换来。此文载于卷 1 第 1 期，第 67 至 82 页。它是本章引以为对照的一条真定理，且偏偏指向反方，提醒我们「最优策略唯一」式的结论需要极强的前提。

25. D. R. Hofstadter (1979). 《Gödel, Escher, Bach: An Eternal Golden Braid》. Basic Books. [②④] 霍夫斯塔特借哥德尔的不完备性、埃舍尔的视觉悖论与巴赫的卡农，编织出一个共同母题：自指与「奇异环」，并由此探讨心智与意义如何从无意义的形式层级中涌现。此书 1979 年由 Basic Books 出版，获普利策非虚构奖。它把自指与递归收束讲成一门艺术，呼应本章那一刻，一本书无法在内部自证，却照样「以它所描述的方式」写自己。

26. E. Nagel & J. R. Newman (1958). 《Gödel's Proof》. New York University Press. [②④] 纳格尔与纽曼用尽量少的技术细节，向一般读者讲清哥德尔第一与第二不完备性定理的证明思路：任何足够强的一致形式系统，都存在它无法证明的真命题，且无法在系统内部证明自身的一致性。此书 1958 年由纽约大学出版社出版。它为本章关于「形式系统的内禀界限」与「自指系统无法内部自证」的论述，提供了可读而准确的依据。

27. G. J. Chaitin (1982).「Gödel's Theorem and Information」. International Journal of Theoretical Physics, 21(12). [②③] 蔡廷用算法信息论重新解读不完备性：一个形式系统的公理蕴含的信息量是有限的，因此无法证明任何复杂度超过该信息量的命题（如某串足够长的随机比特确属随机），不完备性由此被还原为一种信息上限。此文载于卷 21 第 12 期，第 941 至 954 页。它从信息的角度坐实了形式系统的内禀界限，为「无法验证的世界」提供又一层形式根据。

28. M. Mitchell (2009).《Complexity: A Guided Tour》. Oxford University Press. [②③] 米切尔为一般读者梳理复杂系统科学：从信息、计算、演化到网络，介绍涌现、自组织等在生物、计算、社会系统中反复出现的共通主题，并坦诚这一领域尚缺统一理论。此书 2009 年由牛津大学出版社出版。它为本章呈现跨域共通模式的研究现状提供了可靠的入门地图，也提示这些模式至今仍多是经验观察而非定理。

29. J. P. A. Ioannidis (2005).「Why Most Published Research Findings Are False」. PLoS Medicine, 2(8). [③④] 约阿尼迪斯用一个统计模型论证：在小样本、小效应、研究自由度大、利益与偏倚普遍存在的条件下，一个已发表「阳性」结论为真的概率往往低于五成。此文载于卷 2 第 8 期，e124（2005 年 8 月）。它是元科学的经典警钟，提醒本章，看似稳健的经验模式可能只是系统性偏倚的产物，作者没有理由假设自己对此免疫。

30. Open Science Collaboration (2015).「Estimating the Reproducibility of Psychological Science」. Science, 349(6251). [③④] 开放科学合作组织协同上百名研究者，重复了一百项已发表的心理学研究，结果只有约三分之一的重复实验得到与原研究方向一致且显著的效应，且重复效应量普遍小于原始报告。此文载于卷 349 第 6251 期，文章号 aac4716。它把「复制危机」从担忧变成可量化的事实，为本章「强经验模式是否真的稳健」提供了直接而有分量的经验数据。

第 15 章　不靠验证的知识

> **论点**：认识论（epistemology）的落点。如果验证通常不可得，那么我们称之为知识的大部分，都是无验证的知识；而有能力不是「知道自己对」，而是「行动得当，且对自己可能错的方式有良好标定」。

上一章逼出一个问题：如果验证通常不可得，连这本书都只能给一个未经证实的信念，那我们平日称之为「知识」的那一大堆东西，到底是什么？这一章把落点放在认识论上。

重新定义「知道」

哲学课本上，知识是「被证成的真信念」(justified true belief)，最好还附一份证明。对一个有限的主体来说，这个标准在绝大多数有后果的事情上根本达不到，要么没有判定程序，要么代价爆炸，要么状态隐藏，要么时间不够，要么对面有人作对。按这个标准，我们几乎什么都不「知道」。

那就得换一个适合有限存在者的定义。不再是「知道自己对」，而是持有一个标定（calibration）良好的信念，让自己可能犯错的方式处在可控之下。有能力，不是确知真相，而是行动得当，同时对可能的错误方式有清醒的认识。这个转换一旦做出，前面那八招就不再只是工程工具箱，它们升格成了一种认识论，一套「当神谕永不到来时，如何持有信念并据以行动」的办法。

科学，这种姿态的早期原型

这不是什么新发明，人类最严肃的求知建制早就这么干了。科学从不宣称证实，只说「至今尚未被证伪（falsification）」，这正是第 3 章讲过的内容。科学整个就是一台为「在不可验证中持信而行」而优化的机器。哲学家也早把话挑明。杜威 1929 年那本《确定性的追寻》[10]，书名本身就是诊断：人类把太多力气耗在追求一种行动领域里根本不存在的确定上，而知识的真正功能，是引导行动，不是提供保险。詹姆斯的《信仰的意志》[9] 更进一步：有些事，证据齐备之前你就必须表态，而在那种处境下选择相信并下注，是正当的，不是思想上的草率。波兰尼的个人知识（personal knowledge）[8] 则提醒，任何「知道」都含着一份超出可证范围的个人托付。把这些合起来，是一种成熟的姿态：知识不是等来的确定，是被付诸行动的标定信念。

八招，读作一种认识论

于是可以把那八招重读一遍，这次不当工程，当认识。

证书，是把一小块切片彻底弄懂、其余存疑。标定，是老老实实持有分级的信念，而非假装非黑即白。冗余，是用多个互相独立的视角三角定位一件你无法直接看清的事。代理，是借一个可处理的替身去把握那个把握不住的真目标，同时警惕它在 Goodhart 处的背叛。筛查，是把有限的注意力投在最能更新你的地方。神谕，是在自己判断力不及处，求助于更可靠的判断。衰减与留痕，是让自己「行动得当」的底线，把信念付诸实施时，确保万一错了，损失扛得住、错误查得出、还能纠正。合起来，这就是一套有限存在者的可用认识论。

直觉、专长，与判断的真相

落到具体的人身上，技艺高超者究竟怎么做到这一点？这是落足点④最实在的部分，也最容易被神化或一笔抹杀，得说准。

关于专家判断，心理学积累了大量并不总让人舒服的证据。米尔

1954 年[5]、道斯 1979 年[12] 发现，在许多领域，简单的统计模型胜过专家的临床直觉。但另一脉研究给出了互补的图景。克莱因的自然主义决策（naturalistic decision making）[17]、舍恩的「反思性实践者」（reflective practitioner）[13]、德雷福斯兄弟[14] 与埃里克松对刻意练习（deliberate practice）[15] 的研究表明，在反馈充分、规律稳定的环境里，专家能发展出可靠的直觉，那本质是被反馈打磨出来的、标定良好的模式识别（pattern recognition）。吉仁泽的「快而省」启发式（fast-and-frugal heuristics）[16] 进一步指出，简单规则之所以管用，是因为它们吃透了环境的结构（生态理性，ecological rationality）。最持平的综合，来自卡尼曼与克莱因 2009 年那场「未能达成分歧」的对话[24]：直觉值不值得信，取决于环境，高效度、可学习的环境里它可信，低效度、充满噪声的环境里它就是自欺。

这正是本书认识论的人形版本。直觉既非魔法，也非废物，它是一种标定的能力，而标定本身是可以训练的。特洛克主持的「良好判断计划」（Good Judgment Project）[27] 在一场情报界举办的预测锦标赛里，挑出一批被称为「超级预测者」（superforecasters）的普通人，他们既无机密权限，也非领域专家，却靠多角度取证、小步更新、严苛复盘这些可学的习惯，据报道把预测准确度做到了超过能接触机密情报的专业分析师约三成。承认这一点，也就承认了刻意的无知（deliberate ignorance）有时是理性的[28]、承认了在凯恩斯[3] 与奈特[1] 那种根本不确定（凯与金[29] 所谓「彻底的不确定性」，radical uncertainty）面前，按照塔勒布[26] 的思路为稳健与反脆弱（antifragile）进行布局，往往比追求精确预测更加明智。

在不确定中行动的尊严

综合这些观察，浮现出的是一种姿态，它既不是怀疑论的瘫痪（既然什么都不能确定，那就什么都不算数、什么都别做），也不是教条者的假装（供起一个测得出的数字，假装它就是测不出的真相）。它是第三条路：清醒地知道自己不知道，给那份不知道标好刻度，然后照样行动得当。

这里面有一种安静的尊严。承认验证是奢侈品，不是认输，而是认真对待行动的前提。一个好的判断者，不靠确定支撑自己，他靠的

是不给自己的把握注水，以及把这份老实落成行动的那套办法。

合上序里那道弧

回到开篇。古希腊人出征前去德尔斐求神谕，计算机科学家把那个能即时给出答案的黑箱也叫神谕，两者共享同一个幻想：动手之前，先把对错验明。这本书讲的，是这个幻想破灭之后的世界，而它最终的答复是：幻想破灭，并不意味着求知与行动的终结，只意味着它们必须换一种方式进行。

对一个有限的存在者，知道，从来不是「等到了证明」，而是「持着一个标定的信念，把它付诸了行动」。神谕不会回话，可这从不曾、也不该让我们停下脚步。剩下的，是把这话落到一个具体的人身上，那是跋的事。

参考文献

> 落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. F. H. Knight (1921).《Risk, Uncertainty and Profit》. Houghton Mifflin. [②④] 奈特在此划出经典区分：可量化、可投保的「风险」，与无法赋以概率的「不确定性」，真正的利润正源于后者。本章谈「彻底的不确定性」时，这一区分是源头，它提醒读者，许多有后果的决策根本没有概率分布可依。
2. J. M. Keynes (1921).《A Treatise on Probability》. Macmillan. [②③] 凯恩斯在这部早期著作里发展了一种逻辑概率观，并引入「证据权重」的概念：我们对某一概率判断本身的把握，会随证据多寡而变。它为本章「标定信念」一脉提供了哲学根基，说明概率数字之外还有一层对自身知识状态的老实。
3. J. M. Keynes (1937).「The General Theory of Employment」. Quarterly Journal of Economics, 51(2), 209-223. [②④] 这篇为《通论》辩护的文章里，凯恩斯坦言对许多未来之事「我们就是不知道」，没有任何科学依据可形成可算的概率。它把

根本不确定性摆到经济行为的中心，是本章主张「在测不出的真相前照样行动」的重要先声。

4. F. A. Hayek (1945). 「The Use of Knowledge in Society」. American Economic Review, 35(4), 519-530. [②③④] 哈耶克指出，社会运转所需的知识从不集中于任何一处，而是分散在无数个体手中、且多为局部而隐默的。这篇文章关乎有限主体如何在不掌握全局的情形下仍能协调行动，与本章「没人能验明全貌却仍要决策」的处境直接呼应。
5. P. E. Meehl (1954). 《Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence》. University of Minnesota Press. [②④] 米尔系统比较了专家的临床判断与简单统计模型的预测表现，结论是后者往往不逊于甚至胜过前者。这一发现是本章讨论专家直觉的起点，它逼人正视：直觉的可信度需要经验检验，而非想当然。
6. H. A. Simon (1955). 「A Behavioral Model of Rational Choice」. Quarterly Journal of Economics, 69(1), 99-118. [②④] 西蒙在此提出「有限理性」：现实中的决策者算力、信息和时间都有限，于是「满意即止」地选取够好的方案，而非穷举最优。这正是本书全部论证的人类学前提，本章关于「适合有限存在者的认识论」由此立足。
7. H. A. Simon (1956). 「Rational Choice and the Structure of the Environment」. Psychological Review, 63(2), 129-138. [②④] 这篇姊妹篇强调，理性的形态取决于决策者所处环境的结构，简单的决策规则之所以管用，是因为它们契合了环境。本章谈吉仁泽的「生态理性」时，其思想根脉可上溯至此。
8. M. Polanyi (1958). 《Personal Knowledge: Towards a Post-Critical Philosophy》. Routledge & Kegan Paul. [①③④] 波兰尼论证一切「知道」都含有难以言传的默会成分，知者必然投入一份超出可证范围的个人托付，纯粹客观、无主体的知识只是幻象。本章引此说明：即便最严肃的求知，也无法摆脱不可完全验证的个人成分。
9. W. James (1897). 《The Will to Believe and Other Essays in Popular Philosophy》. Longmans, Green. [③④] 詹姆斯主张，面对那些证据不足却又必须表态、且关乎切身的抉择，选

择相信并据以行动是正当的，而非思想上的轻率。本章借此说明：在神谕不回话之前下注，可以是负责任的，而不是认识论上的失格。

10. J. Dewey (1929).《The Quest for Certainty: A Study of the Relation of Knowledge and Action》. Minton, Balch & Company. [③④] 杜威把人类对确定性的执着诊断为一种逃避：知识的真正功能是引导行动、改造处境，而非提供一劳永逸的保险。这本书几乎是本章的题眼，书名本身即点破了全书要破除的那个幻想。
11. A. Tversky & D. Kahneman (1974).「Judgment under Uncertainty: Heuristics and Biases」. Science, 185(4157), 1124-1131. [②④] 这篇奠基之作揭示，人在不确定下的判断依赖少数启发式（代表性、可得性、锚定），它们多数时候够用，却也会系统性地偏离概率法则。本章谈直觉的可靠与不可靠时，它提供了「直觉会犯有规律的错」这一关键背景。
12. R. M. Dawes (1979).「The Robust Beauty of Improper Linear Models in Decision Making」. American Psychologist, 34(7), 571-582. [②④] 道斯证明，即便权重随意设定的简单线性模型，预测也常胜过专家判断，因为它一致地利用了有效线索、不受人类临场波动干扰。它延续了米尔的发现，是本章「简单规则为何稳健」一节的直接支撑。
13. D. A. Schön (1983).《The Reflective Practitioner: How Professionals Think in Action》. Basic Books. [③④] 舍恩提出「在行动中反思」：熟练的专业者并不靠套用既定理论，而是在实践当下与情境对话、即时调整。本章引此刻画专长的另一面，说明在反馈充分的实践里，可靠的判断如何生成。
14. H. L. Dreyfus & S. E. Dreyfus (1986).《Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer》. Free Press. [④] 德雷福斯兄弟提出从新手到专家的技能习得阶段论，认为高阶专长的标志是越过显式规则、转为整体性的情境直觉。本章借此说明专长的成熟形态不是更会算，而是更会看，这也为「直觉是被打磨出来的能力」张本。
15. K. A. Ericsson, R. Th. Krampe & C. Tesch-Römer (1993).

「The Role of Deliberate Practice in the Acquisition of Expert Performance」. Psychological Review, 100(3), 363-406. [②④] 这篇研究主张，造就卓越表现的关键不是单纯的经验累积，而是「刻意练习」，即有明确目标、即时反馈、不断逼近能力边缘的费力训练。本章用它支撑一个要点：可靠的直觉来自反馈的反复打磨，而非时间的自然沉淀。

16. G. Gigerenzer & D. G. Goldstein (1996). 「Reasoning the Fast and Frugal Way: Models of Bounded Rationality」. Psychological Review, 103(4), 650-669. [②④] 两位作者展示，像「认出哪个就选哪个」这类快而省的简单启发式，在合适环境里能匹敌甚至超过复杂的统计推断。本章谈「简单规则为何管用」时，这是直接证据，说明少即是多取决于规则与环境的契合。
17. G. Klein (1998). 《Sources of Power: How People Make Decisions》. MIT Press. [②④] 克莱因通过对消防员、护士等实战者的田野研究，提出「自然主义决策」：专家常不比较选项，而是凭模式识别迅速认出当下属于哪类情境、该怎么办。本章引此呈现专家直觉可信的一面，与统计模型派形成互补。
18. R. M. Hogarth (2001). 《Educating Intuition》. University of Chicago Press. [②④] 霍格思追问直觉从何而来，区分了「友善」与「险恶」的学习环境：反馈准确及时的环境会培养出好直觉，反馈误导或缺失的环境则养出坏直觉。这与本章「直觉是否可信取决于环境」的核心判断高度一致。
19. G. Gigerenzer & R. Selten (Eds.) (2001). 《Bounded Rationality: The Adaptive Toolbox》. MIT Press. [②④] 这部文集把有限理性重述为一套「适应性工具箱」：心智备有多种简单启发式，因情境取用，而非追求全局最优。本章谈生态理性时，它提供了系统化的框架，把零散的启发式研究收束为一种理性观。
20. G. Klein (2004). 《The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work》. Currency. [④] 这本面向实践者的书把克莱因的研究化为可操作的训练：如何积累经验、复盘决策、磨砺并审视自己的直觉。对本章而言，它说明标定良好的直觉不仅可被研究，也可被有意识地培

养。

21. P. E. Tetlock (2005).《Expert Political Judgment: How Good Is It? How Can We Know?》. Princeton University Press. [①②④] 特洛克历经多年追踪大量专家的政治与经济预测，发现其总体准确度堪忧，且越自信、越爱讲大叙事的「刺猬型」专家往往越不准。本章引此既敲打专家的过度自信，也为「预测能力可以被检验、被训练」埋下伏笔。
22. G. Gigerenzer (2007).《Gut Feelings: The Intelligence of the Unconscious》. Viking. [④] 这是吉仁泽面向大众阐释其研究的一本书：直觉并非非理性的冲动，而是无意识地运用了适应环境的简单经验法则，常常又快又准。本章用它支撑「直觉是一种生态理性」的看法。
23. N. N. Taleb (2007).《The Black Swan: The Impact of the Highly Improbable》. Random House. [④] 塔勒布论述那些罕见、难以预测却影响巨大的「黑天鹅」事件，警告人们事后总爱为其编造解释、事前却系统性地低估其可能。本章借此说明：与其追求精确预测，不如为不可预测之事布局，这呼应了后文的稳健与反脆弱主张。
24. D. Kahneman & G. Klein (2009).「Conditions for Intuitive Expertise: A Failure to Disagree」. American Psychologist, 64(6), 515-526. [②④] 分属「直觉多有偏误」与「专家直觉可靠」两派的两位学者，在这篇罕见的对话里达成共识：直觉是否可信取决于环境，规律稳定、反馈充分的环境里它可学可信，低效度、充满噪声的环境里它就是自欺。本章把这视为最持平的综合，是整节认识论的支点。
25. D. Kahneman (2011).《Thinking, Fast and Slow》. Farrar, Straus and Giroux. [②④] 卡尼曼以「系统一」的快直觉与「系统二」的慢推理为框架，总结了数十年关于判断偏误的研究。本章借它把专家直觉放回认知机制的全景中，提醒读者直觉既是能力之源，也是偏误之源。
26. N. N. Taleb (2012).《Antifragile: Things That Gain from Disorder》. Random House. [④] 塔勒布提出「反脆弱」：超越仅仅抗压的稳健，有些系统能从波动、压力与意外中获益。本章引此给出在根本不确定下行动的正面策略，即布置好让

自己从不可预测中受益而非受损的结构。

27. P. E. Tetlock & D. Gardner (2015).《Superforecasting: The Art and Science of Prediction》. Crown. [①④] 本书报告「优秀判断力计划」的发现：少数「超级预测者」的准确度持续高于常人，靠的不是天赋，而是一套可学的习惯，多角度取证、小步更新、严苛复盘。本章引此说明标定本身可以被训练，预测是一门可改进的手艺。
28. R. Hertwig & C. Engel (2016).「Homo Ignorans: Deliberately Choosing Not to Know」. Perspectives on Psychological Science, 11(3), 359-372. [②④] 两位作者梳理人们何以及如何主动选择不去知道某些信息，论证「刻意的无知」常常是理性的应对，而非认知缺陷。本章用它支持一个反直觉的要点：有时不查、不知，恰恰是好的决策姿态。
29. J. Kay & M. King (2020).《Radical Uncertainty: Decision-Making Beyond the Numbers》. W. W. Norton. [②④] 凯与金接续奈特和凯恩斯，批评把一切不确定都强行塞进概率模型的做法，主张面对「彻底的不确定性」时应转而追问「这里到底发生了什么」，靠叙事与稳健的判断行动。本书是本章「彻底的不确定性」一语的直接出处，也是其总论调的当代回响。

跋　学会在没有把握时行动

这本书从一项结构性的研究开始：五种处境、八招、四根杠杆，一张跨领域的对照表，以及一个它自己也承认未被证实的猜想。但它该以一个人来结束，因为这件事归根到底是你的。

你会用一整生把行动投在你无法验证的东西上。你写的代码会带着你查不完的 bug 上线，你信的理论会在你穷尽证据之前就要你投入，你爱的人、你共事的人、你托付的制度，没有一样能在你下注之前验证，而你最重要的那些选择，恰恰是最不可验证的那些。问题从来不是你能不能确定，你不能。问题是，你能不能照样行动得当。

这本书全部的内容，归结起来，就是那个「照样」里蕴含的办法。在你查得动的小块上证个准，剩下的老老实实存疑；用几个独立的视角去三角定位（triangulation）你看不清的事；借一个够用的替身去把握那个把握不住的，同时盯紧它何时开始骗你；把有限的心力投在最能改变你判断的地方；够不着的判断，就向更可靠的人求助；下注的时候，把注下得让自己输得起、错得了也查得出还改得回。这些不是花招，而是一个有限的人在没有神谕的世界里清醒而不瘫痪地活着的全部手艺。

那艘船还在雾里。船长终究没有等来一双能看穿雾的眼睛，海图会旧，罗盘会偏，洋流的估算永远只是估算。可她照样改了航向，不是因为她确信前方没有暗礁，而是因为停在原地同样是一种豪赌，而她已经把能做的都做了：核过图，校过表，留了余量，备好了万一触礁时弃舱的预案。然后她转动舵轮。

学会在没有把握时行动，说到底，就是学会这样转动舵轮。雾不会

散，这正是你必须学会的理由。

参考文献

落足点：① 历史上科学家的判断　② 理论上被研究过的东西　③ 科学如何进展　④ 如何在无法验证的世界里生活。本节经网络逐条核实。

1. Aristotle (约公元前 4 世纪).《Nicomachean Ethics》. [④] 亚里士多德在此提出实践智慧（phronesis）的概念：伦理判断不能化约为普遍规则，而要靠在具体情境中权衡得当的能力，德性正是在反复实践中养成的稳定品格。常用英译本为 R. C. Bartlett 与 S. D. Collins 译本（University of Chicago Press, 2011）。它对本章重要，因为「在没有把握时行动得当」正是一种实践智慧，规则给不出答案时，靠的是经过训练的判断。
2. Epictetus (约公元 125 年).《Enchiridion》. [④] 爱比克泰德这部斯多葛派手册由弟子 Arrian 据其《对话录》辑录而成，核心是「分清什么在我们控制之内、什么不在」，并把心力收回到可控的判断与选择上。它与本章呼应之处在于：面对无法验证、无法掌控的世界，先认清能动的边界，是清醒行动而不瘫痪的起点。
3. Marcus Aurelius (约公元 175 年).《Meditations》. [④] 马可·奥勒留以希腊文写成的私人札记，原题约相当于《致自己》，并非为出版而作，记录了一位斯多葛派统治者如何在权力与无常中自省、克制、尽责。它对本章的意义在于示范了一种姿态：在看不清结局时，仍把当下该做的事做好，与不确定共处而非求一个确定的庇护。
4. C. S. Peirce (1877).「The Fixation of Belief」. Popular Science Monthly, 12, pp. 1–15. [③④] 皮尔士在这篇文章里比较了人们「固定信念」的四种方法：固执、权威、先验合理性，以及科学方法，并论证只有诉诸外部实在、可被公开检验与修正的科学方法，才能让信念经得起怀疑的冲击。它对本章重要，因为它把「在小块上证个准、剩下的老实存疑」的态度追溯到

了源头：信念的价值在于它如何应对怀疑，而非它有多笃定。

5. W. James (1897).《The Will to Believe and Other Essays in Popular Philosophy》. Longmans, Green. [④] 詹姆斯主张，面对那些证据不足以决断、却又必须选择、且事关重大的问题，人有「相信的权利」，因为悬置判断本身也是一种带后果的选择。这正是本章的核心处境：当停在原地与转动舵轮同样是豪赌时，不下注并不等于中立。
6. W. James (1907).《Pragmatism: A New Name for Some Old Ways of Thinking》. Longmans, Green. [③④] 詹姆斯在此系统阐述实用主义：一个观念的意义和真值，要看它在经验中能兑现什么后果、能引导出哪些可行的行动，而非看它是否符合某个抽象标准。它支撑本章的判断观：在无法终极验证的地方，把「够用、能指导行动、可被后果检验」当作务实的真理标尺。
7. S. Kierkegaard (1846).《Concluding Unscientific Postscript to Philosophical Fragments》. [④] 克尔凯郭尔以 Johannes Climacus 为笔名、用丹麦文写成此书，论证关乎生存的真理无法靠客观体系确证，最终须以「信仰的跳跃」在不确定中投入，主观的、激情的承诺不可被客观知识取代。它对本章的意义在于：最重要的选择恰恰最不可验证，到某一步只能在证据穷尽之前先行投入。
8. F. H. Knight (1921).《Risk, Uncertainty and Profit》. Houghton Mifflin. [②④] 奈特在此划出影响深远的区分：「风险」指概率可知、可计算的不确定，「不确定」（后人常称「奈特式不确定」）则指概率本身都无从估计的局面，而利润正源于承担后者。它直接对应本章主题：真正棘手的处境不是赔率已知的下注，而是连赔率都看不清时仍须行动。
9. J. Dewey (1929).《The Quest for Certainty: A Study of the Relation of Knowledge and Action》. Minton, Balch. [③④] 杜威批评西方哲学长期追逐一种「确定性」的幻觉，把不变的知识抬高于易变的行动；他主张知识本就是探究与实验的过程，意义在于改善我们与世界打交道的方式。它为本章提供了思想背景：放弃对确定的执念，转而把判断当作可检验、可修正的实践。

10. R. Niebuhr (约 1943).《The Serenity Prayer》. [④] 这篇广为流传的祷文祈求平静接受不可改变之事、勇气改变可改变之事、智慧分辨二者，作者归属与确切年份均有争议，较可靠的考证可见 E. Sifton (2003).《The Serenity Prayer: Faith and Politics in Times of Peace and War》(W. W. Norton)。它以最凝练的方式说出了本章反复强调的分际：先认清能动与不能动的边界，再把力气用在能改变的地方。
11. F. A. Hayek (1945).「The Use of Knowledge in Society」. The American Economic Review, 35(4), pp. 519–530. [③④] 哈耶克论证社会所需的知识本质上是分散的、局部的、难以集中汇报的，没有哪个中央计划者能掌握全貌，而价格机制恰好是协调这些分散知识、让人各自就地决策的手段。它对本章重要，因为它说明了为何「全局可验证」往往是奢望，以及为何要靠多个局部视角去三角定位看不清的整体。
12. I. Berlin (1953).《The Hedgehog and the Fox: An Essay on Tolstoy's View of History》. Weidenfeld & Nicolson. [④] 伯林借古希腊残句「狐狸知道很多事，刺猬只知道一件大事」，把思想者分为以单一宏大原则统摄一切的「刺猬」与追逐多元、不强求统一的「狐狸」两类。它对本章有用，因为它提醒：在复杂而难验证的世界里，多元视角的「狐狸」式判断常比一元体系更稳健。
13. H. A. Simon (1955).「A Behavioral Model of Rational Choice」. The Quarterly Journal of Economics, 69(1), pp. 99–118. [②④] 西蒙在此提出「有限理性」与「满意化」(satisficing)：真实的决策者受限于信息与算力，不去搜寻最优解，而是寻找一个「够好」、达到可接受门槛即停的方案。它正是本章方法论的理论根基：心力有限，就把它投在最关键处，求够用而非求完美。
14. V. E. Frankl (1959).《Man's Search for Meaning》. Beacon Press. [④] 弗兰克尔以集中营幸存者的亲历为底，提出「意义疗法」：人最深的驱动力是寻找意义，而即使在最无法掌控、最无法验证前景的处境里，人仍保有选择如何面对苦难的自由。德文原著出版于 1946 年。它对本章重要，因为它把「在无从把握时仍能立身」落到了最极端的人类经验上。

15. H.-G. Gadamer (1960).《Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik》. J. C. B. Mohr (Paul Siebeck). [④] 伽达默尔这部哲学诠释学奠基之作（英译《Truth and Method》出版于 1975 年）主张理解总是从「前见」出发、在历史处境中进行的，真理不能被化约为一套方法论程序。它呼应本章对「客观验证」局限的看法：判断离不开立场，承认这一点，才能更老实地与自身视角的有限性打交道。
16. K. R. Popper (1963).《Conjectures and Refutations: The Growth of Scientific Knowledge》. Routledge & Kegan Paul. [③④] 波普尔在这部论文集中阐发其科学观：知识通过大胆猜想与严格反驳而增长，理论的价值在于它可被证伪、敢于冒险接受检验。它对本章重要，因为它把「证伪」立为进步的引擎：好的判断不求自证，而求暴露自己何处可能出错、何时开始骗你。
17. H. A. Simon (1969).《The Sciences of the Artificial》. MIT Press. [②④] 西蒙在此奠定「人工科学」与设计科学的纲领：凡是人造物（包括组织、软件、决策过程）都是为适应目标与环境而设计的，设计就是在受限条件下搜索可行方案的活动。它支撑本章把「下注下得让自己输得起、错得了也查得出还改得回」视为一种可设计的实践：好系统是为应对不确定而造的。
18. C. Argyris & D. A. Schön (1974).《Theory in Practice: Increasing Professional Effectiveness》. Jossey-Bass. [④] 阿吉里斯与舍恩区分了人们「声称信奉的理论」与「实际行动中的理论」，并提出「双环学习」：不只在既定目标下纠错，更回头质疑目标与假设本身。它对本章有用，因为它指向一种自我校准的习惯：行动者要能察觉自己嘴上说的与实际做的之间的落差，并据此修正。
19. A. Tversky & D. Kahneman (1974).「Judgment under Uncertainty: Heuristics and Biases」. Science, 185(4157), pp. 1124–1131. [②④] 特沃斯基与卡尼曼这篇奠基性论文揭示，人在不确定下的判断依赖代表性、可得性、锚定等少数启发式，这些捷径虽常奏效，却会系统性地导致可预测的偏误。

它对本章重要，因为它说明我们对不可验证之事的直觉判断本身就不可全信，故需借外部视角与机制来纠偏。

20. D. A. Schön (1983).《The Reflective Practitioner: How Professionals Think in Action》. Basic Books. [④] 舍恩提出「行动中的反思」：熟练的专业者并非先想清规则再套用，而是在与情境的即时互动中边做边思、随机应变，许多专业知识是难以言传的「默会」知识。它呼应本章对实践判断的看重：在看不全、来不及完全验证时，靠的是一种能在行动中自我修正的现场智慧。
21. M. C. Nussbaum (1986).《The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy》. Cambridge University Press. [④] 努斯鲍姆借希腊悲剧与哲学论证：好的人生与德性本质上是脆弱的，向运气与外部世界敞开，而非自足无虞，企图把善完全置于掌控之内反而会扭曲它。它对本章重要，因为它正面承认了不可控对善好生活的构成性意义：与脆弱共处，本身就是伦理成熟的一部分。
22. J. C. Scott (1998).《Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed》. Yale University Press. [③④] 斯科特考察了诸多宏大社会工程为何失败，提出「清晰化」（legibility）与「高度现代主义」之弊：自上而下的标准化抹去了在地的、难以编码的「米提斯」式实践知识，导致计划脱离现实。它对本章重要，因为它警示了对全局可验证、可量化的迷信，以及由此带来的系统性误判。
23. N. N. Taleb (2007).《The Black Swan: The Impact of the Highly Improbable》. Random House. [②④] 塔勒布论述「黑天鹅」：那些罕见、冲击巨大、事后才被勉强解释的事件，恰恰主导了历史走向，而我们的模型与直觉系统性地低估了它们。它对本章重要，因为它把不可预测、不可验证的极端风险摆到了中心，逼人重新思考该如何在这种世界里下注。
24. G. Gigerenzer (2007).《Gut Feelings: The Intelligence of the Unconscious》. Viking. [②④] 吉仁泽主张，简单的经验法则（启发式）在信息不全的真实世界里常比复杂模型更准、更省，所谓「直觉」其实是适应环境的高效捷径，这与把启发式一概

视为偏误的观点形成对照。它对本章有用，因为它说明在无法穷尽验证时，省俭而稳健的法则往往是更明智的选择。

25. A. Gawande (2009).《The Checklist Manifesto: How to Get Things Right》. Metropolitan Books. [④] 葛文德以医疗、航空等领域为例，论证在高度复杂、易出疏漏的工作中，一份简单的清单能可靠地兜住人会遗忘或想当然的关键步骤，显著降低失误。它直接呼应本章对预案与机制的看重：与其指望临场不出错，不如事先把「万一」固化成可执行的程序。
26. D. Kahneman (2011).《Thinking, Fast and Slow》. Farrar, Straus and Giroux. [②④] 卡尼曼总结其数十年研究，提出快而直觉的「系统一」与慢而费力的「系统二」之分，揭示前者如何在不确定下产生种种可预测的偏误。它对本章重要，因为它系统地说明了我们判断不可靠的机理，从而支持用刻意的、可核查的方法去补直觉之短。
27. N. N. Taleb (2012).《Antifragile: Things That Gain from Disorder》. Random House. [④] 塔勒布在此提出「反脆弱」：有些系统不只是能抵御冲击，还能从波动、压力与无序中获益、变强，与脆弱相对的不是坚固而是反脆弱。它对本章重要，因为它给出了在不可预测世界里下注的正向原则：不求预测准确，而求让自己处于错了也亏得有限、对了则收益放大的位置。
28. P. E. Tetlock & D. Gardner (2015).《Superforecasting: The Art and Science of Prediction》. Crown. [②④] 特洛克基于大规模预测竞赛的研究，刻画了表现最好的「超级预测者」的习惯：把问题拆解、用概率而非笃定来表达、勤于根据新证据小步更新、并事后复盘校准。它直接示范了本章倡导的判断方式：在无法验证的领域，可校准、可追责的概率思维胜过故作确定。
29. A. Duke (2018).《Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts》. Portfolio. [④] 职业扑克手出身的杜克主张把决策看作下注：在信息不全、运气掺杂的世界里，要把决策的质量与结果的好坏分开评判，警惕用结果倒推（resulting）来褒贬当初的选择。它对本章重要，因为它给出了一套与不确定共处的实操语言：以概率下注、以

过程论英雄，而非以一次成败定对错。