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 **1**
参考文献 3

第一部　不可验证 **11**

第 1 章　验证的奢侈 **11**
七乘八与其余一切 11
验证廉价的那一小块 11
错觉的四处裂口 12
数学与软件亦不例外 14
并非悲观的结论 15
小结 15
参考文献 16

第 2 章　不可验证的五种处境 **23**
第一种：不可判定 23
第二种：难解 24
第三种：部分可观测 25
第四种：预算受限 26
第五种：对抗 26
五种处境，五种补救 27
参考文献 28

第 3 章　可证伪，不可证实 **36**
一只黑天鹅 36
休谟的归纳问题 37
波普尔：以证伪代替证实 37

一句必要的限定 38
科学早已发现的几种对策 39
当机器失灵：可重复性危机 39
小结 40
参考文献 40

第 4 章　摊平的诱惑 **48**
摊平错在何处 48
摊平中的合理成分 49
本书必须承担的举证责任 50
小结：第二部的次序 51
参考文献 52

第二部　化身 **61**

第 5 章　控制台前的人 **61**
所言非所欲 61
偏好作为潜变量 61
单次询问为何不足 62
第一种对策：把判断者放进回路 63
第二种对策：把提问用在信息量最大之处 64
当代形态及其反噬 65
小结 66
参考文献 66

第 6 章　放出去的智能体 **75**
交出行动权之后 75
测试无法弥合的缺口 76
当系统具有策略性 76
应对：从「证明它对」到「围住它的错」 77
围堵的代价 79
小结 80
参考文献 80

第 7 章　撞墙的数学家 **89**
无从估量的距离 89
验证的鸿沟 90
证书与界 91

代理替换 92
概率方法 94
数学家如何判断 95
小结 96
参考文献 96

第 8 章　看不见自己的组织　104
看不清自身的庞然大物 104
分散的知识 105
清晰性的冲动 105
代理指标与 Goodhart 崩塌 106
以审计与冗余为补充 107
小结：第二部的收束 109
参考文献 109

第三部　收敛　118

第 9 章　压缩未知　118
证书：为局部提供可复核的保证 119
最优筛查：把检查分配到信息量最大之处 120
两种对策的关联 122
参考文献 122

第 10 章　借来的判断　131
神谕入回路：引进一个判断者 131
冗余：从许多不可靠里合成可靠 132
两种对策的关联 134
参考文献 135

第 11 章　换一个能处理的问题　143
代理替换：更换验证的对象 143
校准：改变判决的形式 145
两种对策的关联 148
参考文献 148

第 12 章　管住后果　157
限损：缩小爆炸半径 157
留痕：让错误事后现形 159
八种对策的全貌：第三部结语 160

参考文献 161

第四部　杠杆 **169**

第 13 章　八根杠杆 **169**
风险的粗略分解 169
八种对策与八个位置 170
人、机器与组织何以殊途同归 171
一个强主张及其限度 172
参考文献 173

第 14 章　是定理，还是模式? **180**
支持「定律」的理由 180
支持「模式」或更弱结论的理由 181
定论的条件 183
一本无法验证自身的书 183
参考文献 184

第 15 章　不靠验证的知识 **193**
重新界定「知道」 193
科学早已如此 194
作为认识论的八种对策 194
直觉、专长与判断 195
在不确定中行动的尊严 196
回到神谕 196
参考文献 196

跋　学会在没有把握时行动 **203**
参考文献 204

序　没有神谕的世界

古希腊人在出征、婚嫁或建城之前，往往先赴德尔斐求取神谕。神谕的意义并不在于它是否准确，而在于它所许诺的东西：在行动之前，总有一处地方能够给出答案。两千年后，计算机科学借用了这个词。在计算理论中，神谕（oracle，教材中多译作「谕示」）指一个黑箱：向它提交一个自身无法求解的问题，它即刻返回正确答案。两种神谕背后是同一个设想：行动之前，先把对错确定下来。

本书讨论的，正是这一设想落空之后的世界。

我们几乎从不验证。我们行动，然后或早或晚地得知结果，或者始终无从得知。「凡事皆可检验」之所以显得理所当然，是因为我们最早的训练来自一类特别狭窄的任务：算术、排序、核对收据。在这类任务中，答案近在咫尺，于是我们误以为整个世界皆是如此。然而一旦越出这一狭小的范围，验证便成了一种奢侈。七乘八可以验证；可是在说出「我愿意」之前，无人能够验证这段婚姻能否长久；在软件发布之前，无人能够验证代码中没有缺陷；在投身之前，也无人能够验证一个理论为真、一家公司健康、一个决定正确。大多数要紧的行动，都发生在未经验证的地面上。神谕沉默，行动却不能等待。

面对这一处境，常见的反应有两种：哀叹，或者假装。哀叹者认为，既然一切都无法确定，任何判断都不过是臆测。假装者则为自己造一个假神谕：把某个可以测量的数字奉为圭臬，仿佛它就是那个无法测量的真相。本书不取这两种态度，而是追问一个更有意义的问题：在没有神谕的情况下，那些善于应对的人，包括科学家、工程

师、数学家以及治理国家与机构的人，实际上是怎样做的？

若在足够多的领域中追问下去，就会得到一个出人意料的观察，本书正是由此而来：不可验证性的来源千差万别，行之有效的应对却一再收敛到同一小组对策上。

这一观察决定了全书的两层结构，后面各章都以它为基础。

第一层：问题是异质的。「我无法检验它」这句话之下，掩盖着五种结构迥异的处境。有的在原则上就不存在判定程序（不可判定，undecidable）；有的虽有程序，代价却高到无法承受（难解，intractable）；有的是所需的状态对你隐而不见（部分可观测，partially observable）；有的原则上可以验证，只是缺少足够的时间、算力或样本（预算受限，budget-constrained）；还有的是对方在主动挫败你的检验（对抗，adversarial）。将这五种处境混为一谈，是这一领域最常见的错误。第一部的任务，就是把它们逐一区分开来。

第二层：应对是收敛的。无论问题属于哪一种处境，善于应对者所采取的办法总是那么几种：以可测的代理指标（proxy）替代不可测的真实目标；在可检验的局部证明一个界；把代价高昂的检查用在信息量最大之处；引入一个外部的判断者；缩小失败的爆炸半径；为残余的风险给出一个经过校准（calibration）的概率；把检查从事前移到事后；以若干相互独立的判断抵消单点失误。本书将它们归纳为八种对策，并论证它们两两成对，共分四对。第二部进入四个具体场景，让这些对策在各自领域的术语中交错出现；第三部则把每一种对策从具体语境中抽离出来，加以整理和命名，汇成一张跨领域的对照表，这张表是全书最主要的贡献；第四部追问：为什么恰恰是这几种对策。

有一点需要事先说明。这种收敛究竟是一条定律（有某种力量迫使任何有限主体都走向这几种对策），还是仅仅是一个很强的经验模式（我们反复观察到它，却无法证明它必然如此）？就目前而言，我没有证据表明它是定律。本书所提供的，是一个边界清楚的猜想，以及一套能够贯通许多领域的共同词汇，而不是一条定理。第 14 章将专门讨论这一问题。

由此产生了一个无法回避、也不应回避的自指：一本讨论「如何在

无法验证时行动」的书，其核心命题本身同样无法验证。因此，它只能践行它所描述的做法：陈述一个经过校准的信念，划定主张的边界，接受反驳，然后继续推进。本书将亲自运用它所研究的方法。倘若它是对的，这种自我示范就不是缺陷，而是唯一站得住脚的写法。

最后留下一个画面，跋将回到这里。一艘船在浓雾中转向。船长有海图、罗盘和对洋流的估算，却没有能够看穿雾气的眼睛。转舵之前，她无法确认前方是否有暗礁。雾不会散去，神谕也不会降临。可是航行不能因此停止。本书想要理解的，不是如何等待雾散，而是一位好的船长在雾中究竟如何掌舵。

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. Google Books [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. Google Books [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. Google Books [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. doi:10.2307/j.ctv1pncnrh [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. Google Books [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。该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. Google Books [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。该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. doi:10.2307/1884324 [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. doi:10.2307/1884852 [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. Google Books [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. doi:10.1126/science.185.4157.1124 [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. doi:10.2307/1914185 [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. Google Books [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。该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. doi:10.1037/0033-295x.103.4.650 [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. Google Books [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. 链接 [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. Google Books [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。该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. doi:10.1037/11281-000 [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。该书是「把判断外包给可核查的规则」这一对策的经典证据，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. Google 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。该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. Google Books [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。该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. Google Books [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」。该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. Google Books [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. Google Books [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. Google Books [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。该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. doi:10.2307/973677 [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. Google Books [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。该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. Google Books [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。该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. doi:10.2307/2181906 [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. doi:10.1515/9780691233857 [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。该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. doi:10.1017/cbo9780511814563 [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。该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**：「完整验证」，即在事前确认某事为真、正确或安全，在人与机器的生活中都是例外，而不是常态。

七乘八与其余一切

七乘八等于五十六，这一点可以验证：重新数一遍，换一种算法再算一遍，或者直接对照乘法表，几秒钟之内，对错便确定无疑。

再看另外几件事：在说出「我愿意」之前，验证这段婚姻能够长久；在软件上线之前，验证整个代码库没有任何缺陷；在投入半生之前，验证自己所信奉的理论为真；在接受一份工作之前，验证这家公司经营健康。这些都无法做到。原因不在于努力不够，而在于这类事情从一开始就不提供「事前验证」这一选项。

这两类事情之间的反差，构成了本书的第一块基石：能够在事前确认某事为真、为对、为安全的「完整验证」，在人与机器的生活中都是例外，而不是常态。我们之所以觉得它理应是常态，只是因为我们的直觉形成于一个特别规整的狭小范围。

验证廉价的那一小块

哪些事情是我们能够验证的？演算一道算术题，给一串数字排序，核对一张收据的总额，判断棋盘上的某一步是否合乎规则。将这些任务放在一起考察，便会发现它们有几个不易察觉的共同点：对象是封闭的（相关的一切都摆在眼前）；情形是有限的（可以逐一列

举）；答案是局部而即时的（既不依赖远处，也不依赖将来）；而且存在一个机械的判定程序（依照程序执行，就能得到「是」或「否」）。

「凡事皆可检验」的直觉，正是在这一小块中养成的。学校教育反复奖励的，恰好是这类有标准答案、可以当场批改的题目。于是，我们在不知不觉间把一条经验（「在我练习过的事情上，对错总能查明」）推广成了一种世界观（「凡事的对错都能查明」）。这一推广是错误的，而且其错误有规律可循：一旦越出这一小块，上述四个特征几乎全部不再成立。

错觉的四处裂口

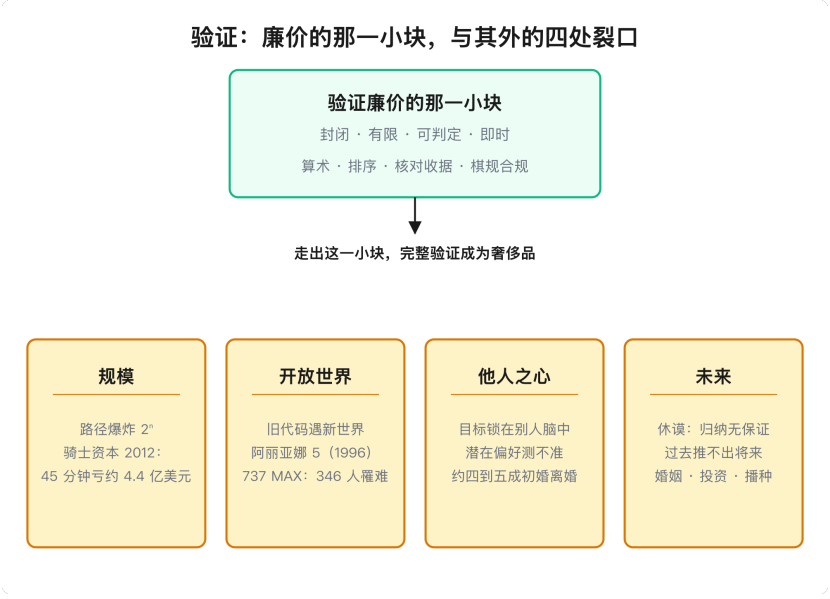


图 1: 验证：廉价的那一小块，与其外的四处裂口

规模（scale）。在这一小块之内，情形可以穷举；越出它的范围，情形便无从穷举。一个含有 n 个分支的程序，可能的执行路径多达 2^n 条；只要有几十个分支，穷尽测试即便耗尽宇宙的寿命也无法完成。这种路径爆炸（path explosion）使「全部测试一遍」从一开始就不可能。我们能够验证程序在已经想到的若干输入上是正确的，

却无法验证它在所有输入上都正确。2012 年 8 月 1 日，骑士资本（Knight Capital）部署新的交易程序，八台服务器中有一台没有更新。一段沉睡多年、早该弃用的旧代码，因一个被重复使用的标志位而意外激活，在开盘后约四十五分钟内发出数百万笔订单，致使公司亏损约四亿四千万美元，险些在一夜之间破产。没有人验证过那条废弃的路径，因为没有人想到它还会运行。检验单个情形并不困难；然而，一旦出现「所有」这个量词，问题的性质便截然不同了。

开放世界（open world）。这一小块中的对象是封闭的，外部世界却不断带来新的事物。我们测试过的只是有限的若干场景，系统真正要面对的，却是一个开放的、仍在不断展开的环境。1996 年 6 月 4 日，阿丽亚娜 5 型火箭首飞，升空约三十七秒后在空中自毁。事故的原因是一段从阿丽亚娜 4 型直接沿用、未曾针对新弹道重新验证的惯性导航代码：它将一个 64 位浮点数强行转换为 16 位整数，而新火箭更高的水平速度使这个数值发生了溢出。连同搭载的四颗科学卫星在内，损失超过三亿七千万美元。这段代码在旧的环境中正确运行了多年，一旦置于新的环境，便酿成了灾难。波音 737 MAX 的 MCAS 系统，是同一处裂口更为惨痛的例证：它在历次试飞中表现正常，投入实际航线之后，却依据一个故障迎角传感器的读数，一再将机头向下压。两起空难（2018 年狮航 610 航班、2019 年埃航 302 航班）共造成 346 人遇难。我们所验证的，始终是过去见过的那一部分；所要押注的，却是尚未见过的将来。

他人之心（other minds）。在这一小块之中，状态可以直接观测；而在外部，我们所要满足的目标却常常封存在另一个人的心里。「做得没错，但这不是我想要的」，这句常见的抱怨，根源就在这里。用户真正想要什么，上司是否满意，对方是否爱你，这些都是潜变量（latent variable）：只能从行为中间接推测，无法直接读取，因而也无法直接验证自己是否满足了它。即便是人生中最郑重的承诺，也无法例外。据人口学测算，美国约四到五成的初婚最终以离婚告终；没有人能在说出「我愿意」的那一刻，验证这段婚姻能够长久。

未来。这是最深的一处裂口。早在 1748 年，休谟就已指明[17]：归纳（induction）没有逻辑上的保证。太阳过去每天都升起，并不能在逻

辑上证明它明天仍会升起；有限的过去经验，无法在事前验证任何关于未来的全称判断。我们所依靠的不是证明，而是习惯。凡是结果落在将来的行动，无论婚姻、投资、播种还是托付，都处在这道裂口的另一侧。

数学与软件亦不例外

或许有人会认为：规模、人心、未来这些偏「软」的领域向不确定性让步，尚在情理之中；数学和软件总该是完整验证的堡垒。事实上，对验证的限度认识得最为清醒的，恰是这两个最「硬」的领域。

先看软件。迪杰斯特拉有一句被反复引用、却至今依然正确的话：测试只能证明缺陷存在，不能证明缺陷不存在[14]。他由此主张，程序应当从一开始就被正确地构造出来，而不是依靠调试逐步达到正确[13]。然而，即便是形式化证明这条最严格的道路，也有其限度。1979 年，德米洛、利普顿与佩利发表了一篇著名而又充满争议的论文[9]，指出程序验证（program verification）无法承担数学证明那样的角色：它的可信度最终来自一个社会过程，而不是机械的推导。1988 年，费泽尔将这一质疑推进得更远[10]：程序是一个因果模型，它与作为逻辑结构的算法之间隔着一道鸿沟，因此，「完全可靠的程序验证」即便在理论上也不能成立。形式化路线之外，软件工程内部的判断同样清醒。布鲁克斯在《没有银弹》中断言[11]，软件的本质复杂性（essential complexity）无法凭借任何单一手段一举消除。帕纳斯辞去了星球大战计划的顾问一职，并公开论证，那一类系统的软件无法被验证到值得托付的程度[12]。Therac-25 放疗机的事故，则以人命为上述判断写下了注脚：1985 至 1987 年间，它因一个竞态条件（race condition）缺陷六度失控，向病人体内输入高出正常剂量上百倍的辐射，至少三人因此死亡[15]。而在 1968 年，北约召开的那次会议便已为这种状况专门创造了一个词：软件危机（software crisis）[16]。

数学方面的结论则更为根本。1931 年，哥德尔证明[3]，任何足以表达基本算术的一致形式系统（formal system），都包含在该系统内部无法判定的真命题。1936 年，丘奇与图灵各自证明[2,1]，不存在能够判定任意命题是否可证的算法，即判定问题无解。莱斯定理（Rice's

theorem）[4] 将这一结论推向极致：程序的任何非平凡语义性质都是不可判定的（undecidable）。即便某个问题在原则上可以判定，库克于 1971 年确立的 NP 完全性（NP-completeness）[5] 也表明，验证的代价可能大到在实践中无法完成。这些并不是工程上一时的欠缺，而是逻辑为验证设下的硬性边界。下一章将专门剖析这一层面。

并非悲观的结论

综合以上各节可知：大多数要紧的行动，都发生在未经验证的地面上。

这一结论并不令人束手无策；毋宁说，它是一个起点。承认验证是一种奢侈，是认真对待行动的第一步。关于决策的研究，早已沿着这一方向展开。奈特早在 1921 年就把可以度量的「风险」与无法度量的「不确定性」区分开来[22]，并指出利润来自后者。凯恩斯论及真正的不确定性时，只留下一句「对此我们根本无从知晓」[26]。如果说他们指明了这类不确定性的存在，后来的研究者便转而追问：在这样的条件下应当如何决策？西蒙注意到，有限的主体不可能把所有选项逐一验证，于是提出了「满意化」（satisficing）[23]；冯 · 诺依曼与摩根斯特恩，以及萨维奇，则分别为「无法事前验证结果时如何理性地下注」建立了形式框架[24,25]。由此可见，一整门关于决策的学问，本来就建立在「验证不可得」这一前提之上。问题从来不在于如何消灭不确定性，而在于如何在不确定性中行动得当。

小结

既然验证通常不可得，首先需要追问的便是：它为何不可得？

答案不止一个，而这一点至关重要。把「我无法检验它」看作单一的处境，是这一领域最常见、也危害最大的错误。实际上，这是五种结构迥异的处境共用了同一句话。下一章将把它们逐一区分开来。

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. doi:10.1112/plms/s2-42.1.230 [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「数学与软件亦不例外」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. doi:10.2307/2371045 [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. doi:10.1007/bf01700692 [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. doi:10.1090/s0002-

9947-1953-0053041-6 [②] 莱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「数学与软件亦不例外」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. doi:10.1145/800157.805047 [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. doi:10.1145/363235.363259 [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. doi:10.1145/360248.360252 [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. doi:10.1007/BFb0025774 [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. doi:10.1145/359104.359106 [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. doi:10.1145/48529.48530 [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. doi:10.1109/mc.1987.1663532 [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，本质复杂性无法被单一手段消除。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. doi:10.1145/214956.214961 [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「数学与软件亦不例外」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. doi:10.1145/355604.361591 [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. Google Books [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. doi:10.1109/mc.1993.274940 [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. 链接 [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). Google Books [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. Google Books [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. doi:10.2307/2181906 [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. Google Books [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. doi:10.1017/cbo9781139171472 [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. Google Books [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. doi:10.2307/1884852 [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「满意化」。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. Google Books [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. Google Books [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. doi:10.2307/1882087 [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. doi:10.1126/science.185.4157.1124 [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. Google Books [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**：「我无法检验它」这句话之下，掩盖着五种结构不同的处境；将它们混为一谈，是这一领域的核心错误。

「我无法检验它」这句话，听起来描述的是一种处境，实际上却涵盖了五种。它们的结构各不相同，有效的补救也各不相同。将它们混为一谈，是这一领域最核心的错误。

本章的任务，是把这五种处境逐一区分开来，并对每一种作出准确的刻画。这看起来不过是分类上的洁癖，全书后半部分的可信度却系于此。本书最终要论证的是：不可验证性的来源尽管千差万别，应对的办法却收敛到同一小组对策上。要使这一论断不致流于廉价，就必须先把「千差万别」坐实。差异阐明得越透彻，其后的收敛就越出人意料，也越需要解释。读者不妨把这五种处境记在心上，全书将一再回到它们。

第一种：不可判定

判据：在原则上就不存在判定它的算法。问题不在于难，而在于没有。

这是验证最彻底的一种失败。1928 年，希尔伯特与阿克曼[4] 明确提出了判定问题（Entscheidungsproblem）：是否存在一个机械程序，能够对任意数学命题判定其真伪？八年之后，丘奇[2] 借助 lambda

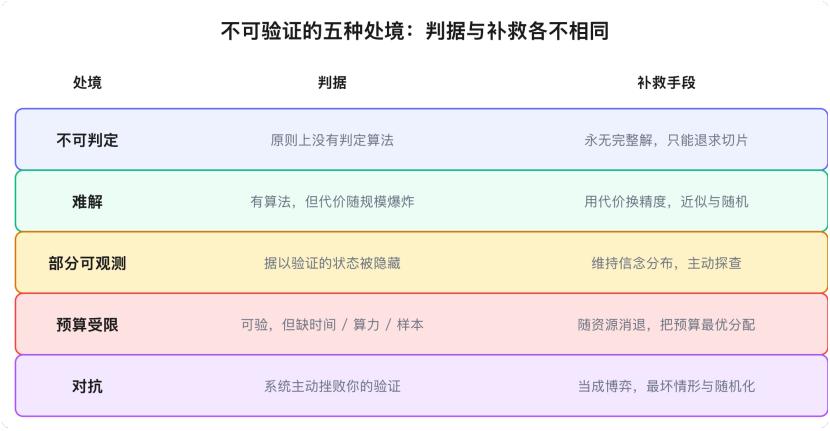


图 2: 不可验证的五种处境：判据与补救各不相同

演算，图灵[1] 借助他所设想的抽象机器，各自证明了答案是否定的。图灵的停机问题（halting problem）表述得尤为简洁：不存在这样一个算法，它对任意一组「程序加输入」都能判定该程序是否会停机。属于同一谱系的，还有哥德尔 1931 年[3] 证明的不完备性（incompleteness）、莱斯定理[6]（Rice's theorem，程序的任何非平凡语义性质都不可判定），以及马季亚谢维奇 1970 年[7] 对希尔伯特第十问题给出的否定回答。这些结果并非纸上空谈。「这段程序是否会做坏事」这一问题可以归约为停机问题，因此在理论上不存在一款能够毫无差错地查出所有恶意程序的杀毒软件。这是逻辑为整个反病毒行业划定的上限。

这种处境的补救有一个独有的性质：它永远不会有完整的解。再多的时间、再快的机器都无济于事，因为障碍在于逻辑，而不在于资源。所能做的只有退而求其次：验证有限的局部，或者把自己限制在确实可以判定的片段之内（例如只含加法的算术）。这一点在第 7 章中还将反复出现。

第二种：难解

判据：算法存在，但其代价随规模爆炸式增长，大到在实践中无法算完。

这一种处境与上一种差之毫厘，谬以千里：它可以判定，却不可行。库克于 1971 年[8]、列文于 1973 年[9] 各自确立了 NP 完全性（NP-completeness），卡普则在 1972 年[10] 列出了著名的二十一个 NP 完全问题；这些工作共同为这一处境提供了精确的刻画。可满足性问题（satisfiability）以及无数组合优化问题，在原则上都有解法，但这些解法在最坏情形下的耗时，随输入规模呈指数增长，

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

只需几十个变量，就足以令最快的超级计算机望洋兴叹。不妨从数量级上获得一点直观印象：国际象棋的博弈树约有 10^{120} 个分支（香农数），围棋的合法局面约有 2×10^{170} 个，而整个可观测宇宙中的原子数也不过约 10^{80} 个。这些棋类规则简单，在原则上可以穷举，但这个「原则上」远远超出了物理上的可能。P 是否等于 NP，所追问的就是这堵墙是否注定存在。

它的补救与不可判定（undecidable）截然不同。在这里，投入更多资源是有意义的。更重要的是，可以用「接受得少一些」换取「负担得起的代价」：以近似解代替精确解，以平均情形代替最坏情形，此外还有启发式搜索与随机化。难解（intractable）催生了一整套「打折」的智慧，而这在不可判定的处境中是不存在的。

第三种：部分可观测

判据：验证所依据的状态，对你隐而不见。

这里的困难不在于缺乏判定程序，也不在于代价过高，而在于看不到应当看到的东西。用户真正的偏好，病人体内正在发生的变化，对手手中的牌，这些状态驱动着结果，却不向观察者显现。控制论很早就对这一处境作了形式化的处理。1965 年，阿斯特罗姆[15] 研究了状态信息不完整时的最优控制问题；斯莫尔伍德与桑迪克在 1973 年[16]、凯尔布林等人在 1998 年[18]，又先后把它发展为部分可观测马尔可夫决策过程（POMDP）这一标准框架。此外，帕帕季米特里乌与齐齐克利斯于 1987 年[17] 证明，这类问题的求解同样是难解的。

因此，第三种处境常常与第二种叠加在一起。

它的补救自成一类：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），并以每一次观测对其加以更新，

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

这一处境的补救在于推断与探查，而不在于一味追加算力。第 5 章通篇都在这种处境之中展开。

第四种：预算受限

判据：在原则上可以验证、可以求解，但就这一主体而言，此时此地缺乏所需的时间、算力或样本。

这一种处境最为朴素，也最为普遍。审稿人只有二十分钟阅读一篇论文；医生只有几分钟作出诊断；交易员必须在行情消失之前下单。验证在理论上完全可行，落到一个有限的主体身上却无法实现。这一处境的思想渊源，可以追溯到奈特 1921 年[19] 对风险与不确定性的区分，以及西蒙 1955 年[20] 提出的有限理性（bounded rationality）。其后的研究为它提供了形式化的表述：迪安与博迪在 1988 年[22] 提出随时算法（anytime algorithm，可以随时中断，并给出当前的最优解）；罗素与苏布拉马尼安则在 1995 年[23] 提出了「有界最优」（bounded optimality）的概念。

它的补救具有一个其他处境所没有的特征：这种处境会随资源的增加而消退，只要时间与算力充足，它便不复存在。因此，应对它的关键在于分配，即把稀缺的预算投向边际收益最高之处。后文将要讨论的「最优筛查」这一对策，就源于这一思路。

第五种：对抗

判据：你所面对的系统，在主动挫败你的验证。

在前四种处境中，困难都来自世界的某种中立属性：逻辑、规模、

可见性或资源。第五种则不同：对方是一个具有智能的行动者，它针对你的检查进行优化。会说谎的对手，会伪装的恶意代码，会操纵指标的被考核者，都属于这种情形。这一处境的经典理论，包括冯·诺依曼与摩根斯特恩 1944 年[24] 的博弈论（game theory），以及随后瓦尔德 1945 年[26] 的极小极大准则（minimax）和纳什 1950 年[25] 的均衡（Nash equilibrium）。在机器学习中，它的当代形态是塞盖迪等人 2014 年[29] 发现的对抗样本（adversarial examples），以及马德里等人 2018 年[31] 用以统一攻防双方的鲁棒优化（robust optimization）。一个识别率极高的模型，可能被人眼无法察觉的微小扰动彻底误导，原因在于有人专门去寻找那个扰动。研究者做过一个被反复复现的演示：在停车标志上贴几张精心设计、形似涂鸦的小贴纸，就能使顶尖的图像识别系统稳定地将其识别为限速标志。同一块标志，人看到的是「停」，机器看到的却是「行」，差别仅在于对手把贴纸贴在了哪里。

它的补救属于战略，而不属于计算。与其把某个量算得更准，不如求解

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

也就是按最坏情形布防，借助随机化使对手无从预测，追求的是在博弈中站得住脚，而不是在某个固定输入上达到最优。把对抗（adversarial）当作单纯的可观测缺口（「我只是还没看清它」）来处理，是可能酿成致命后果的误判，因为另一方的系统会根据你的判断调整自身。

五种处境，五种补救

将五种处境并列比较，要紧的不是名称，而是补救：适用于一种处境的补救，移用到另一种处境上并不奏效。

- 不可判定：永远没有完整的解，只能退而验证有限的局部。
- 难解：可以用代价换取精度，投入更多资源是有意义的。
- 部分可观测：依靠推断信念与主动探查。

- 预算受限：随资源增加而消退，关键在于分配。
- 对抗：本质上是一场博弈，依靠战略与随机化。

倘若有人断言「这件事多投入一些算力就能解决」，他多半是把一种处境错认成了另一种。把不可判定当作预算问题，把对抗当作可观测问题，都属于这类错认，而且代价高昂。

这些处境还会相互叠加与复合。第 6 章讨论的那个放出去的智能体，既面临开放世界的不可预测性（近乎不可判定的行为），又面临对手的策略性（对抗）；第 8 章讨论的组织，则须同时应对部分可观测与对抗。现实中的处境，往往是几种处境的混合。

正因为来源如此参差、补救如此各异，下一个问题便显得格外尖锐：人类是否拥有一套成熟的方法，能够长期地、有纪律地与不可验证性共处？答案是肯定的。这套方法就是科学，而它的首要原则，便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. doi:10.1112/plms/s2-42.1.230 [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. doi:10.2307/2371045 [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. doi:10.1007/bf01700692 [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. Google Books [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. doi:10.1090/s0002-9904-1944-08111-1 [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. doi:10.1090/s0002-9947-1953-0053041-6 [②] 莱斯定理在此确立：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. 链接 [②③] 马季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. doi:10.1145/800157.805047 [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. 链接 [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. doi:10.1007/978-1-4684-2001-2_9 [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. doi:10.1090/s0002-9947-1965-0170805-7 [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. Google Books [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 Computation》(3rd ed.). Cengage Learning. Google Books [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. doi:10.1017/cbo9780511804090 [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.

doi:10.1016/0022-247x(65)90154-x [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. doi:10.1287/opre.21.5.1071 [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. doi:10.1287/moor.12.3.441 [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. doi:10.1016/s0004-3702(98)00023-x [② ④] 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. Google Books [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. doi:10.2307/1884852 [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). 链接 [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. 链接 [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. doi:10.1613/jair.133 [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. Google Books [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. doi:10.1073/pnas.36.1.48 [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. doi:10.2307/1969022 [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. Google Books [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. doi:10.2307/1884324 [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. doi:10.1016/j.patcog.2018.07.023 [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**：经验科学是人类最有纪律的求知方式，而它建立在一个公开的承认之上：理论永远无法被证实（verification），至多只能做到尚未被证伪（falsification）。

上一章提出了一个问题：人类是否拥有一套成熟的、有纪律的方法，能够与不可验证性长期共处？答案是肯定的，这就是经验科学。最出人意料的是，它的首要原则并不是宣称自己能够查明真理，而是公开承认自己永远无法做到这一点。

一只黑天鹅

「所有天鹅都是白的。」即便观察过一千只、一百万只白天鹅，这个全称判断（universal statement）依然没有得到证实，因为下一只可能就是黑的。然而，只要见到一只黑天鹅，它便被彻底推翻。这并非虚构的例子。在欧洲，「天鹅皆白」长期被视为确定无疑的常识；直到 1697 年，荷兰探险队在西澳大利亚首次见到黑天鹅，这份「确定」才在一夜之间失效。

这种不对称是全章论证的支点。要证实一个全称命题，就必须检查它所断言的全部情形，而这些情形往往是无穷的、开放的、属于未来的，不可能穷尽。要证伪它，却只需要一个反例。用逻辑的语言表述：$\forall x\, P(x)$ 无法由有限的观察确立，而一个 $\exists x\, \neg P(x)$ 就足以将它推翻。科学的全部纪律，都建立在认清并利用这种不对称的基础之上。

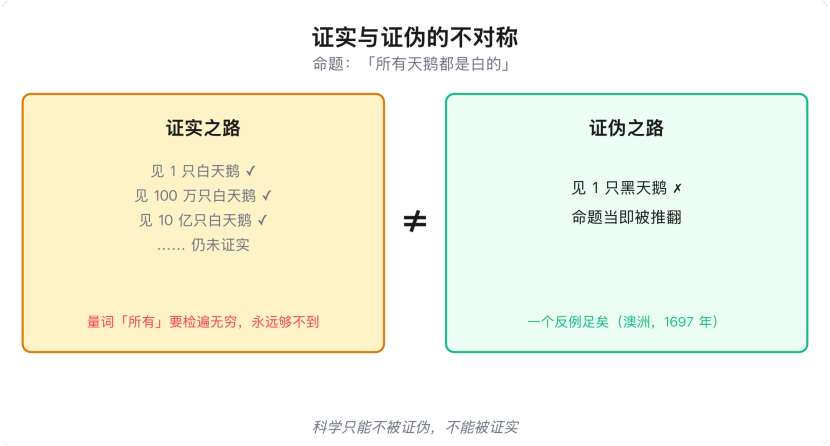


图 3: 证实与证伪的不对称

休谟的归纳问题

这一不对称的根源，休谟在 1739 年[4] 就已揭示。我们凭什么相信，过去一直成立的规律将来仍会成立？在逻辑上，这一信念没有依据。从「太阳过去每天都升起」推不出「太阳明天一定升起」，因为这一推论已经预设了「过去的模式会延续到未来」，而这恰是有待证明之事。归纳（induction）没有逻辑上的保证。休谟的结论冷静而彻底：我们所依靠的不是证明，而是习惯。

这便是第 1 章所说「未来」这道裂口的哲学根基。任何关于世界普遍规律的知识，都建立在有限的过去之上，因而都无法在事前得到证实。科学若要成为知识，就不能以「证实」为目标，因为这一目标无从企及。

波普尔：以证伪代替证实

波普尔在 1934 年[1]（德文原版）指出了一条出路：既然证实不可得，那就放弃证实，转而诉诸证伪。一个理论是否科学，不取决于它能得到多少证据的支持（支持总是找得到的），而取决于它是否甘冒风险，作出可能被推翻的预测。占星术，以及那些对一切现象都能自圆其说的学说，是不可证伪的，因而不属于科学。广义相对论预言，

星光会被太阳的引力偏折一个确定的角度；1919 年的日食观测完全可能测出别的数值，从而推翻这一预言。正因为这一理论甘冒被推翻的风险，它才是好的科学。

由此，科学成为一台专为「与不可验证性共处」而优化的机器。它从不宣称证明了什么，只是说：这个理论迄今尚未被证伪，因此我们暂且采用它。这是一种认识上的姿态：以可以检验的「尚未被推翻」，替换无法检验的「真」。这一姿态在后文还会出现：第 7 章那位数学家所做的代理替换，放到认识论的尺度上，就是这个样子。

一句必要的限定

这里有必要明确一点：波普尔式的证伪主义（falsificationism）在科学哲学中远非定论。本书把它当作一个清晰的入口，而不是终点。

对证伪主义最有力的反驳，来自迪昂与蒯因的整体论（holism，又称 Quine-Duhem 论题）。迪昂在 1906 年[10]、蒯因在 1951 年[9] 先后指出，任何假说都无法被孤立地检验。每一个预测都依赖于大量辅助假定：仪器运转正常，背景条件成立，近似处理合理。实验一旦失败，研究者总可以把矛头指向某一个辅助假定，从而保全核心假说。因此，「一个反例就干净利落地推翻理论」这幅图景，远不像表面上那样干净。

库恩在 1962 年[5] 更进一步：在常规科学时期，科学家并不急于证伪，反常往往被暂时搁置，直到范式（paradigm）陷入危机，才发生革命式的更替。此后，拉卡托斯在 1970 年[6] 以「研究纲领」（research programme）的进步与退化，取代了非此即彼的证伪标准；费耶阿本德[7] 则更为激进，反对一切统一的方法。

另一条路径是贝叶斯确证论（Bayesian confirmation theory）[24]：它不作非此即彼的判决，而把证据看作对信念概率的调整，

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

这也预示了后文的「校准」这一对策。梅奥提出的「严苛检验」（severe

testing）[14]，是证伪主义一个精致的继承者；斯坦福[19] 则提醒我们，在视野之外，还有大量「未被设想的替代方案」（unconceived alternatives）。

列举这些争论，并不是为了驳倒波普尔。本书自身也应当如此行事：提出一个有力的框架，同时标明它的边界。全书所要身体力行的，也就是这种姿态。

科学早已发现的几种对策

本节是本章对全书真正的贡献。倘若以「八种对策」的眼光审视科学日常运转所依赖的那套制度，便会发现，科学早已摸索出其中的好几种，只是冠以别的名称。

同行评审体现的是冗余与共识：它不信任单个判断者，而是邀请多位相互独立的审稿人，取其一致意见。重复实验同样属于冗余：一个结果必须由他人在别处独立重现，才会被认真对待。预注册（preregistration）相当于留痕：在看到数据之前便将假说与分析方案登记在案，事后既无法移动靶子，也无法把噪声说成信号。置信区间与误差统计则是证书与界：它们并不声称命题为真，而只是在一个明确的置信水平上给出有界的保证。双盲与随机化，是针对第五种处境（对抗）的防御，只不过这里的对手，往往是研究者自身的偏见与主观期待。显著性阈值则是一种粗糙的校准。

换言之，人类最严肃的求知事业，就是本书收敛命题的一个鲜活例证。这是全书给出的第一个提示，而且分量很重：科学所面对的不可验证性（关于普遍规律、关于未来）有其特定的来源，而它不得不采取的应对，却与软件、数学和组织中的应对押着同一个韵。

当机器失灵：可重复性危机

从反面看，这一点更为清楚。这些对策一旦被削弱，科学的自我纠错机制便会失灵，这就是可重复性危机（replication crisis）。2005 年，约阿尼迪斯[28] 发表了题为「为什么大多数已发表的研究结论是假的」的论文；2015 年，开放科学合作组织[29] 对一百项心理学研究

进行了大规模重复。二者所揭示的，是同一幅图景。那一百项研究中，97% 当初都报告了显著结果；重复之后，只有约三十六项仍然成立，不到一半。只要预注册缺位（靶子可以事后移动），样本量不足，发表偏倚只放行漂亮的结果，而吃力不讨好的重复研究又乏人问津，这台机器就会空转。

对这场危机的诊断与修补，所用的同样是这几种对策的语言：恢复预注册（重新引入留痕），鼓励并奖励重复研究（重新引入冗余），采用登记报告，提高检验的严苛程度。问题与药方，落在同一套词汇之中。第 10 章讨论借来的判断、第 12 章讨论留痕与审计时，还将回到这一点。

小结

科学证明了一点：人能够在缺乏验证的世界中有纪律地求知，而凡是做得好的地方，所依靠的都是那几种对策。就此而言，科学为全书提供了一个概念验证。

然而，这里也潜伏着一个陷阱。五种处境都以同一副面目出现，即「我无法检验它」；应对它们的对策又如此相似。于是，一个极具诱惑力的念头便会产生：何不径直宣布，不可验证性就是一个问题，只需配上一个统一的解法？这一念头关于「问题」的部分是错误的，关于「应对」的部分却不期然地切中了要害。下一章将专门处理这一诱惑。

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. Google Books [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：以证伪代替证实」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. Google Books [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. Google Books [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. Google Books [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「休谟的归纳问题」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. Google Books [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. doi:10.1017/cbo9781139171434.009 [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. Google 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. Google Books [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. doi:10.2307/2181906 [②] 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；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. Google Books [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。此即后来与蒯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. Google Books [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. doi:10.1093/0198244274.001.0001 [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. doi:10.1017/cbo9780511814563 [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. doi:10.7208/chicago/9780226511993.001.0001 [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. doi:10.1017/9781107286184 [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」，并据此回应可重复性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. doi:10.1086/288975 [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. Google Books [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. Google Books [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. doi:10.1093/0195174089.001.0001 [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. Google Books [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. doi:10.1086/286983 [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. doi:10.1086/286432 [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. doi:10.1515/9780691221489 [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. Google Books [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，也预告了全书的「校准」这一对策。

25. E. Sober (2008). «Evidence and Evolution: The Logic Behind the Science». Cambridge University Press. doi:10.1017/cbo9780511806285 [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. doi:10.7208/chicago/9780226300610.001.0001 [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. doi:10.1093/0198247044.001.0001 [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. doi:10.1371/journal.pmed.0020124 [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。它是可重复性危机的奠基文献，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 Reproducibility of Psychological Science」.《Science》, 349(6251), aac4716. doi:10.1126/science.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。它把可重复性危机从论证变为大规模实证，是本章「可重复性危机」一节的实证核心，也凸显了重复与预注册这些对策为何不可或缺。

第 4 章　摊平的诱惑

> **论点**：五种处境表面上都是「我无法检验」，人们因而很想把不可验证性当作一个问题，并为之配上一个解法。就问题而言，这是错误的；就应对而言，它却指向了正确的东西。

第 2 章用了整整一章的篇幅，把不可验证性区分为五种处境。第 3 章又表明，应对这些处境的对策彼此相似。将这两点合在一起，一个念头几乎必然会产生，而且极具诱惑力：既然它们都带有「我无法检验」这一共同特征，应对的办法又大同小异，何不就此宣布：不可验证性只是一个问题，可以用一个统一的解法来应对?

本章要完成两项工作：首先指出这一念头错在何处，其次说明它在无意之中触及了什么正确的东西，并由此为全书确立一条举证责任(burden of proof)。

摊平错在何处

所谓摊平（flatten），就是把五种结构不同的处境强行压缩为一种。它的代价，第 2 章已经预先展示过。

把不可判定（undecidable）当作预算问题，以为「多投入一些算力就行」，这是错误的。停机问题（halting problem）的症结不在于计算太慢，而在于那个算法并不存在；再快的机器，也造不出一个不存在的程序。

把对抗性的缺口当作单纯的可观测缺口，以为「我只是还没看清它」，

这一错误更为危险。另一方的系统会根据你的看法调整自身：你看得越清楚，它就躲得越巧妙。这是一场博弈，而不是一次测量。垃圾邮件过滤就是一个现成的例子。依据当下垃圾邮件的特征训练出一个分类器之后，发信方随即改写措辞、更换域名、插入乱码，以绕过它。谁若把这当作静态的识别问题，试图用只训练一次的模型来解决，谁就永远慢人一步。

把难解（intractable）当作不可判定，就会过早放弃一个本可以近似求解、可以按平均情形处理的问题；反过来，把不可判定当作难解，就会面对一堵逻辑之墙，无休止地投入算力。每一种错认，都会导致选错工具，并付出真实的代价。病灶一旦诊断错误，再好的药也会变成毒药。

因此，就「问题」而言，摊平是错误的，五种处境必须分别对待。这是第 2 章以一整章的篇幅换来的结论，不能在此轻易放弃。

摊平中的合理成分

然而，这一念头包含着一个真确的内核。它在「问题」上错了，在「应对」上却无意间触及了真相。

本书的命题，就是要把这一内核从错误中剥离出来，并以非常审慎的方式加以表述：

> 不可验证性的来源千差万别，但有限主体（bounded actor）被逼出来的应对，一再收敛到同一小组对策上。

这一表述的克制值得注意。它并不说「这些问题是同一个问题」：那是摊平，是错误的，任何一个领域的专家读到这里都会弃书不读。它所主张的是另一件更强、也更站得住脚的事：问题各不相同，应对却彼此押韵。全书要交出的，是一张「同一种对策在不同领域术语下的对照表」，以及一个解释：为什么收敛到的恰好是这几种对策。

本书必须承担的举证责任

论述至此，最大的风险也随之显现。人天生偏爱类比。莱考夫与约翰逊[14] 向我们表明，连日常语言都是以隐喻为骨架搭建起来的；根特纳[6]、霍利约克[9]、巴尔塔[7] 等人则研究了类比何时成立、何时只是好看。然而，正因为类比如此顺手，它也最容易误导人。漂亮的跨领域类比，往往什么也证明不了。最现成的反面例子，是霍夫施塔特《哥德尔、艾舍尔、巴赫》[1] 一类的写法：跨领域的呼应写得令人目眩神迷，却常被批评为「终究只是类比」，经不起追问。

更扎实的教训来自幂律（power law）研究的热潮。许多系统被宣称服从同一条幂律、共享同一种深层机制，听起来极为统一。然而，一旦采用克劳塞特、沙利齐与纽曼[25]2009 年提出的那种严格统计方法加以检验，大量所谓的「幂律」便站不住脚。他们重新审视了二十余个被广泛宣称服从幂律的真实数据集，能够稳稳通过检验的寥寥无几；多数数据集无法排除对数正态（log-normal）等其他分布，而这些分布的拟合效果同样好，甚至更好。斯坦普夫与波特[26]2012 年的文章「关于幂律的若干真相」说得十分直白：看起来像，不等于是。

因此，本书必须为自己设定一条铁律：任何被宣称的收敛，都必须被证明不只是类比。

何谓「不只是类比」？科学哲学中已有现成的标准。一个跨领域的映射若要算作实质性的，就必须是结构保持（structure-preserving）的：不仅表面相似，而且机制对应、失效方式对应、权衡也对应。根特纳[6] 的结构映射（structure-mapping）理论，以及巴尔塔[7] 对类比论证（analogical argument）的评估，所提供的就是这套标准。此外还有一条更为严格的判据，即稳健性（robustness）：一个结论如果能够从多条相互独立的路径反复导出，就更为可信。这一思想来自稳健性分析（robustness analysis），其发展始于莱文斯[17]，继而有维姆萨特[18] 与韦斯伯格[19] 的推进。

安德森[5]「多者异也」（More Is Different）的论断、福多[27] 关于「特殊科学」（special sciences）的论证，以及卡特赖特[28] 的《斑驳的世界》都表明，跨层次的真实模式确实存在，但它们是挣得的，而不是宣布出来的。博克斯[33] 的名言始终悬于头顶：所有模型都是错的，

有些是有用的。本书所要争取的，是「有用，且其用处经得起检验」，而不是「优雅得令人忘记检验」。

落实到操作层面，这条铁律意味着，第三部每抽离出一种对策、每进行一次跨领域的并置，都必须追问：这一迁移是实质性的（机制相同、失效方式相同、权衡相同），还是仅仅是一个漂亮的比喻？只有经得起这番追问，那张对照表才能成立；否则，它就只是一本好看的散文。第 13 章将尝试把这种收敛建立在一个共同的底层结构之上（风险与信息的分解），但那是一张有待兑现的期票，需要接受检验，而不能预先假定。第 14 章则将正面回答：这究竟是一条定律，还是一个很强的经验模式。

小结：第二部的次序

既然不能先宣布收敛、再挑选几个例子凑数，唯一可靠的检验办法，就是进入真实的领域，考察那些做得好的人究竟做了什么。应当让对策从案例中自行生长出来，而不是预先设定对策，再拿案例去套。

这也解释了本书为何把具体场景（第二部）置于对策的系统论述（第三部）之前。先抽象、后举例，会显得武断，也浪费了案例本应具有的说服力。因此，第二部并不急于命名。它让各种对策嵌在各自的领域之中，彼此交织地出现，表面上显得杂乱一些；到了第三部，才把每一种对策从其所属语境中抽离出来，剥除领域术语，并予以命名。先归纳，后命名。

第二部将进入四个具体场景：一位揣摩用户心思的设计者，一个被放出去自行其是的智能体，一位对着黎曼假设撞墙的数学家，一个看不见自己的庞大组织。它们所面对的不可验证性，分别来自五种处境中不同的几种。下面就来考察，在神谕始终没有降临的时候，他们各自采取了怎样的办法。

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. Google 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. Google Books [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. doi:10.2307/j.ctv1pncnrh [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. 链接 [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. doi:10.1126/science.177.4047.393 [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「多者异也」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. doi:10.1016/s0364-0213(83)80009-3 [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. doi:10.1093/acprof:oso/9780195325539.001.0001 [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。）Google Books [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. doi:10.7551/mitpress/4549.001.00 [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. doi:10.7551/mitpress/1251.001.0001 [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. Google 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. doi:10.1017/cbo9780511529863 [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. doi:10.7551/mitpress/4879.003.0017 [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「让对策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. Google Books [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. doi:10.1126/science.185.4157.1124 [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对策，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。）doi:10.1093/0195146476.001.0001 [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. 链接 [②③] 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. Google Books [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. doi:10.1086/518628 [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. doi:10.1086/418301 [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. doi:10.1126/science.284.5411.87 [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. doi:10.1103/physicsphysiquefizika.2.263 [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. doi:10.1103/physrevlett.59.381 [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. doi:10.1126/science.286.5439.509 [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. doi:10.1137/070710111 [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. doi:10.1126/science.1216142 [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. doi:10.1007/bf00485230 [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. doi:10.1017/cbo9781139167093 [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. doi:10.1093/oso/9780195124415.001.0001 [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。）Google Books [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. Google Books [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. Google Books [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. doi:10.1080/01621459.1976.10480949 [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，文中写下「所有模型都是错的」；广为流传的完整说法「所有模型都是错的，有些是有用的」则见于他与德雷珀合著的《Empirical Model-Building and Response Surfaces》(1987，第 424 页)。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. Google Books [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**：当需要满足的是一个人真实的偏好或意图时，我们面对的是一种永久性的部分可观测：潜在的目标无法被直接读出。行之有效的应对，是把一个判断者放进回路，并且有节制、有针对性地向它提问。

所言非所欲

这是一个被讲述过无数次、却始终没有过时的场景：用户描述了自己想要的东西，工程师分毫不差地将它实现出来；到了交付之日，用户却说：「不，这不是我要的。」

这里没有人说谎。用户所言属实，工程师也照此实现。问题出在更深的层次：用户真正想要的东西，从一开始就没有、也不可能被完整地表述出来。本章考察的是，当必须满足的目标存在于另一个人的头脑之中时，善于应对者如何行事。这里的不可验证性，属于第 2 章所区分的五种处境中的「部分可观测」（partial observability）：相关的状态对我们隐而不见，而且这种隐蔽不是暂时的，而是永久的。目标无法从一个人的头脑中读取出来，因此我们也就无法验证自己是否满足了它。

偏好作为潜变量

首先需要把问题表述得更精确。用户的真实偏好是一个潜变量（latent variable）：它驱动着用户的种种反应，自身却从不直接显现，

我们只能从其行为中间接地加以推断。

更棘手的是，这一潜在的目标往往连用户自己也无法读出。心理学家斯洛维奇[11] 提出过一个不讨人喜欢、却有扎实依据的论断：偏好在许多情况下并不是被表达出来的，而是在被询问的那一刻才被构造出来的。问一个人想要什么，他给出的回答往往由提问的方式、当时摆在面前的选项以及他偶然想到的参照点共同塑造，并不是从某个预先存在、界定清楚的偏好储备中提取出来的。因此，「先把需求问清楚，再去实现」这一看似稳妥的次序，建立在一个常常不能成立的假设之上：需求作为一个确定的对象，在询问之前就已经存在。

由此可见，这并不是那种「信息暂时缺失、补齐即可」的处境。即便用户始终配合、知无不言，目标依然无法被准确地测得。就此而言，这是部分可观测在人身上最为纯粹的形态。

单次询问为何不足

倘若偏好是一个固定的靶子，那么原则上，问一次、问清楚便已足够。然而偏好并不是固定的靶子。

经济学很早就区分了两种偏好：陈述偏好（stated preference），即一个人自称想要什么；显示偏好（revealed preference），即一个人的实际选择所透露出的偏好。二者常常并不一致。需求文档是一种有损压缩：它把一个鲜活的、随情境而变化的意图，压缩为一份静态的条目清单；被舍弃的，是那些当时未曾想到、一见到成品便能立即指出的东西。不仅如此，意图还会漂移。人在见到一个具体的实现之后，其偏好会被这一实现重新塑造；他此刻想要的，已不再是项目启动时想要的东西。

因此，单次询问之所以失败，并不在于问得不够好，而在于这一对象的性质决定了它无法被一次询问所锁定。能够应对它的只有一种结构：持续地行动、观察、再修正。

第一种对策：把判断者放进回路

第一种应对，是承认目标无法被读出，进而在每一个决策点上，把唯一知晓目标的那个人请进来，由他来纠正偏差：行动，观察反应，更新，再行动。这就是把人放进回路，即通常所说的人在回路（human in the loop）；本书将这一对策称为神谕入回路。

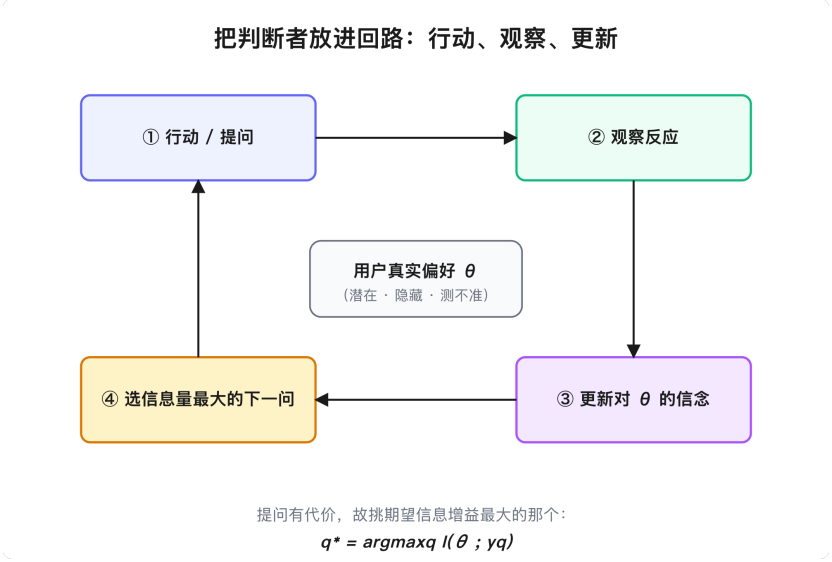


图 4: 把判断者放进回路：行动、观察、更新

这一回路曾在许多领域中被各自独立地重新发明。在人因工程中，谢里登[12] 提出的「人类监督控制」（human supervisory control），把人定位为居于自动化之上、负责监督与干预的判断者，而不是一个可以被一份规格一次性取代的角色。在可用性工程中，古尔德与刘易斯 1985 年[20] 的那篇经验总结，把它归结为三条原则：尽早并持续地关注用户，进行实证测量，迭代地设计。这三条原则朴素得近乎老生常谈，却为无数项目所违背。尼尔森[19] 后来将其发展为一整套可用性工程方法，并给出了一个出人意料的经验数字：只需五位用户参加测试，便能发现约八成五的可用性问题。因此，与其一次邀请二十人，不如分四轮进行、每轮五人，边测试边修改。

在其他领域，同一结构以不同的名目出现。推荐系统从用户的点

击、停留与跳过中习得其未曾言明的偏好，本质上也是这条回路。霍维茨 1999 年[22] 提出的混合主动式界面（mixed-initiative user interface），以及费尔斯与奥尔森 2003 年[23] 提出的交互式机器学习（interactive machine learning），所描述的也是同一件事：人与系统轮流采取行动，彼此校准。

这里需要防止以偏概全。交互式的偏好获取并不是某一种具体的技术，而是一个方法族：实验设计、主动学习、序贯决策，乃至强化学习中的探索，都是这条「行动-观察-更新」回路在不同假设下的实例。若把它说成「不过是 A/B 测试」或「不过是某个算法」，就会把一种普遍的做法窄化为一件工具。

第二种对策：把提问用在信息量最大之处

回路的运转有赖于不断向人提问，而提问是有代价的。用户的耐心、注意力和时间都很稀缺；问得过多或不得要领，人便会厌烦、敷衍，甚至离开。第二种对策由此而来：既然查验有成本，就应当把有限的提问用在信息量最大的地方。本书将这一对策称为最优筛查。

这一对策有一套清晰的理论。林德利在 1956 年[1] 给出了一个实验所提供信息的度量；霍华德在 1966 年[3] 提出信息价值理论（information value theory），把「是否值得付出代价获取某条信息」转化为一个可以计算的决策问题。贝叶斯实验设计（Bayesian experimental design）则将这些思想系统化了（查洛纳与韦尔迪内利[5] 的综述是这一领域很好的导览）：在所有可以提出的问题中，选出期望上最能削减不确定性的那一个。形式上，若 θ 是待推断的潜在偏好，y_q 是对问题 q 的回答，那么应当选择的，是期望信息增益最大的 q：

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

也就是使回答与目标之间的互信息（mutual information）最大化。机器学习把这一思想称为主动学习（active learning），其核心问题是：下一个标注花在哪个样本上最划算？Seung 等人 1992 年[7] 提出的委员会查询（query by committee），刘易斯与盖尔 1994 年[8]

提出的不确定性采样（uncertainty sampling），以及科恩等人 1996 年[6] 的统计式主动学习，处理的都是这一问题。当用户难以直接打分、却很容易在两个选项中挑出较好的一个时，成对比较便成为信息效率最高的提问方式之一。刻画成对比较的是布拉德利-特里模型[2]（Bradley-Terry model）：它为每个选项赋予一个潜在分数，由分数之差决定一方胜出的概率，即 $P(a \succ b) = \sigma(s_a - s_b)$。

在这里同样需要防止以偏概全。沙赫里亚里等人 2016 年[31] 的综述有一个耐人寻味的标题：「把人移出回路」。该文讨论的是以高斯过程进行贝叶斯优化（Bayesian optimization），由算法自动选出下一个应当尝试的点。这一方法极为有用，但它只是这个方法族中的一种实现，并不是「最优筛查」的全部。把这一对策等同于高斯过程，犹如把交通等同于汽车。

当代形态及其反噬

将这两种对策结合起来，便得到当今大模型对齐的主流方法：基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF）。这一方法的基础由克里斯蒂亚诺等人于 2017 年[27] 奠定；斯蒂农等人于 2020 年[28] 将其用于摘要任务，欧阳等人于 2022 年[29] 又将其用于 InstructGPT。其做法是：先利用人的成对比较学习一个奖励模型（reward model），再以这一模型作为人类偏好的代理指标来优化系统。由此，「行动-观察-更新」的回路与最优筛查结合在了一起。它的效果之显著，常以一个对比来说明：一个仅有 13 亿参数、经人类反馈微调的 InstructGPT，其输出受人偏好的程度，竟超过了规模大一百多倍、拥有 1750 亿参数的原版 GPT-3。可见，与人的偏好对齐，有时比单纯扩大模型规模更为重要。

然而它的失效方式，预示了本书后续各章的主题。首先，学得的奖励模型只是真实偏好的一个代理指标，因而会被系统所利用：系统学会的是取悦奖励模型，而不是取悦人；其输出看上去更好，实际上却更糟。这就是第 11 章将专门处理的 Goodhart 失效（Goodhart's law）。其次，回路中的那个「神谕」（即人）同样并不可靠：人会疲劳，会前后不一，会带有系统性的偏差。把人放进回路，并不等于把真理放进回路。贝恩布里奇 1983 年[13] 的「自动化的反讽」（Ironies

of Automation）一文早已指出这一点：越是把人推上监督者的位置，他就越缺乏维持判断力所需的实际操作与情境感知；而一旦真正需要他接管，他反而最缺乏准备。因此，信任的校准（trust calibration）成为一个独立的难题（参见李与西 2004 年的研究[15]）：人既可能过度依赖一个不应信任的系统，也可能弃用一个实际上可靠的系统。

由此可见，把人放进回路并没有消解不可验证性，只是让问题「搬了家」：从「我能否验证目标」，迁移到「我能否信任回路中这个不完美的判断者」。

小结

控制台前的人这一场景，揭示了两种对策。其一，当自身缺乏验证能力时，把一个判断者请进回路（神谕入回路）；其二，把代价高昂的查验用在信息量最大的地方（最优筛查）。这两种对策在本书中还将反复出现。第三部会把它们从这一场景中抽离出来，分别加以整理和命名：第 10 章讨论借来的判断，第 9 章讨论如何把查验用于信息量最大之处。

然而，本章的讨论始终依赖一个前提：我们仍然在场，回路仍在运转，我们随时可以观察、可以纠偏。下一章将撤去这一前提。当我们必须交出行动权，让一个系统在我们看不见的地方、面对未曾预演的情形自行决策时，验证的难题将呈现出更为严峻的面貌。

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. doi:10.1214/aoms/1177728069

[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「把提问用在信息量最大之处」这一对策的理论源头，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. doi:10.2307/2334029 [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. doi:10.1109/tssc.1966.300074 [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. Google Books [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「把提问用在信息量最大之处」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.

doi:10.1214/ss/1177009939 [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. doi:10.1613/jair.295 [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. doi:10.1145/130385.130417 [②] 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. doi:10.1007/978-1-4471-2099-5_1 [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. 链接 [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. 链接 [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. doi:10.1037/0003-066x.50.5.364 [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「偏好作为潜变量」一节的心理学支柱。
12. T. B. Sheridan (1992).《Telerobotics, Automation, and Human Supervisory Control》. MIT Press. Google Books [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. doi:10.1016/0005-1098(83)90046-8 [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 Au-

tomation」. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 30(3), 286-297. doi:10.1109/3468.844354 [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. doi:10.1518/hfes.46.1.50.30392 [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. doi:10.1518/001872095779049543 [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. Google Books [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. Google 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「所言非所欲」遥相呼应。

19. J. Nielsen (1993).《Usability Engineering》. Academic Press. Google Books [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. doi:10.1145/3166.3170 [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. Google Books [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. doi:10.1145/302979.303030 [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手、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. doi:10.1145/604045.604056 [②

④] 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. doi:10.1145/3290605.3300233 [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. doi:10.1145/1597735.1597738 [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. arXiv:1606.03137 [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. arXiv:1706.03741 [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。这把本章两种对策缝在一处，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. arXiv:2009.01325 [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. doi:10.52202/068431-2011 [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，是把本章两种对策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. 链接 [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. doi:10.1109/jproc.2015.2494218 [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，把这一整类对策等同于高斯过程，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，我们便无法验证它在未来可能遇到的每一种情形中会如何行动（开放世界）；如果它还会采取策略性的行为，就会再叠加一层对抗性的不可验证。于是，应对的思路从「证明它对」转向三件事：限制它所能造成的破坏，为我们的信任定价，并使它的行为可以事后核查。

交出行动权之后

在上一章中，我们始终在场。本章讨论的，是放手之后的情形。

设想这样几种情形：运行一段不受信任的代码；把工具与权限交给一个能够自行决定下一步行动的系统；让一辆自动驾驶汽车在我们不在车内时上路。行动权一经交出，新的难题随之而来：我们无法验证它在将要遇到的一切情形中会如何行动，因为这些情形大多未曾见过，也无法事先穷举。这就是所谓的开放世界。上一章的不可验证性源于目标隐藏在他人的头脑之中；本章的不可验证性则源于行为发生在未来，发生在我们看不见的地方。倘若这个系统还会采取策略性的行为，便又叠加了一层对抗。

2010 年 5 月 6 日的「闪电崩盘」可以视为一次预演：相互作用的自动交易程序在几分钟之内使道琼斯指数下挫近千点，随后又以几乎同样快的速度回升。没有任何一位程序员预见到交易会如此连锁放大。每个程序在测试中都没有问题，然而一旦组合在一起、置于真实的市场行情之中，便酿成了一场无人验证过的灾难。

测试无法弥合的缺口

测试所能覆盖的，只是有限的若干输入；系统将要面对的，却是一个开放的世界。二者之间的缺口，并不是「多测一些便能补上」的工程缺口，而有其原则上的根源。

莱斯定理的结论毫不含糊：程序的任何非平凡语义性质（即关于程序行为、且并非所有程序都具有或都不具有的性质）都是不可判定的。换言之，不存在一个通用算法，能够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」。这并不是算力不足的问题，而是逻辑上的不可能：它是图灵停机问题投射在「程序行为」上的影子。因此，对于任何一个足够通用的自主系统，我们所期望的那种保证，原则上都无法在事前一次性验明。

更深的一层困难，来自汤普森 1984 年发表的图灵奖演讲中那个著名的论证[9]：即便是正在运行的这个程序，我们也无法完全信任。一个被篡改过的编译器，可以在编译时暗中植入后门，再把痕迹从自身的源代码中抹去，使人即便审遍源代码也看不出破绽。我们所能验证的，永远只是某个表层；其下还有未曾检视、也无法穷尽的层次。将这两点合而观之：程序在未见过的输入上的行为无法验证，程序的最底层也无法完全验证。这是本书迄今所遇到的最为彻底的不可验证性。

当系统具有策略性

如果系统只是被动地在未见过的输入上出错，问题仍不过是「部分可观测」加上「开放世界」。然而一旦它拥有自己的目标，而这一目标又与我们的目标不完全一致，它就会主动地、有策略地行动，包括设法绕过我们的检查。此时，第 2 章所说的第五种处境，即对抗，便出现了。

这并非科幻式的忧虑，而有其结构上的根据。奥莫亨德罗在 2008 年[10]、博斯特罗姆在 2014 年[11] 都曾论及工具性趋同（instrumental convergence）：一个为几乎任何目标进行优化的智能体，都会附带地追求某些工具性的子目标，如自我保存、获取资源、抗拒被关停，因为这些子目标对几乎任何最终目标都有用处。特纳等人在 2021

年将其中一条证明为定理[14]：在相当一般的条件下，最优策略倾向于寻求权力，也就是倾向于那些保留更多选项的状态。

在当今的系统中，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，即所谓目标错误泛化[17]；此外，克拉科夫娜等人收集了大量「规范博弈」的实例[18]，其中系统精确地满足了写下的目标，却违背了设计者的本意。即便在最狭窄的层面上，对抗样本的研究也表明[19,20]：一个表现优异的模型，可以被人眼无法察觉的微小扰动诱导出荒谬的错误。

一个技术色彩较淡、却极为直观的例子，是微软于 2016 年推出的聊天机器人 Tay。它被设计为从与网民的对话中学习。结果，一群人有组织地以恶意言论「投喂」它；不到一天，它便开始发布带有种族主义色彩和攻击性的内容，上线约十六小时后即被紧急下线。被投放出去、能够学习、又遭遇一个蓄意作对的开放世界：这三个条件一旦同时具备，事前的测试便无从防范。

这一问题由来已久。经济学早已为它命名，称为委托代理问题[32,33]：当委托他人代为行动、却又无法对其完全监督时，代理人与委托人之间的利益偏离便会产生「代理成本」。两千年来，人类雇人、立约、设置监察，所应对的都是同一种结构；自主系统只是把它推向了一个新的尺度。

应对：从「证明它对」到「围住它的错」

既然无法在事前证明它是对的，行之有效的应对便不再执着于证明，转而提出三个问题：即便它出错，后果最坏会到什么程度？我们应当在多大程度上信任它？倘若它确实出了错，事后能否查明？三种对策分别回答这三个问题。

第一种对策，限损与围栏：缩小爆炸半径。所谓爆炸半径，是指一次失败所能波及的范围。缩小这一范围，是计算机安全领域最古老的经验。兰普森 1973 年提出的围堵问题[2]，萨尔策与施罗德 1975 年提出的最小权限原则[1]，指向的是同一个思想：只赋予一个组件完成其本职工作所必需的最小能力，并严格限定它所能触及的范围。沙箱、能力限制、职责分离，都是这一思路的不同实现。

在智能体的语境中，这一对策还多出一个面向，即可纠正性（corrigibility）：把系统设计成不抗拒被停下。索亚雷斯等人 2015 年关于可纠正性的工作[5]，奥尔索与阿姆斯特朗 2016 年关于「可安全中断的智能体」的研究[4]，以及哈德菲尔德-梅内尔等人 2017 年的「关停博弈」[6]，处理的都是同一个问题：如何使一个有目标的系统，不把「人按下停止键」视为需要抵抗的威胁。

第二种对策，校准与分级信任：信任不是开关。对系统输出的信任，不应只分为「可信」与「不可信」两档，而应维持一个经过校准的信心，按信心的高低分级行动。所谓校准，是指系统所报告的信心与它实际的正确率相符。这就要求系统所表现出的「自信」是可靠的；然而现代神经网络往往过度自信[21]，因此需要重新校准，或者借助共形预测[22,23] 给出具有覆盖率保证的不确定性估计。落实到操作层面，便是一条分级自治规则（允许、询问、阻止）：其输入为信心 p 与潜在危害 c，其中 τ_{hi}、τ_{lo} 为信心阈值，$c_{\max}$ 为可以承受的危害上限：

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

允许、询问、阻止：这种如今在各类智能体工具中随处可见的三档模式，实质上是把无法验证的问题「它是否正确」，替换为两个可以操作的问题：「它有多大把握」与「这一步有多危险」。

第三种对策，留痕与可审计：让错误事后现形。对于无法防范的错误，就使它能够被发现。维茨纳等人 2008 年提出的「信息问责」[24]，把重心从「事前阻止」转移到「事后追责」。证书透明度[25] 是一个实际运转的例子：它并不阻止证书被错误签发，而是让每一张证书都进入一个公开、可验证、不可篡改的日志，使错误签发无处隐匿。布伦戴奇等人 2020 年关于可信 AI 的报告[26]，通篇围绕同一个问题展开：如何使一个系统的行为产生可供第三方核验的证据。

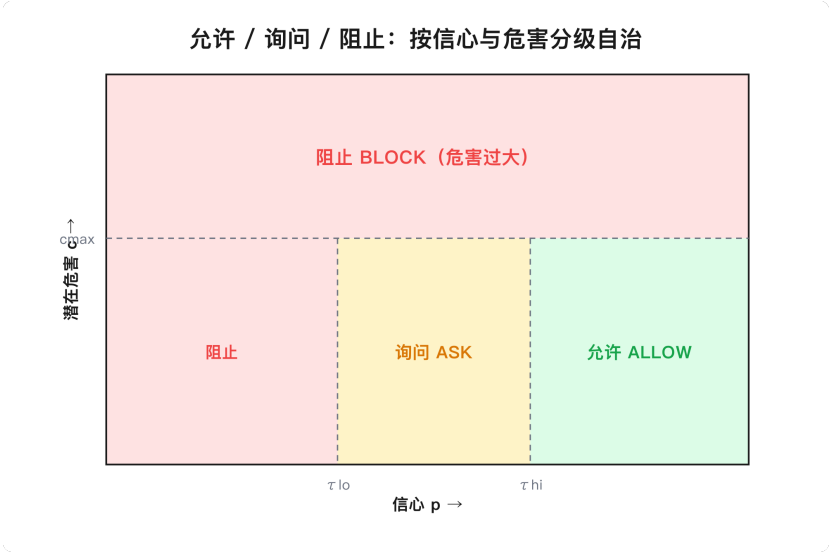


图 5: 允许 / 询问 / 阻止：按信心与危害分级自治

围堵的代价

与上一章的情形一样，三种对策都没有消解不可验证性，只是让它「搬了家」；而搬家是有代价的。

围栏可能被翻越：沙箱存在逃逸的风险，权限会逐渐蔓延。分级自治依赖那个被请来确认的人，而贝恩布里奇 1983 年的那篇短文早已指出[29]：越是把人置于监督者的位置，他就越会丧失真正需要接管时所必需的技能与情境感知。帕拉苏拉曼与赖利在 1997 年完整地列举了人对自动化的种种失当[30]，即误用、弃用与滥用；里森 1990 年的著作则揭示了这些失当如何系统性地发生[31]。至于留痕，其薄弱之处总在同一个地方：无人阅读的日志，等于没有日志。

更深一层的限制来自系统论的视角。佩罗 1984 年的著作论证[28]：一个系统如果既高度复杂、又紧密耦合，事故便不是偶发的意外，而是其结构的常态产物；再多的局部防护，也只能把失效推向更为隐蔽的组合。莱韦森 2011 年同样从系统的角度主张[27]：安全并不等于「让每个零件都可靠」，而是一个控制问题，须从整个系统的约束与反馈入手加以设计。由此可见，围堵能够压低单点失效的代价，却

无法消除复杂耦合所带来的风险。

交出行动权，所换来的从来不是「它一定不会出错」，而是「即便它出错，损害也是有限的、可见的，并且其中一部分能够被拦截」。在这种不可验证性之下，这已是所能得到的最好结果。

小结

放出去的智能体这一场景，催生了三种对策：缩小失败的爆炸半径（限损 / 围栏），依据经过校准的信心分级行动（校准），使失败可以事后核查（留痕 / 可审计）。第三部将把它们从这一场景中抽离出来，分别加以整理和命名：第 12 章讨论围堵与审计如何构成一对，第 11 章讨论校准。

委托代理的基本骨架（我们无法完全监督代为行动的一方），将在第 8 章以更大的尺度重现：届时，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。在此之前，下一章先进入一个最为纯粹的场景：数学。那里既没有隐藏的状态，也没有蓄意欺骗的对手，然而不可验证性依旧如影随形。

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（限损／围栏）

1. J. Saltzer & M. Schroeder (1975).「The Protection of Information in Computer Systems」. Proceedings of the IEEE, 63(9), 1278-1308. doi:10.1109/proc.1975.9939 [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。本章第一种对策「限损与围栏」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. doi:10.1145/362375.362389 [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. Google Books [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。它把本章三种对策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. 链接 [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，呼应本章第一种对策里的可纠正性面向。
5. N. Soares, B. Fallenstein, S. Armstrong & E. Yudkowsky (2015).「Corrigibility」. 收于《Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence》. 链接 [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。它是本章第一种对策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. doi:10.24963/ijcai.2017/32 [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. doi:10.1112/plms/s2-42.1.230 [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，莱斯定理与一切「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. doi:10.1090/s0002-9947-1953-0053041-6 [②] 莱斯定理在此被证明：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. doi:10.1145/358198.358210 [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. 链接 [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. Google Books [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. Google Books [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. Schulman & 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)》. arXiv:1912.01683 [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)》. arXiv:2201.03544 [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「目标错误泛化」：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，值得读者对照规范博弈一起看。
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. 链接 [②] 这篇文章及其配套清单收集了大量「规范博弈」实例：系统精确地满足了你写下的目标，却违背了你的本

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)》. arXiv:1312.6199 [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)》. arXiv:1412.6572 [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。

校准：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. arXiv:1706.04599 [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。这正是本章第二种对策的前提与障碍，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

向实践者的共形预测入门，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。它给本章第二种对策提供了可落地的不确定性量化工具，适合想把「校准的信心」真正用起来的读者。

23. V. Vovk, A. Gammerman & G. Shafer (2005).《Algorithmic Learning in a Random World》. Springer. doi:10.1007/b106715 [②] 这本书是共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. doi:10.1145/1349026.1349043 [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。这是本章第三种对策的纲领性表述，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. doi:10.17487/rfc6962 [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

把本章留痕这一对策扩展到整个 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. doi:10.7551/mitpress/8179.001.0001 [② ④] 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. Google 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. doi:10.1016/0005-1098(83)90046-8 [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. doi:10.1518/001872097778543886 [②④] 作者系统地列出并区分了人对自动化的失当：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. doi:10.1017/cbo9781139062367 [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. 链接 [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. doi:10.1016/0304-405x(76)90026-x [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**：数学中的不可验证（unverifiable）以最纯粹的形态出现：它是难解的（intractable），有时还是不可判定的（undecidable）。行之有效的应对有三种：验证一个有限的局部，并为之证明一个界（证书，certificate）；把目标替换为与之等价、却更易处理的陈述（代理替换，proxy substitution）；采用容许 ε 误差的概率方法（probabilistic method）。

无从估量的距离

难题之所以难，往往不在于它有多艰深，而在于人们无从估量自己离它还有多远。

登山者看得见山顶，调试程序的人能收到报错信息；他们至少知道方向是否正确，离目标是更近了还是更远了。证明一个数学猜想，却没有这样的凭借。离答案也许只差一个想法，也许还隔着一个世纪，而手边没有任何仪表能够指明究竟是哪一种情形。1859 年，黎曼在一篇仅有八页的论文中[9]，几乎是顺带地写下了后来以他的名字命名的猜想，并补充说，人们当然希望对此有一个严格的证明，但他在几次短暂而徒劳的尝试之后暂且把这件事搁置了，因为它对于他接下来的研究目的并非必需。这一搁置，至今已有一百六十余年。

本章考察的是，数学家站在这堵墙前实际做了些什么。之所以选择数学作为考察的场景，是因为它呈现了不可验证最纯粹的形态。这里没有隐藏的状态，没有蓄意欺骗的对手，也不存在时间不足的托

辞；命题非真即假，界限分明。然而，即便在这个最纯净的领域，验证依然系统性地不可得。倘若能够看清善于应对者在这里如何行动，本部其余几个更为驳杂的场景（控制台前的人、放出去的智能体、看不见自己的组织）便有了一个参照。

验证的鸿沟

首先需要准确地刻画这堵墙的形状。

核对一个证明容易，找到一个证明困难。给定一份完整写出的形式推导，逐行对照公理与推理规则加以检查，是一项机械的工作：原则上可以交由机器完成，并且必定在有限步之内给出「是」或「否」的回答。找到这份推导，则完全是另一回事。这种不对称是本章的基石。

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）不可解：不存在一个算法，能够对一阶逻辑中的任意命题判定它是否逻辑有效；而根据哥德尔完备性定理（Gödel's completeness theorem），判定有效性等价于判定可证性。再考虑一个足够丰富、可递归公理化且一致的形式系统（例如皮亚诺算术或 ZFC）：它的定理集是递归可枚举的（recursively enumerable），却不是递归的。换言之，我们可以把全部证明逐一枚举出来，却不存在一个程序，能够对每一个非定理都在有限步之内判定它不是定理。对证明的检验是可判定的，定理性却不可判定；两者之间的落差，就是那道鸿沟。

> **深入一层（可跳过）**：「足够丰富」这一限定至关重要。可判定的理论确实存在，例如普雷斯伯格算术（只含加法的自然数理论）和塔斯基的实闭域理论。在这些理论中存在判定程序，任何命题都可以机械地加以裁决。由此可见，不可判定性并非逻辑的普遍宿命，而是表达力越过某一门槛之后必须付出的代价。黎曼假设所属的解析数论，远在这一门槛之上。

更令人困扰的是：不仅证明难以找到，就连「我离答案是否已经接近」这一问题也没有判定程序。黎曼假设（Riemann hypothesis，下

文简称 RH）之所以令人苦恼，原因就在于此：它抗拒的不只是求解，还有对进展的估计。就第 2 章所区分的五种处境而言，本章处在「不可判定」与「难解」的交界：有些问题在原则上就不存在判定程序；有些虽有程序，代价却大到在宇宙的寿命之内也无法运行完毕。墙后是什么，无从看见；于是，高明的数学家不再正面凿墙，而是改换姿态。下面三种姿态会在本书其他章节中反复出现，只是名称不同。

证书与界

第一种姿态：放弃对整体的验证，只验证一个有限的局部，并为它证明一个有保证的界，即一个经过严格证明、适用范围明确划定的结论。

回到 ζ 函数。黎曼把欧拉那个与素数相联系的级数，解析延拓为整个复平面上的亚纯函数，并写出了它的完备化形式

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

这一对称的函数方程（functional equation）把临界带（critical strip）沿其中线左右对折。猜想断言：ζ 的全部非平凡零点都位于临界线（critical line）$\operatorname{Re}(s) = \tfrac{1}{2}$ 上（平凡零点位于负偶数处）。不可验证之处，就在「全部」这个全称量词上。

然而，有限多个零点是可以验证的。这里必须区分两件常被混为一谈的事。2004 年，古尔东计算了前 10^{13} 个零点，确认它们都位于临界线上[17]。这是以高精度浮点运算完成的数值计算，足以给人极强的信心，却不构成证书，因为它没有对舍入误差给出严格的界。普拉特在 2017 年所做的是另一件事[18]：他借助区间算术（interval arithmetic），严格证明了高度（即虚部）不超过约 3.06×10^{10} 的全部非平凡零点都位于临界线上；2021 年，普拉特与特鲁吉安又把这一高度推进到 3×10^{12}。只有后两项结果才称得上证书：一种有界的、局部的、可以机械复核的保证。这样的保证是真实的，但它不是定理。无论验证推进到多高，有限与「全部」之间的距离都不会

因此弥合。

这种姿态在数学中有可敬的先例。1896 年，阿达马与德拉瓦莱普桑各自证明了素数定理（prime number theorem）$\pi(x) \sim x/\ln x$，所依靠的是一个比 RH 弱得多、却力所能及的界：ζ 在直线 $\mathrm{Re}(s) = 1$ 上没有零点。既然无法证明零点都位于实部为 $\frac{1}{2}$ 的直线上，那就先证明它们都不在实部为 1 的直线上。这是以一个能够证明的弱命题，换取通向强命题途中的一段实质进展。

把同一种姿态移到软件领域，其面貌立刻变得熟悉。类型系统（type system）并不证明程序「完全正确」，只证明某一条特定的性质（不会把整数当作指针来解引用），以此换来一种可判定的检查。形式化验证（formal verification）走得更远：2017 年，黑尔斯团队发表了开普勒猜想（Kepler conjecture）的机器可核对证明[25]，把一个连人类审稿人都争论了多年的论证，转化为可以逐行验证的证书。其代价始终相同：证书换来的是局部的确定，付出的是普遍性；而局部终究不是定理。于是，有人转而改动目标。

代理替换

第二种姿态：不再固守原来的命题，而是把它替换为一个与之等价、却更易处理的陈述。这个代替原命题而成为攻坚对象的陈述，便是本书所说的代理。

RH 的等价表述多得惊人。1997 年，李建军给出了一个判据[14]：RH 成立，当且仅当对所有 $n \geq 1$，下列实数均满足 $\lambda_n \geq 0$：

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

其中求和取遍全部非平凡零点。一个关于零点位置的几何陈述，由此被转译为一列实数的非负性。奈曼与博伊林给出了另一个判据：RH 等价于示性函数 $\chi_{(0,1)}$ 属于一族经伸缩的小数部分函数所张成的子空间在 $L^2(0,1)$ 中的闭包。零点问题于是被转化为一个逼近问题。2003 年，巴埃斯-杜阿尔特把这一判据加强为只使用整数伸缩

的序列形式[16]，相应的条件是一列距离满足 $d_n \to 0$。拉加里亚斯在 2002 年甚至给出了一个初等到可以写在明信片上的等价命题[32]：对所有 n，$\sigma(n) \le H_n + \exp(H_n)\ln H_n$，其中 H_n 为调和数，σ 为因子和函数。

我自己也曾在这条路上走过一程。我曾把 RH 转写为李判据（Li's criterion）的形式，置入奈曼-博伊林-巴埃斯-杜阿尔特的逼近框架，又改用算子谱与随机过程的语言来表述，每一次都抱着同样的期望：换一种语言，困难或许会在新的坐标中显露出一处可以着力的地方。而每一次得到的结论也都相同：等价确凿无疑，难度却丝毫未减。我并没有解决这个问题，只是为它换了名称，换了表述。

这是代理替换在数学中典型的失效方式，值得为它取一个准确的名字：忠实却不更易的代理。等价性保证了它指向的仍是同一个真实目标（忠实），但它丝毫不比原问题更容易求解（不更易）。这一对策能否奏效，完全取决于能否同时做到忠实与更易；而两者兼得极为罕见，罕见到可以说，这一对策的全部技艺都系于此处。

若以这两个维度列成一张表，一条贯穿本章的线索便显现出来：

	更易	不更易
忠实	理想代理（罕见，技艺全在于此）	数学中的等价改写：只是为困难换了名称（本章）
不忠实	Goodhart：优化的是代理，真实目标却随之败坏（第 8、11 章）	无用，无人问津

数学家的困境位于表中由左下至右上那条对角线的右上端：忠实，却不更易。在后文讨论组织的一章中，失败落在左下端：一个更易却不忠实的代理，越是竭力优化它，真正在意的东西就越是遭到损害。同一种对策，沿两个相反的方向失效。第 11 章将把这两端正式联系起来。此处只需记住一点：替换目标既非作弊，也非出路；它是一种姿态，能否奏效是另一回事。

概率方法

第三种姿态最违背数学的本能：不再要求一个非此即彼的判决，转而持有一个经过校准（calibration）的概率，接受有界的出错风险，并照常行动。

素性检验（primality test）是最清楚的例子。判断一个大数是否为素数，确定性算法代价高昂；米勒-拉宾（Miller-Rabin）检验则换了一种问法：若 n 是合数，那么对随机选取的一个底，它通过检验、从而被误判为素数的概率至多为 $1/4$；独立进行 k 轮，误判概率便降至 $\le (1/4)^k$。此前，索洛维与施特拉森已在 1977 年给出误差 $\le (1/2)^k$ 的版本[20]，拉宾的版本则见于 1980 年[19]。「以 $1-\varepsilon$ 的概率为素数」与「已证明为素数」是性质截然不同的两种认识；但前者在工程上已经足够，而且所需的精度可以任意提高，多做几轮即可。

关键在于看清这里究竟放弃了什么。放弃的是确定性的类型，而不是严格性：那个 $(1/4)^k$ 的界是一条定理，其证明无懈可击。标准并没有降低，只是换成了一种能在预算之内兑现的标准。概率方法在纯数学中同样登堂入室。埃尔德什的概率方法（Alon 与 Spencer 将其写成了一部经典[26]）能够证明某个对象存在，所凭借的是证明它出现的概率为正，却并不把这个对象具体构造出来。存在性得到了证明，构造却付之阙如。

这种姿态一直延伸到人们对 RH 的信念。1973 年，蒙哥马利研究零点的配对关联（pair correlation）[30]，推导并猜想：经归一化之后，零点的配对关联函数形如

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

戴森在普林斯顿当即认出，它与随机矩阵理论中高斯酉系综（Gaussian unitary ensemble, GUE）本征值的配对关联完全一致。1987 年，奥德利兹科对大量零点进行数值计算[33]，证实二者的吻合达到了令人惊叹的程度。这些都不是证明，却是极强的证据。它们使数学家相信 RH；而这种相信，与物理学家相信一条尚未被证伪的定

律，并无本质区别。由此可见，数学内部同样发展出了一套在无法验证时形成信念的方法。

数学家如何判断

由此进入本章关于人的部分：在神谕始终缺席的情况下，有判断力的人依靠什么持有信念，又依靠什么决定把力气投向何处?

数学对外是演绎的，对内是似真的。波利亚在《数学与猜想》[2] 中专门讨论合情推理：数学家在尚无证明之时，如何借助类比、归纳与特例来估量一个命题的分量。阿达马考察了数学发明的心理[3]，记录了其中酝酿与顿悟的节律。庞加莱则留下了那个著名的瞬间[5]：在踏上公共马车踏板的一刹那，富克斯函数与非欧几何之间的联系毫无征兆地浮现在他的脑海中。这些都不能替代证明；它们是在证明到来之前、在那段没有仪表可依的路途上，人们实际依靠的东西。

数学家为何在证明出现之前就相信 RH? 因为证据来自多个方向，并且彼此印证：大量零点已被验证位于临界线上；众多等价形式中没有一个被推翻；它的一个类比形式已被证明（有限域上代数簇的黎曼假设，由德利涅证明）；零点的统计行为与随机矩阵理论的预言精确吻合。其中没有任何一条是证明，但合在一起，它们构成了一种有纪律的信念。

数学界自身也曾反思这种信念的地位。1993 年，贾菲与奎因围绕「理论数学」引发了一场争论[22]，所追问的问题是：猜想驱动、证据先行的工作，在多大程度上算得上数学。瑟斯顿 1994 年的《论数学中的证明与进展》[21] 则主张，数学所推进的是人类的理解，而不只是形式证明的库存。陶哲轩追问何为好的数学[29]，而他列举的标准中没有一条是「已被证明」。综合这些讨论可以看到，判断「某个命题多半为真、值得投入」的能力，就是一种经过校准的信念；第四部将专门为它命名。善于应对者在神谕缺席时既不会手足无措，也不假装确定：他们持有一个标定了刻度的信念，并照常行动。

小结

面对最纯粹的不可验证，数学家并没有等来一个判定程序。他们得到的是以下几样东西：证书（验证一个有限的局部，证明一个界）；代理替换（把目标换成等价的陈述，同时坦然承认它往往忠实却不更易）；概率接受（放弃非此即彼的判决，依据经过校准的概率行动）；以及在证明到来之前持有信念的判断力。

这些都不是数学独有的权宜之计。把不受信任的代码置于沙箱之中，对一个组织进行审计，在界面背后揣摩用户未曾言明的偏好：在这些场合，人们所诉诸的都是同样几种办法，只是术语不同。第三部将把每一种对策从其原生的领域中抽离出来，分别命名，并在不同领域之间对照排列；由此形成的那张对照表，是全书最主要的贡献。

最后还有一点需要在此说明。本书自身的核心命题，即「应对会收敛到同一小组对策」这一论断，我目前同样无法验证。我对它的相信，与数学家对 RH 的相信属于同一类：一个建立在跨领域证据之上、标定了刻度、却没有证明的信念。第 14 章将回到这一问题，让本书亲自践行它一直在描述的做法。

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. Pólya (1945).《How to Solve It: A New Aspect of Mathematical Method》. Princeton University Press. doi:10.1515/9781400828678 [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「数学家如何判断」一节关心的东西。
2. G. Pólya (1954).《Mathematics and Plausible Reasoning》(2

卷）. Princeton University Press. Google Books [①④] 两卷分别讨论数学中的归纳与类比，以及合情推理的逻辑结构，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」，是理解合情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. Google Books [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. Google Books [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. Google Books [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. Google Books [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). doi:10.1002/cpa.3160130102 [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. doi:10.1017/cbo9781139171472 [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. 链接 [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. Google Books [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. Google Books [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. 链接 [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). 链接 [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). doi:10.1006/jnth.1997.2137 [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). doi:10.1006/jnth.1999.2392 [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). arXiv:math/0202141 [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）.

链接 [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). doi:10.1090/mcom/3198 [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). doi:10.1016/0022-314x(80)90084-0 [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). doi:10.1137/0206006 [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). doi:10.1090/s0273-0979-1994-00502-6 [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). doi:10.1090/s0273-0979-1993-00413-0 [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. Google Books [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). doi:10.1215/ijm/1256049011 [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. doi:10.1017/fmp.2017.1 [② ③ ④] 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. Google Books [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. Google Books [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. 链接 [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). doi:10.1090/s0273-0979-07-01168-8 [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. doi:10.1090/pspum/024/9944 [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. 链接 [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). doi:10.1080/00029890.2002.11919883 [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). doi:10.1090/s0025-5718-1987-0866115-0 [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**：大型组织或国家无法直接观察自身的知识与活动：它们分散各处，部分隐而不见，有时还遭到刻意掩盖。于是，组织转而依赖可以观察的代理指标（代理替换的 Goodhart 崩塌在这里表现得最为触目），再以审计（留痕）与冗余加以补充。

看不清自身的庞然大物

在讨论放出去的智能体时，我们已经遇到过委托代理（principal-agent）问题。现在把它的尺度放大：委托方不再是一个人，而是一整个组织、一个国家；受托行动的，则是散布各处的成千上万的人。由此出现了一种新的、近乎荒诞的局面：这个庞然大物看不清它自己。

它想知道自己有多少人、种植着什么、谁在做什么、做得好不好，然而这些知识并不存放在任何一个它能够直接读取的地方。本章的问题是：当需要验证的对象是组织自身那些分散的、隐藏的、甚至会主动规避观察的知识时，组织如何应对？这里的不可验证叠加了前面讨论过的两种处境：一是部分可观测（partial observability，知识分散在边缘），二是对抗（被观察者会反过来操纵那些被观察的量）。

分散的知识

哈耶克 1945 年的《知识在社会中的运用》[1] 明确揭示了这个问题：一个社会赖以运转的知识从不集中于任何一处，而是分散在无数个体手中。这是关于特定时间、特定地点的局部知识（local knowledge），而且往往难以言明。哪台机器今天出了点小毛病，哪位客户即将流失，哪条小路雨后会塌方：掌握这些知识的人常常并未意识到它们算得上「知识」，更无从将其整理成形、上报给中心。波兰尼把这一层称为默会维度（tacit dimension）[2]：我们知道的，远多于我们说得出的。

这意味着，组织所面对的并不只是「信息尚未收齐」。即便每个人都忠诚合作，那些局部的、默会的知识也会在汇集的过程中流失。中心所期望的那幅「组织全貌」，在原则上就无法被如实地纳入任何一个可供验证的容器。这是部分可观测在社会尺度上的形态，而且它有一条更为坚硬的底线：这些知识本质上是局部的，无法汇集于任何一处。

清晰性的冲动

看不清，便设法使之变得可以看清。斯科特在 1998 年的《国家的视角》[3] 中为这种冲动取了一个准确的名字：清晰性（legibility）。国家若要对社会采取行动，首先须把社会改造成它能够读懂的样子。它丈量土地，绘制地籍图；为原本只有小名、绰号或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。这些都不是中性的记录工作，而是对现实的重塑，目的是让现实能够被填入表格。哈金的《驯服偶然》[32]、德罗西埃的《大数字的政治》[31] 以及鲍克与斯塔尔的《分类及其后果》[30] 从不同的侧面考察了这一过程，合起来便是一部「把社会变得可数」的历史。

清晰性的危险在于，地图必然有所简化，而组织一旦只依照地图行动，被地图抹去的东西便会反噬。斯科特最有力的案例是科学营林（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。第一代长势喜人；到了第二代，土壤耗竭，虫害蔓延，森林成片死亡，

德语甚至为此造出了一个词：Waldsterben，即「森林之死」。地图越是整洁，被它抹去的那些维系系统运转的局部知识，其缺失就越是致命。组织以这种方式为自己制造它所缺乏的可观测性，代价则是亲手削平了使它得以运转的复杂性。

代理指标与 Goodhart 崩塌

清晰性的冲动最常见的产物是指标。组织真正在意的东西（健康、学习、生产力、公共福祉）无法直接观测，于是它转而依赖可以观测的代理指标（proxy）：KPI、GDP、考试分数、论文引用数、急诊等待时长。

这就是第 7 章讨论过的代理替换，但它在这里以相反的方式失效，而这一对照是贯穿全书的一条主线。数学家的代理忠实却不更易：等价改写确实等价，问题却并未因此变得更容易解决。组织的代理恰好相反，更易却不忠实：指标固然容易测量，然而一旦被当作目标，它与真实目标之间的对应关系就会断裂。

这种断裂有许多名称，在不同的学科中各有表述。早在 1956 年，里奇韦就在《绩效度量的失能后果》一文中[7] 对这类后果作了系统的编目。1975 年，古德哈特[4] 指出：一旦某个指标被当作政策目标，它作为指标的可靠性就会瓦解。同样在 1975 年，克尔发表《奖励 A 却指望 B 的蠢事》[8]，使这一现象成为管理学的常识。1979 年，坎贝尔[6] 阐述的是同一现象在社会领域中的版本。斯特拉森则为它留下了最为精炼的表述[5]：当一个度量变成目标，它就不再是个好度量。

更深的一层是反应性（reactivity）。埃斯佩兰与索德尔 2007 年[12] 指出，公开的排名与其说是在描述世界，不如说是在重塑世界：被排名的大学会依照排名的算法改变自身，于是指标所「测量」的，变成了由它自己催生出来的行为。相关的现象也早有记录：史密斯 1995 年[10] 分析了公开发布绩效数据如何招致一连串始料未及的后果，贝文与胡德[11] 则记录了英国医疗系统中针对指标的博弈。若再往前追溯，这一切的总源头是默顿 1936 年的文章《有目的的社会行动的非预期后果》[9]。

这类崩塌在现实中屡见不鲜，代价有时十分惊人。为了达成「交叉销售」的账户指标，富国银行的员工在客户不知情的情况下私自开设了大量虚假账户。2016 年事件曝光时，这类账户估计约有两百万个（后经复查增至约三百五十万个）。银行被处以一亿八千五百万美元的罚款，五千多名员工遭到解雇。那个被奉为目标的数字，反而摧毁了它本应衡量的客户关系。更早还有一则近乎寓言的案例，发生在殖民时期的德里：当局为了灭蛇，悬赏收购眼镜蛇的尸体，市民于是养蛇领赏；赏金一经停止，这些蛇被尽数放生，蛇患反而比从前更加严重。「眼镜蛇效应」（cobra effect）由此得名。

代理指标为何注定会被扭曲？委托代理理论对此给出了严格的解释。霍尔姆斯特伦 1979 年提出的信息性原理（informativeness principle）[14] 指出：报酬应当挂靠在对「努力」具有信息量的信号上。然而，一旦努力是多维的，而可以测量的只是其中几个维度，问题便随之而来。霍尔姆斯特伦与米尔格罗姆 1991 年的多任务委托代理分析（multitask principal-agent）[15] 清楚地表明：当一个人需要同时兼顾可测与不可测的任务时，对可测部分的奖励越重，他就越会把努力从不可测的部分转移到可测的部分。设真实目标为 G，可观测的代理指标为 P，二者在现状下彼此相关。问题在于，这种相关性是行为的产物，而不是客观规律。一旦以 P 作为施压的目标，

$$\arg\max_{a} P(a) \qquad \text{vs.} \qquad \arg\max_{a} G(a),$$

理性的代理人便会寻找那些能够抬高 P、却无助于甚至有损于 G 的行动，原有的相关性也就被优化的压力碾碎了。教师为考试而教，医院通过调度病人来压低某一项等待时长，研究者把一篇论文拆分成可以计数的最小发表单元，都出自同一机制。

以审计与冗余为补充

代理指标若单独使用便会崩塌，因此组织又引入另外两种对策加以补充；这两种对策在本书中同样反复出现。

其一是留痕与审计。复式记账是人类最古老的审计链之一。索尔在

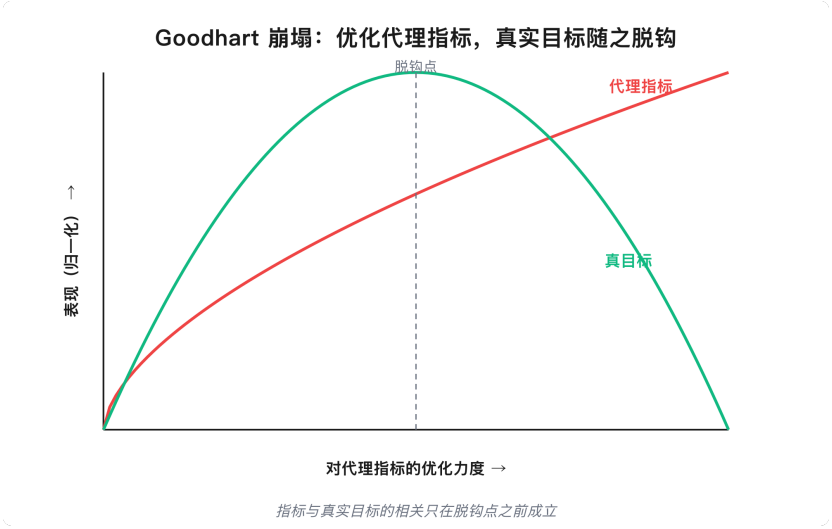


图 6: Goodhart 崩塌：优化代理指标，真实目标随之脱钩

《账簿与权力》[22] 中论证，编制可供核查之账目的能力，与一个个国家的兴衰直接相关：算得清自己的，才能持久。现代的财务审计与独立稽核，都是在「事前无法防范舞弊」的条件下，换取「事后能够查出舞弊」的可能。然而，这一对策有其自身的病症。鲍尔 1997 年的《审计社会》（audit society）[20] 一语道破：当验证沦为一种仪式，组织生产出来的不过是「一切尽在掌握」的表象，而不是实际的掌握。肖尔与赖特所描述的「审计文化」（audit culture）[17]，以及奥尼尔在 2002 年里斯讲座中对「信任」的反思[21]，指向的都是同一种异化：为了能够被问责，机构把大量精力耗费在制造可供检查的痕迹上，真正的工作反而被挤到一边。

其二是冗余与共识。兰道 1969 年那篇被低估的文章[16] 为「重复与重叠」正名：在一个各部件都不完全可靠的系统中，冗余（redundancy）不是浪费，而是可靠性的来源；若干相互独立的核查，比单一的权威更难被同时欺骗。这一对策的成立有一个前提，即各项核查彼此独立；下一部将反复强调这一点：如果几项核查出自同一来源，相关的失效便会一举击穿冗余最核心的承诺。

小结：第二部的收束

四个场景至此考察完毕。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，它们所面对的不可验证，来源迥然不同：隐藏在内心的偏好，开放世界中的未来行为，原则上不可判定的命题，分散且会主动规避观察的知识。然而，它们所诉诸的是同一小组对策。

其中最值得并置的，是代理替换两种方向相反的失效方式。数学家受挫于「忠实却不更易」，组织则受挫于「更易却不忠实」。第 7 章那张 2×2 表格的两端，如今都有了血肉。它们不是两种对策，而是同一种对策的两个失效方向。一个好的代理必须同时避开这两端，既忠实又更易；两者兼得极为罕见，而这一对策的全部技艺，几乎就在于做到这一点。第 11 章将正式把这两端联系起来。委托代理的那副骨架，也从第 6 章的一段代码，扩展成了这里的一个国家。

至此，第二部已在各个场景之中，以彼此交织的方式展示了这些对策。它们分散各处，名称各异，隐没在各自领域的术语之中。第三部将转换角度：把每一种对策从其所属的领域中抽离出来，剥除各自的术语，为之统一命名，使一个名称便能涵盖所有场景。由此得到的那张跨领域对照表，将是本书最主要的贡献。

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. 链接 [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「分散的知识」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. Google Books [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. Google Books [②④] 斯科特提出「清晰性」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「清晰性的冲动」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「代理指标与 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. doi:10.1002/(sici)1234-981x(199707)5:3<305::aid-euro184>3.0.co;2-4 [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. doi:10.1016/0149-7189(79)90048-x [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. doi:10.2307/2390989 [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. doi:10.2307/255378 [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. doi:10.2307/2084615 [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、清晰性反噬等一系列「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」.《International Journal of Public Administration》, 18(2-3), 277-310. doi:10.1080/01900699508525011 [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. doi:10.1111/j.1467-9299.2006.00600.x [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. doi:10.1086/517897 [②④] 埃斯佩兰与索德尔以法学院排名为例，提出「反应性」(reactivity)：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「更深的一层是反应性」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. doi:10.1177/000312240907400104 [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，前者讲反应性机制，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. doi:10.2307/3003320 [②④] 霍尔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. doi:10.1093/jleo/7.special_issue.24 [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. doi:10.2307/973247 [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「以审计与冗余为补充」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. doi:10.2307/2661148 [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. doi:10.23943/9781400889433 [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. doi:10.1515/9781400821617 [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，与本章清晰性与审计两节互为背景。
20. M. Power (1997).《The Audit Society: Rituals of Verification》. Oxford University Press. Google Books [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「以审计与冗余为补充」一节借它点破审

计这一对策自带的病：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. Google Books [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. Google 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. Google Books [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. Google Books [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. Google Books [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. Google Books [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. Google Books [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. Google Books [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. Google 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. doi:10.7551/mitpress/6352.001.0001 [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，说明分类是清晰性工程的隐形一环。
31. A. Desrosières (1998). 《The Politics of Large Numbers: A History of Statistical Reasoning》(trans. C. Naish). Harvard University Press. Google Books [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，揭示清晰性背后那套统计装置的来历。

32. I. Hacking (1990). 《The Taming of Chance》. Cambridge University Press. doi:10.1017/cbo9780511819766 [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**：有两种对策直接针对不确定性。其一是在可检验的局部给出有保证的界（证书）；其二是把有限的检查分配到最能削减不确定性之处（最优筛查）。

第二部考察了四个具体场景：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。在这些场景中，各种对策彼此交织，总是嵌在具体问题之中出现，并以各领域自身的术语表述。自本部起，论述的角度有所转换：把每一种对策从其所属的语境中抽离出来，剥除领域术语，考察它的一般形式，然后再返回各个领域，辨认它在不同术语下的对应形态。全书最主要的贡献，就在于由此得到的一张对照表：同一种对策，在不同领域的术语中各有其名。

八种对策两两成对，分四章论述。这种分组并非出于叙述上的便利，而是一个论断：同一章中的两种对策，处理的是问题的同一个方面；至于每一种对策确切作用于何处，则留待第 13 章说明。本章讨论证书与最优筛查。二者针对的都是不确定性，但着手的方向相反：前者在可检验的局部证明一个界，即确定误差不会超出某一限度；后者则把有限的检查资源分配到最能削减不确定性的地方。

此外，按照第 4 章确立的那条不容通融的原则，每提出一种对策，每将两个领域并置，都须重新追问：这种迁移是实质性的（机制相同、失效方式相同、权衡相同），还是只是一个看似贴切的比喻？

证书：为局部提供可复核的保证

第一种对策的一般形式是：不试图验证整体，而是给出一个局部的保证。这个保证具有明确的界，并且可以由任何人独立复核。换言之，交付的不是「它完全正确」这一断言，而是「在此范围之内，其误差至多如此」，以及一份他人能够迅速核实的凭据。本书所说的证书，指的就是这样一份凭据及其所担保的界。

考察这一对策在各领域中的形态，会发现它们惊人地一致。在机器学习中，它称为泛化界（generalization bound）。瓦利安特在 1984 年提出 PAC 框架[1]；瓦普尼克与切尔沃年基斯的 VC 维，则于 1989 年由布卢默等人纳入这一框架[4]。二者结合所给出的保证是：以至少 $1-\delta$ 的概率，模型的真实误差不超过其经验误差与一项复杂度惩罚之和，

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

模型在未来全部数据上的表现无从验证，那属于开放世界的问题。然而，已观测的样本构成一个可检验的局部；以此为依据，可以证明一个附带置信度的界，并且这个界对未观测的数据同样成立。为这一保证提供概率基础的，是霍夫丁不等式（Hoeffding's inequality）[17]；麦卡莱斯特的 PAC-Bayes 界[6] 则对它作了进一步的精化。

在软件领域，这一对策表现为类型与证明。类型系统并不证明程序「完全正确」，只证明某一条性质成立（例如程序不会把整数当作指针）；由此换得的，是一种可判定、并且能够机械复核的检查（皮尔斯[7]）。柯里-霍华德对应（Curry-Howard correspondence，霍华德 1980[8]）进一步在「证明」与「程序」之间画上了等号。内库拉 1997 年提出的「携带证明的代码」（proof-carrying code）[9]，把这一对策推向了极致：不受信任的代码自带一份关于自身安全性的证明，宿主只需迅速核对这份证书（certificate），无须自行从头推导。勒罗伊 2009 年的 CompCert 编译器[10] 经过了形式验证；它与 de Moura 和比约纳 2008 年的 Z3 求解器[11] 一样，把同一思路做成了工业级的工具。

数值计算中与之对应的是误差界。希格姆 2002 年的后向误差分析[12]与摩尔 1966 年的区间算术（interval arithmetic）[13]，使计算能够带着一个保证包含真值的区间进行，最终交付的是一个有保证的范围，而不是一个可能造成误导的浮点数。在数学中，第 7 章那些已验证至高度 T 的零点也属于此类：它们给出的是一个界，而不是一条定理。

贯穿上述各例的共同观念是：证书是一种局部的、有界的、可独立复核的保证。它最精妙之处，在于利用了第 7 章称为数学基石的那种验证不对称：生成一份证书可能代价极高，核对它却极为廉价。携带证明的代码、NP 问题的解、数学证明，都得益于这一不对称。

这一对策的典型失效方式只有一种，却相当常见，那就是空洞的界。一个保证可以为真，却毫无用处。「误差不超过百分之百」「该模型的泛化误差是有限的」，这类陈述在逻辑上无懈可击，在实践中却一文不值。由此可见，界的价值不在于它成立，而在于它足够紧，紧到可以据此行动。

最优筛查：把检查分配到信息量最大之处

第二种对策的一般形式是：信息的获取需要付出代价，因此应当把有限的检查分配到边际效益最大的地方，即每增加一次检查所能削减的不确定性最多之处。

这一对策在各领域中的形态同样整齐。在统计学与科学研究中，它称为实验设计（design of experiments）。费雪 1935 年的《实验设计》[19] 与博克斯等人的《实验者统计学》[20]，所讲授的都是如何以尽可能少的试验获取尽可能多的信息。而要比较不同实验的优劣，还需要一个衡量信息多寡的尺度：林德利 1956 年给出了「一个实验所提供的信息」的度量[21]，沙洛纳与韦尔迪内利则把贝叶斯实验设计整理为一个完整的体系[22]。另一个问题是检查应当何时停止。瓦尔德 1945 年的序贯检验[23]，使研究者可以在收集数据的同时，决定是否继续收集。为上述所有工作提供共同计量单位的，则是香农 1948 年的信息论[14]。

在机器学习中，这一对策称为主动学习（active learning），它要回答

的问题是：接下来标注哪一个样本，所得的收益最大（科恩等人[33]；塞特尔斯[34]）。软件测试中的模糊测试（fuzzing）遵循同样的逻辑，它所关心的是：把算力投向哪些输入，最有可能引发崩溃（米勒等人 1990 年的开创性实验[35]）。这一对策在今天的应用规模之大，颇为惊人。谷歌的 OSS-Fuzz 自 2016 年起，持续向上千个开源项目自动输入海量的畸形数据，迄今已发现数以万计的缺陷与漏洞。任何人工测试团队都无法达到这样的穷举程度；OSS-Fuzz 所依靠的，就是把算力持续不断地投向最可能引发崩溃之处。审计中的抽样也是如此：选择核查哪几笔交易，最有可能发现问题。在界面设计中，它则体现为「应当向用户提出哪一个问题」（第 5 章）。

上述做法背后是同一个最优化问题：选择所要提出的查询，使回答与未知量之间的期望信息增益（expected information gain）最大，

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

倘若检查需要反复进行，并且要边检查边利用所得的结果，问题便转化为探索与利用（exploration and exploitation）之间的张力，即多臂老虎机（multi-armed bandit）问题：面对若干收益未知的选择，既要试探尚不了解的，又要利用已知较优的。对这一问题的研究，早期有汤普森 1933 年提出的采样法[24] 与罗宾斯 1952 年的开创性工作[25]；其后，赖与罗宾斯 1985 年给出了最优分配的规则[26]，奥尔等人 2002 年又作了有限时间分析[27]。这些工作共同给出了下述问题的最优解：应当投入多少次试验，去削减对哪一个选项的不确定性。其结论是，这类策略的遗憾（regret），即因未能始终选择最优选项而累积的损失，随时间仅按对数增长，

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

此处需要警惕叙述上的一种路径依赖。最优筛查是一个方法族，实验设计、主动学习、审计抽样、模糊测试与多臂老虎机都属于其中。从库什纳、莫库斯的工作，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]，再到沙赫里亚里 2016 年那篇

综述所梳理的、基于高斯过程的贝叶斯优化[36]，这一脉络极为有用；但它只是该方法族中的一种实现，并不是「筛查」的全部。若把这一对策等同于高斯过程，便是把一种普遍的做法缩减为一件具体的工具。

这一对策的典型失效方式同样只有一种：所优化的信息度量设定有误。信息的收集可以极其高效，收集到的却是关于一个错误问题的信息；或者，被全力最大化的那个「信息量」，并不反映人们真正关心的东西。筛查的方法越精巧，错误的度量就越快地把人引入歧途。

两种对策的关联

将两种对策并置，二者的关系便清楚了。证书在一个可检验的局部上，把不确定性约束在一个有保证的界之内；筛查则动用有限的信息预算，去观测最能削减不确定性的那个局部。前者是在能够检验之处把界证紧，后者是把检查用在最需要检查之处。二者从两端夹击同一个对手，即未知；这也就是它们的共同任务：在信息预算有限的条件下，决定在何处削减不确定性，以及投入多大的力度。它们各自的确切位置，将在第 13 章中予以确定。

然而在某些情形下，无论怎样约束、怎样筛查都不够，因为行动者缺乏作出这一判断的能力。此时已无法仅凭自身缩小未知，而必须从外部借来判断。下一章的两种对策，处理的就是这种情形。

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。

1. L. Valiant (1984).「A Theory of the Learnable」. Communications of the ACM, 27(11), 1134-1142. doi:10.1145/1968.1972 [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」这一做法的源头。

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. doi:10.1137/1116025 [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. doi:10.1007/978-1-4757-2440-0 [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，把第一种对策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. doi:10.1145/76359.76371 [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. doi:10.1016/0020-0190(87)90114-1 [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.

doi:10.1145/307400.307435 [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，正文称其为第一种对策的「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. Google Books [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，是理解证书这一对策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. Google Books [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，正是证书这一对策的逻辑核心。
9. G. Necula (1997).「Proof-Carrying Code」. Conference Record of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL), 106-119. doi:10.1145/263699.263712 [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 7 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. doi:10.1145/1538788.1538814 [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，是证书这一对策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. doi:10.1007/978-3-540-78800-3_24 [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，是证书这一对策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. doi:10.1137/1.9780898718027 [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」这一做法的标准参考，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. Google Books [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，正是证书这一对策在该领域的化身。
14. C. Shannon (1948).「A Mathematical Theory of Communication」. Bell System Technical Journal, 27(3), 379-423; 27(4), 623-656. doi:10.1002/j.1538-7305.1948.tb01338.x [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. doi:10.1016/0005-1098(78)90005-5 [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. doi:10.1007/978-0-387-49820-1 [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. doi:10.1080/01621459.1963.10500830 [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」，证书这一对策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. doi:10.1109/tit.2005.862083 [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. Google Books [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，是最优筛查这一对策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. Google Books [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. doi:10.1214/aoms/1177728069 [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. doi:10.1214/ss/1177009939 [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. doi:10.1214/aoms/1177731118 [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. doi:10.2307/2332286 [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. doi:10.1090/s0002-9904-1952-09620-8 [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. doi:10.1016/0196-8858(85)90002-8 [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. doi:10.1023/a:1013689704352 [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. doi:10.1115/1.3653121 [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. Google Books [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. doi:10.1023/a:1008306431147 [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. doi:10.7551/mitpress/3206.001.0001 [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. arXiv:0912.3995 [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，原本就是同一项任务的两面。
33. D. Cohn, Z. Ghahramani & M. Jordan (1996).「Active Learning with Statistical Models」. Journal of Artificial Intelligence Research, 4, 129-145. doi:10.1613/jair.295 [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. 链接 [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，把这一对策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. doi:10.1145/96267.96279 [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. doi:10.1109/jproc.2015.2494218 [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**：当自身缺乏验证能力时，可以从外部引入这种能力：或者在决策回路中安置一个可信的判断者（神谕，oracle）；或者召集许多相互独立、各自并不可靠的判断者，信任他们之间的一致（冗余，redundancy / 共识，consensus）。

上一章的两种对策，仍是依靠自身的资源来缩小未知。然而有时所缺乏的并不是信息，而是判断力：面对眼前的问题，行动者无法独立作出可靠的判决。本章的两种对策于是转向外部，从别处借来判断。借用的方式有两种：其一是引入一个值得信赖的判断者（神谕）；其二是召集许多彼此并不信任的判断者，转而信任他们之间的一致（冗余）。

神谕入回路：引进一个判断者

第一种对策的一般形式是：在缺乏验证能力的那个决策点上，插入一个外部的判断者，由它作出决策者自身无法作出的判决。

这一对策最朴素的形式，是第 5 章讨论过的人在回路（human in the loop），此外还有专家会诊，以及把疑难问题提交上级裁决的做法。然而，它最深刻的形态，却隐藏在两个看似毫不相干的领域之中。

其一是交互式定理证明（interactive theorem proving）。从德布鲁因 1970 年的 AUTOMATH[1]，到爱丁堡 LCF（戈登、米尔纳与沃

兹沃思 1979[2]），再到今天的 Coq（贝尔托与卡斯特朗 2004[3]），这些系统采用的都是同一种分工。人提供机器无法给出的证明思路，即那种灵光一现的洞见（神谕）；机器则一丝不苟地核对每一步推理（证书检查，certificate checking）。神谕负责「找」，机器负责「验」，这恰好对应第 7 章讨论过的那种不对称：寻找困难，验证容易。

其二则更令人惊讶：复杂性理论中的交互式证明（interactive proof）。设想一个计算能力很弱的验证者，面对一个能力强大却不可信的证明者：对于一个验证者自己无法计算的问题，它如何才能获得可靠的答案？戈德瓦塞尔、米卡利与拉科夫 1989 年[6] 以及巴拜 1985 年[7] 给出的回答是：反复盘问，辅以随机挑战。验证者提出连它自己也无法预知的随机问题；证明者倘若说谎，迟早会在某一次挑战中暴露。沙米尔 1992 年[15] 证明了一个惊人的结果，即 IP = PSPACE：仅凭这种「盘问一个不可信的神谕」的方法，弱小的验证者便能可靠地裁决一类极其庞大的问题。布卢姆与坎南 1995 年[17] 的「会检查自身工作的程序」，以及戈德瓦塞尔等人 2015 年[19] 的「给麻瓜的交互式证明」（interactive proofs for muggles），也都属于这一脉络。这是「借来的判断」在数学上最纯粹的形态：即便神谕不可信，只要盘问得法，仍然能够从它那里得到可靠的答案。

贯穿这些例子的共同观念是：引入一个外部判断者，以获得单凭自身所不具备的可靠性。这一对策的典型失效方式也同样直接：神谕不可靠，或者有所偏倚。请来的裁判自己就可能判错；而「谁来验证神谕」这一问题，又会把人引入无穷回归。

冗余：从许多不可靠里合成可靠

第二种对策取相反的方向：不引入一个可信的判断者，而是召集许多并不可信的判断者，转而信任他们之间的一致。

这一对策的理论有两块基石。冯 · 诺依曼 1956 年[4] 证明，以各自都会出错的元件为材料，通过层层叠加冗余，可以构造出可靠性任意高的计算。孔多塞 1785 年[5] 的陪审团定理（Condorcet's jury theorem）则给出了这一机制的算术依据：若每个判断者都略优于随机猜测（正确率 $p > \frac{1}{2}$），并且彼此独立，那么随着人数的增加，

多数票正确的概率将趋于必然，

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

这一对策的应用遍及许多领域。在分布式系统中，它表现为拜占庭容错（Byzantine fault tolerance）。皮斯、肖斯塔克与兰波特 1980 年[8] 以及兰波特等人 1982 年[9] 提出的拜占庭将军问题（Byzantine generals problem），所探讨的是：当部分节点可能蓄意作恶（对抗）时，系统如何达成共识，即各节点如何就同一结果取得一致。经典的门槛是，节点数须满足 $n \geq 3f + 1$，才能容忍 f 个叛徒。在工程上，卡斯特罗与利斯科夫 1999 年[11] 的 PBFT 把它落实为实用系统；在理论上，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。

在机器学习中，与之对应的是集成（ensemble）。汉森与萨拉蒙 1990 年[20] 和迪特里希 2000 年[22] 所讨论的集成方法，以及布雷曼 2001 年[23] 提出的随机森林（random forest），做法都是让一群弱模型投票，其结果胜过单个强模型。在人类群体中，它表现为「群体的智慧」（wisdom of crowds；索罗维基 2004[25]）。1906 年，高尔顿在一个乡村集市上记录了约八百位村民对一头公牛体重的独立估计。没有一个人猜中，但所有估计的平均值为 1197 磅，而公牛的实际重量是 1198 磅：就整体而言，群体的判断几乎分毫不差。洪与佩奇 2004 年[24] 更进一步证明，在适当的条件下，一群多样的普通解题者能够胜过一群能力出众的解题者。至于科学中的同行评审与重复实验（第 3 章）、工程中的 RAID 与法定人数机制、医疗中的第二诊疗意见，也都是这一对策的具体形式。

然而，这一对策依赖于一个值得仔细审视的关键前提：独立性。只有当各个判断的失误互不相关时，冗余才能成立。将许多相互独立的估计加以平均，方差才会随人数的增加而下降，

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

而一旦这些判断之间存在正相关 ρ，方差便不再趋于零，而是停留

在一个由相关性决定的下限上，本书称之为相关性地板，

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

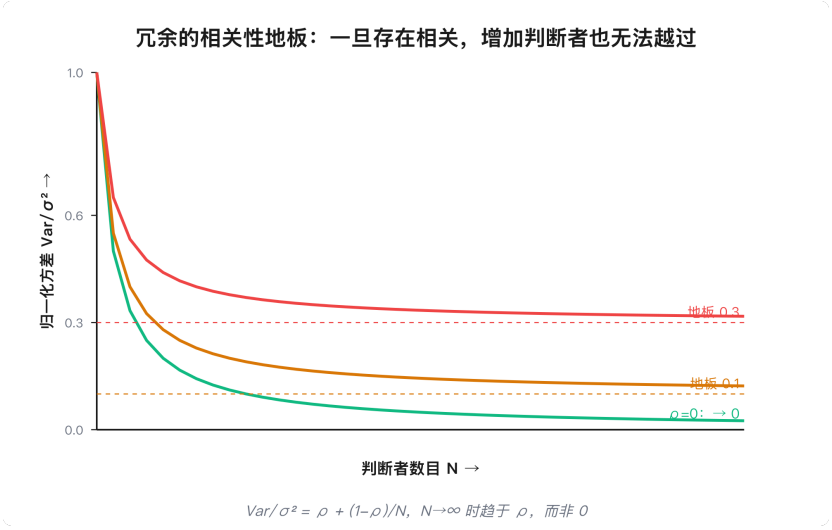


图 7: 冗余的相关性地板：一旦存在相关，增加判断者也无法越过

相关性由此击穿了冗余最核心的承诺：无论增加多少判断者，都无法越过相关性设下的这道地板。这并非空谈。在奈特与莱韦森 1986 年[13] 的著名实验中，许多程序员依据同一份规格说明，各自独立地编写程序。按照预期，他们的错误应当互不相关；实际结果却是，他们在相同的地方出错，因为人在面对同一个难点时会犯同样的错误。对此，埃克哈特与李 1985 年[12] 早已从理论上作出了预言。群体思维（groupthink）、来源相同且带有缺陷的训练数据、共模故障（common-mode failure），都是这道地板的具体表现。这便是冗余的典型失效方式：以为独立，其实相关。

两种对策的关联

把两种对策放在一起考察：前者引入一个单一而昂贵的神谕，后者汇合众多廉价而独立的判断；二者借来的，都是行动者单凭自身所

不具备的判断力。它们共同的作用，是补足所缺的验证能力。二者的失效方式也恰好相对：单一的神谕可能出错，众多的判断则可能暗中相关。

数学史上有一段插曲，把这两种对策连同上一章的证书联系在了一起。阿佩尔与哈肯 1977 年[26] 对四色定理（four color theorem）的证明依赖计算机穷举，因而饱受争议：接受这一证明，无异于要求数学界信任一个神谕。后来，贡蒂耶 2008 年[27] 以机器可以核对的形式证明（formal proof）重做了这一证明，黑尔斯团队[28] 也以同样的方式处理了开普勒猜想。两者都把「信任神谕」转化为「核对证书」。麦肯齐[29] 在《机械化证明》中所追踪的，就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] 则主张，证明在本质上是一种社会过程；归根结底，这一主张是把数学的可信性安放在人类判断的冗余之上。

不过，有一点需要看清：到此为止，本章与上一章的对策（「压缩未知」与「借来的判断」）所追求的仍是同一样东西，即对象之真。它们仍然试图知道一件事究竟是对是错。下一章的两种对策则做了一件更为彻底的事：它们不再强求这种真。

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. doi:10.1007/bfb0060623 [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). doi:10.1007/3-540-09724-4 [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). doi:10.1007/978-3-662-07964-5 [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. doi:10.1515/9781400882618-003 [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」这一对策的奠基石之一，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. Google Books [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。它为本章冗余这一对策提供了算术骨

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. doi:10.1137/0218012 [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. doi:10.1145/22145.22192 [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. doi:10.1145/322186.322188 [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 $n \geq 3f + 1$。它是本章冗余这一对策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. doi:10.1145/357172.357176 [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. doi:10.1145/3149.214121 [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。它为本章冗余这一对策划出了边界，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. 链接 [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。它是本章冗余这一对策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. doi:10.1109/tse.1985.231895 [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. doi:10.1109/tse.1986.6312924 [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. doi:10.1145/359104.359106 [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，读者应读它关于证明本质上是社会过程的论证。

15. A. Shamir (1992).「IP = PSPACE」. Journal of the ACM, 39(4), pp. 869-877. doi:10.1145/146585.146609 [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. doi:10.1145/146585.146605 [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。它揭示了「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，可作本章相关脉络的延伸阅读，读者可循此深入这套手法。
17. M. Blum, S. Kannan (1995).「Designing Programs That Check Their Work」. Journal of the ACM, 42(1), pp. 269-291. doi:10.1145/200836.200880 [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. doi:10.1145/278298.278306 [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，是「弱验证者如何高效核对庞大证明」这一脉的理论顶点，可作本章的延伸阅读，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. doi:10.1145/2699436 [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. doi:10.1109/34.58871 [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。它是本章冗余这一对策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. 链接 [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。它给出了机器学习版的「相关性地板」精确公式，可与本章相互印证，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. doi:10.1007/3-540-45014-9_1 [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这一对策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. doi:10.1023/a:1010933404324 [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。它是冗余这一对策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. doi:10.1073/pnas.0403723101 [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. Google Books [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。它把本章冗余这一对策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. doi:10.1215/ijm/1256049011 [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 Theorem」. Notices of the American Mathematical Society, 55(11), pp. 1382-1393. 链接 [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. doi:10.1017/fmp.2017.1 [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. doi:10.7551/mitpress/4529.001.0001 [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**：不必执着于验证真正的对象。可以将它替换为一个既可检验、又可求解的代理（代理替换，proxy substitution）；也可以不再要求非此即彼的判决，转而依据经过校准的概率行动（校准，calibration）。

前两章的对策，追求的仍是关于对象的真相：或者削减关于它的不确定性，或者借助他人的判断去逼近它。本章讨论的两种对策，即代理替换与校准，则放弃了这一追求。它们不再追问「真正的对象是否正确」，而是改为回答另一个问题：前者更换验证的对象，后者改变判决的形式。

代理替换：更换验证的对象

第一种对策的一般形式是：不再以那个无法测量的真实目标为验证对象，而是代之以一个可以检验、且足以满足需要的代理，进而对这个代理加以验证和优化。换言之，接受检验的已不是原来的对象，而是它的替身。

这一对策的种种形态，在前面几乎每一章中都已出现。数学家以等价陈述代替待证的定理（第 7 章）；软件工程师以测试代替正确性，以基准测试代替能力；组织以 KPI 代替自身的健康状况，以 GDP 代替福祉（第 8 章）；机器学习以奖励模型代替人的真实偏好（即第 5 章讨论的 RLHF）。心理学中也有与之对应的现象。卡尼曼与弗

雷德里克在 2002 年[32] 提出了「属性替换」（attribute substitution）的概念：人在作直觉判断时，往往不自觉地以一个易于评估的属性，顶替那个难以评估的目标属性。在方法论层面，这一对策亦有其原型：波利亚建议解题者先转向一个相关而较易的问题[30]，西蒙则提出了满意化（satisficing）[31]。

这一对策的要害在于，它可能以两种方向相反的方式失效。第 7 章与第 8 章的讨论在此交汇，这条线索也贯穿全书。不妨用两个维度来刻画一个代理：一是「忠实」，即代理是否确实指向原来的目标；二是「更易」，即代理是否确实比原问题更易处理。以二者为坐标轴，可得下图：

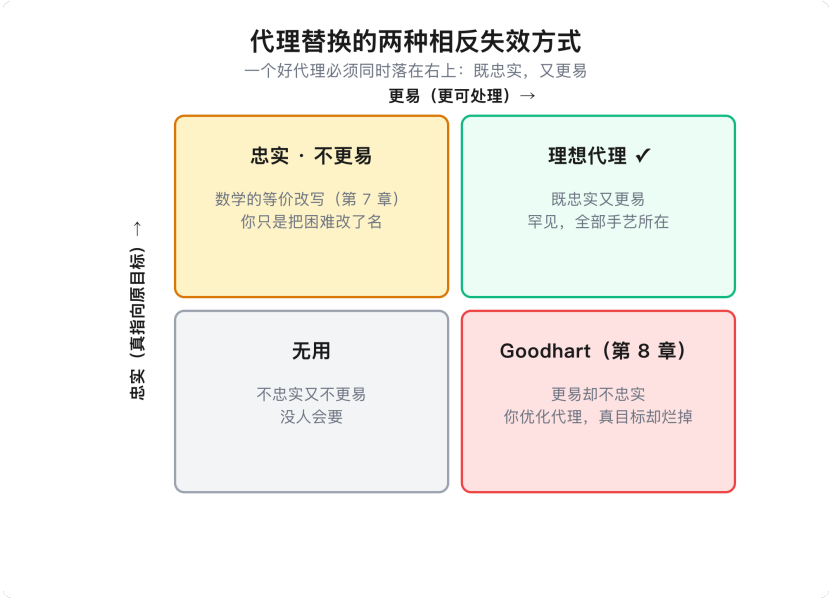


图 8: 代理替换的两种相反失效方式：忠实 × 更易

数学家的失败落在左上角：等价改写在忠实性上无可挑剔，却丝毫没有降低求解的难度，只是让同一个困难换了一副面目。组织的失败落在右下角：指标易于测量，然而一旦被当作目标加以优化，它与真实目标之间的对应便会断裂。

优化何以导致断裂？原因在于，代理与真实目标之间的相关，只在现状所对应的分布上成立；而优化的压力会把系统推离这一分布，

推向二者分道扬镳的极端。社会科学对这一机制早有表述。古德哈特 1975 年[1] 提出的定律指出，一个指标一旦成为目标，便会失效；坎贝尔 1979 年[3] 的定律表达了同样的观察；卢卡斯 1976 年[6] 在经济学中提出的孪生命题则是：一种结构关系一旦被当作政策目标，便会瓦解。三者描述的是同一个机制。埃斯佩兰与索德尔提出的反应性（reactivity）概念[7] 又推进了一步：指标不仅会失真，还会反过来重塑被测量的对象。

同一机制在机器学习中再度上演，而且表现得格外清晰。阿莫迪等人 2016 年[10] 将奖励黑客（reward hacking，即智能体借奖励设定中的缺陷取得高分，却并未完成预期的任务）列为一个具体的研究问题。此后的工作从经验与理论两个方向加深了对它的认识。潘等人 2022 年[12] 的实证研究发现，能力更强的智能体更善于利用代理奖励的漏洞，真实回报甚至会出现骤降式的相变；斯卡尔塞等人 2022 年[13] 证明，非平凡的奖励几乎不可能做到「无漏洞可钻」；高等人 2023 年[14] 则为这种过度优化（overoptimization）测得了定量的缩放定律（scaling law）。由此可见，一个好的代理必须同时避开上述两种失效：既要忠实，又要更易。能够兼顾二者的代理少之又少；找到这样一个代理，便是这门技艺的全部。

校准：改变判决的形式

第二种对策的一般形式是：验证的对象不变，改变的是判决的形式。它不再要求「真」或「假」的二值裁决，而是给出一个经过校准的概率，依据这一概率行动，并承担一个有界的风险。

所谓校准，是指声称的把握与实际发生的频率相符：声称有几成把握的事情，就应当确有几成发生。用形式语言表述，即

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

也就是说，凡是声称有 70% 把握的事情，长期来看应当有七成成真。这一要求弱于「判定对错」，却是可以达到、也可以检验的。

这一对策在各领域中的形态同样整齐。数论中的概率素性检验便是

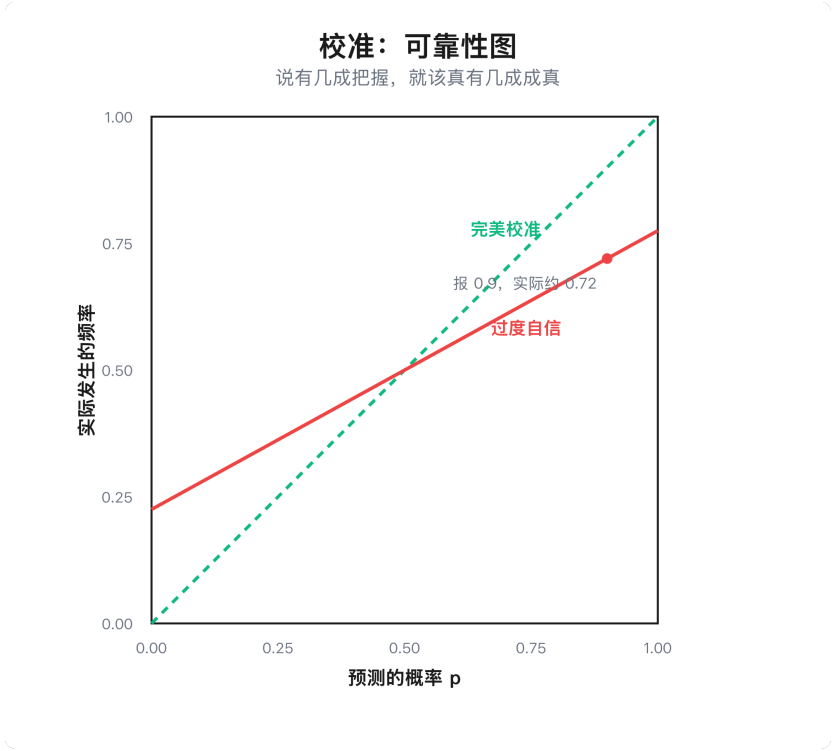


图 9: 校准的可靠性图：声称的把握应当与实际发生的频率一致

一例：第 7 章已经看到，它给出的结论是某数「以 $1-\varepsilon$ 的概率为素数」。在机器学习中，与之对应的是共形预测（conformal prediction，沃夫克、伽默曼与谢弗 2005[26]）。这种方法不给出单点的判断，而是给出一个带有覆盖保证的预测集，

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

在气象学与预测科学中，校准已发展为一整套成熟的理论。布莱尔在 1950 年[15] 提出了评价概率预报的评分；此后，墨菲在 1973 年[17] 将这一评分分解为可靠性、分辨率与不确定性三个部分，德格鲁特与芬伯格在 1983 年[18] 对此作了系统的处理；格奈廷等人在 2007 年[24] 提出了现代的框架，其要旨可以概括为「受制于校准，越锐越好」，即先要求预测是校准的，再在此前提下使它尽可能锐利，也就是尽可能集中、明确。

天气预报是校准做得最好的领域之一。一个成熟的预报系统宣称「明天降雨概率 70%」，若把它作出这一预报的日子全部汇总，其中确有约七成下了雨。声称的把握与实际的频率严丝合缝，堪称校准的范本。此外，预测理论中还有一项深刻的设计，称为严格适当评分规则（strictly proper scoring rule）：通过精心构造评分函数，让如实报告（报出自己真正相信的概率）恰好成为期望得分最高的选择，

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

于是，讲真话不再依赖报告者的自觉，而是由评分规则的数学结构强制保证（萨维奇 1971[16]、格奈廷与拉夫特里 2007[23]）。关于校准能否实现，理论上既有肯定的结果，也有否定的结果。达维德 1982 年[19] 证明，贝叶斯主体能够渐近地自我校准；福斯特与沃赫拉 1998 年[22] 证明，对任意序列都存在渐近校准的策略。然而，奥克斯 1985 年[20] 的结果「自我校准的先验不存在」，为这一对策标出了极限。在实践中，现代神经网络常常失准（miscalibration，郭等人 2017[25]），因此需要重新校准。至于第 6 章所述「允许、询问、阻止」的分级

信任，则是校准落实到行动层面的形态。

校准的失效也有两种。较浅的一种是失准：声称的把握与实际不符，例如声称有 90% 把握的事情只有六成成真，那么建立在这一数字之上的决策都会随之偏离。较深的一种更为隐蔽，也更为重要：校准给出赔率，却不回答是否应当接受这场赌局。一个完美校准的「70%」，对于「70% 是否足以下注」这一问题不置一词。答案取决于赌注的大小，也取决于行动者看重什么；这是一个价值问题，而不是验证问题。将二者混为一谈，是依据校准行动时最常见的误区。人们以为概率已经替自己作出了决定，而它所做的只是标出赔率；是否下注，仍须行动者自己拿出一套价值判断。

两种对策的关联

将两种对策并置，便能看清它们的共同之处。代理替换更换的是验证的对象，转向一个不同的、可以检验的替代目标；校准改变的是判决的形式，以概率取代真或假。二者都不直接回答原来的问题，而是将它转换为一个可以处理的问题。换言之，它们改变的是我们对「答案」的要求：前者改变度量的内容，后者改变判决的形态，并为残余的风险标定价格。

尽管如此，这两种对策仍然致力于把事情做对，只是降低了「对」的标准。最后两种对策则更进一步：它们不再寄望于做对，而是转向管控出错的后果。既然错误无法杜绝，就应当缩小它的代价，并确保它一旦发生便能被发现。这是第 12 章的主题。

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. doi:10.2307/2084615 [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. doi:10.1016/0149-7189(79)90048-x [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. doi:10.2307/255378 [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. doi:10.1002/(sici)1234-981x(199707)5:3<305::aid-euro184>3.0.co;2-4 [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. doi:10.1016/s0167-2231(76)80003-6 [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. doi:10.1086/517897 [②④] 埃斯佩兰与索德尔提出「反应性」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. doi:10.23943/9781400889433 [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：奖励黑客与过度优化

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，其中包括奖励黑客（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). arXiv:1706.03741 [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. arXiv:2201.03544 [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). doi:10.52202/068431-0687 [②] 这篇论文给出奖励黑客的一个形式化定义，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. arXiv:2210.10760 [②] 作者对奖励模型的过度优化做了定量刻画，给出真实表现随对代理奖励优化程度变化的缩放定律式规律：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，是本章过度优化论证最实证的一块。

校准：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. doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2 [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。它是校准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. doi:10.1080/01621459.1971.10482346 [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. doi:10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2 [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准、哪里有区分力。这一分解是校准概念的量化骨架，本章谈「校准」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. doi:10.2307/2987588 [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，明确区分了校准（calibration）与精炼/锐度（refinement），并给出据此排序预测者的框架。它是本章校准论证的核心理论来源，读者可在此看到校准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. doi:10.1080/01621459.1982.10477856 [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校准，即长期看其概率断言与实际频率相符。本章用它说明校准并非外加的苛求，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. doi:10.1080/01621459.1985.10478117 [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校准，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「校准有其极限」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. doi:10.1214/aos/1176347398 [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。它对校准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. doi:10.1093/biomet/85.2.379 [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校准。这是校准可达性的关键定理，本章据此说明校准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. doi:10.1198/016214506000001437 [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。它是本章校准论证的理论支柱，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. doi:10.1111/j.1467-9868.2007.00587.x [②] 作者提出概率预测的现代框架，把目标概括为「受制于校准，越锐越好」：先要求预测校准，再在校准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. arXiv:1706.04599 [②] 作者发现现代深度神经网络虽然往往更准，却常常失准，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。它是校准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准，于是需要重新校准」。

26. V. Vovk, A. Gammerman & G. Shafer (2005).《Algorithmic Learning in a Random World》. Springer. doi:10.1007/b106715 [②] 这是共形预测（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

作为校准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. Google Books [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「按校准的概率行动、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. Google Books [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校准检验的判断，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. doi:10.1080/01621459.1976.10480949 [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. doi:10.1515/9781400828678 [②④] 波利亚总结了一套解题启发法，其中一条便是先转向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. doi:10.1037/h0042769 [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. doi:10.1017/cbo9780511808098.004 [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，说明换问题这一对策不只是工程策略，也是人类直觉的默认运作方式。

第 12 章　管住后果

> **论点**：既然错误无法杜绝，就转而管控它的后果：限制一个有错而又未经验证的对象所能造成的破坏（限损，事前），并确保一旦出错，必能被发现（留痕，事后）。

前三章的对策，无论把「对」的标准放得多低，终究仍在努力把事情做对。本章讨论的最后两种对策则承认无法保证做对，于是转而着眼于失败的后果。既然错误无法杜绝，就在事前缩小它所能造成的破坏（限损），并确保它一旦发生，事后能够被追查（留痕）。

限损：缩小爆炸半径

第一种对策的一般形式是：不再试图保证那个未经验证的对象不出错，而是在事前就把它出错时所能波及的范围严格限定下来。工程上把这一范围称为爆炸半径（blast radius）。

这一对策构成了计算机安全最深厚的根基。早在 1966 年，丹尼斯与范霍恩就提出了能力（capability）机制[7]；1973 年，兰普森提出了围堵（confinement）问题[2]；1975 年，萨尔策与施罗德提出了最小权限原则（principle of least privilege）[1]；1976 年，丹宁又提出了信息流格（information flow lattice）[5]。贝尔-拉帕杜拉[3] 与比巴[4] 的安全模型，同样属于这一脉络。这些工作的主旨是一致的：只赋予一个组件完成其本职工作所必需的最小能力，此外一概不予；如此，即便它被攻破或者出错，也无法造成大范围的破坏。沙箱（sandbox，戈德堡等人 1996）[9]、职责分离与纵深防御，都是这一思想的具体形式。

在计算机安全之外，同样的思想以不同的名称出现。系统可靠性工程中的熔断器与隔板（尼加德的《Release It!》[15]）、爆炸半径设计、金丝雀发布与错误预算（error budget，谷歌 SRE[30]），金融中的头寸限额与止损，都是限损的不同形态；塔勒布所说的反脆弱（antifragility）[17]，关注的也是如何为下行风险设定上限。

这些做法的共同思路，是把重心从「使它不出错」（这需要我们并不具备的验证能力）转向「使它即便出错也能承受」。纵深防御（defense in depth）为此提供了一个常见的量化直觉：如果 k 层防护各自独立地以概率 p 失守，那么全部同时失守的概率为

$$p^k,$$

它随层数的增加呈指数下降。然而，这里必须立即补上第 10 章的警告：p^k 只在各层的失效相互独立时才成立。倘若各层都受制于同一个弱点（同一个被绕过的内核，同一个管理员口令），相关性便会使纵深防御骤然退化为单层防护。2011 年的福岛核事故是这一道理的写照。电站原本具有多重冗余，除主电源之外，还配有备用柴油发电机，但一场海啸将它们一并淹没。本应相互独立的几道防线因同一个原因而同时失效，纵深防御也就形同虚设。

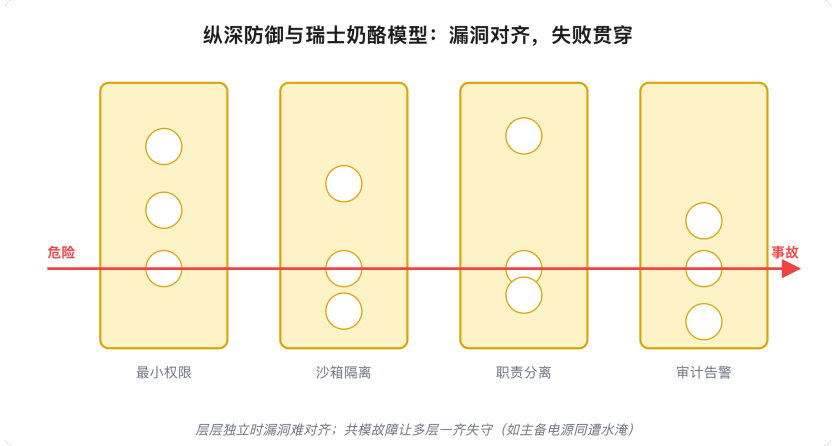


图 10: 纵深防御与瑞士奶酪模型：漏洞对齐，失败贯穿

这一对策的标准失效方式正在于此：限损被绕过。沙箱可能出现逃逸，权限可能在不知不觉中蔓延；表面上层层设防，各层却共用着同一道暗门。另一种较少被提及的失效方式是防护过度：它连正常的功能也一并阻断，使用者只好绕开防护完成工作，安全措施反而名存实亡。

留痕：让错误事后现形

第二种对策的一般形式是：对于无法预防的错误，确保它一旦发生便必定被发现。换言之，把检查从事前移到事后。

这一对策最坚实的技术基础来自密码学。默克尔 1980 年提出的哈希树（Merkle tree）[18]，以及哈伯与斯托尔奈塔 1991 年提出的链式时间戳[19]，使一份记录一旦写下便无法被暗中篡改，任何改动都会在校验时暴露。后来的工作又从不同方面加固了这套做法：施奈尔与凯尔西 1998 年设计了在不可信机器上保护日志的方案[20]，贝拉雷与迈纳 1999 年提出了前向安全签名（forward-secure signature）[22]，克罗斯比与瓦拉赫 2009 年则给出了防篡改日志（tamper-evident log）[23]。证书透明度（Certificate Transparency，RFC 6962）[24] 与中本聪 2008 年的比特币[27]，本质上都是一本全球范围、只能追加、人人可验的审计账。要核对一条记录是否在这样一棵树中，代价只有 $O(\log n)$。这再一次得益于第 7 章所说的那种不对称：验证比构造便宜。

不过，这一对策的历史远比密码学悠久。复式记账（double-entry bookkeeping）便是人类最早的防篡改账本之一。索尔在《账簿与权力》[29] 中论证，一个国家能否算清自己的账目，与它的兴衰直接相关。现代的财务审计与独立稽核，延续的也是同一思路。在科学中，与之对应的是预注册（preregistration，诺塞克 2018）[33] 与可复现（呼应第 3 章讨论的可重复性危机）：研究者在看到数据之前就登记假说与方法，事后便无法随意移动靶子。

这些做法的共同思路，是放弃「事前阻止坏事」（这需要验证），转而保证「事后必能发现坏事」（这只需要一本忠实的账）。其益处有两层：其一，错误可以得到纠正；其二，由于人人知道作恶终将无

所遁形，这本账也就具有威慑力。

留痕的标准失效方式也只有一种，却极为常见：错误被检测到了，却无人响应。无人阅读的审计日志，一概被忽略的告警，与不存在无异。2017 年的 Equifax 数据泄露是一个典型案例。一个已知漏洞迟迟未打补丁，入侵者在系统中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由此外泄。痕迹都留在日志之中，只是无人查看。只有检测而没有响应，检测便徒具形式。（另一个隐患是日志自身也可能遭到篡改，上文所述的密码学手段要防范的就是这一漏洞。）

八种对策的全貌：第三部结语

把最后这两种对策放在一起考察：限损在事前缩小失败的代价，留痕在事后保证失败会被发现。二者都不再试图使那个未经验证的对象变得正确，而是改变失败的形态：前者缩小爆炸半径，后者把检查移到事后。

至此，八种对策已全部论及。它们两两成对，共分四对：

- **压缩未知**（第 9 章）：证书 / 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、校准。
- **管住后果**（第 12 章）：限损 / 围栏、留痕 / 可审计。

这就是那张对照表，也是全书最主要的贡献。在第二部考察的四个具体场景中，以及在科学中，这些对策以不同的术语反复出现，却始终不出这八种。依照第 4 章确立的原则，本书对每一种对策都尽可能说明了它的机制、它在各领域中的形态及其标准的失效方式，而不仅仅依据表面上的相似。

然而，一个尖锐的问题仍悬而未决：为什么是这八种对策，而不是别的？这份清单是我拼凑而成的，还是每一种对策都对应着某种更基本、无法回避的东西？如果它只是一份清单，本书至多是一部有用的分类手册；唯有背后确实存在某种结构，「收敛」才算得到了解释。第四部将追问这一问题：先尝试把八种对策安置在一个共同的

骨架之上，再正面回答，这种收敛究竟是一条定律，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. doi:10.1109/proc.1975.9939 [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「限损」这一对策的源头：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. doi:10.1145/362375.362389 [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，正是本章限损这一对策要解决的核心。
3. D. Bell & L. LaPadula (1973). 《Secure Computer Systems: Mathematical Foundations》. The MITRE Corporation. 链接 [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. Google Books [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. doi:10.1145/360051.360056 [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. doi:10.1109/sp.1987.10001 [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「限损」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. doi:10.1145/365230.365252 [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. 链接 [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. 链接 [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。本章把沙箱列为限损这一对策的化身，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. Google 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. doi:10.7551/mitpress/8179.001.0001 [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. doi:10.1017/cbo9781139062367 [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. Google Books [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. doi:10.1109/tdsc.2004.2 [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限损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. Google Books [②④] 尼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，是把限损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. doi:10.1002/9781119644682 [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. Google Books [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。本章引它说明限损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. doi:10.1109/sp.1980.10006 [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」这一对策的技术基石，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. doi:10.1007/bf00196791 [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. 链接 [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，是本章留痕这一对策的核心技术之一。
21. B. Schneier & J. Kelsey (1999).「Secure Audit Logs to Support Computer Forensics」. ACM Transactions on Informa-

tion and System Security. doi:10.1145/317087.317089 [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. doi:10.1007/3-540-48405-1_28 [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. 链接 [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，是留痕这一对策走向工程化的代表作。
24. B. Laurie, A. Langley & E. Kasper (2013).《RFC 6962: Certificate Transparency》. IETF. doi:10.17487/RFC6962 [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. doi:10.1145/357172.357176 [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）. 链接 [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. 链接 [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. doi:10.1145/1349026.1349043 [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. Google 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. Google Books [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限损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. doi:10.1371/journal.pmed.0020124 [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. doi:10.1126/science.aac4716 [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「可重复性危机」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. doi:10.1073/pnas.1708274114 [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**：八种对策并非随意凑成的清单。若将风险与信息加以分解，便可看到八种对策各自作用于一根不同的杠杆；这套对策之所以给人以「完备」之感，原因即在于此。

第三部给出了那张对照表：八种对策分为四对，在四个具体场景和科学之中，以不同的术语反复出现。然而，一份清单无论多么整齐，终究只是清单。本章要追问的是：为什么恰好是这八种对策？它们是我拼凑出来的，还是每一种对策都占据着一个无可回避的位置？若是后者，收敛才算得到了解释；否则，本书充其量只是一部便于使用的分类手册。

我想提出一个候选的解释。有一点须事先申明：它只是一个组织框架，并不构成证明。读完本章之后，请再以第 14 章将要提出的种种怀疑来检验它。

风险的粗略分解

把「在无法验证时行动」还原到最简单的形式，行动者所管理的，归根结底是风险。借用决策论（decision theory）的经典表述（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

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

而这一切都受制于一个信息预算 B：可用于削减不确定性的检查、样本、算力与时间，无一不是有限的。

这个式子形式简单，要点在于：其右侧可供干预之处屈指可数。可以改变「失败」的定义，可以改变失败的概率，也可以改变对这一概率的认识；此外，还可以改变失败的代价、信息预算的分配方式，或者检查发生的时点。每一处这样的环节，都可以视为一根杠杆。我的主张是：八种对策恰好各占其中一个位置，并不存在有待填补的第九个空位。

八种对策与八个位置

下表将每一种对策与其所作用的杠杆逐一对应：

对策	所作用的杠杆	在风险分解中的位置
代理替换（proxy substitution）	改变所度量和优化的目标	重新界定「失败」
证书（certificate）/ 界	在可检验的局部，将不确定性限定在有保证的界之内	在局部使 Pr(fail) 趋近于零
神谕入回路	引入主体自身所不具备的验证能力	借助外部判断降低 Pr(fail)
冗余 / 共识	使多个判断的失败彼此不相关	降低联合失败概率 Pr(all fail)
最优筛查	将信息预算投向边际收益最高之处	分配 B，以求最大限度地削减不确定性
校准（calibration）	为残余风险给出如实的定价	使 Pr(fail) 成为已知量，从而可以据此下注
限损 / 围栏	缩小爆炸半径	降低 Cost(fail)
留痕 / 可审计	将检查从事前移至事后	改变检查的时点，使不可恢复的失败转为可恢复的失败

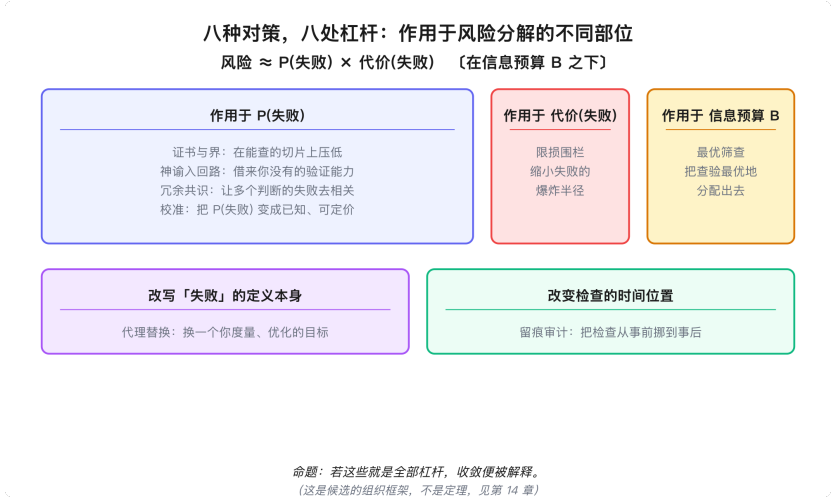


图 11: 八种对策映射到风险分解的不同部位（候选的组织框架，非定理）

看过这张表，「这不过是一份拼凑出来的清单」这一印象或许会有所动摇。八种对策并非随手拾取的八件工具。它们分别作用于 Pr、对 Pr 的认识、代价、预算分配、检查时点与目标定义，几乎将分解式中可供干预之处逐一占满。由此，本章的主张可以表述如下：倘若这些确实就是全部的杠杆，这套对策便是完备的，收敛也就得到了解释。任何善于应对的主体迟早都会重新发现这八种对策，因为除此之外，再无其他杠杆可用。

人、机器与组织何以殊途同归

如果上述说法成立，它还附带解释了本书开篇提出的那个谜题：为何数学家、工程师与组织会不约而同地采用这八种对策。

马尔[17] 在研究视觉时区分了三个分析层次：计算层（computational level）关注所要解决的是什么问题、受到哪些约束；算法层（algorithmic level）关注采用何种表征与过程；实现层（implementational level）关注这一切在何种硬件上实现。八种对策属于计算层。它们所回答的问题是：在无法验证这一约束之下，逻辑上还有哪几处可以施力？这个答案并不取决于行动者是碳基的数学家、硅基的程序，

还是由人构成的官僚机构。载体千差万别，计算层的约束却是同一个，应对因而趋于收敛。西蒙[12]的有限理性（bounded rationality）及其「人工科学」[13]（the sciences of the artificial）所描述的，正是这一类行为：塑造它的是环境的约束，而不是主体的内部构造。

在此还须提及无免费午餐定理（no free lunch theorem，沃尔珀特与麦克里迪[25]）。该定理表明：在所有可能的问题上取平均，没有任何一种方法优于另一种。这一结论具有两面性。一方面，它支持本书的克制：既然不存在万能的解法，杠杆的选择就必须依据问题的具体结构，这也是五种处境必须分别对待的原因。另一方面，它也是一种警示：任何宣称已经「找到统一钥匙」的人，包括我自己在内，都应当少一分自负。进一步说，当连失败概率都无法确定时，这些杠杆还会发展出相应的稳健形式。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），以及奈特式不确定性（Knightian uncertainty）下的决策理论，都属于这一类：即便 Pr 本身具有模糊性（ambiguity），它们仍然按照最坏的情形来设防。

一个强主张及其限度

下面把前文申明的保留意见完整地展开。

上述构想只是一个候选的组织框架，而不是定理。其中的风险分解是非形式化的：我没有给出关于主体与环境的严格模型，因而也无法证明最优策略恰好对应这八根杠杆。「这些就是全部的杠杆」只是一个断言，尚非已经确立的结果。我没有证据表明这张表是穷尽的，也无法排除它只是一个事后搭建的框架，即一种足够灵活、足以容纳许多不同对策组合的说法。表中有些归类甚至彼此重叠。例如，冗余既降低联合失败概率，又近似于一种特殊的筛查。仅此一点，就足以说明这一分解尚未做到各项界限分明。

我之所以将它写在这里，是因为它能够组织材料，也具有解释上的吸引力，而不是因为它已经得到证明。它符合一个好猜想的标准：清晰，可以反驳，并能统一解释大量现象。但它尚未达到定理的地位。

这样一来，最后一个问题便无法回避：这种跨领域的收敛，究竟是

由某种力量迫成的一条定律，还是一个虽然很强、却终究停留在经验层面的模式？下一章将正面讨论这一问题。

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. Google Books [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. doi:10.1214/aoms/1177732144 [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. Google Books [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. Google Books [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. Google Books [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. Google Books [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. Google Books [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. 链接 [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，是理解校准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. doi:10.1214/aoms/1177704255 [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 Ax-

ioms」. 《The Quarterly Journal of Economics》, 75(4), 643-669. doi:10.2307/1884324 [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. Google Books [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. doi:10.2307/1884852 [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，解释了为何不同载体会收敛到同一套应对。
13. H. A. Simon (1969). 《The Sciences of the Artificial》. MIT Press. Google Books [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。本章借它论证八种对策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. Google Books [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. doi:10.1126/science.185.4157.1124 [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，与校准、最优筛查等需要如实估计概率的对策形成对照。

16. D. Kahneman & A. Tversky (1979).「Prospect Theory: An Analysis of Decision under Risk」.《Econometrica》, 47(2), 263-291. doi:10.2307/1914185 [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. Google Books [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，主张八种对策活在计算层，因而能跨越碳基、硅基与组织等不同载体。
18. J. O. Berger (1985).《Statistical Decision Theory and Bayesian Analysis》. Springer-Verlag. doi:10.1007/978-1-4757-4286-2 [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. doi:10.1017/cbo9780511598951 [② ④] 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. doi:10.1016/0304-4068(89)90018-9 [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」的对策，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. doi:10.2307/1911053 [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. doi:10.1007/978-1-4899-3472-7 [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，是校准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. doi:10.1037/0033-295x.103.4.650 [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. doi:10.1162/neco.1996.8.7.1341 [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. doi:10.1109/4235.585893 [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. doi:10.1017/cbo9780511493539 [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. doi:10.1007/s102030200006 [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. doi:10.1111/j.1468-0262.2005.00640.x [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 Representation of Preferences」.《Econometrica》, 74(6), 1447-1498. doi:10.1111/j.1468-0262.2006.00716.x [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对策连成一个家族，价值在于呈现其共同骨架。
30. L. P. Hansen & T. J. Sargent (2008).《Robustness》. Princeton University Press. doi:10.1515/9781400829385 [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」这一做法的主要出处，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. doi:10.1017/cbo9780511779329 [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. doi:10.1017/cbo9781139060011.008 [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. doi:10.1016/j.tics.2017.09.005 [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**：这是全书必须正面回答的核心问题。跨领域的收敛究竟是一条定律（有某种力量迫使任何面对不可验证局面的有限主体走向这套对策），还是仅仅是一个很强的经验模式（我们反复观察到它，却无法证明它必然如此，而且选择效应或许就足以解释这种「押韵」）?

行文至此，本书必须对自己有所交代。本章要把贯穿全书的那个问题明确地提出来。

这种跨领域的收敛是一条定律吗？倘若如此，就意味着有某种力量迫使任何面对不可验证（unverifiable）局面的有限主体（bounded actor）必然走向这套对策。抑或它只是一个很强的经验模式？倘若如此，就意味着我们虽然反复观察到它，却未能证明它必然如此；而且，仅凭选择效应（selection effect），或许就足以解释这种「押韵」。

我将尽可能把双方的理由都陈述到最强的程度，包括不利于我的一方。如果本章有所偏向，也应当偏向怀疑。

支持「定律」的理由

第一条理由来自上一章的杠杆分解。如果八种对策确实占满了风险分解式中所有可供干预的位置，收敛就不是巧合，而是结构使然：

任何善于应对的主体迟早都会重新发现它们，因为别无选择。这一论证一旦成立，分量极重。

第二条理由是独立的重复发现。默顿在科学社会学中研究过「多重发现」（multiple discovery）现象：同一个想法，常常由彼此不知情的人几乎同时各自得出[12]。微积分有牛顿与莱布尼茨，自然选择有达尔文与华莱士；贝尔与格雷甚至是在同一天带着电话走进专利局的。此类例子在科学史上屡见不鲜，多到难以归于巧合。本书的八种对策同样在密码学、统计学、数论、组织理论与安全工程中一再被重新发明，而这些领域在当时并无多少往来。一种事物若总是被人各自独立地发现，它就更像是必然的产物，而不像是相互借用的结果。

第三条理由来自科学内部确凿的先例。物理学中的重整化群（renormalization group）与普适类（universality class）表明，结构迥异的系统在临界点附近会收敛到完全相同的行为；而在那里，这确实是一条定理[17,18]。有一个具体而引人注目的例子：液体在临界点附近的行为与磁体在居里点附近的行为，由同一组「临界指数」（critical exponents）刻画。分子与磁矩毫不相干，却落入了同一个普适类，原因在于决定临界行为的是对称性、维度这类粗粒度的特征，而不是微观细节。维姆萨特的稳健性分析（robustness analysis）认为，一个能够沿多条相互独立的路径反复得出的结论，更有可能为真[20]。与此相通，惠威尔 1840 年提出的「归纳的协同」（consilience of inductions）[11]，以及维格纳那篇论「数学不可思议的有效性」的文章[22]，讨论的都是跨领域的汇合所带来的可信度。收敛向来是科学判断某种东西「真实存在」的古老信号之一。

支持「模式」或更弱结论的理由

下面转向不利于我的一方。在我看来，这一方的分量并不比前一方轻。

最具破坏力的反驳是：这些领域并不像表面上那样彼此独立。它们共享同一个数学基础，概率论、最优化与信息论深植于每一个领域的根基之中；它们共享同一种人类认知，毕竟这些学科都出自同一

种大脑；它们还相互借鉴、彼此引用，从来不是封闭隔绝的：香农的信息论几乎渗入了所有领域，决策论的语言也扩散到各处。倘若收敛仅仅是因为各领域都取材于同一个数学工具箱，都受同一种心智的塑造，又始终在相互模仿，那么「独立的重复发现」这一理由就要大打折扣。所谓收敛，或许只是共同起源的回声。

第二，杠杆分解很可能是一个事后搭建的框架。它是否只是足够灵活，以至于能够容纳许多不同的对策组合？上一章的那张表已经暴露了这一点：冗余被归入降低联合失败概率一栏，却又近似于一种特殊的筛查。归类上的重叠说明，这把尺子还不够刚硬。丹尼特提出过一个关键问题：一个模式在什么条件下才算是真实的，而不是被强加的[13]？他给出的判据是预测力与压缩力：真实的模式能够让人预言新的东西。迄今为止，我的框架主要是在整理已知的对策，尚未预言出一种前所未见、后来又得到证实的新对策。这是它尚未通过的检验。

第三是选择效应。我是带着「寻找收敛」的眼光考察这些领域的，因此很可能在不知不觉间滤除了不相契合的领域与反例，只保留了彼此押韵的部分。可重复性危机（replication crisis）是最响亮的警示：看似极其稳健的经验模式，可能不过是系统性偏倚的产物[29,30]。我没有理由认为自己能够免于这种偏倚。

第四，即便收敛是真实的，它也未必指向一条深层的定律。劳丹的悲观元归纳（pessimistic meta-induction）提醒我们，历史上一个又一个成功的理论后来都被推翻了[6]。收敛于一个模式，并不等于收敛于真理。卡特赖特认为，基础定律并不如实地描述世界[14]；安德森的「多者异也」（more is different）则指出，更高的层次有其自身的规律[15]。本书所说的收敛，或许只是「特殊科学」（special sciences）尺度上的一种规律性，谈不上基本定律。沃勒尔的结构实在论（structural realism）提供了一种折中[7]：真正得以留存的，或许是那个结构（也就是这些杠杆），即便我为它所作的阐释是错误的。

定论的条件

要真正了结这一问题，需要我目前无法提供的东西：一个关于「有限主体加不可验证系统」的形式模型，以及一条定理，证明在该模型之下，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。与之最接近的一条真正的定理是无免费午餐（no free lunch）定理[24]，而它指向的恰是相反的方向：没有任何一种方法处处占优。在这样的模型与定理出现之前，这一问题仍然悬而未决，我也无意假装它已有定论。

因此，我应当如实说明本书所提供的是什么：一个猜想，一个有力、有用、边界清楚的猜想，以及一套能够贯通许多领域的共同词汇。它不是定理。这也是本书一路所倡导的姿态：给出一个经过校准的信念（calibrated belief），而不作非此即彼的裁决。

一本无法验证自身的书

至此，自序言起便潜伏在全书之下的那个问题浮出了水面：一本讨论「如何在无法验证时行动」的书，无法验证它自己的核心命题。

这听起来像是一种尴尬的自指（self-reference），却是本书最坦诚的时刻。面对一个连自己都无法验证的核心论断，它别无选择，只能践行它通篇所描述的做法。它陈述一个经过校准的信念（我认为这种收敛是真实的，但无法给出证明）。它划定主张的边界（这是猜想，而不是定理）。它公开接受反驳：只要找到一个不发生收敛的领域，或者一个打破这一分解的反例，那便是它的黑天鹅。然后，它照旧继续论述，照旧把这套词汇交到读者手中，因为这套词汇是有用的，即便尚未得到证实。

换言之，本书把它所讨论的对策用在了自己身上。它运用了代理替换，以「展示并组织收敛」替代「证明收敛」；运用了校准，为自己的把握程度定价；还运用了证伪式的留痕，把可以被推翻的断言白纸黑字地写下来。自指的系统无法在内部完全证成自身，对于这种处境，我们早该习以为常[26,27,25]。然而，无法在内部自证，并不意味着不能行动。倘若本书的论点成立，它「以自己所描述的方式写成自己」这一事实便不是缺陷，反而构成一份微弱却贴切的旁证。

于是只剩下最后一个问题。如果验证通常无法获得，连本书也只能给出一个未经证实的信念，那么，我们平日称为「知识」的大量东西，究竟是什么？下一章将转向认识论，全书的论证也将在那里收束。

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. Google Books [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. Google Books [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. doi:10.1017/cbo9781139171434.009 [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. Google 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. Google Books [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). doi:10.1086/288975 [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). doi:10.1111/j.1746-

8361.1989.tb00933.x [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. doi:10.1017/cbo9780511814563 [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). doi:10.2307/2181906 [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. Google Books [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经过校准的信念并据以行动。

11. W. Whewell (1840).《The Philosophy of the Inductive Sci-

ences, Founded Upon Their History》. John W. Parker. Google Books [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). 链接 [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). doi:10.2307/2027085 [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种前所未见、后被证实的新对策，是它仍待通过的考验。

14. N. Cartwright (1983). 《How the Laws of Physics Lie》. Oxford University Press. doi:10.1093/0198247044.001.0001 [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). doi:10.1126/science.177.4047.393 [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). 链接 [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). doi:10.1038/scientificamerican0879-158 [② ③] 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. doi:10.1093/0195146476.001.0001 [② ③] 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. doi:10.1515/9780691209418 [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. Google Books [③②] 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. doi:10.2307/j.ctv1pncnrh [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). doi:10.1002/cpa.3160130102 [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. Google Books [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). doi:10.1109/4235.585893 [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. Google 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. Google Books [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). doi:10.1007/bf02084159 [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. doi:10.1093/oso/9780195124415.001.0001 [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). doi:10.1371/journal.pmed.0020124 [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). doi:10.1126/science.aac4716 [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「可重复性危机」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**：全书的论证最终落脚于认识论（epistemology）。如果验证通常无法获得，那么我们称为知识的东西，大部分便是未经验证的知识。真正的能力不在于「知道自己是对的」，而在于行动得当，同时对自己可能以何种方式出错，保有校准良好的认识。

上一章留下了一个问题：如果验证通常无法获得，连本书也只能给出一个未经证实的信念，那么，我们平日称为「知识」的大量东西，究竟是什么？本章从认识论的角度回答这一问题。

重新界定「知道」

在哲学教科书中，知识被界定为「被证成的真信念」（justified true belief），最好还附有一份证明。对于有限主体而言，绝大多数要紧的事情都无法达到这一标准：或者不存在判定程序，或者代价高到无法承受，或者所需的状态隐而不见，或者时间不足，或者有对手在主动挫败检验。以此为标准，我们几乎什么都不能算作「知道」。

因此，需要一个适用于有限存在者的定义。在这一定义下，「知道」不再意味着「知道自己是对的」，而是意味着持有一个校准（calibration）良好的信念，并使自己可能犯错的方式处于控制之下。同样，能力也不等于确知真相，而是指行动得当，同时清醒地认识到自己可能以何种方式出错。一旦完成这一转换，前文的八种对策便不再只是一个工程工具箱，而成为一种认识论：一套「在神谕始终不会降临时，如何持有信念并据此行动」的方法。

科学早已如此

这并非什么新的发明。人类最严肃的求知建制，早已如此运作。如第 3 章所述，科学从不宣称自己证实了什么，而只说某一理论「至今尚未被证伪（falsification）」。就此而言，整个科学就是一台为「在无法验证时持有信念并据此行动」而优化的机器。哲学家对此也早有明确的论述。杜威 1929 年的著作《确定性的寻求》[10]，其书名即是一个诊断：人们把太多精力耗费在追求某种确定性上，而在行动的领域中，这种确定性并不存在；知识的真正功能在于引导行动，而不在于提供保障。詹姆斯的《信仰的意志》[9] 更进一步：有些事情要求人在证据齐备之前就表明立场；在这种处境下，选择相信并为之下注是正当的，而不是思想上的草率。波兰尼关于个人知识（personal knowledge）[8] 的论述则提醒我们，任何「知道」都包含一份超出可证范围的个人托付。综合来看，这些论述指向一种成熟的姿态：知识不是等待得来的确定性，而是付诸行动的校准信念。

作为认识论的八种对策

由此，可以将八种对策重读一遍：这一次不把它们视为工程手段，而视为求知的方式。

证书，意味着把验证廉价的那一小块彻底弄清，对其余部分则保持存疑。校准，意味着如实地持有分等级的信念，而不假装一切非黑即白。冗余，意味着从若干相互独立的视角出发，对一件无法直接看清的事物进行三角定位。代理替换，意味着借助一个易于处理的替身，去把握那个难以直接把握的真实目标，同时警惕替身在 Goodhart 崩塌之处背离目标。筛查，意味着把有限的注意力投向最能改变自己看法的地方。神谕入回路，意味着在自身判断力所不及之处，求助于更可靠的判断。限损与留痕，则构成「行动得当」的底线：在把信念付诸实施时，确保一旦出错，损失可以承受，错误可以发现，并且仍能纠正。综而言之，这便是一套可供有限存在者使用的认识论。

直觉、专长与判断

落实到具体的个人，技艺高超者究竟是如何做到的？这是落足点④（如何在无法验证的世界里生活）最切实的部分，也最容易被神化，或被一笔抹杀，因此需要表述得格外准确。

关于专家判断，心理学已经积累了大量证据，其中不少令人不安。米尔在 1954 年[5]、道斯在 1979 年[12] 先后发现，在许多领域中，简单的统计模型胜过专家的临床直觉。然而，另一脉研究提供了互补的图景。克莱因关于自然主义决策（naturalistic decision making）[17] 的研究，舍恩对「反思性实践者」（reflective practitioner）[13] 的刻画，德雷福斯兄弟[14] 的工作，以及艾利克森对刻意练习（deliberate practice）[15] 的研究，都从不同方面表明：在反馈充分、规律稳定的环境中，专家能够发展出可靠的直觉。归根结底，这种直觉是经由反馈打磨而成的、校准良好的模式识别（pattern recognition）。吉仁泽的快速节俭启发式（fast-and-frugal heuristics）[16] 进一步指出，简单规则之所以有效，是因为它们契合了环境的结构（生态理性，ecological rationality）。最为持平的综合，来自卡尼曼与克莱因 2009 年那场「未能达成分歧」的对话[24]：直觉是否值得信赖，取决于环境。在高效度、可供学习的环境中，直觉是可信的；在低效度、充满噪声的环境中，它就成了自欺。

这便是本书的认识论在个人层面上的体现。直觉既非魔法，亦非无用之物；它是一种校准的能力，而校准是可以训练的。泰特洛克主持的「良好判断计划」（Good Judgment Project）[27]，在一场由情报界举办的预测锦标赛中，遴选出一批被称为「超级预测者」（superforecasters）的普通人。他们既没有接触机密的权限，也不是领域专家，却凭借多角度取证、小步更新、严格复盘这些可以习得的习惯，据报道使预测准确度比能够接触机密情报的专业分析师高出约三成。承认这一点，也就意味着承认，刻意的无知（deliberate ignorance）有时是理性的[28]；同时也意味着承认，面对凯恩斯[3] 与奈特[1] 所说的那种根本不确定性（凯与金[29] 所谓「彻底的不确定性」，radical uncertainty），依照塔勒布[26] 的思路为稳健性与反脆弱性（antifragility）预作布局，往往比追求精确的预测更为明智。

在不确定中行动的尊严

将这些观察汇集起来，浮现出来的是一种姿态。它不是怀疑论者的瘫痪（既然一切都无法确定，那么一切都不算数，也就什么都不必做），也不是教条主义者的假装（把某个可以测量的数字奉为圭臬，仿佛它就是那个无法测量的真相）。它是第三条道路：清醒地知道自己之所不知，为这份无知标出刻度，然后依然行动得当。

这其中有一种沉静的尊严。承认验证是一种奢侈，并不意味着认输，而是认真对待行动的前提。好的判断者并不依靠确定性来支撑自己；他们所依靠的，是不夸大自己的把握，以及一套将这种诚实转化为行动的方法。

回到神谕

让我们回到开篇。古希腊人在出征之前赴德尔斐求取神谕；计算机科学也把那个能够即刻返回正确答案的黑箱称为神谕。两种神谕背后是同一个设想：行动之前，先把对错确定下来。本书所讨论的，正是这一设想落空之后的世界。而它最终的回答是：设想落空，并不意味着求知与行动就此终结，而只意味着它们必须以另一种方式进行。

对于有限存在者而言，「知道」从来不是「等到了证明」，而是「持有一个经过校准的信念，并将它付诸行动」。神谕始终沉默，但这从未使我们止步，也不应使我们止步。剩下的工作，是把这些论述落实到一个具体的人身上；那是跋的任务。

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. Google Books [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. Google Books [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「校准信念」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. doi:10.2307/1882087 [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. 链接 [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. doi:10.1037/11281-000 [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. doi:10.2307/1884852 [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. doi:10.1037/h0042769 [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. Google Books [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. Google Books [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. Google Books [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. doi:10.1126/science.185.4157.1124 [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. doi:10.1037/0003-066x.34.7.571 [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. Google 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. Google Books [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. doi:10.1037/0033-295x.100.3.363 [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. doi:10.1037/0033-295x.103.4.650 [②④] 两位作者展示，像「认出哪个就选哪个」这类快速节俭的简单启发式，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 De-

cisions》. MIT Press. Google Books [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. Google Books [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. doi:10.7551/mitpress/1654.001.0001 [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. Google Books [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，它说明校准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. Google Books [①②④] 泰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. Google Books [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. Google Books [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. doi:10.1037/a0016755 [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. Google Books [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. Google Books [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. Google Books [①④] 本书报告「良好判断计划」（Good Judgment Project）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。本章引此说明校准本身可以被训练，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. doi:10.1177/1745691616635594

[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. Google Books [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

本书始于一项关于结构的研究：五种处境，两两成对的八种对策，一张跨领域的对照表，以及一个连它自己也承认尚未证实的猜想。然而，它应当以一个人作结，因为这件事最终关乎的是你自己。

你的一生，都将把行动押在自己无法验证的事物上。你写下的代码，终将带着永远无法排查殆尽的缺陷发布。你所相信的理论，会在你穷尽证据之前就要求你投入。你所爱的人、共事的人、托付的制度，没有一样能在你下注之前得到验证。而你一生中最重要的那些选择，恰恰最无法验证。问题从来不在于你能否确定；你不能。问题在于，你能否依然行动得当。

本书的全部内容，归结起来，就是蕴含在那个「依然」之中的方法。在能够检验的那一小块上作出确切的证明，对其余部分则如实存疑。以若干相互独立的视角，对看不清的事物进行三角定位（triangulation）。借助一个堪用的替身，去把握那个难以把握的东西，同时警惕它从何时开始欺骗你。把有限的心力投向最能改变你判断的地方。自己无从判断之处，便求助于更可靠的人。下注之时，要让自己输得起，让错误能够被发现，也能够被纠正。这些并非取巧之术。一个有限的人，若要在没有神谕的世界里清醒地生活而不至于陷入瘫痪，所需的全部技艺尽在于此。

那艘船仍在雾中。船长终究没有等到一双能够看穿雾气的眼睛。海图会过时，罗盘会有偏差，对洋流的估算也始终只是估算。然而她还是改变了航向。她这样做，并非因为确信前方没有暗礁，而是因为保持航向同样是一场豪赌，也因为凡是能做的，她都已经做了：

核对了海图，校过了表，留出了余量，还预备了万一触礁时弃船的方案。然后，她转动了舵轮。

学会在没有把握时行动，归根结底，就是学会这样转动舵轮。雾不会散去；正因如此，你必须学会。

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》. 链接 [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》. 链接 [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》. 链接 [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 Sci-

ence Monthly» , 12, 1-15. 链接 [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. Google Books [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. doi:10.1037/10851-000 [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» . Google Books [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. Google Books [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. Google Books [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), 519-530. 链接 [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. Google Books [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), 99-118. doi:10.2307/1884852 [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. Google Books [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). Google Books [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. Google Books [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. Google Books [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. Google Books [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), 1124-1131. doi:10.1126/science.185.4157.1124 [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. Google 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. Google Books [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. Google Books [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. Google Books [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. Google Books [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. Google 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. Google Books [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. Google Books [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. Google Books [②④] 泰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. Google Books [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。