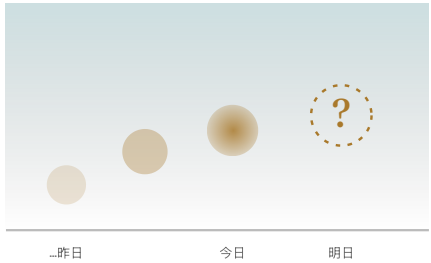
科学方法与划界

太阳四十五亿年来每日东升。那么明日依旧会升起——对吗？



● 每一次过往的日出都是证据——却证明不了下一次日出

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

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

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

地上的裂口

休谟抽去了地基

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

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

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

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

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

关于黑天鹅的注记

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

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

逃遁之路

波普尔的柔道：别再试图证明

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

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

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

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

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

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

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

公允以待弗洛伊德

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

复杂的现实

库恩：但科学并非那样运行

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

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

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

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

一个值得破除的迷思

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

修补

拉卡托斯：理论从不孤身赴死——以及杜恒–奎因的幽灵

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

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

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

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

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

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

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

重锤

费耶阿本德与「那」方法的死亡

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

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

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

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

复活

为何界线依然重要

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

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

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

划界标准表

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

前沿 · 2026

复现危机：划界在现实检验中

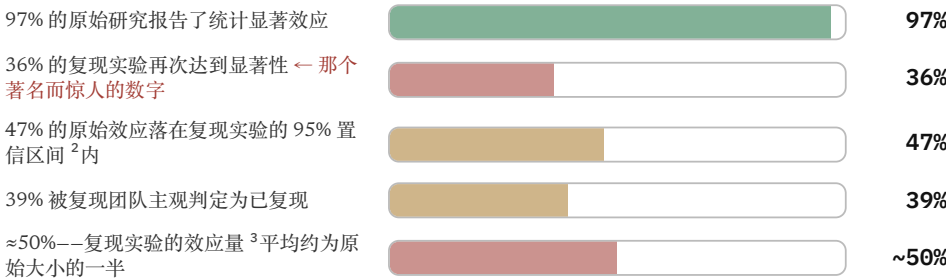
若有一条几乎人人认同的标准——波普尔、库恩、你的高中老师——那便是可复现。真正的结果，当别人照着程序再做一遍时，应当再次出现。它不是侥幸、捏造或风尚。于是在 2010 年代，科学家们做了一件显而易见、令人不安、却从未被系统做过的事：他们取来成堆的已发表、经同行评议、备受赞誉的发现，逐一尝试复现。

结果 01 [已确立] [争议/炒作]

震动心理学的一声枪响

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

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



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

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

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

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

结果 02 [已确立]

这并非一个领域的难堪

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

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

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

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

结果 03 [已确立] [争议/炒作]

那些烟消云散的发现——以及敢于承认的科学家们

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

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

转折 [线索]

这是科学的失败——还是科学在运作？

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

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

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

悬而未决的问题

何谓真正尚未落定

两千五百年过去，「何为科学？」这一问题的审慎回答仍有几条线没有系紧：

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

◆ 一日三句话

核心洞见

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

最佳类比

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

活的争议

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

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

说明

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

来源

来源与延伸阅读

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

可选附录

附录：没有基岩的地基

本节是可选的补充阅读；可以放心跳过，不会影响正文课程。

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

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

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

第一部分 · 深坑愈深

休谟自问自答——古德曼却雪上加霜

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

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

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

四种爬出深坑的尝试

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

斯特劳森

消解问题

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

赖欣巴赫

务实地下注

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

波普尔

否认前提

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

贝叶斯

量化更新

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

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

会变蓝的祖母绿：认识 "grue"

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

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

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

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

下方表格比较观察期、2050 年后的预测，以及古德曼关于可投射谓词的教训。

绿色 vs. 绿蓝，投影表

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

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

第二部分 · 波普尔承认的裂缝

近距离看证伪

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

第一：划界不是关于意义

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

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

第二：最大胆的理论恰恰最不可能为真——而这正是关键所在

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

第三：没有基岩——只有沼泽中的桩子

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

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

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

确证不是真理的首付

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

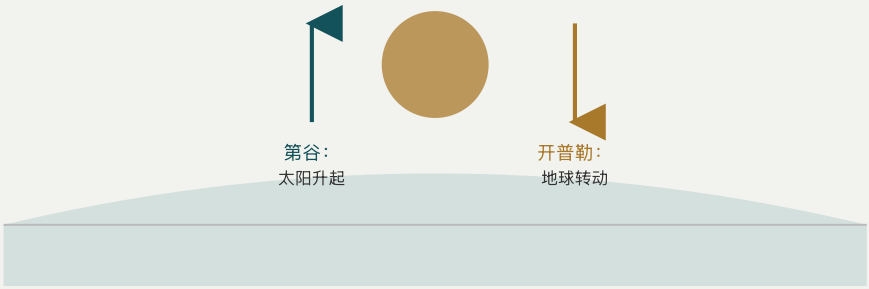
第三部分 · 缺失的中立地带

你们看到的甚至不是同一场日出

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

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

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



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

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

奎因抽出线头，整件毛衣跟着动

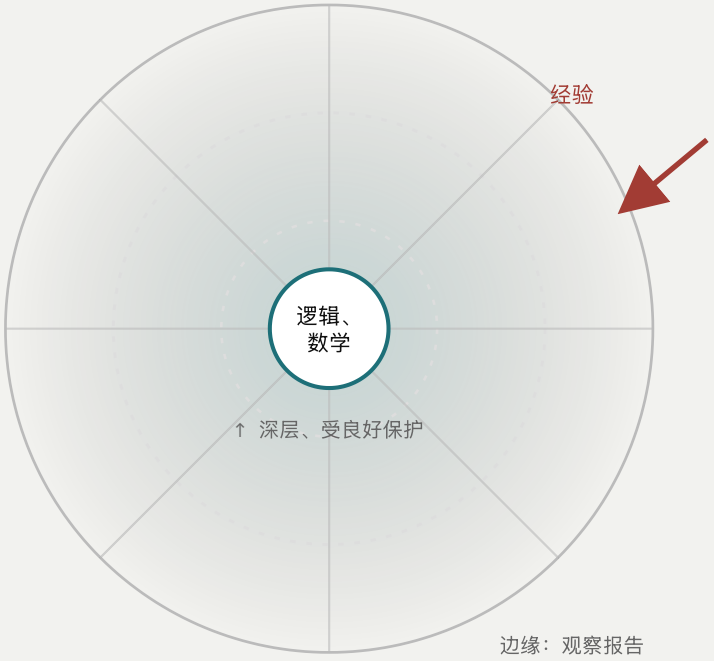
如果说单次观察负载理论，哲学家 W.V.O. 奎因在 1951 年进一步指出，单次检验也负载理论——并据此写成了现代哲学中极具影响力的论文《经验主义的两个教条》。我们在正文中见过它的产物（杜恒-奎因论题：没有假说是被单独检验的）。这里给出的则是它的母体思想，而且更激进。奎因把人类全部知识——从「这里有一只杯子」到逻辑法则——想象成一张巨大的信念之网：

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

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

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

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



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

劳丹踩下刹车：逻辑上可能 ≠ 理性上合理

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

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

第四部分 · 给弗洛伊德一个更公平的审判

格伦鲍姆：精神分析不是非科学——而是失败的科学

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

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

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

第五部分 · 危机的发动机房

为什么大多数研究结果被夸大了：真正的数学

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

齿轮一：基础率陷阱（伊奥安尼迪斯的炸弹）

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

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

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

下方表格列出三个基础率场景及其阳性预测值。

发现纯度引擎，基础率场景

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

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

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

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

齿轮二：灵活性——如何「找到」任何东西（披头士实验）

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

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

最令人不安的部分：你不需要作弊

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

齿轮三：过滤——文献是幸存者展厅

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

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

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

统计学家的判决 [已确立]

p 值不是什么

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

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

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

第六部分 · 定义了一个领域的决斗

伦敦，1965 年 7 月：科学哲学界的一场著名交锋

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

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

一本书里的二十一个范式

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

库恩有两个更深层想法值得从漫画式简化中抢救出来，因为它们都被惯常地夸大了：

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

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

贯穿线

没有底，但它照样运转

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

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

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

◆ 本附录三句话概括

核心洞见

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

最佳类比

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

活的争议

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

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

来源

来源与延伸阅读

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

明日 → 第 03 日

逻辑与有效推理

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

第 02 日终 · 还有 178 日等待深入