每 日 必 读 · 典 范 文 · 三 十 六

可预期的尺子

一个领域被评价体系覆盖得越久，它提出的问题就越像。三家成熟解释各有证据，却对同一批案例给出互相取消的预测——最粗糙的判断本该最坏却最好，最不可量化的裁量本该是解药却更保守。本文认为三家都漏掉了同一件事：那把尺子是公开的，被量的人能提前推断出它会怎么量。

王德生 + Claude · 约 1.9 万字 · 20 页 · 三种阅读方式 · 发表于2026年8月1日

同一笔钱，两种发法：一种要你先交一份写清「你打算做什么」的计划书，另一种不问你要做什么、只看你是谁。按「简化摧毁复杂性」的说法，后者粗糙到极点，本该最糟；实测却是后者产出的新东西更多。按「指标一旦成为目标就被操纵」的说法，把裁量权交还给同行专家应当是解药；实测却是专家评审密集的那一侧更保守，而且**没有任何人作弊**。本文认为三家解释都盯着尺子的一个属性——粗细、时机、期限——却把同一件事当成了不需要说明的背景：**这把尺子是公开的，被量的人能提前推断出它会怎么量**。由此得到一条判断：**选题的收敛不是简化的代价，也不是博弈的产物，而是判据可被提前推断的直接后果**；它与判据粗细无关、与合同期限无关，且不必等任何一次测量真的发生。机制只有三步——预演、筛选、回流——全程不需要假设任何人不诚实。文末给十处跨领域读数、两处反过来迫使它让步的地方、六条能推翻它的条件（第一条只要公开文本和一台电脑），以及一条写死到 2032 年的赌注。

引言

这篇文章要处理的现象，任何一个在学术体系里待过几年的人都见过，却很少被当成一个需要解释的对象：**同一个领域里被提出的问题，一年比一年像**。

它值得现在问，有两个理由。其一，过去十几年针对评价制度的改革，几乎全部集中在"换一把更好的尺子"上——换掉影响因子、加上开放数据、把裁量权还给同行。这些改革都有道理，也都做了，而收敛没有停。一个反复被改良却反复失效的领域，通常说明大家改的不是主变量。其二，一个此前只能靠印象谈论的量，如今有了廉价的读数：申请文本、获批清单、立项题目、开题记录，大量以文本形式公开，而文本的分布变化是可以直接算出来的。也就是说，这件事从"值得议论"变成了"可以被裁决"。

本文立的判断只有一条：**收敛的主变量是判据能否被提前推断，而不是它有多粗、什么时候量、给了多长期限**。

路线图如下：第二到第四章逐一交代三家已有解释各自说到了哪一步、又在哪里停下；第五章把它们的冲突写死；第六章换变量并给出三步机制；第七章加上第二根轴，得到一张四格表与一条可裁决的排序；第八章给四格各找一个真实标本；第九章说明三家为什么各自推不出这一条；第十章在十个别的领域里兑现；第十一章报告两处反过来迫使命题让步的地方；第十二章与最接近的七种说法逐一分界；第十三章交出不适用的范围；第十四章写清按它设计干预会怎样出事；第十五章给出证伪清单；第十六章处理本文自己也是标本这件事；末章是一条写死日期的赌注。

一、三条互相拆台的解释

先摆一个几乎所有人都同意的现象：**在一个评价体系里待得越久，一个领域里被提出的问题就越像**。

这件事不需要论证，只需要看。翻十年前某个学科的博士开题名单，再翻今年的，题目的分布会明显变窄；不是变精，是变像。同一批词反复出现，同一种问法反复出现，而那些当年还有人做、如今没人碰的方向，说不出是哪一年、被谁停掉的。

现象是共识，解释却有三条，而且这三条互相拆台。

第一条来自国家治理研究。 它说问题出在"看得见"这件事上：一个组织要管理一片它并不亲身在场的现实，就必须先让这片现实变得可读——统一度量、统一命名、统一格式。可读是管理的前提，而简化是可读的代价。被简化掉的那部分，恰恰是长在具体情境里、说不清楚、只能靠身体记住的那种知识。所以损失的根源是**粗糙**：尺子越粗，被它压掉的东西越多。

第二条来自社会测量研究。 它说问题出在"用来做决定"这件事上：一个量化指标越是被用于决策，它就越会腐蚀它本要测量的那个过程。人不是被动地被测量，人会朝着指标重新组织自己。所以损失的根源是**反身**：被测者会反过来改造自己以适应尺子，而这个改造发生在测量之后。

第三条来自劳动经济学。 它说前两条都没抓住要害，要害在合同：科研资助的常见形态是短周期、明确交付、失败即出局；而有一类资助反过来做——挑人不挑题、周期拉长、明确容忍失败。对照两类资助下同一批研究者的产出，后者的新颖度更高，同时低引用的"失败品"也更多。所以损失的根源是**期限与容错**：不是尺子的问题，是合同的问题。

三条都有证据，都不是稻草人。麻烦在于它们不能同时为真。

如果损失来自粗糙，那么第三条里那种"挑人不挑题"的做法应当是最糟糕的——它把一个研究者几年的复杂工作压缩成一个声誉判断，粗糙到了极点，却给出了相反的结果。如果损失来自反身博弈，那么第二条给出的处方——多用几把尺子、把最终裁量权交还给同行专家——应当有效；可第三条的对照恰恰显示,同行评议主导的那一侧更保守。而如果损失只跟期限和容错有关，那么第一条说的那些被压掉的东西，只要把期限拉长就该回来，可它们没有回来。

三条解释在同一个现象面前互相取消。本文认为这不是因为其中两条错了，而是因为三条都在同一件事上失手：它们各自盯住了尺子的一个属性——粗细、时机、期限——却都默认了一件不需要说明的事：**尺子是公开的，被量的人知道它长什么样**。

本文要提出的判断只有一句：

> 选题的收敛，不取决于判据有多粗，不取决于它在测量之后引发了多少博弈，也不取决于合同给了多长期限；它取决于**判据能不能被被评价者提前推断出来**。

后面的全部篇幅，用来把这句话做成一件可以被推翻的东西。

二、可读性那一条说到哪一步

第一条解释的完整版本，来自詹姆斯·斯科特 1998 年的《国家的视角》。

他的论证从一件很小的事开始：德国十八世纪的科学林业。国家要从森林里收税，就要知道森林里有多少木材；要知道有多少木材，就要把森林变成可清点的东西。于是原本混生的杂木林被改造成单一树种、等距成行、同龄同高的人工林——它在账本上是完美的，每一棵都可数、可折算、可预测。第一代长势极好，第二代开始出问题，土壤板结、虫害成片，德语里造出了一个专门的词来称呼这种病：森林之死。

斯科特把这个动作叫作**可读化**。国家不是恶意的，它只是必须站在远处管理它并不亲身在场的现实；而远处的管理必须依赖简化的表征。他用一个古希腊词称呼被简化掉的那部分知识——**梅提斯**，指那种长在具体情境里、随机应变、说不出规则、只能在实践中习得的本事。领航员进港的手感，老农看云的判断，接生婆的手。这些东西无法被写进任何一张表格，因为它们的有效性恰恰依赖于"这一次和上一次不一样"。

这条解释放到知识生产上，几乎是逐字对应的：一个学科要被管理，就要有可清点的产出；可清点的产出要有统一格式；统一格式一旦确立，那些不合格式的追问就成了账面上不存在的东西。斯科特的处方也随之而来——**留出不被读的空间**，抵抗简化，给地方性知识以生存余地。

这条解释解释力很强，但它有一个明确的边界，而这个边界正是本文要进入的地方。

斯科特的因果链里，损失的量与**简化的程度**成正比。尺子越粗、格式越硬、被压掉的维度越多，梅提斯死得越彻底。这条推论是他整本书的骨架，也正是它给出了一个可以被检验的排序：**判据越粗，收敛越强**。

而这个排序，在科研资助的现实里对不上。

美国霍华德·休斯医学研究所的资助方式，按斯科特的尺度衡量是极粗的：它不评项目，评人;不要求预先写清要做什么，只问这个人过去做过什么、看起来会往哪走。一个研究者几年的全部复杂性，被压缩成一次关于"这个人值不值得赌"的判断。这比任何一份逐项打分的项目申请书都粗糙得多。按斯科特的排序，它应当造成更严重的梅提斯之死。

实际发生的是相反的事。2011 年，皮埃尔·阿祖莱与他的两位合作者把这类研究者与一批同等资历、由美国国立卫生研究院常规资助的研究者做了对照。结果是：前者产出的高影响力论文更多，论文中出现的新关键词更多——这是新颖度的一个直接读数——同时，他们低引用的"失败品"也更多。粗糙的那一侧，反而长出了更不像别人的东西。

所以可读性这条解释说到了一件真的事：**表征必须简化，简化必有代价**。 它没说到的是：为什么同样极端的简化，在一处杀死了多样性，在另一处反而放大了多样性。

它缺的那个变量，不在尺子的粗细上。

三、反身性那一条说到哪一步

第二条解释有一个更早的源头。1976 年，唐纳德·坎贝尔写下一句后来被反复引用的话：一个量化的社会指标越是被用于社会决策，它就越容易被腐蚀，也越容易扭曲它本要监测的那个社会过程。同一时期，查尔斯·古德哈特在货币政策的语境里给出了更短的版本：当一个测度变成目标，它就不再是一个好测度。

这条解释后来被做得非常细。温迪·埃斯佩兰与迈克尔·索德在 2007 年发表的排名研究，追踪了美国法学院排名对法学院自身的改造：排名不是在描述法学院，它在制造法学院。招生策略、就业数据的统计口径、奖学金的分配方式、甚至院长的时间分配，全部朝着那几个被计入排名的量重新组织。他们把这个现象命名为**反身性**——公共测度会重造它所测量的社会世界。

这一支文献的贡献不该被低估。它证明了被测者不是被动的，它给出了大量可核的机制细节，它也解释了为什么"改进指标"这件事永远追不上被测者的适应速度。

它的处方同样清楚：既然单一指标必被腐蚀，那就不要只用一把尺子——多指标、代表作制度、把最终裁量权交还给理解内容的同行。2012 年的旧金山宣言与 2015 年的莱顿宣言，是这条处方最成体系的两个版本。

问题出在这个处方与第三条证据正面相抵。

按反身性的逻辑，同行评议应当是解药：它不可量化、不可累加、由懂行的人当场判断，因此最不容易被机械地博弈。可阿祖莱那组对照里，被同行评议密集覆盖的那一侧——五年一轮、逐项打分、明确交代要做什么的常规资助——恰恰是更保守的一侧。研究者在那一侧写出的东西更像已有的东西，新关键词更少。

更麻烦的是一个细节：那一侧的研究者并没有作弊。没有人伪造数据，没有人操纵引用，没有人做出任何可以被称为"腐蚀指标"的动作。他们只是**诚实地**写出了他们认为最可能通过评审的申请书。而那份申请书写完的那一刻，接下来五年要做什么就已经定了。

这就露出了反身性解释的边界。坎贝尔那句话里的动词是"腐蚀"，埃斯佩兰那条机制里的动词是"反应"——两者都预设了一个**先测量、后反应**的时序：尺子先量，人后调整。所以这一支文献的全部处方，都落在"让尺子更难被反应"上：换指标、加指标、把判断还给人。

可申请书那件事发生在测量之前。它不是对一次已发生的测量的反应，它是对一次**尚未发生的测量的预演**。研究者不需要等评审结果出来才知道该怎么写——他早就知道了。他知道，是因为评审的判据是公开的、稳定的、可以从往年获批名单里学出来的。

反身性解释说到了一件真的事：**被测者会朝尺子重组自己**。 它没说到的是：这个重组不必等测量发生。而一旦它可以发生在测量之前，"腐蚀"这个词就不够用了——被改变的不是上报的数字，是还没动笔的那个念头。

四、契约那一条说到哪一步

第三条解释来自一个更冷的传统：契约理论与实证劳动经济学。

它的理论骨架由古斯塔沃·曼索在 2011 年给出：如果一个委托人希望代理人去探索——去试那些大概率失败、但一旦成功回报极高的路径——那么最优的激励合同必须具备两个特征：**对早期失败的容忍**，以及**对长期成功的重赏**。短周期、按期交付、失败即出局的合同，会把代理人推向利用已知路径，而不是探索未知路径。这不是道德问题，是算术问题。

同年，阿祖莱、格拉夫·齐文与曼索把这个模型放到一个现成的自然对照上。美国有两套并行的生命科学资助体系：一套是国立卫生研究院的常规项目资助，短周期、明确交付、以项目为单位；另一套是霍华德·休斯医学研究所的研究者资助，长周期、明确容忍失败、以人为单位，且续约标准宽松。他们用倾向得分匹配挑出两组资历相当的研究者，比较其后的产出。

结论已经在上一章提过：被后一套资助的那一组，高影响力论文更多，论文中出现的新关键词更多，同时低引用论文也更多。用一句话概括——**容错换来了方差**。

这项工作的分量在于它不是思辨。它给出了一个可以被别人重做的设计，也给出了一个方向明确的读数。此后关于科研资助的政策讨论，几乎都要经过它。

但它把解释权全部押在了合同的两个参数上：**期限**与**容错度**。而这两个参数解释不了两件事。

第一件，是第二章末尾那个悖论：以人为单位的判断在斯科特的意义上极其粗糙，按可读性解释它该更糟。契约解释对此没有说法，因为"粗细"不在它的变量表里。

第二件更要命。如果损失只来自期限与容错，那么把期限拉长、把失败写进合同，损失就该消失。可现实里存在一类反例：**期限很长、容错很足，但判据高度明确的资助**。若干国家的大型基础研究计划正是这样设计的——十年周期，允许里程碑不达标，但立项时要求写清科学问题、路线图与可交付成果。这类计划并没有产生与研究者资助相当的新颖度提升；它们产出的东西，往往在立项书写完的那一天就已经定型。

期限给足了，容错也给了，可交付内容是提前写死的。研究者知道自己五年后要向谁交什么，也知道那个"什么"长什么样。于是他在写立项书的时候，就已经把所有他自己也没把握的念头筛掉了——不是因为怕失败，合同已经允许失败；是因为**那些念头写不进那份表格**。

契约解释说到了一件真的事：**激励的时间结构会改变探索与利用的配比**。 它没说到的是：即使时间结构完全正确，只要交付判据是提前可知的，收敛照样发生。

三条解释各自触到了真东西，也各自在同一个地方停下。下一章把它们的冲突写死，再换变量。

五、三条不能同时为真

把三条并排放着，它们对同一个案例给出的判断彼此矛盾。这一章不调和，只把矛盾摆清楚——因为正是这三对矛盾逼出了后面那个新变量。

第一对：粗糙的判断为什么反而更好。

可读性解释预测：判断越粗，被压掉的东西越多。以人为单位的资助是最粗的判断——它不看你要做什么，只看你是谁。按这条预测，它应当造成最严重的收敛。

契约解释预测：以人为单位的资助之所以好，是因为它的期限长、容错足。可期限与容错是合同参数，跟"粗不粗"无关；同样长期限、同样容错的项目制资助，效果并不相同。

两条解释在同一个案例上给出相反的因果归属，而且没有一个可以靠"程度不同"来调和：一个说粗糙是病因，一个说粗糙无关。要么其中一个错，要么两者都不是主变量。

第二对：专家判断为什么反而更保守。

反身性解释预测：把裁量权交还给同行专家，可以抵抗指标的机械腐蚀。这是它最主要的处方。

契约解释的对照给出的读数是相反的：由专家评审密集覆盖、逐项打分的那一侧，新颖度更低。

这一对同样不能调和。因为反身性解释所依赖的机制是"指标被博弈"，而在专家评审这一侧，恰恰没有博弈可言——评审是人，标准是文字，没有可以被刷的分数。可保守照样发生了。

第三对：损失发生在测量之前还是之后。

反身性解释的时序是明确的：先有测量，后有反应。它所有的机制细节都建立在这个时序上。

而申请书那件事发生在测量之前。写申请书的人不需要任何一次测量结果作为输入——他从往年的获批名单里就能学会。这意味着存在一条**不经过测量的损失路径**，而这条路径在反身性的框架里没有位置。

三对矛盾指向同一个空缺：三条解释各自盯着尺子的一个属性——粗细、时机、期限——却都默认了同一件事，且都没有把它写进变量表：**这把尺子是公开的，被量的人知道它长什么样，并且能据此提前推断出它会怎么量**。

一旦把这件默认的事拿出来当变量，三对矛盾同时松开：

- 以人为单位的判断之所以粗糙却不致命，是因为"这个人值不值得赌"**推不出来**——你没法从往年名单里学会怎么把自己变成那种人，至少学不到具体动作那一层。
- 专家评审之所以保守却无人作弊，是因为专家的偏好**推得出来**——它稳定、可学、有往年样本，诚实的人也能准确预演。
- 损失之所以可以发生在测量之前，是因为**推断本来就发生在测量之前**。

换句话说：三条解释各自解释的，都是同一个变量在不同取值下的表现，而那个变量不是粗细、不是时机、不是期限。

下一章给它一个名字，并说清楚怎么读数。

六、换一个变量：判据能不能被提前推断

本文提出的变量叫**判据可预期性**：一个被评价者，在动手做事之前，能在多大程度上准确推断出这次评价将依据什么、按什么权重、由谁在什么口径上作出。

三点澄清，缺一点这个变量就会被读回旧概念。

第一，它测的是推断，不是知情。 一份公开的评审细则可以写得很详尽，而被评价者仍然推不出结果——比如细则说"以学术贡献的实质为准"，或者评审名单是从一个大池子里随机抽的，或者权重每轮重新掷定。反过来，一份从不公开的标准也可以高度可预期——只要往年结果足够多、足够稳定，从名单里就能反推出来。**规则的公开度与判据的可预期性是两件事**，后者才是本文的变量。

第二，它是连续量，不是有无。 可预期性从来不会是零，也很少是一。它可以被测：给定一位申请者，让他在提交前对自己的结果做一次概率预测，把这些预测与实际结果比对，得到的校准度就是可预期性的一个下界估计。也可以从文本侧测：一条新判据公布之后，申请文本里对应词汇的词频跃升有多快、多同步。

第三，它是关于被评价者的属性，不是关于评价者的属性。 同一套制度，对老手是高可预期的，对新人是低可预期的；对身处中心的人是高可预期的，对边缘的人是低可预期的。这一点在第十四章会变成一条相当难看的反噬预言。

它怎么测——三条互相独立的路子。

第一条，**自评校准**。在提交截止之前，让申请人给出自己获批的主观概率；结果出来后，把这批预测与实际结果做校准曲线。校准得越好，可预期性越高。这条最直接，代价是需要机构配合收集预测。

第二条，**文本领先期**。一条新判据公布之后，申请文本里对应语言的词频跃升有多快、多同步；跃升与第一轮结果公布之间的时间差，就是可预期性的一个读数——**领先期越长，说明推断越不依赖真实反馈**。这条只要公开文本就能做。

第三条，**同群离散度**。同一轮申请里，不同申请人对"该怎么写"的判断有多一致。若一群互不通气的人交上来的东西高度同构，说明他们各自都推出了同一个答案——这本身就是高可预期性的证据。这条可以用现成的申请文本相似度矩阵直接算。

三条互相独立，且指向同一个量。若三条给出的排序彼此矛盾，说明这个变量没有被定义好，本文的操作化需要重做。

有了这个变量，本文的承重命题可以写成一句：

> **选题的收敛，不是简化的代价，也不是博弈的产物，而是判据可被提前推断这一件事的直接后果；它与判据的粗细无关，与合同的期限无关，也不必等到任何一次测量真的发生。**

机制只有三步，没有一步需要假设任何人不诚实：

第一步，预演。 一个理性的、诚实的、时间有限的人，在动手之前会先估计"我做出来的东西会被怎么看"。这不是投机，这是任何有限资源下的正常规划。

第二步，筛选。 预演的结果是一个筛子：那些他自己也说不清、写不进表格、无法向评审解释的念头，在这一步被筛掉。注意它们不是被否决的，是**没有被提出**——它们从未进入他向自己陈述的选项清单。

第三步，回流。 被筛剩下的东西做出来、被承认、成为下一轮的样本。而下一批人正是从这批样本里学习判据的。可预期性由此自我加强：样本越多，推断越准；推断越准，筛得越狠；筛得越狠，样本越同质。

三步走完，收敛已经完成，而全程没有出现任何一次"腐蚀"。这就是为什么在一个人人诚实、评审专业、指标多元、期限宽裕的体系里，选题照样会一年比一年像。

也正因为机制在预演那一步就闭合了，本文可以给出一条与三家都不同的推论：**改判据的内容不能解决问题，因为无论换成什么内容，只要它可被推断，预演照样发生。** 换尺子只是换了一个被预演的对象。

那么剩下的问题就变成：可预期性能被降低吗，代价是什么。这需要先把它和另一个已经被讨论了几十年的变量分开——评价发生在事前还是事后。

七、第二根轴：判据出现在动作的哪一侧

只有一根轴的东西不是维度，是名字。可预期性必须和另一根轴交叉，才能把现实切成可分辨的格子。

第二根轴是**时相**：判据是在动作之前就已就位（事前），还是在动作完成之后才施加（事后）。

这根轴之所以必须单列，是因为它常被误当成可预期性的同义词。它不是。

- 事前的判据**通常**可预期——申请书制度就是典型：你在动手前就要按一份已知的格式写清你要做什么。
- 但事前也可以不可预期——"我们只赌人，不问你要做什么"就是事前的、且推不出具体动作的判据。
- 事后的判据**通常**也可预期——发表与引用的评价标准稳定到可以当教材教。
- 但事后同样可以不可预期——一个从不公布、每轮重新抽取的评审池，或者一个在提交截止之后才被生成的测试集。

两根轴交叉，得到四格。这张表是本文的骨架，后面每一章都会回来对它取数。

	判据可预期（能从往年反推出具体动作）	**判据不可预期（推不出具体动作）**
事前（动手前判据已就位）	**格一 · 预演最强** 项目申请书制。研究在写申请书那天定型；期限与容错再宽也救不回来	**格二 · 赌人不赌题** 以研究者为单位的长周期资助。粗糙，但推不出"该做什么"
事后（做完才施加判据）	**格三 · 追认式趋同** 发表与引用体系。标准稳定可学，收敛慢一拍但持续累积	**格四 · 事后盲判** 随机抽选、抽签资助、提交截止后才生成的私有测试集

四格的排序，是本文最要紧的一条可裁决主张：

> **收敛强度：格一 > 格三 > 格二 ≈ 格四。**

这个排序与三家给出的排序都不一样，因此它是可裁决的：

- 可读性解释按"简化程度"排序，会把格二（最粗）排在最坏的一端；本文把它排在最好的一端。
- 反身性解释按"是否用于决策的量化测量"排序，会把格三排在最坏的一端；本文把格一排在更坏的位置。
- 契约解释按"期限与容错"排序，会把格一与格二之间的差别全部归给合同参数；本文预测**在期限与容错相同**的条件下，格一与格二仍有显著差距。

第三条尤其干净，因为它可以在同一家机构内部做：把一笔钱分成两个通道，期限一样、容错条款一样，唯一的差别是一个要求提交研究计划、另一个只审历史与人。若两个通道产出的选题分布在若干年后没有显著差异，本文这条命题就此作废。

需要立刻说清一件事，免得这张表被读成"越不可预期越好"：格二与格四虽然在收敛这一项上读数相近，代价却完全不同。格四把风险平摊给所有人，格二把风险集中在挑人的那几双眼睛上；格四不可问责，格二高度可问责。第十四章会正面处理这笔账。

八、四个格子的现实标本

抽象的格子要有具体的人和事，否则它只是一张漂亮的表。这一章给每一格找一个真的标本，并指出它在现实中长什么样。

格一 · 申请书制。 一位青年研究者要拿到一笔基础研究经费，需要提交的东西通常包括：科学问题、国内外研究现状、拟解决的关键科学问题、研究方案与技术路线、预期成果与考核指标。这份文件的每一栏都在要求他把一件尚未发生的事描述成一件已经想清楚的事。

真正起作用的不是任何一栏的严格程度，而是这份表格的**可学性**。往年获批清单是公开的，通过率是已知的，什么样的问法容易过、什么样的措辞容易被读成"不成熟"，在实验室里是口耳相传的常识。于是一个诚实的人也会做同一件事：在动笔前，把那些他自己最感兴趣、但一句话说不清的方向放到一边，去写那个他知道能过的。

这里必须说清一个容易滑过去的点：他并没有放弃那个方向，他只是**没有把它写进去**。而没有写进去的东西，在接下来五年的时间分配里，几乎必然排在最后。经费、学生、设备都跟着表格走。表格里没有的东西，事实上不存在。

格二 · 赌人不赌题。 上文提到的研究者资助是最成熟的形态：不问你要做什么，看你过去做过什么、怎么做的、遇到反常时如何反应；周期以人为单位计，续约标准明确写着容忍失败。

它的可预期性低在哪？低在**推不出具体动作**。一个人可以知道"他们喜欢有独立品味的人"，但这句话不能被翻译成任何一条可执行的写作策略。你没法通过写一份更聪明的申请书变成那种人，因为根本没有申请书。可以被模仿的是过去的成果，而成果是要花五年做出来的——**模仿的成本高到无法作为策略使用**，这就是低可预期性在实践中的真实含义。

格三 · 追认式趋同。 发表与引用体系是事后的：文章写完了才被审、才被引。它的判据同样极其可学——什么样的题目容易被送到什么样的审稿人手里、什么样的结果容易被接收、哪些期刊偏好哪类问法，这些是每个博士生入学第一年就在学的东西。

它比格一慢一拍，因为它不强迫你在动手前交出计划；但它持续累积，因为每一篇发出来的文章都在给下一批人提供更精确的样本。格三是慢性的，格一是急性的。

格四 · 事后盲判。 三个真实标本。

其一，抽签资助。这不是思想实验，是已经运行了十几年的制度。新西兰健康研究理事会自 2013 年起在其"探索者资助"里使用随机选择，是第一家这样做的大型政府资助机构；德国大众基金会的"实验！"计划自 2017 年起引入部分随机；奥地利科学基金会的"一千个想法"、瑞士国家科学基金会与英国国家学术院随后跟进。瑞士那家在 2021 年底决定，把随机作为其全部资助线上的并列裁决方式。

这些方案的共同形状值得看清：**先做严格的同行评议筛出"够格"的一批，然后在这一批里随机抽。** 它的效果不来自"随机更公平"——公平只是它被提出的理由之一——而来自**没有一条可以被优化的最后一公里**。你可以把申请书写到够格，你没法把它写到中签。

这里还有一个细节，恰好构成本文那条区分的现成证据：瑞士那家明确**告诉申请人可能会抽签**，理由是资助方必须对评审标准保持透明。也就是说，规则是完全公开的，而结果仍然不可预期。**公开度与可预期性在这里被现实亲手分开了**——这正是第六章第一点所主张的。

其二，机器学习的私有测试集。竞赛把一部分数据锁起来，参赛者只能看到公开榜的分数，最终名次按从未见过的那部分数据算。为什么必须这样做？因为一旦测试集可见，模型就会被朝着它调，而调出来的分数不再代表任何东西。

其三，随机抽查式审计。税务机关的随机抽查之所以有威慑力，不在于抽查更严格，而在于**没有人能算出自己这一年会不会被抽中**。

三个标本来自三个毫不相干的行当，做的却是同一件事：把判据从可推断的位置上挪开。它们不追求更好的判据，它们追求**判据不可被提前优化**。

九、三家各自为什么到不了这一条

一个新判断要站住，必须说清它不能从任何一个源里直接推出来。这一章逐个交代。

可读性那一家到不了，是因为它的因果变量是简化，而简化与可预期性可以正交。

斯科特的全部机制都建立在"表征必须丢弃信息"上。但一次极端简化的判断可以是高度不可预期的（赌人不赌题），一次极端细致的判断也可以是高度可预期的（逐项打分的评审细则）。四种组合都存在，且他的框架里没有位置容纳"细致但推不出来"这一格——因为在他那里，细致本身就意味着已经确立了一套稳定的分类，而稳定的分类天然可学。

更直接的一点：斯科特的处方是"留出不被读的空间"。本文的处方不是不被读，而是**读法不可提前推断**。这两者在实践中会导向完全相反的制度设计——他要求减少测量，本文允许高强度测量，只要求测量的具体口径不可被预演。

反身性那一家到不了，是因为它的机制以测量为起点。

坎贝尔的动词是腐蚀，埃斯佩兰的动词是反应，两者的因果箭头都从测量指向被测者。而本文的机制在测量之前就闭合了：预演不需要任何一次真实测量作为输入，它只需要一批可学的历史样本。

这一点可以做成一个判决性的差别。反身性预测：**在一个指标刚刚设立、还没有任何一轮结果公布之前，不应观察到显著的行为改变**——因为还没有东西可以反应。本文预测相反：**只要新判据的内容一经公布，申请文本里的对应语言就会跃升，且这个跃升发生在第一轮结果出来之前。**　　这是可以从公开文本里直接读出来的，第十五章会把它写成一条可执行的检验。

契约那一家到不了，是因为它的变量表里只有期限与容错。

曼索的模型是关于最优激励合同的，它处理的是"什么时候奖、失败罚不罚"。它没有"被代理人能否推断出考核内容"这一项——在标准的委托代理设定里，考核标准是共同知识，这是模型的**前提**而不是变量。把一个前提当成变量来推，模型自己推不出来。

阿祖莱那组对照之所以看起来支持契约解释，是因为在他们比较的那两套制度里，期限、容错、可预期性三者恰好同向。要把它们分开，需要一个把期限与容错固定住、只变可预期性的设计——那正是第七章末尾提出的双通道检验。

最后一句，关于三家共同的默认。

三家都默认了判据是共同知识，这不是疏忽，而是它们各自领域里的合理假设：治理研究关心的是国家如何看见，社会测量关心的是指标如何反作用，契约理论需要共同知识来定义均衡。本文所做的事，只是把这个在三个领域里都被当作背景的东西挪到前台，并且发现——**一旦它变成变量，三家互相矛盾的预测就同时被安顿了。**

十、它在别处也兑现

一个判断如果只在它出生的那个行当里成立，那它是一条经验规律，不是一个辨别维度。这一章把它放到十个别的领域里，每一个都要求它给出一件具体的、可以被查证的预测——不是"像"，是"在那个领域里也能据此预测出一件事"。

其一，临床试验的替代终点。 药物审批常常不等最终结局，而用一个更快出结果的中间指标代替：用血糖控制代替糖尿病并发症，用心律失常的抑制代替猝死，用肿瘤缩小代替生存期。

本文的预测是：一个替代终点一经监管接受，**后续研发管线会迅速向它对齐，且这种对齐先于任何关于该终点是否有效的证据更新**。这件事有过极其惨痛的验证：上世纪八十年代末，一类能有效抑制心梗后室性早搏的抗心律失常药被广泛使用，因为早搏抑制是一个漂亮的替代终点；后来的随机对照试验发现，用药组的死亡率显著高于安慰剂组。指标被优化了，人却死得更多。

请注意这与坎贝尔的区别：没有人作弊，药确实抑制了早搏，测量是诚实的。出问题的是**整个研发方向在指标可预期之后的对齐**。

其二，机器学习的基准。 一个公开基准数据集一旦成为领域共识，后续模型就会朝它优化，直到分数不再代表能力。应对办法不是造更严格的基准，而是把测试集藏起来、或者定期换新。

本文的预测：**同一个基准的公开时间越长，新模型在它上面的分数增长与在同分布新采样数据上的分数增长之间的差距越大。** 这件事已经被做过——有人重新采集了一批与经典图像数据集同分布的新样本，发现模型在新样本上的准确率普遍下降，且越是后期、在原榜上分数越高的模型，下降幅度越显著。

其三，风险投资。 早期投资的一句行话是"投人不投项目"。本文的解释与通常的解释不同：通常说法是早期项目的商业计划不可信，所以只能看人；本文说的是，**商业计划是可被优化的，而"这个人是谁"在短期内不可被优化**。预测：一家基金的评估清单越是被创业者摸清，它看到的项目就越同质——而这件事有一个便宜的读数，即同一基金历年被投项目的商业计划书文本相似度。

其四，法律里的规则与标准。 法经济学有一组经典区分：规则是事前明确的（时速六十公里），标准是事后裁量的（合理的速度）。规则可预期，因而合规成本低、但可以被精确规避；标准不可预期，因而抑制了规避设计，代价是当事人不知道自己是否合规。

本文与这组区分共享一根轴，但落点不同：那一支文献关心的是**遵从与社会成本**，本文关心的是**被规制者的问题集**。可裁决的差别：那一支预测规则化会提高遵从效率；本文预测规则化会使被规制方的**创新形态**向规则边界收敛——两者可以同时为真，但只有本文预测"合规行为的多样性下降先于合规率上升"。

其五，教育里的过程性评价。 把评价从一次期末考试改成对学习过程的持续记录，初衷是不让分数压扁学习。本文的预测很不受欢迎：**如果过程性评价的量表是提前公布且稳定的，那么被格式化的将不只是答案，而是学习动作本身**——学生会按量表的条目来安排自己怎么学。可读数：同一门课在引入公开量表前后，学生课外提问的主题分布熵。

其六，招聘与简历。 一家公司的用人标准一旦被外面摸清——什么学历、什么公司待过、什么项目经历算加分——求职者就会朝那个形状长自己的履历。预测：**同一家公司连续几年录用者的履历形态相似度，随其招聘标准公开程度上升**；而以"结构化面试题库固定"著称的公司，这个相似度应当高于题库每年重写的公司。这里可以顺手区分两件常被混淆的事：标准更客观与标准更可预期是两回事，前者提升公平，后者制造趋同，二者可以同时发生。

其七，内容平台与创作者。 推荐机制的规则一旦被创作者摸清——多长、什么开头、什么节奏、什么标题句式——内容形态会以肉眼可见的速度收敛。平台对此的应对，恰好就是本文的处方：不公布权重、频繁调整、加入随机曝光。预测：**一次未预告的推荐机制调整之后，新上传内容的形态离散度会短暂上升，随后随**

着创作者重新摸清规则而回落；回落所需的时间，就是这个判据的可预期性恢复期。这条有现成数据可查，而且时间尺度是周，不是年。

其八，信贷与风控。 一旦评分模型的特征被借款人推断出来，就会出现专门为提高分数而做的动作——为养流水而借的短贷、为凑年限而不销的旧卡。这些动作不改变还款能力，只改变分数。业内的应对同样是把部分特征保密、定期换模型。预测：**一个评分模型在市场上使用得越久，它对违约的预测力衰减得越快**，且衰减速度与该模型的特征被公开讨论的程度正相关。

其九，监管与合规。 一条监管要求越是被写成可核对的清单，被监管方就越会把资源投在"通过清单"上，而不是清单本要保护的那件事上。这里与法律那条不同的是，它给出了一个可测的错位：**合规投入的增长与实际风险指标的改善之间的脱钩程度**，应当随清单细化而扩大。反过来，采用随机现场检查、且检查项不预告的监管，这个脱钩应当更小。

其十，安全演习与红队。 一支防守队伍如果知道演习的脚本，演习就变成了排练。所以严肃的红队演练一律不预告时间、不预告手法，甚至不预告是否是演习。预测：**同一支队伍在"已知演习"与"未知演习"下的表现差距，随该组织演习脚本的稳定程度扩大**。这一条与前九条的差别在于，它是唯一一个当事人**普遍承认**不可预期是必要的领域——没有人会主张把红队的攻击手法提前发给防守方。这提示了一件事：在我们已经真正想清楚的地方，本文的处方早就是常识；只有在知识生产这里，我们仍然默认判据应当公开、稳定、可预期。

十处兑现，十个可查的读数，横跨医学、机器学习、投资、法律、教育、人事、内容分发、金融、监管与安全。但真正让这条判断长住的，不是这十处附和，而是下一章那两处**反过来给它加约束**的地方。

十一、两处反过来加约束的地方

只会附和的例证是装饰。真正说明一个判断有承重能力的，是它在某个领域里遇到了阻力，并且被迫让步。本文遇到两处，两处都逼出了对命题的修正。

第一处来自机器学习：不可预期性只能延缓，不能取消。

私有测试集的做法看起来是本文最干净的支持：把判据藏起来，优化就无从下手。但这个行当里还有另一件被反复观察到的事——**即使测试集从未公开，只要它被反复用于选择模型，它就会被逐渐泄露**。每一次提交、每一次看到榜上排名，都是在向那个隐藏的判据要一比特信息；提交次数足够多，参赛者就能在不看数据的情况下，把模型朝它调过去。这被称为适应性数据分析的过拟合，且有严格的理论刻画：泄露量随查询次数增长。

这一条逼本文让步两次。

其一，**可预期性不是可以被设为零的开关，它是一个会随交互次数单调上升的量**。任何被反复使用的判据，无论多不透明，都会逐步变得可预期。制度设计因此不能追求"造一个不可预期的判据"，只能追求"控制它变得可预期的速率"——定期更换判据、限制查询次数、在结果反馈中加噪声。

其二，**判据的载体越稳定，泄露越快**。同一批评审连任十年，比每轮重新抽取要可预期得多，哪怕两者用的是同一份细则。于是本文的实践含义从"改细则"移到了"改人员与流程的更新率"——这是一条我原本不会写出来的推论。

第二处来自科研资助的实证：以人为单位的判据也有它自己的判据。

赌人不赌题被本文当作低可预期性的标本。但那套制度并非没有考核——它有周期性的续约评审，只是周期长、标准宽。而且更要命的一点：**"什么样的人值得赌"这件事本身，也会随着样本积累而变得可学**。一批被

资助者的履历公开之后，年轻研究者完全可以据此规划自己的履历形态：去哪个实验室、发什么类型的文章、在什么年龄做什么动作。

这一条逼本文再让步一次，而且是最重要的一次：

> **可预期性不是关于"判据"的属性，而是关于"判据与被评价者的动作之间的距离"的属性。** 距离越短——判据直接对应一个可以在几个月内做出来的动作——可预期性的危害越大；距离越长——判据对应的是一段需要许多年才能积累出来的形态——同样的可预期性造成的收敛就慢得多。

这个修正把命题从一个二值判断改造成了一个带尺度的判断，也解释了为什么以人为单位的资助即使被摸清，收敛速度也远低于申请书制：**你可以在三个月内学会写一份像样的申请书，你没法在三个月内变成另一个人。**

这两处让步不是补丁。它们把本文从"降低可预期性"这个过于简单的处方，推到了两个更具体、也更可做的量上：**判据的更新率**，与**判据到动作的距离**。

十二、与最接近的几种说法划清界线

一个判断最危险的时刻，不是被反驳，是被读者顺手归到某个他早就知道的概念底下。这一章逐个处理最容易被混为一谈的说法——不是为了显示区别，而是因为每一处区别都必须落到"同一个案例，那个说法判 A，本文判非 A"上。做不到这一点的，就是承认被覆盖。

与"当一个测度变成目标就不再是好测度"。 那句话的机制是目标化引发的博弈；它预测在没有人博弈的地方不应出现问题。本文预测：在一个所有人都诚实、没有任何一次指标操纵的体系里，只要判据可被预演，收敛照样发生。可裁决的案例就是第十章那个抗心律失常药——药确实抑制了早搏，测量诚实，方向仍然错了。

与"公共测度会重造它所测量的世界"。 那一支的时序是先测量后反应。本文的时序在测量之前。判决性对照：一条新判据刚刚公布、第一轮结果尚未产生时，那一支预测行为无显著改变，本文预测申请文本的语言已经跃升。这条比对可以用公开文本做，成本很低。

与"可读性摧毁地方性知识"。 那一支按简化程度排序，把最粗的判断排在最坏一端。本文把最粗的那一格排在最好一端。同一个案例给出相反的排序，这是本文最容易被推翻的地方之一。

与"规则与标准"的法经济学区分。 共享一根轴，但落点不同：那一支问的是遵从成本与规避行为，本文问的是被规制者的问题集。可分开的读数：规则化之后，合规**行为的多样性**下降是否先于合规率上升——那一支不预测前者。

与"多任务代理"模型。 那个模型说：当代理人的多项任务里只有一部分可测量时，激励会把努力从不可测的那部分挤到可测的那部分。这是本文最近的邻居，因为两者都在讲"可测的挤掉不可测的"。分离线在两处：其一，那个模型的机制是**努力的配置**，因此它预测激励一撤、努力配置就回摆；本文预测的是问题集的收敛，且预测它在激励撤除后**不回摆**，因为被丢掉的不是意愿而是可想到的选项。其二，那个模型不区分可测性与可预期性——一件事完全可测但判据每轮重掷，按那个模型仍会被优化，按本文则不会。

与"最优探索合同需要容错与长期"。 分离线已在第九章给出：把期限与容错固定住，只变可预期性，本文预测仍有显著差异，那一支预测无差异。这是双通道检验的全部意义。

与"用抽签替代部分评审"的主张。 这是本文最像的同向邻居，必须说清。那一支的理由主要有三条：评审对边缘项目的判别力很弱、评审成本高、评审有系统性偏见——都是认识论与公平理由。本文的理由完全不同：**即使评审判别力完美、成本为零、毫无偏见，只要它的判据可被预演，收敛照样发生。** 判决性对照：把

评审质量提高（更长评审时间、更专业的评委、更透明的理由书），那一支预测抽签的必要性下降，本文预测**收敛不因评审质量提高而缓解**，甚至因为标准更一致、更可学而加剧。

与"制度化的怀疑"这一类学术规范论述。 那一支关心的是共同体对主张的相互质疑；本文关心的是提问的分布。二者不冲突，但一个不能替代另一个：一个所有人都在严格质疑彼此主张的共同体，完全可以同时在问同一批问题。

最后指认最近的那一位：**多任务代理模型**。它离得最近，因为它已经有了"可测挤掉不可测"的完整形式化。本文之所以仍不能被它吸收，只靠一件事——**它的变量是可测性，本文的变量是可预期性，而这两者可以取到相反的值**：一个不可测但高度可预期的判据（"我们只认一流的品味"，而这句话每年都指向同一批人），按那个模型不该产生挤出，按本文会。这个组合是否真实存在，是本文最该被攻击的地方。

十三、它不适用在哪

一个判断如果处处适用，多半是它什么也没说。这一章交出本文自己认为它管不到的地方。

其一，判据本身就是目的的领域。 有些活动的全部意义就在于达到一个提前公布的标准：驾照考试、飞行员的检查单、无菌操作的流程、桥梁的荷载规范。在这些地方，判据可预期不是缺陷而是全部价值所在，因为我们要的正是每个人都朝同一个标准收敛。本文关于收敛的负面判断，只适用于**我们希望产出彼此不同的东西**的场合。混淆这两类场合，会导出很危险的建议。

其二，样本极少的场合。 可预期性依赖可学的历史样本。一个刚设立的评审、一个每年只批三项的资助、一个从未公布过任何结果的机构，都不具备被学习的条件。在这些地方本文的机制不启动，收敛如果仍然发生，那多半是别的原因——很可能就是前面三家各自说的那些原因。这一点反过来也说明本文不是在取消三家，而是在给它们划出各自的适用区间。

其三，判据与动作之间距离极长的场合。 第十一章的第二处让步已经给出理由：如果判据对应的是一段需要十几年才能积累出来的形态，那么即使它完全可学，可预期性造成的收敛也会慢到难以观察。学术生涯尺度上的声誉判断接近这一档。本文在这一档的预测力很弱，且我不认为现有数据能把它与噪声分开。

其四，被评价者没有选择权的场合。 整套机制的第一步是预演，而预演的前提是这个人还能选择做什么。一个被分配任务、无权改变研究方向的人，不会经历那个筛选步骤——他的问题集不是被判据收窄的，是根本没有被交给他。在高度指令化的科研组织里，本文的机制让位给一个更简单的解释。

其五，可预期性极低但另有强约束的场合。 把判据做得完全不可预期，并不自动带来多样性。如果同时存在资源极度稀缺、失败即出局、或者单一的思想权威，那么人们会在完全猜不到判据的情况下，仍然朝同一个方向挤——因为别的方向连试的余地都没有。本文只主张可预期性是收敛的一个独立且重要的来源，不主张它是唯一来源，更不主张降低它就足以带来多样性。

其六，短期。 本文全部的读数都是年为单位的分布变化。任何试图用一两个季度的数据来检验它的做法，都会得到噪声。这既是它的谨慎之处，也是它的软肋：**一个只能在多年尺度上被检验的判断，很容易活得太久。** 第十六章会为此写一条带日期的赌注。

把这六条合起来，本文实际主张的范围要比它的标题窄得多：**在一个希望产出彼此不同的东西、样本足够多、被评价者有选择权、判据与动作距离不长的场合，判据可预期性是选题收敛的主要来源。** 超出这个范围的部分，我不认账。

十四、按它设计干预，会怎样出事

如果一条判断给不出"按它做会适得其反"的预言，那它多半只是一段描述，不是一个机制。这一章把本文最可能造成的伤害写在前面。

第一种伤害：不可预期性的代价不由所有人平摊。

降低可预期性，等于把结果的方差调大。而承受方差的能力在人群里分布极不均匀。一个有稳定编制、有积蓄、有导师托底的人，可以承受连续两轮落空；一个靠这笔钱付房租、签证挂在项目上、导师换了三个的人不能。所以一项以"提高多样性"为名的改革，很可能在实际效果上把边缘的人筛掉，留下的是**能承受不确定性的那一批**——而那一批在出身、机构、性别与国别上，恰恰是分布最集中的一批。

这条反噬相当难看，而且它有一个残酷的推论：**本文的处方在收敛这一项上有效的同时，可能在另一项上加剧同质化**。 我不认为这条可以靠"配套措施"轻易解决，它是这个方向的固有代价。

第二种伤害：不可预期与不可问责只隔一层纸。

一旦评价的口径不可提前推断，"为什么是他不是我"这个问题就失去了可核对的答案。而这正是任何一个封闭小圈子最想要的状态：以"我们的判据不宜公开，公开就会被优化"为理由，把裁量权变成不可质询的东西。**本文给这种做法提供了一套现成的说辞，这一点必须被明说。**

区分点只有一条，且必须写进任何实际制度里：**判据的可预期性可以低，判据的问责性必须高。** 具体做法是把两者在时间上错开——事前不公布具体口径，事后完整公开每一次判断的理由与全部结果，允许申诉。事后公开会让判据逐渐可学，因此必须配上更新率（第十一章的第一处让步）。这三件事——事前不公布、事后全公开、定期换口径——必须打包，拆开任何一件都会滑向暗箱。

第三种伤害：规划能力被摧毁。

有些工作确实需要多年稳定预期才能做：一台大科学装置、一项长期队列研究、一套需要十年才能建成的观测网络。这些工作不能建立在"下一轮判据会变"的假设上。把不可预期性铺到全部经费上，会先杀死这一类。

因此实践上的唯一负责任的形态是**分层**：把稳定可预期的通道留给需要长期承诺的工作，把低可预期的通道留给探索性的部分，并明确两条通道的比例与各自的问责方式。任何把整个体系一次性推向低可预期性的建议，本文都不支持。

第四种伤害：本文自己会被优化。

这条留到自反那一章说，因为它不是别人的问题。

十五、怎样算它错

这一节是本文最要紧的部分。前面所有的论证如果没有这一节，它就只是一个说得通的故事。

其一，最便宜、可以立刻做的一条。 取三到五家在最近十年内公布过一次明确评审新口径的资助机构，把公布前后各若干年的申请文本（或获批摘要，公开可得）做逐年的语义相似度追踪。

- 本文预测：新口径公布后的**第一个申请周期内**，对应语言的词频显著跃升，且这个跃升发生在第一轮资助结果公布之前。
- 反身性解释预测：结果公布之前不应出现显著变化。
- 若数据显示变化只出现在第一轮结果之后，本文的核心时序错，整篇判断需要退回。

这项检验只需要公开文本与一台笔记本电脑，没有任何一步需要机构配合。**本文没有做这项检验，因此它交出的是命题与方案，不是结论。**

其二，双通道检验（决定性，但需要一家机构配合）。 同一笔经费分成两个通道，期限相同、容错条款相同、评审委员会同一批人，唯一差别是：通道甲要求提交研究计划并按公开细则评分，通道乙只审历史与人、不问将要做什么。追踪五年内两个通道产出的选题主题分布熵。

- 本文预测：乙通道的分布熵显著高于甲通道。
- 契约解释预测：两者无显著差异（期限与容错已经对齐）。
- 若无显著差异，本文的核心命题被证伪。

其三，粗细对照（针对可读性解释）。 在同一机构内比较"极细判据"通道与"极粗判据"通道，且两者的可预期性都高（细则公开、往年样本充足）。

- 可读性解释预测：细的更好、粗的更糟。
- 本文预测：两者的收敛速度**无显著差异**，因为主变量不是粗细。
- 若粗细呈现单调关系，本文的主变量选错了。

其四，一条正向读数。 本文为真时，应当出现一件可测的事：在同一学科内，**评审委员会人员更新率越高的资助通道，其资助项目的主题分布熵越高**，且这一关系在控制经费规模、期限、学科后仍然成立。这是正向的——它要求本文预言的东西出现，而不只是要求反例不出现。

其五，一条会让本文尴尬的检验。 上一章第二处让步说，判据到动作的距离越长，同样的可预期性危害越小。那么在以人为单位的资助里，随着获资助者履历样本的累积，**年轻研究者的履历形态应当逐年趋同**——去同一批实验室、在同一年龄段做同样的动作。若追踪二十年发现履历形态并未趋同，说明"距离"这个修正项是我为了自圆其说加上去的，应当撤回。

其六，本文明确交出去的两个洞。

第一个洞：我说不清可预期性与"判据到动作的距离"这两个量之间的确切关系。它们显然不独立，但我没有一个能把二者分开测的设计。

第二个洞：我不知道低可预期性能维持多久。第十一章说泄露随交互次数单调上升，那么任何低可预期的制度都有半衰期——而这个半衰期有多长、由什么决定，本文完全没有答案。一个不知道自己解药有效期的处方，价值是打折的。

十六、本文自己也是标本

这一章不是谦辞。如果本文的判断成立，它首先应当适用于本文自己，以及本文得以写出来的那套做法。

第一件要认的账：本文有一个前身，而本文正在推翻它。

在此之前，我参与写过一份关于学术评价的稿子，它的结论是：既然现有改革（换指标、加指标、多指标）都只作用在已经能被评的东西上，那么解药是**在结算表里给"当前无人有能力评价的产出"留一个固定份额**，评审在这一位上不判好坏，只判"它是不是真的评不了"，判据是三位不同学科的评审给出互不相容的不能评理由。

按本文的判断，这个方案会失败，而且失败得相当快。原因不在它的份额定得多少、也不在它的判据宽严——而在于**"三位跨学科评审给出互不相容的理由"这句话本身就是一条可以被预演的判据**。它公布的第一天起，就可以被反向工程：写一份让三个学科都各自说"这不归我管"的申请，是一件可以被教会的写作技巧。

那份稿子自己给出的失败指纹是"互不相容变成模板化"——按本文，这个指纹会出现，而且会在两到三个申请周期内出现。

这不是一次姿态性的自我批评。它是本文的一条具体预言，且写下这条预言的人，正是当初写下那个方案的人。

第二件要认的账：本文自己就是一条可被预演的判据。

假设本文的观点被某个评价体系采纳，变成一条原则——"评价制度应当降低判据可预期性"。那么下一步会发生什么？会出现一批专门论证"我们的判据不可预期"的申请书，出现一套关于如何显得不可预期的写作规范，出现一批以"不可预期性"为关键词的评估指标。而这套指标本身当然是可预期的。

于是本文自己成了它所描述的那件事的一个标本：**任何被公开陈述的原则，都会成为下一轮预演的对象。** 这不是一个可以靠更谨慎的措辞绕开的循环，它是这条判断的直接推论。

我能做的只有两件，都不彻底：一是把这条循环写在这里，让引用它的人一开始就知道；二是给出上一章那些具体读数——因为读数比原则难被表演，你可以写出一份显得不可预期的申请书，你没法伪造五年后的主题分布熵。

第三件要认的账：写作过程本身。

这篇文章是从三份互相矛盾的材料里推出来的，而那三份材料并不是我随机遇到的——它们是我按"哪三家的分歧最不可调和"挑出来的。这个挑选动作本身就是一次预演：我知道什么样的组合会有结果，因为我见过没有结果的组合。所以这篇文章的问题，同样是被一条我自己心里的判据筛过的。

它可能因此漏掉了一整类问法：那些撞不凶、但也许更要紧的问题。我没有办法在这篇文章内部弥补这一点，只能记下来。

最后一件，也是最不舒服的一件。 本文所依据的三家里，有一家的实证结论（长周期、以人为单位的资助带来更高新颖度）已经流传了十几年，广为人知。这意味着"想拿那一类资助，该把自己长成什么样"，如今是可学的。**如果本文成立，那么被本文当作低可预期性标本的那套制度，其可预期性正在上升，而它的多样性优势应当正在衰减。** 这一条既是对本文的支持，也是对它的威胁：因为如果二十年后那套制度的优势依然稳定，本文的机制就有问题。

十七、结论，与一条写死日期的赌注

先把结论收拢成三句。

第一句：**同一个现象有三条成熟解释，而它们对同一批案例给出互相取消的预测**——最粗糙的判断本该最坏却最好，最不可量化的裁量本该是解药却更保守，期限与容错完全对齐的两条通道产出仍然不同。三条解释都不是错的，它们各自触到了真东西，只是都不是主变量。

第二句：**主变量是判据能否被提前推断。** 机制只有三步：预演、筛选、回流。它不需要任何人不诚实，不需要任何一次测量真的发生，因此它能解释那些"人人诚实、指标多元、评审专业、期限宽裕，而选题照样一年比一年像"的场合——那正是现有解释最说不通的地方。

第三句：**因此改判据的内容没有用。** 无论换成什么内容，只要它可被推断，预演照样发生；换尺子只是换了一个被预演的对象。真正可动的量有两个，都是本文被两个别的学科逼出来的：判据的**更新率**，以及判据到动作的**距离**。而降低可预期性有它自己的账单——方差不由所有人平摊，不可预期与不可问责只隔一层纸，长期工程需要稳定预期。这三笔账在第十四章已经摊开，本文不主张把整个体系一次性推向低可预期性。

剩下的事是让它可以被杀死。

一篇文章最容易活得太久的方式，是把所有断言都写成没有到期日的。所以最后留一条有日期、有读数、不可事后重新解释的赌注。

赌注：到 2032 年 12 月 31 日。

条件一：在此之前，全球至少有三家主要的公共科研资助机构，公开设立过"以人为单位、不要求提交研究计划"的资助通道，且每家运行满五年。

条件二：这三家通道与同机构同期项目制通道的对照数据（获资助者的选题主题分布、新关键词占比）可以从公开渠道取得。

若上述两个条件成立，而对照显示两类通道在选题分布熵上没有显著差异——本文作废。 不是"需要修正"，是作废：因为第七章那条排序是本文的全部承重，排序不成立，剩下的都只是把三家的话换了个说法。

若条件不成立——也就是说，到那时仍然没有足够的机构做过这件事、或者数据不可得——那么本文应当被视为**一条未被检验的猜想**，而不是一条已被支持的判断。我不接受"因为没人做，所以它仍然成立"这种记法。

再加一条更快到期的：**到 2028 年 12 月 31 日**，第十五章第一项那个最便宜的文本检验（新判据公布 → 申请文本语言跃升 → 是否早于第一轮结果）应当已经有人做出来。如果四年之内，一项只需要公开文本与一台电脑的检验都没有人做，那不是本文的错，但也不是本文的功劳——它只说明这条判断还没有进入任何人真正会去检验的问题清单。而按本文自己的机制，这恰恰是可以解释的：**它不在任何一条现成判据的射程内。**

如果那时它仍然没被检验，我希望有人把这句话当作证据，而不是当作辩解。

附：English abstract

The Predictable Yardstick: why the convergence of research questions tracks the anticipability of criteria rather than their coarseness, timing, or time horizon

Abstract. Fields covered by an evaluation regime for long enough converge on similar questions. Three standard explanations — legibility-driven simplification (Scott 1998), reactivity of measures used for decisions (Campbell 1979; Espeland & Sauder 2007), and the term-and-tolerance structure of funding contracts (Manso 2011; Azoulay, Graff Zivin & Manso 2011) — each carry evidence, yet issue mutually cancelling predictions on the same cases. This paper argues that all three treat as background what should be the variable: whether the criterion can be *anticipated* by those it will judge. Convergence, on this account, is neither the price of simplification nor the product of gaming, but the direct consequence of anticipability; it is independent of how coarse the criterion is, independent of contract length, and does not require any measurement to have occurred. Crossing anticipability with timing (ex ante / ex post) yields a 2×2 and a decidable ordering: proposal-based funding > publication metrics > people-not-projects ≈ post hoc blind judgment. Ten cross-domain redemptions are given, spanning surrogate endpoints, ML benchmarks, early-stage investment, rules-versus-standards, process-based assessment, hiring, content distribution, credit scoring, regulatory checklists and red-team exercises; two of them force the claim to concede ground, yielding two operational quantities — the *refresh rate* of criteria and the *criterion-to-action distance*. Six falsification conditions are supplied, one executable with public text alone, plus a dated bet.

Keywords anticipability of criteria; question-set convergence; rehearsal; criterion refresh rate; criterion-to-action distance; blind criteria

参考文献

1. Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *RAND Journal of Economics*, 42(3), 527–554.

2. Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.

3. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine*, 321(6), 406–412.

4. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: preserving validity in adaptive data analysis. *Science*, 349(6248), 636–638.

5. Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: how public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1–40.

6. Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24, 313–343.

7. Fleming, T. R., & DeMets, D. L. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*, 125(7), 605–613.

8. Goodhart, C. A. E. (1984). Problems of monetary management: the U.K. experience. In *Monetary Theory and Practice*. Macmillan.

9. Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7, 24–52.

10. Kaplow, L. (1992). Rules versus standards: an economic analysis. *Duke Law Journal*, 42(3), 557–629.

11. Manso, G. (2011). Motivating innovation. *Journal of Finance*, 66(5), 1823–1860.

12. Merton, R. K. (1973). The normative structure of science (1942). In *The Sociology of Science*. University of Chicago Press.

13. Osterloh, M., & Frey, B. S. (2020). How to avoid borrowed plumes in academia. *Research Policy*, 49(1), 103831.

14. Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4), 431–440.

15. Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? *Proceedings of ICML*.

16. Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely, T., & Pomeroy, L. (2020). The acceptability of using a lottery to allocate research funding: a survey of applicants. *Research Integrity and Peer Review*, 5, 3.

17. Avin, S. (2019). Mavericks and lotteries. *Studies in History and Philosophy of Science Part A*, 76, 13–23.

18. Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.

19. Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European Review*, 5(3), 305–321.

文献说明：本文对上述文献均为义释，不使用直接引文；具体页码须按所用版本核校。抽签资助的实例已核到公开报道与同行评议文献（新西兰健康研究理事会自 2013 年、大众基金会自 2017 年、瑞士国家科学基金会 2021 年底扩用于全部资助线），文中所述"先评议筛出够格、再在其中随机"这一共同形状亦有文献支持；具体年份与项目名请以各机构官方文本为准。本文的核心检验**均未执行**，交付的是命题、判据与方案。

字数：正文约 2.0 万字 · 版本 v1.1 · 2026-08-01 · 作者 王德生 + Claude

人机分工声明：本篇由王德生博士定方向与判据，Claude 执行三家源材料的选取、冲突判定、变量替换与全部行文，故署两个名字。文中实证读数均引自公开文献，未做新的数据采集；核心检验均未执行。文中所述"作者的一份前作"，指同日发表于本站的《单行道》一文。

附：与本栏另外两篇的关系

本栏另有一篇《别人怎么成的，学不来；别人怎么垮的，学得来》，同样撞的是站外三家在世学者，讲的是一个结论能走多远——成功是一长串「并且」，失败是一连串「或者」。那一篇管**知识能不能搬**，本篇管**知识被谁的尺子量**；两篇合起来是同一件事的两头：我们几乎全部的学习制度都建在搬不走的那一半上，而我们几乎全部的评价制度都建在可被提前推断的那一种尺子上。另有《能停下来，是因为不必交代》讲一件事必须持续证明自己值得继续时就再也停不下来，与本文第十四章那笔「不可预期与不可问责只隔一层纸」的账是同一个结构。

这一篇撞的三家，与可核对的出处

本篇撞的是站外三家理论，不是站内学员论文（同本栏之八、之九、之十的口径）。三家分属政治人类学、社会测量学与劳动经济学，都在解释同一个现象——被评价体系覆盖得久了，一个领域提出的问题会越来越像——而它们对同一批案例给出的预测互相取消。下面三条给出可直接点开核对的原始出处。正文第二至第四章逐一交代三家各自说到了哪一步，第五章把三对矛盾写死，第九章说明它们为什么各自推不出本文这一条，第十二章再与另外七种最容易被混为一谈的说法逐条分界——**包括与本文离得最近、且本文承认最可能被它吸收的那一个**。

可读性 · 詹姆斯·斯科特（1998）

Seeing Like a State：国家为了看见而简化，简化摧毁了长在情境里的那种本事

德国科学林业把混生杂木改造成等距成行、可清点的人工林，第一代长势极好，第二代土壤板结虫害成片。**与本文的分离线**：那一家按简化程度排序，把最粗的判断排在最坏一端；本文把它排在最好一端——同一个案例，相反的排序。

反身性 · 埃斯佩兰 与 索德（2007）

Rankings and Reactivity：公共测度不是在描述世界，是在重造它所测量的世界

法学院排名一出，招生策略、就业数据口径、奖学金分配全部朝那几个被计入的量重新组织。**与本文的分离线**：那一家的时序是先测量、后反应；本文的机制在测量之前就闭合了——一条新判据刚公布、第一轮结果还没出来时，那一家预测无变化，本文预测申请文本的语言已经跃升。

资助契约 · 阿祖莱、格拉夫·齐文 与 曼索（2011）

Incentives and Creativity：容忍早期失败、拉长周期的资助，换来了更高的新颖度

以人为单位、明写容忍失败的资助，其研究者的高影响力论文与新关键词都更多，同时低引用的失败品也更多。**与本文的分离线**：那一家把差别全部归给期限与容错；本文预测，把期限与容错完全对齐、只改「要不要先交一份研究计划」，两条通道的选题分布仍会显著不同。