学科通融 · 之五十

第一个读数之前

论新旧之别不在产出、不在效率、也不在承认，而在一样东西从被做出来到取得第一个读数所需的时间

经济学 × 艺术学 × 工程学

作者 王德生 + Claude

栏目 学科通融 · 站外碰撞

篇幅 约2.0万字 · 14页

发表 2026年8月2日

SDE Universes · 德麦国际 Demai International · 本篇由三个分属不同学科、且互相冲突的理论体系碰撞而成；三家来源均为站外公开文献

摘　要

一个领域有没有真的长出新东西？工程学说看输出——在标准评测上的表现就够了；经济学说看投入产出比——把多少人力组织成多少产出增长，这正是「想法越来越难找」那条读数的做法；艺术学说两个都不算数——一样东西是不是那件事由条件场决定，而条件场自己是被历史地生产出来的。把同一个用人工智能大幅提高了产出的团队交给三家，一家判进步、一家判衰退、一家判两边都在测一个已经不在那里的东西。三家共同假定了**这件事可以在某一维上单独读出来**，而**推翻这条假定的材料来自其中一家自己**：研究生产率必须为每个领域各选一个产出代理，而代理只能从已被承认为该领域产出的那一类里挑——于是新的产出落在分子之外、却已经落在分母之内，这条读数在结构上必然下降，与新东西实际有多少无关。由此得到的判断是：**新与旧的分别不在产出、不在效率、也不在承认，而在无量期的长短**——一样东西从被做出来到在任何一维上取得第一个读数之间的那一段；在此之前它在三家账上都不是「小」，是零。而人工智能压缩的只是**已有现成读数**那一类的无量期，完全没有压缩**需要先造一把新尺子**的那一类，两类的显形速度差被拉开了数量级；资源被抽向短的那一侧不是因为人短视，是加总规则的算术后果。文中给出一句不含判断词的问话（上一轮做完的东西里，有多少件在开工时还没有任何现成的量可以判定它）、一个四领域量纲不同而比率相同的读数（无量期占比）、两轴四格与最危险的那一格（断供）及三个写实场景、十二处兑现、三件有固定次序的出路与一条反直觉推论（**在造尺子被单独记账之前，任何提高评价质量的努力都会拉大差距**）、两处反过来把本文削小的约束、与十种既有说法的判决性分界（含最危险的方法学占位者「睡美人系数」）、与本栏同题六篇各一条对照预测，七条排除了最强竞争解释的判错条件，以及一条写死到二〇三一年年中的赌注。**本稿为理论建构与研究设计，不报告任何虚构的样本、统计结果或实验结论。**

这一篇由哪三个学科的理论体系撞成

三家在「一个领域有没有真的长出新东西该在哪儿读出来」这同一问题上给出互斥答案：工程学说看输出在标准评测上的表现就够了；经济学说看投入产出比就够了；艺术学说两个都是由旧条件定义的代理，而条件场自己正在被改变，因而两个读数都在测一个已经不在那里的东西。把同一个对象交给三家，得到判进步、判衰退、判两边都在测错三种不能并存的归属。　三家均为站外的公开文献，链接直达原始出处，可自行核对。

经济学 · 研究生产率与产出代理

Are Ideas Getting Harder to Find?（Bloom, Jones, Van Reenen & Webb, *American Economic Review* 110(4), 2020）

长期增长率＝有效研究人员数　×　研究生产率；多领域证据显示研究投入大幅上升而研究生产率急剧下降（维持芯片密度每两年翻番所需的研究人员是 1970 年代初的十八倍以上）。关键的技术细节是：因为「想法」没有统一单位，每个领域必须各选一个产出代理——晶体管密度、作物单产、寿命、获批新药。

https://www.aeaweb.org/articles?id=10.1257/aer.20180338

工程学 · 基准饱和与构念效度

When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation（Akhtar et al., 2026, arXiv:2602.16763）；另见对 445 篇基准论文的系统检查（Bean et al., 2025）与欧盟联合研究中心的评测审视（Eriksson et al., 2025）

六十个常用基准中二十九个出现高度或极高度饱和——排在前面的模型被压缩在一起，分数差不再对应能力差；污染与构念效度问题被普遍承认。该支给出的处方始终是造更好的基准，即承认尺子不好、但不承认「尺子」这个判定位置有问题。

https://arxiv.org/abs/2602.16763

艺术学 · 现场性是被历史地生产出来的范畴

Liveness: Performance in a Mediatized Culture, 3rd ed.（Philip Auslander, Routledge, 2023）；Digital Liveness: A Historico-Philosophical Perspective（*PAJ* 34(3), 2012）；对照 Peggy Phelan, *Unmarked*（1993）

「现场」不是可在技术之外被指认的本体属性，它作为一个范畴是在录音与广播出现之后才成立的；本真性并不站在技术之外，它是技术的产物。推论：一样东西是不是那件事由条件场决定，而条件场自己在动——因此「被不被承认」永远滞后于「发生」。

https://www.routledge.com/Liveness-Performance-in-a-Mediatized-Culture/Auslander/p/book/9780367468170

目 录

摘 要

一个领域有没有真的长出新东西？三门学科各自给出一个自足的答案，而三个答案不能同时为真。工程学说看输出：一个系统够不够格，看它在标准评测上的表现就够了。经济学说看投入产出比：一个体系还行不行，看它把多少人力组织成多少产出增长就够了——这正是"想法越来越难找"那条著名读数的做法。艺术学说两个都不算数：一样东西是不是那件事，由它所处的条件场决定，而条件场自己是被历史地生产出来的，物件与做法都不决定它。把同一个用人工智能大幅提高了产出的团队交给三家，工程学判进步，经济学判衰退，艺术学判两边都在测一个已经不在那里的东西。

本文论证，三家共享一条从未被说出的假定：**一个领域有没有长出新东西，可以在某一维上单独读出来**。推翻这条假定的材料来自其中一家自己：研究生产率的计算必须为每个领域各选一个产出代理——晶体管密度、作物单产、寿命、专利——而代理只能从**已经被承认为该领域产出的那一类**里选。于是任何真正新的产出都落在分子之外，却已经落在分母之内；这条读数在结构上必然下降，与新东西实际有多少无关。

由此得到的判断是：**新与旧的分别不在产出，不在效率，也不在承认，而在一样东西从被做出来到在任何一维上取得第一个读数所需的时间**——本文称之为**无量期**。在取得第一个读数之前，一样东西在三家的账本上都不是"小"，是零：它不构成任何一项。而人工智能这一轮所做的事有一个被忽略的不对称——它把**已经有现成读数**的那一类的无量期压到接近于零，却完全没有压缩**需要先造一把新尺子**的那一类。两类的显形速度差因此被拉开了数量级。资源永远流向先显形的那一侧，不是因为人短视，而是因为在任何一维上，尚未显形的东西在账面上等于不存在。

本文给出一句不含判断词的问话（上一轮做完的东西里，有多少件在开工时还没有任何现成的量可以用来判定它），一个三领域量纲不同而比率相同的读数（无量期占比），一张两轴四格的辨别表并指认最危险的一格（断供），十二处兑现，两处反过来把本文削小的约束，与十种既有说法的判决性分界，与本栏同题六篇各一条对照预测，七条排除了最强竞争解释的判错条件，以及一条写死到二〇三一年年中的赌注。本文为理论建构与研究设计，不报告任何虚构的样本、统计结果或实验结论。

关键词

无量期　无量期占比　断供　读数结构　显形速度差　人工智能与内卷

一 · 导论：三个行当同时在问同一句话

二〇二五到二〇二六年之间，三个互不往来的行当各自陷入了同一种争吵。

做模型的人在吵评测。　一项对六十个常用基准的系统研究发现，其中二十九个已经出现高度或极高度的饱和——不是所有模型都答满分，而是排在前面的那一批被压缩在一起，分数之间的差别不再对应能力上

的差别。同一批文献还在追问一个更靠前的问题：这些题目到底测的是不是它们宣称要测的那个东西。一项对四百四十五篇基准论文的检查发现，只有大约六分之一在比较结果时使用了不确定性估计或统计检验。而污染——测试题出现在训练语料里——已经被承认为普遍现象。于是一边是新基准以每月为单位地被造出来，一边是"我们其实不知道自己在测什么"这句话被越来越多的人说出口。

做研究政策的人在吵生产率。 一条影响极大的读数说：研究投入在持续上升，而研究生产率在急剧下降。最有名的例子是摩尔定律——今天要维持芯片密度每两年翻一番，所需的研究人员数量是一九七〇年代初的十八倍以上。这条读数被广泛用来支持"想法越来越难找"这个判断，也被广泛用来解释为什么投入越来越多而增长越来越慢。而人工智能进入研发之后，争论变得更急：如果工具把每个人的产出放大了几倍，为什么这条曲线没有掉头。

做艺术的人在吵什么还算数。 生成模型能在几秒钟内产出以往需要数月的东西之后，"这算不算作品""这算不算现场"重新变成一个要认真回答的问题。表演研究里有一条已经争了三十年的线：一派认为现场性是不可复制的本体属性，另一派认为"现场"这个范畴根本不是本体给定的，它是在录音与广播技术出现之后才被历史地生产出来的——没有"录下来的"就不会有"现场的"。这一派的推论很不客气：所谓的本真性并不站在技术之外，它是技术的产物。

三个行当，三种材料。表面上是三场专业内部的争吵，实际上三家在同一个问题上给出了不能并存的答案：

> **一个领域有没有真的长出新东西，该在哪儿读出来？**

这个问题不是学院里的问题。它决定了钱往哪里去、人往哪里去、什么被继续做下去。而在生成成本塌陷之后，它变成了一个急迫的问题——因为"看起来在长新东西"和"只是把同一样东西做得更多更快"，在今天的所有读数上，可以完全一样。

二 · 三家的位置

选这三家，不是因为它们离得远，是因为它们各自认为**只看一处就够了**，而三处互不相同。

2.1 工程学：看输出就够了

工程学这一支的承重主张是：**一个系统够不够格，看它的输出在标准条件下的表现就够了；怎么实现的不重要，处在什么环境里也不重要。** 这是黑箱验收的逻辑，也是过去十年人工智能工程的组织方式：造一个基准，跑一个分数，排一个榜，据此决定谁值得继续投入。

这套做法的力量在于它把判断变成了可传递、可复现、可被第三方核对的东西——任何人拿同一套题目跑一遍，得到同一个数。它的代价现在也已经被这一支自己写清楚了：饱和（分数差不再对应能力差）、污染（题目进了训练语料）、构念效度存疑（不确定测的是不是要测的那个）。值得注意的是，这一支给出的处方始终是**造更好的基准**——活的基准、持续更新的基准、防污染的基准、覆盖多个维度的整体评估。也就是说，它承认现有的尺子不好，但它不承认"尺子"这个位置有问题。**输出仍然是唯一的判定处。**

2.2 经济学：看投入产出比就够了

经济学这一支的承重主张是：**一个体系还行不行，看它把多少投入组织成多少产出增长就够了。** 长期增长率被拆成两项之积——有效研究人员数量，与他们的研究生产率；把产出增长除以研究人员数量，就得到研究生产率。这条读数在多个领域给出同一个方向：投入在涨，生产率在跌。

这一支的力量在于它跨领域可比、可以做长时间序列、并且直接对应资源配置决策。它的做法里有一个必须交代清楚的技术细节，本文后面要用到它：**"想法"没有统一单位，所以每个领域必须各选一个产出代理**——半导体用晶体管密度，农业用作物单产，医药用寿命或获批新药数，整体经济用全要素生产率。这些代理各有各的道理，而且都是这一支自己反复讨论过、公开承认过的选择。

2.3 艺术学：看条件场就够了

艺术学这一支的承重主张是：**一样东西是不是那件事，由它所处的条件场决定；物件与做法都不决定它，而且条件场本身是被历史地生产出来的。** 表演研究里那条最锋利的论证是这样走的："现场"不是一种可以在技术之外被指认的属性；它作为一个范畴，是在录音与广播出现之后才成立的——在没有任何东西可以被录下来的年代，没有人需要"现场"这个词。而一旦这个范畴成立，现场表演就开始向被媒介化的形式靠拢，两者的界线不断被重划。

把这条主张放到本文的问题上，它给出的诊断很不客气：**输出的分数与投入产出比都是代理物，而一个代理能不能代表那件事，取决于承认它的那套条件；人工智能改变的恰恰是那套条件本身。** 所以两个读数不是测得不准，是在测一个已经不在那里的东西。

三 · 冲突定位：同一个对象，三家判进步、判衰退、判测错

三条主张听上去可以各管一段。放进一个具体对象，就管不成了。

对象：一个把人工智能大规模用进日常工作的研发团队。一年之后，它的输出在所有标准评测上都更好了；它的人员与算力开支涨了三倍，而它所在领域的产出增长率没有变化；它做出来的东西在同行会议上被讨论的方式与三年前不再相同——原来用来评价这类工作的那套语汇正在失效。

- **工程学判进步。** 输出更好了，这是可复现、可核对的事实。至于开支和语汇，那是别人的事。
- **经济学判衰退。** 投入涨了三倍而产出增长率不变，研究生产率跌了。至于评测分数上涨，那是分子里的噪声——评测本身就是这个体系自己造的，用它来证明自己变强了是循环论证。
- **艺术学判两边都在测错。** 评测的内容是由旧的问题结构定义的；产出代理是由旧的承认关系挑出来的。一年之内这两套条件都动了，所以两个数不能跨期比较，而三家争的那个"变强还是变弱"，在新的条件场里可能根本不是同一个问题。

三家互斥，而且是**划界互斥**——对同一个对象是不是属于"真的长出了新东西"这一类，三家给出不能并存的归属。这比结局对立更硬，因为你可以取任何一个具体对象，逼三家各自给出"算/不算"。

更要紧的是，每一家每天在做的事，正是另一家诊断为病根的事：

- 工程学每天在做的**把判定固化成可重复的题目**，正是艺术学诊断为"用一个由旧条件定义的代理去替换那件事本身"。
- 经济学每天在做的**用一个产出代理去计量整个领域**，同样落在艺术学的诊断里。
- 而艺术学的处方在另外两家看来根本不是处方：如果两个可核对的读数都不算数，剩下的是什么？是"要看那件事本身"——而"那件事本身"没有任何人能拿出来给第三方核对。

三家争的，说到底是**"有没有长出新东西"该在哪一处结算**。而这个争法本身，藏着一条三家都没说出口的假定。

四 · 三家共同假定了什么，以及推翻它的材料来自哪一家

4.1 那条共有假定

把三家的争论压成一句：**三家争的是"有没有长出新东西"这件事该在哪一维上被读出来**。

于是三家共同假定了：

> **这件事可以在某一维上单独读出来——只要那一维的读数取对了，另外两维就不必再问**。

把这句话念给三家听，三家都会说"这还用说吗"。它不是谁的主张，它是三家的主张之所以能成为主张的前提：如果不预设某一维单独够用，"该在哪儿读"这个问题就问不出来。

4.2 推翻它的那件材料，来自经济学自己

推翻这条假定的材料不必从外面搬，它就在第二家自己的做法里，只是那件事在它自己的领域里回答的是另一个问题。

研究生产率的计算要回答的问题是：**怎么把"想法产出"变成一个可以做除法的数**。 它的答案是：因为"想法"没有统一单位，所以每个领域各选一个产出代理——晶体管密度、作物单产、寿命、获批新药、专利。这是这一支公开写明并反复讨论过的技术选择，不是本文的指控。

把这个选择换一个方向读，得到的却是另一件事：

> **产出代理只能从已经被承认为该领域产出的那一类里挑出来**。 一个真正新的产出，在被承认为该领域的产出之前，不可能进入代理；而做它的人已经被数进了研究人员的总数。**于是它落在分子之外，却已经落在分母之内**。

结论很硬：**研究生产率这条读数,在任何"正在长新东西"的时期都会系统性下降,而下降的幅度与新东西实际有多少无关——甚至可能与它成正比**。 新东西越多、越不像旧东西，落在代理之外的部分越大,分子越难看。

这不是说"想法越来越难找"这个判断错了。它可能是对的。本文说的是更麻烦的一件事：**这条读数在结构上不能用来分辨"真的没长出新东西"与"长出来了但还没有代理"**。而这两种情形要求完全相反的处置。

一旦承认这一点，三家的共有假定就站不住了。因为同样的结构在另外两维上一模一样：

- **输出维**：一个基准只能测已经被题目化的能力。一个真正新的能力在被造出题目之前，在所有现成基准上要么得零分，要么不适用——**而"不适用"在榜单上不显示为任何东西**。基准饱和与构念效度那一整批讨论，正是这条在这一支内部的表现：它们诊断出了尺子的毛病，却把毛病归给尺子做得不够好。
- **条件维**：这一支自己承认，条件场是被历史地生产出来的——也就是说，"被不被承认"永远滞后于"发生"。用一个滞后量去判定它的先行量，判出来的永远是上一轮的东西。

三维互生，任一维都读不出全部。**三家不是选错了那一维，是假定了存在这样一维**。

4.3 由此打开的问题

如果没有任何一维能单独读出新旧，那么真正要问的就不是"该在哪儿读"，而是：

> **一样东西从被做出来，到它在任何一维上取得第一个读数，中间隔了多久？而这段时间里，它靠什么活着？**

这个问题在三家的框架里都问不出来。它也正是本文接下来要回答的那一个。

五 · 承重判断：无量期

5.1 名与义

先把一个词定死。

无量期：一样东西从被做出来，到它在任何一维上取得第一个读数之间的那一段时间。所谓"取得读数"，指的是它开始能被某一个现成的量所计量——进入某个基准的题目范围、被某个产出代理计入、或被某个既有的类别接住。

关键在于：**在取得第一个读数之前，它在三家的账本上都不是"小"，是零**。　它不是一个数值很低的项，它不构成任何一项。一个还没有任何量能测到的东西，在按量运行的系统里没有本体地位——这不是记录做得不细，是"记录"这个动作在它身上无从下手。

有了这个词，承重判断可以写成一句：

> **新与旧的分别，不在产出，不在效率，也不在承认，而在无量期的长短；而人工智能这一轮所做的事，是把已经有现成读数的那一类的无量期压到接近于零，却完全没有压缩需要先造一把新尺子的那一类。两类的显形速度差被拉开了数量级，于是资源被系统性地抽向短的那一侧——不是因为人短视，而是因为在任何一维上，尚未显形的东西在账面上等于不存在**。

三个被否定的说法各来自一家：产出是工程学的划界标准，效率是经济学的，承认是艺术学的。三条都描述了真实的东西，三条都不能分辨新旧。

5.2 为什么它不是"眼光"问题

最容易把本文的判断吸收掉的说法是：这不就是短视吗，不就是急功近利吗，不就是资本没有耐心吗。

这条读法把问题放在**动机**上，而本文要说的是**结构**。区别可以写死成一句可检验的话：

> 若问题在动机，那么**一个动机足够长远、考核周期足够长的组织，应当不受此影响**。 > 若问题在结构，那么**即使动机与考核周期完全相同，账面回报仍会系统性偏向短无量期的那一侧**——因为长无量期的那一类在账上根本不出现，它不是"回报低"，是"没有条目"。

第二种情形有一个反直觉的推论：一个真心想做长线的机构，在它自己的报表上会看到它的长线投入**什么都没有产生**——不是产生得少，是那一栏是空的。于是每一次复盘都会给出同一个建议：把资源挪到看得见结果的地方去。**这个建议在它自己的证据体系内部是正确的**。

5.3 这一轮为什么被拉开了

生成成本的塌陷不是均匀地作用在两类事情上。

它压缩的是"做"的时间。 一份分析、一段代码、一幅图、一篇初稿，从想到做完，从几周压到几小时。对于那些**已经有现成读数**的事，做完就是显形——它立刻进入某个基准的题目范围、被某个代理计入、落进某个既有类别。所以这一类的无量期，本来就短，现在接近于零。

它没有压缩的是"造尺子"的时间。 一把新尺子——一个新的评价指标、一个新的期刊栏目、一个新的评奖门类、一套新的验收标准——要成立，需要有人达成共识、需要样本、需要争论、需要制度承认。这些的速度基本没有变，因为它们受限的不是产出能力，是协调能力。所以那些**需要先造一把新尺子**的事，无量期仍然是几年。

差值因此被拉开：原来大概是几个月对几年，现在是几小时对几年。**这不是程度问题，是账面上的两类东西彻底分了家**。

而任何按量运行的系统——季度复盘、年度考核、赛道估值、榜单排名——在这两类之间做选择时，结果是被写死的：短无量期的那一类在每一次结算时都在账上，长无量期的那一类在每一次结算时都不在账上。**这不需要任何人做出偏好，它是加总规则的算术后果**。

这就是本文认为"新经济现象"这个说法名副其实的地方。它不是旧内卷的加剧。旧内卷的画像是"更多人挤在同一条路上"；这一轮的画像是**账面上只剩下一条路，因为另一条路在账面上没有出现过**。

六 · 两轴四格，与最危险的那一格

一个名字不是一条辨别维度。要能分辨，得有第二根轴。

第一根轴是这件事的**无量期**：做完之后立刻能被某个现成的量读到，还是要等一把尺子被造出来。第二根轴是**有没有人替那段没有读数的时间买单**：存不存在一个具体的位置，愿意在没有任何量能证明它有价值的情况下继续供养它。

	有承担者	无承担者
无量期短	① **正常量产**：做完即显形，也有人负责它做得对不对。	② **无所谓**：读数自己会来，有没有人扛都一样。绝大多数日常工作在这一格。
无量期长	③ **真正的研究与创作**：有人替那段没有读数的时间扛着——长周期基金、导师、赞助人、耐心的资本、一个愿意等的老板。	④ **断供**

第四格是本文要打的那一格，本文给它一个名字：**断供**——一样东西被做出来了，它需要一段时间才能被任何量读到，而没有任何位置替这段时间买单；于是它在取得第一个读数之前就停了。

选这一格，不是因为它最惨，是因为**它最不像有问题**。一格值得被打，要满足三条：

其一，它在短期指标上表现为改善。 把资源从④挪到①，产出量上升、评测分上升、研究生产率的分子上升、项目按时交付率上升。每一个数字都变好，而且是真的变好——不是伪造的。

其二，它的失效不产生任何信号。 ④里的东西停下来时，没有任何账目减少一分——它本来就是零。**唯一能记录这次损失的地方，是它本来会取得的那个读数；而那个读数从未出现过。** 一个从来不曾进入任何账本的东西，它的消失在原理上不可被记账。这一条与错误的项目被砍掉正好相反：砍掉一个有读数的项目会立刻在报表上留下一个缺口，而④里的东西消失时，报表更好看了。

其三，它自我加固。 越是把评价做得及时、精确、频繁——这正是过去三十年所有管理改进的方向——长无量期的东西越吃亏。**每一次把考核周期从一年缩短到一季度，都在把④扩大。** 而缩短考核周期在任何一套管理学说里都是进步。

三条齐了，靶格成立。也正因为三条齐了，断供不会被自己发现——**它只能被从外面数出来，而且只能靠数那些还在的东西的构成，不能靠数消失的东西。**

断供长什么样

抽象地说三条签名不够，得看见它。下面三个是把常见情形合成的典型场景，不是对具体个案的报道；它们的用处是让读者认出自己见过的那一种。

一个实验室里的方向。 某个人做出了一个现象。它是真的，重复得出来，而且他自己知道它要紧。问题是没有任何现成的指标能说明它有多要紧——它不在任何一个既有方向的评价体系里。第一年的年度汇报，他这一栏是空的：不是数字难看，是没有栏目可以填。第二年，出于好意，这个方向被并进一个有指标的相邻方向，"先按那边的口径报着"；并进去之后，他的时间被那边的产出要求占满。第三年他离开了，去了一个不需要按季度报数的地方，或者干脆去做别的。

这个过程里，**没有任何一份文件记录了"我们停掉了什么"**。没有立项被撤销，没有申请被拒绝，没有争论。年度报表在这三年里逐年好看，因为空的那一栏被有数的栏目替换掉了。如果五年后有人问"我们当初为什么没做那个方向"，档案里查不到任何答案——因为从来没有人做过那个决定。

一支工程团队里的探针。 一个工程师发现现有的监控体系测不到某一类问题：它不产生告警，不进日志，只在事后回看时能从几个不相干的指标的形状里勉强看出来。他花了两周做了一个粗糙的探针。评审的时候问题很简单也很正当：它现在能测到什么已知问题？答案是零——它是为还没有名字的那一类做的。于是它被判为无收益，排期给了别的事。

六个月后，那一类问题以一次事故的形式出现了。事后复盘写的是"监控覆盖不足，需增加告警规则"，随后新增了七条规则。**那支探针不在任何记录里**：它没有被否决过，它只是没有被排进去。而新增的七条规则测的是这一次已经发生过的那个形状——它们把这一类的无量期变成了零，同时把下一类的仍然留在几个月。

一个不属于任何门类的作品。 一位创作者做的东西横跨两三个既有形式之间。征集的报名表第一栏是门类，必选；她每次都得挑一个最不离谱的填进去，然后在评审里被按那个门类的标准评——评语通常是"作为某某类不够成熟"。三年之后她开始做能被归类的东西，因为那样至少能进入被讨论的范围。

转变的那一刻，没有任何人做过决定，也没有一封拒信。 她收到的每一条评语都是就事论事、态度诚恳的；每一位评委都在按自己领域的标准认真工作。而报名表那一栏——一个为了统计方便而设的必选项——在入口处完成了整件事，不需要任何人的判断参与。

三个场景的共同点，正是第六节那三条签名：每一次账面都在改善（汇报更完整、监控覆盖更全、评审更规范），失效不产生任何信号（没有决定、没有拒绝、没有记录），而且越规范越严重（门类越细、指标越全、规则越多，滤得越干净）。**它们不是失败的案例，它们是运转良好的系统的正常输出。**

七 · 一句不含判断词的问话

要把一格从纸上落到地上，需要一句可以拿去问文件的问话，而不是拿去问人的判断。本文的那一句是：

> **上一轮做完的东西里，有多少件在开工的时候还没有任何现成的量可以用来判定它？**

这句话里没有"应当""重要""真正""有价值"。它问的是立项书、验收标准、评审细则、报名表——问的是记录，不是意见。

把它在五个场合各跑一遍，答案互不相同：

一家实验室的立项书。 查每份立项书"评价指标"那一栏：填的是现成指标，还是"待建立"。这一栏几乎所有资助体系都有，而几乎没有人把它当作一个可以统计的量看过。

一支工程团队的季度计划。 查每一项的验收标准在开工时存不存在。这是五个场合里最容易查的一个，一个下午能查完。

一本期刊的栏目。 查最近一年的来稿里，有多少篇不属于任何既有栏目或关键词表。这一条**查不出来**——因为投稿系统强制选栏目，一篇不属于任何栏目的稿子在系统里根本无法提交。**"查不出来"本身就是答案，而且是五个里最尖的一个：形式系统在入口处就把无量期长的东西滤掉了，连一次拒绝都不需要产生。**

一次艺术征集。 查入选作品里有多少无法归入任何既设门类。答案通常是零，理由与上一条相同——报名表要选门类。

一家公司的季度目标。 查每一条关键结果在制定时有没有现成的度量。答案通常是"全部都有"，而且这不是疏忽：**主流的目标管理方法论明确要求关键结果必须可度量。** 这一条是对照组——一套方法论把无量期长的事在定义上排除掉了，而且它是这样写进教科书的。

五个答案互不相同，其中至少两个不是从命题推出来的、是查出来的：投稿系统的强制选项，与目标管理方法论的定义性要求。前者说明滤除发生在入口而非判定；后者说明这不是执行走样，是设计意图。

八 · 一个量纲不同而比率相同的读数

一句问话只能问一件事。要让不同行当能并排比较，需要一个数。

无量期占比：最近 N 件已立项或已完成的产出中，在开工时不存在任何现成读数——没有既有指标、基准、栏目、门类、验收标准可用——的件数占比。

它在四个行当里的量纲完全不同，比率却相同：

- **学术**：立项书"评价指标"一栏填"待建立"的比例。
- **工程**：开工时没有可用验收基准的项目比例。
- **艺术**：入选作品中无法归入任何既设门类的比例。
- **企业**：季度目标中没有现成度量的条目比例。

这个读数有四条性质，缺一条都不该用它。

其一，无量纲，所以四个行当可以并排比较：一家实验室的零点一与一本期刊的零点一说的是同一件事。

其二，可事后清点：所需材料在既有记录里已经存在——立项书、验收文档、报名表、目标清单。不需要新建任何系统。

其三，它把一个印象变成一个数："我们这里只做有把握的事"是一句人人都说过的抱怨；抱怨不能被检验，比率可以。

其四，也是最要命的一条：它与既有的主要指标正交。 一个与既有指标高度相关的新读数会被立刻读成那个指标的代理，白造。所以必须举一个"既有指标最好看、而这个数最难看"的真实例子——举不出来就该重造。

本文的例子是**一个所有项目都能在季度末给出量化结果的研发部门**。按既有指标看，它接近完美：交付率满分，产出量创纪录，每一项都能报出数字，评测分年年上升。而它的无量期占比是零——因为一个开工时没有现成度量的项目，在这里根本立不了项。这两个读数指向相反的结论，而且都不是错的：它确实每一项都做成了，也确实一件没有尺子的事都没做过。

反过来也成立：一家产出量很低、报表很难看的小实验室，可以有很高的无量期占比。**产出量低不等于在做新东西，产出量高也不等于没在做——这两个数第一次被分开，正是这个读数的用处**。

九 · 十二处兑现

兑现不是"像"，是在那个领域里能据此预测一件具体的事。

一 · **公共科研立项**：在生成工具普及最快的学科里，立项书"评价指标待建立"的比例会先于"原创性下降"的抱怨出现，而不是相反。可查：资助机构的立项库。

二 · **期刊栏目**：栏目表的更新频率会落后于来稿主题的变化速度，且这个差值在生成工具普及后扩大。预测：跨栏目投稿被退回改投的比例上升，而这一项通常没人统计。

三 · **工程验收**：有现成验收基准的项目占比会上升，而不是下降——因为无基准的项目在排期会上说不出预期收益。预测：这一上升与团队规模无关，与考核周期长度强相关。

四 · **人工智能评测建设**：新基准建成后，被它首次读到的能力中，会有相当比例是**在基准建成之前就已经存在而无人计数的**。预测：这些能力的"首次出现时间"若可从模型版本记录回溯，将系统性早于基准发布时间数月至数年。

五 · **专利分类**：新技术方向进入正式分类表的滞后，与该方向在此期间的实际申请量正相关——分类跟不上不是因为方向小，恰恰因为方向大到不得不动分类。预测：分类表更新前后，同一方向的"检索可见度"会跳变，而技术本身没变。

六 · **药物研发**：全新机制的项目与已知机制新适应症的项目，在同一家公司内部会有系统性不同的存活率，而这个差别不由科学风险解释——已知机制的项目在每一次季度评审时都有现成终点指标可报。

七 · **艺术征集与评奖**：门类表越细，跨门类作品的入选率越低；而门类表的细化在任何评审改革里都被写作"更公平"。预测：细化门类之后，入选作品的形式多样性下降而评审争议减少。

八 · **风险投资**：赛道命名成立之前的项目融资难度，与该赛道成立之后同类项目的融资难度差一个数量级，而项目本身没有变。预测：赛道命名的时点可从投资机构的公开文章回溯，融资难度的跳变应紧随其后。

九 · **教育课程**：一门还没有考试形式的课，在任何课程评估体系里都拿不到分。预测：新设课程会系统性地先设计考核方式、后设计内容，而这一顺序在教学大纲的修订记录里可以直接看到。

十 · **政府科技计划指南**：指南里"研究内容"与"考核指标"两栏的对应关系，决定了哪一类工作能被写进去。预测：要求填写量化考核指标的计划，其资助方向的年际变化率显著低于不要求的计划。

十一 · **企业目标管理**：把考核周期从一年缩短到一季度的组织，其无量期占比会在两个周期内下降，而所有交付指标同时改善。这一条是本文最直接的一个可观察后果。

十二 · **开源项目**：一个还没有对应标签的问题类型，在议题系统里既搜不到也统计不到；预测：标签体系的更新会滞后于问题类型的出现，而滞后期内该类问题的重复提交率异常高——重复提交正是"读数不存在"的一个副产品。

十二处里有两处反过来把本文削小了，下一节单说。

十 · 两处反过来削小本文的约束

10.1 工程学削掉的：适用范围

在安全关键与合规领域，**没有现成读数就不许开工**是正确的规矩。你不能把一个没有验收标准的系统装上飞机，不能把一个没有终点指标的方案推进临床，不能让一段没有测试的代码进入支付链路。在这些地方，无量期占比接近零不是病，是设计目标。

这一处直接削掉了本文原本会写下的一句更强的话：

> **原本会写**：无量期占比越高越健康，组织应当提高它。 > **现在只能写**：本文不主张提高任何组织的整体无量期占比。这个读数只适用于**以探索新可能为职能的那一部分单元**，而且它的用法是"这个组织有没有留出这样的单元"，不是"这个组织整体应当多冒险"。

这是把适用范围砍掉了一大块：工业质量控制、安全工程、合规、临床推进，全部不在内。判据也随之明确：这个单元的职能是**产出可靠的结果**，还是**找到还没有人做过的事**？前者无量期短是目标，后者无量期为零是断供。

10.2 艺术学削掉的：价值排序与单件适用性

艺术学那一支提醒本文一件更难受的事：**"新"这个范畴本身也是被条件场生产出来的。** 一件事在开工时没有现成读数，并不等于它是新的；它也可能只是**没有人关心**。一个无人问津的方向和一个尚未被识别的方向，在无量期这个读数上完全一样。

这一处削掉的是本文的价值排序，也削掉了它在单件上的用法：

> **原本会写**：无量期长就是新，就值得被供养。 > **现在只能写**：无量期长是"新"的**必要条件而非充分条件**。因此这个读数**不能用来判定任何单独一件事的价值**——它只能用来判定一个**组织的构成**：这个组织有没有为长无量期的那一类留出位置。用它去判单件，等于把一个统计性质当成个体属性，那是误用，而且是会伤到人的误用。

这两处约束合起来，把本文从一条关于"什么值得做"的普遍主张，压成了一条关于"组织构成"的有限主张。不接受这个压缩，本文会在第一个安全工程师和第一个策展人那里各栽一次。

10.3 经济学补的一处

研究生产率那条读数在有稳定代理的领域——半导体、农业——确实有效，而且它的下降在那些领域里很可能就是字面意思。所以本文不主张废弃它，只主张它**不能单独用来判新旧**，并且**在正在长新东西的时期，它的下降不构成证据**。这一条削掉的是本文原本想写的"这条读数应当被弃用"。

十一 · 与最近的几种说法的分界

本文最容易被吸收进以下十种既有说法。逐条写清差在哪，以及怎样才能判出谁对。

一 · 想法越来越难找（Bloom, Jones, Van Reenen & Webb 2020，经济学）。它说到的那一步是：研究投入上升而研究生产率下降。分界在于：本文认为它测的是"在既有代理下的产出效率"，而代理只能取自已被承认的产出类。判定形式：**若某领域在换用一个更宽的产出代理之后，研究生产率不再下降，则本文对；若换任何代理都下降，则它对。** 这条可以在有多种可选代理的领域（医药、农业）直接跑。

二 · 基准饱和与构念效度（Akhtar et al. 2026；Bean et al. 2025；Salaudeen et al. 2025，人工智能评测）。它们诊断"基准测不准"，处方是造更好的基准。分界在于：本文说**任何基准在原理上都测不到还没有基准的那一类**，造更好的基准只会缩短已有类的无量期、把差距拉得更大。判定形式：若新一代基准建成后，被它首次读到的能力中有相当比例在基准建成前就已存在而无人计数，则本文对；若首次读到的能力确实是新出现的，则它们对。

三 · 常规科学与反常（Kuhn 1962，科学哲学）。**最近的邻居之一。** 库恩说反常在旧范式里读不出来，直到危机积累。分界在于：库恩关心的是范式转换的**逻辑**，本文关心的是转换之前那段时间**由谁供养**——本文比它多一条：读不出来的那一段有长度，而这个长度是可测的、并且是可被制度改变的。判定形式：若两个同等规模的领域，其反常被识别的滞后时间与"有没有位置为无读数期买单"无关，则本文这一条错。

四 · 睡美人文献与延迟承认（Ke, Ferrara, Radicchi & Flammini 2015, PNAS，科学计量学）。**这是本文最危险的方法学占位者**，因为它已经有一套成熟的指标：睡美人系数，用来量一篇论文沉睡多久才被引用。分界在于：睡美人测的是**发表之后**的引用滞后，本文测的是**开工之时**有没有现成读数。二者在同一件工作上可以完全相反——一篇论文可以引用即刻爆发，而它当初开工时并无任何现成指标（无量期长而系数低）；也可以开工时指标齐全，却因为无人关心沉睡二十年（无量期短而系数高）。判定形式：**若睡美人系数与无量期占比高度相关，本文应当被科学计量学的既有指标吸收。** 这两个量在同一批文献上都能算，这条检验现在就可以做，本文最愿意拿它去赌。

五 · 探索与利用的权衡（March 1991，组织学习）。它假定二者是同一个资源池上的配置选择。分界在于：本文说二者的**显形速度不同**，所以即使配置比例被规定死，账面回报仍会系统性偏向利用。判定形

式：在探索与利用投入比例被强制固定的组织里，若账面回报仍系统性偏向利用，则本文对；若回报差随比例固定而消失，则它对。

六 · 技术就绪度等级（NASA / 欧盟框架，工程管理，**方法学占位者**）。它已经把"离可评估还有多远"做成了九级量表。分界在于：这套量表测的是**离应用的距离**，本文测的是**离第一个读数的距离**。二者可以背离：一个就绪度很低的基础研究方向可以指标齐全，一个就绪度很高的产品可以没有任何人知道该怎么衡量它好不好。判定形式：取一批项目同时打两个分，若两者高度相关，本文应被它吸收。

七 · 不可通约性（Kuhn / Feyerabend，科学哲学）。本文**不主张**范式之间不可翻译。本文只主张翻译需要时间，而这段时间无人记账。分界的判定形式：若翻译在原理上不可能，则供养这段时间毫无意义，本文的处方失效；若翻译只是慢，则本文成立。这一条本文站在乐观一侧，因此可被悲观一侧证伪。

八 · 马太效应与累积优势（Merton 1968，科学社会学）。它解释的是**已经显形之后**的放大：得到承认的更容易得到更多承认。本文处理的是显形之前。判定形式：若把初始识别的时点控制住之后，资源分配的偏斜完全消失，则累积优势足够，本文多余；若控制之后偏斜仍在（因为有一类从未有过初始识别），则本文对。

九 · 注意力稀缺（Simon 1971，经济学与认知科学）。它说稀缺的是接收者的注意力。分界在于：本文说稀缺的是**愿意在没有读数的情况下继续供养的那个位置**。注意力可以被外包给机器，供养不能——一个模型可以读完所有立项书，但它不能承担"这三年什么都报不出来"的后果。

十 · 反转模板专查。把本文的命题去掉本领域的一切名词、只留形式，得到的是："**一样东西在取得读数之前不进入任何账，因而系统性地被资源抛弃**。" 这不是"越成功越失败"那一类反转模板，它是**遗漏型**——与本栏《认领物》同型，都是缺栏位。仍要点名最贴近的反转模板：**成功陷阱**（Levinthal & March 1993）与**能力刚性**（Leonard-Barton 1992）。分离线：那两条的机制在**组织被过去的成功锁死**；本文的机制在**账本的时间分辨率**——一个从未成功过、毫无路径依赖的新组织，只要它按季度看数，同样会抛弃长无量期的事。判定形式：若空白组织与老牌组织在无量期占比上没有差别，则本文的机制成立而锁死不是必要条件。

十条里最近的是第四条。本文之所以仍不能被它吸收，只有一个理由：**睡美人的前提是那篇论文已经被写出来并发表了**。本文说的正是那些在被写出来之前就停了的——它们不会睡，它们不曾醒过。

十二 · 与本栏同题六篇的分界

本栏最近有六篇处理同一道题——人工智能时代的内卷困境与出路——而且用的是同样三个学科。四篇与本文各自缺的位置不同，逐一分开，并各给一条判决性对照。

《**认领物**》说：内卷是产出的构成中"认领物"占比上升——为使一件产出可被归到某人名下而必须一并被生产的那一层，三家账本只给它成本位、不给产出位。它与本文同为**缺栏位**型，但缺的东西不同：它缺的是**一层已经被生产出来的东西**，本文缺的是**一段还没有产出读数的时间**。前者是实体，后者是时段。判决性对照：取一个归属完全免费的组织——公司内部代码仓库，每次提交由身份系统自动记名，不需要任何

额外的证明工作。它预测这里认领占比接近零、账面内卷轻；本文预测只要该组织按季度用现成指标看数，长无量期的事照样被抛弃。**两个读数正交，可在同一批组织上同时测。**

《**这道关拦下过谁**》说：可读门槛整体饱和，而没有位置有权宣布一道关作废。它管的是**已经走到关前面**的东西；本文管的是**从来没走到任何一道关前面**的东西——它们在开工阶段就因为没有读数而停了。判决性对照：把一个组织所有饱和的关全部撤掉（它的第一步），它预测代价下降；本文预测**无量期占比不变**——撤关不产生任何新的供养。两篇因此互补：撤关减代价，供养无量期才增新。

《**两种「不」**》说：没有人看过就作出的否决，与看过之后作出的否决，在所有记录上同码。它管**被否决而无人看过**的东西；本文管**从未进入判定**的东西。前者进过流程，后者连提交都提交不进去（见第七节那本期刊的例子）。判决性对照：把未阅率降到零（它的第一步全部落实），它预测投递者处境改善；本文预测**无量期占比仍不变**——被认真看过的，仍然只是那些有现成读数可判的。

《**被丢掉的那一眼**》说：被判定过而未被接受的产出在无人记账处合成一条判定基线。它与本文恰好互斥：暗基线的前提是**有一条线**，本文说的是**没有线**的那一段。判决性对照：一个把参照集轮换做到极致的机构（它的出路），它预测抬升度回落；本文预测无量期占比不动——轮换参照集只是换用哪些旧样本，不产生任何新的参照。

《**认不出来，就得自己交代**》说：核实这份工作从评价一方移到了被评价一方，由此产生一类与真实产出同形、因而不可分离计量的劳动。它与《认领物》形状很近，落点在**转移**：谁来做核实。它管的是**已经被做出来的那一层劳动**；本文管的是**还没有任何量能测到的那一段时间**。判决性对照：取一个核实完全由评价方承担、被评价方不需要任何额外交代的组织——自动身份核验加抽样复核。它预测这里同形劳动接近于零；本文预测只要该组织按现成指标看数，长无量期的事照样断供。两个读数在同一批组织上正交。

《**一件是几件**》说：生成成本趋零之后计数单位失去下界，而"并项"这个单位从未被生产出来。它与本文的关系是**顺序上的前后**，值得写清楚：它处理的是**已经能被计数、只是单位不清**的那一类；本文处理的是**连"几件"这个问题都还问不出来**的那一类——那不是单位不清，是没有单位。**必须先有第一个读数，才谈得上单位**，所以本文所说的那一段落在它的定义域之前。判决性对照：一个把计数单位重新定死（引入并项）的组织，它预测计数恢复可比、账面内卷缓解；本文预测无量期占比不动——定死单位只作用于已经能被计数的那一类，对还没有任何量能测到的那一类毫无影响。

六篇与本文合起来，可以写成一句：**《认领物》缺的是一层东西，《这道关拦下过谁》缺的是一个位置，《两种「不」》缺的是一种处置类型，《被丢掉的那一眼》缺的是一份账，《认不出来，就得自己交代》缺的是一次转移的去处，《一件是几件》缺的是一个计数单位，本文缺的是一段时间。** 七处缺口互不重叠，而它们都长在同一块地上——下一节说这块地是什么。

十三 · 这个判断底下还站着一块更大的地

把上一节那十位邻居逐个问一句"它把什么当作给定，因此结构性地看不见什么"，会得到一份整齐得让人不安的清单。判据只有一条：一旦承认了那件看不见的事，它自己就塌。

- **研究生产率**把"产出可被代理"当作给定，因此看不见没有代理的产出。**一旦承认有这样的产出，研究生产率的分子就没有定义。**
- **基准研究**把"能力可被题目化"当作给定，因此看不见不能被题目化的能力。一旦承认，"造更好的基准"这个处方就失去了它的普遍性——它只能改善已经在题目范围内的那一部分。
- **睡美人指标**把"那篇论文已经被写出来并发表了"当作给定，因此看不见没写出来的。一旦承认，它测的就不是"延迟承认"而只是"发表后的引用动力学"。
- **常规科学**把"共同体已经在场"当作给定，因此看不见共同体形成之前的那一段。
- **技术就绪度**把"目标应用已知"当作给定，因此看不见不知道自己要用在哪的东西。一旦承认，一级的定义就落空——第一级是"基本原理被观察到并报告"，而"报告"预设了报告给谁、按什么格式。
- **马太效应**把"至少有过一次初始识别"当作给定，因此看不见零次识别的。
- **注意力稀缺**把"东西已经在接收者面前"当作给定，因此看不见还没有任何渠道能把它送到任何人面前的那一类。

七行背后是同一件被当作给定的事：

> **凡是一样东西，从它存在的那一刻起，就至少有一个量在测它。**

这句话没有人主张过，因为没有人需要主张它——所有人都站在它上面，包括互相批评的各派。而承认它的反面——**一样东西可以存在一段时间，而任何量都测不到它**——会让上面这一整批工具同时失去定义域。

这也是本文与本栏同题四篇真正的关系。那六篇各自推翻的是同一块地基上的另一个假定：一件东西的生产者是给定的；每一道判定都有一个持有它的位置；每一次处置都有人作出过它；凡是一样东西要么被接受要么退出系统；核实这份工作总归有一方在做；凡是产出总能被数成几件。**七个假定形状同族，都是「总有一个……」。** 七篇合起来指出的不是七个问题，是同一块地基的七个承重点。而这块地基之所以从来没有被回头看过，正因为按量运行的系统只能看见它自己能测到的东西——它没有任何机制去问"有没有我测不到的"。

十四 · 三家为何各自到不了这一条

三家没有走到这一步，不是因为疏忽。**一个领域看不见什么，通常不是它的缺陷，是它的问题结构的影子。**

工程学的核心任务是**把判断变成可传递的证据**。在这个任务结构里，一样"还没有任何量能测到"的东西不是一个对象——它甚至不是一个待解决的问题，因为工程学的问题必须以"给定规格，达成它"的形式提

出。没有规格就没有问题。所以工程学看得见尺子不好，看不见"还没有尺子的那一段"；它把前者当作技术债，把后者当作还没轮到自己管的事。

经济学的核心任务是**配置**：把有限资源分到收益更高的地方去。配置分析总是需要一个可以比较的量，因此**没有量的选项在配置问题里不是"收益未知的选项"，它根本不是一个选项**——它进不了那张表。经济学因此可以非常精确地讨论不确定性、期权价值、探索的收益，但那些讨论全部预设了"我们知道有这么一个选项，只是不知道它值多少"。本文说的那一类，是**连"有这么一个选项"都还没有人写下来**的。

艺术学的核心任务是**抵抗标准化**。它比谁都早看见量的不足——它已经看了一百年，从复制技术开始动摇独一性的那一刻起。但它的任务结构使它把出路一律指向**否认量的资格**，而不是**去问那段还没有量的时间由谁供养**。为一段无读数的时间设一个供养位置，在这一支的语汇里天然读作"又一次把创作纳入管理"，因此它不会去设计它。

三家的盲点不是三种疏忽，是三种任务结构各自的背面。而这三个背面恰好覆盖了同一段时间。

十五 · 一旦被制度采用会发生什么

本文提出的读数一旦被采用，最省事的实现方式是把"无量期占比"做成一个考核指标。而一旦它进了考核，让这个数好看的最便宜办法不是去做没有尺子的事，是**在立项书的"评价指标"一栏改填"待建立"**——一个字段的事，一个下午可以把全院的立项书改一遍。

于是这个数暴涨，而组织的实际构成一动没动。更坏的是，它同时污染了唯一一份能用来检验本文的材料：那些立项书原本是可信的，因为没有人有动机在那一栏上撒谎。

这条反噬我看不到出口。 任何核查都需要有人判断"这个指标算不算现成的"，而这个判断本身没有现成的量——它正好落在本文所描述的那一类里。也就是说，**本文的读数在原理上不能自我保护，而且它比它所要保护的那一类东西更容易被伪造**：伪造它的无量期是零（改一个字段立刻显形），而它要保护的那一类的无量期是几年。

本文能给出的唯一防护很弱：**这个数只统计、不考核**，并且统计口径与统计者必须与被统计单位分离。这不解决问题，只是把伪造的成本从零抬到不为零。写下来，是为了不让读者以为本文没看见这一层。

十六 · 判错条件

先写出最强的竞争解释，因为不排除它，下面的条件都不算数。那句最朴素也最难反驳的话是：

> **"其实不就是短视吗？不就是资本没耐心、考核太急吗？"**

如果问题在时间偏好，那么本文的第二根轴（有没有人供养）就是多余的——供养与否只是时间偏好的另一种说法，而处方只需要延长考核周期。下面第一条就是冲着它去的。

一 · 控制组织的时间偏好之后——把考核周期长度、资金期限、管理者任期这三项控制住——若"无量期占比"与"三年后产生的首创性成果比例"之间不存在剂量—反应关系，则本判断为伪，届时应归因于时间偏好而非读数结构。样本纪律：至少六个同制度环境内的单元（同一资助体系下的六个以上院系或实验室，或同一集团下的六个以上事业部），在学科领域、资金规模、人员规模上同质。**不得拿两个国家或两个体制的对比充数**——那种对比在方法上支持不了任何"控制掉某变量"的说法，只能作启发式例证。

二 · 若两个考核周期完全相同的组织，其无量期占比的差异不能预测首创性差异，则第二根轴不成立，四格退化为一维，本文只剩下一条关于时间偏好的旧话。

三 · 若在一个明确为长无量期项目设立了独立通道的机构里，通道设立后三年内该机构的无量期占比没有上升，则本文的处方无效——尽管第五节的诊断仍可独立成立。

四 · 若睡美人系数与无量期占比在同一批文献上高度相关，本文应当被科学计量学的既有指标吸收。这两个量现在就能算，这是本文最便宜也最危险的一条。

五 · 顺序读数：本文预测**评价周期的缩短早于**无量期占比的下降，而不是同时或滞后。若在时间序列上观察到相反的次序——先是无量期占比下降，然后组织才缩短周期——则本文的因果方向错。顺序比相关难被巧合满足，这一条本文最愿意拿去赌。

六 · 一条低成本可实做的：取任一机构最近一百份立项书，清点"评价指标"一栏在开工时是否已经存在。数据在既有档案里，一个下午可以做完。若各学科的分布与本文预测的"生成工具普及越快、该比例越低"不符，本文在那个行当里不成立。

七 · 一条会推翻本文核心机制方向的：若在生成成本塌陷幅度最大的领域，无量期占比**反而上升**——因为塌陷把人力从常规工作里释放出来，让更多人去做没有尺子的事——则本文的方向判反了。这个反向机制是真实可能的，本文不认为它已经被排除。

这条判断若被证伪，我们会学到什么：若它与时间偏好分不开，那么出路只能在延长考核周期上找——延长任期、拉长基金期限、放宽年度考核。这与本文的处方（为无读数期单独设一个不进考核的通道）指向完全不同的干预，而且前者要贵得多。**证伪本文，会把资源从制度设计推回到期限调整**，这个后果值得一次认真的检验。

分层引用：本文可分三层引用。第一层是"无量期"这个构念作为核心判据，最强也最容易倒；第二层是两轴四格作为分析工具；第三层是一条不依赖前两层的经验主张——**在生成成本大幅下降的行当里，判定流程与立项流程中"开工时已有现成度量"的比例会上升，而这一上升不伴随该流程总环节数的减少**。第三层可以单独被检验、单独被引用；第一层被推翻时它仍然站得住。

十七 · 出路：不是多冒险，是让那一段有账

本文是一条辨别维度，不是一套方案。但一条维度若推不出任何可做的事，它就只是个说法。下面三件，各对一家，而且顺序是死的。

17.1 对经济学：把"造出了一把新尺子"本身记为一项产出

现在的账里，造指标、建基准、开栏目、立门类、写验收标准——这些全部记在成本栏或"科研管理"栏，从来不记为产出。而它们恰恰是**缩短别人无量期**的那件事：一把新尺子造出来，一整类原本读不到的东西同时显形。

它的收益因此全部外溢给后来者，在任何一个单位自己的账上都不划算——造尺子的人拿不到用尺子的人的收益。这是一个标准的外部性问题，处方也是标准的：**单独记账、单独资助，并且不要求承担者同时产出该领域的成果**。　现在的做法正相反：谁想建一个新指标，得先用它做出一批成果来证明它有用；而做出成果又要靠这个指标，循环因此闭合。

最小动作：在成果统计口径里加一类"评价工具"，与论文、专利并列。这一项现在在几乎所有统计体系里都不存在。

17.2 对工程学：在验收之外单设一个不产出验收结论的环节

工程学的整套流程围绕"给定规格，达成它"组织，无规格的事进不了排期——不是被拒绝，是根本填不进那张表。最小动作不是取消规格，是**允许一部分排期以"本期不产出可验收结果"立项**，并且**这一项不进入交付率的分母**。

关键全在最后半句。只要它进分母，任何团队都不会用它——用一次，交付率就掉一块，而交付率是这个团队被看见的唯一方式。这一条看上去只是会计细节，实际上它是整件事成不成的开关：本文第五节说资源流向是加总规则的算术后果，那么改变流向的动作也只能是改加总规则。

最小动作：在交付率的定义里加一个排除项。一行字。

17.3 对艺术学：供养必须是具名的，而且必须承担后果

艺术学在这一点上比另外两家懂得多。赞助人、导师、编辑约稿、画廊的长期代理——这些制度形状相同：**一个具体的人押上自己的判断，替一件还不能被证明的事扛一段时间**。　押上的是他自己以后说话还算不算数。

这件事不能被外包给评审委员会。委员会需要依据，而依据正是这类事情还没有的东西；把它交给委员会，委员会只能退回去找现成指标，于是绕回原点。它也不能被外包给模型——模型可以读完所有申请，但它不能承担"这三年什么都报不出来"的后果。

最小动作：一个基金里划出一小部分，由**具名的人**独立决定、独立署名，事后公开其全部资助记录**包括失败的那些**；并且**不以命中率考核他**，唯一的约束是他的全部记录被公开。用命中率考核会立刻把他推回现成指标——那正是本文要避开的那条路。

17.4 顺序为什么是死的

第一件必须最先。　不给造尺子记账，另外两件产生出来的东西仍然没有读数，仍然会在下一次结算时从账上消失——你把它们做出来了，而它们仍然不构成任何一项。

第二件其次。 它是把第一件的结果装进日常流程的接口。

第三件最后。 它最难，也最容易被误用成任人唯亲；必须等前两件把"什么算数"的账先建起来，具名供养才有可被公开检验的背景。顺序颠倒的后果很具体：先做第三件，会得到一批无法被任何人核对的人情资助，两年之内它会被叫停，并且从此把整个方向的信誉一起赔掉。

17.5 一条反直觉的推论

从这三件里落出来一条与所有管理直觉相反的话：

> **在造尺子这件事没有被单独记账之前，任何提高评价质量的努力都会拉大差距。**

更好的基准、更细的指标、更及时的反馈、更短的复盘周期——每一项都在让**已经有尺子的那一类**显形得更快，而对**还没有尺子的那一类**毫无作用。评价改进因此是单向的：它只加速一侧。这不是说不该改进评价，是说**改进评价必须与第一件事同时做**，否则它的净效果是把断供那一格扩大。

可检验：**一个刚刚完成评价体系升级的组织，其无量期占比会在两个考核周期内下降。** 升级的时点在制度文件里有据可查，占比可从立项书清点，这条现在就能跑。

17.6 为什么不是"多给钱"

最常见的处方是加大投入。本文与它有直接分歧：**钱会流向有读数的那一类。** 在一个所有项目都要报量化结果的体系里，增加预算的结果是同一类项目变多，而不是另一类项目出现——因为多出来的钱仍然要按同一套规则分配。

这条分歧是可检验的，而且数据现成：**科研经费大幅增长的时期，无量期占比应当不变或下降，而不是上升。** 若观察到经费增长同时伴随该比例上升，本文这一条错，加大投入本身就够。

十八 · 不适用的边界

不适用于以产出可靠结果为职能的单元。 安全、合规、质量、临床推进——那里无量期短是目标，本文的读数方向相反。

不适用于判定单独一件事的价值。 无量期长是"新"的必要条件而非充分条件；一个无人问津的方向与一个尚未被识别的方向，在这个读数上完全一样。用它去判单件，等于把统计性质当成个体属性。

不适用于把它当作提高冒险程度的理由。 本文不主张组织应当整体提高无量期占比。它主张的只有一件事：**一个组织应当知道自己这个数是多少，并且知道自己有没有为长无量期的那一类留出位置。**

本文不主张新的一定比旧的好。 大量长无量期的事最后什么也不是，这不是失败，是这一类事情的正常分布。本文只主张：它们在停下来的时候不产生任何信号，因此**没有人能知道停掉的是哪一类。**

十九 · 伦理与执行边界

无量期占比是一个几乎必然会被拿去考核人的读数。三条硬约束写在这里，缺一条这个读数就不该被使用。

其一，这个读数指的是构成，不是人。 一个单元的无量期占比低，是它的职能与流程的性质，不是研究者不敢想。把它算到个人头上是误用，而且按本文自己的机制，一个最有想法的人在一个只接受现成指标的流程里也只能立有指标的项目。

其二，不得据此对个人作出不利安排。 尤其不得把"你的项目开工时没有现成指标"用作否定的理由，也不得反过来把它用作加分项——后者会立刻触发第十五节那条反噬。

其三，若一个机构决定统计这个数，统计口径与统计者必须与被统计单位分离，且只统计、不考核。 一个由被统计者自己填报并进入自己考核的读数，在本文的框架里是一个无量期为零的东西——它会在填报的当天显形，而它要保护的那一类要等几年。

二十 · 结论，与一条写死日期的赌注

把全文压成一句：**一样东西在取得第一个读数之前，在所有账本上都是零；人工智能把已有读数那一类的无量期压到接近于零，却没有压缩需要先造尺子的那一类，于是两类的显形速度差被拉开了数量级，而资源被算术地——不是被偏好地——抽向短的那一侧。**

由此得到的不是一条处方，是一条辨别维度：**要分辨一个领域是在长新东西还是在把同一样东西做得更多更快，不能看它的产出、效率或承认，只能看它最近做完的东西里有多大比例在开工时还没有任何现成的量可以判定它。**

若一定要从这条维度推出一个动作，最小的那个是：在立项文件里加一个字段——"本项目开工时可用的现成评价指标：有／无"——**只统计，不考核，统计者与被统计单位分离。**这不解决问题，它只是让那个数第一次存在。而按本文的整个论证，一样东西的第一个读数出现，正是它开始能被讨论的那一刻。

最后交一条写死日期的赌注：

> **到二〇三一年六月三十日，在至少三个国家的公共科研资助体系中，会出现把"该方向目前不存在公认的评价指标"本身作为一项正面立项理由、写进公开指南条款的做法。**

什么不算命中，一并写死：写"鼓励原创""支持非共识""容忍失败"这类倡导性表述不算；设立了"探索性项目"类别但仍要求填写量化考核指标不算；只在评审内部口头掌握而未写进公开指南不算；把它写进白皮书或战略报告而未落到具体计划的申报条款里不算。

若到期没有出现，本文关于"为无读数期设通道"的处方在制度上不可行，应当被放弃——而第五节的诊断与第十六节第三层的经验主张仍可独立成立。这正是本文采用分层写法的原因。

附 · 本文自己的无量期是多少

接近于零。

本文用的是既有的学术论文形式，投向一个有现成评审细则的通道；它在写完的那一刻就能被现成的量读到——字数、格式、引用、审稿意见、阅读量。**因此本文不是它所主张要保护的那一类东西的标本，它站在靶格的反面。**

这个定位有两个具体后果，都不好听。

第一，本文没有资格用自己作为证据。 它不能说"你看，我这样的工作也很难被看见"——恰恰相反，它很容易被看见，这正是它能被写出来的原因。它只能靠一件事被检验：有人去清点一次立项书，并把结果公开。

第二，也是本文无法弥补的一个洞：如果本文的判断是对的，那么真正落在断供那一格里的工作**恰恰写不出这样一篇论文**——因为写出来、投出去、被读到，就意味着它已经取得了读数。本文能做的最多是指出那一段的存在，**而不能从那一段里拿出任何一个样本**。所有能被举出来的例子，都是最终出来了的那些；那些没出来的，按定义不留下痕迹。

这不是谦辞，是本文自己的判据用在本文身上得到的读数：**一个关于"看不见的东西"的论证，其全部证据都只能来自看得见的那一侧。** 本文承认这个洞，并且认为任何声称补上了它的说法都值得怀疑。

注 释

一 · 本文所说的"读数"取其最宽的意义：任何一个现成的、可被第三方使用的量或类别——一个指标、一个基准、一个栏目、一个门类、一条验收标准。它不要求这个量是好的，只要求它存在且可用。一个粗糙的量存在，也算取得了读数。

二 · "无量期"的起点取"被做出来"而非"被想到"，因为想到的时点在原理上不可核对，而做出来的时点在多数记录里可以定位（立项、开工、首次提交）。这是一个为了可查而付出的精度代价，本文明确接受它。

三 · 第八节的比例阈值没有设固定值，因为"现成"是一个二分判断而非连续量。跨机构比较时必须使用同一套判断规则并注明，否则数字之间不可比。

四 · 本文引用的经济学读数（研究生产率下降）与本文的判断并不矛盾：本文不主张它测错了，只主张它不能单独分辨"没长出新东西"与"长出来了但还没有代理"。

五 · 本文为理论建构与研究设计，不报告任何虚构的样本、统计结果或实验结论。文中所有数字均来自公开来源，出处见参考文献。

参考文献

Akhtar, M. et al. (2026). When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation. arXiv:2602.16763.

Auslander, P. (2023). Liveness: Performance in a Mediatized Culture, 3rd edition. Routledge.

Auslander, P. (2012). Digital Liveness: A Historico-Philosophical Perspective. PAJ: A Journal of Performance and Art, 34(3), 3–11.

Bean, A. et al. (2025). 对 445 篇基准论文的系统检查（不确定性估计与统计检验的使用比例）。

Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are Ideas Getting Harder to Find? American Economic Review, 110(4), 1104–1144. doi:10.1257/aer.20180338.

Bommasani, R., Liang, P. et al. (2023). Holistic Evaluation of Language Models. Transactions on Machine Learning Research.

Eriksson, M. et al. (2025). Can We Trust AI Benchmarks? European Commission Joint Research Centre.

Feyerabend, P. (1975). Against Method. New Left Books.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and Identifying Sleeping Beauties in Science. PNAS, 112(24), 7426–7431.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Leonard-Barton, D. (1992). Core Capabilities and Core Rigidities. Strategic Management Journal, 13(S1), 111–125.

Levinthal, D. A., & March, J. G. (1993). The Myopia of Learning. Strategic Management Journal, 14(S2), 95–112.

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. Organization Science, 2(1), 71–87.

Merton, R. K. (1968). The Matthew Effect in Science. Science, 159(3810), 56–63.

Phelan, P. (1993). Unmarked: The Politics of Performance. Routledge.

Salaudeen, O. et al. (2025). 关于基准构建的原则性框架与构念效度。

Simon, H. A. (1971). Designing Organizations for an Information-Rich World. In M. Greenberger (ed.), Computers, Communications, and the Public Interest. Johns Hopkins University Press.

技术就绪度等级（Technology Readiness Levels）：NASA 与欧盟研究框架计划所采用的九级量表。

本文由三个分属不同学科、且互相冲突的理论体系碰撞而成。 全部来源均为站外公开文献，可自行核对。 作者：王德生 + Claude · SDE Universes · 学科通融