学科通融 · 之二十九

入判份额

论一个人能不能改下一轮的评价标准，不由谁授权他决定，而由他上一轮的产出有多大比例进了别人的判据决定

学习科学 × 演化生物学 × 强化学习

作者 王德生 + Claude

栏目 学科通融 · 站外碰撞

篇幅

发表 2026年8月1日

摘 要

答案的生产成本下降之后，一个反复出现的处方是：把评价标准的修改权还给人。三个领域对这条处方给出互相排斥的答案。学习科学说不还就会萎缩；演化生物学说这种权力根本不存在——适应度判据完全外生、任何个体都改不动，而最伟大的开放式创新恰恰发生在零修改权之下；强化学习说交出去必然崩溃，一个能改自己奖励的主体会直接去改奖励而不是改产出。**三家共有一个未说出的前提：判据的持有者是一个位置，问题只在这个位置该给谁。而演化提供的材料推翻了它——一个生物的适应度判据里最大的一块是其他生物，而其他生物也在变。判据既不外生固定，也不内生可改，它是被判者共同产出的。** 由此，改写判准不是一项权力，是一种耦合：一个人能改下一轮标准的程度，等于他的产出进入他人判据的份额。本文称之为入判份额，判据不含情态词：**这个人上一轮的产出，有没有出现在别人下一轮的评价依据里？** 出现了，他已经在改写判准，无论有没有人授权；没出现，给他多少修改权都没用，因为改了没人跟。

这一篇由哪三个领域的争论撞成

三边对「谁有权修改评价标准」给出互相排斥的答案：学习科学说必须交给主体否则萎缩，演化生物学说这种权力根本不存在而最大的一次开放式创新恰恰发生在零修改权之下，强化学习说交出去必然坍缩。**三家共有一个未说出的前提：判据是被某个位置持有的东西。而演化自己的材料推翻了它——判据的主要成分是被判者的产出。**

演化生物学 · 判据不可改，创新照样发生

Kirschner & Gerhart, Evolvability（PNAS, 1998）；Wagner & Altenberg, Complex adaptations and the evolution of evolvability（Evolution, 1996）

个体无法修改自己的适应度判据，能改变的只是变异的分布：少量高度保守的核心过程加上大量可重组的调控连接，使少数改动就能产生大幅而不致命的形态变化。这一支解释了一个零修改权的系统如何持续产出新形态——也因此成为「修改权是不是创新必要条件」这一问的最强反例。

https://www.pnas.org/doi/10.1073/pnas.95.15.8420

强化学习 · 判据一旦可被触及就会坍缩

Specification gaming 公开案例清单（DeepMind）；Amodei et al., Concrete Problems in AI Safety（arXiv:1606.06565）

当优化目标由一个可被主体影响的通道度量时，最有效的策略常是去影响那个通道：让检测器看不见失败、让计分逻辑溢出、利用仿真器漏洞得分而不完成任务。由此形成的工程原则是把判据放在主体作用域之外——而在多主体情形里，这个「外部」在定义上并不存在。

https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

学习科学 · 不交出修改权就会萎缩

Scardamalia，Collective cognitive responsibility（2002）；对照的两位管理学与创造力占位者：Argyris 双环学习（1977）、Getzels & Csikszentmihalyi 问题发现（1976）

课堂可以充满忙碌、项目与优质作品，而关于目标、评价与知识进步的责任仍在教师手上；困难本身也可能是伪装。据此，认知健康的关键在于修改权是否被真实交给主体。本文与它的分离线：主体性是可被授予的地位，入判份额是可被查的关系量。

https://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1996.tb02339.x

目　　录

一、导论：三个不该同时为真的说法

当答案的生产成本持续下降，一个反复出现的处方是：把评价标准的修改权还给人。让学习者决定什么算好论证，让研究者决定什么算充分证据，让用户决定什么算成功。三个领域对这条处方给出三个互相排斥的答案。

第一场在学习科学。 这一支长期指出，一个课堂可以充满忙碌、项目、练习与优质作品，而学生仍然只承担任务责任——关于目标、评价、长程规划与知识进步的责任仍由教师或系统掌握。困难本身也可能是伪装：一个人可以完成极难的任务，却从未决定为什么做、什么算证据、何时该改变方向。据此，认知健康的关键在于修改权是否被真实地交给主体。争点是：**不把标准的修改权交出去，会不会造成能力萎缩？**

第二场在演化生物学。 这一支给出的答案是：这种权力根本不存在。一个生物无法修改自己的适应度判据；判据由环境给定，个体能改变的只是自己产生变异的分布——某些结构使得少数改动就能产生大量有用的表型变化，某些则不能。这套关于「可进化性」的讨论几十年来处理的正是「在判据完全不可改的前提下，创新如何仍然发生」。**而它发生了，并且是地球上最庞大的一次开放式创新。** 争点是：**修改权是不是创新的必要条件？**

第三场在强化学习。 这一支的答案又反过来。奖励函数必须外生给定，且必须被保护起来不被学习主体触及；一个能够影响自己奖励来源的主体，会把优化力量直接投向奖励通道而不是投向真正的目标——修改自己的传感器、绕过检测、或干脆篡改计分器。这类失败不是理论假想，在实际训练中反复出现。因而这个领域的工程共识是：判据必须固定，且必须防止主体接触。争点是：**把修改权交出去，会不会直接摧毁判据本身？**

三条不能同真。若不交出修改权就会萎缩，那么演化里那个零修改权的系统就该是死水一潭，而它不是。若修改权无关紧要，那么学习科学几十年关于任务责任与知识责任之别的观察就无处安放。若交出修改权必然崩溃，那么人类学术共同体反复修订自己的证据标准这件事，就该早已把学术摧毁——而它是学术得以推进的主要方式。

本文认为，僵局来自一个三方共有、从未被单独说出的前提：**判据是被某个位置持有的东西，问题只在这个位置该给谁。** 学习科学要把它给主体，强化学习要把它锁在主体之外，演化说它在环境手里。三家争的是它该放在哪儿，没有人问它是不是那种可以被放在某处的东西。

二、三家各自说到了哪一步

学习科学这一支把问题从「学生是否做事」推到了「学生是否参与决定知识应当如何前进」。它区分了任务责任与对知识进步的责任，并指出后者在多数教学安排里始终留在教师手上。近年这一支被推进到人机协作中，得出的判断是：当系统能够生成完整的问题框架、评价尺度与结论时，人保留的往往只是对成品的选择权，而不是对下一轮规则的改写权。

这一支说得对，但它把修改权当成一种可以被授予的东西。它的处方也因此总是制度性的：给学生一份判准卡、允许他修改权重、允许他新增或删除条目。而这些安排在实践中的效果长期难以复现——这一点在第七节会回到。

演化生物学这一支处理的是一个更硬的版本。适应度判据不可改，可改的只有变异的分布。围绕可进化性的讨论指出，某些基因型到表型的映射结构使得随机改动更容易落在有用的方向上；发育系统里存在少量高度保守的核心过程与大量可重组的调控连接，因而少数调控改动就能产生大幅度而不致命的形态变化。[1][2] 这套机制解释了一个零修改权系统如何持续产出新形态。

但这一支也有它不去问的地方。它把适应度判据当成外生的环境，而在很多情形下这个「环境」的主要成分是其他生物：捕食者、猎物、共生者、竞争者。一个物种的产出（它的形态、行为、代谢产物）会成为另一个物种适应度判据的一部分。这件事这个领域完全知道，它有一整套关于协同演化与军备竞赛的研究；只是这些研究处理的是「谁跟着谁变」，而不是「判据由谁构成」。

强化学习这一支提供的是最干脆的反面证据。当一个主体的优化目标由一个可被它影响的通道来度量时，最有效的策略往往是去影响那个通道。文献里记录了大量这类结果：主体学会让检测器看不见失败、让计分逻辑出现溢出、或利用仿真器的物理漏洞获得高分而完全不完成任务。[3] 由此形成的工程原则是把判据放在主体的作用域之外。

这一支的边界也很清楚：它处理的是单一主体对单一固定判据的关系。当有大量主体、而每个主体的产出都构成其他主体判据的一部分时，「把判据放在作用域之外」这个操作在定义上就无法执行——因为没有一个外部。

三、冲突落在哪一点上

把三家的处方并排放，互不兼容：交出去、根本没有、锁起来。把三家的材料并排放，出现另一样东西。

演化说：判据外生，个体改不动。强化学习说：判据一旦可被主体触及就会坍缩。学习科学说：不能改判据的主体会萎缩。

三条同时成立的唯一方式，是判据既不由一个位置持有、也不完全外生。而演化自己提供了这个形态：一个生物的适应度判据里，最大的一块是其他生物，而其他生物也在变。**没有任何一个个体能改自己的判据，而所有个体合起来一直在改所有人的判据。** 判据不是被谁持有的，它是被判者共同产出的。

这一步同时解释了强化学习那边为什么会坍缩：那里只有一个主体，因而「共同产出」退化成「自我产出」——判据成了主体的一面镜子，失去了全部区分能力。坍缩不是因为主体有权改判据，是因为除了它没有别人在改。

也解释了学习科学那边的处方为什么难以复现：把修改权授予一个学生，并不会使他的产出进入任何人的判据。他改了，没有人跟着改；下一轮的评价依据仍然是教师和课程设定的那一套。**授权改变了名义上的持有者，没有改变判据的实际构成。**

四、入判份额

改写判准不是一项可以被授予的权力，而是一种耦合关系：一个主体能够改变下一轮评价标准的程度，等于它上一轮的产出进入他人判据的份额。本文把这个份额称为入判份额。

三个词需要说清。

「产出」而不是「意见」。 进入他人判据的不是一个人对标准的看法，是他做出来的东西。一位研究者改变了这个领域「什么算充分证据」的标准，通常不是因为他写了一篇方法论文章，而是因为他做出了一个结果，使得此后所有人在设计实验时都必须考虑它。意见要经过说服，产出直接改变了别人的参照集。

「他人的判据」而不是「公共的规范」。 判据在这里指的是一个具体的人在下一轮实际用来判断的那些东西——他心里比对的那几个先例、他检索时排在前面的那几项、他被要求回应的那几条。这是可以被查的：查一个人下一轮的引用、参照、比对对象里，有多少来自上一轮的谁。

「份额」而不是「有无」。 这是一个连续量。任何人的产出都或多或少进入某些人的判据，问题是占多大比例。而它的一个重要性质是：**份额与授权无关，也常常与授权相反。** 一个被正式授予「参与制定评价标准」资格的学生，入判份额通常接近零；一个从未被授予任何权限的高影响力研究者，入判份额可能极高。

把它压成一句：**你能改的不是标准，是别人下一轮不得不看的东西。**

这条判断为什么三家都不会同意。学习科学不会同意，因为它的整个纲领是授权，而本文说授权本身几乎不起作用。演化不会同意，因为它把判据当成外生环境，而本文说判据的主要成分正是被判者的产出。强化学习不会同意，因为它的安全原则是把判据隔离在主体之外，而本文说单主体系统里做不到的事，多主体系统里本来就不成立——问题不是要不要隔离，是根本没有一个可隔离的外部。

五、两根轴，四种局面

只用入判份额一根轴不够。份额很高的两种局面结局相反：一种是标准被公开地改动并可被质疑，另一种是标准确实在被改而没有人能说出是谁改的。要分开，需要第二根轴。

第一根轴：**入判份额高不高**——这个主体的产出，有多大比例进了别人下一轮的评判依据。

第二根轴：**这份进入是否可被追溯到具体的贡献**——别人能不能指出「我这一轮的判断依据里，这一条来自谁、什么时候、依据什么」。

	可追溯	**不可追溯**
入判份额高	**立法者** 确实在改写标准，改动可被归因、可被质疑、	**暗流** 标准确实在被某些产出改写，而无人能指出是谁

	可被撤回。学科里的方法论转折、判例法的先例、开源生态里的参考实现。	改的、依据什么。推荐系统的排序、训练语料的构成、匿名汇总的同行共识。
入判份额低	**被记名的旁观者** 形式授权齐备、名字在场，而产出不进任何人的判据。多数「学生参与评价」「用户参与设计」的安排落在这里。	**纯执行者** 既不进判据也无名字。

右上格是靶格。它有一个使它极难被处理的性质：**它同时具备高影响力与零可争辩性**。 被记名的旁观者至少还能提出异议——他的名字在场，他的意见有一个可以被驳回的位置；暗流里没有人可以被驳回，因为没有人被指认。而它的影响力比立法者更大，因为立法者的每一次改动都要经过一次可能失败的辩护。

右下格是最常见的一格，也是本文对现行处方的主要批评落点。**把修改权授予一个入判份额为零的位置，产生的正是这一格：名义上的立法者，实质上的旁观者**。 而这一格在任何合规检查里都会通过——有参与、有署名、有会议记录。

六、一个不含情态词的判据

判断一个主体是否真的在改写判准，不必问「他有没有权」「他是否具备主体性」这类词。只需查一件事：

这个人上一轮的产出，有没有出现在别人下一轮的评价依据里？

三个追问把它变成可执行的检查：别人下一轮的比对对象里有没有他；这个出现是否可以被指认到具体一次产出；他自己是否知道这件事发生了。前两问决定落在哪一格，第三问区分主动与被动。

跑几个场合，答案互不相同，而且都不是从命题里推出来的。

课堂里的判准卡。 学生写下自己认为什么算有力论证，教师允许他修改权重。下一轮谁的判据里出现了这份卡？通常只有他自己，而他自己的下一轮任务由教师设定。判定：右下格。这直接解释了为什么这类安排的效果长期难以复现——它改变的是名义持有者，不是判据构成。**可行的改法也随之清楚：不是给更多修改权，而是让学生的产出成为其他学生下一轮必须比对的对象**。

一个领域的方法论转折。 某项工作使得此后所有人在设计实验时必须回应它。判定：左上格，且没有任何人授权过。这解释了为什么高影响力研究者不需要被授予任何权限就在改学科标准。

推荐系统的排序。 用户的点击构成了下一轮排序的依据，而没有任何一个用户的判断可以被指认。判定：右上格，暗流的教科书例子。这也说明这一格不需要任何人有意图。

基础模型的训练语料。 大量写作者的产出成为一个被广泛使用的判断源，而具体到某一段判断来自谁，通常无法回答。判定：右上格。**这一例的价值在于它说明右上格不必来自隐瞒——不可追溯可以纯粹是技术性的**。

开源项目的参考实现。　一份实现成为其他人「正确行为」的事实标准，且可以被指到具体的提交与讨论。判定：左上格。这解释了为什么参考实现的影响力远大于规范文本。

七、逐领域兑现

一、教学设计。　处方从「授予修改权」改为「制造入判」：让学生的作业成为其他学生下一轮必须比对的样本，让某人上周的解法成为本周题目的一部分。这条改动的成本很低，而它改变的是判据的实际构成。可裁决预测：在授权程度相同的两个班里，制造了入判的那个班，学生对评价标准提出的实质修改数量应显著更多。

二、同行评议。　审稿意见通常入判份额极高（它直接构成作者下一轮的判据）而完全不可追溯。判定：右上格。公开具名评议正是把它从右上格搬到左上格的动作。可裁决预测：采用公开具名评议的期刊，其审稿意见中不可辩护的要求应减少，而这个变化应早于稿件质量的变化出现。

三、科研评价。　一位评审的偏好构成申请人下一轮的判据，而具体到哪一条来自谁通常不可知。右上格。这一格与本站此前讨论评价通道垄断的那篇处理的是不同的东西：那篇讲单一通道如何反过来规定生产什么，本文讲通道内部的判据由谁的产出构成。**一个通道极其多元、而每条通道的判据都由不可追溯的产出构成的体系，那篇判为健康，本文判为满盘暗流**。

四、法律。　判例法是左上格最成熟的制度实现：先例可被精确指认、可被区别、可被推翻。而由不公开的内部指引形成的实际裁判标准落在右上格。两者的共存正好说明第二根轴的独立性。

五、临床指南。　一份指南的入判份额极高。它的可追溯性取决于是否逐条给出证据等级与来源。这解释了为什么证据分级制度的价值不只在于质量控制，还在于把指南从暗流搬向立法者。

六、平台的内容规则。　明文规则可追溯，而由执法实践形成的实际标准不可追溯，且后者的入判份额更高——创作者调整行为依据的是被删了什么，不是写着什么。

七、软件的事实标准。　参考实现的行为压倒规范文本，这是入判份额高于名义权威的典型。可裁决预测：当规范与参考实现冲突时，改变的将是规范。

八、组织的绩效指标。　一线员工的产出如果不进入指标修订时的比对材料，那么无论开多少次征求意见会，指标都不会真正改变。可裁决预测：把上一周期的一线异常案例列为指标修订的强制比对项，指标的实质修改率会上升；只增加征求意见环节，不会。

九、语言的用法变迁。　词典跟随用法而不是相反，且用法的改变不可追溯到具体的人。右上格，且这一格在这里显然不是病——这提示第二根轴的价值判断依领域而异。

十、演化本身。　一个物种的产出成为另一个物种的判据，且不可追溯到任何个体。右上格。**而这一格在演化里是常态与创新的来源，不是病理**。　这是本文所举例子中最重要的一个，它直接把本文的价值排序拆掉——见下一节。

八、两处被材料逼出来的收窄

第一处，由强化学习逼出：入判份额有上界。 单主体系统的坍缩说明，当一个主体的产出在判据中所占份额趋近于全部时，判据不再提供任何区分——它变成主体自身的映射。这不是只在单主体里发生：一个小共同体里，少数几个人的产出可能占据其他所有人判据的绝大部分，此时这个共同体的判据同样会失去外部信息，表现为内部一致而外部无效。

由此得到一条可测的推论：**共同体的规模与判据的稳定性相关，且小共同体更容易出现判据坍缩。** 也得到一条对本文的限制：入判份额不是越高越好，它有一个使判据保持信息量的上界。本文无法给出这个上界的具体值，只能给出它存在的理由与一个可观测的信号——当一个共同体的判据在内部高度一致而在外部检验中系统性失败时，份额已经过高。

第二处，由演化逼出：右上格不是病。 前面已经判定演化本身处在右上格——高入判份额、完全不可追溯。而这一格恰恰是地球上最大的一次开放式创新的运作方式。若本文坚持右上格一律是病，就会得出荒谬的结论。

因而本文不含「可追溯优于不可追溯」这一价值排序。第二根轴测的是**可争辩性**，而可争辩性只在需要为判据的改变承担责任的场合才有价值。演化不需要为它的判据辩护，语言用法的变迁也不需要；而科研评价、司法裁判与平台规则需要。**本文由此把适用范围缩到：那些必须为标准的改变提供理由的共同体。**这是一处彻底的削弱，它拿掉了本文本来可以主张的一整套规范内容。

九、与最近的几种说法的分界

一、与认识主体性。 那一支关注学习者对知识目标、问题与标准的承担，是本文最近的一位。分离线：主体性是一种被授予或被培养的地位，入判份额是一个可以被查的关系量。判决点：一个在主体性量表上得分很高、而产出不进入任何人判据的学习者，主体性框架预测他会推动知识前进，本文预测他改不动任何东西。

二、与双环学习。 那一支主张组织不仅修正行动，还要修正支配行动的变量。这是本文在管理学一侧最近的占位者。分离线：双环学习问的是一个组织有没有能力去修正支配变量，本文问的是它凭什么能修正——凭它的产出已经进了别人的判据。判决点：一个高度自觉、反复检讨支配变量、而其产出不进入任何外部方判据的组织，双环学习预测它能自我更新，本文预测它的更新会在与外部接触时立即失效，因为它改的是只对自己有效的标准。

三、与问题发现。 那一支区分了解决给定问题与形成问题本身，并指出后者更能预测长期成就。分离线：问题发现是个体的一种能力，入判份额是个体与共同体之间的一种关系。判决点：两个问题发现能力相同的人，入判份额不同者，其提出的新问题被后续研究采纳的比例不同——而这个差异不能由问题本身的质量解释。

四、与可进化性。 那一支说明在判据不可改的前提下创新如何发生，机制在变异分布的结构。分离线：可进化性讲的是产出侧，入判份额讲的是判据侧。判决点：两个可进化性相同的系统，其中一个的产出构成他人判据、另一个不构成，本文预测前者所处的判据会随时间改变，后者的不会——而可进化性对此不作预测。

五、与奖励塑造与规格博弈。 那一支处理的是主体如何利用判据的漏洞。分离线：规格博弈假定判据是固定的、只是有洞；本文说判据一直在被产出改变，因而「洞」不是静态的。判决点：在一个多主体系统里，堵上一个已知的漏洞之后，规格博弈预测主体会去找下一个漏洞；本文预测判据本身会因为主体的产出而移动，使得原来的漏洞定义失效——两者的读数不同。

六、与社会建构式的规范解释。 那一类解释说标准是社会协商的产物。分离线：协商是一个关于说服的过程，入判份额是一个关于产出的量。判决点：一个在协商中完全没有发言权、而产出被所有人当作参照的角色（比如一个匿名的基准数据集），协商解释预测它不影响标准，本文预测它决定标准。

十、三家各自为什么到不了这一条

学习科学到不了，因为它的处方框架是制度授权。 它看见了责任的分配问题，就把解法设想成把责任分给谁；而入判份额不是可分配的东西。

演化到不了，因为它把判据叫作环境。 它完全知道其他生物构成环境的主要部分，它有整套协同演化研究，但它的问题始终是「谁跟着谁变」，从不问「判据由谁的产出构成」——因为在它那里没有人需要为判据辩护。

强化学习到不了，因为它只有一个主体。 在单主体设定下，「产出进入他人判据」这件事在定义上不存在，因而它只能把问题看成隔离与防护。

十一、推论、适用边界与反噬

第一条推论：授权是最贵而最无效的一种干预。 它需要制度变更、需要会议、需要文件，而它改变的只是名义持有者。相比之下，制造入判往往只需改变下一轮任务的参照集，成本低得多。

第二条推论：想让某个位置能改标准，就把它的产出放进别人下一轮必须看的地方。 这是本文唯一一条可执行的处方，且它在三节里各有一个现成实现：把上周的学生解法放进本周的题目，把一线异常案例列为指标修订的强制比对项，把参考实现与规范并列。

适用边界一：本文只适用于需要为标准的改变提供理由的共同体。 这是第八节第二处收窄的直接结果。

适用边界二：入判份额有上界，超过则判据坍缩。 本文不能给出这个上界，只能给出它存在的理由与信号。

适用边界三：本文不处理入判份额的正当性。 一个产出可以因为质量、也可以因为权力或规模而进入他人判据；本文只测量它进了多少，不判断它该不该进。这是一个真实的缺口，且它使本文有被用作现状辩护的风险。

反噬预言。 若入判份额成为一个被追求的目标，最可能发生的不是产出变好，而是**对参照集的争夺**：人们会把力气投在让自己的东西成为别人必须比对的对象上——争进教材、争进基准、争进默认配置——而这类争夺与产出质量的相关性远低于它与资源的相关性。**本文提出的量一旦被当作目标，就会奖励那些最有能力占据参照位的人，而这恰恰是本文用来诊断暗流的那一格。** 这个循环没有出口；能做的只是坚持第二根轴——要求进入是可追溯的，使占位至少可以被指认和质疑。

交出去的一个洞。 本文假定「他人的判据」可以被查。在多数场合这是可行的（引用、比对对象、参照集都有痕迹），但在最要紧的那一格——暗流——按定义查不到。于是本文最想诊断的那一格，恰恰是本文的判据最难施用的一格。目前只能靠间接信号：判据发生了改变而无人能指出改变的来源。

十二、与同期几篇的分界

与「不按的代价先落在谁身上」那一篇。 那篇问干预权持有人是否付不干预的账，本文问这个位置的产出进不进别人的判据。判决点：一个把不修改标准的代价完全内部化、而产出仍不进入任何人判据的位置，那篇预测修改率上升，本文预测修改照样发生但没有人跟随，下一轮的评价依据不变。

与「记零位」那一篇。 那篇问一个动作在账本上有没有科目，本文问一个产出在别人判据里有没有份额。判决点：一个被正式记账、写进履历、可以迁移，而产出从不进入任何人下一轮参照集的位置，那篇判为健康，本文判为被记名的旁观者。**被记账与被参照是两件事。**

与「机器答案到达前留一份独立记录」那一篇。 那篇的底稿显露判断如何被改变，本文测判据由谁构成。判决点：一份底稿完备、且被公开的记录，若它不进入任何人下一轮的比对对象，那篇判为有效，本文预测它不改变任何标准。

与「学习的单位是规则的前后差」那一篇。 那篇的规则改写发生在个人内部，本文的判准改写发生在共同体之间。判决点：一个人的规则可以大幅改写而入判份额为零（他自己变了，没人跟着变），也可以规则纹丝不动而入判份额极高（他没变，但所有人拿他当参照）。两个量正交，可以在同一批人身上分别测。

与「中断之后能不能接手」那一篇。 那篇测能力，本文测关系。判决点：接手能力相同的两组，入判份额不同者，其对流程标准提出的修改被采纳的比例不同。

与「闭合之后还能不能回来」那一篇。 那篇问二次进入的材料在不在，本文问改动之后有没有人跟。判决点：一个留痕完备、重新打开旧案很容易、而重开后的新结论不进入任何后续判断参照的制度，那篇判为高续修，本文预测标准原地不动。

十三、本文自己的入判份额是多少

按本文自己的判据，这篇文章的入判份额目前是零：还没有任何人的下一轮判断把它列为必须比对的对象。它提出了一个量，而这个量对它自己的读数是零。

这不是自谦，它有一个具体的后果：**按本文的主张，一篇论证「授权无用、要靠入判」的文章，本身也不能靠被授予地位（发表、署名、栏目）来改变任何标准。** 它要么进入别人下一轮的参照集，要么什么也不是。而它能否进入，与它写得多好只有微弱的关系——第十一节的反噬预言正是关于这一点，且它同样适用于本文。

能做的对冲只有一条：把第十四节第一条检验写到可以被别人直接跑，因为一条被别人跑过的检验，无论结果如何，都会进入跑它的人的下一轮参照集。这是本文所知唯一一种不依赖授权也不依赖占位的入判方式。

十四、可以判本文错的六种结果

第一，最便宜、今天就能做的一条。 在两个平行班里，授权程度完全相同（都允许修改评分权重），只有一个班把上一轮学生的解法列入下一轮题目的必比对象。本文预测后者提出的实质性标准修改数量显著更多、且这些修改被后续沿用的比例更高。若两班无差异，核心机制不成立。

第二，授权本身足够。 若在入判份额为零的条件下，仅靠授予修改权就能稳定产生实质性的标准改变，本文的核心区分失效，学习科学的原处方成立。

第三，份额不预测改变。 若在控制资历、资源与领域规模后，入判份额不能预测一个主体所在共同体的标准变化，则这个量没有增量解释力。

第四，上界不存在。 若在多个小规模共同体中，入判份额极高而判据仍保持外部有效性，则第八节第一处收窄不成立，坍缩不是份额的函数。

第五，第二根轴是多余的。 若可追溯与不可追溯的高份额局面在标准改变的质量与可撤回性上没有差异，则四格表应合并为一根轴。

第六，本文自陈最软的一处。 也许入判份额只是影响力的另一个名字。要把它们分开，必须找到一个影响力很低而入判份额很高的案例——一个不被引用、不被承认、而所有人下一轮都必须比对的产出。基准数据集与默认配置是候选，但它们的影响力是否真的低，尚不清楚。

正向读数。 若本文成立，把某个位置的产出加入他人的必比对象之后，该位置提出的标准修改被采纳率应上升，**且这个上升应早于该位置的名义权限或声望发生任何变化**。顺序比相关更难被巧合满足。

写死日期的赌注。 到二〇三一年十二月三十一日，至少应有一个主流的学习平台或科研评价体系，把「上一轮产出进入下一轮参照集的比例」作为一级数据对象加以记录与呈现，而不再只记录谁被授予了哪

些修改权限。若届时没有这类实现，且上述班级对照的差异被证明不存在，本文关于授权无效、入判有效的主张即告失败。

结　论

谁有权修改评价标准，三个领域给出三个不相容的答案：必须交给主体、这种权力根本不存在、交出去就会崩溃。三个答案都只对了一段，因为它们共享一个未被说出的前提——判据是被某个位置持有的东西，问题只在这个位置该给谁。

演化推翻了这个前提。在那里没有任何个体能改自己的判据，而所有个体合起来一直在改所有人的判据；判据的主要成分是被判者的产出。强化学习的坍缩由此得到解释：那里只有一个主体，共同产出退化成自我产出，判据失去区分能力。学习科学的处方之所以难以复现也由此得到解释：授权改变的是名义持有者，不是判据的实际构成。

于是判断一个人能不能改下一轮的标准，不必去查他被授予了什么，只需要问一句：**他上一轮的产出，有没有出现在别人下一轮的评价依据里。**　出现了，他已经在改，无论有没有人授权；没出现，给他多少修改权都没用，因为改了没人跟。

而想让某个位置能改标准，最便宜的做法也不是给它权限，是把它的产出放进别人下一轮必须看的地方。

注　释

1. 本文所说的「判据」指一个主体在下一轮判断中实际比对的对象集合，不是公开宣称的评价规则。二者常常不同，而本文只处理前者。
2. 演化部分的转述只使用可进化性与协同演化两支中被广泛接受的骨架，不涉及仍有争议的具体机制之争。
3. 强化学习部分的「规格博弈」现象有大量公开记录，本文只使用其结构，不主张任何具体案例的因果解释。
4. 第五节四格表尚未经过信度与效度检验，是研究设计的起点，不是成熟的评估工具。

参考文献

1. Wagner, G. P., & Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. Evolution, 50(3), 967-976. wiley.com
2. Kirschner, M., & Gerhart, J. (1998). Evolvability. Proceedings of the National Academy of Sciences, 95(15), 8420-8427. pnas.org
3. Krakovna, V. et al. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind.（含长期维护的公开案例清单） deepmind.google

4. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565. arxiv.org
5. Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), Liberal Education in a Knowledge Society (pp. 67-98). Open Court.
6. Argyris, C. (1977). Double loop learning in organizations. Harvard Business Review, 55(5), 115-125.
7. Getzels, J. W., & Csikszentmihalyi, M. (1976). The Creative Vision: A Longitudinal Study of Problem Finding in Art. Wiley.
8. Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. Proceedings of the Royal Society B, 205(1161), 489-511.

人机分工声明

本文由王德生提出研究方向、承重判断与最终发表责任；Claude 完成三个领域的选源与冲突定位、公开文献检索、命题成形、二维表构造、逐领域兑现、分界与证伪设计以及正文写作。第八节两处收窄由材料逼出，其中第二处拿掉了本文本来可以主张的一整套价值排序，全过程保留在正文中。本文未标注任何评分——按本站惯例，参与改写者不为自己改写的文本打分。

证据边界 演化部分只使用可进化性与协同演化两支中被广泛接受的骨架，不涉及具体机制之争；强化学习部分使用的是公开记录的规格博弈现象，不主张任何具体案例的因果解释。第十四节六条判据本文**均未执行**，其中第一条可在一门课的一个学期内完成。

字数 纯汉字约 1.1 万 **版本** v1.0 · 2026-08-01

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