代理坍缩与空值结算——“学生—AI”为何拒绝被记成一个学生

王德生 · 德麦国际 · 2026 年 9 月 7 日

摘要

生成式人工智能进入教育生产之后，作业、考试与资格认证仍以”学生独立完成”为记账单位，而产出已经由学生与人工智能共同生成，分数、评语与责任归属之间由此错位。既有解释把它分别归入学术诚信、工具素养、过程性评价与测量效度，能够说明学习行为如何变化，却解释不了两件事：为什么评价装置在读数失真之后仍持续运转，以及为什么题目越复杂越容易被转译成新的应试代理。本文提出的承重判断是：学生—AI 既不是”学生加工具”，也不是一个已经稳定成形的新主体，而是一个拒绝被约简的耦合场；教育的危机不在于新单位出现，而在于旧账本把不可约的耦合场继续记作学生个体差异。为说明这一判断，本文命名两个机制：其一是”代理坍缩”，指评价装置把不同生产路径的产出压成同一读数，路径差异因此无法进入分配结果，行为随之沿最低成本方向收敛；其二是”空值结算”，指分数在失去测量含义之后仍被继续用作分配货币，且失真本身反而提高了它的流通量。后一机制是本文与 Goodhart—Campbell 传统的分离线：那一传统预测被博弈的指标会失效并被弃用，本文预测失效的指标会更牢固地留在结算位上。本文还指出新单位的合法化是不对称的：教师—AI 因提高结算效率而在账本一侧悄然合法化，学生—AI 因动摇分数指称而在被记账一侧成为危机。本文以三条已经跑过的历史序列检验这两个机制：合同代写市场对开放性作业的商品化早于人工智能十余年；高利害考试的分数通胀与分数的分配权重长期脱钩；飞行执照与医师再认证把资格的真值从一次考试移到持续承担的执业风险。结论是：教育改革的第一步不是出更难的题，而是建立记录产出来源、生成路径与责任承担的中间账；能够迫使旧账本退场的力量来自校外——职业准入与执业现场以风险为真值，架空学校分数的分配功能。本文同时说明，同日提出的”学席”概念与本文并不冲突：学席是耦合场被转成制度对象后的记账单位，耦合场本身仍不是单位。

关键词：生成式人工智能；教育评价；代理坍缩；空值结算；中间账；职业资格；学生—AI；学席

一、一个已经失效的分数

一名学生提交了一篇结构完整、语言流畅、论证清楚的课程论文，教师按既有标准评分，论文在形式与结果上均达优秀。当教师要求学生在课堂上解释其中一个关键概念、说明论证为何选择某条路径，或者根据一份事先未提供的新材料即时重建论证时，学生做不到。这里的异常不是”学生没有掌握知识”，也不只是”论文可能由人工智能代写”。更深的异常在于，同一个分数同时指向了两个不同的生产过程：一个过程由学生独立生成，另一个过程由学生调用人工智能完成理解、构思、组织与表达。分数仍然存在，分数所指的对象已经不再稳定。

这类场景重要，不在于它揭示了几个学生可能作弊，而在于它暴露出教育评价的一个从未被写出来的预设正在松动。课程设计、课堂教学、作业、考试、学分与文凭，长期以来都以”学生”作为隐含的记账单位：知识被理解为学生拥有的东西，能力被理解为学生能够独立执行的活动，成绩被理解为学生差异的数字表达，责任归学生本人。人工智能进入学习生产之后，产出不再必然由学生单独完成。学生可以让模型解释材料、生成提纲、提出反例、修改表达，甚至承担原来被视为学习核心的理解与推理环节。教育机构仍然在记录”学生的分数”，生产却已转向学生与人工智能的组合。

“学生—AI”因此不是一个简单的新标签。它是由任务、工具、时间、制度、语言和行动共同形成的一个耦合场。同一个学生在不同任务中可能以完全不同的方式使用人工智能：这次作业让它代写，那次课堂练习只要它给反例，某次实验让它清洗数据，某次职业模拟则必须自己下最终判断。每一次组合的配比、边界与责任结构都可能不同。如果把它当作”学生”的技术延伸，评价就会继续假定个体单位没有改变；如果把它当作一个固定的”学生—AI”新单位，又会忽略组合本身的动态差异。

本文要处理的是这个两难背后的结构，并交出三项东西。第一，命名”代理坍缩”：评价装置把不同生产路径的产出压成同一分数，路径差异被抹平，行为沿最低成本方向收敛——这不是作弊增加，而是评价单位与生产单位脱节之后的结构性结果。第二，命名”空值结算”：分数失去测量含义之后不退场，反而流通更广，因为它的结算功能与测量功能的维持条件不同；这是本文与 Goodhart—Campbell 传统的分离线。第三，推翻”用更难、更开放、更深度的考题就能逼出新

单位”这条改革假设，并说明资格重构必须由职业现场的风险承担反向进入校园，而不是由校园内部不断升级题目。

全文的问题可以压成三句：学生—AI 的差异路径与形成土壤如何共同生成，并如何在旧账本里被伪装成学生个体差异；当评价的中间读数已经失真时，教育机构应当先改哪一项配置；为什么职业准入与执业风险可能比校园考题更早推动资格制度改变，而考题的代理化本身反而无法迫使旧账本退场。

二、三条既有解释与它们共享的前提

关于人工智能与教育的既有解释大致落在三条脉络上。

第一条把问题归入学术诚信，关注代写、抄袭、作弊与违规使用。它握住的解释变量是”行为是否违反规定”，适合处理哪种使用被禁止、哪种责任应被追究。这一脉络最强的版本不是道德谴责，而是制度诚实论：学生若隐瞒人工智能的实质参与，教师就无法判断学生是否独立完成，评价失去公平。本文接受这一点，但它只说明了”披露”是必要的，尚未说明披露之后如何处理一个动态耦合场。即使学生诚实声明使用了人工智能，学校仍要回答：这份产出在何种意义上属于学生？学生承担的是选题、判断、核验、修改还是最终责任？这些维度进不了评价结构，诚信规则就只是在旧账本上添一栏，动不了旧账本的单位。

第二条把人工智能视为新的认知工具，强调提示设计、信息核验、批判性思维与人机协作能力，把教育目标从拒绝人工智能转向掌握人工智能。这条脉络比诚信框架更接近现实，因为人工智能已经进入写作、编程、研究与职业实践，学校不可能把所有学习恢复为无工具状态。它的局限恰在于此：当”会用人工智能”成为新的能力标准，能力被重新表述为工具操作熟练度，工具参与所造成的责任分配问题被推到背景。学生可以非常擅长调用模型，却说不出何时应当拒绝模型的建议，也承担不了错误后果。工具素养的提高不等于责任主体的形成。

第三条从教育测量与教学设计出发，主张用过程性评价、口试、现场写作、项目任务与多元证据弥补终结性考试的不足。它握住的变量是”证据生成过程”，能指出单一终稿不足以代表学生的知识与能力。这条脉络已经接近本文，因为它承认结果读数必须与过程读数结合。但它通常仍把过程视为”更好的学生证据”，而不是一种新的生产关系。把作业搬进课堂、增加答辩、要求交

草稿，确实提高可见性，但人工智能会沿注意力梯度移动：终稿外包退回提纲外包，现场表达建立在课前生成的脚本上，课堂写作变成对预制答案的复现。只要过程评价的目标仍是给学生个体一个可比较的分数，过程就会被再次代理化。

三条脉络分别握住诚信、工具能力与评价证据，它们共享一个未经检验的前提：评价对象仍然是学生，人工智能只是影响学生行为的外部因素。这个前提使问题被表述为”怎样区分学生完成的部分与人工智能完成的部分”，或”怎样防止学生借人工智能绕过评价”。可是如果学生—AI 组合的形成受任务、时间、制度激励与既有能力共同影响，且不同组合之间不存在稳定可通约的比例，那么问题就不是测量技术暂时落后，而是测量对象本身无法被稳定约简。

这三条脉络还有一个共同的解释空白：它们都无法说明学校为什么继续使用已知不可靠的分数。若问题只是工具不足，学校在发现测量失真之后应该更换装置；现实是学校继续用旧分数招生、分班、毕业与分配机会。分数不只是测量工具，它还是分配权的通用语言。它可以失去测量含义，却暂时保留分配功能。这个双重性是本文要命名的东西，也是本文与既有理论分离的位置。

三、五位缺席者与分离线

上一节的三条脉络是教育内部的解释。要把本文的判断放到位，还必须对五个更深层的理论位置做划界。它们不在教育文献的常见引用名单上，却各自占着本文论证空间的一角。若不逐一说明本文与它们的分离线，“代理坍缩”和”耦合场”就有被读成旧概念改名的危险。

第一位：分布式认知与延展心智。 Hutchins（1995）对舰船导航的民族志证明，认知任务在人、工具与表征之间分布，导航的”知道”不属于任何一个人。Clark 与 Chalmers（1998）进一步主张，只要外部资源在功能上与内部记忆等价，心智的边界就应延展到工具上。这两家为”学生—AI 是耦合场”提供了本体论的许可：思考确实不必属于皮肤之内。但它们的问题是描述性的——认知如何分布；本文的问题是制度性的——分布之后谁被记账、谁被归责。Hutchins 的导航团队不需要给每个水手打一个”独立完成”的分数，而学校必须。分离线在此：分布式认知说明耦合可以发生，本文说明耦合发生之后旧账本为何仍以个体记账、以及这种记账如何反过来塑造耦合的形态。延展心智没有”责任”这一变量，本文的核心读数正是责任落点。

第二位：Goodhart 定律与 Campbell 定律。 Goodhart（1975）指出一旦统计规律被用作控制目标，规律便告失效；Campbell（1976）指出社会指标被用于决策的程度越高，它越容易被扭曲，且越容易扭曲它本应监测的社会过程；Strathern（1997）把它压成一句——指标一旦成为目标就不再是好指标。Koretz（2008）在教育测量中给出了经验版本：高利害考试的分数在几年内可以上涨一个标准差，而同一群学生在低利害审计测试上的成绩几乎不动。这一传统离本文最近，也是”代理坍缩”最容易被还原进去的地方。分离线在于预测的方向：Goodhart—Campbell 预测被博弈的指标失去信息、被识破、被替换；本文预测失去信息的指标反而更牢固地留在结算位上。Koretz 自己的数据已经暗示了这一点——分数通胀被系统记录二十年，高利害考试的分配权重没有下降。Goodhart 解释了指标为何失真，解释不了失真的指标为何不退场。本文第九节要命名的”空值结算”正是补这个洞。

第三位：文凭主义与信号理论。 Collins（1979）主张文凭的功能不是证明能力而是分配地位；Labaree（1997）指出美国教育的”文凭竞争”目标使学生学会的是获取标记而非学习本身；Spence（1973）的信号模型则说明，当获得信号的成本对所有人同步下降时，信号的区分力下降而信号水平上升。这三家解释了分数为何具有分配功能，但它们把”能力”当作先于信号而存在的私人属性，信号只是能力的不完美代理。本文的分离线是：在学生—AI 耦合场中，能力不是先于任务与工具而存在的固定量，它在路径与土壤中生成，因此信号所代理的那个对象本身已经不稳定。信号理论处理”信号偏离能力”，本文处理”被信号的对象不再是一个对象”。

第四位：行动者网络理论。 Latour（2005）取消了人与非人在行动网络中的先验区分，行动被理解为网络的效应而非主体的属性。这为”学生—AI”提供了拒绝人类中心记账的理由。但行动者网络理论的方法论承诺是”跟随行动者”、拒绝预设结构，而本文恰恰要说明一个结构——旧账本——如何在网络已经改变之后继续以旧单位运转，以及这种运转如何被具体的制度利益维持。分离线是：行动者网络描述网络的展开，本文描述记账结构对网络展开的截断。

第五位：人工智能使用分级量表。 Perkins、Furze、Roe 与 MacVaugh（2024）提出的 AI Assessment Scale 把人工智能在作业中的许可程度分为从”禁止”到”完全使用”的五级，供教师在任务层面声明规则。这是当前教育实践中最直接的制度回应，它承认使用程度必须被写进评价。分离线是：分级量表把”程度”当作变量，本文的变量是”环节”与”责任”——同为第

三级”允许协作”，模型润色一段已独立写成的文字与模型先行组织论点，是两种完全不同的路径结构。程度变量无法替代路径变量，量表因此仍在旧账本内部工作：它规定人工智能可以用多少，不记录关键判断由谁作出。

五条分离线合起来，可以把本文的位置说清楚：本文借用分布式认知的本体论许可，借用 Goodhart—Campbell 对指标失真的机制，借用文凭主义对分数分配功能的揭示，借用行动者网络对人机对称的方法态度，借用分级量表对制度回应的经验形态；本文自己要交出的是它们合起来都没有的一条判断——失真的分数为何不退场，以及旧账本在读数失真后如何反过来塑造耦合场本身。

四、发生学重构：显露、路径与土壤

本文站在发生学与关系本体论的交叉位置上。发生学不问一个对象如何被发现，而问它为何以现在的方式发生：哪些条件使它出现，哪些路径被保留，哪些可能性被排除，形成之后的对象又如何反过来改变原有条件。关系本体论则拒绝把对象理解为先于关系而独立存在的实体，对象的边界、性质与功能都在关系网络中形成。

采用这一传统而不是工具主义、个体主义或制度主义，各有一个理由。工具主义能够说明人工智能提高效率，说明不了学生单位为何失效。个体主义能够保住学生责任，却把人机协作视为个体之外的偶然因素。制度主义能够解释分数的分配功能，却容易把制度当作既定背景，而不是由路径与土壤共同生成、又反过来塑造路径的结构。

为了把这个交叉位置写成可操作的分析层，本文区分三个层次。**显露层**是某个对象在制度表面以何种形式出现：作业、考试、文凭显露为学生个体的产出。**路径层**是对象在一连串选择、试探、保留与排除中如何形成：学生每一次使用人工智能都在试探、避险、保留与迭代中走出一条路径。**土壤层**是制度、资源、语言、关系、时间与心理状态如何共同构成对象的存在条件：评分规则、时间压力、工具可得性、同伴比较、职业竞争与责任结构共同塑造了学生—AI。

三层之间不是相加关系而是互相生成。显露不是路径的简单结果，因为土壤会改变显露方式；路径也不是在既定环境中运行，因为显露出来的结果会反过来改变环境；土壤不是外部背景，因为对象与路径会持续改写土壤。教育评价的假读数正是在这样的情形下发生的：显露层仍然稳定地

显露为学生分数，路径层已经改变，土壤层却继续支撑旧的显露。教育系统看到的是显露层，实际生产发生在路径与土壤的纠缠里，而旧账本把显露层还原成学生分数。

本文因此不问”人工智能是否帮助学生学习”，而问：何种差异路径在何种土壤中被保留下来，为什么最终显露为学生个体差异，以及什么外部事件能迫使制度承认这种显露已经失真。

五、路径与土壤：耦合场的发生序列

第四节给出了三层，这一节把学生—AI 在路径层与土壤层的实际发生序列写出来。只有把序列写具体，才能说明为什么耦合场”不是一个单位”不是修辞，而是可以逐步观察的事实。

路径层的发生从试探开始。学生第一次把课程材料交给模型，通常不是为了代写，而是为了看模型能做什么：解释一个概念、给一个提纲、改一句话。试探的结果是一次局部成功，成本极低而读数不变——教师没有察觉，分数没有变化。第二步是识别风险。学生评估这条路径会不会被发现、被发现的代价是什么、同学是否也在走。第三步是保留。当风险评估为低而收益为稳定时，这条路径被固定为默认动作：先让模型理解，再由自己复述；先让模型生成，再由自己修改。第四步是迭代。学生在不同任务之间调整配比，把模型推向更前端的环节——从表达退到构思，从构思退到理解——每一次后退都是一次新的试探与保留。路径层的关键特征是：每一步都是局部理性的，没有一步是”决定作弊”，然而序列的终点是理解与判断环节被整体让渡。

路径的每一步都需要土壤的支撑，土壤不是背景而是共同作者。时间压力使试探的动机成立：任务在截止日期前必须完成，模型是最快的完成方式。工具零成本使保留的门槛消失：在合同代写时代需要付费的路径，现在不需要。同伴比较使风险评估的结果偏向使用：当同学都在用而分数不变，不用者承担的是相对成本而不是绝对收益。生产力叙事为使用提供正当性：学校、企业与媒体都在说”会用人工智能是未来的核心技能”，学生沿这条路径走时不需要对自己解释。升学竞争与资格分配则把整个序列锁定：只要终局分数决定机会，任何降低取得终局分数成本的路径都会被保留。

序列在群体层面有一个转折点。当使用者比例越过某个阈值，不使用者的相对位置开始系统性下滑，此时”使用人工智能”从个体策略变成群体默认，土壤完成了一次改写：原来支撑路径的条件（低风险、低成本）现在被路径本身加固（大家都在用、不用吃亏）。这就是路径反过来改写

土壤的时刻，也是耦合场”拒绝被约简”的发生学根据——同一个学生的组合方式随任务、随时段、随同伴而变，因为它每一次都是在被上一次改写过的土壤上重新发生的。

显露层在整个序列中始终不动：作业照交，分数照打，学分照记。序列的所有变化都在显露层之下发生，显露层把它们全部压成”学生的分数”。旧账本不是没有看见，而是它的记账单位只允许它看见一种东西。

六、承重命题

本文的承重判断是：

> 学生—AI 不是学生的技术增强，也不是一个已经稳定形成的新单位，而是一个不能被稳定约简的耦合场；教育危机不是新单位取代旧单位，而是旧账本把不可约的耦合场继续记作学生个体差异。

用最短的形式写：学生—AI 既不是”学生加工具”，也不是”可直接替代学生的固定新主体”，而是”在任务、路径与土壤中不断变形的耦合场”。

这一判断包含两个相互关联的否定。第一，学生—AI 不是一个已经完成的新单位，它是差异不断生成的场。第二，考题改变不是单位改变的充分条件；只要旧账本仍把产出压成学生个人分数，新的题目就会成为代理路径的材料。教育改革若只修改题型，就会在内容层面制造变化，在单位层面维持不变。

命题在以下条件下成立：人工智能参与了产出形成而非仅提供机械输入；学生与人工智能之间的分工随任务变化；评价装置只读终局结果而不读来源与路径；不同路径在结果上获得相近判定；低成本代理路径因同价而持续扩张。若人工智能仅承担拼写检查、格式转换等不改变理解与判断的环节，或者评价装置能够稳定、低成本地记录并区分各环节贡献，命题的强度就下降。

命题不成立的条件也要先写出来。若一种新评价装置可以在不依赖学生自报的情况下，稳定记录理解、构思、生成与核验的责任分配，并且这一记录能够进入正式资格决策，学生—AI 就可能被分解为可比较的多维单位。此时它仍然是关系场，但它不再”拒绝被记”，而是被转换成一套新的制度对象——第十六节讨论的”学席”就是这样一个制度对象。若真实执业环境能够直接检验

产出责任，且资格不再主要依赖学校分数，旧账本即使继续存在，也不再承担全部分配功能，本文关于旧账本持续统治的判断随之削弱。

最后要过一道复合测试：本文的核心判断能否由两句现成文献无损重述？一句可以说”工具正在改变学习活动”（分布式认知），另一句可以说”指标一旦成为目标就会失真”（Goodhart—Campbell）。两句合在一起仍推不出：学生—AI 为何不是一个稳定单位，为什么更复杂的考题会被翻译成代理路径，以及为什么旧账本在失去测量含义之后仍保留分配权。本文的增量不在于把两句放在一起，而在于命名把它们连起来的两个机制。

七、代理坍缩：定义、判别式与指数

代理坍缩是评价装置把不同生产路径的产出压缩为同一判定，从而使路径差异无法进入分配结果的过程。它不以学生是否主观作弊为定义。只要独立完成、部分协作、整体外包等路径在评价装置中获得相同或相近读数，并且成本更低的路径因此被保留，代理坍缩就发生了。

这个定义的要害在”坍缩”二字：它不是说学生找到了更聪明的作弊方式，而是说装置本身把所有新差异重新压成一个可比较的分数。压缩是装置的动作，不是学生的动作。学生只是沿着被压平的地面走最短的路。

本文的判别式是：

> 若同一产出在不记录其生成来源与责任承担的情况下获得正式资格读数，并且至少存在一条更低成本的替代路径能够得到相近读数，则该评价装置发生了代理坍缩。

判别式不含”应当”“合理”等情态词，可直接用于流程、日志与记录。它要求观察三个事实：是否记录来源；是否记录责任承担；不同成本路径是否得到相近结果。

把它操作化为读数：设同一任务中存在两条可观察路径，路径 A 为学生独立完成，路径 B 为学生把理解或生成环节外包给人工智能。令评价结果为 R，生成成本为 C。代理坍缩指数

$$AC = 1 - |R_A - R_B| / (R_{max} - R_{min})$$

取值在 [0, 1]，越接近 1，两条路径在评价结果上越难区分。仅有结果相似还不足以构成坍缩，还需记录成本差 $\Delta C = C_A - C_B$，$\Delta C > 0$ 表示独立完成成本高于代理路径。在同一任务、相同时

限与相同评分标准下分别记录两个版本的结果与过程，若 AC ⩾ 0.80 且 ΔC 达到预设的时间或步骤阈值，该事件被标记为”路径同价、成本偏移”。0.80 不是自然常数，是用于提出可复核判据的预设阈值，须经敏感性分析。

判别式在六类场景中的读法如下。课堂写作：学生独立读材料、立论、成文，与学生先让模型生成论点再润色，两者评分相近而评分记录只保留终稿，判据成立。编程作业：学生能运行模型生成的代码却解释不了数据结构与错误处理逻辑，评分只看测试是否通过，判据成立；若评分同时要求学生现场修改一处未预告的错误并记录路径，判据可能不成立。实验报告：模型生成设计与解释、教师只按格式与结论评分，判据成立；学生必须在现场面对设备异常、重新决定变量并承担失败成本，路径差异进入读数，坍缩程度下降。教师批改：模型生成评语、教师仅确认提交、学校把评语视为教师专业判断，“教师—AI”在记账侧先发生。职业准入：资格考试只依据标准化答案发证，执业错误不进入资格维持机制，学校分数具有较高坍缩风险；资格包含持续监督、案例复盘与事故责任，读数便不再只由一次考试承担。

八、教师—AI：新单位先在账本侧出生

学生—AI 不是唯一的耦合场。同一时期，教师也在批改、命题与反馈中调用模型，形成教师—AI。两个耦合场在制度中得到的待遇完全不同，这个不对称本身就是一条证据。

模型生成评语、教师确认并提交，学校把评语视为教师专业判断——这一序列在多数学校里没有引起任何制度反应。它被记录为效率提升：批改时间缩短，反馈周期变快，教师被”解放”出来做更高价值的工作。没有人要求教师申报模型参与了多少，没有人给评语做来源标记，也没有人问一句评语的责任落在谁身上。教师—AI 在记账侧悄悄合法化了。

学生—AI 得到的是相反的待遇：检测器、申报表、口试、惩戒条例。差别不在于人机耦合的深度——教师让模型写评语与学生让模型写论文，在环节结构上是对称的——差别在于两个耦合场与账本的位置关系。教师—AI 站在账本的记账一侧，它提高结算效率而不威胁分配秩序，账本欢迎它；学生—AI 站在被记账一侧，它动摇分数对学生个体的指称，账本视它为风险。新单位因此有不对称的合法化路径：在记账一侧它是效率，在被记账一侧它才成为危机。

这个不对称对本文的命题有两重意义。第一，它说明旧账本不是被动的测量工具，而是一个有自身利益方向的结构：它选择性地承认那些强化它、拒绝那些动摇它的耦合。第二，它说明”教师—AI 先合法化”会反过来改变学生—AI 的土壤。当学生面对的”教师意见”实际上来自模型，学生对反馈的信任锚点已经移动；当学生知道评语由模型生成而教师不必解释依据，学生让模型生成作业的正当性感受也随之上升。账本一侧的耦合为被记账一侧的耦合提供了先例。

这一节的判别式与第八节相同：若模型生成评语、教师仅确认提交，而学校将评语视为教师专业判断，则教师—AI 在记账侧发生了代理坍缩——不同成本的批改路径得到相同的制度读数。若教师保留批改证据、解释评分依据并对错误评语负责，模型可以是工具而不必成为责任主体。这里的证伪条件写在第二十节。

九、空值结算：失真的分数为何不退场

第三节已经指出，Goodhart—Campbell 传统解释了指标为何失真，解释不了失真的指标为何不退场。这一节命名那个被解释空出来的机制。

空值结算是指一个读数在失去其测量含义之后，仍被继续用作分配货币，并且失真本身反而扩大了它的流通量的过程。“空值”指读数与它所指的对象之间的对应已经断开；“结算”指它仍被用来结清资格、名额、机会与地位的分配。

空值结算之所以可能，是因为分数从一开始就有两种功能：测量功能，即分数对某种能力或成就的指称；结算功能，即分数作为不同机构之间可通约的分配语言。两种功能通常被当作一体，测量功能是结算功能的正当性来源。但它们的维持条件不同。测量功能的维持条件是分数与对象之间的校准，这需要不断的外部审计；结算功能的维持条件是分数在机构之间的可通约性，这只需要所有机构继续接受同一种货币。校准可以断，可通约性不会因校准断裂而断。

这就是失真的分数为何不退场：它在机构之间的流通并不依赖它的测量含义，只依赖别的机构继续收它。招生办收学校分数，学校收考试分数，雇主收文凭，每一方都知道分数可能失真，但每一方的替换成本都高于继续接收的成本，因为替换需要另一种所有机构同时接受的货币，而这种货币不存在。分数于是像一种已经脱锚的通货：没有人相信它背后的储备，所有人继续用它结账。

失真反而扩大流通量，这是空值结算与一般”指标失效”的第二个差别。当代理路径使高分变得容易，高分的数量上升，分配需要在更多高分之间做出区分，于是机构对更多分数、更细的分数、更多层次的分数的需求上升，分数的使用密度上升。Spence 的信号通胀模型描述的是信号水平上升而信息含量不变；空值结算描述的是信息含量下降而使用密度上升。两者都在文凭市场上被观察到，前者是均衡的静态描述，后者是失真之后的动态。

空值结算还给出了耦合场为何总被切回一个名字的制度几何。分配容器是单数的：录取录一个人，发证发给一个人，追责追到一个人。耦合场进账时必须通过这些容器，而容器的元数不接受”一个场”，只接受”一个名”。因此单位不是由生产决定的，也不是由测量决定的，而是由分配容器的元数决定的。这一点把”拒绝被约简”的原因从耦合场的任务敏感性挪到了制度几何上：只要结算容器保持单数，任何多元的生产单位在进账时都会被切回单数，测量再精细也改变不了这一刀落在哪里。

空值结算同时解释了第一节引入的现象：为什么题目越复杂越容易被代理化。复杂题目在结算功能上并不比简单题目值更多，它换来的仍是同一种货币。只要货币不变，复杂性就只是提高了代理路径的市场价值——题目越深，越值得被拆解、模板化并做成新题库。开放性本身成了模板化服务的市场机会。改变题目改变的是被测的东西，不改变结算的货币，因此不改变单位。

由此可以给出空值结算的三条可观测征象：其一，分数与外部审计读数的偏离持续扩大，而分数在各类分配决策中的权重不降；其二，围绕分数的代理服务市场规模随题目复杂度上升而扩大；其三，机构在承认分数失真之后，选择的应对是增加分数的种类与层级，而不是替换分数。这三条征象合起来，与 Goodhart 预测的”指标被弃用”方向相反。下面三节用三条已经跑过的历史序列检验它们。

十、跑过的一条：合同代写市场早于人工智能

“深度题会被代理化”的判断不需要等生成式人工智能来验证，它在人工智能之前已经跑了十几年。

合同代写（contract cheating）这个术语由 Clarke 与 Lancaster 在 2006 年提出，指学生付费让第三方完成作业后以自己名义提交。它针对的正是过程性评价路径推崇的那类任务——课程论

文、项目报告、反思日志、学位论文的章节——而不是标准化选择题。Newton（2018）对 1978 年至 2016 年间 71 项样本、六万五千余名学生的自报数据做了系统综述，发现承认过合同代写的学生比例在整个期间平均为 3.5%，但 2014 年之后的样本升至 15.7%，且趋势显著上升。上升发生在人工智能进入教室之前。

这条序列对本文的意义有三层。第一，代理化的对象恰是评价改革试图用来替代考试的开放性任务。任务越开放、越”真实”，越难在考场内完成，越需要课外时间，因此越可外包。复杂性不但没有阻止代理，反而定义了代理服务的产品线。第二，供给方的响应方式证实了模板化预言：代写平台按学科、字数、学位层级与截止日期定价，把开放题当作可重复的产品规格，而不是不可重复的思想事件。第三，制度的响应方式证实了空值结算：英国 2022 年《技能与 16 岁后教育法》把提供代写服务列为刑事犯罪，澳大利亚高等教育质量与标准署自 2022 年起阻断代写网站的访问。两国的应对都是在账本外部打击供给，而不是在账本内部改变记账单位——分数继续被收，只是试图让获得分数的替代路径更贵。

这条序列还给出一个对本文不利的读数，须写下来。合同代写有价格门槛，能够用钱换分的学生是少数，因此在人工智能之前，代理坍缩在总体上被价格限制在一个角落里。生成式人工智能把这个价格门槛降到接近零，代理路径从少数人的策略变成多数人的默认。本文的机制没有变，变的是土壤：价格门槛消失之后，“不使用者承担相对成本”这一条件第一次在全体学生中同时成立。

十一、跑过的第二条：分数通胀与分配权重的脱钩

Koretz（2008）汇集的证据是空值结算最直接的经验形态。在美国若干州引入高利害问责考试之后，州考分数在三到四年内上涨了半个到一个标准差；同一时期同一学生群体在低利害的全国审计测试上的成绩几乎不动，或上涨幅度只有前者的一小部分。分数与它所指的能力之间的校准断了，这是 Campbell 定律的教科书案例。

本文要看的是校准断裂之后发生了什么。在这些证据被系统记录、被测量学界反复引用的二十年里，州考分数在学校评级、教师问责、学生升级与毕业决策中的权重没有下降，多数州反而扩大了考试的科目与年级覆盖。校准断了，结算没有断。这与 Goodhart 传统预测的”指标失效后被弃用”方向相反，与本文第九节的第一条征象一致。

第三条征象也在同一序列中出现。当高分变得普遍，区分能力下降，若干州的应对不是替换分数，而是增加分数的种类：在”及格”之上增设”精通”与”高级”，在州考之外增设课程终结考，把单一分数拆成多层分数。分数的信息含量在下降，分数的使用密度在上升。

这条序列的分离价值在于，它把”指标失真”与”指标退场”在时间上拉开了。如果 Goodhart 传统是完整的，两者应当相继发生；实际观察到的是失真持续二十年而退场没有发生。解释这个时间差需要一个 Goodhart 传统里没有的变量，即分数的结算功能与测量功能维持条件的不同。这就是本文命名空值结算的理由：它不是给已知现象换一个名字，它是给一个已知理论预测失败的位置补一个机制。

十二、跑过的第三条：执照以险为真值

前两条序列说明旧账本如何在校内维持。第三条序列来自校外，说明资格的真值可以不由一次考试承担。

飞行执照是最清楚的例子。取得执照需要通过笔试与飞行考核，但执照的有效性不由那次考核维持：机长必须在规定周期内完成复训与技能检查，必须满足近期飞行经历要求，任何事故与严重事故征候都进入强制报告系统并可导致执照被暂停或吊销。资格在这里不是一次结算的结果，而是一个持续承担风险的状态；执照的真值来自执业现场持续产生的可观察后果，而不是来自入门考试的分数。

医师资格在近二十年里也向同一方向移动。英国医学总会自 2012 年起实施再认证制度，执业医师每五年须基于年度评估、同行与患者反馈、事故与投诉记录重新确认执业资格；美国各专科委员会的”认证维持”制度同样把继续认证与执业记录、持续学习和考核绑定。这些制度的共同点是把资格的一部分真值从入门考试移到执业记录上——移得不彻底、争议很大，但方向是明确的。

这条序列对本文的意义是：在公共风险高、执业许可与资格紧密相连的领域，资格制度已经找到了一种不依赖学校分数的真值来源。当执业现场以风险、错误后果与责任承担建立另一种真值，学校分数在资格分配中的兑换率就开始下降。这正是本文第三个研究问题的答案所在：迫使旧账本退场的力量不来自校内考题的升级，而来自校外资格以险为真值对分数分配功能的架空。校内

考题升级只改变题目表面，只要墙外仍承认分数，墙内土壤便没有自我拆除的动力；只有当学校分数在资格与就业中逐渐无法兑换，而执业风险要求新单位的产出，资格制度才可能先于考题改变。

这条序列的适用边界也须写明。它只在风险后果强、责任不可逆、市场依赖正式许可的行业成立。在风险后果弱、责任可逆或市场不依赖正式资格的行业，雇主可以通过作品集、试做任务或短期试用完成能力识别，资格制度不一定成为最强的外部判决。因此本文不说”职业准入必然先于学校改变”，只说在公共风险高、资格与执业许可紧密相连的领域，职业准入更可能成为外部判决的入口。

三条序列合起来，可以把第九节的三条征象逐一对应：合同代写市场对应第二条征象（代理服务市场随题目复杂度扩大），分数通胀对应第一与第三条征象（校准断裂而权重不降、增加分数种类而不替换分数），飞行执照与医师再认证则给出反向的对照——当真值来源移到执业风险上，空值结算的三条征象同时减弱。三条序列都发生在生成式人工智能之前，因此它们检验的是机制而不是工具；人工智能改变的是土壤中的价格门槛，没有改变机制。

十三、既有解释的判决性读数

第二、三节划出了分离线，这一节把八种最容易占据本文论证空间的解释放进同一批场景，看它们与本文各自读出什么。判决性差异的标准只有一条：同一场景，两种解释给出的读数不同，且差异可观察。

学术诚信解释的最强版本认为核心是学生是否遵守使用规则。课程论文场景中，学生公开声明”使用人工智能协助构思”，论文得高分，课堂追问中学生无法重建论证。诚信解释读为”合规使用后仍需加强学习”，本文读为”来源显露不足，终局分数不能继续代表学生独立差异”。前者记录使用是否被允许，后者记录学生是否承担了理解与判断。诚信解释解释不了的是：规则完全允许使用，评价仍无法区分学生是否拥有论证能力。

工具素养解释认为应培养学生提问、核验、修改与整合的能力。医疗案例模拟中，学生用模型生成诊断建议并按格式完成报告，但没有发现模型遗漏的危险病史。工具素养解释按使用效率与报告完整度判为高水平协作，本文预测责任回收率低，学生没有承担关键判断。两个读数分别是工

具调用成功率与错误发现—修正率。工具素养解释解释不了的是：学生可以熟练操作模型，却无法在错误后果出现前判断何时不应采纳。

过程性评价解释认为终稿不能代表学习，须把草稿、讨论、答辩与课堂表现纳入评价。课堂写作中，学生带着模型预生成的结构进教室，现场完成改写与口头说明。过程评价记录到当场写作与发言，判为过程可见；本文预测生成环节已在课堂之外完成，现场只发生翻译与复现。两个读数分别是”是否当场产生文本”与”关键判断是否当场产生”。过程评价解释不了的是：过程可以被看见，关键生成已在不可见时段完成。

测量效度解释认为问题在于工具缺乏信度、效度或区分度。同一学生在写作、编程与实验中分别让模型参与表达、代码生成与实验设计，测量解释倾向寻找一个统一的”人工智能使用”变量纳入量表；本文预测不同环节的参与改变责任结构，统一变量无法保持指称稳定。测量解释解释不了的是：同样的使用频率在不同任务中对应完全不同的责任后果。

复杂问题解释认为应转向开放问题、真实项目与跨学科任务。项目式课程中，学生完成城市交通方案，题目开放、材料复杂、没有单一答案，但培训机构与模型可以先生成框架、数据解释与展示脚本，学生按模板执行。复杂问题解释判为高真实性，本文预测题目代理化已经发生，路径成本下降而责任读数没有提高。它解释不了的是：开放性本身可以成为模板化服务的市场机会——第十节的合同代写序列已经跑过这条。

资源差距解释认为人工智能造成新的数字鸿沟。若全班都能免费使用同一模型，却同时出现独立理解能力下降、终稿质量上升与现场解释能力下降，资源解释判为差距缩小，本文判为代理坍缩扩大。它解释不了的是：人人获得同一工具之后，评价对象仍可能失去稳定边界。

教师效率解释认为模型减轻批改负担、释放教师专业性。模型生成大部分评语、教师仅确认、学校仍视为教师判断，效率解释记录时间节约，本文记录评分信用锚点已经转移。它解释不了的是：效率提升伴随教师无法说明评分依据，学生面对的”教师意见”实际来自模型。

职业导向解释认为应加强产教融合与职业场景。学校新增人工智能职业项目，企业与准入机构仍以学校考试分数为主要筛选依据，职业导向解释判定课程已接轨，本文判定资格账本没有改变、

外部风险未进入结算。它解释不了的是：学校可以复制职业语言，却不承担职业错误的现实后果。

八种解释分别把问题放在规则、工具、过程、测量、题目、资源、教师效率与职业衔接上。本文不否认这些变量重要；本文的分离线是：这些变量都可以在旧账本中被吸收，而本文要解释的是旧账本为何能吸收它们，并在吸收之后继续把不同路径记作同一类学生差异。

十四、中间账：四类读数与三条自我否决条款

如果代理坍缩是装置的动作，改革的第一步就应当改装置，而不是改题目。本文把需要建立的那一层叫**中间账**：位于产出生成与资格结算之间，记录产出来源、生成路径与责任承担的评价层。它不是最终成绩的附加说明，也不是学生自我申报的道德栏；它的功能是把旧账本读不出的差异显露出来，使不同路径在进入分数、学分或资格之前留下可追踪的痕迹。

中间账由四类读数承载。**来源读数**：独立、协作、外包三类类别变量。**路径读数**：理解、构思、生成、核验、表达五个环节各由谁完成，编码为 0（学生承担）、1（人机共同承担）、2（人工智能承担），一次任务的路径是一个五维序列，两个版本之间的曼哈顿距离可作路径差异的描述，但不直接当作价值判断。**责任读数**：学生能否在限定时间内解释选择、发现错误并承担修正，采用有序等级。**现实读数**：产出在真实或模拟执业环境中引发的可观察后果，采用错误次数、返工次数、风险事件数或任务完成时间。

在这四类读数之上定义三个可观测指标。来源显露率 SR，即正式评价任务中能被记录为独立、协作或外包的任务比例；一次任务计入分子的条件是同时拥有至少一项过程证据与一项责任证据，只有自报而无过程记录的任务不计入。责任回收率 RR，即学生在模型产出出现错误时能够独立发现、说明并修正的比例；只有说明错误来源、指出错误后果并完成修正才计为一次回收，仅仅改对答案不计。资格外部兑付率，即学校评价结果在真实或模拟职业场景中能预测任务承担与错误后果的比例；它不等于就业率，也不等于雇主满意度，而要求把学校读数与后续执业任务建立可追踪的连接。

多少差异才算数，不能只靠一个阈值。本文建议同时满足三个条件：来源显露率在任务层面达到预设比例；路径向量至少在一个核心环节出现稳定差异；责任回收率与最终成绩之间出现可重复

的偏离。若学生成绩高、责任回收率低，且这种偏离在不同任务中持续出现，就说明终局成绩可能已经代理化。

读数必须带自己的否决条款，否则中间账会变成又一套可以被代理的表格。第一，**信度条款**：同一批对象在两周内、难度相近的重测中，来源显露率或责任回收率的组内相关低于 0.70，读数判为不稳定，不得用于高风险资格决策；类别编码的双编码一致性低于 0.80，须回到编码规则修订。第二，**改名嫌疑条款**：若代理坍缩指数与已有的”作业完成度”“学习投入度”“工具使用频率”在控制任务难度后高度重合（相关系数 0.80 以上且增量解释率低于 0.05），本文的读数判为换名而非新构念。第三，**单次事件条款**：一次观测若只能凭学生自报推断来源，不能获得过程记录、现场追问或责任修正证据，该事件不得被标记为代理坍缩或独立完成；概率、倾向和程度只能用于群体估计，不能倒灌为单个学生在单次任务中的事实判定。

十五、辨别格：两轴四格

要给学生—AI 做类型学，先要排除三条看起来自然的候选轴。“人工智能参与程度”是耦合场的成因变量，不是能区分耦合结果的独立轴——参与越多并不必然意味着责任越低，某些任务中模型承担检索或模拟，学生仍承担关键判断。“学习结果好坏”是评价结果而不是结构变量，用成绩作轴会把待解释的结果重新放进解释框架。“学校与企业的距离”能解释外部压力，却不能区分内部责任结构，离企业近的项目也可能把任务模板化。

本文确立的两条轴是：**关键判断是否由学生承担**——不是模型参与多少，而是学生能否在关键节点说明选择、发现错误并承担后果；**产出路径是否被制度记录**——理解、构思、生成、核验与表达的来源是否进入评价。两轴相关但不重合：记录可以很完整而学生仍承担不了判断；学生可能在没有记录的情况下确实独立完成而制度无法识别。

四格如下。**路径记录低、判断承担低**是隐蔽代理化：终稿、分数与资格读数存在，但无法知道理解与生成如何发生，学生也无法在追问中承担判断。它独有的预测是：终局成绩上升或维持而现场责任回收率下降，制度无法解释二者的偏离。**路径记录高、判断承担低**是可追踪外包：学校能记录学生何时调用模型、用了哪些输出、改了什么，学生仍说不出为何接受某个判断。它独有的预测是：来源记录越来越细，学生独立解释能力不随记录增加而改善，制度获得了过程数据但没有获得责任主体。**路径记录低、判断承担高**是未显露的独立责任：学生能在现场面对新材料、解

释选择并修正错误，制度却没有记录路径，最终仍把能力压成一次终稿。它独有的预测是：现场追问显示责任能力高于书面成绩所能表达的范围，成绩与真实表现出现双向偏离。**路径记录高、判断承担高**是责任型人机协作：模型参与若干环节，制度记录来源与路径，学生在关键节点作出判断并承担后果。它独有的预测是：模型参与程度与学生责任能力不再简单负相关，甚至可能在复杂任务中共同提高，但这种提高依赖错误暴露、解释与修正机制。

两轴同时为高的现实案例可以举公开庭审：律师使用模型检索与整理材料，但必须在庭审中对事实主张、证据关联与法律立场作出公开判断，庭审记录、代理关系与责任承担均可追踪。这不是说所有司法场景都已达到高水平，而是说明”人工智能参与”与”责任承担、路径记录”可以同时存在。四格之间的差异不能被”人工智能使用多寡”替代，第四格才是教育改革希望达到的状态，但它不能靠单纯增加模型使用或单纯增加过程记录自动形成。

十六、学席：耦合场如何被转成制度对象

第六节写下了命题不成立的第一个条件：若一种新装置能够稳定记录责任分配并进入正式资格决策，学生—AI 就不再”拒绝被记”，而是被转换成一套新的制度对象。这一节把那个制度对象说清楚，也顺带处理一个可能的自相矛盾。

本文否定了”学生—AI 是一个固定的新主体”，而与本文同日提出的”学席”概念却在给学生—AI 命名并为它设计读数。两者是否冲突？不冲突，前提是分清本体与记账对象。耦合场是本体层的事实：在任务、路径与土壤中不断变形，没有稳定边界。学席是记账层的对象：它是制度在中间账建立之后，为耦合场在某一任务中的一次占位所设的席位。席位有编号、有读数、可比较，但席位不是坐在席位上的那个场。一个学生在编程作业里的学席与他在实验报告里的学席可以有完全不同的路径向量与责任等级，两个学席之间不能通约成”这个学生的能力”，正如两次航班的机长席不能通约成”这个人的驾驶能力”而只能通约成两次记录。

举一个席位读数的例子。某学生在一次编程作业中的学席记为：来源读数”协作”，路径向量 (1, 2, 2, 0, 1)——理解由人机共同承担，构思与生成由模型承担，核验由学生独立完成，表达由人机共同承担；责任读数为三级中的第二级，即学生能解释为何接受模型给出的数据结构，但没有在限定时间内发现模型埋下的一处边界错误；现实读数为该程序在模拟部署中触发一次回退。这一组读数与同一学生在实验报告中的席位读数（来源”独立”，路径向量 (0, 0, 0, 0, 1)，责任第三

级，现实读数零事件）不能相加成一个分数，也不能平均成”这个学生的人工智能素养”。两个席位各自记录一次耦合的形态，学席就是这一次形态的记账名。

这样分层之后，本文与学席的关系变成一种分工：本文说明为什么旧账本不能以”学生”记账，学席说明新账本以什么记账。本文给出学席必须满足的三个条件，它们都来自前面各节。第一，学席的读数必须来自中间账的四类读数而不是来自终局分数，否则它只是把”学生”改名为”学席”，落回旧账本。第二，学席必须允许”无法判定”，不能把模型参与程度直接换算成扣分，否则它会把耦合场再次压成一个数字，只是换了一个更现代的数字。第三，学席的真值必须能与校外的执业风险对接，否则它在校内再精细，也逃不出空值结算——分数的失真不会因为分数被改名而停止。

十七、反噬：申报表如何毁掉中间账

如果照本文推进，最省事的做法是给所有作业增加一张”AI 使用申报表”，让学生勾选独立、协作或外包，并把它作为新的合规指标。这个做法可能恰好毁掉本文想保住的东西，而且它极可能发生，因为它是旧账本吸收中间账的最低成本方式。

三步就够了。学生会把复杂的生产过程压成对自己有利的类别；学校会把来源栏变成新的纪律工具，勾选外包便降低成绩，不问模型究竟承担了什么；教师会把填表数量当作过程性评价。中间账于是成为旧账本的附属票据，来源栏没有打开路径，只增加了一张表。这正是第九节第三条征象的又一次显形：机构在承认失真之后，增加的是分数的种类而不是替换分数。

中间账只有在不把不确定性伪装成精确比例时才有价值。它必须允许”无法判定”，必须记录关键判断与责任承担，而不能把模型参与程度直接换算成扣分。它也不要求课堂成为全程监控空间。教师无法靠注意力覆盖所有生成过程，学生也不应被迫提交无限量的个人数据。中间账应当只记录对责任判断有用的节点：任务从何处开始，关键判断由谁提出，错误如何被发现，最终责任由谁承担。过度记录会制造新的行政负担，也会把教育转化为行为监控。

十八、实践意涵：先改时间配置，不先改题

教育机构若必须在多个改革环节中选第一步，应先改时间配置与中间账，而不是先改题目。理由来自第九节：改题目改变的是被测的东西，不改变结算的货币，因此不改变单位；改时间配置改变的是生成发生的位置，它直接作用于路径层。

把需要观察生成与责任承担的练习移入可见时段，可以增加路径读数：不是把所有作业搬进课堂，而是把关键判断的产生移进课堂。一篇论文的写作可以在课外完成，但论点的选择、对反例的处理、对一份新材料的即时回应，可以安排在课内十分钟完成并记录。这样安排的成本远低于全程监控，读数却落在耦合场最关键的位置。

把来源构成、关键判断与错误修正记录在终结性成绩之前，可以避免所有差异直接坍缩为一个数字。这一步的要点不是记录得多，而是记录得对。中间账应当记录四个节点：任务从何处开始，关键判断由谁提出，错误如何被发现，最终责任由谁承担。四个节点之外的过程可以不记。

第二项实践意涵是，学校不应把”是否使用人工智能”作为唯一政策变量。更重要的是使用发生在哪个环节：模型参与表达修订与参与概念选择不同，参与数据清理与参与研究结论不同。规则应从工具名单转向责任链条：不是列出允许与禁止的工具，而是列出哪些判断必须由学生作出并能被追问。

第三项实践意涵是资格制度必须引入现实风险的反馈。职业资格不能只由入场考试一次性决定，而应通过监督实践、案例复盘、错误报告、持续教育与责任追踪形成动态凭证。这不是取消考试，而是取消考试对资格真值的垄断。第十二节的飞行执照与医师再认证已经给出了可以借的结构。

这三项意涵的顺序不能颠倒。先改题目再改时间配置，新题目会在旧时间配置中被代理；先改中间账再对接执业风险，中间账会在校内自我封闭并被空值结算吸收。改革必须从时间配置进入路径层，从中间账进入显露层，从执业风险进入土壤层，三层同时动，单位才会动。

十九、边界与两个洞

本文的判断有三条边界，每一条都削掉一句原本可能说得过大的话。

第一，不可约性的判断不适用于所有人工智能使用。若模型只承担格式转换、语法校对或机械计算，不改变理解、构思与判断，耦合场对评价单位的冲击有限。本文不能说凡是使用人工智能都使学生单位失效，只能说当模型进入关键差异生成环节时，旧单位面临失配。

第二，职业准入优先改变的判断已在第十二节收缩：只在公共风险高、资格与执业许可紧密相连的领域，职业准入更可能成为外部判决的入口。

第三，题目代理化的判断受制于题目开放程度与模型能力边界。若任务包含无法预先获得的现场材料、不可复制的身体操作或必须承担的实时后果，代理路径可能无法低成本完成。本文不说”复杂题必然被代理化”，只说只要复杂题仍能被提前拆解、模板化或由模型预生成，复杂性就不能保证单位改变。

此外有两个本文解释不了的洞。第一个洞：在长期、低风险、兴趣驱动的学习中，有些学生主动用模型扩展问题意识而不是降低成本，主动增加困难、保留不确定性、追问模型的错误。代理坍缩机制解释不了为什么有些行动者不沿最低成本路径走。这里可能存在一种本文尚未说明的自我界面机制。第二个洞：某些教师—AI 组合可能提高评价公平而不是削弱它——模型帮助教师发现偏差、提供多种反馈、减少疲劳。本文能说明评分信用可能被重新分配，不能判断这种重新分配何时改善了教育正义。后续研究不能预设所有代理化都是负面结果。

二十、证伪条款

以下五条写明各自的状态：哪一条已经由第十至十二节的历史序列跑过，哪一条仍未执行。

条款一（未执行）：若在连续三个学期、不同学科与不同学校中，来源记录、责任追问和过程日志能够以低成本稳定区分独立、协作与外包三类完成，并且区分结果能进入正式成绩与资格决策，则”学生—AI 拒绝被约简”的判断被削弱。既有解释可以预测工具使用与成绩之间的关系，不能独立预测这种稳定分解；若它出现，说明旧账本已经发生结构性升级。

条款二（已由历史序列部分跑过）：若同一复杂题目在连续三届学生中未出现题库、模板、预生成答案或模型代理路径的收敛，且学生必须现场承担不可预先准备的判断，则”深度题会被代理化”的判断被推翻。第十节的合同代写序列已给出反向证据：开放性任务在人工智能之前十余年已被商品化为可定价的产品规格。该条款在生成式人工智能条件下的重跑仍待执行。

条款三（已由历史序列部分跑过）：若指标失真被系统记录之后，该指标在分配决策中的权重在五年内显著下降并被替代读数取代，则”空值结算”的判断被推翻，Goodhart 传统的”失真即退场”预测得到支持。第十一节的分数通胀序列给出反向证据：失真持续二十年而权重不降。该条款需要在其他指标体系（如科研计量、绩效考核）中重跑，以检验它是否只是教育的特例。

条款四（未执行）：若企业招聘、职业准入与学校考试长期保持相同的分数信任结构，即使执业中出现可观察风险，资格制度也不转向持续责任承担，则”外部风险账本将先于校园题目推动资格改变”的判断不成立。第十二节只提供了正向案例，没有提供反向案例的系统排除。

条款五（未执行）：若教育机构在不改变外部就业与资格制度的情况下，仅通过考题内容升级就持续减少代理路径、提高责任回收率并改善真实任务表现，则”考题改变不能自动改变单位”的判断错误。该结果将表明墙内制度具有独立重构土壤的能力。

条款六（未执行）：若教师在使用模型批改之后仍能以独立、稳定且可追责的方式解释每一条评分依据，且教师—AI 组合没有改变评分信用的归属，则第八节”新单位首先在记账侧合法出生”的判断被削弱。既有技术效率解释可以预测批改速度提升，不能预测教师专业责任是否被重新分配；检验方法是抽取模型参与批改的评语，让教师在不看模型输出的情况下复述评分依据，比较复述与原评语的一致性以及教师对不一致之处的归责。

对不利结果的处理原则是：不把未执行条款伪装成结果。过程记录与课堂现场可能确实能在某些学科稳定恢复学生责任；职业准入也可能因政策传统、行业利益或公共风险结构不同而迟迟不改变。本文的命题不是”任何地方都无法测量”，而是”在人工智能深度进入理解与判断、评价继续以终局分数分配机会的条件下，单位分解具有系统性困难”。

二十一、结论

三个研究问题的回答如下。

学生—AI 耦合场的路径由试探、识别、保留与成本比较筛选出来，土壤由时间压力、工具零成本、同伴比较、生产力叙事与升学竞争共同支撑。路径一旦成功，就通过”大家都在用”的现实反过来改变土壤；土壤再通过账本装瞎、分数同价与资格分配回喂路径。学生—AI 的显露不是人工智能单独造成的，而是路径与土壤共同作用之后被旧账本压缩出来的。

教育机构首先应改的配置是中间账，以及把关键产出从不可见时段转移到能观察生成、追问判断和记录修正的时段。课堂练习、现场材料、来源记录与责任回收，比立即设计更复杂的考题更能减少假读数。中间账不要求完美拆分人机贡献，只要求让路径差异不再被终局分数完全抹平。

职业准入与执业风险可能先于考题推动资格改变，是因为学校分数兼具测量功能与结算功能，且两者的维持条件不同：校准可以断，可通约性不断。这就是空值结算——失真的分数不退场，反而流通更广。考题代理化只改变题目表面，若墙外仍承认分数，墙内土壤便没有自我拆除的动力。职业现场以风险、错误后果与责任承担建立另一种真值，只有当学校分数在资格与就业中逐渐无法兑换，资格制度才可能先于考题改变。

本文交出的不是一个已经完成的新单位，也不是一套可以立即替代考试的技术方案。本文交出的是两个被命名的机制、一条判别式、三个指标、三条自我否决条款和五条写明状态的证伪条款。教育危机的核心不是人工智能进入了学生，而是旧账本仍把不可约的人机耦合场记作学生个体；改革若只提高题目难度，会制造更复杂的代理；若只扩大工具使用，会把新单位安置在评分者一侧；若只增加申报表，会把中间账重新压成合规数字。转向必须从”分数代表谁”开始，经过”产出怎样发生”，抵达”谁承担现实后果”。当来源路径、责任承担与职业风险同时进入资格制度，教育是否还能保留一个单一分数作为公共分配语言？如果不能，新的资格结构将以何种形式显露，仍有待现实中的第一本新账被记下。

参考文献

Campbell, D. T. (1976). Assessing the impact of planned social change. Occasional Paper Series No. 8, Public Affairs Center, Dartmouth College. Reprinted in *Evaluation and Program Planning*, 2(1), 67–90 (1979).

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.

Clarke, R., & Lancaster, T. (2006). Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. *Proceedings of the 2nd International Plagiarism Conference*, Newcastle, UK.

Collins, R. (1979). *The Credential Society: An Historical Sociology of Education and Stratification*. Academic Press.

Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. *Papers in Monetary Economics*, Reserve Bank of Australia.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press.

Labaree, D. F. (1997). *How to Succeed in School Without Really Learning: The Credentials Race in American Education*. Yale University Press.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.

Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education*, 3, 67.

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6).

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

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

UK Parliament. (2022). *Skills and Post-16 Education Act 2022*, ss. 32–41 (offences relating to cheating services).

General Medical Council. (2012). *The Good Medical Practice Framework for Appraisal and Revalidation*. GMC.