

被撤销的推定：论学校制度赖以运转的那条从未写下的假设

生成式人工智能撤销的不是任何一项技能，而是一条从未被明文写下、却支撑着作业、学分与文凭全部效力的制度推定

陈晓艳 著 · SDE 学员 · AI教育四联 · 2026年7月28日

摘要 生成式人工智能能够在极短时间内生成答案、文章和解题步骤，这使学校长期依赖的一项隐含判断变得不再可靠：不能仅凭一份完成度较高的认知产出，推断学生经历了与之相称的理解过程。本文把这条从未被明文写下、却支撑着作业、学分、文凭与选拔位次全部效力的假设命名为**产出—过程推定**，并主张：人工智能撤销的不是任何一项具体技能，而是这条推定本身。这一命名的关键后果是把问题从测量层移到授权层——即便检测技术做到完美，只要一份合规产出可以由非本人的过程生成，推定的授权基础就已失效；因此当前以识别、监控和防作弊为主的对策，修复的是测量精度，而失效的是授权资格，两者不在同一层。当前教育改革多从增加人工智能课程、提升数字素养或改进技术监管入手，却较少触及学校日常制度如何分配思考时间、组织课堂对话和处理学生错误。本文提出“自主意义建构能力”概念，指学生在缺少现成答案的情境中，能够形成初步理解、调用既有经验、检验假设、追溯错误并作出有理由判断的综合能力。文章依据有效失败、对话教学、认知能动性教育目的研究，分析三类可能压缩这一能力的制度机制：以知识覆盖率为中心的课时安排，以启动—回应—评价为主导的课堂互动，以及以正确完成为中心的作业评价。本文据此提出三项可试验的制度调整：设置可校准的自主探究时段，建立延迟评价的追问式对话，以及把高信息量错误纳入形成性评价。文章同时澄清，自主探索并不等于取消教师指导，错误也不会自动产生学习；有效的认知挑战必须与适时指导、概念整合和反馈相连接。本文最后提出准实验设计和明确的证伪条件，把“学校应从传递知识转向接生理解”由价值宣言转化为可以被检验的制度假设。

关键词：生成式人工智能；自主意义建构；有效失败；课堂对话；形成性评价；制度设计

1 引言

生成式人工智能进入学校以后，教育讨论迅速集中到几个显眼问题：学生是否使用人工智能代写，教师如何识别机器生成内容，学校是否应开设人工智能课程，以及怎样培养学生的人工智能素养。联合国教科文组织发布的相关指南强调以人为本、保护学习者权利、提升师生能力并审慎治理生成式人工智能（Miao & Holmes, 2023）。这些议题都很重要，但它们仍然可能绕过一个更基础的问题：当机器能够直接生成看似合格的学习成果时，学校如何确认学生真正经历了理解、判断和修正的过程？

过去，作业、论文与考试答案既是评价结果，也是推断学习过程的间接证据。高质量产出通常意味着学生掌握了相应知识，至少完成了相当程度的练习。生成式人工智能削弱了这种证据关系。它没有消除家庭文化资本、教师教学与学生努力的作用，也没有彻底切断学历与社会机会之间的联系；它改变的是“产出质量能够在多大程度上代表学生自身认知活动”这一判断的可信度。由此，学校不能再把提高产出效率视为教育质量的充分证明。

现有回应常采取加法策略：增加人工智能素养内容，增加过程监测工具，增加口头答辩，增加更复杂的评价指标。然而，如果学校日常运行依然要求教师在有限时间内完成高度密集的知识覆盖，课堂对话仍以教师迅速判定答案为终点，作业仍主要奖励正确和完整，那么新增内容很可能被旧的运行逻辑重新加工为新的知识点、新的检查表和新的竞赛项目。

这条被削弱的证据关系，值得有一个名字。本文把它称为**产出—过程推定**：学校在授予学分、评定等级、签发文凭和排定选拔位次时，始终依据一条从未写进任何章程的假设——**从一份合规的认知产出，可以反推出一段与之相称的认知经历**。这条推定之所以长期无需明说，是因为它长期无需辩护：在书写、演算和论证都必须由本人完成的技术条件下，产出与过程之间的连接近乎自动。学校制度不是选择了这条推定，而是建立在它上面。

把它命名出来，是为了看清一件容易被错置的事：**被撤销的是推定，不是技能**。学生仍然可以学会写作与推理，教师仍然可以教；改变的是学校**凭什么**认定某个学生已经做到。这个区分决定了对策落在哪一层。如果问题被理解为“学生用了不该用的工具”，对策就是识别、监控与防范，即修复测量精度；但即便检测做到完美——甚至假设存在一种绝无误判的鉴别技术——只要一份完全合规的产出在原理上可以由非本人的过程生成，推定的授权基础就已经不在了。**测量层的任何改进都不能修复授权层的失效**。本文由此得到一个可以被检验的推论：投入越集中于检测技术，学校越可能维持推定不变而只加固它的外壳，从而把资源持续投在无法解决问题的那一层上。

本文的基本判断是：人工智能对学校提出的根本挑战，不只是“如何使用一种新工具”，而是“如何保护学生不能被工具替代的认知经历”。本文把这种经历概括为自主意义建构，并集中考察三个研究问题：

第一，自主意义建构与批判性思维、创造力、元认知等相邻概念有何联系和区别？

第二，学校的课时、对话与作业制度如何影响自主意义建构的发生概率？

第三，怎样通过不依赖教师个人热情的制度调整，为这类认知活动提供稳定条件，并对其效果进行检验？

“从传递到接生”并不意味着学校停止教授知识。没有学科知识，学生的探索容易退化为无根据猜测。这里的“接生”强调的是：教师不只把既有结论交给学生，还要设计问题、控制提示时机、保护必要的思考时间，并帮助学生把未成熟的想法转化为更有结构的理解。知识传递仍然重要，但它需要嵌入学生真正参与的认知过程。

2 理论基础与研究缺口

2.1 教育不能被压缩为可测量产出

布迪厄与帕斯隆的文化再生产理论揭示，学校并非中立地评价能力，而会把特定阶层的文化经验转换为看似普遍的学业优势（Bourdieu & Passeron, 1990）。伯恩斯坦则说明，不同社会群体所熟悉的语言符码与学校知识组织方式之间存在结构性差异（Bernstein, 1971）。这些研究提醒我们，教育产出从来不只是个人能力的透明反映。

生成式人工智能进一步增加了产出解释的复杂性。资源优势家庭可以更早获得高质量工具、提示策略和人工支持，技术未必消除文化资本差异，反而可能改变文化资本发挥作用的方式。因此，本文不主张人工智能已经终结文化再生产，而主张它使“谁完成了认知工作、完成了何种认知工作”变得更加难以从结果判断。

比斯塔批评把教育简化为抽象的“学习”或可测量表现，强调资格获得、社会化与主体化是不可相互替代的教育目的（Biesta, 2010, 2015）。本文沿着这一思路进一步追问：如果教育还承担帮助学生形成判断和责任能力的任务，那么学校的日常制度是否为这种能力留下了实际空间？价值目标只有进入课时、对话和评价规则，才不会停留在课程文件中。

2.2 有效失败及其边界

有效失败研究表明，在正式教学之前，让学生尝试解决尚未完全掌握的问题，有时能够促进概念理解和迁移。关键并不是“先失败”本身，而是学生在尝试中激活既有知识、暴露知识缺口、比较不同方案，随后再通过教学完成概念整合（Kapur, 2008, 2016）。Loibl、Roll与Rummel（2017）的综述进一步指出，问题解决先于教学何时有效，取决于任务设计、先备知识、错误比较和后续指导等边界条件。

这意味着不能把认知困境浪漫化。缺乏先备知识、任务过难、没有反馈或后续教学的探索，可能只会增加挫败感。反过来，教师过早给出答案，也可能使学生失去识别知识缺口和形成初步表征的机会。真正需要设计的不是“指导”与“无指导”的二选一，而是指导的时间、强度和形式。

2.3 与效度理论的分界：为什么这不是一个测量问题

把产出—过程推定说成“产出的效度下降了”，是最容易、也最误导的一种翻译。效度理论确实提供了最接近的语言：Messick（1989）把效度界定为证据与理论对分数解释与使用的支持程度，Kane（2013）则把效度论证组织为从观察到结论的一串可检验推论。在这套语言里，“产出不再代表学生的认知过程”会被归入构念无关变异——一种污染了分数解释的干扰源。

这个翻译丢掉了要害。效度是**测量的属性**，它的修复手段也在测量之内：改进任务设计、控制施测条件、增加证据来源、收紧推论链。产出—过程推定不是测量属性，而是一条**制度授权推定**——它不回答“这个分数有多准”，它回答“凭什么由学校来认定这个人已经做到”。两者可以彼此独立地失效：一份在测量意义上高度可靠的作业，如果其产出过程在原理上可由他者代行，它在授权意义上已经不作数；反过来，一次测量粗糙但全程可追溯的口头答辩，授权基础完好而测量精度很差。

这条分界不是术语洁癖，它直接决定资源投向。构念无关变异的标准处理是**加**：加监考、加检测、加过程留痕、加更复杂的评分规则。若失效真的发生在测量层，这些做法有效。但如果失效在授权层，同样的加法只会把外壳加厚——学校获得了更高的检测准确率，却没有恢复任何“凭什么认定”的根据，因为根据从来不是准确率给的。本文后续三节讨论的课时、对话与作业安排，其意义正在于此：它们不试图更准地测量产出，而试图让**认知经历本身重新成为学校可以直接组织和直接观察的对象**，从而使授权不再依赖那条已被撤销的反推。

需要说明的是，本文并不主张效度理论在此无用。恰恰相反，一旦学校把观察对象从产出移到过程，Kane 意义上的效度论证立即变得必要且更难：过程证据的评分者一致性、任务间迁移、跨情境稳定性，每一项都需要证据支持。分界要说的只是次序——**先解决凭什么认定，再解决认定得准不准**；把次序颠倒，就会用越来越精密的方法去测量一个已经不再具有授权效力的东西。

2.4 课堂对话与评价惯性

经典课堂话语研究发现，学校互动常呈现启动—回应—评价结构：教师提出一个已知答案的问题，学生作答，教师判断对错并结束回合（Mehan, 1979; Cazden, 2001）。这种结构有助于快速检查知识，却可能使学生把认知努力转向猜测教师期待的答案。

IRE或IRF结构本身并非必然有害。教师的第三步回应也可以用于追问、重新组织学生观点或邀请其他学生参与。问题在于，当“评价”长期成为几乎所有对话的固定终点，学生就很少有机会公开表达未完成的想法。对话教学研究强调，课堂谈话应具有累积性、互惠性和目的性，使观点能够被阐明、质疑和发展（Wells, 1999; Alexander, 2020）。因此，制度改革不应机械禁止评价，而应改变评价出现的时机和功能。

2.5 研究缺口

现有研究分别讨论教育目的、有效失败、课堂对话和技术风险，却较少把它们放入学校制度的同一分析框架。本文的贡献在于，把自主意义建构视为一种需要特定制度条件支持的认知活动，并分析三种日常规则如何共同改变其发生概率。这些制度条件需要通过比较研究不断校准，不能被理解为适用于所有学校、学科与年龄阶段的固定处方。

3 核心概念与理论命题

3.1 自主意义建构能力

自主意义建构能力，是指学习者面对无法直接套用既有答案的问题时，能够形成初步问题表征，调用相关经验和知识，提出并检验假设，识别矛盾、追溯错误，最终作出有理由且可以修订的判断。

这一概念包含五个彼此联系的过程。

第一，进入问题。学生能够承认暂时不知道答案，并把“不知道”转化为需要辨认的具体困难。

第二，生成初步解释。学生根据已有知识形成暂时性理解，而不是立即复制权威答案。

第三，建立联系。学生能够把新问题与既有概念、个人经验或其他学科知识连接起来。

第四，检验与回溯。学生在发现矛盾时，能够返回推理过程，识别假设或证据出现问题的位置。

第五，形成可辩护判断。学生能够在多种可能之间作出选择，说明理由，并保持根据新证据修订判断的开放性。

它与批判性思维、创造力和元认知存在交叉。批判性思维强调评价证据和论证，创造力关注新颖且适切的产出，元认知关注对自身认知活动的监控。自主意义建构聚焦的是这些能力在一个完整问题过程中如何被组织起来。它不是一种神秘的内在属性，而是一组可以通过行为、话语和过程材料观察的活动。

3.2 认知过程与认知产出

认知产出是问题解决后的可见结果，如答案、论文或方案；认知过程是学习者形成、检验和修订理解的行动轨迹。二者通常相关，却不能相互等同。人工智能扩大了二者分离的可能性：学生可以获得高质量产出，却没有经历相应的推理；也可以在尚未形成正确答案时，已经完成了具有学习价值的假设比较和错误识别。

因此，学校评价需要同时关注结果与过程。否定结果会使教育失去知识标准，只看结果则可能误判学习。本文主张的是双重证据结构：既考察学生最终理解是否符合学科规范，也考察这一理解如何形成、是否能够迁移、能否在没有人工智能代答时重新建构。

这里需要挡住一个容易发生的收缩：双重证据结构听起来像是“多加一类证据”，仿佛问题出在证据不够多。但按上文的分层，加证据本身仍是测量层的动作。使双重结构有意义的不是证据的数量，而是**观察对象的更换**——把学校日常直接组织、直接看见的东西，从“学生交上来的东西”换成“学生正在做的事”。只有当认知经历成为学校可以直接安排时间、直接组织对话、直接处理错误的对象时，授权才不再需要那条反推。下文三节各自对应这三个直接性：时间、对话、错误。

3.3 六项理论命题

命题一：推定撤销命题。生成式人工智能撤销的是产出—过程推定的授权效力，而非产出的测量精度；因此检测准确率的提高不会恢复推定，只会提高维持推定的成本。可检验的形式是：在检测技术显著改进的学校，用于识别与监控的资源占比上升，而学生在无人智能辅助条件下的开放问题表现不因此改善。若检测能力提升伴随后一项指标的同步改善，本命题的分层主张即被否定。

命题二：时间条件命题。当课程持续以知识覆盖速度为首要目标时，学生进入开放问题、形成初步解释和修订假设的时间会减少。

命题三：对话条件命题。当课堂第三回合主要承担即时判断对错的功能时，学生公开表达不确定性和推理过程的频率会下降；当第三回合转为澄清、追问和观点连接时，这些行为会增加。

命题四：评价条件命题。当作业只奖励最终正确性时，学生倾向于隐藏或快速抹除错误；当高信息量错误可以获得分析性反馈时，错误回溯行为会增加。

命题五：阶层差异命题。学校提供的自主意义建构机会越少，家庭资源对学生获得这类机会的影响越大；公共学校若能稳定提供相应条件，可能降低而不是扩大这一差异。

命题六：组合效应命题。探究时段、追问式对话和错误分析三项安排的组合效应，应高于任一单项安排，因为它们分别作用于时间、互动和评价三个环节。

4 三个制度节点

4.1 从固定覆盖率到可校准的探究时段

课程需要覆盖核心知识，这是公共教育保证基本质量的重要方式。问题并非“覆盖知识”本身，而是覆盖率一旦成为压倒性指标，教师就会被迫把学生的停顿、绕路和错误视为进度风险。学生刚刚形成一个值得追问的问题，课堂却必须立即进入下一个知识点；久而久之，学生会学会等待标准方法，而不是形成自己的问题表征。

本文建议在具体学科单元中设置可校准的自主探究时段。它不是单独增加一门“思维课”，也不预先规定统一比例，而是在单元设计时根据知识性质、学生先备水平和课时条件，划出一段不以即时正确率和内容覆盖为直接目标的时间。

探究时段应满足四项条件：

问题具有明确的学科边界，但不存在一眼可套用的现成步骤；

学生先独立或合作形成初步方案，再获得教师的针对性指导；

教师记录学生使用的表征、假设和困难，而不只记录最终答案；

探究结束后必须进行概念整合，使学生理解尝试与学科知识之间的关系。

“20%课时”可以作为部分试点的起始参数，但不能被写成普遍标准。不同学科、年级和学生群体所需比例可能显著不同。真正需要检验的是剂量—效果关系：探究时段增加到何种程度能够改善概念理解与迁移，又在何种程度上会损害必要知识的系统学习。

4.2 从即时评价到延迟评价的追问式对话

教师评价并不是课堂对话的敌人。没有评价，学生无法知道自己的理解与学科标准之间的距离。问题在于评价过早出现，会提前关闭尚未成熟的思考。

本文建议建立延迟评价的追问式对话。当学生给出答案后，教师不立即宣布对错，而是至少进行一轮澄清或推理追问，例如：“你依据的线索是什么？”“如果条件改变，这个结论还成立吗？”“谁能用另一种方式解释？”“你现在最不确定的是哪一步？”

此类对话需要防止三种形式化倾向。第一，追问不能变成教师用更多问题把学生逼向预定答案；第二，延迟评价不能变成永不评价，讨论最终需要回到学科证据和概念标准；第三，不能用固定次数替代质量判断。“每学期两次”可以作为最低试点频率，却不足以证明对话结构已经改变。更有意义的指标是：学生是否更频繁地陈述理由、识别不确定性、回应同伴观点，并在没有教师提示时主动修订判断。

4.3 从统一订正到高信息量错误分析

传统作业制度倾向于把错误视为需要尽快消除的缺陷。统一订正可以提高效率，却常把“为什么会错”压缩成“正确答案是什么”。有效失败研究提示，错误只有在暴露知识缺口、进入比较并被后续教学整合时，才具有学习价值（Kapur, 2016; Loibl et al., 2017）。

本文提出高信息量错误分析。教师从学生作业中选择能够揭示概念误解或推理分叉的错误，引导学生重建当时的思考路径：

我当时依据了什么假设？

哪个证据或条件被忽略？

错误最早出现在哪一步？

如果重新解决，我会改变什么？

这一错误能提醒我处理哪一类新问题？

错误分析不宜直接奖励“犯错数量”，否则会产生策略性错误；也不应要求学生写情绪化检讨。评价对象是错误分析的质量，而不是错误本身。教师还可以把匿名化错误用于教研，识别某类误解究竟来自学生先备知识、教材表述还是教学设计。

5 竞争解释与研究设计

5.1 最强竞争解释

本文面临的最强竞争解释是：上述效果无需特殊制度安排，只要教师教学质量提高、学生学习时间增加、班额减少或资源改善，学生的开放问题表现自然会提升。如果这一解释成立，那么自主意义建构只是一般教育质量的副产品。

另一个竞争解释是新奇效应。学生可能因为试点活动不同于常规课堂而短期投入增加，效果并非来自探究、对话与错误分析机制。

第三个竞争解释是教师期待效应。参与试点的教师更认同改革，可能无意中给予实验组更多支持。

5.2 准实验设计

可以在多所背景相近的学校中设置三组。

第一组实施组合制度调整：可校准探究时段、延迟评价对话和高信息量错误分析。

第二组获得等量教师培训、优质练习材料和额外自主学习时间，但不改变课堂对话与作业评价结构。

第三组维持常规教学，用于估计新奇效应和一般时间趋势。

三组投入总量应尽量匹配，教师均接受观察者盲法的视频编码。前测和后测同时测量：

标准化学业成绩；

无人工智能辅助的开放问题表现；

学生独立坚持时间；

理由陈述与假设修订次数；

发现矛盾后的错误回溯行为；

在新情境中的概念迁移。

过程指标应由多名编码者依据公开规则独立评分，并报告一致性。还需记录人工智能使用方式，避免把完全禁止人工智能与允许人工智能但要求过程披露的结果混在一起。

5.3 证伪条件

如果在控制前测水平、教师差异、家庭背景和投入总量后，组合制度调整组在开放问题、错误回溯与迁移指标上并不优于一般教学改进组，本文关于三个制度节点具有特殊作用的主张应被否定或收缩。

如果组合制度调整只提升学生参与感，却持续降低核心知识掌握，说明制度设计没有解决探索与指导的平衡问题。

如果任何单项调整都能获得与组合方案相同的效果，则“三个节点具有组合效应”的命题不成立，学校没有必要同时改变三类制度。

如果效果仅在高社会经济地位学生中出现，则本文关于学校制度能够减少机会差异的判断受到反驳，必须重新分析任务难度、语言要求和支持方式。

6 实施原则与风险控制

6.1 不增加独立行政层级

制度调整应嵌入现有学科教学、课堂互动和作业反馈，而不是新设一个负责“创新能力”的行政部门。新增部门容易制造材料和检查，却不一定改变日常课堂。

6.2 不把试点立即转化为绩效排名

如果教师一开始就需要证明探究时段提高了成绩，他们会倾向选择容易成功的任务并隐藏失败。试点初期应以问题发现和设计修订为目标，待机制稳定后再讨论规模化评价。

6.3 为教师提供对话与诊断训练

追问式对话并非“少讲一点”即可实现。教师需要判断何时保持沉默、何时提供提示、何时要求证据、何时完成概念收束。错误分析

也要求教师具有学科诊断能力。制度调整如果不配套专业学习，可能把教师推入新的无力感。

6.4 保留知识标准

自主意义建构不是“每个答案都有道理”。学生可以提出不同解释，但解释必须接受证据、逻辑和学科规范的检验。学校从传递转向接生，改变的是学生抵达知识的方式，不是取消知识的公共标准。

6.5 明确人工智能的角色

人工智能既可能替代认知过程，也可能支持认知过程。若学生在形成初步解释之前直接索取完整答案，工具容易提前关闭探索；若学生先提交自己的假设，再用人工智能生成反例、比较方案或获得分层提示，工具可能扩大反馈资源。制度设计需要规定使用阶段、披露方式和责任边界，而不是笼统地禁止或放任。

7 讨论

本文对原有判断作出三项重要限定。

第一，人工智能没有切断文化资本与教育结果之间的关系，而是削弱了可见产出对独立认知过程的代表性，并可能重组文化资本发挥作用的方式。

第二，学校制度并非彻底“清除”了自主意义建构，而是在特定压力组合下系统性降低了它出现的概率。不同学校、教师和学科之间存在重要差异。

第三，从传递到接生不是用自主探索取代直接教学。有效学习需要在知识讲解、问题尝试、反馈和整合之间形成合理次序。没有指导的失败不必然有效，过早的指导也可能使学习者失去生成问题表征的机会。

第四，本文只处理了推定被撤销之后学校内部的三个环节，没有处理两件同样要紧的事，它们由同批发表的另外三项研究分别承担。其一，本文提出的三项制度调整，恰恰是最容易被组织重新加工成新任务与新检查表的一类改革；《意义再生循环的锁定与修复》说明了这种加工在什么条件下必然发生——当一所学校的承诺与后果长期失配，任何新增安排都会被成员合理地解释为新一轮考核，探究时段会变成新的备课材料，错误分析会变成新的记录表。本文给出的是应当改什么，那篇给出的是在什么状态的学校里改会失败。其二，产出一个过程推定并不只在学校内部起作用，它同样是家长据以判断孩子的依据；《评价信任的裂变》处理的正是推定在家庭端的分化，并指出一个对本文构成直接风险的机制——本文所倡导的过程证据一旦被赋分并绑定稀缺机会，就会重新压缩为排名，届时“保护认知经历”会变成新的竞赛赛道。其三，本文假定教师有余裕作出情境判断并把它带入正式沟通；《关系基础退化》说明这一余裕本身是有条件的，且正在被数据化管理侵蚀。

这些限定没有削弱文章的核心主张，反而使其更加可检验：学校真正需要改变的，不是一套抽象教育哲学，而是知识、时间、对话和评价的具体组合方式。人工智能越能高效提供结论，学校越需要保护学生形成、检验和修订理解的过程。

8 结论

生成式人工智能使“完成一份好答案”与“形成自己的理解”之间出现更大距离。教育改革如果只增加人工智能课程、监控工具和数字评价，可能提高管理复杂度，却不一定恢复学生的认知参与。

本文把自主意义建构界定为进入问题、生成解释、建立联系、检验回溯和形成判断的连续过程，并分析课时覆盖、即时评价对话和正确完成导向如何压缩这一过程。相应的制度调整不是取消知识教学，而是重新安排指导的时机：给学生一段形成初步解释的时间，用追问延迟过早关闭，以高信息量错误推动概念整合。

“接生意义”不是浪漫隐喻，而是一项严格的专业任务。教师要设计有边界的问题，识别学生当前表征，控制提示强度，组织证据比较，并最终帮助学生进入学科知识。

最后回到那条被撤销的推定。学校长期不必说出它，是因为它长期不会失效；如今它失效了，学校才第一次被迫说出自己凭什么认定一个人已经做到。这个被迫说出的时刻不是危机，而是一次少有的机会：一条无需辩护的推定，其代价是从不接受检查——它把认知经历默认为产出的副产品，因此从不要求学校为这段经历安排任何时间。推定作废之后，学校若要重新获得认定资格，就只剩一条路：把认知经历从被反推的东西，变成被直接安排、直接看见、直接负责的东西。学校未来的不可替代性，不在于比人工智能更快地提供答案，也不在于比它更准地识别答案的来源，而在于它是少数仍然可以为一段理解的发生本身留出时间、并为这段时间负责的地方。

参考文献

- Alexander, R. (2020). A dialogic teaching companion. Routledge.
- Bernstein, B. (1971). Class, codes and control: Volume 1. Theoretical studies towards a sociology of language. Routledge & Kegan Paul.
- Biesta, G. (2015). What is education for? On good education, teacher judgement, and educational professionalism. European Journal of Education, 50(1), 75–87. <https://doi.org/10.1111/ejed.12109>
- Biesta, G. J. J. (2010). Good education in an age of measurement: Ethics, politics, democracy. Paradigm Publishers.
- Bourdieu, P., & Passeron, J.-C. (1990). Reproduction in education, society and culture (2nd ed.). Sage.
- Cazden, C. B. (2001). Classroom discourse: The language of teaching and learning (2nd ed.). Heinemann.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kapur, M. (2008). Productive failure. Cognition and Instruction, 26(3), 379–424. <https://doi.org/10.1080/07370000802212669>
- Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educational Psychologist, 51(2), 289–299. <https://doi.org/10.1080/00461520.2016.1155457>
- Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. Educational Psychology Review, 29(4), 693–715. <https://doi.org/10.1007/s10648-016-9379-x>
- Mehan, H. (1979). Learning lessons: Social organization in the classroom. Harvard University Press.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). American Council on Education & Macmillan.
- Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Wells, G. (1999). Dialogic inquiry: Towards a sociocultural practice and theory of education. Cambridge University Press. <https://doi.org/10.1017/CBO9780511605895>

本次打磨说明

本批打磨未另起附录，改动一律就地落在原文相应位置。作者的论证结构、章节次序与全部经验主张不动，改动只做三件事：把已经写在文中、却被放在次要位置的最锐判断提到主线并给它一个名字；补上原稿放跑的最强近邻并划出分界；把彼此咬合的四篇之间的互引写成可检验的一段分工，而非一句客套。

换题与主线。 原题《从传递到接生：人工智能时代学校教育的制度性复位》以苏格拉底旧喻为主线，而全文最锐的判断其实在命题一：产出不再是认知过程的可靠代理。本次把它提为脊梁并命名为「产出—过程推定」，题目随之改为《被撤销的推定》。「接生」的全部内容保留在第四至第六节，未删一字，只是不再充当主线。

新增 2.3「与效度理论的分界」。 原稿最大的风险是被读成「产出的效度下降了」——若如此，本文即可被还原为 Messick 意义上的构念无关变异，无独立价值。新增一节交代分层：效度是测量的属性，其修复手段在测量之内；产出—过程推定是制度授权推定，回答的是「凭什么由学校来认定这个人已经做到」。两者可彼此独立地失效。这一分界直接决定资源投向：若失效在测量层，加监考、加检测、加留痕有效；若在授权层，同样的加法只把外壳加厚。

命题一改写。 原命题（「证据弱化命题」）只说可靠性下降、程度受调节，实质不可证伪——任何结果都能被读成「调节变量在起作用」。改为「推定撤销命题」，并写死可测形式：检测技术显著改进的学校，识别与监控资源占比上升，而学生在无人工智能辅助条件下的开放问题表现不因此改善；若两者同步改善，分层主张即被否定。

3.2 补一处收缩防线。 「双重证据结构」易被读成「多加一类证据」，那仍是测量层动作。补一段说明：使它有意义的不是证据数量，而是观察对象的更换。

讨论新增第四条限定。 交代本文未处理而由同批另外三篇承担的三件事，并点明其中两条对本文构成直接风险：本文提出的三项制度调整正是最容易被组织加工成新任务与新检查表的一类（见《空转的公开性》），而本文倡导的过程证据一旦被赋分绑定稀缺机会即重新压缩为排名（见《指标化吸纳》）。

结论收口重写。 由「学校要比 AI 更慢更深」这类价值宣示，改为回到被命名的物件：一条无需辩护的推定，其代价是从不接受检查——它把认知经历默认为产出的副产品，因此从不要求学校为这段经历安排任何时间。

本次打磨所核验的文献

- Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13–103). American Council on Education & Macmillan.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

(原稿已引、本次逐条核对无误) Kapur, M. (2008/2016); Loibl, Roll & Rummel (2017); Mehan (1979); Cazden (2001); Biesta (2010/2015); Bourdieu & Passeron (1990); Bernstein (1971); Alexander (2020); Wells (1999); Miao & Holmes (2023, UNESCO).