

观点与分析

编辑日期: 2026年10月29日

错误消失了的学徒

> 问题并不在于机器已经精通新手过去做的事，而在于：一些看似最容易取消的工作，曾经是不写在工资单上的学校，专家正是从那里毕业。

2013 年 1 月 4 日，美国联邦航空管理局（Federal Aviation Administration）发布了一页纸的安全提示，标题是“手动飞行操作”。它没有浪漫地号召飞行员重回操纵杆，也没有把自动驾驶说成一项糟糕发明。相反，文件先确认自动飞行系统改善了安全、工作负荷管理与运行精度，随后指出一个更安静的问题：持续使用这些系统未必能强化手动控制所需的知识与技能，当飞机偏离理想状态时，还可能削弱飞行员迅速恢复控制的能力。1

建议十分具体：在工作负荷较低的适当条件下提供手动练习机会，同时在安全与精度需要时继续使用自动系统。2017 年，管理局再次强调，应在训练和运行中建立并保持手动操作能力，而不是把它变成一场比赛，用飞行员挑战自动化的分钟数来计分。2

这种担忧也不只来自一页提示。2013 年驾驶舱自动化工作组的报告，在更广泛的技术审查中讨论了飞行员训练和飞行路径管理，强调运行与训练设计不能同自动系统本身的特征分开考虑。13

值得思考的不是对机器驾机的恐惧，而是文件把一个问题放到了大多数管理报告不喜欢的位置：如果机器能把日常工作做得很好，人类应对异常时刻所需的经验从哪里来？

在软件办公室、合同审查部门或诊断实验室里，这个问题看起来没有那么戏剧化。这里没有驾驶舱，也没有数千米高度。但相似的东西可能正在这些职业中消失：安全犯错、修正、重试，以及解释第一个方案为什么被拒绝的时间。按企业会计看，它们是缓慢的成本；按职业经济看，它们却是持续积累的资产。

当学徒的错误从工作中消失，随之消失的也可能是让他成为专家的一部分路径。

一项小工作，承担着一项大功能

企业为什么想自动化入门任务，很容易理解。新工程师编写常规测试；法律分析师检索先例；银行员工审查相似文件；实习医生在专科医生监督下起草病例摘要。如果所有这些都能更快、更便宜地完成，谁会为了缓慢本身要求保留缓慢？

但机构里的任务，并不总等同于岗位说明中看到的样子。一项任务可能同时有两种功能：

第一，为今天生产可用输出：能运行的程序、清楚的备忘录或正确的审查。第二，为完成任务的人生产能力：识别模式、形成判断，并学会看穿那些听起来合理却错误的答案。

第一种价值立刻卖给客户，第二种价值多年以后才会显现。第一种出现在经理的周报看板上；第二种往往要等老专家离职，团队才发现没有人学会他如何思考。

这里必须区分转型话语经常混在一起的三件事：完成任务、获得技能、保持技能。一个系统可以在第一项上成功，却对第二项毫无贡献；工具可能帮助人现在获得知识，但我们并不知道六个月后是否仍然保留；员工也可能精通一项工作多年，却因缺少练习而在其中一部分上退步。每种现象都有不同机制，也需要不同证据。

在人力资本（Human Capital）经济学中，Yoram Ben-Porath 早在 1967 年就讨论了个人如何投入时间和资源，以生产未来能力，而不只生产当期收入。[3] 这类职业投资常有一副不可见的面孔：按即时商业标准看似“没有产出”的一部分时间，正在创造明天的生产率。

企业如果删掉全部这类时间，就不只是缩短流程；它可能已经改变了专业能力的生产方程。

反过来，机器辅助下的任务提速也不是幻觉。一项在 2023 年发布初稿的对照实验中，使用 GitHub Copilot 的开发者完成网页服务器编程任务，比未使用者快 55.8%。但实验测量的是特定任务的完成时间，并未设计成测量参与者关闭工具以后学到了什么。[14] 两个问题之间的差别，正需要得到公允对待。

把成果与学习分开的实验

Judy Hanwen Shen 与 Alex Tamkin 在 2026 年 1 月 28 日发布的一篇初步论文中研究了一个具体问题：开发者学习新的软件库时使用人工智能助手，技能形成会发生什么？论文名为 *How AI Impacts Skill Formation*，目前是尚未经同行评审的研究草稿；这一状态对判断证据强度十分重要。[4]

研究者选用了 Python 异步编程库 Trio。参与者并非编程新手；每人都有一年以上 Python 经验，许多人经验更长，但此前没有使用过实验中的软件库。这个细节改变了最顺手的解读：实验比较的不是“会编程的孩子”和“会编程的机器”，而是已有相对经验的人如何面对一项新技能。

主研究共有 52 名参与者，平均分成两组。26 人获得编程助手，26 人没有。所有人都完成限时任务，之后再参加独立测试，测量概念理解、代码阅读和调试能力。

值得注意的不只是辅助组测试分数更低，而是任务完成与学习朝着并不一致的方向移动。人工智能辅助组的技能测试平均约为 50%，另一组约为 67%，相差 17 个百分点；在这项测试中差异具有统计显著性。与此同时，主实验没有发现平均任务完成时间得到显著改善。4

这里，审慎比数字的冲击力更重要。必须记录四项限制：样本相对较小；接触时间较短；测试与训练在时间上很接近；研究领域是学习一个特定软件库，不是程序员整体能力，也不是未来十年编程职业的命运。人工智能组还全部完成了第二项任务，而对照组有四人在时限内未能完成。这意味着，仅仅改变“生产率”的定义，也可能改变结果解读。这项研究不是人工智能毁掉程序员的证书，而是有关学习设计中可能存在权衡的实验证据。

研究者分析使用记录时，还发现完全把工作交给工具的人，与先询问概念解释或错误原因、随后继续思考的人之间存在差异。要求认知投入的互动模式与更好的学习结果相关。但这一部分是辅助组内部的观察性分析，不能单独被提升为适用于所有互动模式的已证实因果效应。

教训比“少用人工智能”更精确。应当问：使用工具以后，还有哪一部分思考留在学习者身上?

错误如何成为训练装置?

设想两名开发者面对同一条报错。第一人把报错复制给语言助手，得到一个能够运行的修复方案。第二人查阅软件库文档，测试两个假设，其中一个失败，最后发现问题不在某一行的写法，而在异步事件的顺序。两人都得到正确代码，认知过程却不同。

这是虚构示例，并非 Shen 与 Tamkin 研究中某位参与者的经历。它的用途是解释机制，而不是扩大研究结论。

第二种情况包含教育心理学所谓的提取练习（Retrieval Practice）：先从记忆中调取知识，再通过反馈比较预期与现实。Henry Roediger 与 Jeffrey Karpicke 2006 年的实验显示，在某些实验条件下，提取测试对后续保持的促进，可能超过即时表现所暗示的程度。5 这不是说每一次失败都有益，而是说提取与验证的尝试不是可以自动删除的学习废料。

Manu Kapur 在“生产性失败”（Productive Failure）研究中说明，如果最初未能得到正确解的尝试经过教学设计，并有后续处理与解释，它可能为更深学习作准备。6 但把这一理念缩成“让员工犯错”，会非常危险。错误只有在成本受控、有人解释、有重试时间，而且学徒能把错误尝试连接到更准确的原理时，才可能有益。把学徒在无人监督下投入高风险案件，不是学习，而是穿着教育勇气外衣的管理失职。

经验不在于一个人犯过多少错误，而在于他理解自己为什么错以后，心智模型改变了多少。

一台新机器里的古老悖论

语言助手出现几十年前，Lisanne Bainbridge 于 1983 年发表了著名论文《自动化的反讽》（*Ironies of Automation*）。其中一个核心观点是：自动化可能把人推离日常练习，却又要求他在最困难的情况中介入。机器负责常规，人负责异常；但常规工作原本提供了应对异常所需的一部分训练。7

这种类比并不完整。驾驶飞机是复杂的感知运动技能，阅读合同是制度性解释技能，调试代码则是逻辑和实验问题。它们不会以同样速度衰退，也不能从一个生动比喻推导出同一种影响。自动化还可能培养新的能力，例如撰写规格说明、审查模型输出。

所以，假设不是所有手工练习都是美德，而是：消失的是哪一种练习，未来要求人作出的又是哪一种判断，二者必须一起研究。

Brooke Macnamara 团队在 2024 年的一篇学术文章中，重新提出自动辅助是否会加速技能流失的问题。这篇文章的价值在于界定研究问题与可能机制，而非已经通过实验证明所有职业都发生普遍退化。12

Stephen Casner 等人的实验发表于 2014 年的 *Human Factors*，为轻率的悲观提供了必要修正。研究者使用波音 747-400 模拟器测试 16 名飞行员，发现基本手动操作表现的若干方面仍保持稳定，一些认知任务则更容易退化。8 如果规律真是“所有自动化都会抹去所有技能”，就不会出现这种差异。更准确的教训是，我们必须知道究竟什么在衰退：动作、注意、诊断、系统理解，还是压力下的工作协调。

如果软件屏幕总能给出正确答案，经理很容易以为自己消灭了错误。他也许只是消灭了那些曾让新手思考方式显露出来的时刻。

医生看得更多，离开屏幕以后呢？

一项 2025 年发表于 *The Lancet Gastroenterology & Hepatology* 的观察性研究，追踪了波兰四家结肠镜中心引入人工智能辅助后的影响。研究不只测量系统在场时的表现，还观察引入人工智能前后，没有自动辅助的常规检查表现。

比较包括 1,443 次未使用人工智能的检查，其中 795 次在技术引入前，648 次在引入后。研究样本中的腺瘤检出率（Adenoma Detection Rate）从 28.4% 降至 22.4%，相差 6 个百分点。研究者谨慎表述：持续接触辅助系统，可能意味着检查者行为发生变化。9

这里必须抵抗轰动。研究不是能够证明机器导致技能下降的随机实验；两个时期的患者构成、中心行为或其他因素都可能发生变化。调整后的分析也不能排除全部竞争性解释。“检出率”也不是临床能力或患者结局的唯一同义词。但它提出了一个值得独立研究的信号：当工具负责提醒人类那些细小目标时，人类观察会发生什么？它可能从专家主动进行的活动，变成他等待出现的结果。

医疗中不能为了训练未来的医生，牺牲眼前患者的安全。这个问题也不能靠笼统歌颂手工操作来解决。替代方案可以是在结果已知的影像病例上训练、进行盲法模拟练习，或协作审查系统捕捉的图像，同时监测独立能力，而不剥夺患者已经证实对其有益的手段。

更重要的是，证据并非只有一个方向。2026 年 6 月发表的一项有关模拟外科技能教学的系统综述与荟萃分析，定量分析纳入 4 项研究和 268 名参与者。人工智能增强教学在多个指标上与专家教学相近，并在一项实际技能指标上呈现小幅优势；但证据确定性较低，研究者并不认为这足以取代人类教师。10

这项相反证据很宝贵，因为它把问题重新表述为：工具究竟是在教学，还是在破坏学习？还不够。真正的问题是：教学方法是什么？技能属于哪一类？反馈如何提供？在不容许频繁犯错的实践中，又由谁判断学徒已经准备好？

在生产中，一条提示可能等于完整答案；在学习中却不等于

2023 至 2024 学年，一项土耳其高中现场实验检验了两种人工智能辅助对数学学习的影响，参与者接近一千名学生。结果于 2025 年发表于 *Proceedings of the National Academy of Sciences*。11

第一种是类似一般聊天界面的 GPT-4 模型，研究者称之为“GPT Base”。第二种“GPT Tutor”在教师指导下设计，通过提示、常见错误与正确解法知识引导学生，而不是急于交出最终答案。还有一组学生不使用这两种助手。

练习期间，第一组相对对照组平均提高 48%，使用教学型助手的一组相对提高 127%。任何只对成果满意的营销演示都可能在这里结束，但实验没有结束。

学生在无辅助条件下接受测试时，一般助手使用者的成绩比对照组相对低 17%。教学型助手把负面影响减弱到与对照组差异不再具有统计显著性，但它也没有在独立测试中取得经证实的优势。

不能把数学实验中相对下降 17%，与Trio 程序员技能测试均值相差 17 个百分点混为一谈。数字相似，并不意味着指标、群体和训练时长相同。

更重要的结果是，学习的基本模型不只是让学生答对。完整答案会立即消除不确定性，提示则迫使学习者自己缩小它。在特定教学环境中，给出更少信息的系统，反而可能保护更大的教学功能。

这不是为故意让人无知辩护。有效提示需要准确知道学生处在什么位置、为什么受阻，也需要可靠的纠错。因此，教学型版本并不只是语气更温和：它针对题目专门配置，并有教师实质参与。把“不要直接给答案”复制进战略文件很容易；难的是建造一个知道何时应停止提示、不让学生困在一个无法理解的错误周围打转的系统。

成本表中看不见的经济

设想一家机构每月有 100 项审查任务，自动化前每项需要新员工工作两小时，使用助手后只需半小时复核。这是一个虚构的算术示例，不是来自真实公司的测量。会计师会看到“节省”了 150 小时，而这些时间确实可能值得节省。

但如果过去的一部分时间还生产了别的东西呢？例如识别异常案例、从相互冲突的来源建立论证，或发现文件缺陷的能力。没有被登记的训练时间，未来可能变成更依赖少数资深员工、审查错误，或不得不从一个不再培养专家的市场购买专家的成本。

可以用一个说明性关系表达这个想法；本文不声称它是经过估计的实证方程：

$$H_{t+1} = (1 - \delta)H_t + \alpha\, P_t\, F_t.$$

其中，H_t 是某一时期个人或团队的技能存量；δ 是技能不用时的流失率；P_t 是相关练习量；F_t 是反馈质量与理解错误的机会；α 是把练习转化为学习的能力。并非每一小时工作都等于一小时训练，也不是每次修正都有用。助手可能提高 F_t，让练习更深入；也可能在替学习者完成全部工作时降低 P_t。它甚至可能改变原本需要的技能。

这个表述可以区分三种情形：

- 自动化重复抄录信息，而这类工作原本不产生真正学习：表现可以提高，H 不会受到显著损失。
- 自动化一项原本要求学徒寻找原因、检验假设的任务：如果练习减少且没有替代方式，就存在风险。
- 助手提供解释、生成测试案例并揭示错误，却不垄断解决过程：即使总用时下降，P_t 与 F_t 也可能同时上升。

这并不是要求企业付钱让员工重新发明轮子。在有些职业中，学徒正是在模拟器上、监督下重新发明轮子的一小部分，才学会这只轮子有一天为什么会掉下来。

可以为这种差距提出一个分析性名称：专业能力再生产赤字（Expertise Reproduction Deficit）。它不是已发表理论，也不是正式标准，而是本文提出的诊断概念。它指对独立专家判断的预期需求，增长得比机构培养能够在没有辅助时作出这些判断的人的能力更快。

在运营中，可以比较现有专家离职或转岗的速度，与特定时期内达到独立能力标准的学徒人数。可检验预测是：如果诊断性练习被取消，又没有有效教学替代，新一批员工在独立判断测试中的差距会持续扩大，即使他们借助人工智能时的生产率指标有所改善。

但如果工作性质已经改变，旧判断不再需要，或工具创造了更好的新学习路径，这个概念就会失效。因此，应当测量对能力本身的需求，而不是神圣化一项已经失去用途的技能。

董事会问错了问题

职业自动化项目通常用一张表展示许可费用、员工工时、任务完成率和发现的错误。这些数字都合理，但表格还需要一列很难投在会议大屏上：员工过去在完成这项任务时学会了什么？自动化以后，谁会学会它？

合同初审可能训练人识别隐藏风险的条款；处理普通银行案例可能训练人连接客户身份与交易原因；检索工业记录可能训练人区分潜在原因与偶然同时发生。这些只是可能机制的例子，不是断言每家机构都依赖它们。作出判断前，必须先分析工作本身。

分析中最大的错误，是从“新手”到“专家”画一条直线。有些专家的核心技能并非来自重复新手任务，而是来自直接教学、困难经历、岗位轮换或模拟练习。人工智能也可能把新员工从多年低价值工作中释放出来，打开更高层次学习的大门。这不是边缘反驳，而是必须认真面对的竞争性假设。

因此，机构重新分配工作时，不应以原样保存“初级岗位”为目标，而应保留能力获得链条，即使工具与任务全部更换。常规报告写作可以消失，但解释证据、反驳假设、面对非典型案例的能力，仍需在某处被学会。

资深经理的角色也会变化：从工作结束后审查数百页的人，变成给新手设计一系列逐步承担认知责任的情境的人。专家用来解释和纠错的时间，眼前看可能比机器几秒钟生成答案更贵；但它也可能是在投资一项不会出现在资产表上的资产。

沙特不必在自动化与训练之间二选一

如果一家沙特软件、银行或工业企业大规模部署智能助手，它不必一开始就假定本地存在技能危机；这需要本地数据。但机构应当检验，工作分配方式是否仍给新加入者建立独立经验的机会。

答案不是给所有人设置“无人工智能日”。在软件开发的某些练习中，这也许可行；在医疗或运营程序中，则可能危险或违反专业政策。更合适的是一种练习预算：安排安全案例，设定明确学习目标，配备专业监督，并定期开展独立能力测试。它应被当作人力资本投资，而不是对使用工具者的惩罚。

技术公司可以让学徒在看到助手建议前，先解释故障原因，再用实际实验比较自己的假设。金融机构可以使用去除个人数据、包含冲突信号的训练案例，让学员分别记录支持和反对某项判断的理由，再接受专家复核。维护工程则可以在模拟中呈现受控故障，避免把真实工业设备变成昂贵课堂。

只记录练习小时数也不够。点击正确答案一百小时，不等于用一小时解释一个看似可信的假设为什么失败。更重要的是，在未见问题、没有辅助、并有与职业风险相称的监督条件下，表现如何变化。

一项能够证明本文错误的机构实验

要证明某家机构确实存在这种风险，采访一名“感觉”员工没有学到东西的经理并不够。在符合专业和伦理要求后，可以开展为期三个月或六个月的对照实验，比较三种安全工作方式：

第一组按机构政策使用助手直接生成方案。第二组使用先提问、给提示、延后展示答案的助手，并接受专家复核。第三组在诊断阶段不直接使用助手，而在模拟案例中学习，随后在比较与解释阶段使用它。

测量不能只有“每组完成多少”。还要测量完成时间、输出质量、监督负担和错误成本，并定期让参与者在无助手条件下解决新问题。还应测试延迟后的知识保持、发现合成输出中故意植入错误的能力，以及从训练过的问题迁移到结构不同问题的能力。参与者既有经验也应记录，避免把自然差异归因于工具。

如果第一组在独立测试中仍保持高质量，我们的假设就会变弱；如果第二组以可接受成本取得更好的独立性，证据就会支持改进辅助设计，而不是禁止工具；如果第三组在机构未来不需要的技能上获胜，也不能把这种优势当作反对进步的理由。

一项预先保证我们偏好观点获胜的实验，没有科学价值。好的机构设计允许公平实验让董事会感到难堪，而不只是让演示文稿说服董事会。

谁负责培养那个能够说“不”的人？

有些职业要求一个人经过多年学习，才能看到非专业人士不会注意的信号。律师发现一段看似完整的措辞藏着例外；工程师察觉某个传感器读数与系统其余状态不符；医生把模糊症状与第一张影像看不出的病史联系起来。这些不是魔法能力，也不全是自然练习的产物。它们通常来自知识、训练、反馈和逐步承担责任的积累。

人工智能系统也可以帮助加速这种积累。它能够生成案例、提出相反解释，并让学徒进行超出人类导师时间容量的尝试。但过度交付答案，可能只让员工学会批准结果，而不会诊断藏在结果中的错误。

战略问题不是企业会不会取消职业阶梯最底层的一部分任务。它们很可能消失或根本改变。问题是：企业会建造一架新的梯子，还是只购买一项把人送到楼上的服务，最后却发现再也没有人知道楼梯如何建成？

回到 2013 年美国联邦航空管理局那一页文件。它要求一种成熟平衡：利用自动驾驶，同时在安全条件允许时保留人的练习机会。不是少让机器飞，也不是拿乘客的安全训练人。

从这个意义上说，一些人工智能项目最大的损失，可能不是机器犯下的错误，而是学徒从未有机会犯下并理解的一次错误。多年以后，当我们需要一个能对运行良好的系统说“这一次，不行”的人，才会知道那些被省掉的错误值多少钱。

参考文献

1. SAFO 13002: Manual Flight Operations。美国联邦航空管理局，2013 年 1 月 4 日；官方一手安全提示。确认开篇事件日期，并在承认自动飞行收益的同时，建议保留手动练习机会。
2. SAFO 17007: Manual Flight Operations Proficiency。美国联邦航空管理局，2017 年 5 月 4 日；官方指引。支持在安全约束下，于训练和运行政策中建立并保持手动操作能力。
3. The Production of Human Capital and the Life Cycle of Earnings。Yoram Ben-Porath，《Journal of Political Economy》，1967 年 8 月 1 日；原创同行评审经济研究。为把技能获得理解为具有未来回报的投资奠定基础；本文说明性方程并非取自该论文。

4. How AI Impacts Skill Formation。Judy Hanwen Shen、Alex Tamkin，初稿提交于 2026 年 1 月 28 日，2 月 1 日修订；尚未经同行评审的初步实证研究。52 名参与者、每组 26 人，测量使用 Trio 软件库练习后的学习；不能证明劳动者的长期影响。
5. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention。Henry Roediger、Jeffrey Karpicke，《Psychological Science》，2006 年 3 月；同行评审实验。在特定条件下测量提取对延迟保持的作用；没有测试人工智能助手。
6. Productive Failure。Manu Kapur，《Cognition and Instruction》，2008 年 7 月；同行评审教育研究。解释最初未成功的尝试经过适当后续处理时，为什么可能成为学习设计的一部分；不授权在工作中制造危险错误。
7. Ironies of Automation。Lisanne Bainbridge，《Automatica》19 卷 6 期，1983 年；原创学术论文。是“抽走日常练习后仍把异常情况留给人”的自动化悖论理论源头。
8. The Retention of Manual Flying Skills in the Automated Cockpit。Stephen Casner 等，《Human Factors》，2014 年 12 月；同行评审模拟研究。反驳所有飞行技能都会普遍退化的说法；样本为 16 名飞行员，情境限于 747-400 模拟。
9. Endoscopist Deskilling Risk after Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study。Krzysztof Budzyń 等，《The Lancet Gastroenterology & Hepatology》，2025 年 8 月 12 日在线发表，10 月见刊；多中心观察性研究。1,443 次无辅助检查，28.4% 对 22.4%；不能隔离所有混杂因素，也不能证明因果。
10. Artificial Intelligence Augmented Tutoring vs Expert Instruction on Learning Simulated General Surgical Skills: A Systematic Review and Meta-analysis。Fouad Hanna 等，《BMC Medical Education》，2026 年 6 月 10 日；同行评审系统综述与荟萃分析。4 项研究、268 名参与者；以低确定性证据说明增强教学的可能性，并不支持取消人类监督者。
11. Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics。Hamsa Bastani、Osbert Bastani、Alp Sungu 等，《PNAS》，2025 年 6 月 25 日在线发表，7 月 1 日见刊；同行评审随机现场实验。比较土耳其近千名学生的练习表现与独立测试，并区分一般辅助和受提示约束的辅助。
12. Does Using Artificial Intelligence Assistance Accelerate Skill Decay and Hinder Skill Development without Performers' Awareness?。Brooke Macnamara 等，《Cognitive Research: Principles and Implications》，2024 年 7 月 12 日；提出假设的研究性分析。适合界定研究问题，不是证明技能流失的实验。
13. Operational Use of Flight Path Management Systems: Final Report。美国安全机构驾驶舱自动化工作组，2013 年 9 月 8 日；官方技术报告。记录训练与系统运行如何进入安全评估，并非有关语言模型的实验。
14. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot。Sida Peng、Eirini Kalliamvakou、Peter Cihon、Mert Demirer，初稿于 2023 年 2 月 13 日发布；初步论文中的对照实验。表明特定情境下编程任务可提速 55.8%，但没有直接测量长期技能形成。