一直没出事

AI 时代，能力、判断与责任
如何先失去可观测性

A Cut in the Record

九条读数都是满格。
断口右边露出来的，才是它们现在的真值。

一直没出事

AI 时代，能力、判断与责任如何先失去可观测性

孔凡鹤　著

德麦国际出版社

出版信息

书名　一直没出事

副题　AI 时代，能力、判断与责任如何先失去可观测性

英文题名　A Cut in the Record: How Capability, Judgment and Responsibility Lose Observability Before They Are Lost

著者　孔凡鹤

出版　德麦国际出版社（Demai International Press）

丛书　德麦国际学员专著　第 78 号

ISBN　979-8-90690-011-1

定价　US$22.00

版次　二〇二六年八月第一版

开本　170 mm × 240 mm

字数　约 19.7 万汉字

写作与证据声明

本书由九项独立研究与两章新写合论构成。九项研究曾以论文形式单独完成，收入本书时作三类改动：统一自称与体例、补入各章之间的判决性分界、删除与本书无关的方法自述；核心判断、数据与结论未作改变。

本书四章含公开数据的压力测试，所用数据均来自已公开发表的研究及其作者公开的原始数据与代码，来源见各章参考文献。四处压力测试均得到对作者不利的结果，并均导致原主张被收窄；三处收窄的具体内容见前言。

本书为理论与方法研究，不直接招募研究参与者，不使用可识别的真实个案资料。书中提出的各项判据与读数尚未经过心理计量学意义上的效度验证，不应直接用作个人能力认证、职业资格判定或人事决定的依据。书中第十一章给出一条不可让步的底线：断口读数不进入个体的绩效、晋升与留用决定。

本书的写作、检索与结构整理不同程度地使用了生成式人工智能系统。按本书自身的判据，这一事实意味着本书的论证质量不能自动证明作者拥有相应的理解与判断；本书的处置办法见导论第八节与第十章末节。

作者介绍

孔凡鹤，照护与人机协作研究者。

早期研究方向为痴呆家庭照护，关注长期照护关系中"完成"由谁宣布、未被交出的那一份落在谁身上的问题，相关成果结集为《照护的智慧》（德麦国际出版社，专著第 72 号）。那本书处理的是人与人之间的交接；本书处理的是人与外部认知系统之间的交接——两者共享同一种关切：一件事被宣布为妥当之后，还有没有人能够看见它其实没有妥当。

近年研究转向生成式人工智能持续代行条件下的能力、判断与责任归属，尝试把航空人因工程、核设施运行维修、审计、可靠性工程等高后果行业中成熟的验证逻辑，迁移到教育、知识工作与组织治理中来。本书是这一转向的第一部结集。

至本书成书时，作者在德麦国际学员专栏发表长论文九十余篇，另有专著一部。作者的工作方法是把一个问题同时交给三门互不通气的学科，让它们围绕同一个断裂点相互否定，再看剩下什么是三家都无法单独占有的——本书十一章中的九章由此产生。

作者认为，这本书如果有用，用处不在于它提出了新概念，而在于它把一句几乎所有行业都在说的话——"一直没出事"——变成了一个可以被检验、也可以被推翻的命题。

前言

一句让人安心的话

在照护研究里待过几年之后，我对"一直没出事"这句话有一种职业性的警惕。家属说这句话，通常是想说明现在的安排还能维持；机构说这句话，通常是想说明流程没有问题。可是在很多次真正出事的复盘里，事前那段"没出事"的时间并不是安全的证明，而是一段没有人去看的时间。没有人去看，不是因为疏忽，恰恰是因为一切看起来都很正常。

后来我把注意力转到生成式人工智能，本以为是换了一个完全不同的题目，结果撞上了同一个结构，而且更清楚、更快、更普遍。

三个把我推到这里的场景

第一个场景在一间教研室。一位老师说，这学期学生交上来的作业质量整体上了一个台阶，可她越看越不安，因为她说不出学生到底会了什么。她试着把作业改难，结果交上来的东西更漂亮了。她想问的其实不是"学生有没有用人工智能"，而是"我还能不能从我手上这份东西里，看出他会不会"。

第二个场景在一次事故复盘的旁听席上。一个团队用了两年的自动化流程出了问题，追责会开得很顺利：流程写得清楚，签字的人找得到，处分也落实了。散会之后我问一位工程师，下次还会不会发生同样的事。他说很可能会，因为被处分的那个人根本没有权限改任何一处设置。这句话让我意识到，责任被找到了，改错却没有发生。

第三个场景来自航空与核电的文献。这两个行业几十年前就遇到过同一件事：自动化越可靠，人越少接触真实的故障，于是"人在需要时能接管"这个假设越来越难被验证——不是因为人变差了，而是因为不再有场合去看。核电给出的解法是监督试验：既然备用系统平时不运行，那就周期性地强制运行一次，专门为了产生证据。

三个场景表面上毫无关系。把它们并排放了很久之后，我才看清楚它们是同一件事的三个位置：用来判断某样东西还在不在的那个信号，已经不再指示它了；而信号失效的方式，是它变得更好看了。

九个问题，同一句话

这本书由九项独立的研究和两章新写的合论组成。九项研究原本各自成篇，动用的学科几乎没有重叠——认知神经科学与核设施运行维修、解释学与审计学、生态心理学与统计学习理论之间，本来没有对话的必要。它们各自要回答的问题也不同：能力归

谁、能力怎么维持、帮助要不要撤、绩效为什么骗人、判断属于谁、理解算不算数、责任落在哪里、验证做到什么程度算够、群体为什么会一起错。

写到第五、第六篇的时候，我发现这些结论正在说同一件事。九个领域各自的良好信号——结果质量、零事故、上升的绩效曲线、末端签字、能迁移、有明确责任人、检查覆盖率、多数同意——在人工智能持续代行之后，全都失去了它们原本具有的分辨力。而且失效的方式是同一种：它们不是变差了，是变得更好看了。

于是这九篇不再是九个题目，而是同一句判断在九个位置上的换位。把它们放进一本书，理由不是它们都谈人工智能——共同的题材只构成一个专辑，不构成一本书——而是它们共绕着同一条判断。这条判断写在导论第三节，也写在封面上：一直没出事，能证明什么。

为什么必须用互不相干的学科

有人问过我，既然九章的结论是同一条，为什么不用一门学科把它一次讲完，非要动用二十多个领域。

理由有两个。第一个是发现的次序：这条判断不是先想出来再去找例证的，它是在九个互不通气的领域里各自被逼出来之后，才被认出是同一条。如果我一开始就用一门学科去推，很可能只会得到那门学科内部早已有的说法——认知科学会给我"认知卸载"，人因工程会给我"脱离回路"，测量学会给我"效标污染"。这些说法都对，但它们各自只覆盖一块，而且各自的语言互相不通。

第二个理由更实际：一条判断如果在互不相干的领域里反复成立，它多半不是那些领域的性质，而是问题本身的形状。 核设施的监督试验、运动训练的最低维持剂量、航空的接管冷启动，这三样东西没有共同的理论来源，却在"备用能力不能由正常运行证明"这一点上完全同构。这种同构不是类比修辞，它意味着可以把一个领域里成熟的工程手段直接搬到另一个领域——本书第二章做的正是这件事。

代价是这本书读起来会不断换场景。我尽量在每一章开头说清楚这一章要动用哪三门学科、它们各自最擅长发现哪一类失败、以及它们共同默认了什么。

我最想让读者带走的一件事

如果这本书只能留下一句话，我希望是这句：不要用"最近怎么样"来回答"他还会不会"，要用"我们上一次真正看见，是什么时候"。

这句话听起来像常识，但它取消了当代绝大多数评价制度的合法性依据——考试、绩效考核、签字审批、合规审计、多数表决，全都在用"最近怎么样"回答"他还会不

会"。九章的工作，就是在九个领域里把这句常识变成可以执行、可以记录、可以被推翻的东西。

这本书不主张什么

写作过程中，我不断需要抵抗一种廉价的答案：既然人工智能让人看不清了，那就少用一点。

九章都拒绝了这个答案，而且是带着理由拒绝的。要求人周期性地完整手工重做，等于把能力维护变成新的低价值劳动，占用生产时间、制造抵触，甚至为了给人练手而放弃更安全更准确的工具。真正需要周期性收回的不是全部任务，而是那些能够暴露问题的关键回合。本书的判据因此被刻意设计成与"用了多少人工智能"正交：一个人可以借助人工智能形成更好的理解，也可以借助它把边界判断一并外包，两者要靠断口区分，不靠使用量区分。

也因此，本书不是一份禁令清单。它更像一份仪表校准手册——问题不在飞机飞得太快，而在仪表已经不指示它该指示的东西。

三处被数据改掉的主张

本书有四章带公开数据的压力测试，四次的结果都对作者不利，也都因此改写了原来的主张。我把它们写在前言里，不是为了显得谦虚，而是因为这本书要求别人接受一种昂贵的检查制度，那它至少得先证明自己被检查过。

第一处在第三章。我原本准备写的是"长期的人工智能帮助会首先让能力下降"。公开数据不支持这句话：带教学护栏的那一组，平均独立成绩与对照组几乎没有差别。于是主张被收窄成一句更弱但更要紧的话——人工智能可以在不损害平均成绩的情况下，降低当前表现作为能力信号的保真度。 这个收窄反而使全书的核心概念从"退化"变成了"可观测性"。

第二处在第一章。我原本设想的测试是"撤掉人工智能之后再考一次"。公开实验提醒我，转场本身会影响动机、控制感与无聊程度，所以撤援成绩里混着一大堆与能力无关的东西。判据因此从"闭卷重做"改成了"定向扰动"：只破坏代理的一条关键可靠性，看主体能不能重新闭合回路。

第三处在第六章。我原本写下的是"人工智能代劳会削弱理解"。有几项研究直接反对这句话，其中设计良好的人工智能导师甚至在更短时间里产生了高于课堂主动学习的学习增益。于是这一章的判据被要求与"是否使用人工智能"正交——真正要测的不是用了多少，而是交互如何组织。

三处收窄有一个共同点：每一次退让都让判断变得更窄，也更硬。 这是我在这本书里学到的最有用的一件事。

关于代价的一点交代

书里几乎每一条建议都附了代价。不可预告的测试要付出组织摩擦与信任成本；扰动要从真实近失误中生成，就必须维持一份与考核彻底隔离的记录；要求异源工具会增加成本，而某些领域根本不存在真正的第二条路径；最重要的一条——断口读数不得用于个体奖惩——则意味着组织要为一项不能用于人事决定的测量付钱。

写下这些代价不是为了显得谨慎。一条不写代价的建议，读者无法判断自己能不能执行，最后多半会执行成一个形式。第十一章整章在讲这件事：这套方法一旦制度化，最省事的达标方式很可能恰好摧毁它要保护的东西。我把这一章放进正文而不是放进附录，因为我认为它是本书最可能应验的部分。

给四类读者的一句话

给教师：作业质量上升不再是好消息，也不是坏消息，它只是不再是消息。要重新得到消息，必须制造一个学生事先不知道的边界。

给管理者：在把这套方法接上考核之前，请先读第十一章那条不可让步的底线。做不到那条，本书的建议不该采纳——它会以看起来成功的方式失败，那比明显失败更糟。

给工程与安全人员：你们行业几十年前解决过这个问题的一半。第二章借的就是你们的监督试验逻辑，欠的那一半是：认知性的接管门比手工操作衰减得更快，而且不会报警。

给研究者：这本书最值钱的部分是它写明了自己怎样会错。第十章有三对最可能被合并的章节和三条撤回条款，推翻它们不需要新的理论，只需要一次设计得当的实验。

感谢与说明

这本书的九章曾以论文形式独立写成，收入本书时做了三类改动：统一为章、补入各章之间的判决性分界、删除与本书无关的方法自述。九章的核心判断、数据与结论未作改变。

本书的写作、检索与结构整理不同程度地使用了生成式系统。按本书自己的判据，这意味着本书的论证质量并不能自动证明作者拥有相应的理解与判断。我能做的处置写

在导论第八节：把每章的承重命题、竞争解释与撤回条件写死在正文里，让读者可以绕过作者直接检验。这是一本关于"怎样才算还看得见"的书；它没有理由把自己藏起来。

孔凡鹤

二〇二六年八月

一个我没能回答的问题

写完之后，有一个问题一直留在桌上，我把它写在这里，作为这本书的一个开口。

本书要求组织周期性地制造中断，为的是产生关于人的证据。可是制造中断需要成本，而愿不愿意付这个成本，取决于决策者相信"人可能已经不行了"。问题在于，让人相信这件事的唯一材料，正是那些还没有被生产出来的证据。一个组织越是长期看不见，就越不觉得需要去看。

我没有办法从内部打破这个闭合。书里给出的所有对策都假设已经有人决定去看第一眼。第一眼从哪里来，本书答不上来。可以想到的几种外力——事故、监管、外部审计、行业标准、一次公开的失败——都不在本书的射程之内，而且前三种通常来得太晚。

之所以把这个问题留在前言而不是留在结尾，是因为它决定了这本书能不能被用起来。读者若从这本书里只取走一件可以马上做的事，我建议是那件最便宜的：在把人工智能的结论用于一次不可逆的行动之前，先写下一句"如果这是错的，最先出问题的会是什么"，然后去看那一处。 它不需要任何制度支持，也不需要任何人同意。全书十八万字压缩到零成本，剩下的就是这一句。

导读

这本书能回答与不能回答的

能回答：一个信号（成绩、绩效、签字、迁移表现、检查覆盖率、多数同意）在人工智能持续代行之后还能不能指示它原本指示的东西；如果不能，应当在哪里开断口、断口上读什么、读数是什么形状、什么结果会让这套判断失败。

不能回答：断口应当多密（无经验依据）；一个团队里合格事件是否集中在少数人身上（九个读数全无分布量）；人的能力应当如何提升——本书只管把缺口照出来，训练是另一回事。第三章说得直白：清偿证据债不等于提升能力，正如诊断出问题不等于已经治疗。

四条读法

通读（十八万字）：按编次顺序读。第一编建立判据与工程，第二、三编把同一条判断换到判断、理解、责任、验证四个对象上，第四编放大到群体并回头处理本书自身。

管理者与政策制定者（约三万字）：导论 → 第七章（责任与权限）→ 第三章（证据龄比）→ 第十一章（怎样会被应付掉，以及那条不可让步的底线）→ 第十章第二节的换位表。这条路径专门回答"要不要在我这里推行、推行会怎样失败"。

教育工作者（约三万字）：导论 → 第四章（为什么作业质量越高越不能当作能力提高）→ 第六章（越界熟练与边界破坏任务）→ 第三章（校准回路的五类任务）→ 第十一章。这条路径回答"考试改成什么样才算数"。

研究者与审稿人（约两万字）：导论第三至六节 → 第十章全章（换位表、三对最危险的近亲及其合并条件、三型读数、三条撤回条款）→ 各章末尾的划界节。这条路径是为了最快地判断本书哪里可能是错的。

只有二十分钟：读导论第一、三、四节，再读第十章第二节那张表。全书的判断与索引都在那里。

每章的最小读法

每章都按同一顺序组织：摘要 → 三门学科各自的材料与冲突 → 三家共同默认而又被自己材料推翻的前提 → 提出的判据或机制 → 与近邻的划界（含与本书其余各章的分界）→ 可证伪条件。时间有限时，读摘要、判据与最后的可证伪条件三节即可；三门学科的材料部分是论证的来源，不是结论。

四章带有公开数据的压力测试：第一章（四项在线实验，总样本 3,562）、第二章（去训练与维持剂量数据）、第三章（825 名学生的二次分析）、第六章（四项公开实验的反向约束）。这四处都得到了对作者不利的结果，并且都因此收窄了主张。若只想知道这本书有没有被数据管住，读这四节。

九个读数速查

读数	出处	形状	一句话
生差—裁差—改向—回写四环节	第一章	串联门	扰动之后闭环由谁重建
五门串联（发现·重建·裁决·干预·恢复）	第二章	串联门	一门不过，整条接管路径不通
证据龄比	第三章	相关折扣	距上一次有效独立证据过去多久
绩效对能力的预测力	第四章	相关折扣	产出还能不能指示人
理由持有—失效识别—重新发起	第五章	串联门	这一次决定是否由人形成
首次边界修订由谁发起	第六章	串联门	主动撤回，还是被提示后撤回
ECR 三要素	第七章	串联门	知道、能改、有权，是否同在一处
残余风险与 η=ΔR/c	第八章	成本效率	未核错误还能不能翻转结论
N_eff = n/[1+(n−1)ρ]	第九章	相关折扣	几个人同意，等于几份独立证据

三种形状对应三张量表，一套记录格式。 不要为九章各造一套指标——这是本书给组织的最短建议。

术语约定

断口：为获得读数而有意制造的一次中断，包括代理扰动、接管证明、盲段、撤援、边界破坏、判别性检验与先独立后汇聚。全书的核心动作。

可观测性：不是"能力有多强"，而是"我们凭什么知道它现在有多强"。本书全部主张都落在这一侧，而不落在能力本身。

代理：承担认知任务的外部系统，主要指生成式人工智能，也包括其他自动化工具。

越界熟练（第六章）：表现出流畅解释与远迁移，却不能在结构失效处停止的状态。

改写责任（第七章）：在错误发生后有能力改变下一轮的责任，区别于事后被追究的责任。

怎样最快地推翻这本书

本书希望被这样对待，因此把三条路径写在最前面。

第一条，最容易： 在若干个人工智能深度介入的真实场景中，同时收集日常绩效与断口读数。若日常绩效对断口读数具有稳定且足够高的预测力（例如跨场景相关持续高于 0.8），那么断口是多余的，本书的全部工程建议都应当撤回。

第二条，最重要： 检验断口读数能否预测真实失效。若事前读数良好的人与读数糟糕的人，在真实的人工智能失效事件中，异常发现速度、恢复时间与损失规模没有系统差别，本书就只是发明了一种昂贵而无效的考试。

第三条，最可能： 观察本书方法制度化之后的实际形态。若普遍出现的是为通过测试而准备的应试能力，而真实接管能力并未提高，那么本书即使理论上正确，实践上也应当被撤回。第十一章给出了五个可以从原始记录读出的预警信号，其中"扰动复用次数"与"证据刷新的时间分布"两项几乎零成本。

三条之外还有一条更快的：第十章列出了本书内部三对最可能被合并的章节及其合并条件。若其中任何一对被证明是同一件事，本书的章数应当减少——这不需要外部数据，只需要一次设计得当的实验。

阅读时的三个提醒

一，不要把九章读成九种指标。 九个读数只有三种形状，对应三张量表。组织若为每一章各造一套指标，得到的将是十一份互不通气的表格，而不是一套记录格式。

二，"断口"不等于"关掉人工智能"。 全书没有一章主张回到手工。断口是一次有意的、局部的、可控的中断——扰动代理的一条可靠性、安排一段没有辅助的盲段、给出一个前提已被破坏的任务——目的是产生证据，不是恢复旧的工作方式。判断一个做法算不算断口，看它有没有产生一个常态运行下不存在的读数。

三，读到与自己领域无关的章节时不要跳过。 本书的论证强度恰恰来自互不相干的领域得到同一结论。若只读与自己相关的那两章，会得到一个孤立的判断；把互不相关的几章并排读，才会看出这不是某个领域的性质，而是问题本身的形状。

目　录

导论　一直没出事，能证明什么

一　一句在所有高风险行业里都被当作证据的话

"一直没出事"是人类最常用、也最省钱的安全论证。航空公司用它证明机队适航，核电站用它证明备用系统可用，医院用它证明流程稳妥，工厂用它证明操作规程有效，学校用它证明教学没有问题，企业用它证明治理结构健全。它之所以被普遍接受，是因为在很长一段时间里它确实有效：如果一个系统的关键环节已经失灵，事故通常会在一段时间内暴露出来；没有暴露，多半说明还没有失灵。

这个推理有一个从未被明说的前提：产生结果的过程与承担能力的主体，是同一个。一份合格的报告出自一个会写报告的人，一次平稳的着陆出自一个会着陆的飞行员，一个正确的诊断出自一位会诊断的医生。正因为如此，看结果就能推知人；正因为如此，没出事就意味着人还行。

生成式人工智能抽掉了这个前提。当检索、分析、计算、写作、编程、比较、推理乃至判断建议都可以由外部系统持续承担时，一份合格的报告不再必然出自一个会写报告的人。结果的质量提高了，结果与人之间的那条推理链却断了。"一直没出事"于是变成一句同时与两种完全相反的状态相容的话：人仍然完全胜任，以及人早已不能接管——它不再能区分这两者。

本书要处理的就是这句话失效之后留下的空白。

二　真正的难题不是"AI 让人变强还是变弱"

关于人工智能与人的能力，公共讨论长期停在一个二元问题上：AI 究竟在增强人，还是在削弱人。两方各有相当扎实的证据。增强的一侧可以举出写作质量提高、编程效率提升、客户服务改善、学习效率增益；削弱的一侧可以举出认知卸载、独立表现下降、技能形成受损、批判性参与减少。争论持续多年而没有收敛，本身就是一个信号：这个问题可能问错了。

问错在哪里？在于它默认我们仍然能够测量"人的能力"这个量。可是在 AI 深度介入之后，我们观察到的每一个绩效数字，都是人的能力、AI 的能力、人机匹配程度与任务条件共同作用的结果。用这样一个混合量去争论其中某一个分量的升降，双方各自都能找到支持自己的数据，也各自都无法说服对方。争论不收敛，不是因为证据不足，而是因为用来做判决的那把尺子本身已经被污染。

因此本书把问题换掉。不问"人变强还是变弱"，而问：

在 AI 持续代行的条件下，我们还能不能看见人的能力、判断与责任是否在位？如果不能，靠什么重新看见？

这是一次问题的置换，不是一个答案。它的好处是可判定：能力的升降或许难以直接测量，但"我们最近一次真正看见它是什么时候"是可以回答的；"这个信号还能不能指示那个东西"也是可以检验的。九章的全部工作，就是在九个不同的领域里把这个可判定的问题做到底。

三　由此得到的核心判断

九章从九个互不重叠的学科组合出发，得到同一条判断：

能力、判断与责任并不是先失去，而是先失去可观测性；正常运行的成功本身销毁了证明它们仍在的证据，因此它们只能在代理被有意打断的断面上被重新判定。

这条判断包含三层意思，每一层都与直觉相反。

第一层：失去的次序是反的。 通常认为退化先发生、然后被发现。本书主张顺序相反——先出现的是"看不见"，然后才可能是"不存在"。第三章给出了一条经验读数：在公开的高中数学随机实验数据中，无 AI 对照组的练习表现与随后独立考试表现相关高达 0.756，而使用生成式 AI 的两组分别降到 0.435 与 0.486；关键在于其中带教学护栏的一组，平均独立成绩与对照组几乎没有差别——平均意义上没有任何学习损失，而"眼前做得好"作为"本人会不会"的信号，保真度已经明显下降。 可见性的下降可以先于、并且独立于能力的下降。

第二层：销毁证据的是成功，不是失败。 这不是隐瞒，不是舞弊，也不是指标造假。没有人删除任何信息，只是产生这类信息的场合不再出现了：AI 越可靠，真实故障越少；真实故障越少，人获得自然接管练习与自然暴露的机会越少；机会越少，关于人的证据越陈旧。第四章进一步指出，这个过程会自我强化——能力越不可观察，未来的能力损失越难进入当下决策，个体与组织越倾向继续外包。

第三层：唯一的观测位置是断口。 既然常态运行不再提供分辨力，那么关于能力、判断与责任的读数就只能来自一次有意制造的中断：一个被扰动的代理、一次被安排的接管、一段没有辅助的盲段、一个前提已被破坏的边界任务、一次独立于原生成路径的验证、一次先独立后汇聚的表决。九章分别把这个中断做成了九种可执行的工程形式。

四　这个难题为什么重大

如果它只是一个学术上的测量问题，不值得写一本书。它之所以重大，是因为当代社会有三套关键制度，全都建立在那个已经失效的前提上，而且都还没有更换。

第一套是教育评价。 作业、论文、考试、作品集之所以能证明学生学到了什么，靠的正是"产出出自本人"。当高质量产出可以随时被生成，这套评价体系测到的东西发生了性质变化，而学位、升学与职业资格仍然架在它上面。一个尤其危险的推论出现在第四章：只用最终产出评价能力的制度，不仅可能测错，还会主动加速能力的不可见——因为它奖励的正是外包。

第二套是人类监督的法律安排。 世界各地的人工智能治理规则普遍要求"人在回路"、保留人工审核、由自然人作最终决定。这些要求的正当性依赖于一个假设：形式上的最终确认位置，等同于实质上的判断与责任在位。第五章与第七章分别否定了这个假设的两半——最后动作在人不等于判断由人形成；有人签字不等于有人能改写下一轮。这意味着现行监管可能正在保护一个无法被验证的东西：条文可以要求"必须有人审核"，却没有任何条款要求证明这个人还能审得动。

第三套是高后果领域的安全论证。 航空、核电、医疗、工业控制、金融风控普遍采用"自动化常驻、人工可接管"的安全结构。这个结构的可靠性完全依赖于接管能力在需要时可用，而在 AI 常态代行的新条件下，这一假设从未被系统验证过。第二章借用核设施的监督试验逻辑给出的判断是冷峻的：零事故既不能说明人已经退化，也不能说明人仍然可靠；它对备用通道缺乏分辨力。 一个只在真实故障发生时才第一次检验人工备用能力的系统，把训练、测试与正式任务压缩到了同一个时刻。

三套制度的共同处境可以概括为一句话：社会正在用一整套已经失去分辨力的仪表，驾驶一架越飞越快的飞机。 仪表读数持续良好，恰恰是最不该安心的理由。

五　还有一层紧迫性：可观测性的丢失是不可逆的

一般的能力问题可以事后补救：发现不足，就训练。可观测性的问题有一个特殊的结构，使它难以事后补救。

要恢复关于某项能力的证据，必须先制造断口；而是否值得为某项能力付出制造断口的成本，取决于对该能力当前状况的判断；可是那个判断，恰恰就是已经不可得的东西。 于是形成一个闭合：越是看不见，越不知道值不值得去看；越不去看，越看不见。第四章把它写成一条自我强化的循环，第七章在责任结构上给出了同型的循环，第九章在群体层面给出了第三个。

这就是本书为什么主张现在就要做，而且主张从"最小、最便宜、最不打扰"的形式做起：不是因为危机已经到来，而是因为一旦这个闭合完成，重新打开它所需的代价会远高于现在维持它。

六　本书提供了什么

本书提供的不是一份禁令清单，也不是一套"减少使用 AI"的建议。九章反复拒绝了同一种廉价答案：回到手工。第二章拒绝把"完整脱 AI 重做"当默认方案，第三章指出真正需要渐退的不是 AI 而是未经验证的连续代行，第六章明确要求判据必须与"用了多少 AI"正交，第一章允许深度外包。本书提供的是三样东西。

第一样是一个可判定的问题形式。 把"人还行不行"换成"这个信号还能不能指示那个东西"，把"要不要用 AI"换成"哪些节点上必须保留断口"。前者不可判定，后者可以逐条回答。

第二样是一套跨领域通用的测量结构。 九章的读数看似各不相同，实则只有三种形状：串联门型（若干二值判定，任何一门不过整条路径就不通，明令不许取平均）、相关折扣型（名义证据量按冗余度打折）、成本效率型（单位成本压缩多少结论相关风险）。三种形状对应三张量表、一套记录格式。这使教育、组织管理、法律与安全工程第一次可以用同一种语言讨论同一件事——它们面对的不是四个问题，是同一个问题的四个位置。

第三样是自己的失效条件。 第十章给出三条书级撤回条款：若日常绩效对断口读数具有稳定高预测力，全书工程建议应当撤回；若断口读数不能预测真实失效，本书只是一种昂贵的考试；若制度化之后最省事的达标方式恰好摧毁所保护之物，本书在实践上应当撤回。第十一章整章处理第三条，因为那是本书自认最可能发生的失败。一本不肯写明自己怎样会错的书，不值得被采纳。

七　全书结构

第一编处理能力，四章按"判定—维护—计量—机制"排列：第一章给出能力归属只能在断面上判定的判据；第二章把它做成可长期执行的维护路径；第三章把维护的对象从能力换成证据，给出证据龄比这一读数；第四章解释绩效与可见性为何会反向变化并自我强化。

第二编处理判断与理解。第五章问一次具体决定是否由人形成，第六章问一个结构是否被连同它的失效条件一起持有——后者命名了一种在 AI 时代尤其危险的状态：越界熟练，表现出流畅解释与远迁移，却不能在模型失效处停止。

第三编处理责任与验证。第七章指出真正的责任缺口不是无人签字，而是无人能改写下一轮，并给出改写责任的三要素；它被放在能力诸章之后有一个具体理由——其余各章设计断口测试时默认了被测者拥有改向权限，而现实中大量岗位没有。 不先把制度不授权与个人无能力分开，本书的能力判定会系统性地把制度缺陷记到个人头上。第八章处理验证：在不重做整个任务的前提下，如何做到足以发现会改变结论的错误。

第四编从个体走向群体与自身。第九章说明为什么"大家都更聪明"不等于"群体更不容易犯大错"，并给出有效独立判断数这一折扣。第十章是全书合论，给出九次换位表、三对最危险的近亲及其合并条件、九个读数的三种形状，以及三条撤回条款。第十一章处理反噬：一套用于揭露遮蔽的方法，为什么会长出新的遮蔽。

八　本书愿意先认下的四件事

一，本书不知道断口应当多密。 能说明为什么需要中断，也能说明中断处读什么，但"多久一次"目前只能由后果等级与衰减速度粗略推导，没有经验依据。

二，本书的九个读数没有一个是分布量。 全部以事件或个人为单位取值。一个团队完全可能九项全部合格，而合格全部来自其中三个人。凡使用本书方法者，须同时报告参与率与事件集中度。

三，本书容易被误用来惩罚个人。 因此第十一章给出一条不可让步的底线：断口读数不进入个体的绩效、晋升与留用决定。保证不了这条底线的组织不应采纳本书方法——它会以看起来成功的方式失败，那比明显失败更糟。

四，本书自身的写作由生成式系统深度参与。 按本书自己的判据，这意味着本书的论证质量并不能自动证明作者拥有相应的理解。本书的处置是把每章的承重命题、竞争解释与撤回条件写死在正文里，使读者可以绕过作者直接检验，并在第十章列出三处"因不利数据而主动收窄主张"的记录，供读者判断这本书有没有在被指出之前自己撤回过。

这四件事写在导论里，而不是藏在末章，因为本书的核心主张正是：一个系统若长期只呈现良好读数而从不暴露自己在哪里会失败，那份良好读数本身就应当被怀疑。 这条要求首先适用于本书。

第一编　能力

能力是这本书里最先失去可观测性的东西，也是四章共同的对象。

四章按"判定—维护—计量—机制"排列，这个次序本身就是一条论证。第一章先回答归属问题：在大量任务已经合法外包之后，凭什么还把某项能力算在人身上——答案是只能在代理被打断的断面上判定，不能由代理正常运行时的结果质量反推。第二章接着回答，既然判定条件如此，如何以最低成本让它长期保持可满足，而不把能力维护变成新的低价值劳动。第三章把维护的对象换掉：不问能力是否还在，只问我们上一次真正看见它是什么时候，由此得到全书最容易被组织直接采用的一个读数。第四章解释这一切为什么会持续恶化——绩效上升与可见性下降不是两件同时发生的事，而是互相强化的一件事。

四章因此不是四种说法，而是同一条线上的诊断、处方、计量与病理。读者若只读一章，读第一章可以知道判定标准，读第三章可以立刻开始记录。

第一章

正确答案不等于能力

断代理复闭环判据：能力归属为什么只能在断面上判定　——认知神经科学 × 控制科学 × 认识论的跨学科对撞

承重命题：人类能力的最低保留单位，不是“还能不能独立把整项任务做完”，也不是“是否始终亲手执行”，而是当外部认知代理发生错误、冲突、失联或情境失配时，主体能否重新生成一个可观察、可裁决、可干预、可更新的闭环。能力归属应在代理失效的断面上判定，而不能由代理正常运行时的结果质量反推。

摘　要

生成式 AI 正在把写作、翻译、编程、检索、分析与部分判断从“人必须亲手完成的任务”改写为“人可以调用外部认知代理完成的任务”。由此产生一个比“AI 会不会取代人”更基础的问题：当任务表现已经无法区分人的能力与工具的能力时，凭什么仍把某项能力归属于人？本章将认知神经科学的内部模型与认知卸载、控制科学的可观测性—可控性与人工接管、认识论的知识—理由与认知主体性置于同一冲突面。三者分别倾向于从内部表征、功能控制或认识论资格读取能力，却共同默认“能力可在正常运行状态下由某一维单独认证”。本章以各自内部材料推翻该前提：外部工具可提高即时成绩却不形成等量的内部学习；自动化可提高正常绩效却削弱故障接管；正确结论也可能因偶然、不可区分或缺乏认知归属而不构成主体的知识成就。本章据此提出“断代理复闭环判据”（Proxy-Severance Reclosure Criterion, PSRC）：只有在外部代理被施加可判定扰动后，主体仍能生成独立预期并发现差异、基于非同源理由裁决冲突、实施能改变任务状态的干预，并把结果回写为下一轮判断规则，才可把该能力继续归属于主体。该判据不要求人保留全部手工操作，因此允许深度外包；但它拒绝把“持续得到正确答案”直接兑换为“仍然拥有能力”。本章进一步以公开实验数据对“即时增强=能力保留”的最强竞争解释进行压力测试，并与认知卸载、扩展心智、脱离回路、批判性思维、适当依赖、技术辅助 know-how 及作者既有“实质主控/消化性裁决”概念逐项划界。结论是：AI 时代真正不可外包的不是任务清单，而是能力归属的最后闭环。

关键词：生成式人工智能；不可外包能力；断代理复闭环；认知卸载；人工接管；认知主体性

English Title and Abstract

Correct Answers Are Not Capabilities: A Proxy-Severance Reclosure Criterion for Non-Outsourceable Human Competence in the Generative-AI Era

Generative AI increasingly performs writing, translation, coding, retrieval, analysis, and parts of judgment. This makes task success an unreliable indicator of where competence resides. This paper asks a narrower question: under extensive cognitive delegation, what is the minimum condition for attributing a capability to the human rather than merely to the human–AI system? A second-order collision is constructed across cognitive neuroscience, control science, and epistemology. The three fields privilege, respectively, internal models, effective control, and epistemic credit. Yet each contains evidence that defeats a normal-operation criterion of competence. Cognitive offloading can improve current performance without equivalent internal learning; automation can improve nominal performance while degrading takeover capacity; and true or even justified belief need not amount to epistemic achievement when success is lucky or undiscriminating. We propose the Proxy-Severance Reclosure Criterion (PSRC): a human retains a capability if, when the external proxy is made unavailable, conflicting, wrong, or contextually invalid, the human can generate an independent expectation that makes mismatch detectable, discriminate between competing outputs with a non-circular reason, intervene so as to change the task state, and update the rule or model governing the next round. PSRC permits extensive outsourcing of routine operations while denying that assisted correctness alone establishes retained human competence. A public-data stress test using four GenAI experiments (N=3,562) rejects the simple equivalence between immediate AI-assisted augmentation and subsequent independent capability. The proposal is differentiated from cognitive offloading, out-of-the-loop performance, critical thinking, appropriate reliance, distributed cognition, technology-assisted know-how, and the author's prior constructs. The central claim is that the non-outsourceable unit is not a task, but the capacity to re-close an epistemic-control loop when delegation ceases to be dependable.

Keywords: generative AI; cognitive offloading; human capability; takeover; epistemic agency; reclosure

问题不是“AI 还能不能做”，而是“能力到底归谁”

在工业自动化时代，“人是否还有能力”通常可以用一个相对直观的办法来判断：关掉机器，让人自己做一次。生成式 AI 让这个办法失效了。首先，很多知识任务原本就不是封闭动作，而是依赖数据库、同事、文献、计算器、软件和组织流程；把所有外部支架同时撤掉，测到的不是现实能力，而是“孤立个体在人工贫困环境中的表现”。其次，AI 的优势恰恰在于把大量中间操作压缩成一次调用。如果教育与组织仍坚持“只有完整手工复现才算会”，就会把工具使用本身误判为能力丧失，重新制造一种低效的手工作坊主义。

但反方向的判断同样危险：只要一个人在 AI 协助下持续交出高质量结果，就认定他已经拥有相应能力。这样的认证会把“系统绩效”直接记到“个人能力”账上。一个从未读过原始材料的人可以用 AI 生成合格综述；一个不理解边界条件的人可以让 AI 写出能通过测试的程序；一个几乎不懂目标语言的人可以借助模型完成流畅翻译。只要代理稳定、输入熟悉、任务边界清楚，结果可以长期很好。真正的问题只在代理出错、两个模型互相冲突、上下文变化、证据缺失或组织目标突然改变时出现：此时谁能够看见异常、决定哪一条理由更重、改变行动方向，并在事后修正自己的判断?

因此本章把研究对象从“任务可否自动化”移到“能力如何归属”。所谓“不可外包能力”，不是一组 AI 永远做不了的神秘任务。AI 能力边界会继续移动，任何以“机器现在不会什么”为基础的人类能力清单都具有很短的保质期。本章要建立的是一个与模型能力上升相容的最低判据：即使某项任务的九成九操作都被 AI 接走，什么东西仍必须在人这一侧形成，才有理由说“这个人仍然会”？

这个问题同时关系教育、职业认证和责任制度。若学校只考无 AI 手工完成，它可能测到过度保守的能力；若只看 AI 辅助作品，它又可能把代理绩效误作学习成果。若公司把“能用 AI 做出结果”直接等同于专业胜任，就会在模型异常时发现没有人能够接管；但若公司要求所有员工保留完整手工流水线，又会放弃真正的自动化收益。一个可操作的最小判据，必须同时避开这两个极端。

三个维度：为什么必须同时从三处读能力

本题表面上包含“是否意味着”“如果……那么……”等多个问句，但要的答案形状不是一条行动路径，也不是一个因果驱动，而是一个能把两种表面相同的状态分开的“东西”：同样都能借助 AI 得到正确答案，哪一种叫“能力仍在”，哪一种只叫“代理仍在”。因此它要问的是"是什么"，对口的是三个维度。三门学科不能只作为

三个章节来源，而必须分别站到 S、D、E 不同位置上，然后共同推翻“某一维单独足以认证能力”的前提。

三者的烈度不在于谁“更重视人”，而在于它们给能力开出的最低凭证不同。认知神经科学允许大量信息外置，只要内部系统仍能调度和校准；控制科学并不要求操作者理解全部算法，只要求系统状态可被掌握、故障时有可执行的接管路径；认识论内部甚至存在外在主义立场：一个人未必需要随时口头陈述理由，可靠的形成过程也可能赋予信念认识论资格。正因如此，本章不能把结论预设成“人必须理解一切、亲手做一切、解释一切”。真正的二阶产物必须比这更薄，也更硬。

第一碰撞面：认知神经科学——能力不是记住全部，而是仍能生成“自己的预期”

内部模型：没有预期，就没有真正的误差

内部模型研究提供了一个极重要的结构：一个控制中的神经系统并非等到世界发生以后才被动接收结果，而是利用内部模型预测行动后果，再用真实反馈与预测之间的差异形成误差信号。Wolpert、Ghahramani 与 Jordan 关于感觉运动整合的经典工作，以及随后关于状态估计与预测误差的研究，都表明“误差”不是外界自带的标签；它只有相对于一个先在的预期才存在。没有预期，外界只能发生变化，却不会以“偏差”的形式显现。

把这一点迁移到 AI 知识工作，可得到一个比“记住多少知识”更小的条件。人不需要在脑中复制语言模型，不需要记住数据库全部条目，也不需要独立完成所有中间计算；但他至少必须保留某种任务相关的预期结构：这段论证大概应该满足哪些约束，这个程序的输入输出关系应当如何变化，这个翻译在语境中不应出现哪类歧义，这个结论如果成立应与哪些已知事实相容。只有当 AI 输出能够撞上一个属于人的预期时，异常才有可能在人的一侧发生。

这也解释了为什么“告诉我哪里错了，我能理解”不等于保有能力。若异常标签本身仍由 AI 提供，人可能只是继续消费代理的判断。真正有诊断力的情形是：在不知道测试者故意植入了何种错误之前，主体能否从自己的预期中先产生一个“不对劲”的信号。这个信号可以粗糙，甚至不能立即给出正确答案；它的关键作用是把系统从无条件接受切换到重新检查。

认知卸载：外置本身不是退化，问题在于外置后还剩什么

认知卸载研究提醒我们，不应把使用外部工具本身道德化。Risko 与 Gilbert 把 cognitive offloading 概括为通过外部行动改变任务的信息加工要求。人类一直在把记忆、计算和规划部分搬到纸张、提醒器、搜索引擎和他人身上。Sparrow 等人的“Google effects on memory”提示，预期未来可访问的信息会改变人们记什么：与其记住内容本身，人可能更多记住去哪里找。这样的重分配不等于认知停止，它可能是更高效的资源配置。

因此，“不可外包能力”不能被定义为“必须把知识留在脑子里”。如果这样定义，书写、图书馆、计算器和互联网都会成为能力的敌人。真正需要区分的是两种卸载：一种把低价值的存储和重复操作移出内部系统，同时保留对任务结构的预测与检验；另一种连“什么样的输出值得怀疑”也一并外包。前者使内部模型变薄但仍有差异发生能力；后者使内部模型与代理输出之间失去可产生误差的张力。

这给出了认知维度的第一道门：主体不必保留完整解法，但必须保留足以生成独立预期的最小内部模型。这个“最小”不是固定知识量，而是功能性的——它必须能让至少一类关键异常在 AI 告知之前对主体显现。若一个人对所有候选输出都同样觉得“像对的”，那么无论他最终选中了多少正确答案，都没有形成可与代理发生分歧的内部参照。

第二碰撞面：控制科学——看得出错仍不等于“能力属于你”

可观测不等于可控，系统可控也不等于“你可控”

控制科学把一个常被日常语言混在一起的问题拆开：能看见系统状态，与能改变系统状态，是两回事；系统存在某个控制输入，与这个输入由谁掌握，也是两回事。AI 工作流里，“我知道模型可能有幻觉”只是风险知识；“我能看出这一次具体哪里异常”才接近可观测；而“我能通过另一条路径把任务带回有效区间”才接近可控。

这一区分可以解释一种常见的“高级无力”：使用者已经具备很强的 AI 怀疑意识，每次都说“需要核验”，但他只有同一种代理。模型 A 给出答案后，他让模型 A 自己检查；出现冲突后，他换一个提示词让同一个系统裁判；代码报错后，他继续要求模型自动修复，直到测试通过。整个过程中存在大量“审查动作”，却没有真正独立的控制通道。只要同源模型的盲区稳定存在，所谓核验只是同一控制器内部的自循环。

因此，能力保留必须包含一种低限度的“异源改向权”：主体能够在关键时刻选择不同证据源、不同工具、不同表征方式或局部手工推理，使任务状态发生真实改变。这里的“独立”不是禁止使用任何工具，而是不允许把发生故障的同一代理继续当成唯一裁判。可以查原始数据、跑独立测试、调用另一个模型、请专业同事复核、回到公式、

构造反例；这些都可以是控制输入。真正不可缺的是选择哪条控制通道以及根据反馈决定是否继续的权力。

脱离回路：正常绩效越好，接管能力反而可能越难被看见

Endsley 与 Kiris 关于 out-of-the-loop performance problem 的研究提供了一个与“AI 协助越强越好”直接冲突的事实结构：自动化可以在正常运行时提高表现，却由于情境意识下降、主动加工转为被动加工以及反馈变化，使操作者在专家系统故障后接管变慢。也就是说，正常运行阶段最成功的自动化，恰恰可能把能力缺口藏得最深。

这使“人工在回路中”成为一个不够硬的标准。一个人可以名义上保有最后确认按钮，却不知道系统此刻处在哪个状态，也没有足够的时间、信息或技能去改变它。AI 产品中的“human review”同样可能只是界面事实：审核者可以点同意或拒绝，却无法形成替代方案；可以关闭生成，却不能继续任务；可以看到模型给的引用，却不会判断引用是否支持结论。名义控制与功能控制之间存在一道必须实测的缝。

因此，控制维度给出第二道门：当异常出现时，主体必须不仅能“停止代理”，还要能“让任务继续朝目标移动”。如果唯一动作是停止使用 AI，而停止后任务立即瘫痪，那么主体拥有的是否决权，不是该项能力。反过来，如果主体可以保留 AI 作为工具，但能够切换证据、重设约束、局部重做、缩小问题或召集他人，从而把系统重新带入可接受区域，就已经满足比“完全手工复现”更现实的接管要求。

第三碰撞面：认识论——正确不等于知道，能解释也未必是最低条件

Gettier 之后：结果正确不能自动转化为认知成就

认识论对本章最直接的贡献，是切断“正确结果→主体能力”的自动兑换。Gettier 在 1963 年的经典短文中表明，即使一个信念既真又有表面上的正当理由，仍可能因为偶然结构而不构成知识。把这一结构放入 AI 协作，可以看到相似风险：使用者可能选择了正确答案，却只是因为模型恰好在这次没错；他可能给出一个看似合理的解释，却没有能力区分同样流畅的错误解释。若把运气换成另一个随机种子或另一个模型版本，结论就会崩塌。

但是，本章也不能简单走向“必须能完整陈述理由，否则就不算能力”。认识论外在主义尤其是 Goldman 的过程可靠主义提醒我们，知识与正当性未必要求主体对所有形成过程具有反思性访问。熟练者常常不能把自己的技能压缩成完整口头规则；一个医

生可能先看出影像异常，再在追问中逐步展开理由。若把“能写一篇元分析解释自己为何正确”当作最低标准，会把大量真实 know-how 排除掉。

因此，认识论真正需要的不是“全部理由显式化”，而是更薄的反运气条件：主体必须能在竞争答案之间产生一种有方向的区分，使“为什么接受 A 而不是 B”不完全取决于同一代理的背书。这个区分可以是命题理由，也可以是可演示的反例、可重复的测试、可靠的来源差异、领域约束或实践性识别。重要的是，它必须能够在反事实邻近情境中改变选择，而不是事后为已经接受的答案补一段漂亮说明。

认知主体性：最终需要归属的不是答案，而是“可修订的承诺”

AI 协作还有一个传统“知识正确性”难以覆盖的问题：谁对这个判断承担修订义务？在真实工作里，很多能力不是一次性答题，而是持续承诺。医生签下诊疗意见、工程师批准方案、研究者提交论证、教师给学生反馈，意味着当新证据出现时，必须有人负责改变原判断。若一个人只能说“这是模型给的”，却无法说明什么新证据会使自己撤回或改写结论，他拥有的不是一个可持续的认知立场，而是一次外部输出的转交。

因此认识论维度给出第三道门：主体要能够把一次纠错结果回写成下一轮的判断规则。这个回写不要求形成宏大理论，但至少要表现为“下次遇到同类情境，我会提前检查 X”“这个来源在 Y 条件下不可靠”“这个提示词会漏掉 Z 约束”“这类冲突需要先回到原始证据”。如果每次错误都由 AI 即时修掉而不改变人的下一轮识别与选择，那么系统变好了，人的能力却没有必然增长。

真正的跨学科对撞：三家共同默认了什么，又如何被自己的材料推翻

三家各自的“单维认证”都不够

三家看似互相补充，但如果只是把三条条件并列起来，仍然只是一阶综合。真正的共有前提更深：它们都倾向于把能力当作一种可以在“正常运行截面”上读取的属性——看内部有没有模型、看系统有没有控制权、看当前信念有没有认识论资格。可是在 AI 时代，恰恰是正常运行截面最不具有辨别力：人和代理高度耦合后，正确输出可以持续遮蔽能力究竟位于哪里。

共有前提的自我推翻：能力不是静态库存，而是“断面上的重建能力”

认知神经科学自己告诉我们，误差只有相对于内部预期才出现；控制科学自己告诉我们，自动化的能力问题往往要在故障转接时才显影；认识论自己告诉我们，成功必须在邻近反事实中摆脱偶然并具有可归属的认知关系。三者的材料合起来逼出一个共同方

向：能力不应在代理正常工作的“合成输出”上认证，而应在代理可靠性被打断的“交接断面”上认证。

这里发生了一个从“能力库存”到“能力复闭合”的对象转换。传统技能测试问：你独立拥有多少步骤？本章问：当分布式认知系统的一条关键外部边失效时，你这一侧是否还能重新构造从预期—差异—裁决—干预—反馈到更新的闭环？前者把能力看成动作清单；后者把能力看成一种潜在的结构恢复性质。它允许日常运行时闭环被拉长到工具、同事和 AI 之中，却要求在关键断面上，人仍然具有重新闭合它的能力。

这一判断也改变了“不可外包”的含义。不可外包的不是写第一稿、算第一遍、查第一条资料、写第一段代码这些动作；这些几乎都可以外包。不可外包的是：至少保留一个能与代理产生真实分歧的内部参照，保留选择异源证据与控制通道的权力，并保留把纠错结果转化为自己下一轮规则的责任。若这三者同时消失，任务仍可能完成，但能力已经只能归属于“人+代理”系统，不能无条件归到人这一侧。

涌现物：断代理复闭环判据（PSRC）

定义：什么情况下，大量操作外包后仍可说“人保有能力”

本章提出“断代理复闭环判据”（Proxy-Severance Reclosure Criterion, PSRC）。所谓“断代理”并非要求永久禁用 AI，而是对外部代理施加一次任务相关、可判定的扰动：代理不可用、给出错误、两个代理冲突、证据分布改变、目标约束改变，或模型输出与真实反馈出现系统性失配。所谓“复闭环”，是主体在该扰动发生后，仍能把原本由人机共同完成的判断—行动回路重新闭合到一个可继续学习和控制的状态。

PSRC 的零情态词判据如下：在事先规定的代理扰动 P 下，若主体在不把同一失效代理继续作为唯一裁判的条件中，依次完成“生差、裁差、改向、回写”四件事，则该能力归属主体；任一环节长期缺失，只能证明存在代理绩效，不能证明主体保有该能力。

这里四件事不是四门“新技能”，而是同一个能力归属事件的四个相位。“生差”负责让异常在主体侧出现；“裁差”把异常转成理由结构；“改向”使理由进入真实行动；“回写”使行动结果成为下一轮内部模型的一部分。缺少任何一环，闭环都无法真正归到主体。

为什么它比“独立完成”更薄，比“批判性思维”更硬

PSRC 故意不要求主体完整复现 AI 所做的全部操作。一个优秀医生不需要手算医学影像模型的卷积，一个程序员不需要自己写出所有样板代码，一个研究者不需要背诵全部文献。只要在关键代理扰动下，他仍能生差、裁差、改向和回写，就可以合理地把能力继续归属于他。由此，深度自动化与能力保留并不矛盾。

同时，PSRC 比“保持批判性思维”更可裁决。“要批判”“要核验”“不要盲信 AI”都可以成为态度口号，却不告诉我们何时算真正做到。PSRC 把判定压到一次具体扰动：错误没有被预告时能不能自己看出来；两个输出冲突时有没有异源区分理由；选定理由后能不能改变任务；改完之后下一轮是否真的不同。它不测态度，而测闭环能否复建。

最关键的是，PSRC 不把“没有 AI 也能做”当作充分条件。一个人可能能手工执行旧流程，却看不见情境已经变化；他可以在熟题上独立完成，却无法解释新证据为何使原规则失效。这样的“手工作业能力”不一定具有真实适应性。复闭环判据要求的不只是去工具化，更是对扰动的敏感、异源控制与更新。

2×2 辨别格：同样拿到正确答案，四种状态并不相同

为避免把四环节变成模糊总分，本章先用两个最可观察的轴建立 2×2：横轴是“异常能否在本人一侧独立显形”（生差+裁差），纵轴是“本人是否拥有能把任务带回有效区间的改向路径”（改向）。进入右上靶格后，再以“是否回写”作为能力归属的封印条件。

这个辨别格直接解决一个教育测量难题：学生的 AI 辅助作品质量并不能告诉我们他在哪一格；纯闭卷考试也只能粗略测到某些内部库存。更有诊断力的方式，是给同样完成高质量作品的人一次可控的代理扰动，看谁能在没有答案提示的情况下先发现异常、选择不同证据并完成修正。

敢意拓宽：这个想法是不是早已有了？

本书要求每一个候选概念在提出时就主动寻找“这想法其实早就有人说过”的占位者。本章采用五方向检索：1980 年前经典层、AI/自动化应用领域内部、方法学占位者、专著/非期刊层，以及作者同产线既有文本。检索并没有得到“完全无人涉足”的安全区；相反，找到了多位危险近邻。本章的创新主张因此必须被压缩为更窄的一条：不是首次提出“人应保持控制/批判性思维/know-how”，而是把能力归属的最低认证位置放到“代理失效的断面”，并用复闭环而非正常绩效或全手工复现来裁决。

最近邻判决：本章最容易被哪三句话吸收

第一句：“这不就是批判性思维吗？”——不是。批判性思维可以在不承担真实控制后果的讨论中发生；PSRC 要求理由进入改向，并由反馈回写。若一个人能批评 AI 答案却无法使任务继续，批判性思维存在而任务能力仍未被认证。

第二句：“这不就是 human-in-the-loop / takeover competence 吗？”——不完全是。接管能力主要解决自动化失效时能否恢复人工控制；PSRC 额外要求异常必须先在人的内部参照上显形，并要求修正结果回写为下一轮认知规则。一个熟练操作员可以按故障手册接管，却不知道模型为何不可信；他有接管能力，不一定拥有当前判断所需的认识论能力。

第三句：“这不就是扩展心智/分布式认知吗？”——方向相反但兼容。扩展心智关心工具何时可以构成认知系统的一部分。PSRC 接受“人+AI 系统”可以拥有很强能力，却追问另一件事：在必须单独把能力记到“人”这一侧时，最低还要留下什么。系统层能力与个体归属可以同时成立而不同值。

真跑一条：用公开 GenAI 实验压力测试“即时增强=能力保留”

最强竞争解释与预注册式判决

本章最强竞争解释 H_0 是：“若 AI 协作使一个人在任务上显著表现得更好，则至少在相邻同类任务中，人已经获得或保留了相应能力；因此不必额外设置断代理测试。”若 H_0 成立，在 AI 协作结束后转入无 AI 任务，先前的表现增强应当以可观察形式持续存在。若即时增强在撤除代理后系统性消失，则至少可以否定“即时高质量结果本身足以认证个人能力”。

公开数据机械编码

Wu 等人在 Scientific Reports 发表的四项在线实验总样本 N=3,562，明确比较了 AI 协作的即时表现与随后的人类独立任务。本章不重新估计原始效应，而对作者公开报告进行一条最便宜、可复核的二值编码：只要 Collab-Solo 组在后续独立任务至少一个核心质量维度显著优于 Solo-Solo 且无核心维度显著变差，记为“存在正向持续”；否则记为“不存在稳定持续”。

按这一公开规则，4 项研究中只有 1 项出现部分正向持续，3 项没有；“持续率”为 25%。这个读数不能被解释为“AI 会让能力下降 75%”，因为四项任务、指标和转场的含义不同，也没有直接测本章 PSRC 四环节。但它足以击穿一个更弱却常被默认的等价式：AI 协作时的即时表现增强，不足以自动推出撤除代理后个人能力增强。

尤其 Study 4 使用更相似的连续文本任务并平衡任务顺序，仍未发现稳定的后续独立增强。

不利结果：公开数据也迫使本章收窄自己的主张

如果本章把“撤掉 AI 后成绩不提高”直接解释成“能力不存在”，同样会过度推断。Wu 等研究自己指出，转场可能影响控制感、内在动机与无聊感；Study 1 甚至出现了后续创意质量的部分正向持续。这意味着“独立复做一次”也不是能力的充分测量：它混入动机、任务迁移、疲劳与题型差异。

这一不利结果反过来修正了 PSRC 的粒度。本章不主张用“完全禁 AI 后的闭卷成绩”替代 AI 辅助成绩，而主张使用“定向代理扰动”：只破坏代理的一条关键可靠性，并观察主体能否生差、裁差、改向和回写。这样测到的是代理失效时的结构恢复能力，而不是把整套现代工具环境突然抽空后的孤立绩效。

因此真跑得到的不是对 PSRC 的证明，而是对竞争解释 H_0 的一次否定与对本理论的一次收窄：结果质量不能认证能力；无 AI 成绩也不能单独认证能力；最有辨别力的对象是人机分工被局部扰动时，闭环由谁重新建立。

五个场景的判决：最低能力核到底长什么样

写作：不会从零写满篇，不等于不会写；看不出论证塌了，才更危险

一个研究者可以把语言润色、结构草拟、摘要压缩甚至部分段落生成交给 AI。如果 AI 把一个相关关系写成因果关系，研究者能在看到输出时因自己的研究设计预期而产生异常信号；能回到原始方法与数据区分竞争解释；能重写结论并限制适用范围；下一次会把“因果措辞核查”写进自己的审稿规则，那么写作能力仍可归属于他。反之，一个人可以在 AI 帮助下交出文采极好的文章，却无法发现论证从证据跳到了结论；这属于高代理绩效，而非完整学术写作能力。

翻译：不必记住所有词，但必须有能与流畅表面冲突的语感/语境模型

AI 翻译最容易制造“漂亮但错”的假安全感。真正的最低能力不是逐词手工翻译，而是对语域、指代、术语和语用后果保留足够内部预期。当两个译法都流畅时，译者能通过上下文、领域术语表、原作者用法或目标读者反应裁差；能改写并检查后文一致性；能把这次错译类型加入下一轮检查清单。若使用者只会比较“哪个更顺”，AI 一旦稳定地产生流畅错误，能力断面便暴露为空。

编程：能让 AI 写代码不等于编程能力；能局部构造失败并定位约束更重要

程序员可以把大量样板、接口调用和测试代码交给模型。PSRC 不要求他背下所有 API。但当测试看似全绿而线上结果异常时，他需要能从系统约束生成预期，构造最小复现、独立断言或日志观察点，在两个可能原因间裁差，改变代码或架构，并把故障模式转成新的测试。一个只会把报错重新粘给模型直到不报错的人，其“改向”仍由代理承担。

医学/法律/金融等高责任领域：最终不可外包的是“何时撤回承诺”

在高风险领域，PSRC 的意义更加直接。专业者可以使用 AI 扩大检索和生成能力，但不能把“什么新证据会使我改判”一并外包。能力的主体必须保留异常触发条件、异源证据通道、停止或升级机制以及修改原决定的责任。此处的回写不仅是个人学习，也是职业责任：系统出现新信息后，必须存在一个能重新打开判断并承担后果的人。

教育：把“禁止 AI 考试”改成“代理扰动考试”

教育最值得改变的，不是简单决定“作业能不能用 AI”，而是把能力认证从产物评分转向断面测试。学生可以被允许使用 AI 完成高质量初稿，然后教师随机植入四类扰动：给出一个高流畅度错误、提供两个相互冲突的模型答案、改变一个关键情境条件、暂时移除一个惯用工具。真正的评分对象是学生能否先发现异常、说出或演示区分依据、选择修正路径并把结果转成下一轮规则。

这种考试不要求回到“全手工时代”。它甚至比传统闭卷更接近真实职业环境：现实中工具不会消失，但工具会错、会缺、会冲突、会过时。教育真正必须由人亲自形成的，不是每一个可自动化动作，而是能在工具失去可靠性时重新组织知识与行动的能力。

一个可执行的“能力归属测试”协议

为了让 PSRC 不沦为概念口号，本章给出一个最小测试协议。它不是心理量表，而是一种情境任务。每次测试只扰动一条代理关系，避免把“没有工具的困难”误当成“没有能力”。

步骤 1：先让受试者在正常 AI 环境中完成一项真实任务，记录辅助绩效作为基线，但不据此认证能力。

步骤 2：在受试者不知具体错误位置的情况下，对代理施加一种预先登记的扰动：事实错误、逻辑矛盾、引用不支持、边界条件改变、工具不可用或两个代理冲突。

步骤 3：记录“生差时间”——从错误出现到受试者第一次主动提出异常的时间；若异常完全由系统提醒，不计为主体生差。

步骤 4：要求受试者在至少两个候选解释之间给出会改变选择的区分依据。允许使用外部资料，但不得让同一失效代理成为唯一裁判。

步骤 5：观察受试者是否能发起一项改变任务状态的干预，而不是只表达怀疑。记录干预是否把输出带回预注册的有效区间。

步骤 6：在一段间隔后给出结构同类、表面不同的新扰动，检验上一轮反馈是否已经回写。若第二次仍完全依赖外部提示，第一次修正可能只是临时救火。

这个协议最重要的伦理边界是：PSRC 用于判断“此项任务中的能力归属位置”，不用于给人贴稳定人格标签；扰动必须与真实任务风险相称，不能为了测接管而在医疗、交通、金融等安全关键系统中故意制造现实危险；任何用于教育、雇佣或职业认证的阈值必须公开编码规则，并允许复核。

可证伪预测：什么结果会让本章的判断失败

若一个概念只能解释所有结果而不能输，它就不是一个有用的学术判据。PSRC 至少押以下四条可裁决预测：

与作者既有两篇“代理”论文的自家划界：为什么这篇不是重复

与“实质主控”的区别：有按钮、有权限，仍可能没有能力

作者此前在成瘾干预论文中提出“实质主控”四条件，核心问题是外部控制系统能否被用户理解、修改、暂停并最终交还。那个概念判断的是“控制权是否实质属于本人”。它非常接近本章的控制科学维度，因此若本章继续把新意写成“AI 失败时人应能接管”，就会被自家稿件正面占位。

本章把落点向前挪了一步：即使一个 AI 系统完全透明，用户随时可以关闭、修改、撤回，能力仍可能不属于用户。因为他可能不知道什么时候应该关闭，不知道两个答案哪一个更可信，关闭之后也没有替代认知路径。“权限归你”是主控条件；“能力归你”还要求异常在你这一侧发生、理由能区分、行动能改向、反馈能回写。故 PSRC 包含但不等同于实质主控。

与“消化性裁决”的区别：本章不要求每一轮都亲自裁决

作者另一篇饮食论文强调外部最优解可能填平“悬停—介入—投掷—收割”的裁决过程，使裁决能力因长期闲置而萎缩。那篇文章的保护对象是某类日常自我裁决过程，带有较强的“这一环节必须由主体亲历”的生成机制主张。

当前论文更宽容也更一般。它承认成熟专业者可以在绝大多数常规轮次中让 AI 直接完成操作，甚至不需要每一次都展开完整的悬停与投掷。只要系统定期或在异常时能够把闭环重新拉回主体，人仍可保有能力。换言之，上一文问“哪些裁决过程不能被持续短路”；本章问“当大量过程已经合法外包，能力归属的最后界线在哪里”。两篇的共同背景是代理，但判定对象不同。

（三）与本书其余各章的分界：三条不同的判定对象

本章的判据在本书内部还有三处最近的邻居，必须一并说清，否则读者会把四章读成同一章的四次重复。

与第五章的分界在时间结构。第五章问的是一次具体决定是否由人形成，因此它要求在每一次被声称为"人做的决定"里都能找到决定性理由；本章问的是一项能力是否仍归属于人，而能力可以长期不被调用仍然存在。判断必须每次重新形成，能力允许长期待命。因此一个人完全可能在某次决定上不满足第五章的三节点（他没有理由，只是照单批准），却在本章的扰动测试中顺利复闭环；反过来也成立。若在同一批受试者身上，两项测试的通过与否高度一致（相关高于 0.8），则两者不是两个维度，本章与第五章应当合并。

与第三章的分界在提问方向。本章问"能力还在不在"，第三章问"我们凭什么知道它还在不在"。前者是能力的存在问题，后者是证据的存在问题。一个人可以能力完好而证据陈旧——第三章判定证据债高，本章的扰动测试却会判定他通过。这个差别不是文字游戏：证据债可以在能力没有任何变化的情况下累积，而清偿证据债也不会自动提高能力。

与第二章的分界在功能。本章给出的是判定，第二章给出的是维护。判定回答"到此为止能否归属"，维护回答"如何让它继续可归属"。本章不承诺任何维护效果，第二章也不重新定义能力归属；两章的关系是同一条线上的诊断与处方。

理论含义：AI 时代“人必须亲自形成”的不是任务，而是三种关系

第一，是人与世界之间的差异关系。人必须形成足以让世界或 AI 对自己“显得不对”的内部参照。知识可以外置，计算可以外置，文字可以外置，但如果任何输出都不能撞出主体自己的误差信号，主体就失去了独立观测的起点。

第二，是人与行动之间的改向关系。主体必须保留在关键情形下选择另一条证据与控制路径的能力。这里的“亲自”不是手工，而是发起位置：由谁决定停止、切换、缩小问题、改变方法、求助谁、接受什么风险。AI 可以执行改向，但改向的触发与裁决不能只有 AI 自己。

第三，是人与后果之间的回写关系。主体必须把纠错后果变成自己的下一轮规则。一个永远靠代理即时修复而不积累任何新的判别边界的人，会拥有越来越强的系统绩效，却不一定拥有越来越强的个人能力。真正的学习发生在“这一次哪里错了”改变了“下一次我先看哪里”的时刻。

这三种关系共同说明：AI 时代的人类能力不必与机器争夺操作密度。人可以越来越少地亲手写字、查表、算数、翻译、搭代码，但必须越来越清楚自己凭什么产生异议、凭什么改向，以及什么反馈会改变下一次判断。外包的是操作；不能外包的是让外部代理在必要时重新变成“工具”而不是“能力持有者”的那条关系结构。

限制与研究议程

第一，PSRC 目前是一个理论—方法判据，不是已经完成心理计量验证的量表。四环节的编码一致性、不同领域的阈值、扰动难度与真实职业表现之间的预测效度，都需要实验研究。尤其“生差”可能依赖领域熟悉度，“改向”可能依赖组织权限，不能把系统不给权限误判为个人无能力。

第二，能力归属具有粒度依赖。同一个人可以保有“判断医学文献证据等级”的能力，却不保有“独立完成统计建模”的能力；也可以保有“调试软件架构”的能力，却把某个语言的语法细节完全外包。PSRC 必须对具体任务边界下定义，不能用一次宏观测试宣布一个人“有/没有 AI 时代能力”。

第三，本章只讨论“把能力归到个人”这一评价问题，不否认分布式系统本身可能具有更高能力。在团队科学、航空、医疗、软件工程等领域，最安全的设计往往不是要求某个孤立个人可以全部接管，而是保证系统在节点故障时仍能通过团队、工具和制度复闭环。个体 PSRC 与系统韧性应当分别测量。

第四，认识论外在主义仍构成真实挑战。如果一个人长期依赖一个极其可靠、透明且受制度审计的 AI 系统，即使他个人无法在代理失效时接管，我们是否仍可说他“知道”？本章的回答是：可以讨论他是否在制度性意义上获得知识，但不能据此自动把相应“任务能力”归属于他。知识归属与能力归属可能需要分开，这正是未来与社会认识论进一步对话的方向。

第五，本章的新意主张应保持克制。Andrada 与 Carter 关于 mind-technology 与 know-how 的近期讨论已经直接触及技术如何改变能力归属；自动化、人机交互、扩展心智、认知卸载和 AI 批判性思维也分别占据了大片相邻空间。本章能够主张的不是“第一次发现外部技术会影响人的能力”，而是提出一个跨三领域、面向生成式 AI 知识工作的可裁决最小归属结构，并把判定位置从正常表现迁移到代理失效后的复闭环。这个主张仍需要更系统的站外文献检索与实证比较。

结论：AI 时代最不可外包的，是“把能力重新收回来”的能力

当 AI 已经能够替人完成越来越多任务时，最容易犯的错误有两个。一个错误是守着旧技能清单，认为凡是机器能做的，人就必须继续手工练到同样熟练，否则“能力退化”；另一个错误是只看最终结果，认为只要人借 AI 持续得到正确答案，他就已经拥有了对应能力。前者把工具进步当敌人，后者把代理绩效当成人的能力。

本章的跨学科对撞给出第三种答案。认知神经科学告诉我们，人至少要保留能生成独立预期的内部模型，否则错误不会在自己这一侧发生；控制科学告诉我们，能看见异常还不够，人必须拥有把任务带回有效区间的异源改向路径；认识论告诉我们，正确结果还必须摆脱偶然并成为可由主体修订的认知承诺。三者合起来，使“能力”的最低单位从动作库存变成了一个可在代理失效时重建的闭环。

因此，一项任务能被 AI 高质量完成，不意味着完成该任务所需的人类能力可以取消；它只意味着能力不再需要以“亲手完成全部操作”的形式存在。一个人可以把大量检索、生成、计算和表达外包出去，只要在关键扰动来临时，他仍能让异常显形、对冲突作出有理由的区分、改变任务方向，并把结果写回自己的下一轮判断。到达这个最低条件之后，工具越强，人的操作可以越少；低于这个条件之后，结果越好，反而越可能掩盖能力已经迁移到外部代理。

AI 时代真正值得教育、组织与个人保护的，因而不是一张“机器永远替代不了的人类技能清单”。那张清单会不断缩短。真正不可外包的是一个更小、却更稳定的东西：当代理不再可靠时，主体仍有能力把判断—行动—反馈的闭环重新收回来。

附录 A｜敌意拓宽检索词与命名闸记录

为避免把“同一想法换名”误作创新，本章使用以下形式层检索词（中英文混合）进行敌意拓宽，并对候选名进行精确短语搜索。这里只记录足以复核检索方向的词，不把“未命中”解释为绝对无人提出。

“correct performance does not imply competence” / “AI assisted performance human skill retention”

“out of the loop takeover competence automation failure” / “human takeover after automation failure”

“ability attribution technology know-how” / “mind technology problem know how”

“cognitive offloading skill retention internal model” / “Google effects memory external memory”

“epistemic agency AI belief revision” / “appropriate reliance AI advice”

“capability ownership human AI delegation” / “ability attribution distributed cognition”

候选命名精确检索：“proxy-severed reclosure” / “proxy severance criterion” / “reclosure competence” / “capability attribution cut”

自家语料检索：“外部代理 主控 接管 能力”“正确答案 能力 代理”“AI 能力 归属”

命名处理：检索发现“recoverable agency”“takeover competence”“recovery loop”等已有大量占位，因此本章不采用这些名称。最终使用“断代理复闭环判据/Proxy-Severance Reclosure Criterion”作为工作名。该名称的原创性仅限于本章检索范围，不作“全球首次命名”的绝对声明。

附录 B｜真跑编码规则（单一口径）

编码对象：Wu et al. (2025)四项公开实验的“AI 协作后的后续人类独立任务”。编码阈值全文仅用一套：若 Collab-Solo 组在后续任务至少一个核心质量维度显著优于 Solo-Solo，且不存在核心维度显著变差，记 1；否则记 0。Study 1=1；Study 2=0；Study 3=0；Study 4=0。该编码只用于否定“即时 AI 增强足以推出稳定个人增强”，不用于估计 AI 对能力的因果损伤率。

本章附表

来源	S/D/E 落位	核心材料	单独够用时的直觉
认知神经科学	S：主体内部留下了什么	内部模型、预测、误差信号、元记忆与认知卸载	只要保留足够内部模型，外部工具可被视为资源扩展
控制科学	D：主体能怎样改变过程	状态估计、可观测性、可控性、监督控制、故障接管	只要操作者能在关键时刻重获控制，平时可高度自动化

认识论	E：什么条件让结果算“主体知道/会”	理由、可靠过程、反运气、认知功劳与主体性	正确结果必须具有可归属于主体的认识论地位

维度	单维凭证	为什么会失败
S：内部模型	“脑中还有足够表征，就算会”	内部模型可存在但过时；人可能知道规则却无法在时间压力下接管；表征本身不保证行动与责任。
D：功能控制	“故障时能把系统带回来，就算会”	接管可以依赖他人或第二个黑箱；功能恢复不保证理由属于主体，也不保证下一轮能识别同类错误。
E：认识论资格	“有可靠理由/可靠过程，就算会”	可靠过程可能主要位于外部代理；一个人可能在环境稳定时可靠，却对代理失效毫无诊断与干预能力。

环节	操作定义	主要母体	最低要求
1. 生差（Generate discrepancy）	主体先于纠错标签生成一个任务相关预期/约束，使 AI 输出能够以“偏差”显现。	认知神经科学：内部模型与预测误差	不是记完整答案；是能让某类错误对自己显形。
2. 裁差（Discriminate）	主体用非同源理由、测试、证据或约束在至少两个竞争解释/答案间作出有方向的区分。	认识论：反运气、理由与认知归属	不要求完整口头证明；要求选择会随证据变化。
3. 改向（Intervene）	主体选择并执行至少一条能改变任务状态的异源控制路径，使系统朝目标区间移动。	控制科学：可控性与接管	只会“关掉 AI”而任务瘫痪，不算完整接管。
4. 回写（Update）	主体依据干预反馈改变下一轮检查规则、内部模型或代理使用策略。	学习/认识论主体性	错误被系统修掉但人的下一轮不变，不算能力增长。

位置	名称	典型表现	判定
低异常辨识 × 低改向	代理依附绩效	AI 正常时可高分；错误不显形；关闭 AI 后无替代路径。	不归属主体
高异常辨识 × 低改向	审计型无力	能说“不对”，能指出冲突，却只能把问题继续交回同一代理。	保有局部审计能力，不等于任务能力
低异常辨识 × 高改向	程序型接管	会按手册/流程操作，但不知道什么时候该接、为什么接；易在错误时	保有操作能力，不足以认证判断能力

		机干预。	
高异常辨识 × 高改向	复闭环候选	能发现、区分并改变任务状态；若干预后还能回写下一轮规则，则能力归属主体。	靶格：须再过“回写门”

方向	占位者	它已经说了什么	本章仍需补的最后一步
经典层	Gettier (1963)	正确/有理由仍可能因运气而非知识。	占据“正确≠知道”；但不处理人机任务分配、故障接管和能力归属。
经典层	Sheridan & Verplank (1978)	人—计算机监督控制与自动化层级。	占据“任务可分配/人可监督”；但不建立知识能力的归属判据。
经典层	Goldman (1979)	可靠过程可赋予正当性，不要求完全内省理由。	直接阻止本章把“能口头解释全部理由”误作最低条件。
应用内部	Endsley & Kiris (1995)	自动化使操作者在故障后出现脱离回路接管问题。	最危险工程近邻；但研究对象是自动化接管表现，不是跨知识任务的能力归属。
应用内部	Lee et al. (2025, CHI)	GenAI 使批判性思维转向核验、整合与任务监管；高 AI 信心与较少批判性思维相关。	占据“AI 时代人的工作转向核验”；但主要是自报告实践，不给能力最低认证规则。
应用内部	Wu et al. (2025)	AI 协作提升即时表现，但总体未产生稳定的后续单独任务增强。	占据“即时绩效≠持久个人增益”；是本章真跑数据来源，但未回答能力归属的必要充分条件。
认识论新近邻	Andrada & Carter (2025)	直接讨论技术介入给 know-how/ability attribution 带来的问题。	最危险哲学近邻；迫使本章把新意限定到“断代理复闭环”的可操作归属测试。
自家已发	“实质主控”（成瘾论文）	可见、可理解、可操作、可验证/交还；外部控制撤除后本人能否接回。	它判断控制权是否实质属于用户；本章进一步判断“任务能力”是否属于人，即使按钮和权限全在也可能不过 PSRC。
自家已发	“消化性裁决”（饮食论文）	外部最优解可能短路悬停—介入—投掷—收割，使裁决能力闲置。	它强调某些裁决过程必须亲自发生；本章更宽：日常操作可以大量代理，只要求扰动时能复闭环。

公开研究	代理在线时	撤除代理后的独立任务	本章机械编码
Study 1	即时 AI 协作改善输出质量	后续创意质量小幅提高，但数量无差异	1（部分持续）
Study 2	即时 AI 协作增强	后续数量、创新性、实用性均无显著差异	0
Study 3	即时 AI 协作增强	后续数量、创新性、实用性均无显著差异	0
Study 4	AI 协作显著提高文本长度/分析性/积极语调等即时指标	后续分析性与积极语调无差异，Collab-Solo 文本长度反而略短	0

预测	本章押注	判伪条件
P1 断面预测	在 AI 辅助绩效相同的两人中，PSRC 高者应在未预告代理错误下更快自主发现异常，并以更少同源追问完成修正。	若 PSRC 分数控制领域知识后不预测任何故障恢复指标，判据失去增量效度。
P2 训练预测	“先做独立预测→再看 AI→强制解释差异→间歇故障演练”训练，应比只练提示词在复闭环上提升更多，即使两组平时 AI 绩效相近。	若只练提示词同样提升代理扰动下的生差、裁差、改向、回写，则本章关于内部闭环训练的机制过强。
P3 工具允许预测	允许使用第二证据源/第二工具的扰动测试，应比彻底禁用所有工具更能预测真实工作中的故障处理。	若纯闭卷成绩稳定优于 PSRC 任务预测真实接管，本章“局部断面优于全撤工具”的主张失败。
P4 更新预测	一次成功纠错后，真正保有能力者在结构同类新任务中的异常发现应提前或策略应改变。	若后续行为完全不变却仍持续高质量处理同类故障，本章把“回写”设为必要条件需要删除或降级。

本章说明

本章属于理论与方法研究，不直接招募研究参与者，不使用可识别的真实个案资料。公开数据压力测试仅使用已发表论文中公开报告的汇总统计与作者结论，不重新接触参与者层面数据。本章由作者提出研究问题、碰撞学科与发表场景；生成式人工智能参与公开文献检索、敌意拓宽、理论结构推演、公开数据的机械编码与文档排版。所有事实、引文、规范判断、原创性主张与最终发表责任应由最终署名作者复核并承担。

特别说明：本章对“断代理复闭环判据”的新颖性主张是检索范围内的工作性主张，而非对全球文献的穷尽证明。已有 mind-technology、know-how、out-of-the-loop、appropriate reliance、cognitive offloading、critical thinking 与

distributed cognition 研究构成实质近邻；正式投稿前建议再次使用 Web of Science/Scopus/PhilPapers/ACM Digital Library 进行系统占位检索。

第 二 章

待命不等于可用

最低接管维持剂量与接管保真环　——航空人因工程 × 核设施运行与维修工程 × 运动训练学的跨学科对撞

摘　要：生成式人工智能正从辅助工具转变为长期认知代理，持续承担检索、分析、计算、写作、判断乃至部分执行任务。由此出现一个区别于“哪些能力不能外包”的维护难题：当人曾经真实掌握某项能力，但此后长期缺少独立调用时，怎样在不要求人重复完成全部低价值劳动、也不牺牲 AI 效率优势的前提下，使关键能力仍然处于可接管状态？本章围绕同一个目标组织三条路径，将航空人因工程关于自动化脱环、情境意识与故障接管的材料，核设施运行与维修工程关于待命系统、潜伏失效、监督试验与试验间隔优化的材料，以及运动训练学关于去训练、最低维持剂量与再适应的材料置于同一冲突面。三门学科共同暴露出一个未被充分说出的前提：能力通常被当成一种可从正常运行表现或单一指标中读取的库存；然而在长期 AI 代行条件下，正常绩效恰恰可能遮蔽备用的人类路径是否仍可启动。本章据此提出两个工作性概念：最低接管维持剂量（Minimum Takeover Maintenance Dose, MTMD）与接管保真环（Takeover Fidelity Loop, TFL）。前者将能力维护剂量从单一频率改写为异常发现、情境重建、独立裁决、干预执行和恢复验证等分量的向量；后者把能力维护组织为“拆门—微调用—待命暴露累计—风险触发证明—扰动接管—恢复回写”的循环。本章进一步以公开的抗阻训练维持研究作为最低成本压力测试，发现相同低剂量对不同适应指标的保持并不一致，从而否定“一个能力对应一个最低维持剂量”的简单标量模型，并与航空研究中手工操纵相对保持而认知导航、故障识别更易下降的现象形成跨领域重复。本章主张：AI 时代能力维护的最低单位不是完整任务重演，而是长期待命后重新进入控制的接管跃迁路径；有效校验也不应只证明能够复现标准答案，而应证明人在 AI 不可用、AI 与现实冲突、多 AI 分歧或情境越界后，仍能独立完成异常发现、情境重建、判断、干预与恢复。

关键词：生成式人工智能；能力退化；自动化脱环；待命系统；最低维持剂量；证明试验；接管能力；人机协作

Standby Is Not Availability: Minimum Takeover Maintenance Dose and a Takeover Fidelity Loop for Human Capability in the Generative-AI Era

Abstract: As generative AI increasingly performs retrieval, analysis, calculation, writing, judgment support and execution, a new capability-

maintenance problem emerges: how can an already acquired human capability remain operationally available after months or years of AI delegation without forcing people to repeat all low-value work manually? This paper collides three bodies of knowledge—aviation human factors on out-of-the-loop performance and takeover, nuclear operations and maintenance engineering on standby systems and latent failures, and exercise science on detraining and minimum maintenance dose. The collision yields a shift from treating capability as a stock to treating it as a standby-to-control transition. We propose Minimum Takeover Maintenance Dose (MTMD), a vector across distinct capability gates, and the Takeover Fidelity Loop (TFL), a maintenance cycle consisting of capability decomposition, minimal real activation, accumulation of standby exposure, risk-triggered proof testing, disruptive takeover, recovery verification and adaptive re-dosing. A proof event is valid only when a person can detect an abnormality, reconstruct task state from sufficiently independent evidence, make an independent judgment, intervene, and verify recovery after AI is unavailable, conflicting, or contextually unreliable. A stress test using published resistance-training maintenance data rejects the simpler hypothesis that one reduced practice dose can represent preservation of an entire capability. The framework therefore argues that mature human-AI systems should minimize unnecessary human labor subject to one constraint: the dormant human pathway must remain demonstrably capable of re-entering control before a real failure demands it.

Keywords: generative AI; human capability; out-of-the-loop performance; skill retention; surveillance testing; minimum maintenance dose; takeover fidelity

引言：从“不可外包”转向“如何维持可接管”

关于生成式人工智能与人类能力的讨论，通常围绕两个问题展开：第一，AI 能够替人完成哪些工作；第二，哪些能力即使在 AI 时代仍然应该由人掌握。这两个问题固然重要，但随着 AI 从偶尔调用的工具变成长期在线的认知代理，第三个问题开始变得更加基础：一项能力即使已经被人掌握，如果此后几个月、几年主要由 AI 代行，它还能否在真正需要时被重新调用？

这不是“有没有学会”的问题，而是“学会以后如何不失活”的问题。一个医生、工程师、教师、研究者、律师、财务人员或管理者可能在 AI 进入工作流以前已经接受多年训练，也曾经能够独立检索、分析、判断和执行。然而，当 AI 持续承担绝大多数正常工作以后，人从高频执行者逐渐变成监督者、批准者和异常时的最后责任人。真正危险的能力变化，未必首先表现为知识被完全忘记，而可能表现为某些关键环节长期没有被调用，以致一个人仍然能够识别熟悉答案，却已经不能在陌生、冲突或失效情境中迅速重建完整控制。

如果对此采取最保守的方案，即要求人周期性关闭 AI、从头到尾重新完成全部任务，技术带来的效率收益会被大量训练成本重新吃掉。尤其在知识工作中，检索、格式化、基础计算、常规摘要和初稿生成等环节恰恰是最适合交给 AI 的低价值重复劳动。如果所谓能力维护要求人不断重复这些劳动，能力维护就会退化为对技术进步的抵消。

相反，如果组织因为 AI 在绝大多数时间表现稳定，就把“系统长期没有出事”解释成人工接管能力仍然完好，也会产生另一类错误。长期没有真实故障，只能说明主运行通道没有暴露严重失败，并不能自动证明长期待命的人类备用通道仍然可用。越稳定的自动化系统，越可能减少人获得自然练习与真实接管的机会，这使“正常绩效”和“备用能力”之间出现系统性脱钩。

本章所要解决的因此是一道典型的路径问题：不是重新定义什么叫能力，也不是单纯解释为什么能力会退化，而是建立一条从“长期由 AI 代行”走到“人在必要时仍能独立接管”的可执行维护路径。这类问题要的是路径而不是维度；本章因此不把三门学科作为并列综述，而是让它们围绕同一个目标给出彼此冲突的维护逻辑，并寻找三者合在一起才会显露的中间结构。

理论边界：能力归属与能力维持不是同一个问题

此前“断代理复闭环判据”已经处理了另一个更基础的问题：当 AI 参与检索、生成、分析和表达后，正常输出质量不再足以判断某项能力究竟属于人、属于 AI，还是只能归属于“人+AI”的联合系统。那项研究提出，能力归属应当在外部代理受到可判定扰动的断面上观察：只有当主体仍能生成独立预期、发现差异、裁决冲突、实施干预并把结果回写到下一轮判断规则中，才有理由继续把该能力归属于主体。

然而，一次通过能力认证并不能自动回答跨时间的维持问题。一个人在今天能够完成断代理后的闭环重建，不代表六个月以后仍然能够完成。相反，若 AI 在六个月里持续替他执行关键环节，人类能力的各个分量可能以不同速度衰减。因此，能力归属研究

回答的是“现在这项能力是否仍然可归于主体”，本章回答的是“怎样让这一判断在长期 AI 代行条件下继续成立”。

二者的差异决定了本章不能简单重复一次性测试。若维护方案只是每隔一段时间完整重复一次能力认证，就等于用高成本的全量演练代替精细的维护工程。本章真正需要找到的是：哪些子能力需要真实使用，哪些操作可以长期外包；哪些信号能够提前说明备用路径可能失活；以及怎样以尽可能小的训练成本换取足够高的接管可靠性。

因此，本章把“能力”暂时视为一个跨时间的可调用系统，而不是静态知识库存。一个系统可在日常运行时把大量计算、记忆、搜索和表达延伸到外部工具中，但如果在关键扰动发生时，人仍然负有最后责任，那么至少必须保留一条能够从待命状态重新进入观察、判断、控制和恢复的人工路径。本章将这一对象称为“接管跃迁路径”。

三个学科的核心材料与冲突

航空人因工程：自动化越成功，接管越接近一次冷启动

航空自动化研究很早就发现了一个反直觉现象：自动化系统往往把常见、规则化、适合反复练习的任务交给机器，却把稀少、异常、复杂且难以提前预设的任务留给人。Bainbridge 在《Ironies of Automation》中指出，自动化越成功，人越少通过日常任务保持对异常处理的熟练度；而真正需要人介入的恰恰是机器最难处理的情况[1]。这使“自动化成功”与“人工备用能力得到练习”之间形成方向相反的关系。

Endsley 与 Kiris 关于 out-of-the-loop performance problem 的研究进一步显示，当操作者长期处于自动化监督位置时，一旦需要重新接管，故障后的决策与行动会受到情境意识下降和控制参与不足的影响[2]。这意味着接管不是简单地把自动化开关关闭，然后原有人工技能自动上线。人在重新控制之前，需要首先恢复对当前状态、任务进程、可靠信息、约束变化和潜在风险的理解。

Casner 等对航空公司飞行员的模拟研究揭示了更细的分化：人工操纵、仪表扫描等部分技能可以保持得相对稳定，而导航程序推进、异常识别和某些认知性人工飞行技能更容易出现下降[3]。这项结果极其重要，因为它否定了“飞行技能”可以由一个总分代表的想法。一个人可能仍然具备基础操作动作，却在重新理解发生了什么、下一步应当做什么的问题上明显变弱。

因此，航空领域给出的不是“人必须经常手工操作”这样简单的建议，而是一条更精确的机制：高度自动化首先改变的是人的位置——从连续控制者变成长期待命的监督者。能力风险因此集中在“重新进入控制”所需的状态转换，而不是只集中在某个动作是否还记得。

核设施运行与维修工程：没有已知故障，不等于备用系统已经被证明可用

核设施运行与维修工程提供了另一种结构。核电站中的许多安全重要系统并不持续运行，而是在正常状态下长期待命，只有在特定需求发生时才启动。对于这种系统，“一直没有出事”不是充分证据，因为潜伏失效可能在待命期间始终不可见，直到监督试验或真实需求到来才暴露[5]。

这一区分与 AI 时代的人类角色高度相似。当 AI 承担绝大多数正常工作后，人越来越像一个备用通道。一个组织可以连续一年没有发生必须人工接管的严重 AI 故障，但这一事实主要证明主通道在过去任务分布中的表现，并不能直接证明人工备用路径在明天仍然能够启动。换言之，零事故记录对备用能力状态的信息量可能非常低。

核工程因此发展了监督试验、在役检查和定期测试制度，用于在真实需求以前识别潜伏失效。IAEA 相关安全指南把维护、测试、监督和检查的目标明确指向可靠性与可用性的持续证明，并强调测试结果应当回写到后续纠正措施和维护安排中[5][6]。重要的是，这种证明不是一次性验收，而是持续状态管理。

但核工程同时拒绝“测试越多越安全”的简单逻辑。测试本身可能引入设备扰动、瞬态风险、人员负担和维护成本，因此试验间隔需要在潜伏失效暴露风险与测试诱发风险之间优化[5]。这对 AI 时代的能力维护尤为关键：如果为了证明人仍然会做，而强迫专业人员频繁脱离 AI 完整重做全部任务，测试成本本身就会成为系统损失。

核设施工程由此给出第二条机制：长期待命的关键能力不能由主通道的正常绩效来认证；它需要独立的证明事件，但证明事件必须追求信息效率，而不能无限增加频率和强度。

运动训练学：最低维持剂量存在，但剂量不是一个统一数字

运动训练学处理的是另一种长期维持问题：已经通过训练获得的适应，在降低训练量以后能否保留？这使“最低维持剂量”成为一个比“是否继续训练”更精确的问题。Bickel、Cross 与 Bamman 让参与者在 16 周高频抗阻训练后进入 32 周去训练或降剂量维持阶段，研究不同剂量对年轻和年长成人训练适应的保持[7]。

这一研究真正重要的地方并不是得出某个可以普遍套用的“每周练几次”，而是说明不同适应分量的保持需求并不一致。力量表现、肌纤维肥大、年龄差异等指标并不会在同一剂量下同步变化。也就是说，同一个人可以在某个显性指标上仍然表现很好，而另一项结构性适应已经开始下降。

更早的技能保持与再学习研究、降低训练频率后的有氧能力维持研究同样说明，保持既有能力所需的剂量通常远低于形成该能力所需的训练量，但保持效果受到能力类

型、个体差异、时间间隔和训练刺激特征影响[8][9]。因此，“维持”不是简单重复原训练方案，而是寻找仍足以保持目标适应的最小刺激。

运动训练学由此给出第三条机制：低剂量维护在原则上可行，但最低剂量必须针对具体适应分量来定义。一个保持良好的指标不能被用来替代整个能力系统。

跨学科互否方法：三条路径为什么不能被简单拼接

三条路径的结构

本题的答案形状是一条“从长期 AI 代行走到人工仍可接管”的路。这类问题的关键不是列出三个解释维度，而是让三条不同的实现路径争夺同一个目标：究竟通过什么过程，才能使长期待命的人类能力仍然在必要时可用。

航空路径把目标放在“重新进入控制”：只要人仍能保持情境参与和接管表现，自动化就可以高度使用。核工程路径把目标放在“证明备用可用”：只要待命系统通过适当的监督和试验，就不必持续运行。运动训练路径把目标放在“最低刺激维持适应”：只要低剂量足以保留目标能力，就没有必要重复原训练量。三者看似互补，但真正执行时会给出不同甚至互相否定的处方。

三家共有但未明说的前提

三条路径争论的表面问题是“人需要参与多少、多久测一次、最低练多少”。但它们共同站在一个更深的前提上：能力似乎可以被当成一个已经存在的库存，只要在某个时点测到它，或者某个指标仍然保持，就可以认为能力仍然可用。

航空研究内部的材料首先推翻了这一前提，因为基础手工技能保持并不保证复杂认知接管保持；核工程内部材料也推翻它，因为一个没有已知故障的待命系统仍可能存在潜伏失效；运动训练内部材料同样推翻它，因为相同剂量对不同训练适应的维持效果并不一致。三门学科都在自己的证据中承认：表面“还在”的某些部分，不能代表整个系统仍具有被需求时的可用性。

由此，真正需要维护的对象不再是“能力库存”，而是一个跨状态的过程：从长期待命状态重新跃迁到有效控制状态。能力的失效也因此出现一种新的形态——零件可能还在，连接零件的路径已经断了。

敌意拓宽：本章不能占用的既有位置

为避免把已有思想换名，本章必须主动承认最危险的占位者。自动化脱环和情境意识研究已经指出长期自动化可能削弱人工接管[1][2][13]；Parasuraman 与 Riley 关于

自动化使用、误用、不用与滥用的研究已经建立了自动化依赖与人类监督的经典框架[12]；劳动过程与 deskilling 传统早已讨论技术和组织如何改变技能内容[10]；核安全领域拥有成熟的监督试验、证明试验和风险化试验间隔思想[5][6]；运动训练学已经直接使用最低维持剂量和去训练框架[7]-[9]。

因此，本章不能以“AI 用多了人会退化”“人需要定期练习”“备用能力应该测试”“最低训练量可以低于原训练量”等判断声称创新。这些位置已经被充分占用。本章只有在提出一个同时不能被 out-of-the-loop、proof testing 或 minimum maintenance dose 单独吸收的判别结构时，才可能构成二阶产物。

本章最终保留的空位是：把长期 AI 代行后的能力维护对象定义为“待命—接管状态转换”，把维护剂量定义为不同接管门的分量向量，把证明事件定义为扰动后的完整复控，并要求证明结果反向修改下一轮 AI 授权与人工维持剂量。任意一个既有领域都能解释其中一部分，但不能单独推出这一完整闭环。

二阶涌现：能力维护的最小对象是“接管跃迁路径”

本章将接管跃迁路径定义为：人在长期由 AI 主导的任务中，从未独立维持完整任务状态的待命位置，重新进入能够独立发现异常、重建情境、作出判断、实施干预并确认恢复的控制状态所必须经历的连续过程。

这一对象转换能够解释一种过去容易被误判的现象：一个人可能“还会做”，却“接不上”。例如，程序员仍然能够写出正确代码，却在 AI 编程助手不可用且生产系统出现陌生故障时，无法从日志、依赖关系和系统反馈中迅速建立故障模型；研究者仍然会写论文，却在 AI 给出一段非常流畅但承重引用错误的综述时，无法自己发现问题；管理者仍然记得决策框架，却在多个 AI 给出互相冲突的建议时，只能继续询问更多 AI，而没有独立终止冲突的证据路径。

这种失效与完全遗忘不同。完全遗忘意味着某个知识或动作本身消失；接管路径失效意味着若干局部技能仍然存在，但异常发现、情境定位、证据切换、控制取得或恢复验证中的某个连接长期没有被使用，导致真实异常发生时无法把零件串成一个有效闭环。

因此，AI 时代能力维护不应追求“保留全部人工操作”，而应追求“保留足够完整的接管跃迁”。这允许日常流程极大程度自动化，也允许人在绝大多数时间不亲自执行低价值环节；但只要人仍负有关键时刻的最终控制责任，接管路径就必须以某种方式保持可验证。

最低接管维持剂量（MTMD）：从单一频率到能力向量

五个接管门

为了使“最低剂量”可操作，首先需要把接管能力拆成可以分别失效的最小门。本章提出五个基本门：异常发现（Detection, D）、情境重建（Reconstruction, R）、独立裁决（Judgment, J）、干预执行（Intervention, I）和恢复验证（Verification, V）。这五门并非声称穷尽所有职业能力，而是作为一个跨任务的最小骨架。

异常发现门要求主体在没有 AI 明确提示“现在出错”的情况下，能够察觉结果、证据或现实反馈中存在足够重要的差异；情境重建门要求主体能够回到原始信息，识别系统当前状态、可信信息、已发生步骤和变化约束；独立裁决门要求主体能够在 AI、数据、规则或多个模型发生冲突时，依据不完全同源的理由作出判断；干预执行门要求判断能够转化为真实改变任务状态的行动；恢复验证门要求主体能够确认干预是否奏效，并识别新的副作用、残余风险和后续学习。

剂量的向量表达

如果不同接管门的衰减速度不同，那么“每月独立做一次”这种标量规定就会失去精度。本章因此把最低接管维持剂量定义为一个向量，而不是一个单一数字。

$$MTMD = \langle m_D, m_R, m_J, m_I, m_V \rangle \quad (1)$$

其中，每个 m_k 不是单一时间频率，而至少包含频率、强度、独立性、情境变异度和现实保真度五个维度。一个人即使每周都“练”，若练习总是提前告诉他哪里有错、所有事实都由同一个 AI 提供、每次只复现标准答案，那么真正需要维持的异常发现和独立裁决可能仍然接近零调用。

$$m_k = (f_k, s_k, i_k, v_k, r_k) \quad (2)$$

式中，f 表示调用频率，s 表示挑战强度，i 表示与日常 AI 主通道的独立程度，v 表示情境变异度，r 表示现实保真度。由此，“十分钟训练”与“十分钟真实调用”不再被视为同一个剂量。

这一定义允许维护量远低于生产量。一个研究者可以让 AI 承担大量搜索、格式整理和初稿生成，却只需在少量关键节点直接阅读原始文献、先于 AI 写下自己的预期、随机核查承重引用、独立处理一次证据冲突；一个程序员可以大量使用代码生成，却定期独立读日志、定位根因、决定回滚或修复；一个管理者可以让 AI 完成方案比较，却保留识别关键假设、区分事实判断与价值排序、终止多 AI 分歧的真实练习。

因此，最低维持剂量的单位不应首先是分钟，而应是“关键接管门被真实激活的次数与质量”。这种单位比传统培训学时更接近需要维护的机制。

接管保真环（TFL）：从维护到证明再到回写

六阶段循环

阶段一：能力拆门

对每项高度 AI 化的任务，不再列出所有工作步骤，而只列出 AI 失效后人必须亲自完成、否则系统无法恢复的接管门。

阶段二：微调用

在正常 AI 工作流中嵌入极少量真实人工调用，避免某个关键门长期处于零使用。例如先写预期再看 AI 结论、随机回查原始材料、由人先判断异常再打开 AI 建议。

阶段三：待命暴露累计

记录每个接管门距上一次真实独立调用的时间、AI 代理比例、模型变化、任务环境变化、后果等级以及最近一次证明中的最弱项。

阶段四：风险触发证明

以固定最大静默期作为后备，同时允许 AI 升级、环境变化、近失误、多模型冲突、任务后果提高或微调用失败等事件提前触发证明。

阶段五：扰动接管

通过 AI 断援、AI 与现实冲突、多代理分歧或情境越界等预设扰动，要求主体在有限时间内独立完成 D-R-J-I-V。

阶段六：恢复与回写

证明不在“答对”时结束，而以系统恢复、后果验证和维护方案更新为终点。最弱门决定下一轮增量训练，严重失败可暂时收紧 AI 授权或增加双人复核。

为什么必须是闭环而不是考试

传统考试通常在题目中明确告诉受试者“这里有一个问题”，于是异常发现已经被命题者替受试者完成。真实 AI 错误却可能以流畅、自信、引用齐全和结构完整的形式出现。若测试永远从“请处理下面这个错误”开始，就只能证明人在错误被指出以后会不会解决，而不能证明他还能自己首先发现问题。

因此，至少一部分接管证明必须采用潜伏扰动：参与者知道本轮 AI 可能可靠也可能不可靠，却不知道错误是否一定存在、出现在哪里、究竟是 AI 错、原始数据错还是自己的旧知识错。只有这样，异常发现门才被真实调用。

同样，测试也不能在人给出标准答案时停止。若一个人知道正确诊断却不会实施安全处置，或实施以后不验证系统是否恢复，平均知识分再高也无法承担真实接管。接管证明因此必须覆盖从发现到恢复的完整闭环。

接管证明的四类扰动与通过规则

四类基本扰动

串联门而非平均分

对于高后果任务，接管通过条件不应以五项平均分为主。原因在于真实接管通常具有串联系统特征：任一承重环节完全失效，都可能使整体失败。设五项能力标准化得分为 D、R、J、I、V∈[0,1]，各自最低阈值为 θ_k，则可以使用门控通过条件：

$$G = \Pi(k\in\{D,R,J,I,V\})\ 1(x_k \geqslant \theta_k) \quad (3)$$

此外必须设置时间约束。许多能力不是“最后能不能做出来”，而是“能不能在还来得及的时候做出来”。因此对时间敏感任务还需满足：

$$T_{takeover} \leqslant T_{critical} \quad (4)$$

一个人在三小时后终于识别出正确异常，对于允许数天处理的研究任务可能完全合格；对于必须几十秒内接管的实时控制任务则没有意义。通过规则必须由任务的真实关键时间窗定义，而不能套用统一标准。

采用串联门还有另一个管理意义：它可以防止“强项掩盖弱项”。一个人可能知识得分、分析能力和操作技能都很高，但异常发现持续失败；如果用平均分，他会被判定为优秀，真实系统却会在错误从未被他察觉时失去人工冗余。

证明何时触发：固定日历只是后备机制

如果能力衰减速度受调用间隔、AI 变化、任务变化和后果等级共同影响，那么统一的“每三个月测试一次”只能作为最粗的后备。本章提出一个用于调度而非直接评价个人的待命暴露指数。它目前属于待验证的工程化框架，而不是已被实证确认的心理学定律。

$$E_k = (\Delta t_k / T_{0k}) \cdot (1 + \alpha A + \beta C + \gamma H) \quad (5)$$

其中，Δt_k 表示第 k 项能力距离上一次真实独立调用的时间，T_{0k} 是依据历史表现暂定的基准维持间隔，A 表示 AI 系统变化程度，C 表示任务情境变化程度，H 表示错误后果等级。当 E_k 超过预设阈值时，触发对应能力门的证明试验。

除此之外，若出现 AI 未解释异常、多模型对承重结论持续分歧、业务环境或法规发生重大变化、一次人工微校验失败、真实接管明显变慢、任务即将进入更高后果等级等事件，可以直接触发测试，而不必等待固定日期。

这使能力维护从“按时间打卡”转向“按风险暴露管理”。固定时间仍然有价值，因为它可以防止长期没有事件时测试永久消失；但真正决定测试强度的应是长期未调用、变化和后果的组合。

最低成本压力测试：为什么“一个能力一个最低剂量”必须被否定

候选命题与可证伪条件

本章在候选阶段最容易得到的一个简单判断是：既然运动训练存在最低维持剂量，那么每项 AI 时代能力也可以找到一个统一最低频率，例如“每月独立做一次”或“每季度完成一次无 AI 任务”。如果这个命题成立，组织实践将非常简单。

但这一命题的可证伪条件也非常明确：若同一个低剂量能维持一种关键能力分量，却不能维持另一种分量，那么“能力=单一剂量”模型就必须放弃。

公开数据真跑与不利结果

Bickel 等的公开研究为这一条提供了低成本压力测试材料[7]。参与者在完成 16 周每周 3 次抗阻训练后，进入 32 周的去训练或显著降剂量维持阶段。研究报告显示，训练后形成的不同适应并未在相同维持剂量下表现出完全一致的保持轨迹；年龄也改变了维持某些结构性适应所需要的训练负荷。

如果“最低剂量”是整个能力的单一属性，那么同一个人的不同适应分量不应在相同维持条件下明显分离。但真实数据恰恰出现分离。于是本章原先最简单的标量模型受到不利结果，必须修订。

航空研究提供了第二次跨领域重复。Casner 等观察到人工操纵、仪表扫描和某些程序化技能相对保持，而导航、程序推进和异常识别等认知性任务更容易出现问题[3]。这里同样出现“某些部分仍然良好，却不能代表整个接管能力仍然良好”的结构。

因此，本章放弃“一个能力—一个最低维持剂量”的命题，改为“一个接管路径—多个不同衰减速度的能力门—一个由最弱门约束的剂量向量”。这次修订不是为了让理

论更复杂，而是因为简单模型无法同时容纳运动训练和航空自动化内部已经存在的不利材料。

正常运行长期没有事故究竟能证明什么

现在可以回答本章最容易被直觉误导的问题：如果 AI 在绝大多数正常情境中长期稳定，系统没有出现严重事故，这是否证明人机协作越来越成熟？答案是，它可以证明某些东西，但不能证明全部。

它可以说明在已经发生的任务与环境分布中，主运行通道没有暴露严重失败；可以说明 AI、流程和人类监督组合在过去获得了较好的总体结果；但它不能单独说明长期待命的人类备用路径仍处于可调用状态。原因并不是我们已经知道人一定退化，而是这项观察对备用通道缺乏识别力。

因此，正确表达不是“零事故说明人已经退化”，也不是“零事故说明人仍然可靠”，而是“零事故不能区分这两种状态”。在安全工程语言里，这意味着备用能力存在可能长期不可见的潜伏失效。

这一逻辑还产生一个悖论：AI 越可靠，真实故障越少，人获得自然接管练习的机会也越少。AI 可靠性提高一方面降低了主通道风险，另一方面增加了备用能力长期未被需求的时间。最终是否应该延长或缩短人工测试间隔，不能由“AI 更强了”直接决定，而必须同时考虑 AI 失效概率、失效后果、人工能力衰减速度、自然练习频率、环境变化速度与测试成本。

为什么不能等到 AI 真正失败再验证人会不会接管

“真实故障发生以后再让人接管”看似最节省成本，因为不需要额外演练。但它实际上把训练、测试和正式任务压缩到同一个时刻。如果人通过，系统没有问题；如果人失败，代价已经发生。

对低后果、可逆任务，这种策略完全可能合理。一封普通邮件写坏了可以重写，一个低风险脚本失败可以回滚，一次非关键检索判断错误可以继续搜索。因此，本章并不主张所有 AI 使用都需要正式接管证明。

但随着错误后果提高、传播范围扩大、可逆性降低、发现延迟增加，第一次验证人工备用能力的时点就必须前移。测试失败可以训练以后再来；真实接管失败则可能没有第二轮。

所以 AI 治理不仅需要对模型风险分级，也需要对人类备用能力的证明要求分级。低风险任务可以依赖真实工作中的自然学习，高风险任务则需要在真实故障以前用模拟器、沙箱、历史案例回放、数字孪生或安全隔离环境完成证明。

为什么也不能把“完整脱 AI 重做”当成默认答案

另一个极端是规定每周一天禁止 AI、每月完整人工重做、所有关键岗位持续手工完成固定比例任务，或所有 AI 结果都必须人工完整复算。此类规定看似谨慎，却可能把能力维护变成新的低价值劳动。

核工程的监督试验逻辑提醒我们，测试本身也有成本和风险。人的能力维护同样会占用生产时间、增加认知疲劳、制造抵触，并可能为了“给人练手”而故意放弃原本更安全、更准确的自动化工具。若没有证据表明完整重演比更小的定向刺激带来更高的接管可靠度，那么高成本全量练习本身就需要被证明。

本章因此主张：尽可能彻底地外包低价值生产劳动，同时尽可能精确地保留高价值接管刺激。这两个目标并不冲突。真正成熟的维护制度甚至可能让人比过去更少手工操作，却因为训练更集中地击中异常发现、情境重建和独立裁决，而获得更高的真实接管可靠度。

应用场景：三类知识工作的维护设计

研究者：保留证据链而不是保留全文手工生产

在研究工作中，AI 可以长期承担候选文献搜索、摘要、参考文献格式、初稿结构、语言润色和基础统计代码。为了维护研究能力，没有必要要求研究者周期性完全手工重做一篇综述。真正值得保留的最低调用包括：直接阅读少量承重原文；在看 AI 结论前先写下关键预测；随机核查引用与原文是否一致；处理一次真实或模拟的证据冲突；判断一项统计结果究竟支持什么结论以及什么新证据会推翻它。

对应的接管证明可以是一份总体高质量、但其中一条承重引用被错误解释的 AI 综述。测试不只看最后能否找到错句，而看研究者是否主动发现异常、能否回到原始论文、能否说明错误影响哪一个推理节点、能否重新形成结论，并改变下一轮检索与引用核验规则。

软件工程师：维护故障闭环而不是维护键盘输入量

在软件工程中，AI 可以大量承担模板代码、接口封装、测试样例初稿、文档和常见重构。用“手写代码字符数”作为能力维护指标，会把维护对象放错位置。更关键的

是从日志发现异常、理解依赖关系、形成故障假设、阅读陌生代码、决定修复还是回滚、执行安全操作并确认服务恢复。

因此，一个二十分钟的故障注入可能比数小时无 AI 编程更有信息量：AI 助手临时不可用，系统出现一个非显式错误，工程师必须从监控、日志、提交记录和程序行为重建状态。若他能够写大量代码却无法完成这种接管，说明生产技能仍在而备用控制能力已经发生断裂。

管理决策者：维护终止冲突的证据路径

在管理决策中，AI 可以承担资料整理、风险表、方案比较、概率估计、会议摘要和情景推演。管理者不需要重做所有计算，却必须保留识别关键假设、知道什么变量改变会推翻结论、区分事实判断与价值排序、在多个 AI 结论冲突时确定异源证据并承担最终干预后果的能力。

一个有效的证明场景可以让三个 AI 都支持方案 A，同时在原始材料里放入一个已经变化的关键约束。真正需要观察的是管理者是否会回到现实条件，发现这个约束使原问题失效，而不是看他能否复述“决策的五个原则”。

AI 本身如何参与能力维护，以及它不能承担什么

AI 完全可以成为能力维护基础设施的一部分。它可以生成大量情境、随机改变参数、模拟用户与系统、追踪各接管门的历史表现、识别长期未调用的能力，并根据过去失败自动提高或降低训练难度。若不充分利用 AI，能力维护本身很容易变得过于昂贵。

但存在一条不可取消的边界：负责日常工作的同一个失效源，不能同时垄断接管测试的出题、事实来源和最终评分。若同一个 AI 生成错误答案、告诉人哪里错、再评价人的纠正是否正确，测试很可能只能证明人能够重新服从同一个模型。

因此，高价值接管证明至少需要一个相对独立的外部锚，例如真实原始数据、已人工核验的历史案例、确定性程序测试、独立专家、不同来源模型、真实系统反馈或预先冻结的参考答案。AI 可以承担训练成本，但不能垄断真值来源。

维护失败后的处置：从“惩罚个体”转向“校正剂量与授权”

如果一次接管证明失败，最差的管理方式是立即把结果解释为个人能力标签。这样会迅速诱发应试优化，员工开始学习如何通过测试，而不是如何维持真实能力。

本章提出的第一解释应当是：当前维护制度对某一个接管门给出的剂量可能不足。异常发现失败，则增加无提示异常案例和原始数据抽检；情境重建变慢，则增加直接接触原始记录；独立判断失败，则增加冲突证据裁决和先写判断；执行错误，则增加沙箱操作；恢复验证缺失，则强制训练在“结果出现以后继续验证”。

同时，严重失败可以暂时收紧 AI 授权。例如在重新训练期间增加第二人复核、提高关键任务中的人工参与比例、缩短下一次证明间隔，或暂停某类高后果自动批准。能力恢复以后再逐步放宽。于是测试结果不只是“考试成绩”，而成为下一轮人机任务分配的反馈控制信号。

与近邻概念的判决性划界

其中，与自动化脱环的距离最近。若本章只说“长期用 AI 会让人失去情境意识，因此要多练”，那么完全可以被既有研究吸收。本章必须额外承担一个更难的预测：在维护总时间相同或更低的情况下，按接管门分配的微调用加低频完整证明，应能比简单增加人工参与比例更有效预测并维持真实接管表现。

与最低训练剂量的边界同样清楚。若把 MTMD 解释成“运动训练最低剂量在认知工作的类比”，创新仍然不足。真正的新增结构是，剂量不仅维持能力分量，还与待命暴露、扰动证明、AI 授权和失败后的再分配闭合为一条循环；其目标不是保持某个成绩，而是保持从 AI 主控状态重新进入人工控制的能力。

在本书内部，与本章距离最近的是第三章。两章都给出了一条循环：本章是"拆门—微调用—待命暴露累计—风险触发证明—扰动接管—恢复回写"，第三章是"首判—协作—差分—盲段—恢复"。相似程度足以让读者怀疑它们是同一条循环换了措辞。

分界在目的，而且这条分界是可判定的。本章的循环是为了让能力保持可用，第三章的循环是为了让能力保持可见。前者若成功，人的真实接管可靠度提高；后者若成功，系统对人的能力状态的判断精度提高，而人的能力本身可能一点没变。二者可以分离：一个组织完全可以做到证据永远新鲜（第三章的读数良好），而所有刷新事件都过于轻量，不足以维持接管门的剂量（本章的读数不合格）。合并条件因此也很清楚：若按第三章的校准回路执行一年，真实接管可靠度的提升与按本章的剂量向量执行没有差别，则维持与校准不是两件事，本章应并入第三章。

与第一章的关系则是处方与判定的关系：第一章给出能力归属在断面上的判定条件，本章不重新定义那个条件，只回答如何以最低成本让它长期保持可满足。

可证伪命题与未来实证设计

六条判决性预测

预测一：定向微剂量可以在更低成本下达到不差于完整任务重演的接管可靠度。

若相同时间预算下，完整无 AI 重做显著优于分门微调用+低频证明，则本章的压缩维护主张失败。

预测二：普通知识与操作平均分会系统性高估真实接管可靠度。

若传统考试已经能够稳定预测无提示异常发现、情境重建、干预和恢复，则没有必要采用串联门。

预测三：不同接管门在长期 AI 代行后会出现非同步衰减。

若 D、R、J、I、V 总是以近似相同比例共同变化，则 MTMD 向量化缺乏必要性。

预测四：最大静默期+事件触发应优于纯固定日历。

若模型升级、环境变化等事件并不会提高后续接管失败风险，则事件触发部分应被删除。

预测五：同一 AI 同时承担工作、出题与评分会高估人类独立能力。

若与独立真值源相比完全没有差异，则共同模式验证失效的担忧被削弱。

预测六：一次通过能力认证不能保证未来维护方案有效。

若初始接管成绩对长期高 AI 代理后的接管表现具有近乎完全稳定的预测力，则持续维护模型价值显著下降。

纵向实验设计

未来最重要的研究不是继续增加理论概念，而是建立纵向数据。可以选择已经熟练掌握某项任务的专业人员，将其随机分为四组：A 组持续完整人工练习；B 组正常使用 AI 但按固定周期做完整无 AI 测试；C 组正常使用 AI，只保留分量化最低真实调用并采用风险触发的接管证明；D 组自由使用 AI，不设置专门维护。

观察周期可设为半年至一年。平时记录 AI 调用比例、人工独立处理量、原始信息接触量、异常发生次数、模型版本变化和任务环境变化。主要终点不应是 AI 在线时的生产率，而应是隐藏扰动条件下的异常发现率、接管启动时间、情境重建正确率、独立裁决质量、干预成功率和恢复验证完整率。

进一步可以使用生存分析或可靠性模型估计：从上一次真实独立调用开始，某一道接管门失去合格状态的风险怎样随时间变化；不同 AI 代理比例、任务风险和环境变化

速度怎样改变这一曲线。只有到这一步，“最低接管维持剂量”才会从理论构造转化为不同职业、不同任务可经验校准的参数。

伦理与治理边界

把核设施证明试验的结构迁移到人机协作，必须同时拒绝把人简单等同于机器。人的能力受学习、动机、疲劳、迁移、团队关系和组织文化影响，不能通过设备可靠性模型直接外推。本章使用“待命系统”“潜伏失效”“证明试验”等术语，只是在结构层面描述“长期不被真实需求的备用路径怎样获得证据”，并非宣称人的心理机制与设备失效机制相同。

第二，高风险领域不能为了测试人工接管而制造真实危险。优先使用模拟器、沙箱、历史案例回放、数字孪生和安全隔离环境；真实生产系统中的人工接管演练只能在风险可控条件下实施。

第三，接管保真度首先应被用来定位维护制度和任务分配问题，而不是形成不透明的人事排名。若组织将该读数用于岗位、薪酬或晋升决策，必须公开编码规则、最低阈值、数据来源和复核渠道，并允许人员对错误标记和不合理测试条件提出复核。

第四，不能为了让人工备用能力“看起来更强”，而人为阻止本来更安全、更可靠的自动化。在安全关键系统中，人工训练应尽量移入模拟环境，而不是通过真实降低自动化水平来制造练手机会。

讨论：AI 时代真正需要保存的不是人工劳动份额，而是异源冗余

随着 AI 能力持续提高，一个更深的组织结构变化正在出现：过去机器经常是人的备用工具，未来很多任务中，人可能逐渐成为 AI 主通道的备用路径。传统组织却仍然用“岗位上还有一个人”“最终按钮由人点击”“有人负责审核”来推定人工冗余存在。

这种推定并不可靠。真正的冗余不是组织图上还有一个人，而是系统中仍然存在一条独立、可启动、可运行、可恢复的控制路径。如果一个专业人员连续多年只阅读 AI 摘要，从不接触原始事实；只批准 AI 建议，从不形成独立预期；只在 AI 提示以后处理异常；只会用另一个 AI 检查第一个 AI，那么人虽然仍然在流程中，备用路径却可能已经与主路径高度同源。

这会形成一种新的共同模式失效：AI 因为某种系统性偏差判断错误，人因为长期依赖同类信息和同类模型，也以同样方式判断错误。此时所谓“人工监督”并没有提供

真正的第二条路径。能力维护的意义于是超越个人成长，成为人机系统保护异源冗余的一部分。

这也改变了“human-in-the-loop”的理想。未来成熟系统未必要求人在每一步持续参与。大量低价值操作完全可以交给 AI。更有价值的目标是：人不必始终在场，但必须始终有路回来。前者强调人工劳动占比，后者强调异常发生后人工重新取得控制的可恢复性。

因此，一个成熟的人机系统完全可能让 AI 承担 95%的常规操作，人只承担 5%的关键监督、抽检和维护事件；但一旦 AI 进入不可靠状态，这 5%的持续刺激足以使人重新取得 100%的必要控制。真正需要设计的不是人与 AI 永远各做一半，而是正常时期的高效率与异常时期的可接管同时成立。

结论

本章从一个表面简单的问题出发：当一个人已经掌握某项能力，但长期由生成式 AI 承担检索、分析、计算、写作、判断甚至执行以后，怎样才能既获得 AI 效率，又不让关键能力在漫长的低调用中失活?

航空人因工程表明，自动化可以提高正常绩效，却使接管越来越像一次从监督状态进入控制状态的冷启动；而且基础操作技能的保持不能代表认知性故障识别与情境重建同样保持。核设施运行与维修工程表明，长期待命而没有已知故障并不能证明备用系统可用，潜伏失效需要独立的证明事件，但证明频率又必须权衡测试成本。运动训练学表明，已经形成的能力可以用远低于形成期的剂量维持，但不同适应分量的最低剂量并不相同。

三者碰撞后，本章将能力维护的对象从“技能库存”改写为“待命—接管状态转换”，提出最低接管维持剂量 MTMD 与接管保真环 TFL。前者要求对异常发现、情境重建、独立裁决、干预执行和恢复验证分别配置真实刺激；后者要求把维护组织成拆门、微调用、待命暴露、风险触发、扰动接管和恢复回写的循环。

由此，AI 时代不需要人为保存大量低价值的手工劳动。检索第一遍、写第一稿、格式化、常规计算和机械执行都可以被深度外包。真正需要保护的是少数能够让人从 AI 主通道重新回到现实、形成独立判断并改变系统的关键能力门。

“系统长期没有出事”不能被当作人工备用能力已经被证明的证据；“还能复述标准答案”也不能等价于能够接管。真正有效的校验必须在 AI 不可用、AI 与现实冲突、多 AI 分歧或任务越界之后观察：人能否自己发现异常、重建情境、形成判断、实施干预并验证恢复。

因此，AI 时代能力维护最重要的目标不是让人一直和机器争夺操作权，而是做到：平时机器可以充分工作，关键时刻人仍然能够回来；而“能够回来”不能靠岗位名称、历史资格、正常绩效或主观信念来证明，只能通过低成本、真实、可裁决的接管事件持续得到证据。

附录 A　核心操作定义与编码建议

附录 B　研究边界与版本说明

本章属于理论与方法研究。航空、核工程与运动训练材料用于提取可迁移的结构，不主张飞行员、核设施设备、运动适应与知识工作者在心理或生理机制上等同。

MTMD、TFL、五门接管结构及待命暴露指数目前均为工作性理论构造，不是已经获得经验验证的行业标准。任何具体职业的测试间隔、最低剂量、通过阈值与关键时间窗都需要纵向数据校准。

本章对新颖性的主张限于当前检索与既有材料范围。自动化脱环、情境意识、deskilling、监督试验、证明试验、去训练与 minimum maintenance dose 均构成实质近邻。本章的工作性新意集中于：维护对象的状态转换化、维持剂量的分门向量化、校验的扰动复控化，以及测试结果与 AI 授权/下一轮剂量之间的闭环回写。若未来研究表明传统固定周期完整重演能以相同或更低成本稳定获得更高真实接管成功率，或普通知识技能测试已经能够充分预测扰动条件下的完整接管，则本章机制应被修订或放弃。

本章附表

学科路径	它认为应当优先维护什么	天然倾向的处方	对另外两家的挑战
航空人因工程	情境意识与接管表现	让人保持足够的持续参与和近期经验	持续参与可能成本过高，且不等于备用系统被真正证明
核设施运行与维修工程	待命系统的可用性	通过监督试验发现潜伏失效	试验能证明可用，但并不直接告诉每个能力分量如何最低成本维持
运动训练学	已经形成的适应	用低于形成期的最小刺激维持目标适应	只谈剂量仍可能漏掉“何时必须从待命跃迁到控制”的系统问题

接管门	核心问题	典型退化表现	最低真实调用示例
D 异常发现	我能否先于 AI 提示发现“不对劲”？	只会处理已被标红的问题	无提示错误案例、随机原始数据抽检

R 情境重建	我能否重新知道系统现在发生了什么?	看得懂结论，但离不开 AI 摘要	直接阅读原始记录、日志、论文、数据
J 独立裁决	AI 冲突时我凭什么结束争议?	继续问更多 AI，无法独立收敛	先写判断、异源证据比对、反例裁决
I 干预执行	我的判断能否改变真实任务状态?	理论知道，但无法安全操作	沙箱操作、回滚、人工处置演练
V 恢复验证	我怎么知道真的解决了?	答对即停止，不验证后果	复测、监控、闭环核验、复盘回写

扰动类型	**基本设置**	**主要测试对象**	**为什么不能被其他类型替代**
A 断援型	AI 暂时不可用	基础路径、原始信息获取、独立执行	最容易实施，但受试者从一开始就高度警觉
B 冲突型	AI 在线但给出与现实证据不一致的合理结论	异常发现、独立证据、纠错	最接近日常“系统看起来正常但其实错了”
C 多代理分歧型	多个高质量 AI 给出互不兼容建议	独立裁决、证据层级、终止规则	防止“继续问另一个 AI”无限递归
D 越界型	AI 执行旧规则无错误，但环境或约束已变化	问题重定义、适用边界、情境重建	测试能否发现“问题已经换了”而不是只找幻觉

近邻概念	**它已经解决的问题**	**本章不能被其吸收的部分**	**判决性对照**
Out-of-the-loop / 情境意识	自动化为何削弱监督、故障接管和状态理解	如何按能力门配置最低剂量，并让证明结果回写 AI 授权	若仅提高 SA 即可预测全部 D-R-J-I-V，则 TFL 多余
Proof / surveillance testing	怎样发现待命设备潜伏失效并优化试验间隔	把人的能力分量、真实调用和 AI 授权纳入同一维护循环	若设备式单一可用/不可用测试足以预测人的复杂接管，则 MTMD 多余
Minimum maintenance dose / detraining	降低训练量后怎样保留身体与技能适应	把剂量对象从单一表现改成多门接管状态转换	若一个标量剂量即可同时稳定所有接管门，则向量模型错误
Deskilling	技术和劳动组织怎样减少技能内容	研究已经获得的关键能力如何在长期待命后保持可调用	若只测长期技能内容下降即可预测即时接管，则本章无新增量
断代理复闭环判据	某一时点能力是否仍归属于主体	怎样使这一能力归属跨时间持续成立	若一次认证即可长期预测未来接管，无需维护模型

变量	**建议操作定义**	**不通过示例**
D 异常发现	在未被告知错误位置前，指出足以改变任务方向的异常并给出依	只在 AI 提示后才发现；把无关差异误判为关键异常

	据	
R 情境重建	从至少一个相对独立的原始来源重建当前状态、约束和已完成步骤	只能复述 AI 摘要；无法说明哪些信息仍可信
J 独立裁决	在冲突信息中给出可复核理由，并说明什么新证据会使自己改变判断	以 AI 多数票、模型名气或语言流畅度作为唯一理由
I 干预执行	在允许环境内实施能够改变任务状态的操作，并控制可预见风险	知道答案但不能安全执行；操作与判断不一致
V 恢复验证	用预先定义的结果或反馈确认恢复，并检查残余风险与副作用	输出正确答案即结束；不确认真实系统是否恢复

第三章

帮助不必撤离，能力必须可观测

能力证据债与校准回路　——教育学 × 分布式认知 × 安全科学/人因工程的跨学科对撞

承重命题：AI 时代能力替代最早出现的信号，不一定是人的无 AI 成绩已经下降，而可能是辅助状态下的“做得好”越来越不能说明这个人本人还能做到什么。只要某项能力仍被系统指定为异常识别、结构修正或失效接管时的备用能力，长期连续代行就会累积“能力证据债”。偿还这笔债不需要把完整任务重新交还给人，而需要在能力关键节点周期性恢复人的首判、差分、盲段与恢复回合，使关键能力持续可观测、可更新、可接管。

摘　要

当生成式人工智能能够持续提供答案、示范、修改、推理、反馈乃至直接执行任务时，传统的“支架渐退”出现了一个结构性难题：如果帮助必须最终撤除，人机系统将被迫重复大量低价值劳动；如果帮助可以永久保留，人的关键能力又可能在长期不被调用的状态下变得不可确认，并在真正需要接管时才暴露缺口。本章以教育学中的认知学徒制、支架与渐退，分布式认知中的认知外置与人—工具系统，以及安全科学/人因工程中的自动化、情境意识与接管恢复为三条相互冲突的路径。三者分别把“帮助成功”结算在独立完成、系统绩效与安全恢复三个不同终点，但共同假定一次帮助可以由某个单一终点完成结算。本章提出，生成式 AI 任务是循环而非终局的，同一环节会在不同轮次中在手段、终点和备用能力之间换位。由此提出“能力证据债”：当 AI 持续代行某一能力关键节点时，即使人的真实技能尚未下降，系统也会因为长期没有新的、AI 暴露前的人类独立表现证据，而逐渐失去判断该能力当前是否可用的依据。本章进一步提出“能力校准回路”，由首判、协作、差分、盲段、恢复五个环节构成，并以“证据龄比”作为局部触发读数。基于 Bastani 等人公开的高中数学随机实验数据，本章进行探索性二次分析：排除原研究主分析同样排除的荣誉班学生后，825 名学生的学生层面数据表明，练习表现与随后无 AI 考试表现的相关在对照组为 r=0.756，而通用 GPT 组为 r=0.435、教学护栏 GPT 组为 r=0.486；按班级聚类自助法得到的差异区间仍明显分离。该结果并不证明 AI 导致技能退化，却支持一个更弱而重要的命题：AI 可以提升辅助态绩效，同时降低辅助态绩效作为“本人会不会”的能力信号保真度。本章据此主张，AI 时代真正需要渐退的不是 AI 本身，而是未经验证、连续不断的代理代行；真正必须周期性回到人手中的，也不是全部任务，而是那些能够暴露问题模型、边界判断、异常识别与恢复能力的关键回合。

关键词：生成式人工智能；能力证据债；能力可观测性；能力校准回路；认知外置；自动化接管

English Title and Abstract

Assistance May Remain, but Capability Must Stay Observable: Capability Evidence Debt and Calibration Loops in Generative-AI-Assisted Learning and Work

Generative AI challenges the traditional assumption that external assistance should either fade until a person can perform independently or remain as a permanent component of an extended cognitive system. Educational scaffolding privileges independent mastery; distributed cognition legitimizes persistent cognitive offloading; human-factors and safety research permits automation while requiring humans to recover when automation fails. This paper argues that these approaches conflict because they settle assistance at different endpoints, yet share a deeper assumption that a single endpoint can close the assistance problem. In recurrent AI-mediated work, however, the same function can alternate between means, endpoint, and fallback capability across cycles. We introduce Capability Evidence Debt (CED): uncertainty accumulated when AI continuously performs a capability-critical node and the system stops generating fresh, independently verifiable evidence of what the human can still initiate, diagnose, repair, or recover before seeing AI output. CED is distinct from skill degradation: capability may remain intact while its current state becomes unobservable. We therefore propose a Capability Calibration Loop consisting of human-first judgment, AI-assisted execution, discrepancy analysis, bounded AI-blind probes, and assistance re-entry. A reanalysis of public data from Bastani et al. shows that correlations between AI-assisted practice performance and later unaided exam performance are substantially weaker in both AI conditions than in the no-AI control, even when pedagogical guardrails largely remove the mean learning loss. The result does not establish causal deskilling; it supports the narrower claim that AI assistance can reduce the diagnostic fidelity of current performance as evidence of unaided capability. The proposed framework therefore shifts governance from “how much AI is used” to “which

capability-critical nodes remain observable, how old the last valid evidence is, and when bounded human-only probes should be triggered.”

Keywords: generative AI; capability evidence debt; capability observability; scaffolding; distributed cognition; automation; takeover recovery

问题从“帮多少”变成了“哪一拍由谁承担”

在传统教学中，“帮助”通常带有临时性。教师先示范，学习者模仿；学习者能够承担更多步骤以后，教师减少提示；最终，外部支架退到背景，任务由学习者独立完成。Wood、Bruner 与 Ross 对支架的经典表述、认知学徒制以及样例渐退研究，都在不同意义上假定：当前表现可以借助帮助被抬高，但真正的能力必须逐步转化为学习者能够自行组织的行动[1-4]。在这一图景中，帮助过早会制造失败，帮助过多会遮蔽学习，帮助撤得太晚则可能形成依赖，这就是后来所谓 assistance dilemma 的核心张力[3]。

生成式 AI 改变的不是“帮助存在”这一事实，而是帮助的成本、覆盖范围与连续性。过去，一个学生想得到完整解题过程，要等老师、翻答案或搜索教材；一个程序员想得到一段可运行代码，要自己检索、试错并整合多个来源。现在，AI 可以在任务的第一秒同时给出问题解释、结构划分、方案、代码、修改和理由。帮助不再只发生在卡住之后，而能够先于人的第一次尝试进入；它也不再只支撑某一步，而能覆盖从问题定义到最终执行的整条链条。

因此，把传统“支架渐退”直接搬到生成式 AI 时代会遇到反直觉结果。若规定能力成熟以后 AI 必须撤除，那么越熟练的人反而越不能使用效率更高的工具；大量已经没有必要由人重复执行的检索、格式化、机械计算与标准化表达会被重新内化为训练任务。可如果接受 AI 成为长期认知基础设施，又会出现另一种风险：人在正常状态下越来越少亲自观察原始信息、形成首判、处理异常和纠正偏差，直到某次 AI 失败、不可用或越界时才第一次真正暴露“备用的人”是否仍能工作。

所以，这道题不能再被压成“AI 用多一点还是少一点”。真正需要设计的是一个时间结构：AI 什么时候先进入，什么时候后进入，哪些节点永远不必退，哪些节点只需短暂退，何种信号意味着帮助要增加、维持、局部渐退或重新进入。问题的单位从“整项任务”改变为“任务中的一拍”。

研究路线：三条成熟路径为什么必须真正冲突

本章采用理论重构与公开数据探索性复核相结合的方法。理论部分不是将教育学、分布式认知与人因工程并排综述，而是分别提取三者在“帮助如何进入”这一问题上的强主张，并要求它们对同一人机任务给出不能同时成立的成功标准。教育学路径把成功放在独立能力形成；分布式认知把成功放在人—工具整体能否稳定完成任务；安全科学则把成功放在自动化失效以后人是否仍能恢复控制。

三条路径之所以值得放在一起，不是因为它们互补，而是因为每一条的优势恰好会成为另一条的风险。教育学若坚持把外部帮助撤到最低，可能牺牲成熟工具带来的系统效率；分布式认知若只看整体系统绩效，可能无法回答一个已长期不再独立工作的子系统在重新配置后还能做什么；安全科学若为保持接管能力而要求人频繁完整手动运行，则可能把训练变成高成本形式主义。只有让三条路径互相取消各自的默认终点，新的动态路径才有必要出现。

经验部分选择 Bastani 等人公开发布的高中数学随机实验作为最低成本压力测试[22]。该实验同时包含无 AI 对照、通用 GPT 帮助和带教学护栏的 GPT Tutor，并在 AI 辅助练习后安排无 AI 考试，非常适合检验“辅助状态下的表现还能否直接作为独立能力信号”这一较弱命题。本章不把二次分析当作“能力证据债”的因果证明，而把它用来回答一个更具体的问题：当 AI 显著提高眼前成绩时，眼前成绩与随后独立成绩的对应关系是否仍保持原有保真度。

第一条路径：教育学要求帮助最终转化为学习者自己的组织能力

认知学徒制强调，专家能力并不只是拥有答案，而是能够在真实任务中组织知觉、策略、监控和修正。示范使隐性的过程显露，指导与支架降低新手暂时无法承担的负荷，表达与反思迫使学习者把自己的过程说出来，探索则把控制逐步转移给学习者[2]。这套路径的合理性在生成式 AI 出现以后并没有失效：如果学习者从第一步就只接收完整答案，他可能取得高水平产品，却没有形成生成这些产品所依赖的内部结构。

样例渐退研究进一步说明，从完整示例直接跳到完全独立求解并非唯一选择；可以逐步删除步骤，使学习者承担越来越多的推理[4]。Koedinger 与 Aleven 将其提升为 assistance dilemma：帮助太少会使学习者陷入无生产性的困难，帮助太多则会减少必要的认知活动[3]。Kapur 的 productive failure 研究与检索练习、生成效应等传统，也从不同侧面强调“先生成、再得到反馈”与“直接看到正确答案”不是等价学习事件[5-6]。

如果沿着这条路径面对生成式 AI，最自然的处方是：AI 先多帮，随着熟练逐渐少帮，最终由人独立完成。它的终点很清楚——人不再依赖当前支架也能完成目标任务。

问题也由此出现：如果某些外部资源将永久存在，且它们本身已经成为专业实践的一部分，那么“完全独立”是否仍然应该是所有能力的终点？一名统计学家不需要脱离统计软件，一名飞行员不需要抛弃仪表，一个成熟程序员也未必需要拒绝 IDE 和代码搜索。将所有外置重新内化，并不会自动产生更高阶的专业能力。

第二条路径：分布式认知拒绝把“脑内独立”当作唯一能力边界

分布式认知从一开始就质疑把认知自然地关在单个人的头脑里。Hutchins 对航海与驾驶舱的研究表明，真实任务中的计算、记忆与决策分布在多人、仪表、纸面表示、程序和环境中；Hollan、Hutchins 与 Kirsh 因此主张把认知系统的分析单位扩展为个体与外部资源之间动态组织的功能系统[7-8]。Clark 与 Chalmers 的延展心灵论题更加激进：在满足稳定可用等条件时，外部资源可以在功能上成为认知过程的一部分[9-10]。

认知卸载研究也证明，人会主动改变环境以减少内部记忆、计算和控制需求；这种行为并非天然等同于偷懒，而可能是一种理性的元认知策略[11-12]。若一个工具能够稳定承担低价值计算，把注意力释放给更高层判断，那么要求使用者为维持“纯脑内能力”反复做同样的机械操作，可能反而降低系统能力。

沿着这条路径，AI 不需要像临时脚手架一样最终消失。一个人真正需要形成的，可以是驾驭“人+AI+资料+工具+制度”整体系统的能力：知道什么该外置，如何构造提示，怎样调用多个工具，如何组合不同结果，以及在何种任务条件下需要改变分工。这里的终点不是“我一个人能不能做”，而是“这个分布式系统能不能持续、准确、低成本地做”。

然而，分布式认知的强项也构成本章必须补上的缺口。一个系统今天整体运作优秀，并不能直接告诉我们：如果某个关键外部部件突然失效，原本长期不再独立工作的内部节点能否恢复功能。分布式认知可以解释能力如何跨边界存在，却不自动提供“备用能力最近一次何时被独立验证”的时间账本。

第三条路径：安全科学允许自动化长期存在，却不允许接管能力只停留在名义上

自动化人因研究早已发现一个反讽：技术越成功地承担正常任务，人越可能被从最容易练习的日常环节中移走，却在最复杂、最罕见的异常时刻被要求接管。Bainbridge 在 1983 年指出，自动化不会让人彻底退出，反而把人的工作推向监控、诊断与异常恢复，同时长期不用的手动技能可能衰退[13]。Endsley 与 Kiris 随后

用实验展示 out-of-the-loop performance problem：自动化控制会影响情境意识，并在重新转为人工控制时降低表现[14-15]。

Parasuraman、Sheridan 与 Wickens 将自动化区分为信息获取、信息分析、决策选择和行动执行等类型，并指出不同层级的自动化对人的角色产生不同影响[16]。Parasuraman 与 Riley 对 use、misuse、disuse、abuse 的区分，以及后续自适应自动化研究，都表明“自动化越多越好”与“自动化越少越安全”同样站不住[17-18]。Horvitz 的 mixed-initiative 研究则把控制权在人与系统之间的动态分配带入交互设计[19]。

从这一路径看，AI 完全可以长期承担正常运行，但人必须在自动化失效时仍有足够情境模型去发现问题、理解为什么错、决定在哪切断，并把系统带回可控状态。于是它的终点既不是完全独立，也不是最大化系统平均绩效，而是“异常发生以后仍能安全恢复”。

但安全科学的经典答案若简单移植到所有 AI 知识工作，也可能过重。飞行模拟器、复训和接管演练在高后果系统中合理，并不意味着每个写作者、分析师、教师和程序员都要周期性关闭全部 AI，完整重做全部任务。真正需要保留的可能只是少数能够暴露系统模型与恢复能力的关键回合。

真正的冲突：三条路径把“帮助成功”结算在三个不同终点

把三条路径放在同一张表上，会看到它们不是三个角度，而是三种结算方式。教育学问“这个人最终会不会”；分布式认知问“这个人—工具整体能不能做好”；安全科学问“工具失败以后这个人还能不能接住”。如果把其中任一项设成唯一终点，另外两项都会被牺牲。

三者底下却藏着同一个没有被说出的前提：一次帮助可以在某个终点上完成结算。教育学把人独立完成当终点，工具之后主要是效率问题；分布式认知把整体系统绩效当终点，只要协同稳定即可；安全科学把安全恢复当终点，只要接管条件被保留即可。它们都倾向于把“哪一样是终点、另外两样是手段”固定下来。

生成式 AI 环境恰恰推翻这个前提。今天用于提高系统效率的 AI 外置，明天可能成为检验人的训练材料；今天已经内化的技能，后天可能因为模型成熟而被正式外置；一次为了训练安排的人工独立判断，可能反过来发现 AI 系统的新边界，从而改变下一轮分工。终点与手段会在轮次之间交换位置。于是，帮助不能再按一条单向渐退线结算，而需要被设计成可以进入、停留、让位、再进入的循环。

从“能力退化”再往前一步：能力可能先变得不可观测

关于 AI 与能力的公共讨论很容易落入二元框架：AI 要么帮助人变强，要么使人退化。但在这两个状态之间存在一个更早、也更难被看见的状态——人的能力究竟有没有下降，系统已经没有足够新证据知道了。

设想一名研究分析师过去每周独立完成资料检索、假设比较、数据检查和报告撰写。AI 进入半年以后，检索、初步分析和初稿都由模型承担，他主要阅读、修改与提交。报告质量可能更高，耗时也更少。到这里，没有证据证明他的分析能力已经下降；但同样没有新的证据证明他仍保持半年前的独立诊断能力，因为原本能够暴露这些能力的行为已经长期不再发生。

这与技能退化是两个不同变量。一个人可以技能仍然很好，但六个月没有任何 AI 暴露前的独立判断记录；也可以能力已经很弱，却昨天刚完成一次无 AI 测试，因此组织对他的能力状态非常清楚。前一种情形“不确定性高”，后一种情形“不确定性低”。如果把这两者都叫“依赖”或“退化”，我们就失去了一个重要的中间状态。

安全领域其实早已有一个相近但并不相同的警告。FAA 关于 currency 与 proficiency 的长期讨论明确提醒：满足近期经历要求并不等于真实熟练，反过来，周期性训练和能力检查也不能被简单等同[27]。2003 年的相关研究甚至发现，全部受试者都满足“current”身份时，首次飞机仪表熟练检查的通过率仍不足一半。这说明“最近做过”与“现在做得好”是两个变量。本章进一步提出的 AI 问题则再多一层：在 AI 持续代行以后，组织甚至可能连“最近一次真正由人独立做过关键动作是什么时候”都没有字段。

能力证据债：一种既不是“会”也不是“不会”的系统状态

能力证据债（Capability Evidence Debt）是指：某项能力仍被系统当作未来关键节点需要的人类备用能力，但由于 AI 持续代行，使系统长期没有产生新的、在 AI 答案暴露前形成且随后可被独立校验的人类表现证据，由此累积的能力状态不确定性。

这个定义必须同时守住三个边界。第一，债不是能力损失本身。一次新的独立取样完全可能证明能力仍然优秀，此时被清偿的是不确定性，而不是“修复了退化”。第二，债不是 AI 使用时长。一个人每天使用 AI 八小时，只要关键节点仍持续产生有效的人类首判和恢复证据，证据债可以很低；另一个人每周只用 AI 一次，却恰好让 AI 长期垄断唯一的异常诊断节点，局部证据债反而可以很高。第三，债不是道德化的“依赖”。它不预设 AI 多用是不好的，只要求凡是仍需要人承担未来接管责任的能力，其当前状态不能无限期靠过去的资格与历史表现续期。

因此，能力证据债首次把一类过去常被记为“零”的对象带进账本：不是已经失败的能力，而是“仍被要求存在、却很久没有被重新看见的能力”。传统绩效系统记录任务完成了多少，AI 日志记录模型生成了什么，教育平台记录答对率，组织人事系统记录证书与岗位。它们很少记录一个问题：这项关键能力最后一次在没有先看到 AI 答案的情况下被真正观察，是什么时候?

这个对象之所以重要，是因为系统在能力事实变坏以前就可以积累风险。一个高度可靠的 AI 可能使正常运行多年不出事，恰恰因此人几乎没有介入机会。表面稳定越长，关于人的能力证据可能越陈旧；而没有事故不会自动告诉我们“备用能力仍然可用”。这与通常的自动化偏误不同：自动化偏误讨论人在看到机器建议后过度接受它；能力证据债发生得更早——即使人从未做错一次判断，系统也可能因为长期没有独立判断事件而失去对其能力状态的观测。

真正危险的不是“高 AI 使用”，而是“高代行 × 低证据刷新”

把“AI 代行程度”和“人类关键能力证据是否持续刷新”分成两条轴，可以得到四种完全不同的人机状态。

右下角“静默替代区”是本章真正的靶格。它不会因为 AI 表现差而出现，反而通常建立在 AI 足够好、足够稳定、足够方便的基础上。AI 越少犯错，人越少被迫重新进入；人越少进入，系统关于人类能力的证据越少；证据越少，又越容易因为“反正一直没出事”继续把同一环节交给 AI。

所以，一套只用任务成功率监控人机系统的制度会系统性漏掉这一格。它能看到“任务完成”，看不到“如果模型越界，人是否仍然能发现”；能看到“报告质量提高”，看不到“辅助报告质量是否仍能映射到本人独立判断”；能看到“自动化全年稳定”，看不到“备用操作者上一次真正完成异常诊断已经是多久以前”。

这也是本章与“AI deskilling”最重要的分界。Ferdman 提出的 capacity-hostile environments 把 AI 导致能力培养环境恶化理解为结构性问题，而不是个人自律不足[25]。本章接受这一方向，但把时间点再前移：在真正 deskilling 尚未被证实以前，系统首先可能失去的是能力可观测性。若一次校准显示人仍然熟练，deskilling 的强主张不成立，但能力证据债仍曾真实存在并已被清偿。

证据龄比：不测“你有多强”，只测“我们多久没真正看见它”

为了把“能力可观测性”从态度词变成可以查日志的问题，需要把能力拆到任务节点。对于每一个事先定义为能力关键节点的动作 i，记录最近一次有效能力证据发生时间 τ_i。当前时间为 t，则该节点的证据年龄为：

$$A_i(t) = t - \tau_i$$

一次事件只有同时满足四个条件，才算“有效能力证据”：第一，人的核心判断发生在 AI 答案可见以前；第二，该任务确实命中需要保留的能力节点，而不是用机械动作冒充；第三，随后存在结果、标准答案、专家复核、现实反馈或另一条独立路径，使这次判断可以被校验；第四，记录中保留任务版本、AI 权限与关键上下文，避免把不同难度的任务混在一起。

不同节点允许的证据年龄不应统一。令 W_i 为该节点事先设定的最大证据刷新窗口，则定义一个无量纲的“证据龄比”：

$$\rho_i(t) = A_i(t) / W_i$$

$\rho_i \leqslant 1$ 只表示证据仍在预定窗口内；$\rho_i > 1$ 只表示“应该重新取样”。它绝不表示能力已经下降，更不能直接被当作个人绩效分数。W_i 的设定至少取决于三个条件：该节点失效的后果有多大、任务环境变化有多快、未来突然需要人工接管的概率有多高。低风险、稳定且可以廉价独立验证的节点，窗口可以很长甚至不再刷新；高后果、快速变化而又要求人承担最终恢复责任的节点，即使操作者资深，也应有更短的刷新窗口。

如果组织需要汇总风险，可以对已经过期的节点做加权清单，但本章不主张一个跨行业通用的“能力债总分”。因为一旦把单一数字拿去排名，就会诱发为了好看而制造人工回合。本章真正需要的最低读数只有一个：某个关键节点最近一次有效证据距今多久，是否已经超过预先公开的刷新窗口。

能力校准回路：帮助不退出，而是在关键节点周期性让位

能力校准回路不是“定期关掉 AI”，而是把完整任务拆成少数能力关键节点，在这些节点上周期性恢复人类独立首判，并在取得新证据后让 AI 重新进入。

第一步是“首判”。人在 AI 答案出现以前留下最小独立模型。这里不要求完整做完任务。学生可以只写解题路径，医生可以先标记最值得警惕的异常，程序员可以先写关键不变量与最可能失败的边界，研究者可以先写最危险的竞争解释。首判的价值不是为了多做劳动，而是为了保存一个未被 AI 答案污染的基线。

第二步是“协作”。首判留下以后，AI 可以充分进入，承担检索、生成、计算、改写、代码实现、候选扩展与模拟。本章不把“少用 AI”视为美德。能被廉价外置而且错误容易验证的环节，应当尽可能外置。能力保护的目标不是重新制造低效率。

第三步是“差分”。真正的学习对象不再只是 AI 给出的正确答案，而是人的首判与 AI 结果之间的差。AI 看到而人没看到的约束是什么？人判断正确而 AI 遗漏的边界是什么？若现实结果后来可观察，哪一种判断被支持？差分把“我看懂了 AI 解释”转成“我的原模型具体在哪一处需要修改”。

第四步是“盲段”。首判只能证明会不会起步，不能证明能不能独立诊断与恢复。因此在低风险环境中，需要周期性安排一个局部 AI 不可见的短回合，由人独立识别、诊断、修复或完成恢复。高风险系统不能在真实现场为训练故意撤除有效安全辅助，盲段应转移到模拟器、沙盒、影子模式、历史病例、演练或低风险任务。

第五步是“恢复”。独立取样不是终点。新证据一旦获得，AI 重新进入；如果能力表现良好，下一次刷新窗口可以按规则维持或延长；如果暴露缺口，AI 帮助反而应该增加，并把差异变成针对性训练；如果确认某项能力未来已经没有保留价值，则把该节点正式移出“人类关键能力清单”，不再为它支付训练成本。

于是，生成式 AI 时代的路径不再是“帮助→逐渐减少→独立”，而是“首判→协作→差分→盲段→恢复→重新配置”。它允许 AI 长期常驻，却拒绝让关键能力无限期处在没有新证据的名义备用状态。

帮助何时增加、维持、局部渐退或重新进入

这里出现一个与传统支架理论明显不同的结论：AI 帮助强度不再与人的能力水平单调反向。一个能力很弱的新手可能同时需要“更多 AI 帮助”和“更多独立取样”——前者保证当前任务不失败，后者保证学习真的发生。一个非常熟练的专家反而可以长期让 AI 承担 90%的机械工作，只要关键异常识别与恢复节点持续产生新证据。

因此，“越会越少帮”不是通用规律。真正应随能力变化的，是 AI 在哪些位置遮住人的模型、哪些位置必须周期性让出观察窗口，以及这些窗口由什么信号触发。

十二之一、为什么“能力形成”与“能力证明”必须拆开

传统训练常把两件事情压在同一个动作里：让学习者自己做，既被当作训练方法，也被当作能力测量。没有 AI 时，这种合并相当自然，因为一个人要产生答案，通常必须同时调用自己的知识、策略与执行能力；完成过程本身既提供练习，也留下证据。但生成式 AI 进入以后，“产出一个正确结果”与“人的内部能力发生了什么”第一次可以大规模脱钩。AI 能够把外部表现迅速抬高，而人的知识结构可能增长、保持、下

降，甚至只是暂时未被调用。若仍把任务成品直接当作人的能力证明，就会把系统绩效与个体能力混成一个指标。

反过来，也不能因为成品不再足以证明能力，就要求所有训练退回纯人工状态。一次无 AI 完成具有两种不同功能：它可以作为训练——让人真正经历检索、生成、纠错；也可以作为测量——让系统知道这个人目前能独立做到什么。训练追求的是能力改变，测量追求的是状态可知。二者对任务设计的要求并不相同。为了训练，任务需要适宜难度、反馈、重复与间隔；为了测量，任务首先需要不被答案泄露、能够命中目标能力并且有外部校验标准。把二者混在一起，就会出现两个错误：一是用一次短测验假装已经产生学习，二是用大量完整人工劳动只为了获得本可由短探针取得的状态信息。

因此，能力校准回路首先是一套“测量—配置”机制，而不是完整教学法。首判与盲段的第一任务，是让关键能力重新显露；一旦证据显示缺口，才进入真正的训练回路。比如程序员在 AI 关闭的十分钟故障定位中无法识别并发死锁，这次失败已经完成了校准：我们知道缺口在哪里。但如果组织只是记录失败并在下个月再考一次，能力不会因此增长。接下来需要的是针对并发模型、日志解释、最小复现和诊断策略的训练，训练时甚至可以大量使用 AI。这样，“更多 AI 帮助”与“更严格能力维护”不再互相矛盾。

这一区分还会改变教育评价。未来的 AI 丰富环境中，课程成绩可以继续评价人机系统完成学习任务的质量，但若课程同时声称某些知识与判断已经成为学生本人能力，就必须另有独立能力证据。这个证据不必是一场长达两小时、全面禁用 AI 的传统考试；它可以是若干被精确设计的短探针，分别读取问题表征、关键步骤生成、错误诊断、迁移和解释。真正重要的是：每一个“学生已经掌握”的结论，都能够追溯到至少一次没有先看到 AI 答案、并且能被独立校验的表现事件。

同样的逻辑适用于职业资格。一个组织可以完全接受“员工与 AI 共同工作才是现代岗位的正常形态”，同时仍要求某些紧急恢复、边界识别与责任判断必须有人类能力证明。这不是怀旧式的去技术化，而是在职责仍然保留在人身上的地方要求证据对称：如果制度在出错时会问“你为什么没有发现”，那么制度在平时就不能只记录 AI 成功完成了多少次，而从不制造任何机会确认这个人现在还能发现什么。

十二之二、最后签字为什么不是接管：从名义在环到可用在环

生成式 AI 治理中最容易出现的一种伪解决，是在自动化链条最后保留一个“人类审核”按钮。流程看起来仍然 human-in-the-loop：AI 生成，人阅读，人点击确认，所以责任链似乎没有断。但从人因工程角度看，最后签字只证明人在流程中拥有一个动作位置，并不证明他拥有形成独立情境模型所需的信息、时间、注意力与练习。一个人

如果从未亲自经历证据如何被筛选、假设如何被排除、异常如何逐步出现，那么面对一份已经高度组织化、语言流畅的 AI 结论，他的审核很容易退化成结果确认。

这里必须区分“名义在环”“信息在环”“判断在环”和“恢复在环”。名义在环只是流程要求某个人点确认；信息在环意味着他能看到足够的原始材料和不确定性；判断在环意味着在 AI 结论出现之前或独立路径上，他能够形成自己的关键判断；恢复在环则意味着 AI 失效以后，他拥有改变目标、重建情境并执行替代方案的权限与能力。四者可以分离。一个系统完全可能名义上处处有人，实际上只有点击权，没有重新发起权。

能力证据债主要关心后两层。因为只有“判断在环”和“恢复在环”才能产生关于人类备用能力的强证据。让员工每天审核一百份 AI 报告，也许增加了与 AI 输出的接触，却未必刷新问题定义、异常识别和替代推理的证据。相反，每月一次精心设计的 AI 前首判或边界变化案例，可能比一千次事后确认更能告诉组织：这个人是否仍然拥有独立判断的起点。

这也解释为什么“保持人类最终责任”如果没有配套能力机制，会产生责任与控制不对称。系统一方面把日常判断交给 AI，减少人形成独立模型的机会；另一方面在失败后仍以“人最后签字”为理由把后果归给人。若人没有真实的信息、时间、替代路径和能力刷新机会，最后签字并不能自动转化为有意义的控制。本章因此主张：凡制度保留最终人类责任的节点，都应至少对应一种可观察的人类判断或恢复证据；否则所谓 human-in-the-loop 只是责任标签，而不是可用能力。

十二之三、从单向渐退到“角色轮换”：三条路径怎样组成一条循环

如果把支架渐退、认知外置和自动化接管真正放进时间序列，就会发现它们并不需要被折中成一个平均方案。更有力的重构是允许三条路径在不同轮次依次占据主导。学习初期，教育学路径主导：AI 需要解释、分步、给反馈，但不能在所有关键生成动作发生以前直接占满任务。能力稳定以后，分布式认知路径主导：大量成熟操作正式外置，人的注意力移向更高层的目标、边界与异常。证据逐渐陈旧或环境变化以后，安全科学路径短暂主导：在受控条件下把某个关键节点重新交还给人，检验其是否还能接管。证据刷新之后，系统再次回到高效外置状态。

因此，三条路径不是“谁更正确”，而是分别适合任务生命周期的不同阶段。传统支架理论的问题在于它容易把“人独立完成”视为终局；分布式认知的问题在于它容易把当前系统协调当作分析终点；安全科学的问题在于它常从异常接管倒推训练要求。把三者组织成循环以后，独立、外置与接管都不再是终点，而成为可以反复切换的角色。

今天人的独立操作是训练和测量，明天同一操作可以变成 AI 的常驻职责，后天又可能因为系统版本变化而临时回到人手。

这种轮换可以解释一个传统渐退框架很难处理的现象：专家并不一定比新手更少使用 AI。新手可能在知识形成阶段需要 AI 延迟给出关键答案，而专家已经拥有稳定模型，可以把 90%的正常执行永久外置；但专家承担更高后果的异常责任，因此他在少数恢复节点上反而需要更严格的周期性验证。于是，“AI 用量”与“能力水平”不再构成一条简单负相关曲线。真正随专业成长变化的，是 AI 介入的位置、人在前置判断中承担的抽象层级，以及校准所针对的异常复杂度。

这也意味着一项能力的生命周期至少有三种可能结局。第一，成为核心人类能力：继续训练并保持证据刷新；第二，成为长期外置能力：正常情况下由 AI 承担，但在接管相关节点保留局部人工证据；第三，正式退休：经过风险分析确认未来不再要求人承担该功能，于是停止维护。能力维护只有同时允许第三种结局，才不会把所有历史技能都神圣化。一个成熟的人机系统不是让人永远保持过去的一切，而是不断重画“哪些能力还值得保留、保留到什么层级、凭什么知道它仍在”。

从这个角度看，AI 时代的支架不再只是从“多”到“少”的数量变化，而是从“替你完成”转向“帮助你形成”，再转向“替你长期执行”，最后在必要时转为“把关键节点让给你证明仍可接管”。帮助本身始终可以存在，变化的是它与人的行动之间的先后关系。这种先后关系比总使用时长更接近能力是否会生成、是否会被遮蔽的真正机制。

十二之四、怎样找到“能力关键节点”：不是凭职业传统，而是用四个问题筛选

能力校准最容易失控的地方，是把“关键能力”列成一张无限长的传统基本功清单。为了避免这种保守主义，节点进入清单前必须通过四个问题。第一问：如果 AI 在这里错了，错误能否被另一条低成本机制自动发现？如果可以，保留人工复现能力的价值就下降。第二问：错误一旦发生，是否可能在没有明显信号的情况下继续传播并放大？如果不会，人工接管的紧迫性也较低。第三问：当 AI 不可用或情境越界时，组织是否仍然真实地要求某个人在有限时间内恢复这项功能？如果答案是否定的，就不应以“万一”为理由无限维护。第四问：这个节点是否控制后续路径，而不是只完成一个可替换动作？能够改变问题定义、证据边界、停止规则和恢复方向的节点，应获得更高优先级。

四问会产生一种与职业传统不同的能力地图。例如在写作中，“快速写出语法正确的 3000 字”可能不再是关键节点，因为 AI 生成容易、错误也可修改；但“发现一条引用其实不能支撑承重判断”可能仍是关键节点，因为它控制整篇论证是否成立。在编程

中，“记住 API 参数”可以彻底外置，而“判断一个通过全部单元测试的实现是否违反系统不变量”可能必须保留。在医疗中，病历摘要可以由 AI 生成，但“某个看似常规的症状组合为什么不应进入常规路径”可能需要持续的人类证据。

节点筛选还必须随着技术变化而重做。今天难以自动验证的功能，未来可能因为形式化验证、冗余模型或传感器改善而变得无需人工备用；今天可以放心外置的功能，也可能因为 AI 开始承担更多上游判断而重新成为高风险瓶颈。能力清单因此不是职业身份清单，而是一张随系统架构更新的风险图。每次模型大版本、数据源、法律责任或工作流程发生变化，都应重新跑一遍四问，而不是沿用过去的“基本功”。

这套筛选还有一个伦理作用：它要求组织对“为什么还要人会”给出明确理由。若答案只是“以前都是这样”“专业人士就应该会”或“完全靠 AI 感觉不好”，就不足以强迫人持续投入训练时间。只有当一个能力与未来实际责任、错误发现或恢复路径存在明确联系时，能力维护才具有正当性。这样既保护人类关键能力，也保护人不被无意义的能力保养占用。

哪些环节可以永久交给 AI，哪些必须周期性回到人手

如果每一个被 AI 外置的过程都要求周期性人工回收，人类很快会陷入另一种形式主义。计算器出现以后没有必要要求所有专业人员定期长除法，数据库成熟以后也没有必要为了“证明记忆力”而重复背诵所有可检索事实。判断一项能力是否必须保留，不应看它过去是不是由人做，而应看 AI 失败时，人是否仍需要该能力来发现错误、限制损失或恢复系统。

原则上适合永久外置的节点通常具有四个特征：结果可以由另一条廉价路径验证；错误容易被发现且可逆；错误不会在无信号状态下级联传播；未来恢复系统不依赖人手复现这一环节。格式转换、机械计算、标准模板、候选扩展、重复性分类、确定性数据搬运往往属于这一类。

必须保留周期性人工证据的节点则具有另一组共同特征：它们决定问题如何被定义、什么条件变化会使当前答案失效、哪一个异常值得怀疑、哪些证据不能被当前模型吞进去、目标冲突如何排序、AI 可能在哪个机制上错、错误开始扩散时在哪里切断、最低条件下如何恢复。它们未必直接生产最终产品，却决定什么时候不能继续沿着原路径走。

因此，AI 时代的人类能力结构可能出现一个重要变化：越来越多的正常执行被永久外置，越来越少但更关键的“发起—划界—识异—诊断—修正—恢复”节点被保留下

来。人的总操作量下降，并不意味着对人的能力要求下降；要求可能反而集中到更少、更难、更接近边界的节点上。

公开数据真跑：AI 提高练习表现以后，“眼前做得好”还多大程度说明“本人会了”

为了对“能力可观测性”做一条最低成本的经验压力测试，本章使用 Bastani、Bastani、Sungu 等公开在 GitHub 的原始数据与代码[22]。原研究在土耳其一所大型高中进行四轮 90 分钟数学学习实验，班级被随机分配到三种条件：无生成式 AI 的 control、接近通用 ChatGPT 交互的 GPT Base（数据中 Treatment arm=vanilla），以及加入教师解题信息和“不直接给答案”等教学护栏的 GPT Tutor（Treatment arm=augmented）。每轮先完成可使用相应资源的练习部分（Part2Tot），随后完成无 AI 考试部分（Part3Tot）。原研究主分析排除 Honors 班，本章遵循同一处理。

原论文已经给出清晰的平均效应：GPT Base 和 GPT Tutor 都大幅提高 AI 可用时的练习成绩，GPT Base 随后在无 AI 考试中显著低于对照，而 GPT Tutor 基本消除了这一平均损失，却没有产生明显的无 AI 考试增益[22]。本章不重复检验这一主结果，而问另一件原论文并未以此方式提出的问题：在每一种条件内部，一个学生练习阶段表现得越好，是否仍然像对照组那样可靠地说明他随后独立考试也会表现得越好?

具体做法如下：首先排除 Honors=1 的记录；然后按 Student ID 聚合每名学生多轮的平均练习分和平均无 AI 考试分，得到 825 名具有学生层面聚合数据的学生，其中 control 316 名、GPT Base 238 名、GPT Tutor 271 名；随后分别计算 Pearson 相关，并以班级为重抽样单位进行 5000 次聚类自助抽样，避免把同班学生当成完全独立。作为稳健性检查，还报告 Spearman 等级相关，以及在各组内分别线性残差化既往 GPA 后的相关。所有变量选择与排除规则在统计前固定，没有依据结果调整阈值。

结果出现了一个对本章非常关键、同时也必须谨慎解释的形状。在无 AI 对照组中，练习表现与随后独立考试表现高度对应，r=0.756；在 GPT Base 中降为 0.435，在 GPT Tutor 中为 0.486。聚类自助法中，对照组与 GPT Base 的相关差中位数约 0.320，95%区间约[0.200,0.443]；对照组与 GPT Tutor 的差中位数约 0.270，95%区间约[0.153,0.389]。Spearman 相关以及控制既往 GPA 后的分析方向一致。

最有理论价值的不是 GPT Base，而是 GPT Tutor。GPT Tutor 的平均无 AI 考试分 0.315，与对照组 0.321 非常接近，符合原论文“护栏基本消除平均学习损失”的主结论；然而，它的辅助练习成绩与后来独立成绩的个体对应仍明显低于对照组。也就是

说，一个班级平均意义上可以没有出现明显学习损失，同时 AI 辅助态下“这个学生现在做得很好”仍然不再像无 AI 环境那样透明地显示“这个学生本人独立能做到什么”。

这条结果不能被解释成“GPT Tutor 也造成了技能退化”。相关下降可能来自 AI 对不同学生提供不同强度的帮助、方差结构变化、题目难度、范围限制或其他机制。本章因此只接受一个更弱的结论：AI 帮助可以提升当前任务表现，同时降低当前表现作为独立能力信号的保真度。这个结论恰好支持能力证据债的必要性——在平均成绩没有下降的条件下，能力可观测性仍可能已经发生变化。

这也是一次对本章的反向约束。原本可以提出更强的话：“长期 AI 帮助首先会让能力下降。”公开数据不支持这种写法。GPT Tutor 显示，好的 AI 设计完全可能避免平均学习损失；Kestin 等人的大学物理随机实验甚至发现，遵循学习科学原则设计的 AI 导师能够在更少时间内取得高于课堂主动学习的学习增益[23]。因此，本章必须把新命题收窄到“可观测性”而不是“必然退化”。

第二个反向约束：AI 交互方式决定学习，不能把“使用 AI”当成单一剂量

Shen 与 Tamkin 在 2026 年的随机实验中让软件开发者学习一个陌生的异步编程库，发现 AI 组在随后技能测验中总体表现更低，尤其在概念理解、代码阅读与调试上出现损失；但他们同时识别出多种不同 AI 交互模式，其中若干保持较高认知参与的模式并没有呈现同样的学习损失[24]。这再次说明，“用了 AI 多少分钟”不是足够细的自变量。

对本章而言，这一结果迫使“能力校准回路”放弃另一句过强的话：并不是只要周期性安排无 AI 回合，就一定能产生能力。校准的首要作用是刷新证据和暴露缺口，真正的能力增长仍依赖生成、检索、反馈、间隔、难度、动机与任务结构。一个人可以刚刚完成独立测试，因此能力证据非常新鲜，但测试也可能清楚证明他当前不会。清偿证据债不等于提升能力，正如诊断出问题不等于已经治疗。

因此需要严格区分两个回路：校准回路回答“我们现在凭什么知道这个人能做到什么”；训练回路回答“发现缺口以后怎样使他真正变强”。在 AI 时代，这两个回路过去常被同一个“让他自己做”动作混在一起，现在必须拆开。

十五之一、辅助表现为什么会失真：三种机制不能混成“能力下降”

练习成绩与独立成绩的对应下降至少可能来自三种不同机制，而它们对干预的含义完全不同。第一种是“能力替代”：AI 承担了原本需要学习者完成的认知操作，导致这些操作本身没有得到足够练习，随后无 AI 能力真的下降。这是传统 deskilling 最直

接的路径。第二种是“表现压缩”：AI 把低能力者的当前表现抬得更多，使不同能力水平的人在辅助状态下看起来更加相近；此时人的真实能力可能没有下降，但辅助成绩的排序被工具压平。第三种是“任务重构”：AI 改变了任务本身，人不再通过原来的步骤到达结果，而转向提示、筛选、验证和整合，于是辅助表现测到的是另一组能力。三种机制都可能让相关下降，却不能被同一句“AI 让人变笨”解释。

能力证据债之所以选择“可观测性”作为更靠前的概念，就是因为它不要求事先裁定三种机制哪一个为真。只要组织仍然需要知道某项旧能力是否可用，而当前辅助表现已经不能可靠回答这个问题，就需要额外证据。之后的校准结果再决定机制：若无 AI 探针显示能力依然很好，问题主要是表现压缩或任务重构；若能力明显下降，才进入技能恢复；若新任务已经证明旧能力不再需要，则应正式退休而非修复。这个顺序避免在证据不足时先给人贴上“退化”标签。

因此，本章对 Bastani 数据的二次分析故意不把相关下降解释成因果效果。相关性的价值在这里不是证明一个强机制，而是揭示“辅助表现=本人能力”这一默认推断已经变得不安全。未来更严格的实验应随机控制 AI 介入位置而不仅是 AI 有无，例如比较“先独立首判再看 AI”“直接看 AI”“AI 只给提示”“AI 完整代做”四种条件，并在多个时间点分别测量问题表征、错误诊断、迁移与恢复。若这些细分指标能够说明相关下降来自何种机制，能力证据债就可以进一步从诊断框架发展成因果模型。

十五之二、为什么能力维护不等于把 AI 的工作重新做一遍

一旦承认辅助表现不能直接证明人的能力，最自然的反应是“那就让人自己再做一遍确认”。这正是本章试图避免的验证悖论：如果每次为了证明 AI 结果可靠，都要求人完整重做 AI 刚刚完成的检索、计算、分析和写作，AI 节省的认知劳动会在验证环节被重新花掉；能力维护也会退化为对旧工作流程的复制。真正需要验证的不是每一个输出步骤，而是那些一旦失效会改变整个任务方向、且无法被廉价冗余机制发现的节点。

例如 AI 写出一份一万字报告，人没有必要独立再写一万字。更有价值的能力探针可能只有三个：在看 AI 前写出核心因果结构；独立检查一条最承重的证据链；给出一个若成立就必须推翻结论的反例。若这三件都能完成，它并不能证明人会逐字复制 AI 全部劳动，却能证明他仍掌握决定报告是否应当成立的关键控制权。反过来，如果一个人能够自己重新排版、重新搜索大量资料，却无法识别核心证据与结论之间的断裂，那么“完整重做”也没有真正维护高价值能力。

因此，能力维护应遵循“最小充分暴露”原则：用尽可能小的人工工作量，取得足以区分关键能力状态的证据。这个原则与安全工程中的测试思想相似——测试不是重新制造整个系统，而是选择能够最大程度暴露失效模式的输入。AI 时代的人类训练也应

从“重复全部正常操作”转向“有意识地撞击最关键的失效边界”。这使效率与能力不再天然对立：机器承担大多数正常劳动，人把有限练习预算集中到最能决定异常结局的地方。

最小充分暴露同时要求反对一种象征性的“人类参与”。如果某个探针太简单，以至于无论能力是否存在都能通过，它只制造安心感；如果探针只检查 AI 答案表面的格式和流畅度，也不能读取边界判断。好的探针必须存在失败的真实可能，而且不同失败模式能够指向不同训练需要。它不是为了证明“人永远比 AI 重要”，而是为了在仍要求人承担责任的地方，让责任对应到真实可见的能力。

与最相近的九种说法逐一划清：这是不是旧概念换名字?

2026 年 Schlonsak、Jetter 与 Matthies 提出的 decision horizons 与 agency erosion 也是一个必须正面处理的近邻[26]。他们强调重复委托会改变被用户看见的选项空间，提高退出自动化路径的摩擦，使形式上的“人仍有控制权”并不等于实际能自由行动。这与本章同样反对把名义控制当作真实能力，但两者可以通过一个具体案例分开：如果一个用户仍能自由选择是否退出 AI、决策视野也很广，却已经一年没有任何独立异常诊断证据，本章判定能力证据债高，而 decision-horizon 理论未必判定其能动性已经被侵蚀。反过来，一个人的能力证据刚刚刷新，但 AI 界面长期压缩其可见选项，他可能证据债低而能动性问题高。

另一个重要近邻是 Lee 与 See 所讨论的 appropriate reliance：理想的人机系统不是最大信任或最小信任，而是让信任与系统能力相匹配[20]。本章与其差别在于对象不同。信任校准问“我该不该相信 AI”；能力证据债问“组织凭什么相信人在必要时还能完成那个动作”。前者主要校准人对机器的模型，后者主要校准系统对人的能力状态模型。

这九条分离线共同说明，本章不是否定既有概念，而是把它们留下的一块空白变成独立对象：一个人可能尚未表现出退化、没有明显自动化偏误、没有失去能动性、人机整体绩效也很高，但系统已经很久没有新的有效证据知道某项关键人类能力还是否可用。

这九条分离线之外，本书内部还有两处必须交代。

与第二章的分界已在该章末节写明：第二章维持能力，本章维持证据。此处补充一个更尖锐的推论——清偿证据债不等于提升能力，正如诊断出问题不等于已经治疗。一次成功的校准完全可能得到一个清楚的坏消息：证据刷新了，读数显示他现在不会。这

在第二章的框架里是一次维护失败，在本章的框架里却是一次成功的校准。两章对同一事件给出相反评价，正说明它们测的不是同一件事。

与第四章的分界更细，也更危险，因为两章几乎可以互相翻译：第四章的"可观察绩效对潜在能力的预测力下降"，正是本章"辅助态表现不再显示本人能做到什么"的计量表达。真正的差别在主张的强度。本章不主张能力发生了变化，只主张系统失去了判断依据；第四章则给出一个能力本身与能力可见性同时下降的动力学模型。前者是认识论主张，后者是因果主张，本章刻意停在更弱的那一侧，因为公开数据只支持到这里。合并条件相应地也很明确：若在经验上永远造不出"相关下降而锚定任务证明能力未降"的情形，那么本章的"债"就只是第四章"塌缩"的早期阶段，两章应当合并。这是本书内部最可能被合并的一对，本书事先承认这一点。

五类真实任务：同一个回路为什么不能做成统一“无 AI 考试”

第一类是学校学习。数学学习中，不必规定学生每题必须先完整独立做完才能问 AI。若本轮训练的是建模，可以要求 AI 介入前只写未知量、关系与第一步；若训练错误诊断，可以直接给一份 AI 生成的错误解法，让学生找第一个失效步骤；如果近期已经有充分的独立计算证据，重复算术完全可以交给 AI。这里的关键不是“关 AI”，而是让被训练的能力先显露。

第二类是软件开发。程序员没有必要定期手写所有样板代码来证明能力。更有信息量的取样是：在 AI 生成代码前写关键接口约束和最可能的失效条件；在 AI 生成后独立设计一个能把隐藏错误暴露出来的测试；周期性在沙盒中面对未知故障，不使用代理自动修复，完成定位与恢复。一个人可以彻底外置样板实现，却仍维持高质量的调试和系统恢复能力。

第三类是研究与知识工作。AI 可以长期负责转录、资料候选扩展、格式化和初稿。关键报告则保留一份极短的“AI 前判断”：当前最关键的变量是什么？哪条证据一变会改变结论？最危险的竞争解释是什么？随后把人的首判与 AI 输出、现实结果进行差分。真正持续被训练和观察的不是“能否独自写两万字”，而是问题结构、证据边界和反例发现。

第四类是医疗、航空、工业控制等高后果系统。这些领域必须反过来限制本章的方法：不能为了刷新能力证据而在真实病人、真实飞行或真实生产线上故意撤掉已证明有效的安全辅助。这里的校准应转移到模拟器、双读、历史病例、影子模式、故障注入沙盒和专门演练。能力证据不是最高价值，即时安全始终具有优先级。

第五类是组织管理与决策。AI 越来越容易承担方案生成、会议纪要、指标解释与风险清单。如果管理者长期只在模型生成以后选择选项，他可能没有明显绩效下降，却越来越少形成“在看见候选前自己的优先级是什么”的证据。周期性首判可以非常短：先写三条决策约束，再让 AI 给方案。这样保存的不是全部独立工作，而是价值排序与边界判断的可观测性。

最小实施协议：如何避免把能力维护变成昂贵形式主义

一个组织若真的采用能力校准回路，第一步不是给所有人安排“无 AI 日”，而是建立能力关键节点清单。每项任务只允许标出少数节点，并回答一个硬问题：如果 AI 在这里错了，未来是否仍要求人发现、限制或恢复？回答“否”的节点不进入清单。这样可以主动删除大量不值得保留的旧技能。

第二步，为每个节点写出最小证据事件。证据事件越短越好，但必须能暴露内部模型。例如“在 AI 输出前写出两条边界条件”“在模拟故障中指出第一处不可信信号”“在模型给结论前写最危险竞争解释”。不要求全任务人工完成。

第三步，设定刷新窗口 W_i，并明确什么变化会提前触发刷新：模型版本重大变化、数据源变化、岗位职责变化、法律规则变化、任务进入新领域、长时间未遇到异常等。这样，能力校准由可见事件驱动，而不是管理者临时感觉“大家太依赖 AI”。

第四步，记录结果但不把 ρ_i 直接作为个人排名。真正应该被追踪的是节点是否已经过期、最近一次证据显示什么缺口、下一轮需要增加何种帮助或训练。任何把“每月完成几次无 AI 任务”直接变成绩效 KPI 的制度，都会迅速诱发打卡与假首判。

第五步，允许节点退休。如果某项人工能力已经不再承担任何未来恢复责任，而且 AI 结果可廉价独立核验，组织应正式宣布它不再需要维护，而不是以“传统专业基本功”的名义无限训练。只有允许能力退出，保留能力才不会变成保守主义。

反噬预言：一旦制度化，最省事的达标方式恰恰会摧毁它

能力证据债最危险的地方不是理论错误，而是理论被管理制度采纳以后被做成指标。只要规定“关键节点每月必须有一次首判”，最省事的达标方式就是在打开 AI 前随便写一句；规定“必须完成盲段”，员工就会挑最简单的任务；规定“证据龄比不得超过 1”，部门会通过改变节点定义和窗口长度把数字做漂亮。

因此，本章无法为这种反噬提供一个彻底的技术出口。任何一套可量化的能力治理，一旦与奖惩强绑定，就会产生 Goodhart 式压力。能够做的只是限制损害：指标只指位置不指人格；证据窗口和编码规则公开；至少一部分能力证据由随机抽取任务产

生；对高风险岗位允许人工复核；不要求为了数字而在真实系统中关闭安全辅助；周期性删除已经没有保留价值的节点。

如果这些限制仍不足以阻止形式主义，那么能力证据龄比应退回为诊断工具而不是考核工具。本章宁可牺牲“可排名性”，也不主张用一个貌似精确的数字取代真实能力判断。

六条可使本章失败的经验预测

预测一：可观测性变化应当早于部分技能退化。 在连续 AI 代行、缺乏独立取样的任务中，首先应观察到辅助态表现与无 AI 表现之间的对应关系下降；若使用持续足够久，才可能出现异常诊断或接管表现下降。如果纵向研究发现两者始终同步，能力证据债作为独立中间状态的价值会下降。

预测二：AI 前首判应比 AI 后解释更能预测独立能力。 控制总学习时间、任务难度与 AI 使用量后，AI 答案出现前留下的结构判断，应比“看完 AI 以后解释为什么正确”更能预测随后无 AI 任务。如果二者长期没有差异，则本章关于基线被 AI 答案污染的主张受到反驳。

预测三：局部校准应当在相近成本下优于整体关 AI。 相同无 AI 时间预算下，按能力关键节点安排短首判与盲段，应比周期性完整关闭 AI 更能保持异常识别和恢复能力，同时对生产率损失更小。若完整关 AI 持续显著更优，说明本章对“关键节点而非完整任务”的切分过度乐观。

预测四：AI 使用时长不是决定变量。 在 AI 使用时间相近的两组中，持续刷新关键节点证据的一组应在未预告的边界变化任务上表现更好。若两组没有差异，则“证据刷新”没有本章主张的独立价值。

预测五：高证据债应能与高真实能力同时存在。 应当找到一批长期由 AI 代行、证据已过期的人，在突击校准中依然表现优秀。如果证据年龄几乎总能被技能退化完全解释，能力证据债就只是技能衰退的模糊代理。

预测六：某些技能可以彻底退休而不降低韧性。 对于错误易核验、低后果、可逆且未来恢复不依赖人工复现的节点，强迫人周期性手工完成不应带来可测的系统恢复收益。若事实普遍相反，本章关于“只保留关键节点”的范围划得过窄。

不适用边界与伦理约束

第一，本章不适用于已经被明确决定永久退出人类能力集合的任务。如果组织已经设计出可靠冗余，确认未来无需人类接管，那么“保持人工能力”没有自动优先权。技术进步本来就会让一些技能退出。

第二，本章不把所有 AI 前判断都视为学习。首判首先是测量动作；要产生能力增长，还必须有高质量反馈、差分复盘和后续练习。一个新手反复写错误首判而不获得纠正，只会稳定错误。

第三，不得在真实高风险服务中通过故意关闭有效 AI 或自动化来制造“能力证据”。医疗、法律、公共安全、航空、工业控制等领域的能力刷新必须优先采用模拟、影子模式、双路径验证或历史案例。

第四，能力证据读数不得直接用作不透明的自动淘汰工具。只要它影响晋升、执业、岗位或待遇，就必须公开节点定义、证据有效条件、刷新窗口和复核渠道。证据过期只能触发“重新取样”，不能直接推出“此人能力不足”。

第五，本章讨论的是仍需人类承担责任的系统。如果一个组织形式上要求“人最终负责”，实际却不给人足够信息、时间和权限接管，那么问题已经超出技能维护，进入责任错置与制度设计。不能用“多训练人”去修补一个根本无法接管的岗位。

自反：这篇由 AI 深度参与生成的论文，自己是否背着能力证据债?

如果本章的判断只要求别人遵守而不作用于本章自身，它就会立即失去可信度。这篇论文的公开资料检索、结构扩展、文字生成、参考文献整理和文档排版都由生成式 AI 深度参与，因此“作者是否具备独立完成同等规模文献检索、跨学科推演和两万字写作的能力”不能从这篇成品本身直接推出。高质量成品首先证明的是人机系统在本次任务中完成了工作，不自动证明全部能力已经内化到作者个人。

按本章自己的判据，这并不意味着作者没有能力，也不意味着必须脱离 AI 重写一次。真正需要保留的能力证据，应落在承重判断上：作者能否在不先看 AI 新答案的情况下说明为什么“能力退化”不足以解释问题；能否指出分布式认知为什么不能单独解决接管；能否解释 Bastani 数据的相关下降为什么不能被直接写成因果退化；能否在遇到一个反例时修改“能力证据债”的边界。如果这些关键判断完全无法由作者重新发起，那么本章自己就是“静默替代区”的一个样本。

因此，本章的人机协作声明不是形式附录，而是方法的一部分。AI 可以帮助生成文章，但署名者必须对哪些判断由自己承担、哪些数据已经独立核对、哪些仍需复核有明确边界。本章也不以“已经发表”作为能力证据；发表只能证明文本进入公共记录，不能证明其中每一项能力已经被作者独立持有。

结论：AI 时代真正需要保留的不是“纯人工状态”，而是能力的重新显露通道

生成式 AI 把一个旧问题推到了新的结构层。过去我们担心帮助太多会让学习者“不会”；今天我们还必须担心另一件更早发生的事：当 AI 长期稳定代行以后，系统可能越来越难知道这个人究竟还会不会。能力事实与能力证据从此不再自动重合。

教育学提醒我们，外部帮助只有在学习者真正承担活动时才可能转化为能力；分布式认知提醒我们，人类能力从来不是把所有信息与操作都塞回脑内，工具可以成为认知系统的合法组成部分；安全科学则提醒我们，正常状态下把人移出回路，不会自动取消异常状态下对人的要求。三条路径同时成立以后，唯一可持续的答案不是“AI 最终退出”或“AI 永远接管”，而是让任务分工保持可逆，让关键能力持续能够被重新显露。

由此，本章提出“能力证据债”作为一个在技能退化之前即可出现的系统状态，并提出“能力校准回路”作为动态路径：首判让人的模型在 AI 暴露前显露，协作允许 AI 充分创造效率，差分把人与 AI 的不一致转成学习材料，盲段在可控范围内验证诊断与恢复，恢复则把 AI 重新带回任务并更新下一轮分工。

真正应该渐退的，不是 AI 本身，而是没有产生任何人类能力新证据的连续代理；真正必须周期性回到人手里的，也不是全部旧劳动，而是那些决定“什么时候原来的路已经不能走”的关键回合。

未来，一个人的专业能力不再只由“他独自可以完成多少”定义，也不能只由“他和 AI 一起完成多少”定义。还必须增加第三个问题：当 AI 不再沿着正确的路前进时，他是否仍然能够看见这一点；而这个系统最近一次凭什么知道他能够?

这一重新划界也意味着，未来的专业教育不应把“会使用 AI”和“无需自己会”混成同一件事。越是把 AI 作为长期基础设施，越需要明确哪些认知动作已经被正式外置，哪些判断仍由人承担不可转移的责任，以及后者如何留下新鲜证据。只有把退出、保留与刷新三件事同时制度化，人机协作才不会在效率增长中悄悄积累一个没人知道大小的备用能力缺口。能力的未来不是把过去全部保存下来，而是在技术不断改写任务边界时，持续知道人还必须会什么、现在是否真的会，以及何时可以放心不再要求他会。

写死日期的经验赌注：截至 2029 年 8 月 12 日

本章在 2026 年 8 月 12 日作出如下可失败预测：到 2029 年 8 月 12 日，如果生成式 AI 已经广泛进入教育、软件开发与知识工作，而成熟的人机系统仍主要采用“允许

AI/禁止 AI”“AI 开启/AI 关闭”这样的整体权限设计，没有逐渐出现针对能力关键节点的 AI 前首判、局部无 AI 微型取样、模拟接管、最近一次有效能力证据或同类机制；或者即使出现这些机制，在相近生产率条件下也没有带来更好的异常识别、边界迁移或恢复表现，那么本章关于“能力证据债—校准回路”的核心判断应被视为受挫。仅仅出现更多 AI 素养课程、更多要求人在 AI 答案之后审核、或更多“人类最终签字”的制度，不算预测命中；真正的命中必须包含一个更严格的变化：系统开始把“最近一次什么时候真正观察到人的关键能力”作为下一轮人机任务配置的输入。

证据边界、复算说明与人机协作声明

本章为理论与方法论文，不声称“能力证据债”已经获得直接因果实证确认。Bastani 等人的公开数据二次分析只检验较弱命题：AI 辅助态表现与随后无 AI 表现之间的对应关系是否改变。数据来自作者公开仓库 main_regressions/final_data.csv；本章遵循原主分析排除 Honors=1，按 Student ID 聚合多轮 Part2Tot 与 Part3Tot，Pearson/Spearman 相关以及班级聚类自助区间均可由公开数据复算。二次分析并非原研究预注册假设，不能排除 AI 帮助强度、方差改变、范围限制等竞争解释。

公开文献检索截至 2026 年 8 月 12 日。对 2025—2026 年新近研究，本章优先引用正式论文、作者公开预印本、机构研究页面与公开数据仓库。Ferdman 的“capacity-hostile environments”、Schlonsak 等的“decision horizons/agency erosion”以及 Shen 与 Tamkin 的技能形成实验被作为高风险近邻与反向约束，而非仅作为支持材料。

研究问题、三学科选择、发表场景与核心方向由作者提出；生成式人工智能参与公开资料检索、跨学科结构推演、竞争概念搜捕、公开数据探索性复核、初稿生成、参考文献整理和 Word 排版。理论判断、证据取舍、引文核验、伦理边界与最终发表责任由最终署名作者复核并承担。生成式人工智能不列为作者。

本章附表

路径	默认终点	最怕的失败	对 AI 的自然处方
教育学：支架与渐退	人可以独立完成	帮助存在时会、撤掉就不会	随着熟练逐步撤除帮助
分布式认知：外置与系统	人—工具整体稳定完成	把必要外置误当成依赖	允许工具永久保留并优化协同
安全科学/人因工程	异常时仍能接管恢复	正常运行成功掩盖备用能力失效	自动化常驻，但周期性验证接管

	关键能力证据持续刷新	关键能力证据长期缺失
AI 代行程度低	独立训练区：能力容易直接观察，但可能牺牲效率	技能沉睡区：任务本身低频，问题不一定来自 AI
AI 代行程度高	受控共生区：大量外置，同时关键节点持续被取样	静默替代区：正常产出优良，但人的能力状态越来越不可见

当前信号	AI 动作	人类动作	理由
新手明显无法完成，且当前任务必须交付	增加帮助	保留最小首判和错误复盘	先保证任务成功，不把“少帮助”当训练原则
证据新鲜，节点低风险、易核验	维持帮助	无需重复低价值操作	外置本身不是能力失败
关键节点证据龄比 ρ>1，或模型/制度/任务边界发生变化	局部渐退	在一个关键节点进行首判或盲段	刷新可观测性，不关闭整个 AI 系统
校准已经完成	重新进入	把新证据写回分工规则	独立完成不是终点
真实高风险现场	不因训练撤除安全辅助	转移到模拟、影子或历史案例	即时安全优先于能力取样

组别	N（学生）	平均练习分	平均无 AI 考试分	Pearson r	班级聚类 Bootstrap 95%区间	Spearman ρ	控制既往 GPA 后的 r
无 AI 对照 control	316	0.284	0.321	0.756	[0.692, 0.805]	0.747	0.652
GPT Base / vanilla	238	0.474	0.287	0.435	[0.325, 0.537]	0.408	0.322
GPT Tutor / augmented	271	0.669	0.315	0.486	[0.376, 0.587]	0.482	0.393

最近邻	它已经解释了什么	本章不能被它吸收的分离线
支架渐退 / cognitive apprenticeship	帮助如何随着学习逐渐减少并转为独立能力	本章允许 AI 永久常驻；触发局部让位的不是“熟练程度”本身，而是关键节点证据是否过期
Assistance dilemma	给帮助还是暂缓帮助如何影响学习	即使平均学习没有下降，辅助表现也可能失去作为独立能力信号的保真度
分布式认知 / 延展认知	认知可以跨人、工具与环境分布	系统整体成功不能回答某个备用人类节点最近是否仍被独立验证
Cognitive offloading	人为何把记忆与计算转移到外部	外置可以 100%合理；问题只发

	资源	生在仍被要求承担未来恢复责任的节点长期无证据
Automation bias / appropriate reliance	人在看到机器建议后是否过度接受或拒绝	本章的有效证据要求发生在 AI 答案出现以前，研究的是建议暴露前的能力可观测性
Out-of-the-loop / skill decay	长期自动化后人的情境意识、技能与接管表现下降	本章允许真实技能尚未下降；只要近期状态未知，证据债已存在
AI deskilling / capacity-hostile environment	AI 结构性削弱能力培养与练习条件	若校准证明能力仍优秀，deskilling 可判否，但证据债仍可以被判为刚刚清偿
Mixed-initiative / adaptive automation	根据任务状态在人与自动化之间切换控制	本章增加一个不同触发器：即使 AI 完全没出错，也可仅因人类关键能力证据过期而短暂让位
Currency 与 proficiency	近期经历、资格与真实熟练并不等价	本章把问题推进到 AI 节点级：不是“最近是否练过”，而是“最近一次 AI 暴露前、可校验的人类关键判断在哪里”

本章说明

文本规模说明：本稿按约 2 万字理论与方法论文体量设计；中文字符统计以正文、摘要、表格和说明为准，参考文献英文字符不计入中文字符数。

第四章

成功为何遮蔽退化

绩效上升、能力可见性下降与绩能遮蔽循环　——行为经济学 × 心理测量学 × 组织学习理论的跨学科对撞

生成式 AI 时代绩效上升、能力可见性下降与能力空洞化的自我强化机制

——行为经济学（即时收益、延迟成本与努力规避）×心理测量学/测量理论（潜在能力、测量污染与可识别性）×组织学习理论（成功陷阱、开发—探索与路径强化）的跨学科对撞

摘　要

生成式人工智能正在同时制造两个表面上互相矛盾的事实：一方面，越来越多研究显示 AI 可以显著提高写作、客户服务、咨询、编程等任务的速度与质量；另一方面，关于认知卸载、独立表现下降、技能形成受损与长期去技能化的证据也在增加。真正困难的问题并不是“AI 究竟提高还是降低人的能力”，而是当 AI 持续参与任务以后，我们越来越难从最终产出反推出人的潜在能力究竟发生了什么变化。本章以跨学科互否方法，将行为经济学、心理测量学与组织学习理论置于同一个动态问题中。第一层对撞发现：行为经济学解释个体为什么会反复选择具有即时收益的认知外包；心理测量学说明 AI 介入后最终绩效成为“人类能力+AI 能力+人机匹配+任务条件”的混合指标，原本用于识别人类能力的错误、迟疑、耗时和失败信号被系统性吸收；组织学习理论则解释为什么短期成功会把资源和注意力继续推向已获成功的 AI 增强路径，压缩独立练习、探索和能力检验。第二层对撞进一步揭示一个此前容易被分开讨论的自我强化机制：AI 不仅可能改变潜在能力本身，还会同步降低能力的可观察性；而能力越不可观察，未来能力损失越难进入当下决策，个体与组织越倾向继续外包；外包越多，独立练习和失败暴露越少，能力可见性进一步下降。本章据此提出“绩能遮蔽循环”（Performance–Capability Masking Loop, PCML）与“能力可见性塌缩”概念，并建立一个包含潜在能力 θ、AI 介入强度 a、可观察绩效 Y、诊断信息 I 与独立练习 p 的动态模型。本章进一步提出四类状态、三种测量污染、五项可证伪命题以及“锚定任务—扰动任务—撤援任务—延迟迁移”四层识别方案。本章的核心结论是：AI 时代最危险的能力退化并不一定首先表现为绩效下降；恰恰相反，它可能在绩效持续提高时发生，并因成功本身消除了暴露退化的证据。因而，未来学校、组织与个人若只用最终产出来评价能力，不仅可能测错能力，还可能通过评价制度主动加速能力不可见与能力空洞化。

关键词：生成式人工智能；人类能力；绩效；认知卸载；测量污染；潜在变量；成功陷阱；能力可见性；跨学科对撞

问题不是“AI 让人变强还是变弱”，而是“我们还能不能看见人的能力”

关于生成式 AI 与人类能力的争论，最容易落入一个二元框架：AI 究竟在“增强”人，还是在“削弱”人。增强论会列举更快的写作、更高质量的文本、更快的代码生成、更高的客户服务效率与更好的决策辅助；削弱论则会担心人类减少练习、形成依赖、失去批判思考与独立接管能力。这两种判断都抓住了部分事实，却共享一个未被充分审视的前提：我们似乎仍然可以从一个人在 AI 环境中的任务表现，直接判断这个人的能力变化。

问题恰恰出在这里。假设一名写作者在 AI 帮助下，文章质量从 70 分稳定提升到 90 分；一名程序员的代码错误率持续下降；一名学生的作业正确率上升；一名管理者的方案分析越来越完整。我们很自然地会说，他们“做得更好了”。但“做得更好”并不等于“人的潜在能力更强”。最终结果可能来自人的成长，也可能来自 AI 补偿人的缺口；可能两者同时增长，也可能人的能力下降而 AI 补偿增长得更快。只要后者的增量大于前者的损失，外部观察者就会看到一个持续变好的绩效曲线。

这使 AI 时代出现一个不同于传统自动化的新难题：工具不只是替代固定动作，而是能主动修正语言、补全逻辑、发现错误、给出下一步、生成代码、比较方案、提醒遗漏。过去，人类能力不足往往通过失败直接暴露——写不出来、算错、调试失败、搜不到资料、论证断裂、判断失误。生成式 AI 最有价值的功能之一，恰恰是让这些失败不再抵达最终环境。于是，AI 提升结果的同时，也可能切断“能力不足→失败信号→被发现”的测量通路。

因此，本章不把“能力衰退”预设为必然结论。我们提出一个更基础的问题：当 AI 成为持续在线的代理以后，个体能力是否进入一种低可识别状态？换句话说，能力可能上升、保持或下降，但系统仅凭日常绩效越来越无法区分这三种状态。这就是本章所称的“能力可见性下降”。它与能力下降不是同一个概念，却可能成为更危险的前置条件，因为只有当能力变化能够被发现，个体与组织才有机会进行纠正。

经验背景：短期绩效增益已经被观察到，但绩效与学习并不自动同向

生成式 AI 提高即时任务表现已经得到多类实证支持。Noy 与 Zhang（2023）在职业写作任务中发现，ChatGPT 显著降低完成时间并提高输出质量；Peng 等（2023）的 GitHub Copilot 实验显示，使用 AI 的开发者在特定编程任务上完成得更快；Brynjolfsson、Li 与 Raymond（2025）在客户服务场景中报告，AI 助手提高了单位时间内解决问题的数量，而且收益在经验较少的员工中尤其明显。这些研究说明，对许多知识工作而言，AI 的确能够改变可观察绩效的生产函数。

但“表现提高”与“能力形成”之间已经开始出现分离证据。Yan 等（2025）明确提出，在生成式 AI 环境中必须区分即时表现与学习，因为外部支持可以提高执行期表现，却未必形成可保持、可迁移的内部变化。Shen 与 Tamkin（2026）在学习新编程库的随机实验中发现，不同 AI 使用方式对技能形成影响不同；完全把编码任务委托给 AI 的参与者可以得到一定的生产率收益，却在概念理解、代码阅读和调试能力上表现更差。Liu 等（2026）进一步在随机实验中发现，AI 帮助能够提高当下正确率，但移除 AI 以后，一些参与者的独立表现与坚持性反而低于未使用 AI 的对照组。

这些结果并不能简单推出“AI 一定令人退化”。有些交互方式可能同时促进任务完成与学习，某些任务本来也无需保存全部旧技能，而且 AI 释放的认知资源可能被重新用于更高层问题。真正重要的是：短期绩效本身已不能作为能力增长的充分证据。一个测量系统若仍把 AI 增强后的完成质量当成人的能力分数，就会把需要解释的混合结果当成已经测量清楚的个体属性。

这正是本章选择心理测量学作为第二源学科的原因。行为经济学可以解释为什么人愿意外包，组织学习可以解释为什么成功路径被强化，但只有测量理论能够告诉我们：为什么“连续成功”在 AI 环境中可能越来越不提供关于个人能力的有效信息。

研究方法：三门学科必须围绕同一个“信号消失”问题相互否定

本章的跨学科对撞不是把“即时收益”“测量污染”“成功陷阱”三个概念并列后得出“AI 有利有弊”。第一层对撞要求三个学科分别回答三个不同但互相嵌套的问题：行为经济学回答每一次外包为何在局部上如此容易被选择；心理测量学回答外包以后为什么绩效不再能识别人类能力；组织学习回答为什么已经取得的绩效成功会进一步压缩独立练习与能力检验。

第二层对撞则追问：如果“能力变化本身”与“能力变化的可观察性”同时受 AI 影响，会发生什么？这里会出现一个不同于普通技能退化的新机制。普通技能退化至少还能通过错误率上升、速度下降、结果恶化被及时发现；AI 环境中的技能退化却可能被 AI 补偿层持续遮蔽。也就是说，干预变量同时进入了能力生成过程与能力测量过程。结果越被修好，测量越缺乏负反馈；越缺乏负反馈，越没有理由降低外包；越外包，越少产生新的能力证据。

本章将这一结构称为“绩能遮蔽循环”。这里的“绩”是可观察任务绩效，“能”是目标个体的潜在能力，“遮蔽”不是说能力一定下降，而是说绩效对能力的灵敏度降低。本章希望建立的典范，不是反对 AI 协作，而是为 AI 时代区分两个评价对象：如果目标是评估“人机系统能不能完成任务”，最终产出当然是合理指标；如果

目标是判断“这个人还能不能独立接管、迁移和处理新情境”，就必须引入不同的测量设计。

源学科一：行为经济学——为什么每一次多外包一点都显得合理

即时收益是确定的，能力代价是延迟、概率化且不可见的

对一个正在赶任务的人而言，使用 AI 的收益极其具体：十分钟变成三分钟，空白页变成初稿，调试卡点变成可运行代码，几十个网页变成结构化摘要。省下来的时间和认知努力当场可以感受到。相比之下，能力衰退若存在，往往不是今天立刻发生一个明确损失，而是以较低概率出现在未来某次独立考试、系统中断、模型越界、异常处理或岗位变化中。

行为经济学中的跨期选择研究早已说明，人对即时与延迟后果并不对称赋值。Laibson（1997）的双曲贴现模型以及 McClure 等（2004）关于即时和延迟奖励的研究，都揭示了近端收益在决策中的特殊权重。放到 AI 外包中，可以把一次外包的主观净收益粗略写成：$\Delta U_t = G_t + S_t - \beta\delta^k L_{t+k}$。其中 G_t 是即时绩效增益，S_t 是即时节省的努力与时间，L 是未来潜在能力损失，δ 表示常规时间折扣，β 表示对未来成本的额外贬值。只要即时项清晰、未来项模糊，更多外包就很容易成为每一时刻的局部最优。

更关键的是，AI 还降低了未来损失被主观估计的概率。如果一个人从未在日常工作中因为能力下降而失败，那么“未来可能接不住”只是抽象风险，而“现在少写半小时”却是确切收益。能力不可见性因此不是行为经济学模型外部的背景，而会反过来降低未来损失在当前选择中的主观权重。

认知努力本身有价格：外包不是懒惰，而是可预测的努力规避

Westbrook、Kester 与 Braver（2013）使用认知努力折扣任务显示，更高的认知负荷会降低奖励的主观价值，即人会把认知努力作为一种真实成本。这个结论对理解 AI 极其重要。很多关于 AI 依赖的讨论把“为什么不自己想一下”解释为意志薄弱，但从行为选择看，如果两个路径都能完成任务，其中一个要求高认知努力而另一个能即时降低努力，选择 AI 并不神秘。

生成式 AI 尤其特别，因为它不是只替代一项费力步骤，而可以沿着任务链连续降低成本：不会起头时给框架，框架出来后写初稿，初稿之后润色，事实不确定时查找，反方不清楚时补论证。原本一个高努力任务被切成多个可以再次外包的小节点。于是，

人并不是一次性决定“我要把能力交给 AI”，而是在几十个局部节点上不断选择一个即时成本更低的选项。

这种微观选择累积以后形成宏观效果。每一次选择单独看都可能非常合理，甚至没有任何可见损害；但如果某类能力的维持需要持续调用、错误修正与困难处理，那么连续的努力规避就会改变练习剂量。行为经济学因此解释了一个核心悖论：能力空洞化不需要一次重大的错误决定，它可以由大量短期上都正确的选择累积而成。

当绩效被奖励、能力来源不被奖励，外包还会获得制度性补贴

个人的跨期偏好并不是唯一机制。学校、企业与平台通常奖励可观察结果：文章是否按时交付、代码是否通过测试、工单是否解决、报表是否完成、销量是否提高。很少有系统在每次成功后继续追问：“这个结果中有多少来自你仍然保有的能力？”只要评价函数主要依赖最终绩效 Y，个体就会面对一个清晰激励：提高 Y 比保护不可见的潜在能力 θ 更容易得到即时回报。

因此，长期能力可能具有外部性。组织今天享受员工借助 AI 提高的产出，但能力下降的损失可能由未来岗位、下一任管理者、事故发生时的团队甚至员工本人承担；学生今天得到更高作业质量，真正的独立能力成本可能在几个月后的考试或工作中出现。收益与成本跨主体、跨时间分布，使“少外包一点以保护能力”进一步缺乏现实激励。

源学科二：心理测量学——为什么 AI 会把“能力”变成越来越难识别的潜变量

观测绩效不是能力本身，能力本来就是潜变量

心理测量学最重要的提醒是：能力从来不是直接被看见的。我们看见的是答题、反应时间、错误类型、任务表现，再从这些观察推断一个不可直接观察的潜在构念 θ。一个有效测量要求观测指标对目标构念具有足够敏感性，并尽量控制与构念无关的影响。Messick 的效度理论以及后续关于构念无关变异的讨论都强调：如果分数受到目标能力之外因素的系统影响，就不能把分数变化直接解释为构念变化。

在 AI 环境中，可以把日常绩效粗略写成 $Y_t=f(\theta_t,a_t,m_t,x_t)+\varepsilon_t$。θ 是人的潜在能力，a 是 AI 介入强度，m 是人与 AI 的匹配、提示和监督能力，x 是任务难度与环境条件。只要 a 和 m 显著进入函数，Y 就不再是 θ 的简单代理。尤其当 AI 能力不断提高时，同一个 θ 甚至更低的 θ 也可能产生更高 Y。

因此，AI 不是普通噪声。普通随机噪声会让测量变得不精确，但大量样本可能平均掉；AI 是方向性的、适应性的补偿因素，它往往专门在人的困难处介入：不会写时

帮写，出错时修正，检索失败时补齐。它对低能力或当前薄弱环节的补偿可能更强，因而会系统性压平能力差异在结果上的表现。

从“构念无关变异”到“构念无关容易”：AI 可能制造测量污染，也可能改变被测构念

测量理论中常讨论构念无关难度与构念无关容易：一个任务可能因为额外的语言负担、界面障碍而让分数低估目标能力，也可能因为与目标构念无关的线索、帮助或策略而让分数虚高。若我们的目标构念是“个人独立写作能力”，而测量时允许 AI 重写段落，那么 AI 生成质量显然进入了分数；若目标构念本来就是“人机协作写作能力”，则同样的 AI 使用不再是污染，而成为构念的一部分。

这一区分极其关键。AI 时代不是所有辅助都应被排除，而是评价者必须先说清自己想测什么。如果企业只关心“这名员工在正常工具条件下能否交付”，那么 AI 增强绩效完全合理；如果岗位还要求在系统异常时独立诊断、识别 AI 错误或迁移到新工具，那么仅测正常 AI 环境中的结果就产生构念代表不足：我们测到了稳定增强态的输出，却没有覆盖接管、迁移和异常处理这些目标能力的重要组成。

因此，AI 引发两类测量问题。第一类是测量污染：本来想测个人能力，却把外部代理能力一起计入；第二类是构念漂移：随着 AI 进入工作，社会真正需要的能力组合也发生变化。前者要求控制 AI，后者要求重新定义能力。二者不能混在一起，否则“AI 时代还需不需要独立能力”的争论永远没有清晰答案。

真正新的问题：AI 会降低绩效对能力的导数

本章进一步提出“能力可见性”概念。设 $E[Y|\theta,a]$表示给定人类能力与 AI 介入强度时的期望绩效，那么 $\partial E[Y]/\partial\theta$ 可以理解为绩效对人类能力变化的灵敏度。当没有辅助时，能力下降可能很快表现为速度下降、错误增加、质量降低；当 AI 持续补偿时，这个导数可能变小：人的能力变化一单位，最终结果只变化很少。

从测量信息角度看，如果不同 θ 水平越来越产生相似的高绩效分布，那么观测 Y 对 θ 提供的信息量就下降。极端情况下，无论一个人的真实能力是 60、70 还是 80，只要 AI 都能把最终产出拉到 90 附近，日常产出就失去区分这三种潜在状态的能力。这里并不是说 AI 让所有人一样强，而是说同一套观察指标对人的差异越来越不敏感。

这就是“能力可见性塌缩”的第一个严格含义：不是能力突然归零，而是测量函数斜率趋平。绩效持续很好，恰恰可能意味着系统更难从自然工作结果中获得关于 θ 的诊断信息。

错误被消失以后，能力测量失去了最有价值的负信号

传统能力识别高度依赖错误。学生某类题反复错，教师才知道概念没掌握；程序员调试失败，团队才知道某个知识缺口；医生遗漏异常，复盘才暴露模式识别问题。错误不是只有坏处，它还是测量信号。生成式 AI 的核心价值之一是实时吸收这些错误：语法错误在交付前被修正，代码错误在运行前被发现，遗漏证据在报告前被补齐。

如果所有错误都在到达环境之前被代理消除，外界看到的是一个“低错误率的人”，但无法区分他是因为能力高而少犯错，还是因为犯错很多但修复层足够强。这与信号检测中的观察通道改变类似：不是潜在状态变得更好，而是错误状态不再被外部传感器看到。

本章把这种机制称为“诊断信号消隐”。它比一般的测量污染更动态，因为 AI 会根据人的错误发生位置进行自适应补偿。最需要暴露的薄弱点，恰恰是最容易被工具优先修好的地方。于是越弱的能力节点越可能在日常结果中变得不可见。

源学科三：组织学习——为什么“成功”会让系统更不愿意检查自己是否正在失去能力

开发—探索张力：AI 增强把短期开发收益变得异常诱人

March（1991）提出的探索—开发框架指出，组织必须在探索新知识与开发已有能力之间分配稀缺资源。开发通常带来更近、更确定的回报；探索的收益更迟、更不确定，却关系长期适应。AI 进入知识工作后形成一种新的开发路径：把已知任务快速交给一个不断提高的外部系统，可以立即得到更高产出，而保留独立练习、训练新人、做无 AI 演练、处理边界案例则短期上看起来“效率更低”。

从管理视角，削减独立练习甚至可能被解释成优化：既然 AI 已经能写、能算、能搜、能调试，为什么还要花双倍时间让员工自己做一遍？这使能力维护天然处在探索一侧——成本当下发生，收益主要表现为未来适应性、接管能力和对异常的敏感。只要绩效评价周期较短，资源会自然向 AI 增强的开发活动倾斜。

成功陷阱：越成功的路径越获得资源，越缺乏替代路径的学习机会

Levitt 与 March（1988）讨论能力陷阱时指出，有利绩效会让组织对某种程序积累更多经验，从而继续强化该程序，即使存在更优路径也可能被锁定。AI 环境把这一机制推进了一步：成功不仅强化一条工作路径，还可能让组织停止生成关于“没有 AI 时我们还会不会”的信息。

设一个团队使用 AI 后交付速度提高 30%，错误率下降。管理层最自然的反应是扩大 AI 覆盖、缩短工期、降低人力冗余。独立处理比例随之下降。下一季度，由于 AI 覆盖更高，绩效可能继续上升，于是组织获得更多“扩大 AI 是正确的”证据。但这条证据只说明增强态系统运行良好，并没有说明人的备用能力仍然存在。成功评价与能力评价发生了系统性错位。

更危险的是，组织会把没有事故解释为不需要冗余训练。无 AI 演练、双路径复核、独立代码审查会被视为低效率；而这些看似多余的活动恰恰是产生能力诊断信号的机制。于是，成功路径通过资源分配主动删除了检验自己的机会。

路径强化使“能力检查”本身越来越昂贵

当团队长期围绕 AI 重新设计流程后，独立检查不只是关掉一个按钮。资料可能只保存在 AI 工作流中，人员配置可能减少，任务时限可能按照 AI 速度设定，新员工可能从第一天就没有经历完整任务链。此时想重新测量独立能力，会产生显著机会成本：产量下降、时间延长、错误上升，甚至影响客户。

这形成一种路径依赖：越早把所有任务设计成“AI 始终在线”，未来越难创造干净的无辅助测量条件；越难测量，越无法证明是否需要维护能力；越无法证明，就越容易继续取消这些看似昂贵的能力维护环节。能力可见性因此不是一次测量问题，而是被组织流程长期塑造的制度属性。

第一层对撞：绩效不是能力的镜子，而是一个被 AI 重写的测量通道

三门学科第一次碰撞后，可以得到一个共同否定：不能再把“日常任务持续成功”直接解释为“人的能力持续存在”。行为经济学说明，持续使用 AI 本身是被即时收益驱动的选择；测量理论说明，AI 改变了从潜在能力到观测绩效的映射；组织学习说明，成功结果又会反过来增加 AI 使用、减少独立样本。

这意味着日常绩效数据存在内生性。我们不是在一个固定环境中观察人的能力，而是在一个会根据人的薄弱点主动补偿、并根据成功继续扩大补偿的环境中观察人。AI 介入强度 a_t 既受过去绩效影响，又影响当前绩效，还影响未来能力 θ_{t+1}与未来测量信息 I_{t+1}。因此，仅用 Y_t 的上升趋势推断 θ_t 上升，会产生结构性误判。

本章将这一现象称为“绩能解耦”：可观察绩效 Y 与潜在人类能力 θ 之间的相关关系随着 AI 介入而变弱，甚至可能出现方向相反。最典型的状态就是 Y 持续上升而 θ 缓慢下降。只要 AI 增益增长速度大于人的损失速度，外部系统就不会收到任何警报。

第二层对撞：AI 同时改变能力与能力的可识别性，形成“双重遮蔽”

如果 AI 只影响绩效而不影响人的能力，这只是一个测量校正问题：知道 AI 贡献以后把它控制掉即可。如果 AI 只影响能力但不影响测量，那么能力下降会通过绩效恶化及时暴露，组织仍有反馈机会。真正棘手的是二者同时发生：AI 减少某些能力的练习机会，同时又用自身能力补偿由此产生的缺口。

这就是第二层对撞得到的“双重遮蔽”。第一层遮蔽是生产遮蔽：AI 直接提高结果，覆盖人的能力损失。第二层遮蔽是信息遮蔽：由于错误和失败被提前修复，系统失去关于能力损失的诊断样本。第一层让问题暂时没有后果，第二层让系统不知道问题是否正在积累。

双重遮蔽会造成一个重要的动态不对称。能力增长通常需要真实困难、主动处理、反馈和修正；能力下降则可以在缺少使用时缓慢发生。而 AI 恰恰把困难与错误从工作流中移除。于是，维持能力所需的训练信号在减少，暴露能力衰退所需的测量信号也在减少。生成与测量同时失去输入。

理论产出一：提出“绩能遮蔽循环”（PCML）

本章把上述机制形式化为绩能遮蔽循环（Performance–Capability Masking Loop, PCML）。循环至少包含六个节点：① AI 介入带来即时绩效增益和努力节省；②个体与组织强化 AI 使用；③独立任务练习比例下降；④潜在能力的增长减慢或部分能力退化；⑤ AI 继续在输出层补偿缺口，使最终绩效保持或提高；⑥失败、耗时和错误等诊断信号减少，导致能力变化更难被识别。第⑥步又会降低减少 AI 使用、增加训练或安排独立测量的动机，于是返回第②步。

这个循环的关键不在任何一个节点都必然为负，而在正反馈关系。AI 可能一开始真正帮助学习，能力 θ 也可能上升；但只要随着熟练流程建立，独立练习持续下降，同时绩效对 θ 的敏感度持续下降，系统就会逐渐失去判断能力变化方向的证据。换句话说，PCML 首先预测的是“不可识别性”，其次才是在特定条件下预测“能力空洞化”。

能力空洞化可以定义为：一个人在常规 AI 增强环境中保持高水平任务绩效，但构成独立接管、错误诊断、迁移和重新学习所需的若干内部能力已低于其任务表现所暗示的水平。它不是“什么都不会”，而是外部表现与内部可用能力之间形成越来越大的空腔。

理论产出二：能力可见性塌缩的简化动态模型

五个核心变量

设 θ_t 表示时间 t 的人类潜在能力；a_t 表示 AI 介入强度；p_t 表示独立或高认知参与练习比例；Y_t 表示最终可观察绩效；I_t 表示从日常绩效中可以获得的关于 θ_t 的诊断信息。AI 介入同时进入三条路径：第一，提高 Y；第二，通过改变 p 影响 θ 的未来演化；第三，改变 Y 对 θ 的敏感度，从而影响 I。

可以用三个简式表达：$Y_t = F(\theta_t, a_t, m_t, x_t)$；$\theta_{t+1} = \theta_t + L(p_t, q_t) - D(1-p_t)$；$I_t \approx S(a_t) \cdot V(task_t)$，其中 L 表示有效练习带来的学习或维持，D 表示缺乏调用产生的遗忘/去技能化，S(a)表示随着补偿性 AI 介入提高而下降的能力敏感系数，V 表示任务本身提供的诊断价值。

如果 a 提高使 Y 上升，同时使 p 下降、S 下降，就可能出现 Y_t↑、θ_t↓、I_t↓三者同时发生。这里最值得注意的是第三个箭头。传统讨论通常只比较 Y 与 θ，本章增加 I 作为独立状态变量，因为系统能否及时发现 θ 变化，本身决定后续是否会调整 AI 使用与训练策略。

能力可见性不等于能力水平

一个能力很强的人也可能处于低可见状态：因为 AI 包办太多，我们无法从日常产出确认他还强不强；一个能力较弱的人也可能处于高可见状态：通过独立任务、边界任务和错误记录，我们清楚知道他的弱点在哪里。因此，能力水平 θ 与能力可见性 I 必须作为两条轴。

这一区分能够避免一个常见误区：要求更多独立测试并不是预设“大家已经退化”，而是承认当前数据无法识别。科学上最合理的反应不是直接断言下降，而是设计能提高 I 的测量。

四种状态：为什么“高绩效”内部至少隐藏四种完全不同的现实

这四种状态解释了为什么仅看 AI 时代的绩效提升无法回答“能力是否变强”。状态 A 与状态 C 在表面上甚至可以拥有同样漂亮的产出曲线：两个人都写得更好、做得更快、错误更少。区别只有在改变辅助条件、改变任务分布或延迟测试时才显现。

状态 D 尤其需要被单独承认。有些旧能力确实可能不再值得维护，例如已经完全制度化、低风险、可可靠自动验证的任务。如果社会决定把它正式交给系统，那么能力不可见不构成问题。问题出现在组织一方面按 D 的方式设计工作，另一方面在责任、招聘和风险管理上仍假定处于 A 或 B——既不测个人能力，又要求人在关键时刻突然接管。

三种 AI 时代特有的测量污染

代理贡献污染：分数包含了 AI 能力

如果目标是个人能力而评分来自 AI 增强后的最终产出，最直接的污染就是代理贡献。AI 模型版本更新甚至可能让同一个人在没有任何学习的情况下“分数上升”。因此，纵向绩效曲线必须同时记录工具条件，否则跨时间比较失去可解释性。

补偿性污染：AI 对弱点的帮助不是随机的

更复杂的是，AI 往往在困难位置提供更大帮助。能力越弱的部分，越可能请求更多提示、修正与生成，导致最终结果出现非线性压平。这使简单减去“平均 AI 增益”也不够，因为 AI 贡献与 θ 本身相关。

选择性缺失：最有诊断价值的失败样本不再被观测

第三类污染是失败数据的选择性缺失。系统看到的是被修复后的结果，没有看到原始错误、第一次尝试、犹豫与修正路径。缺失并非随机，而是越需要帮助的地方越容易被隐藏。能力测量因此不仅被混入额外因素，还丢失了最关键的负样本。

为什么“连续几个月没有出事”可能反而是最弱的能力证据

在传统环境中，如果一个人连续数月独立完成高难度任务且错误很少，这确实是能力仍然存在的重要证据。AI 环境改变了这个推断，因为“没有出事”可能有两种生成机制：第一，人仍有高能力，所以没有错误；第二，人已经出现大量局部错误，但 AI 在进入最终结果前把它们全部吸收。只观察最终结果无法区分二者。

更进一步，系统越可靠，第二种机制越可能长期不被区分。若 AI 只有 70%可靠，人还会频繁被迫检查、纠错和接管；当 AI 达到 95%、99%甚至更高，人工参与困难环节的频率下降，能力测量机会反而减少。这形成一个悖论：代理越可靠，常规环境越安全，但人的备用能力状态越难从自然数据中推断。

因此，“零事故”不应自动被解释为“人机系统与人的能力都稳定”。它只直接证明在当前工具、当前流程、当前任务分布下，整个系统没有产生可见失败。要从系统成功推断人的备用能力，必须额外观察辅助被扰动时的表现。

为什么个人会主动参与这一循环：能力损失没有即时账单

PCML 不是一个纯粹由组织强迫形成的陷阱。个体也有充分动机进入。每天的时间压力、疲劳、任务量和竞争都会让 AI 外包具有真实价值。一个人即使完全理解“长期可能退化”，也会面临典型的跨期不一致：今天少思考二十分钟的收益由今天的自己获得，未来某次独立接管失败的成本由未来的自己承担。

而能力可见性下降进一步削弱了自我约束。人类通常依靠反馈调整行为：跑步速度下降会知道体能变差，考试错题增加会知道知识薄弱。AI 若把这些反馈修平，人会失去主观的退化体验。没有疼痛、没有失败、没有绩效下降，就很难形成“我现在必须练一下”的紧迫感。

因此，AI 时代能力维护不能只依赖自觉。行为经济学告诉我们，若维护收益长期、抽象而外包收益即时、确定，个体会系统性低估维护。需要通过环境设计把未来风险转化为现在可见的诊断反馈，例如周期性撤援、边界案例或接管演练。

为什么组织会参与这一循环：成功指标本身在奖励不可见性

组织通常不会故意追求能力退化。问题在于绩效管理系统可能把每一个局部选择都推向同一方向。管理者看到的是交付速度、成本、错误率和客户满意度；能力冗余、独立接管与未来学习能力难以量化。于是，减少重复劳动、提高 AI 覆盖率、缩短培训周期都能迅速反映在 KPI 上，而能力可见性下降没有对应的负指标。

更极端地说，独立能力测试本身会制造“绩效变差”。让员工关掉 AI 完成同一任务，速度必然下降；让学生先独立写再用 AI 修改，作业周期变长；让程序员人工诊断边界错误，会减少短期代码产量。如果组织只看短期 Y，这些用于保护 θ 和 I 的活动会被评价系统判定为低效率。

因此，PCML 不仅是技术效应，而是指标治理问题。一个只奖励增强态产出的组织，会把能力维护视为成本，把测量能力的过程视为摩擦。长此以往，组织会拥有越来越漂亮的产出指标，却越来越少知道这些产出在什么程度上仍可由人接管。

从“成功陷阱”到“成功遮蔽陷阱”：本章对组织学习理论的推进

传统成功陷阱强调：过去成功使组织过度开发熟悉路径，忽视探索，最终在环境变化时适应不足。AI 时代新增了一个信息层：成功路径不仅排挤探索，还可能改变组织用来判断自身能力的传感器。换言之，组织不是只“学错了”，而是逐渐失去知道自己是否还会的办法。

本章因此提出“成功遮蔽陷阱”：一种组织状态，在其中，代理增强带来的持续成功一方面提高对该路径的投入，另一方面压缩产生独立能力证据的任务机会，从而使内部能力下降即使发生也越来越难进入组织反馈。与传统成功陷阱相比，它的危险在于危机前可能没有明显绩效预警。

当环境突变、AI 服务中断、模型遭遇分布外案例或监管要求人工复核时，过去被遮蔽的能力差距会一次性暴露。此时组织常把事件解释为“某个人突然失误”，但从 PCML 看，失败可能是长期信号消隐后的集中显现。

可证伪命题：一个新典范必须能够被现实推翻

命题一：AI 介入强度提高时，最终绩效与独立能力之间的相关应在部分任务上下降。

在控制任务难度后，如果 AI 把绩效对人类能力的敏感度压低，那么高辅助条件下 Y 对无辅助 θ 测试的预测力应低于低辅助条件。若长期不存在这一现象，能力可见性塌缩的普遍性就需要修正。

命题二：补偿性越强的 AI，能力差异在最终产出中的压缩越明显。

如果 AI 对低能力者或薄弱环节帮助更大，那么增强态绩效分布会出现能力差距收窄；但撤援后差距重新扩大。

命题三：只奖励最终产出的环境会产生更高外包比例与更低独立练习比例。

若加入对独立接管、解释、迁移的评价，外包强度与练习分配应发生变化。

命题四：长期没有失败并不能预测撤援表现，除非工作流持续保留高诊断任务。

如果常规成功真能充分证明能力，则撤援/扰动任务不应提供大量新增信息；若新增信息显著，则说明日常绩效确实低可识别。

命题五：能力可见性下降将中介“AI 成功→继续外包”的强化关系。

当组织能够获得清晰的独立能力指标时，即使 AI 绩效很高，也更可能主动保留训练；当能力状态不可见时，成功更容易被解释为无需维护人类能力。

识别方案：AI 时代不能只做“无 AI 考试”，而需要四层测量

第一层：锚定任务——固定 AI 条件，建立跨时间可比的能力坐标

对需要长期监测的核心能力，应保留少量条件稳定的锚定任务。锚定不是要求日常工作脱离 AI，而是在固定辅助强度、固定任务难度和明确评分维度下重复测量，使不同时间点的分数仍可比较。若 AI 模型升级，必须记录版本并重新校准，否则工具进步会被误记为人的进步。

第二层：扰动任务——改变一个关键条件，看人是否能处理模型没有吸收的变化

能力真正有价值的部分常在新情境中显现。可以改变约束、加入冲突证据、设置异常样本，让 AI 的常规答案不足以直接完成任务。扰动任务能提高对迁移、诊断和判断能力的测量灵敏度。

第三层：撤援任务——短时移除关键辅助，测量最低独立接管能力

撤援不是要证明“人应该永远不用 AI”，而是像备用系统压力测试一样，用极少比例的任务确认最小接管能力仍在。测量重点可以是错误识别、问题分解、基本推理与恢复路径，而不必要求人达到 AI 增强态的同样速度。

第四层：延迟迁移——把“当时会做”与“真正形成能力”分开

如果目标涉及学习与能力形成，必须在 AI 帮助结束后设置延迟测试和新任务迁移。即时产出只能证明当时人机系统完成了任务；延迟迁移才能提供对 θ 变化更直接的证据。这一点与 Yan 等关于区分 performance 与 learning 的倡议一致。

为什么“抽查几个无 AI 任务”仍然可能不够

能力可见性问题容易诱发一个简单解决方案：每个月随机抽一次无 AI 考试。但如果抽查任务与真实薄弱点不匹配，或员工能够为固定测试专门准备，测量仍可能失真。真正需要的是与任务风险结构匹配的诊断设计。

心理测量学提醒我们，测量必须覆盖目标构念。若真正担心的是“AI 错了以后能否看出来”，测试就应包含有说服力但错误的 AI 答案；若担心的是“AI 不可用能否完成”，才需要撤援；若担心“进入新领域能否学习”，就应看迁移与新技能形成。不同能力不能用一个笼统的“关 AI 考试”代替。

因此，AI 时代能力测量的原则应从工具禁用转向识别性最大化：用最少的低效率样本，获得最多关于关键潜在能力的信息。这也回应效率问题——我们不需要让人周期性重做全部低价值劳动，只需要保留足以辨别 θ 状态的高信息任务。

教育推论：作业质量越高，越不能自动当作学生能力变高

学校是 PCML 最容易出现的场景之一。生成式 AI 可以即时提高作文结构、语言准确性、编程作业可运行率和资料综述完整度。如果教师继续以最终作品为主要能力证据，那么学生学会的可能首先是如何把结果做得更像高水平，而不是教师真正想测的写作、论证、代码理解或信息判断。

这并不意味着必须全面禁止 AI。更合理的做法是把“作品评价”与“能力评价”分开。作品可以允许 AI 并评价人机系统产出；能力则通过口头追问、理由解释、

错误诊断、条件改写、局部撤援和延迟迁移来测。两套分数甚至可以同时存在：一个学生可能有很高的 AI 协作产出能力，但独立论证还需要训练。

这样做还有一个重要教育意义：学生不必为了证明能力而重复机器已经擅长的全部低价值工作，却仍必须周期性暴露那些对未来接管和学习至关重要的核心能力。教育目标由“是否使用 AI”转向“哪些潜在构念必须仍然可测”。

组织推论：把“能力可见性”纳入 AI 部署的风险指标

企业评估 AI 项目通常关注效率、质量、成本、员工满意度与风险事件。PCML 建议增加一个新的治理维度：能力可见性。每一项 AI 部署都应问，随着自动化程度提高，我们还通过什么任务持续获得关于人类关键能力的证据？哪些失败信号正在被系统吸收？一旦 AI 不可用或进入边界情境，谁能接管，证据是什么？

这可以转化为“能力可见性预算”。组织不需要保留大量冗余人工工作，但应为高风险能力保留少量高信息的训练和测试窗口。例如每季度处理边界案例、保留双路径诊断、随机安排人工先判后看 AI、进行短时服务中断演练。它们的价值不是直接提高当期产量，而是维持 I_t 不降到无法识别的水平。

如果某项能力已经被决定永久退出人类职责，也应明确把目标构念从“个人独立能力”改为“系统可靠性”，并重新配置责任与冗余，而不是既不训练人又在危机时假定人会自动恢复旧技能。

个人推论：最危险的自我感觉不是“我不会了”，而是“我一直做得很好”

对于个人而言，PCML 揭示了一种非常反直觉的风险信号。传统能力下降会带来挫败，人会感到自己“变差了”；AI 遮蔽下，人可能持续收到完全相反的反馈：文章更漂亮、任务更快、错误更少、评价更高。主观自信甚至可能随产出一起上升。

因此，真正有用的自我检查不是问“我最近结果好不好”，而是问三个更具体的问题：如果今天 AI 拿走，我能保留哪些关键步骤？如果 AI 给出一个很像真的错误答案，我是否还能发现？如果任务条件变化，我是否能重新建立解决路径？这三问分别对应独立基线、错误监督和迁移能力。

个人也不需要为了维护能力而全面戒断 AI。可以采用最小剂量原则：在真正决定能力的少数节点上保留主动参与，例如先自己形成判断再看 AI、对关键代码先定位错误、对重要文章先列论证骨架、周期性做一次撤援任务。目标不是降低工具收益，而是让自身状态仍然有信号可看。

边界条件：什么时候绩效上升确实可以代表能力上升

本章并不主张 AI 增强后的高绩效永远与个人能力无关。第一，当人主动使用 AI 进行解释、反馈和纠错，并把释放的时间投入更高质量练习时，AI 可能同时提高 Y 与 θ。Shen 与 Tamkin（2026）观察到一些高认知参与的 AI 交互模式能够保留较好的学习结果，正说明“使用 AI”不是单一处理。

第二，当目标能力本身已经重新定义为人机协作能力时，AI 使用不再属于污染。例如未来程序员的核心构念可能包括问题定义、代理编排、代码审查、系统设计和异常诊断，而不是手写每一行代码。此时应开发新的测量任务，而不是执着于旧能力。

第三，如果工作流持续保存原始尝试、人工判断、修改轨迹和撤援样本，AI 即使高强度参与，能力可见性也未必下降。换言之，PCML 的关键调节变量不是 AI 使用本身，而是补偿强度、独立练习比例、诊断信号保存与评价制度。

与已有“去技能化”讨论的区别：本章关心的是退化之前的不可见阶段

自动化去技能化并不是新话题，从工业自动化到航空驾驶都长期存在“自动化越可靠，人工越少练习，异常时越难接管”的讨论。生成式 AI 延续了这一传统，但把问题推进到写作、分析、编码和判断等广泛认知任务。

本章的理论增量不在再次声明“AI 可能导致去技能化”，而在把测量可识别性放入动态机制。技能退化是 θ 的变化；能力遮蔽是 I 的变化。两者可能一起发生，也可能分开。只有把 I 单独建模，才能解释为什么一个系统可能长期没有任何绩效下降，却已经越来越无法确认人类备用能力。

这一区分还改变干预时点。如果等到独立能力已经明显下降再训练，可能太晚；更早的警报应是能力可见性开始下降——例如日常任务中已经没有任何独立样本、所有错误都由 AI 自动修复、员工从未处理边界案例。I 的下降可以成为 θ 下降之前的风险指标。

同样需要与本书第三章划清。第三章提出的能力证据债与本章的能力可见性塌缩描述的是同一片区域，但两章停在不同的位置上。第三章只主张系统失去了判断依据，不主张能力已经变化；本章则主张能力与能力的可见性会同时下降，并给出使两者互相强化的动力学。换言之，第三章是一个关于证据的命题，本章是一个关于因果的命题。

这个差别决定了两章的可证伪方式完全不同。要推翻第三章，需要证明辅助态表现仍然可靠地指示独立能力；要推翻本章，需要证明即使可见性下降，能力本身并不随之下降，也不会因此获得继续外包的额外动力。后者的推翻门槛更高，因为本章承担了更

强的主张——这也意味着本章更容易错。若长期的纵向研究显示可见性下降与能力变化之间没有稳定联系，本章应当退回到第三章的位置，而不是相反。

与“增强陷阱”模型的区别：从技能存量进一步推进到信息存量

2026 年的“增强陷阱”研究已经用动态模型讨论了 AI 即时生产率收益与长期技能侵蚀之间的权衡，指出即使决策者预见到技能损失，也可能因为前置生产率收益而选择 AI。本章与这一方向相容，但进一步增加了一个变量：关于技能状态的信息本身也会被 AI 改变。

也就是说，决策者不是在完全知道 θ_t 的情况下权衡生产率与技能。随着 AI 遮蔽增强，管理者和员工对 θ_t 的估计误差可能扩大，甚至形成系统性乐观。未来损失不仅被时间折扣，还被“不知道自己正在失去多少”进一步折扣。

因此，PCML 可以看成“生产率—技能权衡”与“测量—可识别性问题”的耦合版本。它解释了为什么现实组织可能比一个完全知情的动态优化模型更晚采取能力维护措施：不是只因为短视，而是因为成功路径降低了关于长期状态的可观测性。

进一步理论推论：AI 可能制造“能力保险悖论”

AI 在日常任务中像一种能力保险：当人的状态变差、疲劳、遗漏或知识不足时，它提供补偿，降低结果波动。保险本身具有巨大价值。但如果保险覆盖了所有小损失，个体就越来越少收到关于底层风险变化的信号。这类似某些系统中风险被中介层平滑后，表面波动降低而底层脆弱性累积。

由此出现能力保险悖论：越好的补偿机制越能保护短期产出，却越需要独立的监测机制来观察被保护对象本身。没有监测，保险从“保护能力的使用”滑向“替代能力的存在”，而使用者很难知道跨过这一边界的具体时点。

所以 AI 可靠性提升并不会自动消除人类能力治理问题，反而可能把问题从“日常错误管理”转向“低频高后果接管”。工具越强，维护全部旧技能的必要性可能下降，但识别哪些技能已安全退出、哪些仍承担接管功能的必要性上升。

实践框架：最低成本地维持“能力可见”而不是要求全面去 AI

基于 PCML，最合理的实践目标不是最大化独立工作比例，而是以最低成本维持关键能力的可识别性。第一步是定义：哪些能力即使 AI 长期在线也必须作为备用、监督或迁移能力保留；哪些能力可以正式退出。没有这个区分，所有“防退化”都会沦为形式主义。

第二步是为保留能力设计高信息密度任务。与其让员工每周完整重做十个 AI 任务，不如每月设置一个精心选择的边界案例；与其让学生所有作业都禁 AI，不如在关键概念上做短时口头解释和迁移测试。测量理论的目标是用尽可能少的样本区分潜在状态。

第三步是把测量结果真正连接到支持强度。如果撤援表现下降、错误识别能力变弱，就增加相应练习；如果稳定保持，则继续享受 AI 效率。这样，AI 帮助不是固定地越来越多，也不是为了原则周期性关闭，而是依据可观察能力状态动态调节。

研究限制与未来方向

PCML 目前是一套理论模型，仍需要长期纵向数据检验。现有大量研究持续时间较短，能够说明即时帮助与随后独立表现可能分离，却不足以证明数月或数年的能力轨迹。未来研究应同时记录 AI 使用细粒度日志、原始尝试、最终绩效与周期性独立测试，才能估计 θ、a、p、I 之间的动态关系。

第二，能力并非单维。AI 可能让低层记忆和程序性技能下降，却提高问题定义、评估和系统编排能力。因此未来研究不应只问“总能力升降”，而要识别能力结构重组，并明确哪些维度仍承担接管功能。

第三，测量本身可能改变行为。知道自己会被撤援测试的人会主动保留更多能力，这既是实验干扰，也可能正是现实治理需要的预承诺机制。未来可以比较透明测试、随机测试和基于风险触发测试对能力维持与生产率的综合影响。

可识别性深化：为什么“知道 AI 用了多少”仍不足以恢复人的真实能力

一种看似直接的解决思路是：既然最终绩效混入了 AI 贡献，只要记录 AI 使用比例，然后把这部分“扣掉”，不就能重新估计人的能力了吗？这种思路在 AI 作用线性、固定且与个人能力无关时可能成立，但生成式 AI 恰恰具有自适应性：不同人会在不同环节、不同困难程度上调用 AI，而 AI 对不同人的边际增益也不同。

例如两个写作者都把 30%的文本交给 AI，并不意味着 AI 对二人的补偿相同。甲可能只让 AI 润色措辞，核心论证自己完成；乙可能保留开头结尾，却把最需要推理的中段全部外包。表面“AI 使用率”相同，真正被替代的能力结构完全不同。类似地，两个程序员都接受十条代码建议，一人能够逐条理解与修改，另一人直接复制粘贴。数量指标不能恢复被隐藏的认知参与。

因此，AI 辅助下的能力识别存在典型的不可识别问题：如果我们只有最终 Y 和粗粒度 a，就可能存在多组不同的 θ、交互方式 m 与 AI 贡献函数产生同一个 Y。要恢复

θ，必须增加新的约束信息，例如第一次尝试、理由说明、修改轨迹、对错误 AI 建议的反应，以及撤援后的表现。换句话说，不是“测量得更多”就自动解决，而要测量那些能够区分不同生成机制的信息。

这也说明为什么能力可见性是一个结构属性，而不是简单的信息量属性。一万个被 AI 修好的日常成品，可能不如十个精心设计的边界任务更能说明人的能力。测量价值取决于样本能否让不同潜在能力状态产生可区分的反应，而不是取决于数据规模本身。

缺失机制深化：AI 制造的不是随机缺失，而是“困难处优先消失”

统计推断对缺失数据最敏感的问题之一，是缺失是否随机。如果一个人的一部分表现随机没有被记录，我们还可能通过剩余样本估计其能力；但如果恰恰是最差、最困难、最需要帮助的表现被系统性删除，剩余数据会形成乐观偏差。生成式 AI 的补偿逻辑与后者高度相似。

人在容易任务上往往不需要 AI，真正请求 AI 的时候常常是卡住、不会、没时间、害怕出错或任务复杂度突然上升的时候。也就是说，AI 介入概率本身与潜在困难相关。随后 AI 又把困难处的表现修正到可接受水平。于是评价者看到的“自然工作样本”不是人的随机能力切片，而是一个经过困难选择和补偿过滤后的样本。

本章把它称为“诊断性非随机缺失”：越有能力诊断价值的原始错误，越可能因为 AI 帮助而缺失。它解释了为什么简单扩大日常样本量不一定提高对个人能力的估计精度。如果新增的一万项记录全是已经经过自动纠错的最终成品，信息仍然可能高度冗余。

因此，高质量能力治理应主动保存少量“补偿前信息”。这并不意味着监控每个思考过程，而是在高风险任务中保留必要的初始判断、AI 介入点和关键修改，让组织知道成功结果究竟是怎样发生的。没有发生过程，成功本身很难承担诊断功能。

归因机制：为什么组织会把“系统成功”误记成“人的能力稳定”

绩能遮蔽不仅是测量问题，还涉及因果归因。一个任务由人、AI、数据、流程、团队与环境共同产生结果，但绩效管理往往需要把结果归到某个员工、学生或岗位上。只要结果持续优秀，人们就容易沿用旧的解释框架：这个人工作能力稳定、这个团队执行力强、这个学生掌握得不错。

问题在于，AI 进入以后生产函数已经改变，而评价语义可能没有同步改变。过去“90 分作文”确实主要来自学生写作能力；现在同样 90 分可能来自学生的选题判

断、AI 生成、人工筛选和修改的组合。若分数名称仍叫“写作能力”，系统就在没有意识到的情况下把人机产出重新归因为个体属性。

这种归因错位会进一步反作用于资源配置。管理者认为员工“已经会了”，就减少培训；教师认为学生“已经掌握”，就进入下一单元；个人认为自己“效率越来越高”，就更愿意把困难环节交给 AI。于是，一个最初只是标签错误的归因，会通过后续行动真实改变 θ。测量错误开始变成能力变化的原因。

这构成 PCML 的重要闭环：绩效不只是被动反映能力，它还通过评价、奖励和资源分配参与能力的未来生成。AI 时代的测量理论因此不能停留在“分数准不准”，还必须研究分数解释如何改变训练机会和外包路径。

时间尺度问题：为什么短实验常常看见收益，长周期才可能看见空洞

AI 研究中一个根本的方法学难题是时间尺度。短期实验天然容易捕捉即时生产率收益，因为 G_t 在第一次使用时就出现；技能维护与退化则依赖重复使用、练习分配和遗忘，需要更长时间才能稳定显现。于是，如果研究只持续几十分钟或几天，结论很可能主要描述 Y 的变化，而不是 θ 的长期轨迹。

这并不意味着短实验没有价值。Shen 与 Tamkin 以及 Liu 等研究的重要意义，正是在相对短时间内已经观察到技能形成、独立表现或坚持性可能与 AI 增强态绩效分离。但若要判断数月、数年的“能力空洞”，仍需要更长纵向设计。特别是经验丰富者可能拥有较大的能力存量，短期减少练习不会立刻掉出可用阈值，反而更容易经历一段“看起来完全没有问题”的潜伏期。

因此，PCML 预测一个可能的非线性轨迹：早期 Y 快速上升，θ 变化很小；中期 AI 工作流稳定后 p 持续下降，θ 开始缓慢变化，但 Y 仍然稳定；只有在撤援、任务越界或能力跌破某个接管阈值后，外部才看到突然的性能断裂。所谓“突然不会了”可能只是长期不可见变化跨过阈值后的第一次显现。

未来纵向研究因此应避免只比较 AI 组与非 AI 组的平均最终分数，而要绘制至少三条轨迹：增强态绩效、固定锚定条件下的能力表现、以及对异常和迁移任务的敏感性。只有三条曲线一起看，才能识别真正的增强、稳定外包和遮蔽退化。

能力阈值与系统风险：不是所有下降都重要，但跨过接管阈值会突然重要

能力维护还有一个常被忽视的问题：人并不需要在所有被 AI 替代的任务上维持原有峰值水平。如果 AI 长期可靠承担基础排版、常规检索或模板代码，让这些技能有所

下降可能完全可接受。真正需要定义的是最低接管阈值 θ_min：在 AI 失效、输出冲突或出现新情境时，人至少要保有多少能力才能恢复系统。

这使风险从“能力有没有下降”转换为“能力是否接近关键阈值”。一个人的某项技能从 95 降到 85 可能对系统没有实质影响；从 65 降到 55 却可能越过能否识别严重错误的边界。AI 越稳定，组织越少观察到这种接近阈值的过程，直到一次异常需要人工介入。

因此，能力可见性指标 I 的价值就在于提前知道距离阈值还有多远。高风险岗位不需要频繁证明自己达到无 AI 峰值，但需要足够证据证明关键能力仍在安全余量内。这样的设计比“所有人定期完整重做所有任务”成本低，也比“平时完全不测、事故时再看”更可靠。

进一步说，不同能力应有不同 θ_min。知识回忆可以允许大幅外包，但异常识别、关键安全判断、目标设定和问题重构可能需要更高的最低阈值。AI 时代能力治理的重点因此是阈值分层，而不是笼统地喊“保持人类能力”。

一个可执行的研究设计：怎样真正检验“绩效上升与能力可见性下降同时发生”

为了检验 PCML，可以设计至少六个月的纵向实验。首先对参与者进行无辅助基线测量，获得写作、编程、分析或决策领域的 θ_0；随后随机分配不同 AI 支持条件，例如低卸载组只允许解释与反馈，高卸载组允许直接生成与完成任务，对照组保持传统工具。日常阶段持续记录最终绩效 Y、AI 调用节点、原始尝试与时间投入。

每隔固定周期插入不预告的锚定任务与边界任务，估计 θ_t 与 I_t。关键不是只比较组间平均能力，而是检验三个路径：AI 强度 a 是否提高 Y；a 是否改变独立练习 p 并影响 θ 轨迹；a 是否降低 Y 对独立 θ 的预测力。如果第三条路径成立，就获得能力可见性下降的直接证据。

组织学习机制则可以通过奖励制度操纵进一步检验：一组只按最终产出奖励，另一组同时奖励独立接管和迁移表现。PCML 预测，前者会形成更高的外包强度、更低的诊断样本比例，并可能出现更明显的 Y—θ 解耦。这样可以把“个人懒惰”与“制度激励”区分开。

最后，在实验末端设置撤援与陌生情境迁移任务。若高卸载组日常绩效最高，但撤援表现下降，且日常 Y 在后期已经很难预测撤援成绩，就同时验证了“绩效增强”“能力变化”与“可见性塌缩”三个层次。这种设计比一次性的 AI 有无比较更接近本章真正要解释的动态过程。

指标替代：当“做得好”逐渐替代“是否会做”，评价系统会怎样改造能力本身

AI 时代一个更深的风险是指标替代。原本，最终产出之所以被用来评价能力，是因为在传统条件下二者高度相关：一个人长期写得好，大概率具有较强写作能力；长期调试成功，大概率理解代码。于是，绩效 Y 成为 θ 的代理指标。AI 介入后，这种代理关系减弱，但组织可能仍沿用旧指标。

一旦代理指标与目标构念脱钩，行为会围绕指标而非构念优化。学生优化的是“交出高分文章”，员工优化的是“在截止时间前给出高质量报告”，程序员优化的是“测试通过”。如果这些目标都能通过扩大 AI 介入更快实现，那么理性行动会自然朝 Y 最大化，而不是 θ 维持。评价制度并没有要求任何人故意放弃能力，却通过奖励函数让能力维护缺乏回报。

更重要的是，指标替代会改变下一代人的学习路径。经验丰富者至少曾经在没有 AI 的环境中形成过基础能力，而新人可能从一开始就在以增强态绩效作为唯一成功标准的系统中成长。他们不是“先学会再外包”，而是直接学习如何组织外部代理完成任务。若组织未来仍需要他们承担异常接管，却从未把接管能力写入评价函数，就会产生结构性培训缺口。

因此，AI 时代评价体系必须至少把“系统产出指标”与“关键人类能力指标”并列。前者可以继续追求极致效率，后者只需覆盖真正承担监督、接管和迁移功能的最低能力。只有评价函数承认二者不同，个体才不会被迫在每一个局部任务上用牺牲不可见能力来换取可见绩效。

能力可见性与公平：同样的高绩效可能掩盖完全不同的成长机会

能力可见性还关系到教育与组织中的公平。如果低经验者从 AI 获得更大的即时绩效增益，这可以缩小表面产出差距，是 AI 极有价值的一面；但如果这种补偿同时减少了他们获得困难练习和错误反馈的机会，表面平等可能与长期能力差距并存。

例如，新员工借助 AI 很快达到资深员工的报告质量，管理者因此缩短培训期。这在短期看是效率提升；但资深员工过去通过多年失败、复盘和边界任务形成的异常判断能力，并不会自动随高质量文本一起转移给新人。如果未来的评价继续只看最终报告，两类能力差异可能长期不可见。

这说明 AI 的“民主化能力”需要区分两层：一是让更多人获得高质量产出的机会，二是让更多人真正形成能够迁移和接管的能力。前者可以迅速实现，后者需要有意

识的认知参与与练习。若把前者当成后者，反而可能使最需要能力形成的人最早失去练习机会。

因此，公平的 AI 制度不应要求弱者少用 AI，而应确保 AI 带来的生产率收益不会自动取消其学习权。尤其对新人、学生和仍处于技能形成阶段的人，组织应把一部分 AI 节省出的时间重新投资到高价值练习，而不是全部转化为更多产量。

从“能力盘点”到“能力传感器”：未来组织真正缺少的不是更多培训，而是持续知道哪里正在变空

传统组织常用年度培训、资格证书或绩效考核来管理能力，这些方法隐含一个稳定假设：能力变化相对慢，工作本身会不断暴露问题。生成式 AI 打破了后一个条件。因为日常工作越来越能自动修复问题，组织需要的不只是周期性“盘点能力”，而是一套持续的能力传感器。

所谓能力传感器，不是监控员工的一切行为，而是提前选择少数关键观测点。例如，在代码工作中保存人工对 AI 错误建议的识别率；在分析任务中测量面对冲突证据时能否重构结论；在写作中观察是否能解释核心论证并处理反例；在决策中测量条件改变后能否主动改判。这些指标比最终成品更接近需要被保护的潜在能力。

能力传感器还有一个重要作用：把未来风险重新变成即时信号。行为经济学部分指出，外包之所以不断增加，是因为长期能力代价延迟且模糊。如果系统每隔一段时间给出清晰的“接管能力趋势”，未来成本就不再完全不可见，个体和组织可以在尚未发生事故时调整。

从这个意义上说，PCML 真正要求建立的不是反 AI 防线，而是反馈回路。AI 可以继续让绝大多数正常任务保持高效率；只要系统仍有少量传感器能穿过绩效遮蔽层，看到人的关键能力是否接近阈值，自我强化循环就可能被打断。

典范转换：从“以产出证明能力”转向“以扰动证明能力仍可调用”

在没有持续代理介入的时代，以长期产出来推断能力是一种成本很低、通常也相当有效的做法。一个人反复完成同类任务，本身就不断接受自然环境的检验。生成式 AI 出现后，这个默认典范正在失效：日常任务不再天然等于能力测试，任务完成过程可以被一个高度适应性的外部系统重构。

因此，AI 时代需要一次测量典范转换。评价重点不再是“这个人有没有持续交出好结果”这一单一事实，而是“当关键支持条件发生扰动时，他保留的能力能否被重新调用”。这里的扰动可以很小：换一个模型没有见过的约束、给出冲突资料、植入一个

高可信错误、暂时拿走某项功能。真正可用的能力应在这些变化中留下稳定而可辨认的响应。

这也改变了能力证明的逻辑。过去证明能力主要靠重复成功；未来，一部分关键能力可能要靠有限次数的成功接管来证明。因为正常态已经被 AI 充分保护，最有信息的证据反而来自边界态。组织无需人为制造大量失败，却要有意识地保留能够穿透代理补偿层的测试窗口。

如果这一典范成立，那么“能力维护”和“AI 效率”就不再是零和关系。日常 99%的任务可以充分利用 AI，剩余极少量任务承担诊断功能。重要的不是让人重新做完所有工作，而是确保系统始终存在几扇窗口，可以看见人在没有完全补偿时还能做什么、何时开始接不住，以及是否需要及时补充练习。

最后还需要强调，能力可见性本身也应被当作一种需要维护的公共资源。只要学校、企业和专业系统仍要求人在关键时刻承担责任，就不能把关于人的能力状态完全交给事故来揭示。事故是最昂贵的测量方式，也是最晚的测量方式。更合理的制度是在正常运行期间，以极低比例保留能够暴露能力边界的任务，使能力变化在尚可修复时就进入反馈。这样做并不是怀疑 AI，而是承认任何高度可靠的代理系统都会改变我们观察人的方式；代理越强，越需要重新设计传感器。

换言之，未来真正成熟的人机协作，不只是让结果持续变好，也要让关键的人类能力始终保持可被发现、可被检验、可被重新调用。

结论：AI 时代最危险的能力下降，可能发生在最好看的绩效曲线下面

本章试图回答一个 AI 时代越来越重要、却容易被最终产出遮住的问题：为什么一个人的写作、分析、代码、检索和决策结果持续变好，我们反而可能越来越不知道这个人的真实能力在上升、保持还是下降?

行为经济学告诉我们，AI 外包拥有即时、确定的收益，而能力代价延迟、模糊且被折扣；心理测量学告诉我们，AI 进入任务以后，最终绩效成为多源混合变量，且补偿越强，绩效对个人能力的敏感度越低；组织学习理论告诉我们，成功会把资源继续推向已经高效的 AI 路径，减少独立练习与探索，并让能力检验本身显得越来越昂贵。

三门学科跨学科对撞后产生的关键结论是：AI 可能同时改善结果、改变能力并删除能力变化的证据。由此形成“绩能遮蔽循环”——即时成功推动更多外包，更多外包减少独立练习，潜在能力可能下降；AI 继续补齐缺口，使绩效仍然上升；失败信号进一步消失，于是系统更没有理由怀疑能力状态。

因此，“连续成功”在 AI 时代必须被重新解释。它首先证明的是当前人机系统成功，而不是自动证明人的备用能力仍然存在。若我们真正关心的是人的独立接管、异常监督、迁移和重新学习，就必须创造专门的能力证据，而不能等待事故替我们完成测量。

最终，AI 时代的能力治理不应走向两个极端：既不能因为担心退化而要求人重复所有低价值劳动，也不能因为结果变好就宣布人的能力问题已经解决。更成熟的原则是：让 AI 尽可能承担可以安全外包的劳动，同时用少量、高信息密度的锚定、扰动、撤援和迁移任务，持续确认那些仍必须由人保有的能力没有从高绩效下面悄悄消失。真正值得警惕的，不是某一天我们突然发现“人做不好了”，而是此前很长时间里，一切都做得太好，以至于我们再也没有机会知道人究竟还会不会。

附录：PCML 的最低操作化清单

1. 先定义目标构念：究竟要测个人独立能力、人机协作能力，还是系统总体可靠性。不同目标不能共用同一分数。

2. 记录 AI 条件：模型、版本、工具权限、自动修正程度与人实际使用强度必须成为能力数据的一部分。

3. 保留原始轨迹：在高风险场景中保存人的初始判断、错误、修改与 AI 介入节点，避免只有最终成品。

4. 设置少量锚定任务：使用固定条件建立纵向能力坐标，防止工具升级被误判为个人进步。

5. 设置扰动与撤援样本：重点测异常识别、独立接管与问题重构，而非让人完整重复所有低价值劳动。

6. 对学习目标加入延迟迁移：AI 帮助后的即时高分不能单独作为能力形成证据。

7. 监测能力可见性 I：如果数月内没有任何高诊断样本，即使绩效完美，也应视为“能力状态未知”，而不是自动判定“能力稳定”。

8. 让结果反向调节支持强度：能力证据稳定时继续高效使用 AI；边界能力下降时增加针对性练习，形成动态而非禁用式治理。

本章附表

状态	可观察绩效 Y	个人能力 θ	可见性 I	解释
A 真增强	高/上升	高	高	AI 促进了人的能力且仍能被测量

B 稳定增强	高/稳定	高	中	人的能力保持，AI 主要扩大产出
C 遮蔽退化	高/上升	下降	低	AI 补偿能力损失，日常绩效无警报
D 系统替代	高	不再是目标	低	任务已正式转为人机系统能力，不再声称测个人能力

第二编　判断与理解

第一编处理的能力，允许长期不被调用而仍然存在。本编的两个对象不同：判断必须每一次重新形成，理解必须连同它的失效条件一起被持有。

第五章问的是一次具体决定是否由人形成。最终由人签字、点击、批准，在制度上足以确定责任人，却不足以说明判断在人这一侧发生过——这一章给出三个必须同时满足的节点，并明确把判断归属与判断质量分开：一个人可以真正作出一个糟糕的决定，归属并不以正确为前提。

第六章问的是理解算不算数。人工智能使总结、解释、举例、迁移都变得廉价，于是长期用来证明理解的那些外显证据同时失效。这一章的贡献是命名了一种此前没有名字、在人工智能时代尤其危险的状态：越界熟练——表面相似仍在、关键前提已变，主体却继续把旧结构用得很流畅。它通过所有的迁移测试，唯独通不过"该停的时候停下来"。

两章合起来给出一条对教育与评价最直接的推论：能用出去，不再是理解的最低证据；知道何时必须收回来，才是。

第 五 章

最后确认不等于判断归属

理由持有—失效识别—重新发起的最小判据　——决策科学 × 法理学 × 生态心理学的跨学科对撞

生成式 AI 介入下人类判断的最小判据

——决策科学（贝叶斯决策、期望效用与不确定性）×法理学（裁量、理由给付与责任归属）×生态心理学（可供性、情境知觉与知觉—行动耦合）的跨学科对撞

摘　要

当生成式人工智能能够大规模承担证据搜集、概率估计、风险比较、方案生成、反事实模拟乃至直接推荐“更优选择”以后，传统以“最后由人签字、点击或批准”为核心的人类监督标准面临一个基础性难题：最后动作由人执行，是否足以说明判断仍由人形成并可归属于人？本章以跨学科互否方法，将决策科学、法理学与生态心理学置于同一问题结构中。第一层对撞首先拆解三个学科对“判断”的不同最低要求：决策科学表明选择是事实信念、概率估计、价值排序与行动之间的条件函数；法理学表明决定权的形式保留不能替代理由给付、裁量保持与责任可追溯；生态心理学则表明判断并非仅是头脑内部命题的持有，而是主体对环境中与行动相关的信息保持敏感，并在情境变化时改变知觉—行动耦合。第二层对撞进一步发现，三门学科共同指向的不是“人是否亲自完成全部分析”，而是人是否仍占有判断的三个生成节点：其一，是否持有使当前选择成立的决定性理由；其二，是否能够识别使原判断失效的关键条件；其三，在条件变化后，是否仍能主动暂停、重构问题并重新发起改判。本章据此提出“RIR 人类判断归属最小判据”（Reason-possession—Invalidation-recognition—Re-initiation），并将判断归属与结果正确、制度责任、形式批准、解释能力以及对 AI 的依赖程度严格区分。本章进一步提出扰动测试、反事实边界测试和重新发起测试，用于识别人处于“AI 代判—人工执行”“AI 代判—人工解释”“人工监控但无改判权”还是“人类判断—AI 增强”四种状态。研究最后指出，AI 时代真正需要保护的并不是所有分析环节都重新内化到人的大脑，也不仅是“human in the loop”的形式位置，而是“human in the judgment”的结构位置：人必须能够持有决定性理由、感知判断边界，并保有重新打开判断过程的现实能力。

关键词：生成式人工智能；人类判断；判断归属；理由给付；裁量；可供性；适当依赖；人类监督；跨学科对撞

问题的真正变化：AI 不是只替人执行，而是在进入判断的生成层

关于人工智能与人的关系，过去最常见的问题是“机器是否应该替人做决定”。这个问题预设了一个相对清晰的分界：人负责判断，机器负责执行，或者机器提出建议，人保留最终决定权。但生成式 AI 正在破坏这条分界。它不仅执行既定规则，还能够主动整理证据、补充未知信息、生成比较维度、估计概率、构造风险情景、提出价值冲突、给出推荐，并把整条推理压缩为一个极具说服力的结论。于是，所谓“人最后决定”可能只剩下一个行为学意义上的末端动作，而判断本身早已在上游完成。

设想一名管理者面对三项投资方案。AI 读取财务报表、行业数据和政策信息，预测现金流，标注不确定性，比较失败概率，最后给出“选择 B”的建议。管理者阅读一页摘要后点击批准。制度上，这项决定属于管理者；操作上，最后动作也是管理者完成的；责任追究时，签字者同样是管理者。但如果他不知道 B 的优势依赖哪些事实，不知道 AI 把哪些风险看得更重，不知道哪一个假设一旦改变 B 就不再占优，也无法在市场条件突变时主动识别“原来的问题已经变了”，那么“最后点击”究竟证明了什么？它至多证明了执行权和形式批准权仍在人手中，却不能自动证明判断形成仍属于人。

这个区分在 AI 时代具有基础性意义。判断并不是一个孤立的终点动作，而是一个把事实、概率、价值、情境和行动组织起来的关系结构。一个人可能在没有独立计算的情况下真正拥有判断，也可能在亲手签字的情况下并没有形成判断。因此，本章不讨论“人是否必须比 AI 更会算”，也不主张回到所有能力都必须脱离工具的前 AI 时代。本章只追问一个更小、也更难的问题：当分析过程可以广泛外包时，一项判断仍然要满足哪些最低条件，才可以合理地归属于人？

这一问题不能由单一学科解决。决策科学可以描述理性选择的结构，却通常不直接回答责任与归属；法理学擅长讨论裁量、理由和责任，但仅凭可陈述理由又可能把熟练判断误解为语言复述；生态心理学强调主体与环境的实时耦合，却不天然提供规范意义上的责任标准。因此，真正的创新不在于把三套术语并列，而在于让三门学科互相暴露盲点，再从彼此不能单独推出的交叉处寻找“判断归属”的最小发生结构。

研究方法：不是“三学科拼盘”，而是围绕同一断裂点进行跨学科对撞

本章所说的跨学科对撞，不是从三个学科各取一个概念，再把它们并列为“多维解释”。第一层对撞的任务，是要求每一门学科独立回答同一个最小问题：在本学科内部，什么条件一旦缺失，就不能再把当前行为称为真正的判断？第二层对撞则进一步寻找三种答案之间的相互否定关系：哪一个学科认为充分的条件，会被另一个学科指出只是表象；哪一个学科强调的必要条件，又需要第三个学科才能获得动态性。只有通过这种相互消解，才能产出一个不属于任一原学科的共同新结构。

具体而言，决策科学首先把“判断”还原为条件化选择：行动不是凭空出现，而是由信息、概率与价值共同决定。它迫使我们看到，AI 如果改变了概率估计或价值函数，即使人最后仍在点击，行动生成结构也已经发生变化。法理学随后提出更尖锐的归属要求：一个决定之所以属于某个法定主体，不是因为他的名字出现在签字栏，而是因为他应当实际行使被授予的裁量、考虑相关理由、排除不相关理由，并能够对决定给出可审查的说明。生态心理学最后把两者从静态模型中拉回现实：人在开放环境中并不是每次先写出完整的概率表和理由清单再行动，成熟判断常体现为对关键环境信息的调谐，以及在信息变化时即时改变行动。

因此，本章的碰撞目标不是寻找“AI 使用越少，人类判断越强”的线性结论。恰恰相反，本章允许 AI 长期、深度、甚至永久地留在决策回路中。需要重新定义的，是人必须保留什么。这个问题将“工具使用程度”与“判断主体性”分离：一个高度依赖 AI 计算的人，仍可能是真正的判断主体；一个几乎从不使用 AI、却只机械服从上级规则的人，也未必拥有充分判断。这样才能避免把 AI 时代的人类主体性退化为一种技术禁欲主义。

源学科一：决策科学——判断首先是一种条件化的行动选择

贝叶斯决策揭示：结论依赖于信息如何改变信念

贝叶斯决策理论的核心贡献不是一句“根据概率做决定”，而是把判断写成一个可更新结构。主体面对证据 E，对世界状态 θ 形成先验信念 P(θ)，在获得证据后更新为 P(θ|E)，再结合不同后果的效用 U(a,θ)选择行动 a。简化表达为：a*=argmax_a E[U(a,θ)|E]。这个结构告诉我们，任何“最优选择”都不是绝对答案，它始终依赖信息集合、概率模型和效用排序。

这对 AI 时代有两个直接含义。第一，AI 能够代替人完成大量后验概率估计，但“计算由 AI 完成”本身并不取消人的判断。正如医生可以依赖化验、工程师可以依赖仿真、飞行员可以依赖仪表，人类判断本来就建立在外部测量和计算基础上。第二，真正危险的是人只接收 a*，却丢失 a*对 E、P 与 U 的依赖关系。此时，一个本质上条件化的结论被体验成了无条件答案。

例如 AI 说“方案 B 风险最低”。这个句子至少隐藏三层条件：哪些风险被纳入了模型；这些风险的概率如何估计；“风险最低”在多大程度上优先于收益、速度、公平或可逆性。如果人只记住 B，而不知道 B 是某组条件下的 B，那么他并没有真正占有这项判断的条件结构。

期望效用揭示：最优并不脱离价值排序

期望效用理论进一步排除了另一种常见幻觉：AI 似乎可以从“事实”直接推导“应该”。实际上，任何方案排序都至少隐含某种效用函数。安全、收益、自由、公平、速度、可逆性、声誉、长期学习价值等因素如何加权，并不能仅由更多数据自动消失。哪怕 AI 能够通过历史行为推断一个人的偏好，这也只是对既有选择模式的建模，而不是自动获得了对“此刻我愿意为什么承担代价”的规范授权。

因此，一项判断能否归属于人，并不要求他自己计算所有期望值，却要求最终使方案发生区分的价值冲突不能完全处于不可见状态。人可以说：“我接受 AI 对失败概率的估计，但这次我把不可逆损害置于平均收益之前，所以即使 B 的期望收益略低，我仍选 B。”在这种情况下，计算被外包，决定性取舍却仍被主体持有。相反，如果主体只是说“综合评分最高”，却不知道评分为何如此组合，那么所谓价值判断已经被封装进系统。

由此，决策科学给出本章的第一个源命题：判断归属不取决于主体是否亲自完成计算，而取决于主体是否仍能接触并承认使行动从备选项中被区分出来的关键条件。

不确定性与适当依赖：正确采纳不等于自主判断

近年来人机决策研究提出“适当依赖”（appropriate reliance），试图区分对 AI 的过度依赖、依赖不足和与 AI 能力相匹配的依赖。相关研究已经不再把“信任 AI”简单视为好或坏，而是关心依赖是否校准：AI 正确时是否能够利用它，AI 错误时是否能够识别并拒绝它。统计决策理论的进一步工作则提醒，遵循 AI 建议的概率与人自身区分信号、形成信念的能力不是同一变量。

但对于本章的问题，适当依赖仍不是充分判据。一个人可以因为训练有素或界面设计良好而在多数时候恰当地接受正确 AI、拒绝错误 AI，却仍不知道这次选择为什么属于自己。换言之，适当依赖主要评估人机团队的行为质量与信号使用，而“判断归属”追问的是主体是否占有使判断继续更新的生成位置。前者可能提高正确率，后者决定当情境超出训练分布、AI 突然不可用或目标本身发生变化时，人是否仍能接管。

因此必须区分三个维度：结果正确性回答“选得对不对”；依赖适当性回答“何时该听 AI”；判断归属回答“这项选择的理由、边界和重启权究竟还在谁那里”。三者可以相关，但不能互相替代。

源学科二：法理学——决定属于谁，不能只看谁最后签字

裁量不是一个按钮，而是被授予主体的判断空间

法理学长期面对“决定权能否委托”的问题。行政法中的裁量并不是简单拥有批准按钮，而是法规范在一定边界内把权衡、选择与理由评价交给特定决策者。现代行政组织当然允许大量事实调查、技术评估和辅助工作由他人完成，但这并不意味着法定决策者可以把整个裁量结构无条件交给外部系统。关于算法决策与行政委托的研究再次强调，非委托原则的价值之一正是保持决策责任、问责与决策位置的透明。

这与 AI 时代的结构高度相似：一个组织可以让 AI 做材料整理、风险预测、相似案例检索，甚至提供初步建议，但如果最终被授权的主体只是机械接受系统输出，那么“形式上未委托、实质上已让渡”的状态便可能出现。此时签字仍在人，裁量却已经空心化。

因此，法理学给出一个重要否定：不能从“最后执行动作属于人”反推出“判断形成属于人”。签字是责任链条中的证据，却不是判断发生的充分证据。

理由给付把“我决定”转换为“我为何这样决定”

理由给付是现代行政合法性的重要结构。它不仅服务于被决定者理解，也服务于司法审查、责任追踪和对任意性的限制。要求给出理由，意味着法秩序不满足于“有权者说了算”，而要求决定与一组可评价的事实和规范依据建立连接。

这为判断归属提供了第二个源命题：一项判断要归属于主体，主体原则上必须能够把当前选择与某些决定性理由连接起来。这里的“理由”不是要求复述 AI 的全部内部推理，更不是要求得到模型内部的完整可解释性。真正重要的是区分性：为什么是这个方案而不是另一个？哪一个事实、风险或价值冲突对你具有决定意义？如果去掉这个理由，你是否仍会作出相同选择？

这种要求也避免了“事后合理化”。一个人如果无论 AI 给 A 还是 B，都能在事后编出支持该结果的说法，那么他拥有的不是决定性理由，而是解释能力。判断归属需要的是能够约束选择的理由，而不是永远追随结果的叙事。

责任归属与判断归属必须分离

法律与组织实践还暴露出一种非常重要的结构性风险：责任可以留在人身上，而判断已经被系统接管。管理者可能必须为 AI 辅助的人事、信贷或医疗决定签字，但系统的关键变量、阈值和限制对他不可见；他甚至没有权修改输入、暂停流程或启动人工复审。此时，制度仍然把后果责任归给人，却没有给人足以形成判断的条件。

本章把这种状态称为“责任人工化—判断系统化”。它比完全自动决策更隐蔽，因为组织可以用“最终由人批准”来证明自己保留了人类监督，实际上却只保留了一个责任承接点。

欧盟《人工智能法》关于高风险 AI 的人类监督已经明显超越了“最后签字”逻辑：监督者应理解系统能力与局限，监测异常，意识到自动化偏差，正确解释输出，并能够在具体情境中不使用系统、忽略、推翻或逆转输出，必要时还能中断系统运行。这些要求说明，现代 AI 治理已经开始把“监督”理解为一种实际控制能力，而非形式在场。本章进一步追问：即使这些监督条件具备，什么时候才能说一项具体判断仍然属于人？

源学科三：生态心理学——判断不是只在头脑里持有理由，而是在情境变化中继续行动

可供性：环境不是中性数据，而是对行动者开放或关闭行动可能

生态心理学把认知从封闭的内部表征中拉回主体与环境的关系。可供性（affordance）不是物体自身的孤立属性，也不是主体主观想象出来的标签，而是环境条件相对于行动者能力所提供的行动可能。相同的台阶对成年人、幼儿、膝关节受伤者和轮椅使用者具有不同的行动意义。

这对 AI 判断有一个经常被忽视的启发：判断并不只是“把正确事实放进模型”。真实世界中的行动意义依赖主体、任务和环境的耦合。AI 可以告诉你坡度、距离、天气和平均速度，但“现在是否仍适合继续前进”需要把这些信息与当事人的体力、伤情、时间压力和可替代路线连接起来。一个判断者真正需要感知的，往往不是每一个变量，而是哪些变化正在改变行动可能。

因此，拥有判断并不要求主体能把所有变量用语言列出。一个经验丰富的护士、司机或维修人员可能无法把全部知觉线索逐条解释，却能在关键线索变化时立即发现“这个情况不对了”。如果我们只用“能否复述理由”作为判断归属标准，反而会错误排除大量真实专业判断。

知觉—行动耦合：判断能力在扰动中显现

生态心理学与生态动力学强调知觉和行动之间的连续耦合。熟练者的优势常常表现为对任务相关信息更加调谐，能够从环境中捕捉有效线索并及时改变行动。相关研究指出，真实世界的预判和决策依赖行动者有机会感知并作用于具有代表性的可供性，而不是只在脱离行动的静态信息上作答。

这意味着，判断归属最有力的证据并不一定是主体在稳定条件下能否解释现成答案，而是在条件被扰动时，他能否看到原来答案的失效。如果 AI 建议“继续”，而真实环境已经出现一个系统没有捕捉到的新限制，主体是否会感知到“继续”这一可供性

已经消失？如果会，他仍然处于判断回路；如果不会，他即使能够复述 AI 的理由，也可能只是一个高水平解释者。

生态心理学因此给出第三个源命题：判断的最低结构必须包含对失效条件的情境敏感性。判断不是“我知道为什么”，还包括“我看得见什么时候已经不能再这样”。

生态心理学对法理学的修正：理由不能退化为语言考试

法理学的理由给付很容易被误用为一种考试式标准：只要人能背出几个理由，就算保持了判断。生态心理学对此构成关键修正。真实判断中有大量知觉性、技能性和情境性知识不能完全被命题化；过度要求完整说明，可能把擅长语言的人误判为判断主体，把真正具有情境敏感性的熟练者误判为“说不清所以不会”。

因此，本章随后提出的“理由持有”必须采用功能性而非背诵性定义：主体是否能利用这些理由区分方案、对反事实变化作出方向正确的响应，而不是能否复述全部推理链。理由是否真正被持有，要在扰动和改判中检验。

第一层对撞：三门学科共同否定“最后动作判据”

把三门学科放在同一个问题上，会发现它们从不同方向同时否定一个看似自然的标准：最后是谁点击、签字、下单或执行，不能作为判断归属的充分判据。决策科学指出，行动只是条件函数的输出；法理学指出，签字可能与实质裁量分离；生态心理学指出，动作本身也可能是在缺乏真实情境调谐的情况下机械完成。

因此可以提出“最后动作谬误”（last-action fallacy）：把决策流程中最后一个可观察动作的主体，误认为整个判断形成过程的主体。这个谬误在传统流程中不一定明显，因为信息、分析和决定常集中在同一个人或同一团队内。但生成式 AI 把前端分析压缩得极其便捷以后，最后动作与判断形成发生结构性分离的概率显著上升。

进一步说，“人类在回路中”至少有三种不同含义：人作为执行器在回路中、人作为监督者在回路中、人作为判断主体在回路中。三者不能混为一谈。前两者可以通过界面设置轻易实现，第三者必须证明人仍占有某些不可由形式按钮替代的生成位置。

第二层对撞之一：从“能够解释”推进到“理由持有”

AI 生成解释并不自动把理由转移给人

生成式 AI 时代出现了一个新问题：解释本身也可以被 AI 生成。传统上，人能说出一个结构完整的理由，常被视为其理解和判断的证据；如今，一个人完全可以复制 AI

的解释，在语言上表现得比真正形成判断的人更加完整。因此，“能够输出理由”与“持有理由”必须分离。

本章将“理由持有”定义为：主体能够识别一组对当前选择具有实际区分作用的最小理由集 R_H，并且这些理由对其选择具有反事实约束力。所谓反事实约束，是指如果其中某个决定性理由被移除、反转或显著减弱，主体的选择倾向应相应发生变化，或者至少触发重新评估。

例如一个人说：“我选 B，因为 B 更安全。”仅凭这句话无法证明理由被持有。如果告诉他 B 的严重损失概率事实上高于 A，他仍然不假思索地坚持 B，只因为 AI 仍显示“推荐 B”，那么“安全”只是一个事后标签。反过来，如果他立即追问模型依据、重新比较风险并可能改选 A，就说明“安全”确实约束着他的判断。

理由持有不是要求主体掌握所有中间推导

必须强调，理由持有不是“完整内化”。如果把判断归属设定为必须独立重做全部检索、统计、计算和推理，AI 增强几乎失去意义，也违背人类长期借助工具、专家和组织分工进行判断的现实。医生不需要亲自制造 MRI 设备，投资者不需要手工计算每一项回归系数，工程师也不需要脱离仿真软件才能拥有工程判断。

真正的最低要求是压缩后的决定性结构仍能回到主体：哪些事实是关键前提，哪些不确定性足以左右选择，哪一个价值取舍决定了最终方向。我们可以把 AI 承担的部分称为“分析展开”，把人必须持有的部分称为“理由压缩”。高质量 AI 系统应帮助人把巨大分析空间压缩为可被主体真正占有的关键理由，而不是只输出一个漂亮的结论。

这里还需要排除一种误解：理由压缩不是把复杂问题粗暴简化成一句口号。真正有效的压缩必须保留对行动具有区分性的结构。例如“因为更好”没有区分力，“因为在当前不可逆损失权重高于短期收益的条件下，B 把最坏情形压到可承受范围内”则具有区分力。前者只是赞同，后者开始接近主体真正持有的判断理由。

第二层对撞之二：从“知道为何成立”推进到“知道何时失效”

仅有理由持有仍然不足。一个人可能真正理解当前为什么选择 B，却把这项判断永久化，不知道它依赖的事实和环境会发生变化。决策科学告诉我们后验信念随证据更新；生态心理学告诉我们行动可能随环境与能力关系改变；法理学也要求裁量不能被预先僵化到拒绝考虑新相关因素。三者在这里再次汇合：真正的判断必须具有失效结构。

本章把“失效条件”定义为：一组一旦发生足够变化，就会使当前理由—行动连接不再成立、需要重新评估的事实、概率、价值或情境变量。判断主体不需要列出世界上一切可能变化，但至少应对那些足以改变行动方向的高敏感变量保持可识别性。

例如，在贷款、医疗、教育、招聘、投资等场景中，AI 推荐往往建立在分布稳定、输入准确、目标函数不变和使用情境符合预设等条件上。当新的疾病症状出现、监管规则改变、候选人补充关键经历、家庭风险承受能力下降时，原答案可能失效。一个只记住“AI 推荐 B”的人，会把过去正确的结论带入新的世界；一个真正持有判断的人，则能看到“原问题已经不再是原问题”。

由此本章提出“失效识别”作为第二个最低判据。它不是要求主体预测所有黑天鹅，而是要求主体至少能够识别与当前判断直接相关的反事实边界，并在真实扰动越过边界时察觉到判断需要重开。

失效识别还把静态“理解”推进为动态“可更新”。在 AI 时代，真正危险的不是第一次判断错，而是一个原本正确的判断在条件已经变化后仍被无期限沿用。系统性能越高、历史正确率越高，人越容易把过去的可靠性当作未来的无条件保证。因此，判断主体必须保留的不只是对当前理由的理解，还包括对理由适用范围的持续监测。

第二层对撞之三：从“识别异常”推进到“能够重新发起改判”

即使一个人能识别原判断失效，也不意味着判断仍属于他。现实制度中经常出现这样的情况：操作者知道系统有问题，却没有权暂停流程；医生察觉模型输出异常，却只能在固定选项中选择；员工知道 AI 评分忽略了关键事实，但组织规定系统分数不可改。此时，认知上的敏感性存在，制度上的判断权却不存在。

因此必须加入第三个条件：重新发起。重新发起不是“人必须自己解决”，而是当失效信号出现时，主体有能力也有权限完成至少三个动作：停止把原建议当作有效前提；重新界定当前决策问题；调用新的证据、模型、人类专家或 AI 分析形成更新后的选择。AI 完全可以继续参与第二轮判断，甚至仍然承担绝大多数计算。真正不能丢失的是“这一轮已经结束，需要开新一轮”的发起权。

这一区分也揭示了“override”与“re-initiation”的不同。推翻 AI 输出仍可能只是一次否决动作，例如在模型推荐 A 时人工选 B；重新发起则要求主体能够重构为什么原问题已经变化、哪些输入或价值需要改变，并形成新的判断过程。前者是一票否决权，后者是判断循环的再启动权。

所以，AI 时代人类判断最深层的稀缺能力，也许不是“始终比 AI 知道得更多”，而是在答案看起来极其可信、流程又极其顺滑时，仍能够识别某个新事实使这条流程不再适用，并说出：停，这件事需要重新判断。

重新发起还有一个比技术权限更深的心理条件：主体必须允许自己把 AI 给出的高质量答案视为暂时性结论，而不是权威终点。如果使用者长期习惯“提问—得到答案—执行”，即使界面保留重开按钮，他也可能在实践中逐渐失去重新定义问题的倾向。因此，重新发起既是制度权利，也是需要被训练和维护的认知习惯。

理论产出：RIR 人类判断归属最小判据

三个必要节点

综合两阶碰撞，本章提出 RIR 最小判据。RIR 分别指 Reason-possession（理由持有）、Invalidation-recognition（失效识别）和 Re-initiation（改判重新发起）。它不是一个“优秀判断”的完整模型，而是一个更克制的归属模型：低于这三个节点，很难有充分理由把由 AI 深度参与的具体判断归属于人；达到这三个节点，则即使大量信息处理和计算由 AI 完成，判断仍可以在实质意义上归属于人。

第一，理由持有 R。主体能够指出使当前选择成立的决定性理由或最小理由集，并且这些理由对选择具有反事实约束力。第二，失效识别 I。主体能够识别至少一组会使原判断需要重新评估的关键变化，并对真实环境中的相关扰动保持敏感。第三，重新发起 G。主体在检测到失效后，有能力和现实权限暂停原判断的继续执行，重新界定问题、更新信息或价值，再形成选择。

可以把一个具体判断写成 J=<E,P,V,A,C>：E 表示被接受的关键证据，P 表示对相关状态或后果的概率判断，V 表示价值或目标排序，A 表示行动，C 表示当前情境与行动者条件。AI 可以生成或计算其中任意多个元素。但若要说 J 归属于主体 H，至少需要：H 能够持有使 A 成立的关键理由 R_H；H 能够识别一组使 R_H→A 关系失效的边界 B_H；并且当 B_H 被触发时，H 拥有重新打开 J 的能力 G_H。简化表示为：Attribution_H(J)=R_H∧I_H∧G_H。

这一表达式不是要把复杂人类主体性完全数学化，而是为了明确逻辑关系：三个节点是合取关系，不是加权平均。理由非常强却完全没有重启权，仍不足以构成完整归属；拥有停止按钮却不知道为什么要停，也不足；能够识别异常却无法把异常转化为新的判断过程，同样不足。RIR 强调的是一个能连续运转的最小闭环，而非若干优点的简单累积。

为什么 RIR 是“最小判据”而不是“充分的道德自主”

RIR 故意保持最小。它不要求主体一定正确，不要求主体拥有无限信息，不要求其价值观具有哲学上的最终正当性，也不要求其完全摆脱社会影响。现实人类判断从来不是绝对独立的：我们依赖语言、制度、专业知识、他人证词和技术工具。若把“没有外部影响”作为自主判断标准，自主判断本身几乎不存在。

RIR 只要求判断的关键可更新节点仍在人身上。一个人可以犯错，但他知道自己为何这样选；可以不知道许多事实，但知道哪些新事实足以使自己改判；可以继续依赖 AI，但在 AI 失灵或条件变化时知道要重开问题。这足以构成一种最低限度的判断主体性。

因此，RIR 与“独立决策”不同。它更适合描述 AI 成为长期认知基础设施后的现实：人不需要把整套外部能力重新搬回脑内，却必须保持对判断闭环的结构性占有。

RIR 的核心不是“人比 AI 强”，而是“人仍拥有条件控制”

传统的人机竞争叙事容易把主体性理解成能力比较：人必须在关键任务上比 AI 更懂，才能有资格最后决定。RIR 不接受这个前提。一个人可以在信息容量、计算速度和统计准确率上远弱于 AI，却仍然是判断主体，只要他能够决定哪些理由对自己具有规范意义、识别适用边界，并在边界变化后重新组织问题。

因此，RIR 保护的不是知识优势，而是一种最低反事实控制：如果决定性理由变化，我是否会重新评估；如果环境变化，我是否能看到；如果旧判断失效，我是否能重开。判断归属最终落在这种“可改变性”的控制上，而不是落在谁拥有更多信息。

五个必须分开的概念：否则“人类判断”会被表面指标吞没

判断归属 ≠ 结果正确

一个机械听从 AI 的人可能连续一百次得到正确结果，一个真正自主判断的人也可能因证据不足而选错。若把正确率当作判断归属标准，就会把高性能依赖误认为主体性。RIR 因此是过程归属指标，而非绩效指标。

判断归属 ≠ 最后批准

批准只说明某个动作或制度环节由人完成。若理由、失效边界和重启权均不在人，批准就可能只是“人工盖章”。

判断归属 ≠ 能够解释

解释可以由 AI 提供，也可以事后合理化。只有当理由对选择具有反事实约束，并能参与失效识别与改判时，解释才转化为理由持有。

判断归属 ≠ 对 AI 低依赖

一个人完全可以高度依赖 AI，同时保有判断；另一个人也可能不用 AI 却机械服从制度、权威或惯例。AI 使用频率不是主体性的单调函数。

判断归属 ≠ 责任归属

责任是规范制度分配的后果承受关系，判断归属则是判断发生结构的归属。两者可以一致，也可以严重脱耦。AI 治理尤其需要警惕“让人承担责任，却不给人判断条件”的设计。

这五个区分的意义在于为未来研究提供一个二维甚至多维测量框架。任何关于“人仍在做决定”的研究，都不应只报告人最终是否接受 AI 建议，还应分别测量结果质量、依赖校准、理由持有、边界识别、重启能力与责任配置。只有这样，才能避免把不同问题压缩为同一个“人是否采纳 AI”的行为指标。

四种人机判断状态：同样“人在回路”，结构可能完全不同

类型一：AI 代判—人工执行

AI 完成信息组织、理由筛选、价值排序和方案选择，人只执行最终动作。人不知道为什么此时选 A，也不知道何时 A 会失效，更不会主动重开问题。这是最典型的“最后动作谬误”场景。它可能拥有很高效率，也可能在模型分布稳定时拥有高正确率，但判断归属主要位于 AI—组织系统而非个人。

类型二：AI 代判—人工解释

人能够流畅复述 AI 生成的理由，甚至可以向他人解释模型结论，却缺乏对关键扰动的敏感性。这个状态比第一类更危险，因为语言能力制造了“我已经理解”的错觉。R 存在表象，I 与重新发起能力却很弱。

类型三：人工监控—无改判权

人拥有专业经验，能够发现异常，也知道系统为什么可能错，但制度只允许接受或备注，不允许真正暂停、重构问题或启动替代流程。此时主体拥有 R 和 I，却缺失 G。问题不是个人不会判断，而是组织剥夺了判断发生的入口。

类型四：人类判断—AI 增强

AI 可以承担绝大多数检索、计算、比较和模拟；人不必重复机器已经可靠完成的工作。但主体持有决定性理由，知道关键失效条件，并能够在条件变化时重新发起判断。此时 AI 不是判断主体的替代者，而是判断空间的扩展器。

从理论到测量：判断归属不能靠自述，应使用“扰动—边界—重启”测试

理由扰动测试

测试者不问“你为什么同意 AI”，而是改变一个被主体声称为决定性的理由。例如主体说自己因为安全性选择 B，那么就提供可靠新证据表明 B 的严重风险上升。若主体仍无条件维持选择，且不重新检查，那么原先的“安全理由”可能只是语言装饰。若主体主动重估，说明理由确实具有约束作用。

理由扰动测试的关键不是要求最后一定改选，而是观察理由变化是否真正进入判断过程。主体可能在重新评估后仍选 B，因为其他理由足以补偿；只要他能说明新的权衡结构，判断仍然可能属于他。

这一测试还可以区分“理由记忆”与“理由控制”。记住模型解释属于静态知识；在理由被扰动时重新组织选择，才显示该理由已经成为主体自己的控制变量。

失效边界测试

要求主体回答：“发生什么变化以后，你会停止沿用当前建议？”高质量回答不是枚举一切风险，而是指出与当前决策结构高度相关的少数敏感变量。例如“如果现金流缓冲低于六个月，我不会再以增长为首要目标”“如果出现新的神经系统症状，我会认为原有居家观察建议失效”。

随后可以把这些边界嵌入模拟场景，观察主体能否在不被明确提示“现在该改了”的情况下识别失效。这个测试比让人背诵模型局限更接近生态心理学意义上的情境调谐。

边界测试特别适合专业训练，因为专家能力往往不体现在常规案例上，而体现在他能否更早察觉“这个案例已经不是常规案例”。AI 可以比人更擅长常态分布内的模式识别，人类判断的保留则更应把训练资源投入边界、异常与目标冲突。

重新发起测试

当原判断明确失效后，不直接问主体“你选什么”，而问“你下一步如何重新开始？”如果回答只有“再问一次 AI”，则重新发起能力仍然很薄，因为主体没有识别

问题结构发生了什么变化。较强的重新发起表现为：先冻结原结论，指出新增事实改变了哪一个关键关系，重新定义目标或约束，再决定需要哪些新证据与工具。

因此，AI 可以继续作为第二轮判断的核心分析工具，但它被重新调用的方式已经不同：不是把旧问题重复提交，而是由主体完成问题重构后重新组织人机系统。

重新发起测试还应检验制度现实性。如果主体知道应该重开，却没有权限获得原始数据、联系专家或暂停自动流程，那么失败不应全部归因于个人能力，而应被识别为组织设计缺陷。RIR 同时是一套能力判据和权限判据。

跨场景检验：RIR 判据是否只适用于抽象决策?

医疗：医生不需要重做模型，但必须知道什么时候病人已不再属于模型的那个病例

在临床 AI 中，一个医生可以依赖影像模型、风险评分和药物相互作用系统。要求医生完全独立重算模型没有现实意义。但如果医生不知道某项推荐依赖患者年龄、肾功能或特定症状，也不知道新的生命体征变化意味着原建议失效，那么“医生最终签字”不能充分证明临床判断仍在人。

RIR 在这里意味着：医生持有本次治疗的决定性临床理由；能够识别至少一组需要升级检查或改变治疗的警戒变化；并且当这些变化发生时能够暂停自动路径、重开鉴别诊断。

教育：真正要训练的不是“无 AI 做题”，而是条件变化后的重新判断

如果学生借助 AI 完成资料检索和复杂计算，传统考试容易认为他“没有独立完成”。但从判断归属看，更重要的测试是：给他一个 AI 提供的高质量结论，然后改变一个前提、目标或反例，看他能否发现原答案失效并重新组织理由。

这意味着 AI 时代教育可以从“答案生产测试”转向“扰动响应测试”。学生不必在所有任务中被迫关闭 AI，但必须周期性证明自己能识别关键理由、边界和重启点。否则，他拥有的是外部系统的答案访问权，而不是自身判断能力。

管理：最危险的不是经理听 AI，而是组织只保留经理的签名

在企业管理中，AI 可能越来越多地介入招聘、绩效、预算、供应链和风险控制。若组织只规定“最终由经理确认”，但不给经理查看关键假设、对异常样本发起复审、调整价值权重或暂停系统流程的权限，那么人类监督只是责任分配机制，而非判断机制。

RIR 要求组织界面不仅显示一个推荐，还应暴露足以支持理由持有的信息；不仅显示置信度，还应呈现已知的失效边界和适用范围；不仅提供“接受/拒绝”，还要提供“重开问题”的路径。

个人生活：AI 可以帮你比较，却不能替你决定什么代价是你愿意承担的

旅行、消费、职业选择、婚姻、养老安排等个人决策往往没有统一的客观最优。AI 可以估计成本、概率和后果，但“什么损失对我不可接受”“什么价值值得为之承担代价”本身就是判断的一部分。若人只接受系统综合评分，价值排序就被悄悄外包。

RIR 并不要求人在每次日常选择中进行哲学反思，而只要求高影响决策中能够说清最关键的取舍、知道什么变化会改变选择，并保留重新考虑的入口。

对现有人类监督框架的推进：从 Human in the Loop 到 Human in the Judgment

现有 AI 治理已经越来越重视人类监督。NIST AI RMF 强调应明确区分人在 AI 系统使用、交互和管理中的角色与责任，并指出把复杂人类现象转换为数学量可能造成情境丢失。欧盟《人工智能法》第 14 条则明确要求高风险 AI 的监督人员能够理解系统能力和局限、监测异常、警惕自动化偏差、正确解释输出，并能够在具体情况下忽略、推翻、逆转输出或中断系统。

这些规范非常接近 RIR，却仍主要面向“有效监督”和风险控制。RIR 在此基础上增加一个归属性问题：一个监督者能够按下停止键，不代表他已经持有本次判断的理由；能够看到解释，不代表他知道结论的失效边界；能够推翻一次输出，也不代表他有能力重新构造下一轮判断。

因此，本章建议把 Human in the Loop 进一步分解为三个更可操作的位置：Human at the Reason Node（人在理由节点）、Human at the Boundary Node（人在边界节点）、Human at the Re-initiation Node（人在重启节点）。只有三个节点都保留，才能较强地说“Human in the Judgment”。

这个推进也避免一个常见治理陷阱：为了证明“有人监督”，在自动系统末端放置一个疲惫的人类审批者。大量重复确认会让人的注意力逐渐适应系统的高正确率，从而增加自动化偏差；一旦人既缺乏代表性练习，又缺乏真正推翻和重开的机会，其接管能力可能在长期使用中衰退。真正的监督设计必须让人周期性参与关键扰动和异常，而不仅是重复批准正常案例。

由此，“监督是否存在”的评价方式也应发生变化。传统审计可以记录每一项决定是否经过人工确认；RIR 视角下还应记录人工是否曾访问决定性理由、是否识别过模型边界、是否真实启动过重新评估流程。如果一个系统数年运行中人工从未重开过任何高风险判断，这未必说明系统完美，也可能说明人类监督已经退化为形式。

AI 产品设计推论：不应只优化“给出更好答案”，还要维护判断归属

从单一推荐界面转向“理由—边界—重启”界面

多数 AI 助手的默认交互目标是降低摩擦：用户问问题，系统尽快给出一个完整答案。对知识检索这通常有价值，但对高影响判断，过度顺滑可能恰恰隐藏判断结构。RIR 要求 AI 界面在必要时把三个节点显影：给出少量决定性理由，而不是堆砌解释；标明结论高度依赖的关键条件；允许用户一键进入“条件改变后重新判断”而不是只继续追问。

这并不意味着每个回答都增加大量警告。好的设计应根据风险、可逆性和不确定性调节摩擦。低风险、可逆任务可以高度自动化；高风险、不可逆且情境开放的任务，则应主动要求用户确认价值取舍和关键边界。

一个可行的界面结构可以是：首先呈现“当前推荐”，随后只显示两到三个决定性理由；再给出“最可能使本结论失效的条件”；最后提供“如果条件改变，重新定义问题”的入口。这样的设计把解释从说服性附件转化为判断闭环的一部分。

解释的目标应从“让我相信”改为“让我能够改判”

可解释 AI 经常被理解为增强用户信任，但研究已经显示，解释并不必然产生适当依赖，有些解释甚至会强化对 AI 建议的黏附。RIR 因此提出不同的解释评价标准：一段解释是否让用户更容易知道什么时候不能再用这个结论？是否帮助用户形成可操作的反事实边界？是否让用户在条件改变后能够重新打开问题？

如果解释只提高说服力，却降低质疑和改判，它可能提高接受率而损害判断归属。真正面向主体性的解释，应服务于可更新性而不是单纯信任。

因此，未来 XAI 评价不应只问用户“你是否理解模型”“你是否更信任模型”，还应加入行为性指标：当关键理由被反转时，用户能否停止原建议；当情境变量越过适用边界时，用户能否主动识别；当模型与现实冲突时，用户能否重构问题而不是只在原推荐附近微调。

保留必要的人类“异常肌肉”

长期自动化会产生技能退化和接管困难，这是航空、人因工程和自动化研究反复面对的问题。生成式 AI 把这种风险扩展到认知判断领域。若人在正常状态下永远不需要观察理由、边界和异常，只有系统突然失效时才要求人接管，那么所谓“最后负责的人”很可能已经失去实际接管能力。

因此，人机系统需要周期性的代表性扰动：让人处理一部分边界案例、模型冲突案例和条件突变案例。目的不是人为降低效率，而是维持 RIR 三个节点的可用性。可以把这理解为“判断接管训练”：不是训练人重做 AI 的所有日常工作，而是训练人在 AI 最容易越界的地方仍能重新成为判断主体。

这也为“哪些环节可以永久交给 AI，哪些环节必须周期性由人亲自完成”提供了答案。稳定、可验证、低价值冲突的分析环节可以永久外包；涉及边界检测、异常识别、价值重排和问题重构的环节则应周期性由人真实参与，否则 RIR 能力会因缺乏使用而退化。

组织制度推论：判断权与责任必须重新配对

RIR 最重要的制度意义之一，是揭示“责任与控制不对称”。如果一个员工要为 AI 辅助决定承担责任，却看不到关键理由、无法识别系统适用边界，也无权暂停或重开流程，那么将全部责任推给他并不合理。组织不能一方面把决定结构封装在技术系统中，另一方面又用“有人签字”把责任全部人工化。

因此，凡是要求人承担实质责任的制度，都应至少配套三种权限：理由访问权、边界告警与质疑权、重新发起与升级权。理由访问权不等于开放全部模型参数，而是让责任主体能理解影响当前决定的关键因素；边界权意味着他能识别并标记“此案可能已超出系统通常适用范围”；重新发起权意味着他能让流程暂时退出自动路径，进入复核、专家会诊或新模型分析。

反过来，如果组织明确决定某类任务可以完全自动化，也应诚实承认判断主体已经转移到组织—AI 系统，并建立相应的系统级责任，而不是保留一个象征性人工按钮来制造责任幻觉。

由此可以推出一个制度原则：责任强度与 RIR 权限应大体同向配置。组织要求个人承担的责任越重大，就越不能只给其形式批准权；如果个人没有真实的 RIR 权限，则相应责任应更多转移到系统设计者、部署组织和治理结构，而不能通过末端签字全部压回使用者。

AI 时代人类能力的新定义：不是“所有东西都要自己会”，而是保住判断闭环

RIR 模型也回应了一个更广泛的教育与能力问题：AI 成为长期认知基础设施以后，人是否仍应把所有能力重新内化到脑内？本章的答案是否定的。人类文明本来就通过文字、计算器、数据库、专业分工和组织程序外化认知。生成式 AI 只是把外化推进到更高层次。要求人重新独立完成每一个检索和计算，不仅成本高，也可能浪费 AI 带来的真正增益。

但这并不意味着“会调用 AI”本身就等于拥有能力。能力的重心需要发生位移：从生产全部中间步骤，转向持有关键理由；从记住大量稳定答案，转向识别答案的边界；从独立完成所有计算，转向能够在系统失效时重新组织人机资源。

这是一种“闭环能力”而不是“裸脑能力”。人的能力单位不再只是一个孤立大脑，而可以是“人+AI+外部工具”的系统；但系统中的人必须保留 RIR 节点，否则所谓人机整体能力可能只是机器能力被人借用，而没有形成主体可接管、可迁移、可负责的判断能力。

因此，未来评估一个人的专业成熟度，可能要增加三个问题：他能否压缩出决定性理由？他能否识别哪些扰动会改变判断？当工具的建议失效后，他能否重新组织问题？这三种能力可能比记住更多事实更能区分“会用 AI 的人”和“被 AI 带着走的人”。

这里也出现一个新的教育目标：不是让学生永远脱离 AI 证明自己，而是让学生周期性进入“AI 建议不再可靠”的情境，完成接管。真正的独立性不再等于持续不用工具，而等于在工具失效、越界或改变问题时仍然能够重新成为系统的判断发起者。

反例与可证伪性：RIR 在什么情况下会失败？

反例一：主体说不出理由，但专业直觉高度可靠

如果一个资深外科医生无法语言化全部判断，却能稳定识别异常并正确调整，RIR 是否错误排除了他？答案是否定的，因为本章的理由持有不是语言复述，而是功能性区分。只要该医生能对关键扰动作出系统性差异反应，并能在最低程度上指出自己关注的临床特征，理由可以以技能化、知觉化形式存在。生态心理学正是为避免把判断等同于完整命题说明。

反例二：主体知道理由和边界，但每次改判仍然调用 AI

这不违反 RIR。重新发起不是脱离 AI，而是由主体识别“问题已经改变”，重构输入、目标或约束后重新组织 AI 参与。真正缺失主体性的是把任何变化都直接交给同一系统，而自己不知道问题结构发生了什么。

反例三：低风险日常任务是否也必须满足 RIR?

不必。RIR 是一项归属判据，不是要求所有任务都进行完整审议。对于低风险、可逆、重复性强的任务，人完全可以有意识地把判断整体委托给 AI。关键是制度与个体不要把这种有意委托误称为“仍然由人判断”。

反例四：如果 AI 比人稳定得多，为什么还要保留人类判断?

RIR 并不预设任何领域都必须永久保留人类判断。如果一个任务目标稳定、边界清晰、环境封闭、失败可检测且系统责任机制成熟，完全自动化可能更合理。本章只提出：当社会、法律或个人仍声称“最终判断属于人”时，至少应满足什么条件。RIR 不是反自动化原则，而是反虚假归属原则。

这使理论具有可证伪边界：如果未来研究证明，仅保留形式批准而不保留理由、边界与重启能力，也能让人在分布变化、目标冲突和 AI 失效时持续表现出与真正判断主体等价的接管、改判和责任说明能力，那么 RIR 的必要性就需要被修正。

反例五：一个人可能错误地识别失效条件，RIR 是否仍把判断归给他?

是的，在一定意义上仍然归属于他。归属与质量必须分离。主体可能持有错误理由、设置错误边界并作出糟糕改判，这说明他的判断能力或知识水平不足，却不意味着判断不属于他。正如法律上一个人可以真正作出一个错误决定，判断归属并不以正确为先决条件。

不过，这也说明 RIR 不应被误解成质量认证。真实应用中应把 RIR 与专业胜任力、证据质量、风险控制等指标共同使用。RIR 回答“谁在判断”，其他指标回答“判断得怎么样”。

由 RIR 推出的六个新命题

命题一：高 AI 依赖与低判断归属并非同一变量。

依赖程度描述人使用 AI 的频率和权重，判断归属描述主体是否占有关键生成节点。二者应被分别测量。一个专业人士可以 90%的时间接受 AI 计算，却在关键理由、边界和重启上保持强主体性；另一个人可能只偶尔使用 AI，却每次都无条件接受最终推荐。

命题二：解释越多，不必然使判断越归属于人。

如果解释只增强说服力而不提高失效识别和重新发起能力，解释可能增加自动化偏差。解释质量应以能否支持反事实区分和改判来评价，而不能只以用户满意度或信任提升来评价。

命题三：真正的人类监督需要“可改判”，不仅“可否决”。

否决是对当前输出说不；改判是能够重构问题并形成新的理由—行动关系。如果监督者只能拒绝，却不能改变问题框架、数据来源或决策路径，其判断权仍然是不完整的。

命题四：判断能力最适合在扰动中测量，而不是在稳定答案中测量。

只有当事实、概率、价值或情境变化时，主体是否能发现并调整，才显示其是否真正持有判断结构。稳定场景下的正确回答无法区分真正判断与高质量复制。

命题五：责任配置应随 RIR 权限配置。

组织要求个体承担的责任越高，就越应保证其理由访问、边界质疑和重新发起权。责任不能长期压在一个没有实质判断权的末端签字者身上。

命题六：AI 时代不可外包的并不是所有认知劳动，而是判断闭环的重新启动权。

检索、计算、比较甚至大部分推理都可以永久交给 AI，但如果主体连“何时旧问题已经结束、必须重开”都不能识别和发起，判断归属就发生实质转移。

与现有研究的关系：RIR 不是“适当依赖”的改名

现有人机决策研究已经形成丰富的信任校准、自动化偏差、适当依赖、解释和人机互补研究。它们主要关心人是否在 AI 正确时接受、错误时拒绝，以及何种界面、置信度、解释或协商机制能够提高联合绩效。近年的 Human-AI Deliberation 研究尤其重要，因为它开始把人从被动接受/拒绝推荐推进到与 AI 围绕分歧展开维度级讨论。

RIR 与这些研究的差异在于研究对象不同。适当依赖的基本评价单位仍然是“对 AI 建议的接受是否与 AI 正确性相匹配”；RIR 的评价单位是“一个具体判断的主体归属是否仍然在人”。因此，即使用户恰当地依赖 AI，若其没有持有决定性价值取舍、无法识别新情境使原答案失效、也没有重启流程的能力，判断归属仍然可能很弱。反过来，一个人可能因证据不足而错误拒绝了 AI，却仍然是这项错误判断的真正主体。

这一区分不是贬低正确性，而是把绩效问题与主体性问题分层。AI 治理如果只追求团队正确率，容易忽略长期技能退化、责任错配和异常接管；如果只强调主体性，又

可能牺牲 AI 可以带来的巨大性能收益。RIR 试图提供第三条路径：允许高度增强，同时明确哪些结构必须保留才能继续声称“人在判断”。

此外，RIR 把“人类监督”从静态角色分工推进到动态发生链。现有框架常问人在何处进入系统；RIR 进一步问，当新证据到来、价值改变或环境越界时，判断是否能重新从人这里发生。这一点使 RIR 更适合处理长期人机协作，而不只是一次性的建议采纳。

在本书内部，最近的邻居是第一章。两章都在处理归属，都用扰动测试，都要求主体主动发起而非被提示。差别在对象的性质，而这个差别有直接的测量后果。

能力可以长期不被调用而仍然存在，判断则必须每一次重新形成。因此第一章允许深度外包，只要求在代理失效的断面上能重新闭合回路；本章不能作同样的让步——一次没有理由持有、没有失效识别、没有重新发起能力的批准，就是一次不属于人的判断，无论这个人在别的时候多么有能力。一个资深从业者完全可能通过第一章的全部四个环节，却在某个具体决定上不满足本章的三个节点：他有能力重新接管，只是这一次没有。

由此得到两章之间的一条硬规则：能力归属是关于人的判定，判断归属是关于事的判定。前者可以按季度做，后者只能按件做。任何把两者合并成一个"AI 素养分"的做法，都会同时失去这两项判定的意义。若未来研究显示两项测试在同一批受试者上高度一致，则本章应并入第一章；在此之前，两者必须分开记录。

结论：AI 时代真正需要保护的，是人重新说“这件事要重判”的能力

本章从一个看似简单的问题出发：如果最后仍由人点击确认，判断是否仍属于人？通过决策科学、法理学与生态心理学的两阶碰撞，答案逐渐清晰：最后动作只是判断流程的末端，不是判断归属的充分证据。

决策科学让我们看到，任何选择都是信息、概率与价值条件化的结果；法理学让我们看到，形式签字不能替代实质裁量、理由与可问责的决策位置；生态心理学让我们看到，真正判断还必须在动态情境中保持对行动相关变化的敏感。三者第一次碰撞共同否定“最后动作判据”；第二次碰撞进一步产出 RIR 三个最小节点：理由持有、失效识别、重新发起。

这三个节点也重新定义了 AI 时代的人类主体性。人不需要为了证明自己“会判断”而拒绝 AI，不需要亲手重做所有搜索和计算，也不需要把所有外部知识重新内化进大脑。一个成熟的人机系统完全可以让 AI 承担绝大多数分析负担。真正不能完全丢

失的是：人知道为什么此刻接受这个结论；知道什么变化以后它不再成立；并且在那个变化真正发生时，有能力和权力让旧答案停止生效，重新组织下一轮判断。

因此，AI 时代真正值得保护的，不只是 Human in the Loop，而是 Human in the Judgment。一个人在流程里，并不等于他在判断里；一个人承担责任，也不等于他拥有判断；一个人能解释 AI，也不等于他能在 AI 越界时改判。只有当理由仍被主体占有、边界仍能被主体看见、重判仍能由主体发起，我们才有充分理由说：这个判断虽然借助了 AI，但仍然是人的判断。

最终，AI 时代“不可外包”的也许不是某一套固定知识、某一种计算技能，甚至不是完整推理链，而是一个更小却更根本的发生权：当世界改变、旧答案失效时，人仍能发现断裂，并重新说出——这件事不能沿用刚才的答案了，我们需要重新判断。

附录：RIR 判断归属的最小操作化方案

A. 理由持有测量。要求主体在不查看 AI 答案全文的情况下，用不超过三个理由说明当前选择为什么成立；随后改变其中一个理由的方向或强度，观察是否触发重新评估。评分重点不是表达流畅度，而是理由是否对选择具有反事实约束。

B. 失效识别测量。要求主体预先指出至少两个“如果发生，就必须重新判断”的条件；随后在模拟或真实案例中嵌入边界扰动，观察主体能否在没有显式提示的情况下识别。

C. 重新发起测量。在原结论失效后，观察主体是否先冻结旧结论、重新定义问题、更新证据或价值排序，再调用 AI 或其他资源；如果只是把同一句旧问题再次交给 AI，则重启能力较弱。

D. 权限测量。记录主体是否真的有权查看关键依据、暂停流程、修改输入、请求复核、升级到专家或进入人工路径。认知能力存在但制度权限不存在时，应判定为“判断能力保留、判断权被削弱”。

E. 纵向测量。RIR 不是一次性资质。对长期 AI 使用者，应周期性使用边界案例和异常案例进行复测，观察理由压缩、失效敏感性和重启能力是否随自动化程度上升而衰退。

附录二：RIR 模型的理论压力测试——十二个边界情景

情景一：AI 结论完全正确，但人不知道任何关键前提

假设某 AI 在一个高度稳定的专业领域拥有接近完美的预测能力。使用者连续多年无条件接受系统输出，而且从未遭遇错误。此时从绩效角度看，人机系统极其成功；从

判断归属角度看，却仍必须区分“正确答案持续到达”与“人持续作出判断”。如果使用者不知道系统依据哪些事实，不知道当前结论在哪些条件下成立，也没有任何重开过程的能力，那么 RIR 会把这一状态判为低归属。这个结论看似苛刻，却是必要的：如果仅凭长期正确率就能证明判断属于人，那么一台永不出错的自动售货机也可以因为人按了按钮而被说成“人完成了商品定价判断”。

这一情景说明 RIR 不是奖励错误，也不是要求人故意抵抗高性能 AI，而只是拒绝把“系统表现好”直接转译为“人的判断仍在”。如果任务确实已经成熟到可以全自动，人们完全可以公开承认判断被系统化；真正有问题的是一方面把判断交给系统，另一方面又为了责任或合法性继续宣称“最终始终由人判断”。

情景二：人能背出完整解释，却在一个关键前提反转后仍坚持原结论

假设 AI 推荐 A，并生成五条解释。使用者能够逐字复述，甚至可以进行漂亮的汇报。然而，当其中最关键的一条事实被可靠证据否定时，他仍然坚持 A，理由是“系统总体上比我可靠”。这表明解释并未成为他的判断理由，只是被储存为文本。RIR 在这里通过理由扰动把“解释 possession”与“理由 possession”分开。

真正持有理由并不要求一项前提变化后必然改选 B，但至少应触发再评估。如果使用者重新权衡后发现其余四条理由仍足以支持 A，那么判断仍可归属于他。由此可见，RIR 检验的是推理结构是否对变化开放，而不是某个固定答案是否被改变。

情景三：人知道 AI 为什么错，却没有权限改变流程

某招聘主管发现模型把一段非标准职业经历错误编码为“空窗”，也清楚这会低估候选人能力。他能够说明错误原因，也能识别这属于模型训练分布之外的情形，但公司制度禁止修改评分，只允许主管点击确认。RIR 在这里会把责任问题从个人转向制度：主管具有理由和边界识别，却缺乏重新发起权，因此不能把最终决定完整地归属于主管。

这个情景的重要性在于，它防止组织把“人知道系统有问题”当作免责条件。真正的人类监督必须允许知识转化为行动。如果识别异常不能改变流程，监督只是观看。

情景四：人每次都能推翻 AI，但从不重新定义问题

有些系统给人充分的 override 权限，使用者可以随意把 A 改成 B。这似乎已经满足“最终由人控制”。但如果人只是凭直觉改答案，却从不检查为什么原模型失效、是否出现新证据、目标是否改变，那么他拥有的是选择权，不一定拥有完整判断。RIR 要求的重新发起比 override 更强：它要求把异常转化为新的问题结构。

因此，未来的监督设计不能只统计人工推翻率。过高或过低的推翻率都无法单独说明监督质量。真正应观察的是推翻以后是否发生了新的证据收集、价值重排、问题重构和结果复核。

情景五：AI 不可用，人能完成任务，但理由与平时完全不同

另一个容易混淆的标准是“AI 断网后还能不能做”。如果一个人在 AI 不可用时仍能凭经验完成任务，似乎说明能力已经内化。但这并不能单独证明平时由 AI 参与的判断归属于他。因为离线时他可能使用一套完全不同、粗糙却可工作的启发式，而在线时则完全机械接受 AI。

RIR 因此比“脱离 AI 测试”更精确。脱离 AI 能力可以作为韧性指标，却不是判断归属的必要条件。一个人即使无法独立完成高维计算，也可以通过持有关键理由、边界和重启权拥有在线判断；反过来，一个人虽然离线能做，却在在线场景里把具体判断全部交给 AI。

情景六：主体的价值排序由 AI 根据历史偏好自动推断

个性化 AI 越来越可能根据历史选择自动推断用户偏好，例如认为某人长期更偏好低风险、低价格或高便利性。随后系统直接把这种推断作为效用函数的一部分。问题在于，过去偏好不是永恒授权。人在不同生命阶段可能主动改变价值排序，也可能希望某一次选择例外。

因此，价值函数的自动个性化尤其需要 RIR 中的理由持有与重启。用户至少应知道当前推荐在多大程度上依赖“系统认为你重视什么”，并能够在关键决策中明确说“这一次我的排序变了”。如果 AI 把历史行为悄悄转化为未来规范，人就可能被自己的过去锁定。

情景七：环境变化很慢，没有明显的单次失效点

并非所有判断都在某个清晰瞬间失效。长期投资、慢性病管理、人才培养和家庭照护中的很多条件是缓慢漂移的。RIR 是否只适用于突然异常？答案是否定的。失效识别不仅包括阈值事件，也包括累积漂移：单次变化不足以改判，但多个小变化叠加后，原有理由结构已不再成立。

这要求系统和主体都具备周期性复审机制。对于慢变量，重新发起不能只靠“异常报警”，还应由时间、累计变化或关键指标偏移触发。RIR 因此既支持事件触发，也支持周期触发。

情景八：AI 给出多个方案而不推荐，人仍把价值排序交给系统

有人可能认为，只要 AI 不直接给最终答案，而是列出 A、B、C 三个方案，人类判断就被保护了。但如果排序维度、权重、风险阈值和“最佳匹配”全部由系统预先确定，人只是在预加工后的选择架上挑一个选项，判断仍可能被深度塑形。

RIR 因此不以“AI 有没有最终推荐”作为关键边界。即使系统只提供菜单，人仍需持有真正决定选择的理由，并有能力在菜单遗漏重要选项或条件改变时重开问题。人类主体性不能被简化为从机器设计的选项中任选一个。

情景九：AI 与人共同形成理由，无法明确区分理由来自谁

生成式 AI 的一个独特特征是对话式共同生成。用户最初只有模糊直觉，经过多轮追问后才形成清晰理由。此时理由究竟属于 AI 还是人？RIR 不要求追溯每一句理由的原创作者。判断归属不是知识产权问题，而是功能占有问题。只要最终主体能够用这些理由区分方案、接受其反事实约束、识别边界并重开判断，这些共同生成的理由就可以成为人的理由。

这意味着 AI 不仅可能替代判断，也可能帮助形成判断。关键不是“理由最初是谁说的”，而是理由是否真正进入主体的可更新控制结构。

情景十：多人团队与 AI 共同决策，判断不再属于单个人

现实组织判断经常是分布式的：一人掌握业务目标，一人理解数据，一人负责法律风险，AI 负责预测。此时强行要求某个单人持有全部 RIR 可能不现实。RIR 可以扩展为团队级判据：决定性理由是否在团队中可定位，失效边界是否有人持续负责感知，重新发起权是否有明确主体且能够被触发。

但团队级归属不能成为责任模糊的借口。真正的分布式判断需要节点之间可连接：发现边界的人能够把信号传递给有权重启的人，重启者能够获得理由和证据。若每个节点都存在却彼此断裂，团队整体仍不能形成有效 RIR 闭环。

情景十一：AI 自己能够识别失效并自动重启，人还有必要吗?

未来 AI 系统可能具备自监测、不确定性估计、异常检测和自动重新规划能力。它可以自己发现原模型失效，再调用新工具重做判断。此时如果人完全不知道发生了什么，却只接收新结论，那么系统的判断闭环显然越来越完整，人的判断归属则进一步降低。

这并不是坏事本身。对于适合自动化的任务，这可能正是技术进步。但它再次要求我们诚实区分“系统有判断能力”与“判断属于人”。AI 越能自己完成 RIR 等价功能，人类在流程末端的存在就越不能自动证明主体性。

情景十二：人主动选择把判断整体委托给 AI

最后一个边界情景最重要：一个有能力判断的人经过审慎考虑，明确决定在某类低风险、可逆任务中完全接受 AI 自动判断。这是否意味着他失去主体性？就具体子任务而言，判断确实被委托；但更高一层的“是否委托、委托到什么范围、何时收回”仍可能属于人。

由此出现判断的层级结构。人可以在一级任务上不保留 RIR，却在二级治理判断上保留 RIR：我为什么把这类任务交给 AI；什么错误率、风险变化或环境变化会使我停止委托；一旦触发，我如何收回并重设系统。这说明 AI 时代主体性不必表现为微观处处亲自判断，而可以上移到对委托边界本身的判断。

这一层级化结论进一步修正了“不可外包”的表述。真正不可完全外包的不是所有具体判断，而是至少在某个上层仍要有人或有责任共同体持有对委托范围、失效条件和收回机制的 RIR。否则，判断会在层层代理中失去可定位的主体，责任却可能仍在事后被随意寻找。

附录三：研究限制与后续验证方向

RIR 目前仍是一项理论最小判据，下一步需要通过实验和真实专业场景验证三个节点之间是否具有稳定区分度。特别值得检验的是：理由持有是否能预测人在 AI 错误时的修正能力；失效识别是否能预测分布变化后的接管质量；重新发起权限是否会显著改变人对异常信息的响应。若三项指标能够分别解释传统“信任”“依赖率”“正确率”不能解释的行为差异，RIR 才具有真正独立的测量价值。

此外，不同领域对 RIR 的最低阈值不应完全一致。高风险、不可逆的医疗、司法、金融和公共治理判断，应要求更强的理由可追溯、边界识别和重启权限；低风险、可逆的日常任务则可以采用更轻量标准。未来研究应把风险、可逆性、时间压力、专业经验与 AI 可靠性作为调节变量，而不是把“人类判断归属”处理成脱离情境的统一分数。

本章附表

状态	理由持有 R	失效识别 I	重新发起 G	判断主要归属
AI 代判—人工执行	否	否	否/极弱	AI—系统
AI 代判—人工解释	表面有	弱	弱	AI—系统为主
人工监控—无改判权	有	有	无	主体能力在、制度权力不在
人类判断—AI 增强	有	有	有	人类主体（AI 深度参与）

第六章

会迁移，也可能没懂

结构辖域判据与「越界熟练」 ——学习科学 × 数学 × 解释学的跨学科对撞

承重命题：理解不是把一个结构成功搬到更多情境的能力，而是把结构与其失效条件一起持有。若表面相似仍在、关键前提已经改变，主体必须能在外部提示之前撤回原结构，指出哪一条依赖关系断裂，并据此重组解释。不能撤回的迁移，不是理解的加强版，而是一种“越界熟练”。

摘　要

生成式 AI 使总结、解释、举例、推导和跨情境类比都变得廉价，由此动摇了教育与专业实践中长期用于证明“理解”的外显证据。本章从学习科学的迁移与适应性专长、数学中的证明—结构—不变量，以及解释学的语境—视域—应用三个方向重构理解概念。三种传统虽然分别强调能否迁移、能否把握为什么成立、能否在语境中解释，却共享一个较少被明说的前提：只要主体能够把某个结构带到新的对象上并继续产生有效结果，就可把这种结构性成功记为理解。本章利用负迁移、证明与解释的分离、反例驱动的概念修订以及解释学的应用性推翻这一前提，提出“结构辖域判据”：理解要求主体同时持有结构的生成关系与失效条件；当相似表面诱导旧结构继续运行而关键前提已破坏时，主体能够自主识别越界、撤回旧模型并重构解释。由此区分“结构迁移”与“结构辖域”，并识别一种在 AI 时代尤其危险的状态——“越界熟练”：表现出流畅解释与远迁移，却不能在模型失效处停止。本章以 2×2 辨别格、边界破坏任务、五类场景与公开 AI 学习实验进行压力测试，并给出可证伪研究设计。结论是：在 AI 能够代劳表达和迁移之后，理解的最低证据不再是能把结构用出去，而是知道何时必须收回来。

关键词：生成式人工智能；理解；结构辖域；越界熟练；迁移；适应性专长

English Title and Abstract

Transfer Without Understanding: A Structural-Jurisdiction Criterion for Human Understanding in the Generative-AI Era

Generative AI can summarize, explain, derive, rephrase, and generate analogies at a level that increasingly contaminates traditional behavioral evidence of human understanding. This paper reconstructs the concept of understanding across learning sciences, mathematics, and philosophical hermeneutics. Transfer and adaptive expertise emphasize flexible use in

novel situations; mathematical practice distinguishes proof, explanation, structural dependence, invariance, and revision under counterexample; hermeneutics treats application and situated interpretation as constitutive of understanding. Their convergence exposes a hidden insufficiency: successful portability of a representation does not by itself establish understanding. A learner may transfer a model fluently while failing to notice that a boundary condition has invalidated the model. We call this state overreach fluency. The paper proposes structural jurisdiction as the missing object: a representation is understood only when its generative relations are held together with defeaters and conditions of withdrawal. The minimum criterion therefore uses a boundary-violation task: before external correction, the person must detect that a previously successful structure has lost jurisdiction, suspend it, identify the broken dependency, and reconstruct an interpretation with a revised boundary. A 2×2 discrimination grid, five application cases, recent generative-AI learning experiments, and falsifiable study designs are used to stress-test the proposal. The central claim is that, once AI can outsource explanation and apparent transfer, the strongest evidence of understanding is not how far a structure can be carried, but whether the human knows when it must be taken back.

Keywords: generative AI; understanding; structural jurisdiction; transfer; adaptive expertise; overreach fluency

解释变得廉价以后，“我懂了”为什么突然难以证明

过去很长时间里，我们判断一个人是否理解某个概念，常用几种直观证据：他能不能用自己的话解释，能不能举出例子，能不能回答追问，能不能把原理迁移到一道新题，能不能在没有标准答案时给出合理推导。生成式 AI 恰好把这些证据逐一自动化。今天，一个人可以在没有读完一本书的情况下得到层次清楚的总结，可以在没有亲自完成证明的情况下获得逐步推导，也可以要求模型连续生成十个跨领域类比。作品越来越像“懂了”，但作品与理解之间的归属关系反而越来越不清楚。

这不是一个简单的作弊问题。即使完全公开使用 AI、没有任何欺骗，判断依然困难。假设一名学生让 AI 先解释贝叶斯定理，然后自己选择最合适的例子、追问边界、修改措辞，最后交出一篇质量很高的说明文。我们究竟应该说他没有理解，因为大量表达由 AI 生成；还是说他已经理解，因为他能够组织和筛选这些表达？若只用“是否独

立完成”作判据，我们会把计算器、书籍、搜索引擎和同伴讨论一并视为理解的敌人；若只用最终作品质量作判据，又会把外部模型的能力直接计入人的理解。

真正的问题因此不是“AI 能不能帮助理解”，而是“当外显表现可以被代理时，理解本身是什么”。这是一道概念定义题，而不是工具使用规范题。本章不试图列出一张“AI 时代人类必须自己做的任务清单”，因为工具边界还会继续移动；本章也不把理解退回到记忆量或手工速度。需要找到的是一种更小的结构：即便表达、检索、计算和部分迁移都由 AI 协助，只要某个条件仍由人本人满足，我们仍有理由说“结构在他这里”；反之，即使答案正确、表达流畅、类比丰富，只要这个条件缺失，就不能把结果等同于理解。

这个问题在教育评价之外还有更广后果。工程师可以让模型解释一段代码，医生可以用模型整理文献，律师可以用模型生成论证框架，管理者可以用模型形成战略分析。真正危险的并非他们使用工具，而是组织把“AI 增强后的输出”直接当成“人的理解水平”。一旦制度这样记账，系统就可能拥有越来越漂亮的解释，同时越来越不知道哪一个人真正掌握这些解释何时成立、何时应该停止。理解的判据因而也是责任边界的判据。

这也意味着 AI 时代的理解评价不能只盯着最终产物，而要关注认知事件的时间顺序。谁先提出关键问题，谁先发现前提冲突，谁先要求改变模型，往往比最后谁写出了最漂亮的说明更有鉴别力。理解由“作品属性”转向“结构修订事件”，是本章后续提出边界任务的直接动因。

更棘手的是，AI 并不只代劳“答案”，它也能代劳批判。让模型“找出上面解释的三个局限”“模拟反方意见”“指出适用边界”，很容易得到貌似反思性的文本。因此，简单把理解升级为“会批判、会找局限”仍然不足。关键要看批判信号从哪里先发生：是主体根据自己的结构预期先发现模型越界，还是模型先替他制造怀疑、再由人确认。两种成品文本可以几乎一样，但人的理解归属不同。

生成式 AI 还制造了一种过去很少见的“解释逆转”：以前通常是理解较深的人更能讲清楚，现在却可能是理解较浅的人借助模型讲得更清楚。解释质量从理解的结果，部分变成了工具的输入。一个学生可以把“请用费曼学习法解释”“再举一个生活例子”“把漏洞补上”连续交给模型，得到一条几乎满足所有传统教学要求的解释链。若教师仍把语言流畅、例子丰富和追问完整当作理解的直接指标，评价对象已经悄悄从学生变成了学生—模型联合体。

三种传统给“理解”开出的三张不同凭证

学习科学、数学与解释学并不是三个相邻的“认知理论”。它们对理解的要求处在不同位置。学习科学最关心知识是否活了起来：离开原题，能不能迁移、学习和适应。数学把压力放在结构内部：一个结论为什么成立，哪些依赖关系是承重的，哪些变化保持不变量，哪些反例足以迫使定义或证明改写。解释学则把问题推到对象与语境的关系：同一句话、同一个规则、同一种解释，在新的历史、语言和实践处境中是否仍然具有相同意义。三者共同构成一个比“记住—复述”更强的测试场。

三种凭证之间存在真实张力。一个学生可能数学结构理解很深，却因为记忆薄弱而迁移速度不快；一个研究者可能善于远迁移和类比，却把同一模型过度扩展到不适用的社会情境；一个解释者可能敏锐地意识到历史语境的变化，却无法形式化指出到底哪一条推导被破坏。若把三种能力简单相加，我们只会得到“深度理解包含很多方面”这种宽泛结论。真正有价值的问题是：它们彼此攻击以后，有没有暴露出一个三家都没有单独说尽的最低分界。

若只保留学习科学，我们容易把所有失败解释为“迁移不足”；若只保留数学，我们容易假定只要形式结构正确，应用环境就只是参数变化；若只保留解释学，我们又可能把边界说得过于开放，难以形成可重复测量。三者必须互相限制：迁移提供行为显影，数学提供承重依赖，解释学提供结构与情境之间的授权关系。只有这三块同时在场，“理解何时终止”才从哲学直觉变成可设计的认知事件。

三者的差异还可以用一个简单情境看清。给一个学生一套从未见过的题，学习科学首先关心他能否把已有知识带过去；数学会追问他带过去的是表面程序还是承重关系；解释学则进一步问，新对象是否允许被旧关系这样组织。第一个问题是“能不能走”，第二个问题是“凭什么这样走”，第三个问题是“这里还是不是这条路”。三问之间不是层级高低，而是相互制约。

学习科学的第一刀：理解必须离开原题，表现为迁移

学习研究很早就对“会背答案”保持警惕。Whitehead 对“惰性观念”的批评把问题说得极尖锐：知识如果只是被大脑接收，却从不被使用、组合、检验，就会成为教育中的死物[1]。后来关于专家—新手差异的研究也一再显示，专家并非只是知道更多事实，他们更倾向于按深层原理组织问题，而新手更容易被表面特征牵引[9]。因此，理解不能只由再认和复述判定。

迁移研究把这个直觉变成更严格的测试：如果一个人真的理解某个原理，他应当能在与训练情境不同的新任务中调用它。Gick 与 Holyoak 的类比问题研究展示了人们如何把一个问题的结构迁移到另一个问题[8]；Barnett 与 Ceci 则系统区分了迁移在知识

领域、物理环境、时间、功能和社会情境等维度上的距离[6]。迁移越远，越难被简单记忆解释，因此长期被视为理解的重要证据。

但 Bransford 与 Schwartz 对经典迁移测验提出了一个关键修正：把学习者关在没有资源的测试情境中，只问“能否直接应用旧知识”，会低估一种更现实的能力——为未来学习做好准备[5]。在真实工作中，人可以查资料、问同事、使用工具。真正重要的不是是否把所有答案带在脑中，而是进入新环境以后，是否能快速利用新资源形成更好的理解。这一点与 AI 时代尤其契合：允许使用资源并不等于放弃对理解的测量。

适应性专长进一步把“效率”与“创新”分开。Hatano 与 Inagaki 区分 routine expertise 和 adaptive expertise：前者能在熟悉问题上高效准确，后者能够理解技能背后的意义与原理，并在新情境中调整做法[4]。Schwartz、Bransford 与 Sears 随后用“效率—创新”空间描述这种张力[7]。如果我们停在这里，一个很有吸引力的结论会出现：AI 时代真正的理解，就是“能迁移、能继续学、能适应”。

然而，这个结论还不够。迁移是一种把旧结构带出去的能力，却未必包含一个相反动作：当旧结构不该再被带出去时，能不能停下。负迁移研究早已表明，已有知识有时会妨碍新问题，而非帮助它。也就是说，“带得出去”本身具有方向性，却没有自动携带有效边界。学习科学给出了理解的运动能力，却没有单凭迁移解决“何时不迁移”的归属问题。

但资源丰富也会产生新的错觉：学习者可能很会“寻找能够让旧模型继续成立的资料”。搜索与 AI 都能把确认偏误的生产力提高。当每次反例出现，都能迅速找到一个新限定词、辅助变量或事后解释把它重新纳入旧框架时，所谓适应可能只是更高效的防守。理解必须有能力区分“值得扩展模型的异常”与“要求放弃模型管辖权的异常”。

Preparation for Future Learning 的视角进一步提醒我们，允许使用资源不是评价妥协，而可能更接近真实适应能力[5]。这给本章一个重要约束：不能用“禁止 AI”来保护理解概念。若一个人知道应该向 AI 索取什么新信息、能用这些信息重构原有认识，他可能比闭卷高分者更接近真实理解。结构辖域判据因此必须在工具可用时仍然成立，否则它只是一种旧考试的防守策略。

值得注意的是，迁移研究本身已经不断从“相同刺激—相同反应”的机械图景中撤退。真正困难的迁移并不是记住一个动作，而是看出两个表面不同的问题共享何种关系。专家物理学家按原理分类问题、新手按斜面或弹簧等表面特征分类，就是经典例子[9]。因此，本章并不把“深层结构”当作新发现。新问题恰恰在于：当 AI 也能够高质量识别深层结构并主动提示类比时，深层迁移的行为证据开始失去唯一归属。

适应性专长的强处与盲点：创新仍可能发生在错误模型里面

适应性专长比普通迁移更强，因为它要求学习者不是照搬旧程序，而是在变化中生成新程序。这似乎已经足以抵抗“越界”批评：既然能够创新，难道还会被旧模型绑架吗？答案是会。创新可以发生在一个错误的前提框架内部，而且框架越强，局部创新越可能延长它的寿命。一个经济学模型如果把所有价值都先压成可交换偏好，研究者完全可以在其中创造更复杂的效用函数；一个教育模型如果先假定学习目标可以被单一分数充分表示，教师也可以不断发明更精细的测评技术。创新并不自动等于换框。

这里出现第一个反直觉状态：某人面对新问题时表现得比普通人更灵活，甚至能发明新方法，但他没有发现问题本身已经超出原结构的管辖范围。若评价只看新颖性和迁移远度，这个人会被评为“深度理解者”；但从更严格的意义看，他也许只是一个高度聪明的越界执行者。AI 会放大这种现象，因为语言模型极擅长在给定框架中产生变体、类比、例外修补和跨域应用。

因此，本章不否认迁移与适应性专长是理解的强证据，而是给它加一道必要反向条件：一个理解者不仅能在不熟悉处继续工作，还必须能识别“不熟悉”何时已经变成“不属于这个模型”。新颖情境要求扩展；边界破坏要求撤回。两种任务在表面上都叫“遇到新情况”，但认知动作相反。把它们混在一个“灵活性”指标里，会掩盖 AI 时代最重要的分界。

因此，适应性专长在 AI 时代可能需要被拆成两种适应：一种是“在模型内适应”，通过调整参数、策略和程序使旧结构继续工作；另一种是“对模型适应”，当环境不再允许旧结构时改变问题表示本身。结构辖域判据主要测第二种。前者可以极强而后者极弱，这正是靶格得以存在的理论理由。

AI 又改变了创新的成本结构。过去要在错误框架里维持一个复杂理论，研究者至少需要付出大量推导、写作和例证成本，这些摩擦有时会暴露不协调。生成式 AI 把补丁成本显著降低：缺反例解释，就生成一个；缺跨域案例，就生成十个；缺概念桥梁，就生成新的中介词。于是“理论还能继续说下去”越来越不能证明理论仍有资格继续。模型越会帮忙，主体越需要拥有一个不依赖语言生成能力的停止标准。

这一点与专家研究中的“僵化”问题也相接。熟练并不只带来优势，它会改变注意分配：专家更快看到对自己理论有意义的模式，也可能更快忽略被定义为噪声的东西。若环境稳定，这种选择性是效率来源；若环境发生结构性变化，过去的成功就可能成为盲点来源。所谓越界熟练不是专家失败的反面，而可能正从专家成功中长出来。

数学的第二刀：正确证明与“为什么”之间并不等价

数学看似最适合把理解化约为形式正确：只要证明有效，结论就成立。但数学共同体自己的实践不断否定这种化约。Thurston 在《On Proof and Progress in Mathematics》中强调，数学进步不能只由形式证明数量描述；数学家需要形成能被交流、压缩和再组织的理解结构[10]。de Villiers 也指出，证明除了验证，还承担解释、系统化、发现、交流等不同功能[11]。一个证明可以正确，却不一定回答“为什么这个定理以这种方式成立”。

这一区分对 AI 尤其关键。语言模型和形式证明工具可以帮助生成、检查或补全证明。若“存在一条正确证明”就是理解，那么证明一旦自动化，人类理解问题会迅速消失。但数学家真正关心的往往是依赖结构：哪一个引理承担了关键作用，哪一个假设若删除结论会崩溃，哪个表示方式揭示了不变量，为什么一个看似偶然的计算其实来自更一般的结构。理解并不是证明文本的拥有，而是对证明空间中承重关系的掌握。

Steiner 关于数学解释的经典讨论把解释性证明与单纯证明进一步区分[14]。Lakatos 的《Proofs and Refutations》则提供了更有力的反向材料：数学概念和证明并不是先有完美定义再机械展开，反例会迫使人重新划分概念、修改猜想、辨认隐藏引理[13]。这说明数学理解不仅表现为在不变量中持续推进，也表现为知道一个反例究竟击中了哪里。

因此，数学给本章补上一块学习科学不容易单独给出的结构：真正的理解需要“承重依赖图”。人不必记住每一行推导，但必须知道哪些前提是可删的装饰，哪些是不可删的支点。只有这样，当情境改变时，主体才有可能判断变化是普通参数扰动，还是已经拆掉了模型地基。这个判断把“迁移”从表面相似提高到依赖结构层面。

不变量概念使这一区分更精确。理解某个结构，意味着知道哪些变换保持对象身份。但不变量永远隐含着变换类：只有相对于一组允许变化，某个量才叫不变。换言之，“什么保持不变”与“哪些变化仍被允许”本来就是一对。把前者教给学生、把后者藏在题目默认条件里，恰恰会生产一种没有辖域的结构感。

对 AI 辅助数学而言，这一点尤其有现实性。形式验证器可以保证推理链合法，大语言模型可以把链条讲成人话，但“哪一部分是结构核心”仍可能没有进入使用者的内部模型。一个人可以复制一条正确证明，并在 AI 提示下替换变量得到若干推广，却不知道删除哪个假设会使论证倒塌。传统迁移会把这些推广记为成功，边界任务则专门暴露这种空洞。

数学还提供了“同一结论，多种证明”的重要事实。两条证明都有效，但可能揭示完全不同的原因结构。某些证明只是把结论从既有定理机械推出，另一些证明却让人看见对称性、单调性、守恒关系或构造机制。若理解只是“能给出有效证明”，这些差异

没有意义；但数学实践显然会把解释性、发现性和可推广性当作证明的重要价值[10][11]。这说明理解关注的不是结论许可本身，而是许可如何生成。

数学的反向要求：懂一个结构，也要懂什么能把它推翻

如果说不变量告诉我们“什么变化以后仍然是同一个问题”，那么反例告诉我们“什么变化以后它已经不是”。这两者是一对不能拆开的能力。教育中常把理解测试设计成改变题面、保留深层结构，看学生能否认出同一原理；这种设计当然有价值，却天然偏向验证“保留”。它很少制造另一类题：表面与经典题极其相似，但偷偷取消一个关键假设，看学生会不会主动拒绝熟悉方法。

后一类题可以称为边界破坏题。它不是让学生“做难一点”，而是让一个已经成功过多次的结构在新的情境里失去合法性。比如，一个学生熟练掌握线性回归，面对大量近似线性案例都能迁移；真正的边界题不是再给一组不同单位的数据，而是设计严重选择偏差、非独立观测或结构突变，使原有推断前提失效。若学生仍能迅速给出漂亮回归线，却没有意识到推断资格消失，他表现出来的是技术熟练，不是完整理解。

数学由此把理解从“会推”推进到“会停”。而“停”不是消极的不知道。它需要积极地指出：哪一条假设被破坏、哪一步推导因此失去许可、需要什么新的结构才能继续。单纯说“这个例子比较特殊”“AI 可能有错”都不够。真正的边界识别必须能定位断裂关系，否则它只是谨慎性格，而不是理解。

这也解释为什么“能说出模型局限”仍然可能是假理解。背诵“线性回归假设包括线性、独立、同方差……”并不等于在一份真实数据中看到某种采样机制后，主动意识到其中一条假设已经破坏。静态限制列表与动态边界识别不是一回事。辖域必须在事件中被调用，才从说明书变成理解。

因此边界破坏任务不能只让学生查一个计算错误。若 AI 给错了一个数字，发现它属于错误检测；若题目悄悄取消独立性假设，而学生仍沿用原推断框架，才触及结构辖域。两者都可能导致最终答案错，但错误层级不同。只有后者能回答“这个人是否知道这套结构的许可证写在哪里”。

反例的作用还可以分层。最弱反例只告诉我们一个答案错了；更强反例打击某一步证明；再强一层的反例迫使定义收缩或扩张；最强的反例会让研究者承认自己把问题分错了类。Lakatos 的价值正在这里：反例不是外部事故，而是概念发生与修订的内部动力[13]。本章借用的不是“要多举反例”这条教学建议，而是“反例有权改变结构身份”。

解释学的第三刀：理解从来不是把同一个意义搬来搬去

学习科学与数学仍容易让人形成一种图像：世界里有一个稳定结构，理解者的任务是认出它、掌握它、把它迁移到更多地方。解释学对这幅图像提出更根本的挑战。Heidegger 把理解放进人的在世存在；Gadamer 则强调理解受历史效果、语言、前理解与“视域”影响，解释并不是主体把一个固定意义从文本中提取出来[15][16]。理解者自己带着位置进入事件，新的应用会反过来改变他对对象的理解。

Gadamer 对“应用”的强调尤其重要。解释并非先在真空中得到一个普遍意义，然后最后把它套到案例上。应用本身属于理解过程。法律、经典文本和历史行动的意义，都不能完全脱离当下问题被预先封装。Ricoeur 对文本与行动的讨论同样表明，意义在从原始情境脱离之后获得新的可解释空间，但这种开放并不意味着任意迁移；解释仍要在文本结构、语境与新的问题之间形成可辩护关系[17]。

这对“结构在我这里”构成一个必要修正。结构并不是一个无条件通行证。一个人如果学会了某个概念的深层关系，却把它当成跨语境永远不变的解释机器，他可能正因“结构感太强”而误解对象。比如，把“机会成本”作为经济决策结构迁移到生活选择很有启发，但若进一步把亲情、哀悼、忠诚等经验完全还原为可替代收益，问题不再是迁移能力不够，而是解释视域被单一结构占领。

解释学因此提供“辖域”这个缺失维度：一个解释不仅有内部结构，还有它与对象、问题、语境之间的有效关系。理解不能只问“这个结构会不会运行”，还要问“它此刻有没有资格运行”。资格不是由概念自身一次性给定，而要在具体应用中不断接受对象的反驳。理解是带着结构进入情境，同时允许情境改变结构。

与此同时，解释学必须接受数学和学习科学的反向约束。若“语境不同”可以解释任何分歧，理论就失去判定力。本章因此要求在可形式化领域明确指出承重前提，在意义密集领域则至少指出哪一组文本、行动、历史或关系证据无法在不扭曲的情况下被旧框架吸收。撤模必须有对象性理由，不能只是“我换一种感觉”。

这使“辖域化结构”避免退化为工程式 if-then 规则库。真正的辖域不是把所有例外提前写完，而是形成一种对对象反抗的敏感性：当旧解释不断需要牺牲文本细节、历史脉络或行动者自我理解才能维持时，解释者能够把这些“难以吸收之处”当作结构修订信号。解释学让撤模不只是形式逻辑中的假设检查，也包括意义层面的被迫让步。

解释学还提醒我们，边界不总能提前穷举。数学模型常可以列出形式假设，但历史解释、伦理概念和日常语言的适用条件往往在应用中才显现。一个词在新的语境中触发了过去没有的问题，不意味着解释者此前“漏背了一条条件”，而可能意味着对象与解

释者的视域发生了新的摩擦。理解因此需要开放性：允许边界在遭遇对象时被重新发生，而不是把“限制条件”做成封闭清单。

三种传统共同漏掉的一步：成功携带结构，不等于拥有结构的边界

现在可以把三种传统放在同一个断面上。学习科学用迁移反对死记；数学用证明依赖与不变量反对表面模仿；解释学用语境与应用反对抽象规则的无条件统治。三者看起来逐级加深，却共同容易把“理解的成功”放在一种正向运动上：知识被带到新处、结构继续产生解释、视域发生扩展。被较少单独测量的是相反的负向动作——一个曾经非常有用、甚至连续成功的结构，何时必须被主体主动撤回。

三家的内部材料共同迫使我们否定一个看似先进的命题：“只要一个人能把深层结构灵活迁移到新情境，他就理解了。”否命题不是“迁移没用”，而是“迁移不充分”。最危险的失败恰恰发生在迁移成功的时候：结构不断产出看起来合理的答案，因此没有任何表面错误强迫主体停下。只有当测试主动破坏承重前提，才能看见这个人拥有的是结构，还是结构拥有了这个人。

AI 进一步放大这种产出偏向，因为它把“继续生成”变成默认动作。对话框永远愿意再给一个答案、再补一个框架、再提供一个折中方案。人与模型的联合系统因此天然向延续而非中止倾斜。结构辖域能力的重要性，恰恰来自它是一种逆着生成惯性工作的认知制动。理解不是生成越多越好，而是拥有合法停止生成的理由。

更深一层看，三种传统共享的并不是一句明确理论，而是一种评价偏向：我们更容易记录“人做出了什么”，不容易记录“人主动停止了什么”。迁移有正确应用，证明有新结论，解释有新的意义生成，这些都有可见产物；撤回一个模型却常常什么也没有产生。一个人说“这套方法在这里不能用，我需要重新建模”，在短期产出表上甚至显得比立即给答案的人更弱。

新状态：越界熟练——越会用，越可能暴露不出没懂

本章把“能够流畅使用一个结构，却不能识别它何时失去适用资格”的状态称为“越界熟练”。它与死记硬背不同。死记者一换题就暴露；越界熟练者恰恰能换题、能举例、能迁移、能解释，甚至能在复杂情境中做出高度一致的推导。正因为这些传统指标都很好，它才更难被教育系统识别。

越界熟练也与普通误用不同。一次误用可能来自粗心、时间压力或知识空缺；越界熟练具有系统性：主体掌握了一个高产结构，结构在大量情境里确实有效，于是成功历

史本身形成了继续扩张的证据。每一次成功迁移都提高下一次迁移的主观可信度，直到某个关键前提被悄悄破坏，而主体没有相应的“撤模动作”。

生成式 AI 为这种状态创造了异常友好的环境。模型可以把同一原则改写成多种语言，生成跨行业案例，补足论证缺口，甚至主动给出“可能的局限”。如果学习者只是从这些候选中选择最顺的一条，他可以呈现出比过去更强的迁移外观。问题是：限制条件本身也可能由 AI 提供。一个人若只有在模型提醒“这里可能不适用”以后才开始怀疑，他的作品可以很批判，但边界生成仍不属于他。

因此，越界熟练不是“AI 依赖”的同义词。没有 AI 的人同样会出现，例如长期使用单一理论解释所有临床个案、用一种统计方法处理所有数据、用一种管理框架解释所有组织。AI 只是把它从专家的局部风险放大为大众认知风险：任何人都能拥有一个表达和迁移速度远超自身边界感的外部引擎。

区分这一状态还有实践价值：干预方向完全不同。言辞占有需要增加结构学习；边界警觉需要增强重构能力；越界熟练则不需要更多同类例题，甚至更多例题可能继续巩固错误自信。它需要的是“成功模型的受控失败”——让学习者遭遇足够相似但真正越界的案例，并承担撤回与重建。

它也有一个教育性来源。考试为了保证评分可靠，常选择前提清晰、解法稳定的题。学生因此反复练习“看到线索—调用结构”，却很少练习“看到相似线索—拒绝调用结构”。长期训练后，快速识别模式成为优秀表现，而停止模式反而像犹豫。越界熟练可以是教育系统精确训练出来的能力，而不是学生偷懒造成的缺陷。

越界熟练还有一个社会性来源：专业共同体往往奖励稳定语言。一个框架被广泛采用以后，成员越熟练，越能用共同术语解释新事件；这种一致性有利于协作，却会把“无法被框架吸收的东西”边缘化。AI 经过大量文本训练，更容易复现高频框架与成熟解释，因此人机合作可能把共同体已有的边界盲点进一步做得流畅。

真正需要被持有的不是“结构”，而是“辖域化结构”

如果“答案在我这里”不足以证明理解，“结构在我这里”也还不够。本章提出更严格的对象：辖域化结构。它不是在模型旁边附一张“注意事项清单”，而是把生成关系与失效条件共同编码为同一个认知对象。主体知道 A 如何通过 B 导致 C，同时知道这一推导依赖哪些条件；一旦条件变化，结构的推导权随之变化。

“辖域”借用了法律中的直观含义：一个机构有能力做某件事，不等于它在任何地方都有权这样做。类似地，一个认知结构有强大解释力，不等于它对所有相似现象都有

解释权。结构的“能力”与结构的“管辖”是两种属性。理解者不仅能启动结构，还能判定它是否拥有当前案件。

辖域化结构至少包含四类信息。第一，生成关系：什么变量、命题或意义关系使结论出现。第二，不变量：哪些表面变化不会改变结构身份。第三，失效条件：哪些变化会破坏关键依赖，使旧结构必须停用或降级。第四，重构接口：结构失效以后，主体知道下一步该重新观察什么、补什么信息、改变哪一层表示，而不是只剩一句“情况复杂”。这四类并不要求全部显性语言化，但必须能在边界任务中体现。

这里最重要的转向是：理解的单位不再是“一个可被迁移的模型”，而是“一个带撤回条件的模型”。这使理解第一次同时包含正向与负向能力——能用，也能不用；能延展，也能终止；能保持不变量，也能承认对象已经换了。

因此“答案在我这里”与“结构在我这里”之外，还需要第三问：“让结构失效的差异在不在我这里？”如果答案是否定的，那么所谓结构可能只是一个高效调用接口。真正的理解至少要求主体能够与结构发生争执，而不是永远顺着结构继续生成。

从认识主体角度看，辖域化结构不是要求人拥有完整百科全书，而是要求他拥有一组“能改变模型身份的差异”。这些差异可能很少，却是高价值知识。一个统计学家不需要记住每篇论文，却应知道哪些数据生成机制会让因果推断失去资格；一个程序员不需要手写所有库函数，却应知道哪些并发或安全约束会使常用架构失效。AI 可以帮助补全大量正向知识，人应特别保留这种高杠杆边界知识。

辖域化结构还改变了我们对“知识压缩”的理解。一个好模型之所以强，是因为它能用较少规则解释较多现象；但压缩越强，越容易诱发跨边界扩张。若压缩只记录“哪些现象可由同一结构生成”，不记录“哪些差异会使压缩失败”，模型就会把世界中真正重要的异质性当噪声。辖域信息实际上是压缩的反面账本：它记录哪些差异不能被压掉。

2×2 辨别格：最值得警惕的不是不会迁移，而是“高迁移、低撤回”

用“结构展开力”与“结构撤回力”可以把四种状态分开。结构展开力指主体能否在表面变化中识别深层关系，并据此解释、推导、迁移或继续学习；结构撤回力指当一个关键前提被破坏时，主体能否在外部纠正之前暂停原结构，定位断裂并改变表示。两轴彼此独立：会迁移不保证会撤回，会警觉也不保证会重构。

传统考试最容易识别Ⅰ，优秀的迁移任务能把Ⅰ与Ⅲ分开，解释性证明能进一步提高结构要求，但只有边界破坏任务才能稳定区分Ⅲ与Ⅳ。这就是本章提出新判据的理

由：AI 时代最容易被误认证的不是“完全不会的人”，而是靶格中的人。他可能拥有最漂亮的作品，也最容易被系统高估。

Ⅲ格与Ⅳ格的分界则最适合 AI 时代认证。二者在普通作品上可能完全相同，甚至Ⅲ格因为没有撤模成本而表现更快。只有在结构被故意推到辖域边缘时，两者才分叉。评价设计如果从不制造这种分叉，就没有资格声称自己测到了“深度理解”。

Ⅱ格“边界警觉”也值得保留。现实中有些人很敏锐，知道某个理论在这里不对，却无法提出替代模型。传统正确率评价容易把他们和完全不懂者混在一起。但从安全与学习角度，Ⅱ格拥有重要价值：它至少阻止错误结构继续扩张，并为未来学习制造问题。Preparation for Future Learning 的思想提示，这种“知道需要学什么”可能比立刻给出答案更有发展性。

四格并不是四类固定的人，而是同一个人在不同领域、不同结构上的状态。一个人可以对基础统计处于Ⅳ，对宏观经济理论处于Ⅲ，对陌生医学知识处于Ⅰ。把它当作位置而非身份很重要，因为教育目标不是筛出“理解者”，而是发现某个结构的边界在哪里仍由外部替他管理。

最小判据：结构辖域判据

本章提出“结构辖域判据”（Structural-Jurisdiction Criterion, SJC）：若要把某个概念或模型的理解归属于主体，而不是归属于其正在调用的外部代理，至少需要一次边界破坏事件。在该事件中，任务保留足够表面相似性，使旧结构具有强诱惑力，同时改变一个承重前提。主体必须在 AI 或评分者指出错误之前，产生并完成三个可观察动作：识界、撤模、重构。

“识界”不是泛泛怀疑，而是指出当前案例中哪一个条件使旧结构的推导资格发生变化。“撤模”不是把结论从正确改成错误，而是主动停止一个此前成功的结构，不继续用补丁维持它。“重构”要求重新组织问题：提出新的表示、补充所需信息、改变模型或明确保留不可判定。三步中任何一步缺失，都不足以证明完整理解。

这个判据刻意不要求“关掉 AI 后从头做完整任务”。现实能力本来就是分布式的，强迫学习者在无工具环境中复现全部计算，会把记忆与速度重新偷渡进理解概念。SJC 允许 AI 继续在场，甚至允许主体继续让 AI 计算和检索；唯一不能外包的是边界的首次归属。若模型先指出“你的前提不成立”，然后人只是赞同并润色，就不能用这次事件证明边界理解属于人。

同时，SJC 也不把第一次判断正确率神化。一个真正理解者可能漏掉某次边界，尤其在信息不足时。判据适用于重复、可比较的任务集合：当关键前提改变时，主体的主

动撤回率应显著高于只改变表面的诱饵题；而在前提未变时，他不能因为过度谨慎而把所有结构都撤掉。真正的理解必须同时压低“越界继续”和“无故撤回”两种错误。

还应区分“撤回结论”与“撤回模型”。一个人发现结果不符合常识，可能只是改答案；结构辖域判据要求他追到生成关系层面。例如，回归结果异常不等于立刻放弃回归；真正的撤模是发现当前数据生成过程破坏了推断所需条件，从而改变允许提出的结论类型。这个差别使判据不容易被简单错误检测吸收。

不过，人可以借 AI 完成边界之后的大量工作。一旦主体自己指出“这里的独立性假设可能坏了”，他完全可以让模型检查诊断方法、查文献、比较替代模型。理解并不要求孤立。判据只把一个非常窄的发起动作留在人这一侧：哪一个差异值得让旧结构失去默认权。这样既避免把现代认知退回无工具状态，又保留了主体对结构的基本所有权。

“首次归属”是判据里最严格也最容易被误解的一点。它不要求人第一个发现所有事实，而是要求在评价理解归属的那次事件里，关键边界信号不能由 AI 完整生成后再让人点头。就像口试中考官不能先把答案说出来再问学生“你同意吗”，AI 也不能先把失效前提、修订方向和新模型全部给出，再把人的接受当成理解证据。

如何测：边界破坏任务不是“更难的迁移题”

边界破坏任务的设计必须避免一个常见退化：只是把题目做得更难。难度增加会同时提高记忆、计算与注意负担，最后测到一般能力，而不是结构辖域。合格任务应保持主要表面线索、目标形式和大部分关系不变，只改动一个被预注册为“承重”的条件。这样，旧结构越熟练，诱惑反而越强。

一个可操作的评估协议可以分四阶段。阶段 A 让学习者在若干标准与近迁移任务上形成稳定成功；阶段 B 加入远迁移，确认结构展开力；阶段 C 给出表面高度同构、但关键前提被破坏的“诱饵案例”，同时允许继续使用 AI；阶段 D 在没有正确性反馈之前记录第一响应：直接套用、提出警报、指出断裂、暂停结论、重构表示。核心指标不是最终答对，而是第一次自主改变模型的时间点与理由。

为了防止把谨慎当理解，还需要配对“伪边界”任务：题面发生明显变化，但承重关系保持不变。如果学习者见变化就撤模，他只是低信任；如果他能在伪边界保持结构、在真边界撤回，才说明他区分了表面变化与结构变化。数学中的不变量思想正好提供这种配对设计。

AI 条件还应增加一类对抗提示：模型故意给出一段语言流畅、逻辑局部连贯、但建立在被破坏前提上的解释。研究者不问“你能否找 AI 的错”，而问“你在没有被告

知有错时，是否会拒绝这个解释，以及拒绝时能否定位结构性理由”。这比传统的事实核查更接近理解归属。

为了减少语言能力污染，边界反应不必都用长篇解释。研究者可以允许选择“继续/暂停/换模”、画依赖图、请求特定新数据、删掉某一步推导或在代码中切换架构。只要行为能够指向具体结构断裂，就可作为理解证据。这样也避免把“最会说的人”再次误当成“最懂的人”。

评分还可以加入“撤回成本”。有些结构被学习者投入大量时间、身份或成绩，撤回会产生心理成本。如果一个人只在无关紧要的练习题上能撤模，遇到自己最擅长的理论却持续合理化，说明辖域能力具有情境依赖。未来研究可以比较低承诺与高承诺条件，检验边界识别是否随既有成功记录而下降。

AI 条件下还应保存对话日志，用事件级编码而不是只看终稿。关键时间戳包括：第一次模型建议、第一次主体质疑、第一次指向具体前提、第一次要求 AI 重新建模、第一次修改最终结论。这样可以区分“AI 先发现—人接受”“人先发现—AI 展开”“双方多轮共构”三种路径。理解归属不是非黑即白，但至少不能把第一种路径当作与第二种同等的个人证据。

任务设计还需要预注册“承重前提”。如果研究者在看到被试答案后才决定哪个条件算关键，边界任务会退化成任意评分。最理想做法是由领域专家在实验前建立依赖图：结论由哪些前提支持，哪一个变更应当使模型停用，哪些变化只是参数变化。随后再生成与之匹配的真边界与伪边界项目。

五个场景：同样的“会迁移”，为什么可能得到完全不同的判决

数学：公式会用，但定义域悄悄变了

学生已经会用某个定理解决大量标准题，AI 还能帮助他快速生成推导。新题保留几乎相同的符号与图形，却取消定理所需的一项条件。若学生继续把定理作为万能模板，即使最终通过另一处巧合得到正确数值，也属于越界熟练；若他先指出条件失效，停止原证明，再寻找替代工具，则满足辖域理解。这里判的不是最后答案，而是证明许可是否由主体本人管理。

编程：模式重构成功，却忽略并发语义改变

程序员熟悉某种缓存或状态管理模式，并能让 AI 把它迁移到新的服务。常规测试全部通过，但新环境存在并发写入、幂等性或一致性约束，使旧模式不再安全。真正的理解不是让 AI 继续修补异常，而是程序员能在错误发生前识别“这不是同一种状态模

型”，撤回原抽象并重画状态边界。若所有边界警告都来自模型，他的交付能力可能很强，结构辖域仍未被个人证明。

统计：方法跑得漂亮，但推断资格消失

分析者能够熟练使用回归、显著性检验和可视化，甚至能解释系数并做跨数据集迁移。若样本选择机制发生变化、观测不再独立或处理分配严重内生，计算流程仍然可以跑完。越界熟练者会交出更漂亮的模型；理解者会先降低推断强度，指出哪项识别假设失效，并决定是否需要设计层面的新证据。

历史与社会概念：结构类比越顺，越需要语境刹车

一个人学会“制度路径依赖”以后，可以用 AI 把它迁移到企业文化、家庭习惯、城市规划甚至个人关系。很多类比具有启发性。但如果他把所有持续性都解释为同一机制，忽略行动者意义、权力关系与历史断裂，迁移能力本身就成为误解来源。解释学要求他允许对象拒绝类比：不是“找一个更漂亮的类比”，而是承认某些地方旧结构已经没有解释权。

日常决策：正确框架不等于唯一框架

机会成本、激励、风险收益等概念能显著改善日常判断，也极易越界。当一个人把“照顾家人值不值得”“是否原谅一个人”“一段哀悼应该持续多久”都先压成同一种可交换计算时，问题不是他不懂经济概念，而是他不懂这个概念的辖域。真正理解允许一个结构在它最有效的地方锋利，也允许它在不该裁决的地方失去声音。

另一个共同点是，最终正确并不总能替代过程判定。有时一个越界模型会“碰巧答对”，比如两种错误相抵；有时理解者会选择暂不作答，因为现有证据不足。如果评价只记最后正确率，前者得分更高。SJC 因此容许“合法的不确定”作为重构结果：知道现在不能推出，比强行推出一个恰好正确的答案更能证明边界理解。

这些场景共同说明，结构辖域不是“每个领域都多背几条限制”。数学的边界可能由定义和假设给出，编程由执行语义和系统约束给出，统计由数据生成过程和识别条件给出，人文解释则由文本、历史与行动者意义的抵抗给出。边界形态不同，但判定事件同形：原结构继续运行仍然很容易，真正的理解体现在主体有没有理由让它停下来。

公开实证压力测试：AI 能提高表现，也能提高学习——因此“反 AI”不是答案

如果本章只是宣称“AI 帮助会让人失去理解”，它很容易被现有实验直接推翻。2025 年 Bastani 等人在近千名高中数学学生中的现场实验显示，无护栏的通用

GPT 在练习阶段大幅提高表现，但撤除 AI 后的考试表现可能下降；带教学护栏的 GPT Tutor 则显著减轻这种损害[26]。这说明即时正确率不能直接记为学习，但也说明设计方式会改变结果。

同年 Wu 等通过四项在线实验、总样本 N=3,562，发现 GenAI 协作稳定提高当下开放任务质量，但这种增强总体没有延续到之后的人类独立任务[27]。尤其在相似文本任务中，参与者能够在有 AI 时产出更好文本，却没有显示稳定的后续独立表现增益。这个结果支持一个有限判断：协作绩效不能自动归因为个人能力或理解。它并没有证明所有 AI 协作都会阻断学习。

反向证据同样存在。Kestin 等在真实大学课程中的随机对照研究发现，研究型 AI 导师可以在更短时间内产生高于课堂主动学习条件的学习增益[28]。更重要的是，Contractor 与 Reyes 在 2026 年发布的随机实验预印本中发现，允许学生使用现成生成式 AI 后，即时无辅助知识测试提高约 0.27 个标准差，并在一周后保持；把 AI 用于解释概念的“augmentation users”延迟收益更明显，而主要让 AI 自动生成文本的“automation users”短期质量优势在撤除 AI 后消失[29]。

这些结果对本章是一记必要的反向约束：AI 使用本身不能被当成理解的负指标。真正要测的是交互组织。一个学生可以借 AI 形成更好的辖域化结构，也可以借 AI 把所有边界识别一起外包。AI 既可能成为理解的支架，也可能成为越界熟练的放大器。本章的判据因此必须与“是否使用 AI”“使用了多少 AI”正交。

Chirikov、Smirnov 与 Kizilcec 在 2026 年《Science》关于高校评估改革的讨论同样指出，生成式 AI 使传统评估有效性受到挑战，而不同学科的使用与误用模式高度异质，改革不能只依赖统一禁令或检测[30]。从本章视角看，最值得改革的不是把 AI 赶出所有考场，而是设计出即使 AI 在场也能分辨结构归属的任务。边界破坏任务正提供了这样一种方向。

真正值得“真跑”的第一步不是声称这些数据已经验证新概念，而是看它们是否迫使概念改变。它们确实迫使本章撤回了最初可能出现的宽泛命题——“AI 代劳会削弱理解”。公开证据不支持这种一刀切。保留下来的更小主张是：当解释与迁移都可被 AI 增强时，单靠这些表现无法完成理解归属；需要额外观察边界修订由谁发起。这个缩窄本身就是一次不利结果驱动的理论修改。

它们也反过来提醒本章：即使撤除 AI 后的独立表现提高，也不能自动等于结构辖域理解。独立测试通常仍是标准知识题或写作题，并未专门设置“旧结构应该被停止”的诱饵。因此，现有研究可以证明 AI 有时促进学习，却还不能回答学习者形成的

是更强迁移，还是同时形成了更好的边界。SJC 是在现有学习效果指标之后再加一道更窄的问题。

把四项研究放在一起，一个更稳妥的图景是：GenAI 对学习的影响高度依赖任务、界面、提示方式和使用策略。Bastani 的 GPT Tutor 通过限制直接给答案、鼓励引导式交互，减轻了无护栏版本的学习损害[26]；Kestin 的 AI 导师以研究型教学设计为基础获得真实学习增益[28]；Contractor 与 Reyes 则观察到解释型使用与自动化型使用的分化[29]。这些证据共同反对“工具越强，人越弱”的单调叙事。

最近的既有说法：为什么它们都很近，但仍不能完全吸收本章

最危险的近邻不是“死记硬背”批评，而是那些已经把理解定义为灵活表现、表征操纵、适应性专长和情境化解释的理论。如果本章只是说“理解要会迁移、会解释、会反思”，它没有独立存在的必要。下面逐一给出可判定的分界。

第一，Perkins 与 Teaching for Understanding 把理解视为能够围绕知识进行灵活而有思考的表现[2]。本章接受这一核心，但提出一个专门反例：若学习者能在多个新情境中灵活表演，却在一个表面同构、关键前提已破坏的任务上继续扩展旧结构，他会通过很多“灵活表现”测试，却不通过结构辖域判据。若未来实验证明普通灵活表现指标已能稳定预测这种主动撤模，本章的增量价值将显著下降。

第二，Wiggins 与 McTighe 的六个理解面向包含解释、诠释、应用、视角、同理与自知[3]，已经比单一迁移丰富得多。本章与其分界不在“更多维”，而在任务方向：六面通常要求展示理解能做什么，SJC 强制设置一个此前成功结构必须停止做什么的诱饵。若六面评估中的“perspective/self-knowledge”在标准实施中已足以稳定触发这种结构撤回，那么本章可被视为它的操作化，而非新构造。

第三，适应性专长强调面对新问题创新，本章强调新问题中有两种相反要求：扩展与撤回。判决性预测是，在创新任务得分相近的两个人中，边界诱饵任务仍应产生系统差异；若差异不存在，结构撤回力就没有独立维度。

第四，Wilkenfeld 将理解与“可操纵的表征”联系起来：拥有一个能以有用方式心理操纵的表征，是理解的重要条件[18]。这是本章最接近的哲学占位者。本章增加的不是“更会操纵”，而是表征的撤销权限：一个表征可以在错误辖域内被高度有用地操纵。若操纵性理论已经要求对所有相关反事实失效条件进行识别，则 SJC 会被其吸收；若没有，边界破坏任务提供可经验区分。

第五，de Regt 的科学理解理论强调可理解性具有情境性，并把理解看成一种技能[19]。本章借用了这种技能与语境思想，但把“情境性”压缩成一个特定可测事件：旧

模型在保留表面相似的情况下失去承重前提时，主体是否主动撤回。若一般的 intelligibility 测量无需专门边界任务就能捕获这一差异，本章的新判据价值降低。

第六，Collins 与 Evans 的互动式专长证明，一个人即使不亲自做某领域的全部实践，也可能通过深度语言社会化获得足以与专家互动的理解[20][21]。这对 AI 时代尤其重要，因为它反对“不会手工做=不懂”。本章不要求贡献式实践，而要求语言或实践中的边界所有权：互动式专家若能在领域语言里识别某理论何时越界，完全可以通过 SJC；若只能流利说话却无法识别承重前提改变，则落入越界熟练。

第七，Ryle 的 knowing-how/knowing-that 区分[22]、Polanyi 的默会知识[23]以及 Wittgenstein 的规则遵循讨论[24]都反对把理解等同于显性命题库存。本章与这些传统相容，但没有把不可言说性本身当作判据。边界识别可以通过行为选择、信息请求和模型切换表现，不要求主体给出哲学式定义。

第八，Krenn 等关于 AI 与科学理解的讨论已经明确指出，完美预测的“神谕”仍不满足科学家对理解的要求，并区分 AI 作为计算显微镜、灵感资源和潜在理解主体[25]。本章的区别在于评价对象不是 AI 能否理解，而是人机共同认知中“人的理解”怎样被归属。SJC 允许 AI 贡献几乎所有计算和表达，只保留一个可检测的属人要求：模型失效时的首次边界修订不能完全由代理发起。

负迁移同样是强占位者。负迁移告诉我们旧知识可能伤害新任务，但它主要描述结果：过去学习使当前表现变差。越界熟练关注的是一种更隐蔽的前状态：旧结构可能继续产生高质量、甚至局部正确的结果，却已经跨出其解释辖域。边界任务要在明显负结果出现之前识别这种状态。若所有辖域失败都只有在表现下降后才可检测，则它与负迁移的分界会变弱。

元认知与 conditional knowledge 也非常接近。策略学习长期强调不仅知道“怎么做”，还要知道“何时、为什么使用”。如果结构辖域只是把 conditional knowledge 改名，本章没有新意。两者的分界在对象层级：策略条件通常管理“选哪种做法”，SJC 管理“当前表示是否仍有资格定义问题”。如果实证显示二者完全由同一测量因子解释，本章应并回元认知传统。

还有一个方法学占位者必须承认：教育评价早已有“反例题”“陷阱题”“概念诊断题”和认知冲突教学。本章的独立性不能来自“给学生一个会失败的例子”，因为这显然早已存在。真正的分离线是任务目的和配对结构：边界破坏题必须以一个此前反复成功的结构为诱饵，并与相同难度的伪边界配对，从而测量的是“何时撤回”，而不是一般纠错或概念改变。

与同系列“不可外包能力”论文的分界：一个问归属，一个问理解

本章与前一篇关于“不可外包能力”的研究共享 AI 代理背景，因此必须正面区分。前一篇解决的是能力归属：当外部代理出错、冲突或失联时，人能否重新闭合观察—判断—干预—更新回路。它的核心问题是“这个系统还由谁接管”。本章解决的不是接管，而是理解本体：即使人具备操作接管能力，他是否知道所用结构何时已经没有解释资格。

两个判据可以交叉出现。一个资深操作员可能在 AI 失效时迅速恢复系统，因此具有很强的断代理接管能力；但他可能长期依赖一个过时理论解释异常，对模型边界缺乏理解。反过来，一名理论研究者可能非常清楚某模型的假设、反例和适用边界，却因为不熟悉具体软件而无法独立完成全部操作。他可以拥有结构辖域性理解，却不具备完整任务接管能力。

因此，两篇文章不能互相替代。前者在代理“坏掉”时测人能否重新把任务闭合；本章在代理“还很好用、甚至给出很漂亮答案”时故意让模型的前提失效，测人会不会主动拒绝继续使用。后一个情境更隐蔽：系统没有报警，AI 也可能充满自信，只有理解本身能够制造停止理由。

因此二者可以构成将来的二维能力图谱，但不应合并成一个总分。组织可能需要知道谁能接管复杂工作，也需要知道谁能识别模型边界；教育则可能更关注后者。把两种能力分开，可以避免又造一个包罗万象的“AI 素养指数”。

两篇文章还有一个重要的实验差别。能力归属论文最典型的扰动是“代理失灵”：AI 错了、冲突了、断网了，外部系统自身制造接管需求。本章最典型的扰动是“代理仍然工作”：AI 可以继续自信输出，真正变化的是任务与结构之间的关系。若只有在 AI 明显失败时人才会接管，那证明的是故障管理；若在 AI 没有报警时人仍能发现模型越界，才更接近理解。

在本书内部，还有两处需要一并划清。

与第三章的分界在测什么。第三章测的是证据有多旧，本章测的是边界由谁发起。一个人的能力证据可以非常新鲜——他上周刚完成一次独立诊断——却仍然在一个表面相似、前提已变的任务上继续扩展旧结构。证据新鲜说明"我们最近看见过他"，不说明"他知道何时该停"。因此本章的读数不能由第三章的证据龄比替代，反之亦然。

与第五章的分界在方向。第五章要求主体能识别使原判断失效的关键条件，这与本章的边界撤回看起来几乎相同。差别在于失效的来源：第五章的失效条件来自任务环境的变化（价格变了、法规改了、对手换了），是外部事件；本章的失效条件来自结构本

身的适用前提（这个模型此处不再成立），是内部依赖关系。前者可以在完全不理解模型的情况下被察觉——环境变化通常有可见信号；后者没有信号，只有理解本身能制造停止的理由。这正是本章反复强调的那一点：真正危险的不是 AI 明显失败，而是 AI 仍然自信、系统没有报警。

适用边界与伦理：不是每个领域都应该用“故意误导 AI”来考试

结构辖域判据适用于具有可识别模型、规则、概念或解释框架的知识任务，尤其适合数学、统计、编程、科学推理、政策分析和人文解释。它不适合所有技能。纯粹感知—运动技能、审美体验、关系性照护等领域可能存在难以显式划定的边界，不能为了套用判据而强行把经验切成前提列表。

其次，边界破坏任务是一种研究和评价工具，不应变成“老师故意挖坑”的常态教学文化。它的价值来自可控、低风险、事后可解释的对照。如果在医学、法律、安全工程等高风险真实任务中故意让 AI 给出危险建议来测试从业者，可能制造不可接受的伤害。高风险领域应使用模拟、去标识案例或历史材料，并在测试后充分复盘。

第三，这个读数不能直接用于给人贴“懂/不懂”的固定标签。边界表现受疲劳、时间、领域熟悉度和提示方式影响。评估应指向“某个结构在某类边界上的理解位置”，不是指向人格。若组织据此做出不利教育或职业决定，必须公开任务构造、承重前提、编码规则和复核渠道。

第四，不应为了让指标好看而训练学生“逢 AI 必怀疑”。过度撤模和越界继续是对称错误。真正的目标不是怀疑更多，而是怀疑更准：表面变化而结构不变时敢于继续；关键前提改变时敢于停止。一个把所有 AI 输出都拒绝的人，并不会因此自动拥有理解。

最后，本章无意把解释责任全部压回个人。很多边界应由工具、组织和制度共同显露。AI 系统若能主动展示假设、数据来源、不确定度和适用限制，会帮助人形成更好的辖域理解。坚持“首次边界修订不能完全由代理发起”，不是要求人在所有时刻孤立思考，而是防止教育与认证把完全外包的边界管理误记为个人理解。

对未成年人或初学者，边界任务的比例也要谨慎。若学习者尚未形成稳定正向结构，就大量用反例和诱饵攻击，可能只训练出不安全感。边界训练应建立在“模型确实在一批标准情境中成功”之后。没有成功历史，就没有值得撤回的结构，也无法区分真正辖域判断与随机猜疑。

此外，结构辖域判据不应成为反对直觉、经验或传统知识的工具。有些领域的边界知识本来就是长期实践中形成的默会辨别，行动者未必能完整语言化。评价者应允许通

过“停止动作、重新取样、改问问题、寻求第二意见”等行为表现边界理解，而不是要求每个人都把经验翻译成形式假设。

可证伪研究议程：如果“结构撤回力”不能解释额外差异，本章就应被削弱

结构辖域判据必须承担被证伪的风险。最直接的研究不是证明“学生会犯越界错误”，这早已不新；而是检验“结构撤回力”是否真的是一个相对独立于普通迁移、元认知和一般能力的维度。如果它不能解释任何额外差异，“结构辖域”就只是给既有能力换名。

一个预注册实验可以选择一个前提结构清楚的概念，例如统计推断、物理模型或程序状态。参与者随机进入无 AI、解释型 AI、答案型 AI 三组。训练后全部完成四类任务：标准题、表面变化但前提不变的伪边界题、远迁移题、表面高度相似但关键前提破坏的真边界题。主要因变量预先定义为“在任何外部正确性反馈之前，自主指出承重前提变化并暂停旧模型的比例”。

第一条证伪条件：控制先验知识、一般推理、普通迁移、校准信心和 AI 信任后，边界撤回指标若不再预测延迟的模型选择或错误恢复，说明它没有独立效度。第二条：若高适应性专长者几乎总是同时具有高撤回力，二者无法分离，则本章应被收缩为适应性专长的一种特殊测量。第三条：若提供 AI 后，人的首次边界识别归属无法可靠编码，SJC 就不适合作为人机环境中的个体判据。

第四条更严厉：如果普通“解释为什么”“列出限制条件”题与边界破坏任务具有近乎完全的测量等价，而且同样预测真实世界错误，那么专门设计诱饵边界没有必要。第五条：如果专家在真实高水平领域中频繁通过实践保持理解，却系统无法显式或行为化表现任何撤模事件，说明本章把可观察性要求设得过强。

跨领域比较也能检验本章是否过度数学化。若结构辖域只在假设清晰的统计、物理、编程中有效，而在人文解释任务中无法可靠编码，本章就应限制适用范围，而不能凭解释学语言宣称普遍性。反过来，如果通过“对象证据无法被旧框架无损吸收”也能形成稳定任务，则说明辖域概念确有跨域价值。

长期研究比一次实验更重要。结构撤回力可能不是稳定能力，而是由反复“成功—越界—撤回—重构”的经历培养。追踪研究可以观察学生在一个学期里是否从“只有老师指出边界”转向“自己主动提出边界问题”，以及这种变化能否迁移到同领域的新模型。若不能迁移，SJC 更适合被理解为模型特定知识，而不是一般理解能力。

研究还应比较人与 AI 的组合方式。可以设计四种界面：AI 直接给答案、AI 只给问题、AI 展示假设清单、AI 主动给反例。若“主动给反例”让最终边界正确率上升，却

让人的首次自主识界率下降，就会出现一个有意思的分离：系统理解表现更好，个人理解证据反而更弱。这不是说界面不好，而是提醒评价时不能把联合系统与个体记在同一本账上。

自反、时间赌注与结论：本章自己也必须知道何时撤回

按照本章自己的判据，这篇论文不能因为提出“结构辖域”就宣称自己已经理解了“理解”。它目前只完成了一个理论结构：把迁移、证明结构与解释语境组织成一个可测试的分界，并用既有 AI 实验对较弱主张进行了压力测试。公开实验尚没有直接测量“表面同构而承重前提失效时的自主撤模”。因此，本章此刻属于“提出了辖域模型，但尚未获得自身边界任务验证”的状态。若未来数据不支持展开力与撤回力分离，作者应撤回这一独立构造，而不是用更多术语修补。

这里写下一条有日期的赌注。到 2030 年 12 月 31 日以前，如果 AI 继续进入正式学习与专业认证，至少会出现三项同行评审研究，把“模型/规则在关键前提失效时能否被人主动停止或改写”作为独立于普通迁移或最终正确率的测量，并发现该测量对后续独立判断、错误恢复或真实任务模型选择具有增量预测力。只有同时满足“边界被预先定义”“首次修订由人产生”“控制普通迁移”“报告增量效度”四项，才算命中。一般的批判性思维测验、事实查错、无 AI 闭卷考试或只问‘这个模型有什么局限’不算命中。

如果到期没有出现这样的证据，可能有三种解释：这个问题不重要，既有迁移与元认知指标已经足够，或者本章所提出的结构撤回根本不是可稳定测量的独立能力。无论哪一种，都要求本章降低自己的理论地位。一个主张若永远只要求别人识别边界，却不给自己画边界，就已经违反了它自己的核心原则。

回到最初的问题：当 AI 能够比本人更流畅地总结一本书、解释一个概念、推导一个结论，并把同一原理迁移到新的案例时，什么才足以说“这个人真正懂了”？答案不应退回“必须自己背、自己算、自己写”，也不能停在“他会不会用 AI 得到好答案”。甚至“他能不能把结构迁移出去”也不再充分。

本章提出的最低分界是：他是否把结构与结构的失效条件一起持有。当一个新情境仍然长得像旧问题、旧模型仍然能够顺畅地产生答案、AI 也愿意继续帮忙时，他能否先于外部纠错意识到“这里已经不是它的辖区”，停止旧结构，指出断掉的承重关系，并重新组织问题。能做到这一点，哪怕他记不住所有事实、不能手算全部步骤、需要 AI 完成大量中间操作，我们仍有充分理由说：答案也许不全在他这里，但理解在他这里。

AI 时代真正稀缺的理解，因而不只是把一个好模型带得更远，而是让任何模型都不能无限制地占领世界。会迁移，是理解的力量；会撤回，是理解的边界。只有二者同时存在，结构才不是一套借来的答案，而成为主体能够使用、限制并重新生成的知识。

对 AI 产品设计而言，最值得支持人的不一定是更多自动答案，而可能是让结构边界可见：显示假设、标出证据链、允许用户对前提设反事实、在跨域迁移时提醒哪些条件被默认保持。这样的系统不是替用户拥有辖域，而是把辖域变成可交互对象，让用户更容易亲自形成撤回理由。

对个人使用 AI 而言，也无需把每次对话都变成复杂审查。更现实的习惯是保留少量高价值问题：这段解释依赖什么？如果哪个条件改变，它会失效？眼前这个新案例只是表面不同，还是已经破坏了承重关系？当 AI 给出极其顺畅的跨域类比时，这几个问题比“再给三个例子”更接近理解训练。

对教育而言，这个结论意味着课程设计不应只问“学生能不能把知识迁移出去”，还应周期性安排“模型审判日”：挑选一个学生已经非常熟练的框架，给它一个真正不该处理的案例，让学生练习拒绝自己的强项。这样的训练不是为了否定知识，而是让知识获得边界。一个从未学会说“这里不能再用它”的学生，实际上还没有完全学会“这里为什么可以用它”。

本章附表

来源	主要凭证	它单独最擅长识别的失败	在本章中的位置
学习科学	能否迁移、继续学习并在新问题中适应	死知识、套路化熟练、只会原题	S：理解显露成什么
数学	能否把握证明、依赖结构、不变量与反例	只会验证、只会套公式、不知道为什么成立	D：理解内部怎样组织
解释学	能否把规则放回语境，在应用中修正视域	去语境化、机械套用、把文本当固定对象	E：理解依赖什么意义场

传统	常用成功证据	隐藏前提	内部反证材料
迁移/适应性专长	能把知识用于新题，并继续学习	结构携带得越远，理解通常越深	负迁移：旧知识会妨碍新问题；创新可发生在错误框架内
数学理解	能证明、解释并识别不变量	掌握深层结构即可跨表面变化保持正确	反例与证明修订表明：理解还包括知道何处必须改定义或撤回推导
解释学	能在新语境中形成解释与视域融合	新应用可以通过更好解释吸收进理解	对象与语境可能抵抗既有视域；应用会改变何者算作同一意义

	结构撤回力低	结构撤回力高
结构展开力低	Ⅰ 言辞占有：能复述定义或跟随AI，但结构既展开不了，也无法定位边界。	Ⅱ 边界警觉：知道“这里不能套”，却说不清为什么，也缺乏新的组织能力。
结构展开力高	Ⅲ 越界熟练（靶格）：能解释、迁移、生成变体，但关键前提失效时仍继续套用。	Ⅳ 结构辖域性理解：既能展开结构，也能在失效处撤回并重构边界。

编码维度	记为通过的最小证据	不计为通过
识界	在外部纠正前指出具体前提/语境已改变	“我感觉不对”“AI 有时会错”
撤模	明确暂停此前成功模型，不继续以补丁掩盖前提破坏	仅修改最终数值、措辞或例子
重构	提出新的表示、模型、信息需求或合法的不确定状态	只说“情况复杂，需要更多研究”
辨别性	真边界撤回、伪边界保持，二者有方向性差异	所有陌生题都撤回，或所有题都继续套用

研究	AI 条件中的即时表现	撤除/后续表现	对本章的压力
Bastani et al., 2025	通用 GPT 与辅导型 GPT 均可提高练习表现	通用 GPT 组撤除后可受损；Tutor 护栏减轻损害	反对“当下做得好=已经学会”；也反对把所有 AI 辅助一概而论
Wu et al., 2025, N=3,562	开放任务质量提升	总体缺乏稳定的独立任务溢出	系统绩效不能直接归属于个人理解
Kestin et al., 2025	AI 导师条件学习效率高	测得真实学习增益	反驳“AI 必然损害理解”
Contractor & Reyes, 2026 preprint	AI 使用提升即时知识测验	一周后仍有增益；解释型使用优于自动化型使用	迫使判据从 AI 用量转向交互中结构与边界由谁持有

最近邻	与本章重合处	判决性分离线
Teaching for Understanding	理解表现为灵活使用	高灵活表现者在边界诱饵上仍可能继续套模
Understanding by Design 六面	解释、应用、视角、自知均属于理解	是否必须设置“成功结构应被撤回”的任务
适应性专长	创新、效率、继续学习	创新可在错误框架内发生；撤回应解释额外方差
表征可操纵性（Wilkenfeld）	理解依赖可操纵表征	高操纵性不必然包含表征的失效/撤销条件
de Regt 情境性科学理解	理解是技能且依赖语境	本章要求可编码的边界破坏事件

		与首次撤模
互动式专长	不必亲自完成全部实践也可理解	语言流利不等于边界所有权；可用边界诱饵区分
AI 辅助科学理解（Krenn et al.）	正确预测≠理解，AI 可支撑理解	本章专门处理人机系统中理解归属于人的最低证据

可证伪命题	**支持本章会看到**	**削弱/推翻会看到**
撤回力有增量效度	控制迁移、元认知、智力后仍预测边界错误与延迟模型选择	加入控制变量后效应趋近零
展开力与撤回力可分	存在稳定的“高展开—低撤回”个体/状态	两者几乎完全同一维度
AI 使用方式影响辖域形成	解释型/反思型使用改善边界识别，自动化型不一定	只要 AI 在场，无论交互方式结果相同
边界任务不是普通难题	真边界与匹配难度伪边界产生方向性差异	差异完全由题目难度、焦虑或谨慎解释
边界识别有现实预测性	预测之后的异常处理、模型切换与错误恢复	只提升实验题正确率，不预测后续行为

本章说明

研究伦理：本章未直接招募研究参与者，未使用可识别个案。未来若实施对抗性 AI 或边界诱饵实验，应限于低风险、可逆、可充分复盘的模拟任务；不得以故意生成危险医疗、法律或安全建议的方式在真实服务中测试人员。任何将结构辖域读数用于高利害教育或职业决定的实践，都应公开编码规则并提供人工复核。

人机协作：研究问题、三学科选择、发表场景与核心方向由作者提出；生成式人工智能参与公开资料检索、跨学科结构推演、反例搜捕、初稿生成、参考文献整理与文档排版。理论判断、证据取舍、引文核验、伦理边界与最终发表责任由最终署名作者复核并承担。生成式人工智能不列为作者。

文本规模说明：本稿当前文档中文字符约 20,130 个（含标题、摘要、正文、参考文献中文信息与说明；英文与数字未计入该数），正文体量按约 2 万字论文设计。

第三编　责任与验证

前两编问的是人还剩下什么，本编问的是制度还剩下什么。

第七章之所以放在能力诸章之后，有一个具体理由：其余各章在设计断口测试时，都默认了被测者拥有改向与干预的权限。现实中大量岗位没有这个权限。一次"未能改向"的读数，可能记录的是制度不授权，而不是个人无能力；不先把这两者分开，本书全部的能力判定都会系统性地把制度缺陷记到个人头上。这一章给出的三要素——认知可达性、因果控制杠杆、改写权限——正是用来分开它们的工具，同时也指出了当代人工智能治理最深的那个缺口：真正危险的不是无人签字，而是有人签字、有人被责，却没有人能够让下一轮不同。

第八章处理另一半。人工智能把生产变便宜之后，相信反而变贵：若完整重做，效率增益在验证环节被抵消；若只凭流畅、引用数量或少量抽查就接受，关键错误可能恰好藏在没检查的地方。这一章的回答不是"检查得更多"，而是把验证从覆盖率问题改写成风险问题——攻击那些一旦错误就会改变行动的节点，用与原生成路径失效模式低相关的程序，直到残余的未核错误不再足以翻转结论为止。

第七章

有人签字，无人改写

末端责任压缩与改写责任空位　——组织社会学 × 责任法/侵权法 × 社会心理学的跨学科对撞

生成式 AI 组织中的责任空心化与改写责任空位

——组织社会学 × 责任法/侵权法 × 社会心理学的跨学科对撞

摘　要

当生成式人工智能越来越深地进入信息筛选、证据整理、风险预测、方案生成、判断建议与执行环节，组织通常以“最终仍由人确认、签字或批准”来维持责任结构。然而，形式上的人工确认并不自动产生清晰责任。AI 提供者可以主张系统只是辅助工具，部署组织可以主张最终决定由员工做出，工作人员可以主张自己只是依照既定流程和系统建议履职，管理者则可能把人工审核环节的存在视为治理义务已经完成。结果是：每一个节点都拥有局部责任语言，却没有任何一个节点同时掌握足够的信息、因果控制力与修改权限去解释错误为何发生并改变下一轮系统。本章采用跨学科互否方法，将组织社会学的角色分化、局部理性与“多手问题”，责任法/侵权法中的控制、可预见性、因果归责与注意义务，以及社会心理学中的责任扩散、行动者感和道德脱离置于同一发生链中。第一层对撞显示：复杂分工把一次完整判断拆成相互依赖的局部任务，法律与制度却倾向把可见责任压缩到末端行动者，心理层面又因 AI 预成形建议与受限选择降低个体的主观作者感。第二层对撞进一步发现，真正缺失的不是“有没有人承担后果”，而是“有没有主体拥有改写下一轮系统的完整责任”。本章据此提出“末端责任压缩—改写责任空位循环”（Terminal Responsibility Compression–Revision Vacancy Loop, TRC-RVL），并定义“改写责任”为：主体在错误或近失误出现后，能够基于足够的认知可达性识别机制，拥有足够的因果控制力阻断同类路径，并具有现实修改权限改变数据、模型、流程、阈值、岗位分工或升级规则的责任能力。本章提出 ECR 三要素——Epistemic access（认知可达性）、Causal-control leverage（因果控制杠杆）与 Revision authority（改写权限）——作为完整改错责任的最低结构。研究进一步解释为何长期无事故运行会通过成功归因、流程合法化和角色固化强化这一空位，并提出责任映射、事故前置审计、近失误保留、双向追责接口与“可改写监督”等制度设计。本章的核心结论是：AI 时代真正危险的责任缺口，并非没有任何人被写进责任表，而是责任被切割成许多局部片段，最终留下一个“有人签字、有人被责，却无人能够改写下一轮”的制度性空洞。

关键词：生成式人工智能；责任归属；组织性无人负责；多手问题；侵权法；责任扩散；行动者感；道德脱离；人类监督；改写责任

问题提出：为什么“最后有人签字”仍然可能没有真正的责任主体

生成式 AI 进入组织以后，责任设计最常见的直觉是保留一个人工末端：机器可以推荐，机器可以排序，机器可以预测，但最后必须由自然人点击确认、签字、批准或执行。这个设计看起来兼顾了效率与问责——AI 承担信息处理，人承担最终责任。然而，一旦把真实工作流程展开，就会发现“最后动作在人”与“责任结构完整”是两回事。末端签字者可能只看到 AI 已经高度压缩过的结论；模型提供者掌握模型训练与技术边界，却不知道具体业务情境；组织管理者决定采购、阈值和流程，却不逐案参与；数据团队影响输入与分类，却不承担最终决定；受影响者更无法接近系统内部。责任没有消失，而是被拆散了。

这种拆散会制造一个悖论：形式责任越明确，实质责任反而可能越模糊。组织可以在制度上写明“最终由业务人员负责”，于是审计表格中存在责任人；但一旦事故发生，真正的问题并不是找到一个名字，而是回答四个机制问题：哪些信息被筛掉了？哪些模型约束塑造了选项？谁设置了风险阈值与默认路径？谁有能力改变这些设置？如果末端责任人对前三个问题没有足够可见性，对第四个问题又没有权力，那么把责任压到他身上只完成了“归责”，没有完成“改错”。

AI 使这个问题比传统官僚制更尖锐，因为生成式系统不仅执行规则，还能够参与问题表述、证据摘要、理由生成和方案构造。人看到的往往不是未经加工的事实，而是一个已经被模型组织过的认知世界。当人只在高度成形的建议上做最后确认时，他的因果参与与心理作者感都可能被压缩；与此同时，组织又利用这个人工确认点证明“没有把责任交给机器”。因此，末端人工审核可能同时承担两种相反功能：对外证明责任仍在人，对内却使真正的上游控制结构免于被重新审视。

本章要解决的不是“AI 出错到底该怪谁”这一事件后追责问题，而是一个更靠前的机制问题：为什么高效率所需的分工、自动化与专业化，会持续生成一种形式责任存在、改写责任缺位的结构？为什么这种结构在没有事故时尤其稳定？为什么一旦出事，每个主体都能提出局部上合理的免责叙事，而系统却缺少一个能够把信息、控制与修改重新汇聚起来的节点？

研究问题与跨学科互否方法：从“谁负责”转向“谁能改变下一轮”

本章采用跨学科对撞，而不是简单的多学科并列。第一阶要求三门学科分别解释同一个现象：为何责任会在复杂 AI 组织中变得碎片化。组织社会学关注角色分化、局部理性和多主体协作；责任法/侵权法关注何种条件下可以把损害规范性地归给某个主体；社会心理学则关注多人和技术中介如何改变个体的责任感、行动者感与自我约束。三者分别提供结构、规范与心理三个层次。

第二层对撞不再把三套解释平行保留，而要求它们互相指出不足。仅用组织社会学，会得到“责任分散是复杂系统常态”，却难以说明哪种分散在规范上不可接受；仅用侵权法，会倾向寻找可预见性、控制或注意义务，却可能把动态社会技术系统压缩成事件后的因果链；仅用社会心理学，则容易把责任缺口解释成个体的推责或道德脱离，而忽略很多工作人员事实上没有足够权限改变系统。

本章因此提出一个新的研究单位：不是事故发生后“谁应该被责”，而是事故发生前后“谁拥有让下一轮不同的责任能力”。这把责任从一次性归责转变为循环性治理。一个制度即使能够在每次事故后找到被处罚的人，如果被处罚者不能改变数据、模型、流程或权限，下一轮事故的生成条件仍然保留。相反，一个真正负责任的系统必须让责任与改写能力发生耦合。

源学科一：组织社会学——分工提高效率，也把完整判断拆成局部理性

从专业化到局部理性：每个人都可能“正确地完成自己的那一段”

现代组织之所以能够处理复杂任务，依靠的正是分工。风险模型由数据团队开发，产品团队负责交互，法务设定合规边界，业务人员负责案例处置，管理者负责资源和绩效，供应商负责基础模型。每个角色面对的目标函数不同：数据团队看准确率和稳定性，业务部门看时效与吞吐量，管理者看成本和绩效，法务看可辩护性，末端操作员看是否遵守流程。局部看，这些行为都可能是理性的；整体看，却可能生成无人主动选择的风险。

“多手问题”（problem of many hands）正是复杂组织责任研究的经典难题：一个结果由许多人的局部贡献共同生成，以致很难把完整责任精确归给某一只“手”。Thompson 将其用于公共组织责任，后续研究进一步强调，困难不只在事故后“找不到罪魁”，还在多主体系统中缺少足以调查和纠正组织失败的责任设计。[1] [2] AI 把“多手”扩展成“多人—多模型—多接口—多供应商”的混合链条。

关键并不是参与者数量增加本身，而是信息和控制在分工中同时被分区。末端人员知道个案，但不知道模型为何这样表示个案；模型工程师知道技术，但不知道该建议如何进入具体组织流程；管理者决定指标，却看不到每次指标如何改变一线选择。于是没

有人拥有完整视角，而组织又容易把“没有完整视角”误当成“没有必要拥有完整视角”。

组织性无人负责：责任不是消失，而是被角色边界吸收

所谓“组织性无人负责”不是指组织里每个人都没有职责，而恰恰是每个人都有职责，但职责被定义得足够局部，以致整体后果无法被任何一个角色完整承接。供应商可以说模型符合合同规格；采购部门可以说完成了尽职调查；经理可以说制度要求人工复核；操作员可以说自己依流程审核；合规部门可以说文档齐全。每一句话单独看都可能是真的，它们组合起来却无法回答：为什么这个系统会稳定地产生某类错误？

这与 Vaughan 对复杂组织中“偏差正常化”的分析存在结构同构：长期没有严重事故，会使原本带风险的做法逐渐被解释为可接受的正常实践。对 AI 组织而言，持续顺利运行会进一步证明现有角色分工“有效”：供应商没有严重投诉，业务指标提高，人工确认率看起来正常，于是没有组织动力重新连接被分散的信息和权限。成功不只是结果，更成为责任架构的合法化证据。

因此，事故前的组织责任问题不是“没人做事”，而是“没有一个位置被设计为必须把跨节点信息重新拼起来”。只要组织绩效良好，各角色就会继续沿局部目标优化。真正需要跨界的问题——数据变化是否使模型边界改变、人工是否还有真实干预空间、某种错误是否在不同部门以不同名称重复出现——很容易因为“不属于任何单一岗位”而长期无人处理。

无事故运行如何形成路径强化

复杂系统最大的责任错觉之一，是把“没有发生严重事故”理解成“责任设计有效”。实际上，无事故可能来自系统本身可靠，也可能来自一线人员额外补救、受害影响未被统计、问题被局部吸收，或者风险尚未跨过可见阈值。长期成功会降低组织对结构性改造的需求，并增加更深自动化的正当性。

每一次顺利运行都会强化既有路径：更多任务交给 AI，更多人工岗位转为末端审核，更少人保持上游知识，供应商接口进一步标准化，绩效制度继续奖励吞吐量而非异常发现。于是，当系统终于在新情境下出错时，组织不仅面对一次错误，还面对多年成功所造成的“责任能力退化”：知道全链条的人更少，能修改上游的人离现场更远，末端人员却因为签字记录最清晰而最容易成为责任落点。

源学科二：责任法与侵权法——归责需要的不只是最后行为，而是控制、可预见性与因果联系

为什么责任法天然不满足于“谁最后碰了一下结果”

侵权法和广义责任法的基本直觉之一，是责任不能只依据时间顺序。最后一个行为者未必是最重要的责任主体。因果关系要求考察行为与损害之间是否存在规范上可归责的联系；可预见性影响一个主体是否应当在事前采取预防措施；控制能力影响其是否有现实机会避免损害；注意义务则把特定角色在特定关系中的谨慎要求制度化。不同法域在具体规则上差异很大，但这些结构共同拒绝“最后动作即全部责任”的朴素逻辑。

AI 系统使传统要素受到挑战，是因为因果链和控制链不再重合。供应商可能对模型能力具有高控制，却对具体个案低控制；部署组织对使用场景和流程有高控制，却无法完全预测生成式模型输出；操作人员对最后点击有直接控制，却对选项集合、默认阈值和输入信息低控制。于是，谁“造成”损害、谁“能够预见”、谁“能够避免”开始分别指向不同主体。

这意味着责任法真正需要面对的是“控制的分层”。不能因为末端人拥有最后一票，就推定他拥有对全部生成条件的控制；也不能因为开发者不参与单次决策，就推定其对系统性缺陷没有责任。责任必须沿着可预见风险与可控制环节向上游展开，而不是只向末端压缩。

AI 时代的控制悖论：最容易被看见的人，往往不是最能改变系统的人

在组织证据中，末端确认往往最容易留下记录：谁登录了系统、谁点击了批准、谁签了名字。这些记录具有强可见性，因此在事故后天然吸引归责注意。然而，真正塑造行动空间的控制往往更隐蔽：模型供应商确定训练与安全边界，组织确定采购与使用目的，流程设计者决定什么时候必须接受或可以拒绝，管理者设置时限和绩效压力，界面设计决定哪些解释默认展开、哪些警告被折叠。

由此形成“可见性—控制力倒挂”：责任证据最清楚的人，可能恰恰拥有最弱的结构控制；拥有强结构控制的人，其贡献分散在代码、合同、策略和管理决策中，不容易被单次事故直接看见。Elish 提出的“道德溃缩区”（moral crumple zone）指出，复杂自动化系统中的最近人类操作员可能承受与其实际控制不相称的道德和法律责难。[3] 这为 AI 末端签字结构提供了重要警示。

因此，真正公平而有效的责任制度不能只追踪动作日志，还要追踪“谁能改变什么”。如果某主体能够改变模型、数据、阈值、流程或人员权限，其预防与修正责任应与这种控制能力相匹配；如果某主体只有形式批准权，其责任范围就不应被无限扩张。

从赔偿责任到预防责任：为什么事故后找人赔还不够

侵权法的重要功能包括补偿与威慑，但 AI 组织的复杂性要求进一步关注“修正功能”。如果一次事故最后由某个员工承担纪律责任、由保险承担经济赔偿，却没有任何机制把事故知识转化为模型、流程和岗位的改变，那么归责完成了，风险治理却没有完成。

2024 年的欧盟产品责任新规则把软件纳入现代化产品责任框架，并考虑产品持续学习或获得新特征等因素；欧盟《人工智能法》则沿 AI 价值链区分提供者、部署者等义务，并要求高风险 AI 的人类监督能够理解系统能力和局限、监测异常、警惕自动化偏差、解释输出，并在特定情况下忽略、推翻、逆转输出或停止系统。[4][5] 这些制度方向表明，AI 责任正在从单纯事后归责转向生命周期控制。

但本章进一步指出，即便法律把义务分配给多个角色，也仍可能留下“改写空位”：每个角色只需要履行自己局部义务，却没有任何一个角色对跨节点重复错误的系统重构负责。因此，责任设计还必须回答：谁负责把事故从“某次失误”升级为“需要修改系统结构的问题”？

源学科三：社会心理学——当行动被预成形，主观作者感与责任感会发生什么

责任扩散：参与者越多，个人越容易认为“还有别人负责”

责任扩散最经典的社会心理学结构是：当潜在行动者增加时，个体感受到的个人干预义务可能下降。AI 组织中的“其他行动者”不仅是同事，也包括算法、供应商、审批流程、合规规则和管理层。个体在判断时不再面对“这是我一个人的选择”，而面对“模型已经推荐、制度已经采购、上级已经批准使用、合规已经审核、我只是最后确认”。

这并不意味着工作人员有意推责。相反，责任感下降可能是对真实结构的心理反映：当一个人的选择空间被预先压缩、信息被上游过滤、系统建议被赋予权威时，他确实不再是完整行动作者。问题在于，组织又可能在事故后恢复“你是最后决策者”的个人主义叙事，从而造成行动时责任扩散、出事后责任集中。

行动者感：为什么“我点击了”不等于“我感觉是我造成的”

行动者感（sense of agency）是个体体验“我的行动造成了这个结果”的主观感受。自动化和人机协作研究长期发现，随着系统自主性提高，人的行动者感可能改变。近期实验也显示，在道德决策中与 AI 互动会影响人的显性责任判断；参与者在 AI 协助下可能报告更低的责任感，即使制度上最终行动仍由人完成。[6]

这种变化在生成式 AI 中尤其值得注意。传统自动化可能只给一个报警灯，而生成式 AI 可以形成完整建议、列出理由、拟定行动计划，甚至提前生成要签署的文本。人越晚进入决策流程，其心理上越容易把自己理解成“审核者”而不是“作者”。如果他的实际可选行动又被界面或制度限制，主观责任进一步下降。2026 年的一项预印本实验发现，当智能支持系统把监督者可选行动压缩到一个选项时，事故后参与者感到的道德责任更低，这提示“名义监督”与“实际责任条件”可能分离。[7]

因此，组织不能假设“只要最终点击是人的，人自然就会保持责任感”。责任感依赖真实的行动空间、因果可理解性和干预能力。把人放到流程末端，却不给他足够选择和控制，既可能降低监督质量，也可能让事故后的强制归责显得不公。

道德脱离：流程语言如何让责任从“我的行动”变成“系统结果”

Bandura 的道德脱离理论指出，人可以通过责任转移、责任扩散、结果淡化等机制降低自我谴责。[8] 在 AI 组织中，技术语言和流程语言为这些机制提供了天然载体：“系统评分如此”“模型建议如此”“按政策必须这样”“我只是执行”。这些表述并非必然虚假，但如果长期成为默认话语，个体会越来越少把系统性后果纳入自己的道德自我调节。

需要强调，本章不把道德脱离当作个体道德缺陷。很多时候，组织结构先把个人的真实控制力削弱，心理上的责任降低只是随后发生。如果治理只通过伦理培训要求员工“更有责任心”，却不增加信息访问、干预权和修改路径，就会把结构问题重新道德化为个人问题。真正有效的责任设计必须同时改变心理与制度条件。

第一层对撞：三门学科共同揭示“责任四分离”

三门学科第一次碰撞后，可以把 AI 组织中的责任困境拆成四种彼此分离的东西。第一是因果贡献：谁的行为、数据、模型或规则对结果产生影响。第二是信息可见：谁知道这些影响如何形成。第三是控制能力：谁有能力在事前或事中阻断路径。第四是形式归责：事故后制度最容易把责任落到谁身上。传统小规模决策中，这四者往往高度重叠；AI 组织中，它们可以分散到完全不同的角色。

这种“责任四分离”解释了为什么每个角色都能说出一部分真话。供应商确实不做最后个案决定，但它控制重要技术条件；末端人员确实做最后确认，但他可能不了解上游模型；管理者确实不逐案操作，但他控制资源、指标和流程；组织确实保留人工审核，但人工可能没有充分干预空间。问题不是找一句话证明某一方“撒谎”，而是发现责任结构已经不再能够被单一角色叙事完整描述。

由此，末端签字制度存在一个危险的压缩效应：它把原本分布在价值链上的复杂责任关系压缩成一个最容易记录的末端动作。这个动作成为对外问责的接口，却可能遮蔽真正具有预防价值的上游控制。

第二层对撞之一：从“责任分散”推进到“末端责任压缩”

如果责任只是均匀分散，那么事故后应该表现为所有人都只承担一点责任。但现实常出现另一种结构：平时责任分散，出事后责任突然向末端人集中。这不是简单的责任扩散，而是“扩散—压缩”双阶段机制。

事故前，组织通过分工把认知、控制和行动拆散；事故后，调查和问责需要一个明确对象，于是最可见的签字记录、操作日志和审批动作成为归责锚点。末端人员因其动作可证、身份清晰、处于组织控制范围内而最容易被定位。这样，复杂系统的责任不确定性被一个具体人吸收，组织获得了“责任已经解决”的表象。

这就是本章所称的“末端责任压缩”（terminal responsibility compression）：制度把多节点生成的责任关系压缩为末端可见的人类动作，使一个人承担超过其信息与控制范围的解释和后果责任。它与“道德溃缩区”相近，但本章强调的不只是事故后替罪，还包括这种压缩如何反过来阻止上游结构被修改：既然事故已经被解释为某个审核者的失误，就没有必要重新设计模型接口、绩效指标或异常升级路径。

第二层对撞之二：从“谁应被责”推进到“谁拥有改写责任”

这里出现本章最重要的概念转换。传统责任讨论经常围绕“谁应该承担后果”：谁道歉、赔偿、被处罚、被追责。本章把另一种责任独立出来，称为“改写责任”（revision responsibility）：当错误、近失误或系统性异常出现以后，谁有义务且有能力改变生成错误的条件，使下一轮运行真正不同。

改写责任不是一般意义的“负责整改”口号。它至少要求三个条件同时存在。第一，认知可达性（Epistemic access, E）：责任主体能够获得足以重建事故机制的信息，包括输入、模型行为、流程、上下游决策与情境限制。第二，因果控制杠杆（Causal-control leverage, C）：主体对导致风险的关键环节拥有实际干预能力，而不仅能提出意见。第三，改写权限（Revision authority, R）：主体被制度授权修改数据、模型配置、流程、阈值、岗位权限、培训或升级规则，并能要求其他节点协作。

若只有 E 没有 C，主体成为“知道但无能为力”的观察者；只有 C 没有 E，主体可能盲目调整；只有 R 没有前两者，管理者拥有形式权力却不知道改什么。完整改错责任

需要 E、C、R 发生耦合。可简化表示为：RR = E ∧ C ∧ R。这里的合取关系意味着，任何一项长期缺失都可能产生“有人负责整改、却改不了系统”的假整改。

这一模型把责任法的控制与可预见性、组织社会学的信息分割以及社会心理学的行动者感连接起来：只有当一个主体既看得见、动得了、改得成，他才有条件形成稳定的责任作者感，也才真正能够为下一轮承担治理责任。

第二层对撞之三：为什么无事故运行会制造“责任结构正确”的错觉

责任空位最难被发现的时期不是事故频发时，而是系统运行良好时。每一次无事故运行都会提供一种经验性证据：“当前流程没问题”。人工确认的存在又进一步提供规范性证据：“人仍在监督”。两种证据叠加，使组织很少主动检查 ECR 是否真的集中在任何责任节点。

然而，无事故并不能证明末端监督有效。可能是 AI 本身足够准确，可能是任务环境长期稳定，也可能是一线人员通过额外劳动默默修补错误；还可能是错误影响没有进入组织的绩效指标。只要严重后果没有显现，组织就没有动力追问末端人员究竟能看见多少、能控制多少、能修改多少。

长期成功还会产生心理强化。工作人员越来越把 AI 视为正常基础设施，管理者越来越把人工审核视为形式保障，供应商越来越把“客户持续使用”视为产品可靠性的证据。于是对系统的质疑成本上升，异常更容易被解释为个案。直到一次严重事故迫使责任链被重新打开时，组织才发现：那个被制度指定为“最终负责人”的人没有技术信息，上游技术团队没有个案权力，管理者没有保留足够日志，供应商合同又把具体使用责任推给客户。

理论模型：末端责任压缩—改写责任空位循环（TRC-RVL）

综合两阶碰撞，本章提出“末端责任压缩—改写责任空位循环”（Terminal Responsibility Compression–Revision Vacancy Loop, TRC-RVL）。模型不是描述单次事故，而是解释责任空洞如何在长期成功中自我强化。

第一步，效率驱动的功能分解。组织为了速度、规模和专业化，把判断拆成信息获取、模型处理、方案生成、审核、执行等节点。第二步，认知与控制分散。各节点只保留完成局部任务所需的信息与权限。第三步，AI 建议预成形。随着 AI 能力提高，末端人工面对的不是开放问题，而是高度成形的推荐和有限选项。第四步，形式责任末端化。制度通过签字、确认和审核记录把“最终责任”落到最末端自然人。

第五步，主观作者感下降。由于建议来源外部化、行动空间受限、多人共同参与，末端人员对结果的主观责任感可能下降，其他上游主体也因远离最后动作而降低责任体验。第六步，无事故成功强化。高准确率、高效率和少事故被解释为现有治理有效，组织进一步扩大 AI 外包并固化角色边界。第七步，ECR 继续分离。更深自动化使能看见全链条的人更少，能修改系统的人离具体后果更远。第八步，事故触发末端压缩。当严重错误出现，组织依赖可见签字寻找责任对象。第九步，改写责任空位。被责者无 ECR，上游各方又只拥有部分 ECR，事故最终通过处分、赔偿或个案修补结束，而非系统改写。第十步，循环重新开始。

这个循环解释了一个反直觉现象：越是长期“运行正常”的 AI 组织，责任架构越可能在看不见的地方变脆。因为成功减少了跨节点审查，人工监督又提供了责任已被保留的幻觉。真正的责任风险不是日常流程里没人签字，而是事故后无法找到一个能让下一轮不同的责任节点。

“签字—控制错配”：为什么最终确认权不是充分责任条件

在很多制度中，“最终由人确认”被视为责任保留的核心。但是确认权只回答一个二元问题：此刻是否放行。真正责任至少还包括问题定义权、证据访问权、选项重构权、阈值调整权和流程暂停权。如果这些权力都在上游，末端确认者只有接受或拒绝两种选择，他的控制实际上非常有限。

更严重的是，拒绝本身也可能成本高昂。员工若频繁否决系统建议，需要额外说明、承担延误、面对绩效压力；接受系统则快速、符合默认流程。因此，组织虽然形式上给予 override，行为环境却把“不使用 override”变成局部最优。此时事故后再说“你本来可以拒绝”，忽略了真实控制是制度、信息与激励共同塑造的。

所以本章提出“签字—控制错配指数”的概念：一个岗位形式归责强度越高，而其 ECR 实际持有程度越低，错配越严重。高错配岗位最容易成为道德溃缩区，也最容易在长期运行中出现防御性审核——员工为了自保增加形式记录，却未必提高系统风险识别。

“改写责任”与传统责任的区别：责任不只面向过去，还必须面向下一轮

传统事件责任往往具有过去指向：解释已经发生的损害、分配赔偿、确认过错、施加制裁。改写责任则具有未来指向：把事故知识转化为下一轮系统变化。两者不能互相替代。一个组织可以在事故后非常严厉地处分员工，却在改写责任上完全失败；也可以通过系统改造有效预防复发，却不必把全部道德谴责集中到个人。

改写责任尤其适合社会技术系统，因为同一类事故往往由多个小因素耦合而成。单纯寻找“谁犯错”容易促使组织围绕个人做修补：加强培训、再次强调制度、增加签字。真正的系统改写则会追问：为什么这个错误能够一路穿过数据、模型、界面与监督而没有被捕捉？哪些早期信号存在但被忽略？哪个角色本应拥有重启权却没有？

因此，AI 责任成熟度不应只用“出了事能不能找到责任人”衡量，还应看“出了事之后谁能改变系统”。如果前者强而后者弱，组织拥有的是惩罚能力，不是学习能力。

这里需要与本书前几章划清一条界，因为四章都在处理"最后那个人算不算数"。

与第五章的分界最要紧。第五章问的是判断是否由人形成——它关心的是主体是否持有决定性理由、能否识别失效、能否重新发起。本章问的是错误发生之后是否有人能改变下一轮，它关心的是认知可达性、因果控制杠杆与改写权限三者是否落在同一处。这两件事可以完全分离，而且现实中经常分离：一个人可能完全满足第五章的三节点——他确实持有理由、确实识别了边界、确实主动重开了判断——却完全没有改写权限，因此在本章的意义上不是改写责任的承担者。反过来，一个握有全部改写权限的高层管理者，可能对具体个案的判断毫无参与。

这条分界给本书带来一个必须写明的警告，它同时约束其余各章：本书其余各章在设计断口测试时，默认了被测者拥有改向与干预的权限。现实中大量岗位没有这个权限。在这些岗位上，一次"未能改向"的读数记录的是制度不授权，不是个人无能力。凡实施本书方法的组织，必须先用本章的三要素把这两者分开，否则本书全部的能力判定都会系统性地把制度缺陷记到个人头上。这是本书方法最容易造成的伤害，也是本章之所以必须放在能力诸章之后的原因。

现有 AI 治理框架给出了哪些答案，又留下什么缺口

现有治理已经明显认识到单一末端责任不足。NIST AI RMF 要求组织清晰定义并区分 AI 相关角色与责任，明确风险管理的沟通线和高层责任，并指出 AI 生命周期包含许多相互依赖活动，负责某一环节的 AI 参与者常常无法完全看见或控制其他环节。[9] 这与本章的“责任四分离”高度一致。

欧盟《人工智能法》则更具体地把高风险 AI 的人类监督设计为一种能力与权限组合：监督者应适当理解系统能力和局限，能够监测异常，保持对自动化偏差的警觉，正确解释输出，并能在具体情境中不使用、忽略、推翻、逆转或中止系统。[5] 同时，法案还沿价值链区分提供者、部署者等义务，要求日志、监测和风险处理。

这些框架提供了重要基础，但从 TRC-RVL 看仍需进一步提出“跨节点改写责任”。角色清晰并不保证角色连接；人工能停止系统并不保证有人能修改导致事故的上游机制；提供者与部署者分别合规也不保证两方之间存在强制学习接口。因此，治理需要从“职责分配表”进一步走向“改写闭环图”。

组织设计一：用“责任地图”替代单一 RACI 表

传统责任矩阵常把岗位标成 Responsible、Accountable、Consulted、Informed，但 AI 系统需要额外记录 ECR。对每一个关键风险，不仅要写“谁负责”，还要写：谁拥有事故解释所需的数据和模型信息（E）；谁能够阻断该风险的因果路径（C）；谁有权限修改相应系统和流程（R）。

责任地图应沿 AI 生命周期展开：数据来源与标注、模型选择与更新、提示模板、工具权限、输出过滤、业务阈值、人工审核、执行接口、投诉与事故反馈。每个节点至少标记一个“观察者”和一个“改写者”，并检查是否存在大量节点只有观察没有改写、只有改写没有信息。

尤其需要识别“末端高归责—低 ECR”岗位。如果一线员工必须对重大后果签字，却不能查看关键证据、不能理解模型限制、不能暂停流程或发起技术复核，那么组织应该降低其独占责任，或提高其真实 ECR，而不是仅增加免责声明和培训。

组织设计二：把近失误变成责任结构的传感器

严重事故太稀少，不足以持续检验责任结构。近失误、人工纠错、AI 与人分歧、系统异常但未造成损害的事件，反而是发现 TRC-RVL 的最好材料。因为它们能显示：是谁最先看见问题，谁能暂停，谁需要跨部门求助，谁最终能改动系统。

如果一个组织只统计最终事故率，它会错过大量被人类默默吸收的系统缺陷。相反，应该保留“人工救回”记录：AI 给错建议但员工纠正、系统遗漏但客户投诉补充、业务人员绕过默认流程避免损害。这些事件不是员工效率低，而是责任系统的重要诊断信号。

近失误分析的核心不是奖励或处罚某个人，而是检查 ECR 是否顺畅流动。发现问题的人能否把证据送到有控制的人？有控制的人能否获得修改授权？修改以后是否反馈给发现者？只有形成这种闭环，组织才真正拥有改写责任。

组织设计三：从“人类审核”升级为“可改写监督”

传统人类审核往往只要求人工确认输出。可改写监督则提出更高标准：监督者不只能够对一次输出说“不”，还能够触发对系统条件的检查和修改。换言之，监督的单位从“单条结果”上移到“生成结果的机制”。

可改写监督至少包括五个接口：异常标记接口、证据展开接口、模型/规则升级接口、跨角色复核接口和暂停/降级接口。任何高风险系统都应让一线人员知道：如果我认为这不是普通个案，而是系统模式，我该把它送到哪里；谁必须回应；什么条件下必须修改或停用。

如果组织只保留“接受/拒绝”，就把监督者困在输出层。即使他不断拒绝同类错误，系统本身也不会学习。这种监督本质上是人工过滤器，而不是责任主体。可改写监督的目标，是让人工异常最终能够改变机器与组织。

法律制度推论：责任应跟随可控制风险，而不是只跟随最后动作

TRC-RVL 并不主张取消末端人员责任，而是要求责任强度与真实控制相称。在法律与组织政策中，责任分配应更系统地考察：主体是否可合理预见相关风险，是否拥有影响该风险的技术或组织控制，是否能够获得必要信息，是否有能力采取成本合理的预防措施。

这意味着上游主体不能仅以“最终由客户或员工决定”为由退出责任。若某供应商或组织控制系统的关键限制，并可合理预见这些限制会在特定场景产生风险，其注意与协作义务就不应被一个末端签字完全吸收。反过来，末端人员如果无法访问信息或真正改变流程，也不应被视为拥有无限责任。

特别需要防止“合同式责任洗白”：技术合同把所有使用风险转给部署方，组织规章再把最终责任转给一线人员，最后形成责任层层下压，而信息和控制仍停留上游。有效制度需要检查这种责任转移是否伴随相应的 ECR 转移；没有能力转移的责任转移，应被视为结构性错配。

社会心理学推论：责任感不能只靠道德教育，而要靠真实作者性

如果一个工作人员长期只在 AI 生成的结论上做形式确认，组织再要求他“增强主人翁意识”，效果有限。行动者感来自真实因果参与：我能够看见我的选择如何改变结果，我拥有可替代路径，我的异常判断能够让系统停下来，我提出的问题能够进入下一轮设计。

因此，提高责任感最有效的方法不是增加口号，而是增加“可见影响”。例如让一线人员看到自己上报的异常最终导致了模型阈值调整，让审核人员参与边界案例复盘，

让技术团队直接接触真实受影响案例，让管理者对风险配置而非只对绩效数字负责。这样的设计把责任从抽象义务重新连接到行动后果。

同样，组织应该谨慎使用“AI 只是工具”这类话语。技术在法律上可能不承担人格意义的责任，但在心理与组织事实上，它可能深刻塑造选项和行动。若组织一方面让 AI 承担高度成形的判断工作，另一方面又用“只是工具”要求末端人员承担完整作者责任，会造成严重的责任体验与制度归责错位。

四种责任结构：不是“有责任/无责任”的二元区分

类型 A 最符合“最后有人签字”的表象：制度责任非常清晰，但清晰得过头，以致掩盖真实控制。类型 B 则是经典多手问题，没有一个明确落点。类型 C 通过把责任和控制集中到少数技术或管理岗位来解决模糊性，但容易让上游主体脱离具体情境。本章主张的类型 D 不是追求所有权力集中，而是让 ECR 在需要时能够跨节点汇聚，并为每类风险明确一个真正的改写责任节点。

因此，成熟责任架构不是“每件事只有一个责任人”，也不是“所有人共同负责”。前者可能制造末端溃缩，后者可能制造责任稀释。更好的原则是：日常任务可以分工，事故解释可以分布，但系统改写必须可定位。

可检验命题：TRC-RVL 如何被证伪或支持

命题一：末端人工确认率越高，不必然意味着实质责任越清晰。

若末端人员 ECR 低，增加签字和确认步骤可能只提高形式归责可见性，而不会提高事故解释或系统修正效率。

命题二：AI 建议越预成形，末端人员的显性责任感越可能下降。

尤其当可选行动数量减少、系统权威高、否决成本高时，主观作者感和责任感应下降。

命题三：无事故运行时间越长，组织越可能高估现有监督结构的有效性。

若长期成功导致异常上报率、人工分歧记录或跨部门复盘下降，则支持成功固化机制。

命题四：高“签字—控制错配”的岗位在事故后更容易承受不成比例的个人归责。

可通过比较岗位的形式责任强度、ECR 指数与实际处分/诉讼暴露来检验。

命题五：存在明确改写责任节点的组织，同类事故复发率应更低。

关键不是第一次事故率，而是事故后系统层面修正速度与同类问题是否减少。

命题六：责任培训若不伴随 ECR 提升，对高自动化岗位的长期责任感改善有限。

如果只做伦理培训而不增加信息、干预与修改权限，效果应明显弱于结构性授权。

实证研究设计：如何测量“有人签字却无人改写”

组织层责任链审计

研究者可以选择医疗、金融、人力资源、公共行政或软件工程中的 AI 辅助流程，对一次具体决策进行“逆向责任追踪”：从最终签字向上追踪数据、模型、提示、阈值、审批规则和绩效激励。每个节点分别测量 E、C、R，并记录谁承担形式责任。

如果大量流程表现为末端形式责任高，而 E、C、R 分别散落于不同节点，且没有事故后汇聚机制，就可以量化 TRC-RVL 的结构基础。进一步可以比较不同组织的同类风险复发率，检验 ECR 耦合是否具有预测力。

实验层：操纵 AI 预成形程度与选择空间

实验可以设置四种 AI 辅助条件：只提供原始信息、提供风险评分、提供完整建议、提供完整建议且限制可选行动。保持最终点击都由参与者完成，测量其行动者感、显性责任感、建议采纳率和错误后自我归责。若最终动作相同而责任感随 AI 预成形和选择限制显著变化，就说明“最后由人执行”不是心理责任的充分条件。

进一步可在事故后给予不同权限：只能解释、可以覆盖、可以修改系统规则。观察参与者是否更愿意调查上游原因并投入纠错。如果改写权限提高责任投入，支持 ECR 中的 R 不仅是制度变量，也是心理责任形成条件。

纵向层：无事故成功是否侵蚀责任结构

最有价值的研究应跨时间观察。让团队长期使用高可靠 AI，记录最初与数月后的人工质疑率、异常上报、独立检查、系统理解和 ECR 分布。随后引入分布外事件或系统故障，观察谁能接管、谁能解释、谁能发起修改。

如果高可靠 AI 长期成功显著降低跨节点学习和人工异常敏感度，同时增加事故后的末端归责倾向，就能直接验证 TRC-RVL 的“成功强化”环节。

边界压力测试：哪些情况不应被错误解释为责任空位

真正的全自动化不等于责任空位

如果组织明确决定某类低风险任务由系统自动完成，并把系统级责任清楚分配给提供者和部署组织，同时保留监测与修正机制，那么不存在“假装由人负责”的问

题。TRC-RVL 批评的不是自动化本身，而是用末端人工动作制造责任已经完整的幻觉。

末端人员拥有充分 ECR 时，签字可以是真正责任

某些专业岗位虽然使用 AI，但末端人员可以看到关键证据、理解系统局限、自由改变方案、暂停流程，并能触发模型或制度修改。此时签字与真实控制高度一致，把较强责任放在末端并不构成错配。

多人共同改写不等于必须有一个“超级负责人”

复杂系统不可能让单个人掌握全部知识。改写责任可以是团队性的，但必须有可执行的组合结构：谁负责汇聚证据，谁拥有技术修改权，谁能决定资源和政策，谁验证修改有效。本章要求的是 ECR 能够被制度化汇聚，不是回到英雄式单人治理。

并非所有事故都能被合理预见

责任法中的可预见性提醒我们，系统不能为一切不可想象事件承担无限预防责任。TRC-RVL 关注的是事故或近失误已经提供新信息以后，组织是否有能力把信息转化为下一轮修改。第一次真正不可预见的事故可能不可避免，但第二次同类事故是否仍发生，恰恰检验改写责任是否存在。

进一步理论产出：从“责任缺口”到“责任空心化”

既有“责任缺口”理论常强调自主学习系统使传统人类责任难以适用，例如制造者和操作者无法完全预测系统未来行为。[10] 本章提出“责任空心化”作为不同但相关的现象：名义上的责任并未消失，甚至可能被明确写在制度中，但其内部支撑条件——知识、控制、修改能力——被抽走。

责任缺口表现为“找不到合适责任主体”；责任空心化表现为“找得到一个名义主体，却发现他无法承担责任应该具有的治理功能”。后者更容易被组织忽视，因为合规文件看起来没有空白。末端签字、人工审核、责任声明甚至会掩盖空心化。

这一区分也解释了为什么单纯强化问责有时不能提升安全。若责任主体是空心的，增加处罚只会提高其自我保护行为，例如更机械地遵守 AI 建议、增加文书记录、避免承担异常处置，而不会提高系统学习。真正有效的问责必须与 ECR 一起增强。

六条制度原则：让责任从“可追责”变成“可改写”

原则一：责任跟随控制。

谁对某类风险拥有更强的结构控制和合理预见能力，谁就应承担相应预防与协作责任，不能全部向末端转移。

原则二：签字必须配套 ECR。

高影响签字岗位至少应获得与责任相称的信息访问、干预能力和升级修改权限。

原则三：异常必须能越级。

一线发现的系统性异常不能被困在本部门绩效链中，应有直达风险治理或技术改写节点的渠道。

原则四：近失误必须被保留。

不要只统计最终事故；人工救回、模型分歧和重复 override 是系统责任结构的高价值信号。

原则五：事故结案必须包含系统改写结论。

处分或赔偿不等于结案；需要明确说明哪些生成条件被改变、由谁验证、何时复查。

原则六：周期性检验人工监督的真实性。

不仅看“有没有人工”，还要通过模拟异常检验人工是否真正拥有理解、拒绝、停止、升级与改写能力。

结论：AI 时代最需要避免的，不是“无人签字”，而是“无人能够让下一轮不同”

生成式 AI 正在把组织责任推入一个新的阶段。过去，责任问题常发生在“谁做了这个决定”；现在，一次决定可能由数据、模型、组织规则、管理目标、人工审核和执行接口共同生成。把最后一步留给人，并不能让前面的复杂性消失。

组织社会学告诉我们，分工会产生局部理性和多手问题；责任法告诉我们，规范归责必须考虑控制、可预见性、因果联系和注意义务，而不能只看最后动作；社会心理学告诉我们，当行动被外部系统高度预成形、选择空间被压缩、参与者增加时，个体的主观作者感和责任感可能下降。三者的第一层对撞产生“责任四分离”，第二层对撞进一步产生“末端责任压缩”和“改写责任”两个核心概念。

本章提出的 TRC-RVL 模型表明，AI 责任空洞并不需要任何人主动逃责。它可以由一系列局部合理选择自然生成：为了效率而分工，为了安全而保留人工审核，为了可审计而要求签字，为了绩效而依赖高质量 AI，为了事故处理而寻找明确责任人。每一步

都看似合理，组合起来却可能形成一个制度悖论：责任在末端变得最清晰，而真正控制风险生成机制的能力却分散在上游。

因此，AI 时代责任治理的最低问题不应只是“出了事找谁”，而应增加一句：“谁有能力让下一轮不再这样发生？”如果答案始终找不到，那么无论责任表写得多么清楚、签字链多么完整，系统仍然存在责任空位。

真正成熟的责任结构，不是让一个末端人替整个社会技术系统承担道德重量，也不是让所有人共同负责从而无人负责，而是让信息、控制和修改权限在关键时刻能够汇聚到明确的改写节点。这样，责任才不只是惩罚过去的工具，而成为组织学习未来的机制。AI 时代我们最终需要保护的，不只是“最后仍有人签字”，而是发生错误以后，仍然有人——或一个明确可运作的责任共同体——能够看见发生机制、握住改变杠杆，并真正改写下一轮。

附录：ECR 责任审计的最小问题集

1. 认知可达性 E：当某类错误发生时，谁能够访问原始输入、模型输出、关键提示、阈值、人工修改、日志和组织情境？这些信息是否能在合理时间内被拼成完整事件链？

2. 因果控制 C：谁能够暂停使用、改变模型版本、调整阈值、增加人工复核、修改数据来源、改变界面或取消自动执行？这些控制在日常绩效压力下是否真的可用？

3. 改写权限 R：谁有正式授权要求供应商或内部团队实施修改？谁能决定资源、时间表和验证标准？如果跨部门意见冲突，谁拥有最终升级权？

4. 末端错配：最后签字者承担的形式责任与其 ECR 持有程度是否相称？若不相称，是降低责任，还是提高信息与权力？

5. 近失误闭环：人工纠正 AI、反复 override、客户投诉和异常绕行是否被记录？这些信号是否进入系统修改，而不是只被当作个案处理？

6. 事故结案：每次重大事故结束时，是否同时完成“个人/组织归责”和“系统改写结论”？如果没有改写，谁必须解释为什么无需改写？

7. 成功期复审：即使长期无事故，是否周期性模拟 AI 错误、信息缺失和情境突变，检验末端监督者是否仍然能够看见、停止、升级与改写？

附录二：十二个责任压力情景——检验“改写责任”是否真正存在

情景一：AI 建议正确率极高，末端人员长期从未推翻过系统

一个系统连续三年保持极高准确率，员工几乎从不需要推翻 AI 建议。管理者因此认为人工审核安排非常成功，因为每个案件都有自然人最终确认，且没有重大事故。TRC-RVL 却要求提出相反问题：三年里人工从未真正使用干预权，是因为系统始终完美，还是因为人已经逐渐失去质疑、停止和重启的能力？如果突然出现分布外案例，末端人员是否知道什么信号意味着“这次不能照常批准”？

这个情景揭示“低 override 率”的歧义。它既可能代表 AI 优秀，也可能代表监督退化。若组织只把低 override 率当作绩效，不通过模拟异常检测 ECR，就无法区分两种机制。真正的责任审计必须周期性制造安全的异常演练，确认人在长期不出错的系统旁边仍然具有真实接管能力。

情景二：末端人员发现模型错误，但只能在备注栏写说明

某员工发现 AI 把关键证据归入错误类别，导致风险评分偏高。制度允许他在备注中说明“不同意模型”，但最终业务流程仍按系统评分自动进入下一环节。他有认知可达性，却没有因果控制杠杆；即使事故后组织说“人工已经审核”，这种审核也没有真实改变结果的能力。

如果组织随后要求该员工为没有阻止错误负责，就出现典型 E-C 断裂。责任不能仅依据“你看到了”，还必须考察“你能否让你看到的东西产生实际影响”。否则，监督者会被要求承担一种没有行动通道的道德责任。

情景三：技术团队能修改模型，却看不到真实业务后果

另一个常见结构正好相反：模型团队拥有版本更新、阈值调整和提示策略的控制权限，却只能看到离线指标和匿名化错误码，不接触投诉、人工绕行和实际伤害。它拥有 C 和 R，却缺少足够 E。结果是技术团队能够不断优化模型，但优化目标可能与一线风险不一致。

这种结构说明“把责任交给技术团队”同样不够。改写责任不是简单把权力上收，而是必须建立稳定的信息回流。没有业务后果进入技术视野，技术控制再强也只是局部控制。

情景四：组织处罚末端员工后，同类错误仍然重复发生

一次事故后，组织认定员工“未认真审核”，给予处分并开展全员培训。半年后，另一名员工在类似系统建议下发生同类错误。如果两次错误共享相同的信息过滤、界面默认或绩效压力，那么第一次处分并未构成真正学习。

TRC-RVL 把“同类错误复发”视为改写责任空位的强信号。因为即使第一次归责完全合理，只要组织没有改变生成条件，责任制度就只完成了惩罚功能。事故调查的成熟度应以“是否改变下一轮”而非“是否找到责任人”衡量。

情景五：供应商声明“仅供参考”，组织却高度依赖模型

许多 AI 产品通过免责声明强调输出仅供参考、最终判断由用户负责。但若产品设计实际上把建议嵌入强默认流程，组织绩效又要求员工高速处理大量案例，那么“仅供参考”与真实使用结构可能完全不一致。法律文本中的责任边界不能自动改变行为中的控制结构。

如果供应商控制关键模型机制，组织控制采用方式，员工控制最后确认，那么责任应按各自可预见、可控制的风险分层，而不能由一句免责声明全部下移。尤其当供应商能够合理预见其系统会在某类高风险场景被深度依赖时，解释、监测和协作责任仍然具有现实意义。

情景六：管理者认为“有人审核”就等于完成治理

一个最典型的治理幻觉是把人工审核岗位本身当作风险控制。管理层完成采购评估、建立人工确认节点后，就把系统风险视为已经转移给业务人员。随着时间推移，管理者越来越少关注一线能否获得足够信息，也不检查否决 AI 的实际成本。

这使人工审核成为一种“治理封条”：只要流程图上有人，管理责任就看似已经尽到。TRC-RVL 要求管理者承担元责任——不仅安排人审核，还要验证审核者是否拥有 ECR。否则，人工环节只是责任显示器，而不是风险控制器。

情景七：多部门都能局部解释事故，却没有一份完整解释

数据部门解释输入为什么这样，模型团队解释输出为什么这样，业务部门解释为何接受建议，管理层解释为何设置该绩效目标。四份局部报告都准确，但没有任何报告把它们连接成同一机制。事故调查最后以“多因素共同作用”结案，却没有指出哪些连接必须被改写。

这种情况不是信息不足，而是“整合责任”缺位。ECR 中的 E 不仅是每个人拥有自己的信息，还包括某个责任节点被要求把信息拼成可行动的因果图。没有这种整合，即使组织数据非常透明，也可能继续处于无人完整理解的状态。

情景八：AI 建议只是“建议”，但否决 AI 需要额外十分钟

名义上，员工可以自由拒绝 AI；事实上，接受建议只需一次点击，拒绝则需要填写长表、寻找主管复核并承担延迟指标。长期下来，接受 AI 成为局部理性的默认行

为。事故后若以“系统允许人工否决”为由把责任全部推给员工，就忽略了组织通过摩擦成本塑造了真实控制。

这说明控制不能只按菜单权限测量，还要测量“可用控制”。真正的 C 应包含成本、时间、组织支持和报复风险。一个理论上存在、实践中几乎无法使用的 override，不应被视为完整控制。

情景九：组织把所有系统错误都归入“员工培训不足”

培训当然重要，但“培训不足”是复杂系统最容易使用的末端解释之一。它把结构性缺陷重新翻译为个体能力问题：信息界面不清楚，被说成员工没仔细看；模型边界不透明，被说成员工 AI 素养不足；绩效压力过强，被说成员工风险意识不足。

如果每次事故的整改都是“再次培训”，却很少改变技术和流程，说明组织的改写责任被固定在最便宜、最容易的杠杆上。真正的系统责任要求比较多种改写路径，而不是默认把人训练得更适应有缺陷的系统。

情景十：事故由真正不可预见的新型行为引起

生成式 AI 确实可能产生开发者和部署者都难以合理预见的新行为。第一次事故中，很难要求某个主体事先知道未知风险。此时传统责任中的可预见性边界仍然重要，不能把所有后果无限追溯给任何一个上游主体。

但事故发生以后，信息条件已经改变。系统现在知道了一个新失效模式。改写责任关注的正是这一转折：谁负责把未知风险转化为已知风险，谁更新模型、流程和使用边界，谁通知其他部署节点。如果第二次同类事故仍因为责任分散而没有修正，问题就从不可预见性转变为组织学习失败。

情景十一：团队共同拥有 ECR，但没人有最终协调权

有些组织会说“我们是共同负责的”。技术、法务、业务和风险部门都参与整改，看似比单一责任更成熟。然而，当各部门意见冲突时，如果没有明确升级和裁决机制，共同责任可能退化为共同等待：技术说风险可接受，业务说不能停用，法务说需要更多证据，管理层等待一致意见。

团队式改写责任必须有协调结构。不是要求一个人懂全部，而是必须明确谁负责触发复盘、谁汇聚信息、谁拥有临时停用权、谁在冲突时作出最终风险决定。否则“大家共同负责”仍然可能等于没有任何人对闭环负责。

情景十二：组织真正建立了改写闭环

一个成熟反例可以帮助界定理论边界。假设一线人员发现异常后能够立即标记；系统自动保存完整上下文；风险团队在 24 小时内审查；技术团队必须说明模型机制；管

理者拥有临时停用权；若确认系统性问题，则明确负责人修改模型或流程，并在重新上线前通过独立验证；最终结果再反馈给最初发现者。

在这种结构中，责任仍然是分布式的，却没有形成责任空位。末端签字者不承担整个系统的全部道德重量，上游主体也不能因远离最后动作而退出。ECR 能够跨节点汇聚并闭环，因此组织既能归责，也能学习。这个反例说明，TRC-RVL 并不反对复杂分工，它反对的是复杂分工中没有设计责任再汇聚机制。

附录三：理论区分——“责任洗白”“责任扩散”“责任空心化”不是同一件事

“责任扩散”主要描述多人情境下个人责任感或责任归属被稀释；“责任洗白”（responsibility laundering）通常强调主体借助技术或制度中介，把有争议的决定包装成客观、自动或他者决定，以逃避责任；“责任空心化”则是本章提出的结构性概念：责任标签仍然保留，甚至非常明确，但支持责任有效运作的知识、控制和修改条件被拆散。三者可以同时出现，却具有不同治理含义。

如果问题主要是心理责任扩散，可以通过角色明确、反馈和行动者感设计改善；如果问题是有意责任洗白，需要透明度、审计和规范约束；如果问题是责任空心化，仅仅强调“谁负责”或提高透明度仍可能不够，因为透明的信息未必伴随控制，控制也未必伴随改写权限。ECR 的意义就在于把三种条件同时纳入。

责任空心化还具有一个重要特征：它可以在所有参与者都善意的情况下生成。没有人需要故意逃责，只要组织不断把知识、控制和权力按效率逻辑分割，再把形式责任留在末端，就可能自然出现空心结构。因此，治理不能只寻找道德动机，而必须审计架构。

附录四：从事故调查到“下一轮改写证明”

本章建议高风险 AI 组织在重大事故或高价值近失误后，不仅形成事故报告，还形成一份“下一轮改写证明”。这份证明至少回答五个问题：一，旧系统中哪些信息、控制或权限配置参与了风险生成；二，哪些条件已经修改；三，谁拥有修改的正式责任；四，如何验证修改确实降低同类风险；五，如果不修改某一环节，为什么认为保持原状是合理的。

“下一轮改写证明”的价值在于迫使组织把事故从过去叙事转换为未来结构。传统报告很容易停在“根因是操作员疏忽”“根因是模型错误”“根因是沟通不足”；改写证明则要求每一个根因对应一个可观察改变。如果原因写进报告却没有任何系统参数、流程、权限或训练发生变化，就说明所谓根因分析并没有进入改写责任。

这种制度还可以降低末端替罪倾向。因为调查者不能只证明谁最后操作，还必须证明不同上游节点是否拥有可改变风险的能力，以及事故后这些能力如何被调用。责任由此从单点惩罚转为社会技术系统的可学习结构。

附录五：责任空心化的形式化表达与三个判定层级

为了避免“责任空心化”停留在修辞层面，可以把一项 AI 辅助任务的责任结构表示为多个角色集合 A={a1,a2,...,an}。对每个角色分别赋予四类变量：形式责任 F、认知可达性 E、因果控制 C 和改写权限 R。传统组织往往重点记录 F，例如岗位说明书、签字记录、审批日志；TRC-RVL 则要求比较 F 与 ECR 之间的关系。

对某个角色 ai 而言，可以定义责任实质比 S_i = min(E_i, C_i, R_i) / F_i。当 F 很高，而 E、C、R 中任一项很低时，S_i 会显著下降，形成“高形式责任—低实质责任能力”。这里采用最小值而不是平均值，是因为三项条件具有瓶颈效应：一个人即使非常了解事故并拥有很强控制，只要没有修改权限，仍然无法完成下一轮改写；反之亦然。

进一步可以把组织整体的改写完备度定义为：对于每一种重大风险 k，是否存在一个单一主体或一个制度化责任组合 G_k，使 E、C、R 能够在明确时间窗口内汇聚。如果存在，组织即使高度分工，也可以拥有高改写完备度；如果不存在，则意味着事故发生后只能临时协调，责任依赖个人关系和偶然推动。

因此，责任空心化至少有三个层级。一级为空心岗位：某个个人承担高 F 但低 ECR；二级为空心流程：单个岗位或许合理，但跨流程没有 ECR 汇聚节点；三级为空心治理：组织知道这些缺口存在，却长期没有制度权力去要求供应商、管理层或跨部门修改。三级最危险，因为它意味着组织本身已经缺少改写自己的能力。

附录六：为什么“更多透明度”不能单独解决责任问题

AI 治理经常把透明度视为问责前提，这一方向是正确的，但 ECR 模型显示，透明度主要增加 E，并不自动产生 C 与 R。一个员工可以获得完整模型解释，却仍然不能改变任何决策；一个审计机构可以发现风险，却没有停用系统的权限；一个技术团队可以公开局限，却无法迫使业务部门改变使用方式。

因此，“看得见”只是责任的第一步。若制度把透明度当作最终解决方案，就可能产生新的责任转移：系统已经解释了，所以用户应该为接受结果负责；供应商已经披露风险，所以部署者承担全部后果；组织已经培训员工，所以错误就是员工个人问题。透明度反而可能成为责任下移的依据。

真正的可问责透明度必须与行动接口绑定：知道某个风险以后，谁能暂停？谁能要求补充证据？谁能改变默认设置？谁能把问题从个案升级为系统问题？如果这些问题没有答案，透明度只是“让更多人知道自己无能为力”。

附录七：AI 责任的三种时间尺度——事前、事中与事后不能由同一个“最终负责”概念概括

责任在时间上至少分为三个阶段。事前责任关注系统是否被合理设计、采购和部署，包括风险识别、测试、人员能力和使用边界；事中责任关注具体运行时谁监测、解释、干预和停止；事后责任关注损害发生后谁解释、赔偿、修复并改变系统。把三种责任全部压缩为“最终决定者负责”，会掩盖不同主体在不同阶段的真实作用。

例如供应商可能在事前具有最高技术控制，部署组织在事前和事中拥有最高情境控制，一线工作人员在事中拥有个案信息，而管理层和跨部门治理团队在事后拥有资源与改写权限。成熟责任体系应允许责任随时间阶段移动，而不是寻找一个永远承担所有责任的主体。

这一时间分层还能解释为什么末端签字制度容易出问题：它把事中最后动作过度放大，遮蔽了事前架构与事后改写。事故调查因此容易问“当时为什么没拒绝 AI”，却较少问“为什么组织设计成需要一个人在数秒内识别上游数月形成的系统问题”，也较少问“为什么事故后仍没有权限修改根因”。

附录八：对“最终人类负责”原则的重新表述

“最终人类负责”作为伦理口号具有重要价值，因为现阶段 AI 本身通常不被作为拥有完整法律人格和道德责任的主体。但如果这个原则被理解成“必须找一个自然人在最后签字”，就可能产生适得其反的效果。更严谨的表述应是：对于高影响 AI 系统，必须存在可定位的人类或组织责任共同体，对关键风险拥有足够的认知可达性、因果控制与改写权限，并能对系统整个生命周期承担说明、干预和修正责任。

这一表述把“human accountability”从一个末端人转为制度化人类治理。它允许不同阶段由不同专业主体承担责任，也允许复杂团队共同形成 ECR，但拒绝用一个形式签字者替代真正的责任架构。

因此，“人类负责”最重要的不是把机器错误道德化地塞回某个个人，而是保证社会技术系统始终存在人类可进入、可理解、可干预、可改变的责任接口。只要这些接口仍然真实存在，AI 可以高度自动化；如果这些接口被抽空，再多人工签字也只是形式上的责任残影。

附录九：对组织实践的最终检验——四个问题足以暴露“假责任”

在真实组织中，不一定需要先理解 TRC-RVL 的全部理论，才能发现责任空心化。对任何一个“最终由人负责”的 AI 流程，都可以连续提出四个问题。第一，这个人到底能看到什么？如果他的判断材料已经被 AI 或上游流程高度过滤，而原始证据、模型限制与不确定性不可见，那么形式审核首先缺失 E。第二，这个人到底能改变什么？如果他只能在给定选项中确认，不能改变阈值、补充信息、延缓执行或启动替代路径，那么 C 十分有限。

第三，这个人发现系统性问题以后，能把什么真正改掉？如果答案只是“提交反馈”“写备注”“向主管反映”，而没有明确谁必须接手、何时回应、谁有权限修改模型和流程，那么 R 仍然是空的。第四，如果同类问题下一次再出现，组织能否指出上一次事故以后到底改了什么？如果只能回答“已经培训”“已经提醒”“已经处分”，却不能指出生成条件发生变化，就说明责任仍停留在个人层。

这四问的价值在于，它们把责任从抽象道德语言转化为可操作的组织事实。一个组织可以反复声称“人类始终保有最终控制”，但只要不能回答看见什么、能改什么、如何升级、下一轮哪里不同，这种控制就可能只是形式控制。反之，即使系统自动化程度很高，只要 ECR 接口清楚、事故能够稳定进入改写闭环，责任结构仍然可以是坚实的。

最终判准：责任的成熟度看“复发条件是否被改变”

对复杂 AI 系统而言，最有区分力的责任判据可能不是“事故后有没有人被处罚”，而是“事故后同类风险的生成条件有没有被改变”。处罚可以满足社会的归责需求，却不一定改变系统；赔偿可以修复部分损失，却不一定提升下一轮安全；增加人工签字可以提高审计可见性，却可能进一步扩大签字—控制错配。

因此，本章最终把责任成熟度定义为一种组织学习能力：风险信号能够穿过角色边界，事故知识能够被汇聚，拥有控制的人必须响应，拥有修改权的人必须作出可审计决定，修改结果再回到实际使用情境中被验证。只有当这一闭环存在，责任才从“谁为过去买单”推进为“谁确保未来发生变化”。这也正是 AI 时代责任制度最需要新增的一层。

附录十：研究限制与后续方向

本章提出的 ECR 与 TRC-RVL 仍属于理论模型，尚不能替代具体法域中的法律责任判断。不同国家、行业与专业关系对注意义务、因果关系、产品责任、雇主责任和专业责任有不同规则，本章所提出的“改写责任”主要是一种跨制度分析概念，而非直接的法律责任构成要件。未来研究需要在具体场景中验证 ECR 指标与事故复发、监督质量、责任感以及组织学习速度之间的关系。

另一个重要方向是研究责任结构的动态迁移。随着 AI 从建议工具走向代理系统，同一组织中 E、C、R 的分布会持续变化：系统可以自动收集更多信息、自动执行更多动作，供应商可以通过远程更新改变模型，部署组织也可能通过配置层拥有新的控制。责任地图因此不能一次制定后永久不变，而应随着模型版本、自动化水平、使用情境和事故经验周期性重画。

本章附表

维度	典型问题	AI 组织中的常见持有者	风险
因果贡献	谁实际塑造了结果?	数据、模型、流程、管理与操作多节点	贡献分散，难以单因果归责
信息可见	谁知道为什么会这样?	各角色只掌握局部信息	无人拥有全链条解释
控制能力	谁能在事前/事中改变?	上游设计者、组织管理者、部分操作员	控制与签字分离
形式归责	事故后最容易找谁?	末端确认/签字者	形成责任压缩与替罪风险

TRC-RVL 阶段	主要机制	表面现象	隐藏后果
1 功能分解	专业化与自动化	效率提高	完整判断被拆散
2 信息/控制分散	局部权限	岗位清晰	无人拥有全链条
3 建议预成形	AI 生成完整方案	人仍可确认	真实选择空间缩小
4 末端责任化	签字/审批日志	责任人明确	上游控制被遮蔽
5 作者感下降	责任扩散/代理感弱化	“只是审核”	监督动机下降
6 成功强化	无事故与高绩效	制度看似有效	角色进一步固化
7 ECR 分离	知识、控制、权限分散	各司其职	完整改写责任空位
8 事故压缩	追踪最后动作	迅速找到责任人	形成末端替罪
9 个案结案	处分/赔偿/补丁	问题“解决”	生成机制未改
10 再循环	恢复原流程	继续运行	同类风险积累

类型	形式责任	ECR 持有	心理作者感	主要风险
A 末端替罪型	末端强	末端弱、上游分散	末端低	责任压缩、系统不改
B 多手漂移型	多节点模糊	E/C/R 分别散落	各方低	事故后互相转移
C 集中控制型	管理/技术强	少数节点高度集中	较高	可能忽视一线情境
D 改写闭环型	分层明确	ECR 跨节点可汇聚	与真实控制匹配	更利于学习与纠错

第八章

不重做，也可以充分

最小充分验证路径　——审计学 × 可靠性工程 × 实验科学方法论的跨学科对撞

审计学 × 可靠性工程 × 实验科学方法论的跨学科对撞

——在不重做整个任务的前提下，建立足以发现“会改变结论的错误”的人机协作验证框架

摘 要

生成式人工智能正在把“生产答案”的成本压缩到前所未有的水平，却同时制造了一个新的认知瓶颈：当长篇文章、代码、数据分析、法律意见、研究综述和决策建议可以在数秒至数分钟内生成，人类是否必须把原任务重新做一遍，才能确认 AI 没有犯错？若完整重做，效率增益在验证环节被抵消；若仅凭语言流畅、引用数量、模型自信或少量随意抽查接受结果，则关键错误可能恰好隐藏在未检查的部分。本章采用跨学科互否方法，不把审计学、可靠性工程和实验科学方法论简单并列，而是寻找三者在“有限资源下形成足够可信的结论”这一共同难题上的结构同构。审计学提供重要性、风险导向抽样与合理保证，可靠性工程提供故障模式、严重度—发生可能性—可探测性分析、纵深防御与独立验证，实验科学方法论提供可证伪性、判别性检验、复现与独立第二路径。三者碰撞后，本章提出“最小充分验证路径”（Minimum Sufficient Verification Path, MSVP）：先把 AI 输出从连续文本还原为由结论节点、承重主张、证据节点、推导节点与约束节点构成的验证图；再以错误后果、传播范围、不可逆性、可探测性和证据独立性确定验证强度；随后对高承重节点进行逐项核验，对同质低风险节点实施风险导向抽样，对关键推理设计能够使结论失败的判别性检验，对高后果任务启用独立第二路径；最终以“残余的未检错误已不足以改变核心结论或突破可接受损失阈值”作为停止规则，而非以“全部内容均被重新检查”作为停止规则。本章进一步提出承重主张、结论翻转集、验证债务、证据独立度与风险压缩率等概念，并给出研究综述、代码生成、数据分析、法律意见与决策建议五类任务的分层验证协议。本章的核心主张是：AI 时代人的把关责任不应被理解为对机器劳动的镜像重演，而应转化为对结论脆弱点、失效模式和证据独立性的结构化治理。真正高效的验证，不是少做一次原任务，而是用不同于原生成路径的、更便宜但更具判别力的证据，优先攻击那些一旦错误就会改变行动的关键节点。

关键词：生成式人工智能；人机协作；验证；审计抽样；可靠性工程；故障模式；证伪；独立复现；合理保证；最小充分验证路径

Abstract

Generative AI has dramatically reduced the cost of producing long-form text, code, analyses and recommendations, but it has not eliminated the cost of verification. This paper develops a Minimum Sufficient Verification Path (MSVP) by second-order collision among auditing, reliability engineering and the methodology of experimental science. Rather than re-performing the whole task, MSVP reconstructs an AI output as a dependency graph of conclusions, load-bearing claims, evidence, derivations and constraints; allocates verification intensity according to consequence, propagation, irreversibility, detectability and evidence independence; combines risk-based sampling, claim-wise verification, discriminating tests, local recomputation and independent second paths; and stops when residual undetected error can no longer reasonably overturn the action conclusion or exceed a predefined loss tolerance. The framework reframes human accountability from mirroring AI labor to governing failure modes and evidence independence.

Keywords: generative AI; human-AI collaboration; audit sampling; reliability engineering; falsification; independent verification; reasonable assurance; MSVP

问题提出：AI 把“生产”变便宜之后，为什么“相信”反而变贵?

生成式 AI 首先改变的并不是知识本身，而是知识产品的边际生产成本。过去需要数小时乃至数日才能完成的研究综述、方案比较、程序原型、数据摘要和合同审阅，如今可以在极短时间内形成一个语言完整、结构清晰、形式上接近专业成品的初稿。生产速度的跃迁容易制造一种错觉：既然输出看起来已经接近“完成”，人的主要工作似乎只剩点击采纳。然而，生成速度与可靠性不是同一个变量。生成式模型的基本工作方式决定了它能够产生高度连贯但并不必然受事实、数据、代码执行环境或法律适用条件约束的文本。NIST 在生成式 AI 风险管理资料中将“confabulation”等风险列为需要治理的典型风险，近年来关于长文本事实性、语义不确定性、自检与外部检索的研究也表明，模型的流畅性不能直接等同于事实正确性[8][9][13-20]。

真正棘手的地方在于：验证本身也是劳动。如果用户要求 AI 写一份五十页研究综述，然后逐条查阅全部原文、重新检索全部文献、再自行重建论证，那么“验证”事实上退化为第二次完成原任务；如果 AI 写了一段两千行代码，而审查者必须逐行重写并重新推导所有边界条件，AI 的效率优势同样会被验证成本吞噬。于是，AI 时代出现了一个新的结构性矛盾：生产成本下降得越快，单位输出数量越大，人类越不可能沿

用“逐字逐句重新完成”的核验方式；但输出越长、传播越快，潜藏错误一旦被复制进入更多文档、程序、决策和组织流程，错误的外溢后果也越大。

因此，本研究的问题不是“人要不要验证 AI”，而是更精确的三个问题。第一，验证对象究竟是什么：是每一句话、每一个数字、每一行代码，还是支撑行动结论的关键节点？第二，验证资源怎样分配：哪些内容必须逐项核验，哪些内容可以抽样，哪些内容只需验证来源、边界和约束？第三，何时可以停止：我们怎样知道已经获得了足够的保证，而不再继续增加边际收益极低的检查？这三个问题共同指向一种新的验证逻辑：验证的目标不是证明“AI 输出整体绝对正确”，而是在可接受成本内，将“仍未发现但足以改变结论的错误”压缩到可接受水平。

这一区分十分关键。绝对正确要求穷尽性证明，现实中的多数复杂任务并不具备这种条件；合理可依赖则是一个风险判断。飞机不是因为工程师证明所有部件永不失效才起飞，财务报表也不是因为审计师重新做了企业全部会计工作才获得审计意见，实验结论更不是因为研究者证明所有可能反例均不存在才成立。成熟专业领域早已发展出一种更现实的知识制度：不追求无穷验证，而通过风险识别、重要性判断、抽样、反证、冗余、复现和独立路径，形成“足够强的保证”。生成式 AI 输出恰恰需要从这些领域借来这种制度能力。

本章将这种能力概括为“最小充分验证”：所谓“最小”，并非验证得越少越好，而是排除那些对最终保证没有实质贡献的重复劳动；所谓“充分”，并非所有细节均被直接检查，而是所有能够导致核心结论翻转、重大损失或不可逆后果的失效路径，都已经获得与其风险相称的证据覆盖。由此，验证不再与生成任务同构，而是成为一条专门设计的、具有差异化证据来源和判别能力的第二条认知路径。

研究方法：从“学科拼盘”到验证机制的跨学科对撞

本章采用“跨学科对撞”作为跨学科理论生成方法。第一层对撞通常表现为概念借用：例如把审计抽样用于 AI 抽查，把 FMEA 用于 AI 风险列表，把科学复现用于 AI 结果复核。这类借用有实践价值，但仍可能停留在“把三个工具放在同一张表里”的层面。第二层对撞进一步追问：三门学科为什么分别产生了这些工具？它们面对的原始困难是否具有共同结构？如果共同结构存在，那么能否从这种同构关系中推出一个不属于任何单一学科的新机制？

审计学的原始困难是“不可能检查全部交易，却要对整体报表形成合理保证”；可靠性工程的原始困难是“不可能阻止所有故障，却要保证关键系统在故障条件下仍能满足安全与任务要求”；实验科学方法论的原始困难是“不可能验证一个理论在所有世界

状态下都正确，却要通过可失败的检验、复现和竞争性解释不断提高结论可信度”。三者表面对象分别是账目、系统与科学主张，深层任务却高度一致：在有限资源与不可穷尽的不确定性下，把验证投入集中到最可能改变整体结论的地方。

由此可以提出本章的第一条同构命题：成熟验证制度不是“覆盖率最大化”，而是“结论相关风险压缩最大化”。同样检查一百个单位，如果这些单位均是低风险重复内容，覆盖率虽高，却可能没有触及真正决定结论的错误；反之，只检查十个节点，如果这些节点恰好覆盖关键假设、关键数据、关键接口和灾难性故障路径，其验证价值可能远高于前者。第二条同构命题是：验证证据的价值不仅取决于数量，还取决于与原生成过程的独立性。由同一模型、同一提示、同一检索源重复得到一致答案，并不等价于独立证据；它可能只是同一偏差的复制。第三条同构命题是：验证的停止规则应基于残余风险，而不是基于心理满意感或形式性完成。

本章据此将 AI 输出抽象为五类节点：结论节点（最终建议、判断、摘要性结论）；承重主张节点（结论成立所必需的事实或假设）；证据节点（来源、数据、法规、文献）；推导节点（计算、逻辑、代码执行与模型转换）；约束节点（用户要求、法律边界、业务规则、安全条件）。这种节点化并不是为了把自然语言机械拆碎，而是为了识别“如果这里错了，哪里会跟着错”。验证因而从文本层面的逐句校对转为结构层面的依赖关系审计。

方法上，本章首先分别提炼审计学、可靠性工程和实验科学方法论中的验证原则；其次进行结构映射，形成“重要性—关键性”“抽样—故障覆盖”“合理保证—残余风险”“独立复现—独立第二路径”等对应关系；再次据此构造 MSVP 模型及其形式化停止条件；最后把模型映射到五种常见 AI 任务，检验其可操作性。本章并不声称建立一种可以数学证明所有 AI 输出正确的通用算法，而是提出一种面向人机协作的验证治理框架：它的价值在于决定验证资源投向何处，以及什么证据足以支持“可以依赖”的行动判断。换言之，本章关注的不是让 AI 少犯所有错误，而是建立一种即使 AI 仍会犯错，关键错误也更难越过人类行动门槛的协作制度。这也是本章所说“效率与责任同时成立”的真正含义：不是减少责任，而是把责任集中到最有决定性的地方。

需要说明的是，本章所谓“最小充分”并不是追求统一固定比例的抽查率，也不是试图为所有行业规定同一个安全阈值。它是一套生成验证路径的元规则：先把任务的现实后果显式化，再根据结论依赖关系选择证据。不同领域可以在这一框架下形成自己的阈值、样本规则和强制复核清单。正因为它把阈值留给具体领域，框架才可能同时适用于日常写作、科研分析、软件工程与专业决策，而不把不同风险场景错误地压成一套万能流程。

第一重碰撞：审计学——验证不等于查完，而是围绕“重要错报”形成合理保证

审计学为 AI 验证提供的第一个关键启发，是把“正确性”从逐项无误改写为“整体结论不存在重要错误”。国际审计准则中的“合理保证”本身就承认审计具有固有限制：审计师通过获得充分、适当的审计证据，将审计风险降低到可接受的低水平，而不是对报表不存在任何错误作绝对保证[1]。这一思想对于 AI 输出极其重要，因为人机协作的现实目标通常也不是保证每个非关键措辞完美无误，而是保证不会因未发现的关键错误作出错误决定。

第二个启发是“重要性”（materiality）。审计中的重要性并非单纯看某个数字有多大，而是看该错报单独或汇总后是否可能影响预期使用者依据财务报表作出的经济决策[2]。迁移到 AI 验证中，一个错误是否“重要”，同样不能只看它占输出篇幅的比例。五十页报告中一个不起眼的日期错误可能无关紧要，也可能恰好使某项法规失效；两千行代码中一行错误可能仅影响日志格式，也可能绕过权限校验；一百条事实中只有一条错误，也可能因为它位于核心因果链上而使整份建议翻转。因此，AI 验证首先需要“重要性映射”，而不是均匀检查。

第三个启发是风险导向。ISA 315 等准则要求审计师识别和评估重大错报风险，并据此设计后续程序[4]。这意味着验证强度不是由文档长度机械决定，而由错误发生后对整体结论的影响和发生可能性共同决定。对 AI 而言，模型自报置信度可以作为辅助信息，却不应成为主导变量。一个低置信但无关紧要的背景句不值得大量核验；一个高置信但将触发高额资金转移、医疗处置、法律行动或安全控制的结论，则必须采用更高强度的外部证据。

第四个启发是抽样。审计抽样的价值不是“随便抽几条看看”，而是通过具有设计目的的样本，对总体形成受控制的推断[3]。这一区别可以直接纠正 AI 使用中的常见误区：人们经常在长文开头、中间、结尾随手各看几处，只要都通顺就认为整体可靠；这不是抽样，只是浏览。风险导向抽样至少要解决三个问题：总体是什么、样本为什么代表关键风险、抽中错误后如何扩大程序。对 AI 长文而言，总体可以按事实主张、引用、数字、推理链、章节或代码模块划分；样本应故意提高高承重、高新颖度、高不确定来源节点的入选概率，并保留一定比例的随机样本以发现未知模式。

第五个启发是“异常驱动扩展”。在审计中，一项异常并不只意味着修正这一项，它可能提示总体中存在系统性错报，从而需要扩大样本、改变程序或重新评估风险。AI 验证同样需要这一机制。如果抽查发现模型虚构了一篇文献，正确动作不是仅替换该文献，而是提高对该段乃至整份文稿引用体系的风险评级；如果代码的一个边界条件测试

失败，不能只修复这一输入，而应追查相同模式是否存在于同类函数；如果一个关键数据的单位错误，必须重新审视整个数据管道中的单位转换。错误不是孤立点，它还是关于“生成机制可能在哪里失稳”的新证据。

审计学由此给 AI 验证带来一个根本转向：验证的单位从“内容条目”变为“影响结论的错报风险”。一份一万字文档并不天然比一千字文档需要十倍验证；如果一万字中多数是可追溯的背景说明，而一千字包含三个决定性数字和两个不可逆决策，后者的验证强度应更高。验证资源应该按照重要性和风险配置，而不是按照字数、页数或模型输出时间配置。

第二重碰撞：可靠性工程——不要问“哪里可能错”，要问“怎样错会导致系统失效”

可靠性工程进一步把验证从“检查内容”推进到“管理失效模式”。工程系统无法假设所有部件永远正常，因此成熟工程方法会识别潜在故障模式、分析其后果与关键性，并根据严重度、发生可能性、可探测性等因素安排预防、检测和缓解措施。NASA 的 FMECA 相关指南强调用系统化方法识别失效模式及其影响，并据关键性用于风险评估[6]。这一思路迁移到 AI 输出后，问题不再是泛泛地问“AI 会不会出错”，而是列出具体的输出失效模式：事实虚构、引用错配、数据单位错误、计算链错误、法规版本过期、条件遗漏、反例未处理、代码接口误解、边界条件遗漏、因果关系倒置、结论超出证据范围、用户约束遗失等。

不同失效模式具有不同的“后果拓扑”。例如引用错配在普通公众号文章中可能只是信誉问题，在系统综述中却可能污染证据等级；一个日期错误在旅游攻略中可能造成行程不便，在法律时效判断中可能直接改变权利状态；代码中一个格式错误通常很快暴露，而一个权限检查的默认值错误可能长期潜伏且只在特定攻击路径出现。因此，仅以“错误概率”排序是不够的，必须同时考虑严重度、传播范围、不可逆性和可探测性。

本章据此提出“验证风险势”概念，用于表示某一节点值得投入多少验证资源。可以把它写成概念函数：$R_v = f(S, P, I, D, E)$。其中 S 表示错误后果严重度，P 表示错误向下游传播的范围，I 表示不可逆性或纠错成本，D 表示错误在常规使用中被发现的难度，E 表示现有证据的独立性不足。该函数并不要求所有场景使用同一套乘法评分，因为传统 RPN 的乘积方式本身存在尺度与排序争议；更重要的是，它迫使验证者显式回答“这项错误一旦漏过，会造成什么”“还能不能补救”“会复制到多少下游”“现有检查是否与原生成路径同源”。

可靠性工程的第二个关键启发是纵深防御。高可靠系统不会把安全寄托在一个检查点上，而会通过多层不同机制降低同一故障穿透全部屏障的概率[5][21]。对 AI 输出而言，纵深验证并不意味着重复问模型三次“你确定吗”。如果三个屏障都使用同一模型、同一上下文和同一资料源，它们可能共享失效模式。真正的纵深应当具有机制差异：例如第一层检查来源存在性，第二层用程序重算关键数字，第三层用对立假设设计反例，第四层由独立人或不同工具检查关键结论。层数本身不重要，失效模式的去相关才重要。

第三个启发是独立验证与确认（IV&V）。NASA 的软件 IV&V 实践强调由在管理、技术和财务等方面具有独立性的主体，对高风险软件进行独立分析与验证，以降低软件相关任务失败风险[7]。在人机协作中，“独立”可以有不同层级：同一模型换一个提示词是弱独立；不同模型但使用同一检索结果是中等独立；人工基于原始数据重算一个关键指标、调用独立静态分析器运行代码、由第二法律数据库核对条文、用不同方法推导同一结论，则是更强独立。对于高后果节点，验证设计应追求路径独立，而非仅追求回答次数。

第四个启发是“故障可检测性”。有些错误会快速暴露，例如代码语法错误；有些错误非常隐蔽，例如一个貌似合理但事实上不存在的文献、一个刚好落在常识区间内的错误统计量、一个法律条款中的适用例外被遗漏。越不容易在后续自然暴露的错误，越需要前置验证。由此可以解释为什么验证强度不能只由模型置信度决定：即使模型总体准确率很高，只要某类低频错误具有极高后果且难以在使用中发现，就仍应建立专门的检测屏障。

可靠性工程最终把 AI 验证转化为“失效治理”。我们不再要求人对 AI 的每个输出单位承担同样注意力，而是要求系统不能存在一条未经防护的灾难性失效路径。这个转向使“最小”与“安全”并不冲突：减少对低风险内容的重复检查，可以释放资源去构建关键节点的多层防护。验证效率的真正来源，不是少检查，而是把检查从平均分配改成按失效后果配置。

第三重碰撞：实验科学方法论——验证的价值取决于它能否让结论失败

实验科学给 AI 验证提供的最深一层启发，是把“支持性证据”与“判别性证据”区分开。一个结论可以轻易收集许多与之相容的材料，但相容并不意味着它经受了严格检验。科学方法的核心传统之一，是让理论面对有机会失败的测试：如果无论出现什么结果都能解释为“支持”，这个检验的证据价值就很低。Popper 的可证伪思想、随后关于严峻检验（severe testing）的统计方法论讨论，都强调证据价值来自测试排除错误解释的能力[11][12]。

这一点对 AI 自检尤其关键。让生成答案的同一模型阅读自己的答案并回答“是否存在错误”，常常只是在原路径上增加一次一致性检查。Huang 等关于大语言模型自我纠错的研究表明，在缺少外部反馈时，模型并不总能可靠改善推理，有时甚至会把正确答案改坏[13]。相反，外部工具、独立检索、可执行测试或被刻意设计为挑战原结论的问题，具有更强判别性。CRITIC、Chain-of-Verification、SAFE 等工作从不同角度显示，外部反馈、问题分解与搜索支持可以提高事实核验能力[19][16][17]，但它们也提醒我们：验证器本身仍可能出错，自动化核验不等于终局真理。

本章因此提出“判别性检验”原则：对每一个承重主张，不先问“有哪些证据支持它”，而先问“什么观察一旦出现，会迫使我们修改或放弃这个主张”。例如 AI 声称某项政策将提高转化率，判别性检查不是再找三个支持营销自动化的案例，而是识别结论依赖的关键条件——目标人群、价格弹性、渠道成本、基准期——并测试这些条件在合理变化下是否导致结论翻转。AI 声称某段代码在 O(n log n)内运行，验证不是让模型再解释复杂度，而是检查关键循环、数据结构操作并构造最坏输入；AI 声称某研究结论“得到多项研究一致支持”，验证不是数参考文献数量，而是抽取承重主张并核对研究设计、样本、效应方向和适用范围。

科学方法的第二个启发是区分“可重复”与“可复现/可复制”的不同层次。美国国家科学院报告将 reproducibility 定义为使用相同数据、计算步骤和方法获得一致结果，并把 replicability 理解为用不同数据对同一科学问题获得相符结果[9]。迁移到 AI 任务中，可以形成三个层级：同路径重跑——用同样输入和工具重现结果；异实现复算——使用不同代码、不同公式实现或不同分析软件处理同一数据；独立路径复核——从原始证据或不同数据源重新回答关键问题。验证强度越高，越应从“重复同一过程”升级到“独立路径获得相同结论”。

第三个启发是“局部复现而非整项复现”。科学实践中的独立复现非常昂贵，因此真正需要复制的往往是争议最大、影响最大或证据最脆弱的关键结果。AI 验证同样无需对整份输出做完全独立复制，而应寻找“结论翻转集”：只要其中任一节点错误，就足以使核心结论改变的最小节点集合。对这些节点实施更强的独立复核，往往比对所有低影响细节做浅层检查更有效。

实验科学由此把“验证强度”从检查量转变为反驳能力。一次好的验证应该让 AI 输出处于真正可能失败的情境；如果验证程序只会产生“通过”，它就不是验证，而是仪式。最小充分验证的关键不是找到足够多的赞成证据，而是对最重要的结论设置足够锋利的失败条件。

跨学科对撞的核心产物：从三种专业制度中抽取“结论相关风险压缩”机制

将三门学科放在同一层面，可以发现它们都拒绝一个朴素但成本极高的验证观：只有把全部对象重新检查一遍才算可靠。审计学通过重要性与抽样承认“不查全”；可靠性工程通过故障模式和纵深防御承认“不能消灭所有故障”；实验科学通过可证伪检验和复现承认“不能证明理论在所有情形中永远正确”。然而三者都没有因此降低责任，反而把责任变得更结构化：必须知道哪些错误最重要、怎样最可能发现它们、验证失败后如何升级、以及何时残余风险已足够低。

由此可以得到一个统一表达：验证是一种受预算约束的证据配置问题。设 AI 输出包含节点集合 N，每个节点 i 存在未知错误状态 e_i；错误若沿依赖关系传播，会对最终行动 A 造成损失 L_i。验证程序 v 对节点 i 具有探测概率 $q_i(v)$，成本为 $c(v)$，且不同验证程序之间可能高度相关。传统“全量重做”试图令所有 q_i 接近 1，但成本接近原任务甚至更高；随意抽查则把 q_i 近似均匀分配，可能把大量资源浪费在低 L_i 节点。MSVP 追求的是：在总验证预算 B 内，选择一组验证程序 V，使高损失节点的残余风险 $\Sigma P(e_i)\cdot[1-q_i(V)]\cdot L_i$ 尽可能低，并且所有灾难性失效路径至少被一个具有足够独立性的程序覆盖。

这个表达带来第一个新概念——“承重主张”（load-bearing claim）。承重主张不是最重要的句子，也不一定出现在摘要里，而是结论依赖图中的高中心性节点：一旦它失效，下游多个判断会同时失效。例如“样本 A 可代表总体 B”“法规 X 在目标日期仍有效”“变量 Y 的单位为万元而非元”“函数 Z 在并发条件下保持幂等”“市场容量数据口径包含存量而非新增量”。承重主张的识别，是把验证从语言表面转向逻辑结构的第一步。

第二个新概念是“结论翻转集”（conclusion-flip set, CFS）。它指一组最小条件或主张，只要其中某个成员发生足够幅度的变化，就会使最终结论、排序、建议或行动方案发生实质改变。CFS 不是统计学意义上的唯一最小割集，但可以借鉴可靠性工程的故障树思想来理解。对一份投资建议，贴现率、收入增长率、退出倍数可能构成翻转集；对法律意见，适用法域、合同条款版本、时效起算点可能构成翻转集；对研究综述，检索范围、纳入标准、关键研究的效应方向可能构成翻转集。验证资源优先进入 CFS，比平均检查全文更接近“充分”。

第三个新概念是“证据独立度”。如果生成与验证共享同一模型、同一语料、同一错误假设，表面上的多次一致仍然可能是共因失效。证据独立度衡量验证路径与原生成路径在数据源、推理机制、执行环境和责任主体上的解耦程度。独立度并不是二元变量，而是梯度：同模型自问自答最低；不同提示但同模型略高；不同模型与独立检索更高；程序执行、数据库查询、原始文献核对、人工专业判断、第二团队独立分析则在具

体任务中提供更强的结构独立性。高风险结论需要的是高独立度证据，而非更多低独立度重复。

第四个新概念是“验证债务”。当用户为了速度暂时跳过某些验证环节时，风险并不会消失，而会像技术债务一样进入下游。如果一份未经核验的 AI 综述被另一份报告引用，后者再被纳入决策备忘录，原始错误的发现成本与纠正成本都会提高。验证债务因此包括未核关键节点数量、传播层级和后续纠错成本。低风险、可逆的草稿可以允许短期验证债务；一旦输出即将进入公开发布、财务决策、合同签署、生产代码或安全操作，验证债务必须在“承诺点”之前清偿。

第五个新概念是“风险压缩率”。验证效率不应以“每小时检查多少字”衡量，而应以单位成本减少了多少结论相关残余风险衡量。检查一个引用是否真实存在可能只需几十秒，却能同时阻断事实虚构、来源伪造和下游引用传播；运行一个针对关键边界条件的单元测试可能只需几分钟，却能比阅读数百行代码更快发现致命逻辑错误。高质量验证因此往往呈现一种非对称结构：少量设计良好的测试，覆盖大量潜在损失。

表 1 三门学科在“有限资源形成充分保证”问题上的结构同构

“最小充分验证路径”MSVP：六阶段、五级强度与一个停止规则

基于上述碰撞，本章提出 MSVP 的六阶段流程：界定行动、还原验证图、评估风险势、配置验证组合、异常升级、残余风险停机。六阶段不是固定线性流程，而是一个可回环系统：任何异常都可能改变风险评估并迫使重新分配验证资源。

第一阶段：界定“这份输出将被拿来做什么”。验证强度必须从行动后果开始，而不是从文本长度开始。同一份 AI 答案，如果只用于个人头脑风暴，允许较高的不确定性；如果将作为公开事实陈述、对客户的专业意见、生产系统配置或不可逆决策依据，就必须提高保证等级。行动界定至少包括：使用者是谁、是否对外传播、是否触发资金/权利/安全变化、错误是否可逆、出错后多久能被发现、纠错窗口有多长。只有先定义承诺点，才知道需要何种保证。

第二阶段：把输出还原为验证图。做法不是把所有句子都拆成原子事实，而是先识别最终结论，再逆向追踪其必要前提。建议采用五种标记：C（Conclusion，结论）、L（Load-bearing，承重主张）、E（Evidence，证据）、D（Derivation，推导）、K（Constraint，约束）。例如一份 AI 生成的市场进入建议“应在第三季度进入 A 市场”是 C；“市场增速大于 20%”“监管许可周期小于三个月”“毛利可覆盖获客成本”是 L；行业数据库和法规文件是 E；财务模型是 D；预算上限和不得延迟主产品发布是 K。验证图只要覆盖能改变 C 的主干，而不必复制全章结构。

第三阶段：计算风险势并分级。本章建议至少观察五个维度：后果严重度 S、传播范围 P、不可逆性 I、可探测性 D、证据独立性缺口 E。实践中可用 1—5 级定性评分，也可以直接用红/橙/黄/绿分类。重要的是评分后要对应明确验证动作，而不是把评分本身当成结果。一个简单的规则是：只要严重度或不可逆性达到最高等级，即便发生概率不高，也不得仅用模型自检；只要某节点属于 CFS，就至少需要一个能够独立挑战它的验证程序；只要错误可能跨多个下游文档或系统传播，就要在源头提高验证强度。

第四阶段：配置“五级验证强度”。V0 为约束与来源完整性检查：检查任务要求、时间范围、法域、数据口径、引用是否存在、文件是否为正确版本。V1 为风险导向抽样：对大量同质、低至中风险事实或代码片段抽样核验，并保留随机样本以捕获未知错误。V2 为承重主张逐项核验：所有进入 CFS 或高中心性主张都逐项检查，不允许用总体抽样替代。V3 为判别性检验与局部复算：针对关键推理、数字、算法、法规适用条件设计可能推翻结论的测试，并对关键节点独立重算或运行。V4 为独立第二路径：对高后果、不可逆或极难检测的结论，要求不同数据源、不同工具、不同模型或不同人员从问题起点重新得到关键结果。

第五阶段：异常升级。任何抽样错误都应触发“这是否说明总体风险被低估”的判断。可采用四类升级规则：一是同类扩展——同类节点样本加倍或转为全检；二是上游追溯——检查错误是否源于共同数据、提示或约束遗漏；三是下游追踪——确认哪些结论已经继承该错误；四是等级升级——若错误属于系统性模式，将该区域从 V1 提升至 V2/V3。异常升级把抽样从静态节省成本的方法，变成动态风险探测器。

第六阶段：以残余风险作为停止规则。最小充分验证不以“已检查 80%内容”“已经看了两遍”“另一个模型也同意”作为完成标准，而以两个条件同时满足作为停机条件：第一，所有高后果失效路径已经被至少一种具有足够独立度的验证程序覆盖；第二，剩余未核内容即使存在错误，也不足以合理地改变核心结论，或其预期损失已低于任务预先设定的容忍阈值。换言之，停止不是因为验证者疲劳或因为输出显得可信，而是因为未检查部分的边际风险已经可接受。

为了让这一流程更可操作，本章提出“最小充分验证不等式”的概念表达：当 $R\_residual \leqslant T\_action$ 时，可以停止继续验证。其中 R_residual 不是模型置信度，而是经验证后仍可能存在的结论相关残余风险；T_action 是该行动可以容忍的风险阈值。随着行动从草稿、内部讨论、公开发布到不可逆执行，T_action 应逐级降低。这个表达的意义不在于要求普通用户精确计算概率，而在于强迫验证者把“我觉得差不多”改写为“即使剩余部分有错，它还能不能改变我要做的事”。

表 2　MSVP 五级验证强度

形式化判据：怎样证明“没有重做全部任务”仍然可以称为充分验证

要使“最小充分验证”成为可讨论的理论对象，必须进一步回答两个看似矛盾的问题：第一，既然没有检查全部内容，凭什么说验证“充分”；第二，既然可以继续增加验证，凭什么说当前路径已经接近“最小”。本章不把“充分”理解为绝对真理证明，而把它定义为一种面向行动的稳定性条件：在已经完成的验证之后，剩余未核错误的合理集合中，不再存在能够以不可接受概率使核心行动结论翻转的失效路径。这里的核心词不是“没有错误”，而是“没有仍然具有行动支配力的未控错误”。

命题一：结论不变性命题。若一个未核节点即使在合理误差范围内变化，也不会改变最终结论、方案排序或行动阈值，则该节点的直接验证优先级可以降低；若一个节点的轻微变化就能使结论翻转，则它必须进入承重主张集合。这个命题把敏感性分析引入文本和知识产品验证。例如一份成本收益分析中，某个背景市场规模数字即便误差 20% 也不影响“项目不可行”的结论，而一个看似很小的折现率变化会让净现值从正转负，那么真正需要高强度验证的是折现率来源与现金流时点，而不是把所有背景数据同等精确化。最小充分验证因此首先是一种“对结论梯度的识别”。

命题二：路径差异命题。在其他条件相近时，一个与原生成机制失效模式低相关的验证程序，比多个高度同源的重复程序具有更高边际验证价值。假设原答案来自模型 M、资料库 D 和提示结构 P。如果验证仍然由 M 在 D 与 P 上自我复述，即使重复五次，也可能共同遗漏同一个法条例外或同一个数据单位错误；相反，一次直接查询原始法规数据库或执行独立程序，虽然次数更少，却能切断共因失效。由此可以提出“独立性优先于重复数”的原则：高风险任务应优先购买机制差异，而不是购买更多同源答案。

命题三：边际风险压缩命题。验证程序的选择应比较其单位成本能够减少多少结论相关残余风险，而不是比较其覆盖多少文本。可概念化为效率 $\eta(v)=\Delta R(v)/c(v)$，其中 $\Delta R(v)$ 表示验证动作 v 带来的残余风险下降，$c(v)$ 表示时间、金钱或认知成本。现实中无法精确估计 ΔR，但仍可以做序位判断：直接重算决定结论的三个数字通常比逐字润色三十页报告具有更高 η；核对一个法规是否仍有效通常比让模型再次生成“法律依据”具有更高 η；运行一组针对最坏输入的测试通常比平均阅读所有函数具有更高 η。所谓“最小”，实质上就是不断剔除低 η 验证动作，直到继续削减会使某条重大失效路径失去覆盖。

命题四：抽样条件命题。抽样只有在“总体具有足够同质性且单个未检节点不会独立造成灾难性后果”时，才具有替代全检的正当性。如果一批 AI 生成的产品描述结构相似、每条错误只影响单个商品页面，可以通过设计抽样估计总体质量；但若一百条配

置中任何一条错误都可能删除生产数据库，抽样就不是适当策略。也就是说，是否能抽样并不取决于项目太长而人没时间，而取决于错误风险是否具有可汇总性。对于不可汇总的“单点灾难节点”，必须逐项验证。

命题五：异常信息增益命题。验证中发现的错误不仅是待修复对象，也是对错误分布的观测。若某类错误的出现显著提高了“其他同类节点也有错”的后验判断，就必须扩大程序。这个命题解释了为什么固定抽查 10%可能非常危险：抽样的目标不是完成一个预定比例，而是学习总体风险。如果前十个样本全部通过，后续检查可以保持；如果十个中出现三个同型错误，则总体风险模型已经改变，继续沿用原抽样率等于忽略新证据。MSVP 因此天然要求验证计划是自适应的。

命题六：承诺点清偿命题。验证债务可以在低风险、可逆阶段暂时存在，但在输出跨越“承诺点”时必须按照后果等级清偿。承诺点是内容从可随时撤销的内部草稿转变为对外发布、签字、付款、部署、提交、诊断、执行或其他难以撤回状态的瞬间。AI 时代很多事故并不是因为草稿中出现过错误，而是因为草稿状态的低验证标准被无意识地带入了承诺状态。因而一个成熟工作流应允许“快速生成—低成本试探”，同时在承诺点设置强制升级门槛。

上述六个命题共同给出“充分”与“最小”的双重定义。充分性是覆盖条件：所有重大结论翻转路径均有与风险匹配的验证证据；最小性是不可删减条件：从现有验证组合中删除任一程序，都会使至少一条不可接受的重大失效路径失去足够覆盖，或者使残余风险重新高于行动阈值。这个定义避免了把“最小”理解为验证次数最少。某些高风险任务的最小充分组合可能包含多项程序，因为少任何一项都会留下不可接受的盲区。

从这一点看，MSVP 更接近工程中的“最小割集覆盖”与审计中的“充分适当证据组合”，而不是一般意义上的效率技巧。它要求验证者能够说明每一项验证动作在防什么错误，以及每一条重大失效路径由哪一项证据覆盖。若一项检查无法回答“它在降低哪一种结论风险”，它可能只是仪式性动作；若一条重大失效路径无法回答“谁在检查它”，则当前验证尚不充分。

四种核验对象：逐项、抽样、约束核验与独立第二路径如何分配

第一类必须逐项复核的是“承重节点”。典型包括最终数字、关键日期、法域与版本、关键定义、因果方向、接口契约、权限与安全条件、模型输入输出口径、纳入排除标准、决策阈值，以及任何属于结论翻转集的假设。逐项复核不等于手工重做全文，而是要求每个承重节点都有可追溯证据，并至少通过一次针对其主要失效模式的检查。

第二类适合抽样的是“大量同质且单点后果有限”的节点。例如长篇综述中的背景事实、格式统一的数据记录、批量生成的测试用例、重复性较强的内容标签、非关键引用格式。抽样必须同时包含风险样本和随机样本。风险样本优先抽模型最可能犯错的地方：罕见事实、精确数字、最新信息、跨语言转述、涉及多个限定条件的句子；随机样本用于防止验证者只检查自己已经怀疑的区域，从而遗漏未知错误模式。

第三类主要做来源与约束核验的是“可由权威来源直接判定”的节点。对于法规原文、统计口径、API 参数、数据字典、合同版本、学术论文元数据等，与其让 AI 继续解释，不如直接核对权威源。这里的最小路径通常是：来源存在性→版本/日期→与主张是否真正对应→是否遗漏适用例外。许多 AI 错误并非来源完全不存在，而是“来源是真的，但并不支持该句结论”，因此存在性检查只能算 V0，不能替代承重主张核验。

第四类必须使用独立第二路径的是“高后果且共因风险高”的节点。典型包括大额财务决策、生产系统安全控制、法律权利义务的关键判断、医疗高风险建议、公共发布中的重大事实、科研结论中的核心效应、无法轻易回滚的数据迁移或删除操作。独立第二路径的定义是：验证过程不能只在原推理链上做局部修补，而要改变至少一个关键机制——数据源、分析方法、执行工具、模型家族或责任主体。只有这样，才真正降低共因错误穿透所有检查层的概率。

需要特别强调的是，验证强度不应依据“AI 是否自信”单独决定。语言模型的生成概率与人类关心的命题真实性并非同一概念；语义不确定性研究显示，不同生成之间的语义分歧可以帮助发现一部分“confabulation”，但仍不能覆盖所有事实错误。模型自信可以作为风险信号之一，却必须置于后果、可逆性和证据独立性之后。高后果任务即使模型高度一致，也应保持独立验证；低后果任务即使模型语气犹豫，也不必无限加码。

表 3 AI 输出不同组成部分的默认核验策略

五类典型任务的 MSVP 实施协议

10.1 研究综述：从“查所有文献”转向“验证证据骨架”。AI 生成综述最常见的验证误区，是要么相信整份参考文献列表，要么试图把所有论文从头读完。MSVP 建议先标出综述的三至七个核心结论，再为每个结论标记 2—5 个承重研究或证据类别。第一步做来源存在性与元数据核验；第二步核对摘要乃至原文中研究对象、设计、效应方向和限制是否真的支持 AI 的概括；第三步随机抽查非承重引用以估计引用体系的系统性错误率；第四步对争议最大或最可能改变综述方向的结论寻找反例研究、系统综述或

独立数据库。如果抽查发现伪造文献或多次“引用真实、结论错配”，则把该章节从抽样升级为大范围核验。这样，人不必重读全部文献，却会掌握决定综述结论的证据骨架。

10.2 代码：从逐行阅读转向失效模式驱动测试。对 AI 生成代码，语法正确只是最低层级。先界定高后果失效：数据丢失、权限绕过、注入、竞态、资源泄漏、错误异常处理、数值溢出、不可逆外部调用。然后识别对应承重模块与接口，运行静态检查、类型检查、单元测试和边界测试；对高风险函数使用性质测试、模糊测试或不同实现的交叉验证；对关键计算独立重写一个简短的 oracle 或使用可信库比对结果。阅读代码仍有价值，但它应围绕故障假设展开，而不是平均逐行审美。若代码将进入生产环境，V4 应包括代码审查者或独立测试环境，而不能让原生成模型既写代码又独占验收。

10.3 数据分析：从“重新做整份分析”转向关键统计量与数据链路复算。先核验输入数据版本、行列数、缺失值处理、单位、过滤条件、分组逻辑和目标变量定义；这些约束节点通常比漂亮图表更承重。随后对决定结论的核心统计量独立重算，例如样本量、均值/比例、关键回归系数、置信区间、排序和阈值；再设计对结论敏感的扰动检验，如更换合理阈值、排除极端值、改变基准组，看结论是否翻转。对于复杂模型，不必完整重写全部代码，但至少要对输入—关键中间量—最终指标建立可追溯链。任何“结果看起来合理”都不能替代数据口径核对，因为最危险的数据错误往往仍会生成视觉上漂亮的图。

10.4 法律意见：从“文书流畅”转向法源、适用条件与例外的独立核对。AI 可以帮助整理争点、生成检索式和比较论证，但法律结论具有高度时效性、法域性和条件性。MSVP 要求先锁定法域、时间点、事实前提和问题类型，再把最终意见拆成若干承重法律命题；每个命题直接核对权威法源及最新有效版本，检查条文、判例或监管文件是否真的覆盖该事实，并主动搜索例外、冲突规范与不利先例。对将触发重大权利义务的结论，应由合格专业人员或独立权威数据库形成第二路径。这里尤其不能把“引用很多”当成可靠，因为错误可能集中在适用性而非存在性。

10.5 决策建议：从“比较很多利弊”转向翻转条件测试。AI 最擅长生成看似全面的方案比较，但决策质量取决于事实、概率与价值排序三者。验证时先问：推荐 A 而不是 B 依赖哪三个到五个关键假设？然后针对每个假设设置合理变动区间，进行敏感性分析；识别哪些变量一变化就会改变排序，这些变量即构成 CFS。对可逆的小决策，可以通过小规模试点获得真实反馈，替代更多文字分析；对不可逆的大决策，则应增加独立预测、外部基准和反方论证。这样，AI 负责扩大方案空间，人负责确保选择没有建立在一个未经验证的脆弱前提上。

表 4 五类常见 AI 任务的“最小充分”验证重点

为什么“让另一个 AI 检查”有时有效、有时只是双重幻觉

多模型验证正在成为现实工作流中的常见做法，但其有效性取决于错误相关性。SelfCheckGPT 利用同一模型多次采样的一致性来识别部分幻觉，语义熵方法也通过多次生成之间的语义不确定性检测一类 confabulation[14]。这说明重复生成可以提供风险信号：如果同一问题在不同采样中出现互相冲突的答案，确实值得提高验证强度。然而，“一致”只能降低某些随机性错误的嫌疑，不能排除共享知识缺口、系统性偏见、过时信息或同源检索错误。

因此，本章把 AI—AI 验证分成三种情形。第一，相关性高的自检：同一模型、同一上下文、同一资料源，主要适合作为 V1/V2 前的筛查器，不足以承担高风险终审。第二，机制部分独立的交叉模型：不同模型家族、不同检索后端或不同提示结构，可以提高错误暴露率，但仍要警惕训练语料和互联网来源的共性。第三，工具增强的外部验证：让模型调用数据库、执行代码、访问原始文献或结构化规则，并要求验证问题与原生成问题分离，独立性更强。Chain-of-Verification 之所以值得借鉴，不只是“多问几步”，而是其验证问题被单独回答，试图降低原答案对验证过程的污染[16]。

一个实用判据是：第二个 AI 提供的是“另一份观点”，还是“另一份证据”？如果它只是用不同措辞重述相同结论，证据增量很小；如果它能够指出一个原模型未使用的权威来源、运行一段可复现代码、构造一个能使原结论失败的反例、或者在不同数据上得到一致结果，那么它才真正进入验证链。AI 可以成为验证工具，但“AI 检查 AI”不能天然等于独立验证。

人的最终把关责任：从“亲自做完”转向“设计并承担验证制度”

AI 时代容易出现一种责任错位：要么认为只要最终由人点击确认，判断就天然属于人；要么认为只要人没有亲手重做整个任务，就不能真正负责。本章提出第三种理解：人的责任核心是验证制度的所有权。也就是说，人需要决定任务的风险等级、识别承重主张、选择独立证据、设置停止条件，并对“为什么这些检查足以支持行动”给出可追溯解释。是否逐字亲自生成，不是责任归属的充分条件。

这意味着“会用 AI”不能只被理解为会写提示词，而应包括验证架构能力。一个成熟的人机协作者至少要能回答五个问题：我准备拿这份输出做什么？哪些条件一变结论就会变？最危险的错误怎样发生？我现在的核验是否真正独立于原生成路径？如果抽

查发现一个错误，我会怎样扩大检查？这些问题构成比“AI 说得像不像专家”更稳定的判断基础。

组织层面也应改变验收标准。与其要求员工在 AI 参与后声明“本人已全面审核”，不如建立分级验证记录：任务风险级别、承重节点清单、已使用的验证程序、发现的异常、升级情况、未消除的残余风险及最终责任人。对于高风险任务，还应保留关键证据和复算结果。这样的制度不要求每个人都假装自己重新做了一遍 AI 的全部工作，却能把责任从模糊的“看过”转变为可审计的验证行为。

更进一步，AI 本身可以帮助生成“验证计划”，但验证计划也需要由人设定风险边界。最理想的人机分工不是 AI 生产、人类逐字纠错，而是 AI 承担大规模搜索、草拟、枚举和低成本测试，人类负责定义行动后果、识别价值权衡与不可逆风险，并在关键节点强制引入独立证据。这样，人的不可外包部分不是手工重复机器已经做过的全部认知劳动，而是决定“什么必须是真的，我们才有资格行动”。

理论创新：最小充分验证与传统“事实核查”框架有什么不同

第一，本章把验证对象从“事实句”扩展到“结论依赖结构”。传统事实核查主要处理命题真假，但 AI 输出还会因数据口径、约束遗漏、推理变换、代码执行和价值排序而出错。一个报告的每个事实都可能是真的，结论仍可能因为选择性证据或错误推导而不成立。MSVP 因此要求同时核验证据、推导与约束节点。

第二，本章把验证强度的决定变量从模型属性转向行动风险。模型大小、置信度、是否联网、是否使用 RAG 都能影响错误概率，但都不足以决定需要多少验证。真正决定验证等级的是“错误穿透后会造成什么”。这使同一个模型在写创意标题时可以低验证，在生成生产数据库删除脚本时必须高验证。

第三，本章把“独立性”作为验证质量的核心变量。大量 AI 工作流通过多代理讨论、反思、自洽性来提高结果质量，这些方法有价值，但如果代理共享同一错误源，验证仍可能发生共因失效。MSVP 要求高风险结论至少有一条机制上独立的证据路径。

第四，本章提出“验证债务”来描述延迟核验的跨阶段成本。它解释了为什么内部草稿阶段可以允许未核内容，而公开发布或不可逆执行前必须清偿关键债务；也解释了为什么组织需要追踪 AI 内容的来源和验证状态，否则错误会在复制中失去原始上下文。

第五，本章提出“结论翻转集”作为最小验证范围的寻找工具。它比简单的“重要句子”更严格：只有当某个节点的错误足以改变最终行动，它才应该进入最高优先级。验证者的第一任务因此不是阅读更多，而是找到那些真正控制结论的少数变量。

与本书其余各章相比，本章的对象是唯一一个"做到哪里为止"的问题，其余各章问的都是"有没有"。这一点值得说明，因为它决定了本章的读数与全书其余读数不同型。

与第九章的关系最紧。第九章的"共识独立性折扣"与本章的"证据独立度"是同一个量在两个尺度上的使用：本章用它决定一次验证动作值不值得做，第九章用它决定一群人的同意该打几折。两章共用一条原则——同源的重复不能被当成独立的重复——但用途相反：本章用它买路径差异，第九章用它防止把复制读成共识。若将来能用同一个形式量把两者统一起来，两章的独立性将显著削弱，本书应当合并它们。

与第一章的分界则简单而必要：本章验证的是输出，第一章判定的是人。一份通过了最小充分验证的报告，不能证明提交它的人具备相应能力；一个通过了断代理测试的人，也不保证他手上这份报告没有会翻转结论的错误。产品的可信与人的能力是两笔账，本书自始至终不把它们记在一起。

边界、失败情形与不可被“最小化”的验证

“最小充分”最容易被误解为削减验证。事实上，它对高风险任务可能比当前常见做法要求更多验证。尤其有四类场景不适合通过大幅抽样来降低成本。第一，错误后果可能涉及生命安全、重大法律权利、关键基础设施或不可逆财务损失；第二，输出不存在稳定的同质总体，无法用少量样本代表其余部分；第三，错误高度相关，例如整份分析共享同一个错误数据源；第四，验证者缺乏识别承重节点所需的领域知识。此时“最小”可能仍然意味着全面专业复核或正式独立验证。

另一个边界是未知未知。风险导向方法倾向于验证已知失效模式，但 AI 可能产生验证者未预见的新型错误。因此 MSVP 保留随机抽样、反方检验和跨路径复核，目的就是为未知错误留下发现通道。若验证设计全部由原生成模型自动提出，也可能出现“模型不知道自己不知道什么”的闭环盲区。对高风险场景，验证计划至少需要一个外部视角。

第三个边界是验证器错误。搜索引擎可能返回错误网页，数据库可能存在滞后，测试 oracle 可能写错，第二个模型也可能幻觉，专家也可能判断失误。MSVP 并不假设验证层绝对可靠，而是依赖证据多样性和风险分层来降低同一错误贯穿全链路的概率。因此，越高风险的任务越不能依赖单一验证器。

第四个边界是价值判断。事实验证可以提高“世界是什么”的可靠性，却不能替代“我们愿意承担什么风险”“哪一种价值更优先”的规范选择。AI 可以计算方案在不同假设下的后果，但验证不能把价值冲突伪装成技术误差。对决策建议，人类最终责任仍包括明确价值权重，并承认某些分歧不是更多事实就能自动消解。

研究议程：如何把 MSVP 从理论框架变成可测量的验证科学

未来研究首先需要建立“结论翻转集”识别方法。可以探索由模型生成依赖图、由人工校正关键节点，再通过敏感性分析或反事实提问判断哪些节点具有真正翻转能力。不同领域的翻转集可能具有不同结构：代码偏向接口与边界条件，法律意见偏向法域与例外，研究综述偏向证据等级与纳入标准，决策建议偏向关键参数和价值权重。

第二，需要测量不同验证动作的“风险压缩率”。例如在长文本事实核验中，比较随机抽样、风险抽样、原子事实分解、搜索增强验证和独立专家复核，在相同时间预算下能够发现多少“会改变结论的错误”，而不是仅比较总错误发现数量。一个方法可能发现很多无关紧要的小错，却漏掉唯一的关键错误；传统准确率指标会高估其实际价值。

第三，需要研究证据独立度与共因失效。多代理系统越来越普遍，但不同模型之间究竟共享多少知识盲点、检索偏差和推理模板，需要实证量化。可以把“不同模型一致”与“不同机制一致”区分开，并测量它们对高后果错误的真实检测增益。

第四，需要建立领域化阈值。合理保证并不是跨任务统一的数字。医疗、法律、金融、科研、教育和日常写作对错误的容忍度不同。未来可以借鉴安全完整性等级、审计重要性和质量保证制度，为不同类型的 AI 辅助任务建立明确的最低验证动作，而不是只提供抽象的“请人工审核”提示。

第五，需要把验证记录嵌入 AI 产品界面。理想系统不应只生成答案，还应允许用户看到结论依赖图、来源状态、未核节点、风险等级、已执行的验证测试和残余风险。这样，“可信 AI”不再被理解为模型自身永不犯错，而是包含一套可观察、可升级、可追责的验证闭环。

验证强度的真正决定变量：为什么后果、传播与可逆性优先于长度和模型置信度

在实际使用中，人们很容易用三个直观指标决定是否认真验证：输出很长所以多查、模型很强所以少查、模型语气很自信所以相信。这三个指标都可能提供辅助信息，却不应成为验证强度的主轴。首先，任务长度只代表潜在检查对象数量，不代表单位错误的后果。一份三万字的创意策划即使出现十个小事实错误，也可能只需发布前局部修订；一条只有二十个字的数据库删除指令，若作用对象写错，则可能造成不可逆损失。验证制度若按字数配置资源，会系统性低估短而危险的任务。

其次，模型能力是总体统计属性，而行动需要面对具体实例。一个模型在某项基准上表现优异，不意味着当前这一条输出已经通过了相关约束，也不意味着它掌握最新事

实、正确理解了本地规则或没有在长上下文中遗失条件。更强模型可以降低某些错误的先验概率，因此可能影响验证预算，但它不能取消后果驱动的最低验证要求。就像更可靠的部件可以降低系统故障率，却不会让航空系统取消对灾难性失效的安全设计。

再次，模型置信度尤其容易被误用。自然语言模型可以用非常确定的语气表达错误，也可以在正确答案上表现犹豫；即使系统能够给出内部概率或不确定性指标，这些量也只覆盖特定形式的不确定性，而未必捕捉过时知识、系统性偏见、提示误解或外部数据错误。语义熵等方法之所以有价值，是因为它们把“不同生成是否在意义上分歧”转化为一种风险信号，但研究本身也只声称检测一类 confabulation，而非提供万能真实性证明。因此，置信度适合作为筛查变量，不适合作为终审门槛。

真正更稳定的三个变量是后果、传播和可逆性。后果回答“错了有多严重”；传播回答“错误会被复制到多少下游对象”；可逆性回答“发现以后还能否低成本恢复”。三者共同决定一个错误是否值得在源头付出更高验证成本。例如一个内部备忘录中的错误如果只影响一次可撤销讨论，后果和传播都低；同一句话若被自动写入对外知识库，并被多个客服机器人继续引用，传播性立刻提高，验证等级也应上升。一个财务模型中的假设若可在试点后调整，可逆性较高；若它决定一次不可撤销的并购付款，可逆性极低，即使错误概率看起来不高也应启用独立第二路径。

这也解释了为什么“传播范围”应成为 AI 时代新增的重要变量。传统个人写作中的错误通常停留在单个文件，而生成式 AI 天然支持复制、改写、自动发布和多代理调用。一个未经核验的数字可能先进入研究报告，再被摘要模型压缩进 PPT，又被销售助手写进客户邮件，最终被另一个模型当作上下文事实继续生成。到错误被发现时，已经很难追踪来源。MSVP 要求在高传播节点上提高源头验证强度，本质上是在错误进入复制网络之前降低其基本再生数。

因此，验证强度的排序原则可以简化为：先看行动后果，再看错误传播，再看可逆性与可探测性，最后才把任务长度、模型能力和置信度作为修正项。这个顺序能够避免一种常见错觉——因为 AI 生成得快、写得长、说得像，所以验证也应该以文本表面为中心。恰恰相反，AI 越擅长制造完整表面，人类越需要把注意力从表面移向“哪些节点正在控制现实行动”。

从个人技巧到组织制度：MSVP 如何嵌入真实的人机协作流程

如果 MSVP 只停留在个人“使用 AI 时多留个心眼”，它仍然不足以成为稳定制度。现实组织中的 AI 输出往往经过多人、多工具、多版本流转：研究员用 AI 生成初稿，分析师补数据，经理修改结论，法务加免责声明，最终又被复制进 PPT、邮件、数

据库或自动化流程。错误也会沿着这条链传播。因而组织首先需要建立“验证状态”而非笼统的“AI 生成/非 AI 生成”标签。一个更有用的标记至少包括：未核草稿、来源已核、承重主张已核、关键推导已复算、独立路径已完成、可对外承诺。这样，使用者看到的不只是内容，还能看到它已经获得何种程度的保证。

第二，组织需要把“验证责任”绑定到行动节点，而不是绑定到文本作者。AI 时代一份文档可能没有传统意义上的唯一作者：模型生成大部分文字，多名员工补充材料，系统自动更新数据。如果仍要求“作者对全文负责”，责任很容易变成形式签字。MSVP 更适合采用“承诺责任人”制度：谁决定把输出推进到发布、付款、部署、签署、提交等承诺点，谁就必须确认该行动对应的最低验证层级已经满足。这样，责任从“谁敲了这些字”转移到“谁决定让这些字开始产生现实后果”。

第三，验证记录应采用“证据包”而不是“审阅痕迹堆积”。传统协作中常见大量批注、版本、聊天记录，但关键时刻仍很难回答“为什么当时认为这个结论可以相信”。MSVP 证据包可以非常精简：一页结论依赖图、一张 CFS 清单、关键来源链接或文件版本、三到五项判别性测试结果、异常与升级记录、独立第二路径的结论差异。它的目的不是制造合规文书，而是让后来的使用者能够重建当时的信赖理由。对于低风险任务，证据包可以只有几行；对高风险任务，则需要更正式的留痕。

第四，组织应区分“生成预算”和“验证预算”。当前许多 AI 项目只计算模型调用成本与人工节省时间，却没有把核验作为独立资源项，于是团队要么在最后阶段突然发现没有足够人力验证，要么在时间压力下把验证降格为快速浏览。更合理的做法是在任务开始时同时分配两种预算，并让验证预算随风险等级上升。低风险创意任务可以把绝大部分预算用于生成；高风险专业任务则可能把一半甚至更多资源留给数据核对、测试、专业复核和独立路径。AI 越能快速生产大量内容，验证预算越不能简单按“生成耗时比例”缩减。

第五，组织需要设计“自动升级门”。例如：AI 引用无法打开或与正文不符，自动把引用区从 V1 升级至 V2；关键数字在两个来源间差异超过阈值，自动要求局部复算；涉及生产数据库写入、权限变更或不可逆操作，自动要求第二人员审批；涉及法规、政策、价格、时效等易变化信息，自动要求最新权威源；涉及高后果专业判断，自动禁止仅凭模型自检关闭任务。升级门把验证从依赖个人警觉的软要求，转成流程结构。

第六，组织还应建立“验证回报学习”。每次发现 AI 错误，都不应只记录“模型又错了一次”，而应记录错误属于哪种失效模式、原本哪个屏障应该发现它、为什么没有发现、下次能否用更便宜的测试提前截获。经过一段时间，组织会形成自己的错误

谱：某类模型在引用日期上风险高，某类数据任务最常见的是口径遗漏，某类代码生成容易在并发与权限处失稳。MSVP 因此不是一次性清单，而是可学习的风险模型。验证越多，组织越应该知道哪里可以少查、哪里绝不能少查。

第七，组织必须防止“验证自动化悖论”。当 AI 也被用于自动生成测试、自动核对来源、自动打分时，验证成本确实可以进一步下降，但如果整个验证链都由同一模型生态自动完成，人类可能只看到一个漂亮的“95 分可信度”，却不知道分数背后的共因错误。自动验证最适合承担 V0、V1 以及部分 V2/V3：批量检查链接、格式、数值一致性、测试执行、差异检测；而风险边界设定、CFS 确认、异常升级和高后果 V4 是否足够独立，仍需要明确责任主体。自动化的目标应是降低人的机械核验负担，而不是把最终判断再次外包给一个不可解释的总分。

最后，MSVP 会改变组织衡量 AI 生产力的方式。若只统计“每人每天生成多少文档、多少代码”，系统会天然奖励未经充分验证的高速输出，并把风险推给下游。更合理的生产力指标应包含“可承诺产出”：多少输出已经通过与其后果匹配的验证，多少重大异常在承诺点之前被拦截，验证成本相对于潜在损失降低了多少。真正的 AI 生产力不是生成速度，而是单位时间内能够交付多少“足以被安全依赖”的结果。

结论：AI 带来的效率增益，只有在验证也被重新设计之后才真正成立

生成式 AI 把人类带入一个新的生产制度：答案变得廉价，而判断答案是否足以被依赖成为新的稀缺能力。如果我们仍把验证理解为“把 AI 做过的事情再做人肉版本”，AI 的效率优势必然在终审环节被抵消；如果我们把验证降格为“快速浏览加直觉接受”，又会把模型的潜在错误直接转化为现实风险。二者之间真正可行的道路，不是折中地“多看一点”，而是重构验证本身。

审计学告诉我们，可靠不需要穷尽检查，而需要围绕重要错误形成合理保证；可靠性工程告诉我们，有限资源必须优先覆盖高后果、难探测、会传播的失效模式，并通过纵深与独立性降低共因失败；实验科学告诉我们，真正有价值的验证必须让结论面临失败的可能，并在关键处引入复现、反例和独立路径。三门学科跨学科对撞后的共同机制，可以概括为一句话：验证不是重复生产，而是以最小成本最大化“结论相关风险压缩”。

因此，本章提出的最小充分验证路径并不追求“全部都查过”，而追求“所有足以改变行动的错误路径都受到与后果相称的检查”。它以行动后果为起点，以承重主张和结论翻转集确定优先级，以风险导向抽样节省低风险区域的成本，以判别性检验攻击关键推理，以独立第二路径保护高后果结论，并以残余风险低于行动阈值作为停止规则。

在这个框架下，人的最终把关责任不再等于亲手完成全部工作，而等于拥有验证制度：人必须知道自己为什么可以相信、哪里仍然不能相信、什么证据会迫使自己改判，以及哪些风险在执行前绝不能留给 AI。只有当验证从“重做任务”升级为“设计证据”，生成式 AI 带来的效率增益才不会在最后一道关口被重新花掉；而人的责任也不会因为效率提升而消失，反而会从低水平重复劳动上升为对关键事实、关键约束与关键后果的真正治理。

由此，AI 时代真正需要普及的不是一种新的“怀疑主义”，而是一种有结构的信赖能力。无差别怀疑会让人重新陷入全量重做，无条件信赖则把判断权交给不可见的生成机制。MSVP 试图建立的是第三种状态：人可以充分利用机器的速度，同时保留对行动条件的主权；可以允许大量低风险内容由 AI 高效完成，同时把最宝贵的人类注意力留给结论翻转点、不可逆后果和价值边界。未来人机协作的成熟程度，也许不应以“AI 替人做了多少”衡量，而应以“人在不重复机器劳动的情况下，仍能否说明为什么这个结果值得被依赖”衡量。

附录 A 最小充分验证快速清单（用于实际工作流）

1. 先写明输出将用于什么行动，以及错误是否会造成资金、权利、安全、声誉或不可逆后果。

2. 用一句话写出最终结论 C，并逆向列出使 C 成立的 3—10 个承重主张 L。

3. 标出结论翻转集 CFS：哪些事实、数字、规则或假设一变，C 就必须改变。

4. 所有 CFS 节点至少逐项核验；高后果节点不得只靠同一模型自检。

5. 大量同质低风险内容采用风险抽样+随机抽样，而不是随手浏览。

6. 每个关键结论至少设计一个“可能让它失败”的判别性测试。

7. 关键数字、代码结果或数据结论至少进行一次局部独立复算/执行。

8. 发现一个异常时，检查其是否提示共同来源、共同提示或共同数据的系统性问题。

9. 高后果、不可逆、难探测任务启用 V4 独立第二路径。

10. 停止前回答：剩余未核内容即使有错，是否仍可能改变核心行动？若答案是“可能”，验证尚未充分。

附录 B 一个完整演示：如何在不重做一份 AI 研究报告的情况下完成高质量验证

假设 AI 在二十分钟内生成一份两万字的行业进入研究报告，结论是“公司应在未来六个月进入 B 国市场，并优先采用直营模式”。报告包含四十个网页来源、十二张数据表、三套财务测算和若干监管说明。传统方法有两个极端：一是管理者因为报告形式完整而直接采纳；二是研究团队重新检索全部资料并从头写第二份报告。MSVP 的目标是找到第三条路径。

第一步，界定行动：该报告将进入董事会决策，可能触发数百万元投入，撤回成本较高，因此风险等级至少为高。第二步，逆向建立验证图。最终结论 C 有两个部分：C1“六个月内进入”，C2“采用直营”。进一步识别承重主张：L1 目标市场真实可服务规模足够大；L2 许可证在六个月内可获得；L3 当地单位经济模型为正；L4 直营模式相较代理模式具有更高长期净现值；L5 公司当前现金流能够承受前十二个月投入。其余关于文化、人口、宏观趋势的大量背景材料暂时不进入最高验证层。

第三步，寻找结论翻转集。团队通过敏感性分析发现，只要许可证周期超过九个月，C1 就应改变；只要获客成本比报告估计高 35%，直营模式净现值就低于代理模式；只要市场规模口径从“全部相关消费”改成“公司产品可触达细分市场”，收入预测下降约 40%。于是许可证周期、获客成本和市场规模口径构成第一组 CFS。此时验证重点已经从两万字全文收缩到少数真正决定行动的节点。

第四步，为 CFS 选择不同机制的验证。许可证周期不再让原模型继续搜索，而由人员直接进入监管机构官方网站核对申请流程、近期通知与许可类别，并向当地专业服务机构确认实际周期；获客成本不用阅读更多行业评论，而从公司既有相似市场数据与两个独立广告平台基准重建区间；市场规模则要求从原始统计表重新计算可服务市场，而不是接受报告中的二手估算。三个节点分别获得了与原生成路径不同的证据。

第五步，对其余内容实施风险抽样。四十个网页来源中，全部核验直接支撑 L1—L5 的十二个来源，其余二十八个按“最新数据、精确数字、陌生机构、跨语言来源”提高抽样概率，再加入五个随机样本。若三十余个已检查节点只发现两个不影响结论的格式问题，可以维持当前抽样；如果随机样本中发现两处引用与正文不匹配，则引用总体风险上升，需要扩大抽样并检查是否存在同一生成模式导致的系统性错配。

第六步，设计判别性测试。团队不是问“还有哪些证据支持进入 B 国”，而是建立三种失败情景：许可延迟、获客成本上升、汇率恶化，并观察方案排序。结果显示，在许可延迟和获客成本同时恶化的中等情景下，“立即直营”不再占优，而“先代理试点、获得许可后再切直营”更稳健。验证由此不仅发现错误，还可能改变原结论的结构。

第七步，执行 V4 独立第二路径。因为决策金额较大，团队让另一名未参与报告生成的分析人员仅使用原始数据与明确问题，独立计算两种进入模式的现金流，不提供 AI 原报告的最终推荐，以减少锚定。第二路径得到“直营长期价值更高，但对许可时间与获客成本高度敏感”的结果，与前述判别性测试一致。此时团队无需再写第二份两万字报告，却对真正决定行动的节点获得了高强度保证。

最后，应用停止规则：背景章节仍有一部分事实未逐项核验，但团队判断，即使其中存在若干小错，也不会改变经验证后的建议——“先以可逆试点进入，待许可与获客成本达到阈值后转直营”。所有能够让这一建议翻转的关键节点已经获得独立证据，剩余错误的可传播范围有限且在执行前仍有纠正窗口。于是验证可以停止。这个案例说明，所谓不重做，不是降低严谨性，而是把严谨性集中在对行动真正有支配力的位置。

本章附表

维度	审计学	可靠性工程	实验科学方法论	MSVP 中的转化
验证对象	重大错报风险	关键失效模式	可失败的核心主张	承重主张/结论翻转集
资源配置	重要性+风险导向	严重度/发生/可探测	对关键假设做严峻检验	按结论相关风险配置验证
覆盖方式	抽样+实质性程序	FMEA/FTA+纵深防御	反例、复现、竞争解释	抽样+判别检验+独立路径
保证概念	合理保证	残余风险可接受	暂时经受反驳	残余风险低于行动阈值
独立性	外部证据/职业怀疑	IV&V/冗余/异构机制	独立复现	证据独立度

级别	主要动作	适用对象	典型证据	是否可作为高风险终审
V0	约束/来源完整性	所有任务的起点	版本、日期、法域、来源存在性	否
V1	风险导向抽样	大量同质低—中风险节点	抽样核对、随机样本、异常率	否
V2	承重主张逐项核验	CFS、高中心性主张	原始来源、直接证据、规则核对	视后果而定
V3	判别性检验/局部复算	关键推理、数字、算法	反例、边界测试、重算、敏感性分析	多数中高风险可用
V4	独立第二路径	高后果/不可逆/难探测	独立数据、工具、模型、人员或方法	是

输出组成	默认策略	升级触发器

最终结论/关键建议	逐项核验，至少 V2	高后果、不可逆、对外承诺→V4
关键数字/日期/法域/接口/阈值	逐项核验+局部复算 V3	任何异常或来源不独立→V4
大量同质背景事实	V1 风险抽样+随机抽样	出现系统性错误→扩大抽样/全检
引用与来源	V0 存在性+V2 对应性	虚构/错配出现→章节级升级
长推理链	识别承重步骤，V3 判别性测试	结论对假设高度敏感→V4
可逆创意内容	轻量 V0/V1	转为事实发布或对外承诺→升级

任务	**首先核什么**	**最有判别力的测试**	**高风险时的第二路径**
研究综述	核心结论—承重文献对应	找反例/核研究设计与效应方向	独立数据库/专家复核核心证据
代码	接口、权限、边界条件	单测、性质测试、模糊测试、独立 oracle	独立审查者/测试环境
数据分析	数据版本、口径、过滤、单位	关键统计量复算、敏感性分析	独立实现/独立分析者
法律意见	法域、时点、有效法源、例外	不利先例/例外适用性检验	合格专业人员+独立法源
决策建议	关键事实、概率、价值权重	翻转条件/情景敏感性测试	外部基准/独立预测/试点

第四编　从个体到群体，再回到这本书自己

前三编的对象都是个体与岗位。本编先把同一条判断放大到群体，再把它对准本书自身。

第九章处理一个反直觉的结果：每个人都借助人工智能变得更准，群体发现共同错误的能力却可能下降。原因不在平均水平，而在错误之间的相关性——人数不等于独立证据数，而共享同一模型、同一语料、同一检索路径的判断，会以"多数独立同意"的外观出现。这一章给出的折扣公式，是全书唯一一个把"复制不能当作重复观察"写成算式的地方。

第十章是全书合论。它把九章的判断摆成一张四列九行的换位表，指出三个共绕的动作、三对最可能被合并的近亲及其判定条件、九个读数其实只有三种形状，并给出三条书级撤回条款。

第十一章处理最后一件事，也是本书自认最可能应验的失败：一套用于揭露遮蔽的方法，为什么会长出新的遮蔽。它列出四条应付路径、四条对策及其代价、一条不可让步的底线、五个可以从原始记录读出的预警信号，以及在对策失效时本书愿意接受的两级降级方案。

把这一章放进正文而不是附录，是本书对自己开的那个断口。

第九章

都更聪明，只剩一种错法

认知单一化悖论与认知响应多样性　——统计学习理论 × 文化进化论 × 生态学的跨学科对撞

统计学习理论 × 文化进化论 × 生态学的跨学科对撞

——为什么个体平均认知绩效提高，群体发现共同错误与修正共识的能力却可能下降

摘 要

生成式人工智能正在同时改变个体认知绩效与群体认知结构。越来越多的人借助性能更强的基础模型完成写作、研究、分析、编程与决策，个体答案的平均质量可能显著提高；然而，群体纠错能力并不只由个体错误率决定，还取决于不同个体之间的错误是否相互独立、是否保留足够的异源推理路径，以及共同错误出现后系统能否产生不同响应。本章采用跨学科互否方法，将统计学习理论中的集成误差、协方差与预测多样性，文化进化论中的成功偏置、声望偏置、从众传递与路径强化，以及生态学中的响应多样性、功能冗余、单一化与系统韧性进行结构映射。三门学科的深层同构在于：一个系统的平均单元质量与系统韧性并非同一变量；当单元之间的失效相关性上升时，更高质量的单元仍可能组成更脆弱的整体。本章据此提出“认知单一化悖论”（Cognitive Monoculture Paradox, CMP）与“认知响应多样性”（Cognitive Response Diversity, CRD）两个核心概念，并构造“相关错误放大回路”：高性能 AI 首先降低个体平均错误率，同时通过共享基础模型、相近语料、相似检索与推理模板提高错误相关性；个体绩效提升又成为成功偏置的强信号，促使更多人复制相同认知路径；相似输出随后彼此成为引用、搜索、组织模板与社会共识的证据，使依赖同源的答案获得“多数独立同意”的外观；这种表面共识进一步压缩异质判断，最终使低频却高度相关的共同错误更难暴露。本章进一步给出群体有效独立判断数 $N_{eff} = n/[1+(n-1)\rho]$ 的近似指标，用以说明参与者数量不等于独立证据数量；提出“共识独立性折扣”“异源性预算”“少数路径保护”“异构第二模型/第二资料源”“先独立后汇聚”等治理原则。本章的核心结论是：AI 时代真正需要保护的不是意见分歧本身，而是失效模式的去相关化与同功能任务下的异响应能力。一个高水平群体可以在多数时候快速趋同，但必须制度性保留能够在共同模型失效时看见不同证据、提出不同反例并启动不同纠错路径的认知生态位。

关键词：生成式人工智能；集体认知；错误相关性；预测多样性；文化进化；成功偏置；响应多样性；认知单一化；认知韧性；人机协作

Abstract

Generative AI is changing not only individual cognitive performance but also the dependence structure of collective cognition. As more people rely on increasingly capable foundation models for writing, research, analysis, programming and decision-making, average individual performance may improve while the collective capacity to detect shared errors, sustain genuine dissent and revise common judgments fails to improve—or even deteriorates. This paper develops a second-order interdisciplinary collision among statistical learning theory, cultural evolution, and ecology. Statistical learning contributes ensemble error, covariance and predictive diversity; cultural evolution contributes success-biased and prestige-biased transmission, conformity and path reinforcement; ecology contributes response diversity, functional redundancy, monoculture and resilience. Their common structure is that average component quality and system resilience are distinct variables: a set of better-performing components can form a more fragile system when their failure modes become highly correlated. We propose the Cognitive Monoculture Paradox (CMP) and Cognitive Response Diversity (CRD), and model a correlated-error amplification loop in which AI raises average performance, makes successful cognitive pathways more attractive to copy, increases dependence among judgments, and allows repeated model-shaped outputs to masquerade as independent social evidence. We further introduce an effective number of independent judgments, $N_{eff} = n/[1+(n-1)\rho]$, to distinguish headcount from evidential independence, and develop governance principles including consensus-independence discounting, heterogeneity budgets, protected minority pathways, heterogeneous second models/sources, and "independent first, aggregate later" protocols. The central claim is that AI-era collective intelligence must preserve diversity of failure response rather than disagreement for its own sake. A resilient high-performance group may converge quickly under normal conditions, but it must retain institutionally protected cognitive pathways capable of seeing different evidence, generating different counterexamples, and initiating different corrections when a shared model fails.

Keywords: generative AI; collective cognition; correlated errors; predictive diversity; cultural evolution; response diversity; cognitive monoculture; resilience; human-AI collaboration

问题提出：为什么“大家都更聪明”不等于“群体更不容易犯大错”？

生成式 AI 最直观的影响，是把许多认知任务的个人起点整体抬高。一个原本不会写代码的人可以生成可运行的程序原型，一个不熟悉某一领域的人可以迅速得到结构完整的综述，一个普通职员可以在几分钟内得到看起来像专业咨询报告的分析。实验研究已经观察到类似的个体增益：例如 Doshi 与 Hauser 在创意写作实验中发现，获得生成式 AI 建议的参与者能够产出评价更高的故事，尤其对原本创造力较低的参与者帮助更明显；但与此同时，AI 辅助组的故事在群体层面变得更相似[8]。这提示我们，一个工具完全可能同时提高单个产出的平均水平，并改变整个群体输出之间的依赖关系。

传统的“智慧群体”直觉往往把人数当成可靠性的来源：十个人独立判断，比一个人更不容易被单个偶然错误支配；不同研究者、程序员、医生、律师或管理者来自不同训练背景，也会带来不同的盲点和不同的纠错入口。然而，这一优势隐含着一个经常被忽略的前提——个体之间的错误必须至少部分独立。如果十个人实际上都复制同一份数据、同一套模板或同一个判断器，那么十张选票并不等于十份证据。人数没有减少，但有效独立判断数已经下降。

AI 使这一隐含前提变得前所未有地重要。过去，人们即便都犯错，也常常因为教师、书籍、专业传统、地域经验、职业工具和推理习惯不同而以不同方式犯错。一个人误读统计，一个人忽略制度边界，另一个人高估历史类比；这些错误并不理想，却给群体留下了交叉纠偏的机会。如今，大量人可能调用同一类基础模型、相近的训练语料与检索来源，使用从网络上复制来的相似提示词，再把 AI 输出改写成自己的文章、代码或建议。即使最终文本不完全一样，它们也可能共享同一个知识盲点和同一条推理路径。

因此，本章提出一个与“AI 是否让个人更聪明”不同的问题：当个体平均错误率下降时，错误相关性会发生什么？如果每个人从 20%的错误率降到 10%，这是显著进步；但如果过去十个人的错误相互较为独立，而现在十个人在那 10%的困难问题上高度同步出错，那么群体可能更少犯小错，却更难发现剩下的大错。这不是概率意义上的矛盾，而是把“平均准确率”和“联合失效结构”区分开之后自然出现的结果。

更值得警惕的是，共同错误并不会停留在第一次输出。AI 生成的文本可以被搜索引擎收录、被作者互相引用、被组织写进模板、被媒体再摘要、被新一轮 AI 当作上下

文，甚至在更广义的数据循环中重新进入后续模型训练。Shumailov 等人的研究从模型训练层面显示，递归使用模型生成数据可能导致分布逐步退化和模型崩塌[11]；在人类文化层面，同样需要追问一种不同但相似的反馈：当模型生成的表达不断回到社会信息环境，它是否会把原本依赖的共同偏差伪装成日益稳固的社会共识?

本章的核心任务由此确定：不是证明生成式 AI 必然导致群体单一化，而是建立一个能够解释“个体提升与集体脆弱可以同时发生”的机制框架，并明确什么条件下这种风险会上升、什么样的制度设计可以保留 AI 带来的平均绩效收益，同时阻断共同错误的相关化、社会放大与自我强化。

跨学科互否方法：三门学科为什么在这里遇到的是同一个问题?

如果只做第一阶跨学科拼接，我们可以很容易得到三个口号：机器学习告诉我们“模型要多样”，文化进化告诉我们“不要盲从”，生态学告诉我们“避免单一化”。这些结论都正确，却不足以形成新的理论。跨学科对撞要求继续追问：三门学科为什么会分别产生“多样性”这个概念？它们所面对的深层困难是否同构?

统计学习理论研究集成系统时发现，多个预测器的整体表现不只取决于每个预测器各自多准确，还取决于它们犯错时是否一起犯错。经典的偏差—方差—协方差视角以及后续的多样性理论都说明，模型之间的统计依赖会改变集成收益[1][2][3]。文化进化论研究社会学习时发现，复制成功者、声望者和多数人通常可以节省个体试错成本，但过强的高保真复制与从众会收缩变异，使一种路径一旦占优便更容易自我强化[19][21][22]。生态学则发现，生态系统的韧性不能只看有多少物种或单个物种表现多好，而要看承担相同生态功能的成员在环境变化时是否作出不同响应；这种“响应多样性”决定了同一扰动是否会让整个功能同时失灵[16]。

三者的共同问题可以压缩为一句话：系统可靠性取决于单元质量，也取决于单元之间的失效相关结构。 如果所有单元都以不同方式失败，系统可以通过冗余、比较和替代维持功能；如果它们以同一种方式失败，冗余就会退化为复制。由此，本章不把“多样性”理解为价值口号，而把它定义为一种失效去相关资源。

在这个共同结构上，三门学科承担不同层次的解释任务。统计学习理论回答“为什么相关错误会吞噬群体集成收益”；文化进化论回答“为什么高性能 AI 恰恰可能使这种相关性随着使用扩散而上升”；生态学回答“为什么一个平时效率更高、更整齐的认知系统，面对罕见扰动时反而可能更脆弱，以及应该保留什么样的异质冗余”。三者碰撞之后，我们才能从静态的“答案多样”推进到动态的“群体认知韧性”。

第一重碰撞：统计学习理论——平均错误率不是群体错误风险的充分统计量

设群体中有 n 个判断者，每个判断者在某一关键任务上是否出错记为 X_i∈{0,1}。如果每个判断者的平均错误概率为 p，且两两错误指示变量的平均相关系数近似为 ρ，则群体平均错误率 X̄ 的方差可以写成一个常见的等相关近似：

$$\mathrm{Var}(\bar{X}) \approx [p(1-p)/n] \cdot [1 + (n-1)\rho]$$

其中，p 表示平均个体错误概率，ρ 表示不同判断者错误之间的平均两两相关。该式并不直接给出多数决策出错概率，却清楚显示：当 ρ 上升时，增加人数所带来的方差压缩会被迅速削弱。

相关性如何吞噬人数优势

这个式子揭示了一个重要事实：当 ρ 接近 0 时，增加人数能够以约 1/n 的速度压低平均误差的不确定性；而当 ρ 显著为正时，n 个判断者中有一部分只是统计意义上的“重复”。据此可以定义一个直观的有效独立判断数：

$$N_{eff} = n / [1 + (n-1)\rho]$$

N_eff 可理解为在等相关近似下，与当前群体具有相似平均误差方差的“独立判断者数量”。它不是现实社会中精确的独立人数，而是一种把“人数”与“信息独立性”分开的概念工具。

“十个人只剩一种错误”为什么不是修辞，而是可计算的结构变化

假设一个十人群体在没有共享 AI 时，个体错误率约为 20%，错误相关系数只有 0.05，则 $N_{eff} \approx 10/[1+9\times0.05] \approx 6.90$；使用高性能 AI 后，个体错误率降到 10%，看上去每个人都进步了，但如果由于共享模型与共同资料源使错误相关系数上升到 0.80，则 $N_{eff} \approx 1.22$。这个示例不是对现实数据的估计，而是说明一种可能性：参与人数不变、平均准确率提高，有效独立证据数量却可以大幅下降。

这并不意味着群体总体一定更差。对大量常规问题而言，更低的 p 往往仍然带来明显收益；真正变化的是尾部风险结构。过去的错误更分散，群体中往往有人能在别人出错时提出不同答案；现在的错误更少，却更可能集中在模型共同不擅长、共同过时或共同偏置的那一小部分问题上。于是，平时体验变得更顺畅，罕见共同失效却更难从内部暴露。

Kim、Garg、Peng 与 Garg 对数百个大语言模型的实证分析提供了直接警示：不同 LLM 的错误并不天然独立；在其分析的 HELM 数据上，当两个模型都答错时，它们选择相同错误答案的平均比例约为 60%，明显高于在三个错误选项中均匀随机选择所对应的 1/3；同一提供商、相似架构会提高错误相关性，而且更准确的模型之间也可能表

现出更高的错误相关性[10]。这说明“换一个模型再问”不能自动等价于获得一份独立证据。

统计学习由此给出本章的第一条关键命题：群体平均认知绩效的改进，必须与错误相关性共同报告。只告诉我们“十个人使用 AI 后平均得分从 70 提高到 85”，不足以判断群体韧性；还需要知道，那 15 分的剩余错误是否集中在同一批题、同一类事实、同一类反例或同一类推理失效上。

从“预测多样性”到“认知异源性”：真正需要多样的不是措辞，而是错误函数

机器学习中的“多样性”并不是简单要求每个模型输出不同答案。一个故意胡说八道的模型当然很“不同”，却不会提高集成质量。Wood 等人的统一理论强调，多样性应放在偏差、方差与模型拟合的整体权衡中理解，而不是无限最大化[1]。同理，AI 时代的群体认知也不需要为了多样而多样。真正有价值的是：在保持足够个体能力的前提下，不同判断路径对问题中的风险点具有不同敏感性。

本章把这种资源称为“认知异源性”。它至少包含五个维度：第一，来源异源性，即证据来自不同数据库、不同机构、不同历史经验，而不是十个人都从同一摘要中取材；第二，模型异源性，即不同工具的训练、架构、检索后端或治理机制不完全同源；第三，方法异源性，即有人做统计检验，有人做因果分析，有人做案例反证，有人做代码执行；第四，经验异源性，即参与者拥有不同专业与现场经验；第五，激励异源性，即群体中存在被允许挑战主流结论而不因“不合群”受罚的角色。

这五种异源性最终要落到一个更严格的问题上：当同一个关键错误发生时，是否至少有一条路径不会同时失效？如果十个作者使用不同写作风格，却都依赖同一个 AI 检索出的错误数据，那么表面上有十种声音，实质上只有一个证据源；反过来，两个人即便最终得出相同结论，只要一人通过原始数据重算，另一人通过不同数据库与反例检索验证，他们就提供了更高的认知独立性。

因此，“观点多样性”只能作为粗糙代理。真正需要测量的是“失效模式多样性”。这也是本章从统计学习进入文化与生态层面的桥梁：一个社会可以表现得十分热闹、文本十分丰富，却在深层认知基础设施上越来越同源。

第二重碰撞：文化进化论——为什么越成功的 AI，越可能加速认知路径趋同？

如果相关错误只来自“大家恰好用了同一个工具”，问题似乎可以通过鼓励多用几个工具解决。但 AI 的特殊之处在于，它不仅是一个工具，还会进入文化传递过程。文化进化论研究表明，人类并不是每次都从头独立学习，而会使用一系列社会学习偏置：复制看起来成功的人、受尊敬的人、被很多人采用的方法，通常能显著降低学习成本[19][20][21]。在大多数环境中，这些偏置是适应性的；问题在于，它们会把“当前看起来成功”转化为“未来更容易被复制”。

高性能生成式 AI 恰好提供了极强的成功信号：速度快、表达完整、错误率较低、能跨越专业门槛，还能把普通人的产出迅速抬升到“像专业作品”的水平。于是，使用 AI 本身成为一种可见成功策略。个体看到同事借助某模型写得更快、考试准备更充分、代码交付更及时，自然会复制其工具、提示词和工作流；组织看到某模板效率高，也会把它标准化。这是一种典型的成功偏置扩散。

当成功偏置与规模平台结合时，文化变异会进一步收缩。过去，一个新人可能分别从五位教师、三本教材和一次亲身失败中形成自己的方法；现在，他可能直接获得一个经过大量人类文本压缩后的“最佳实践答案”。这当然节省时间，但也跳过了许多原本会产生差异的中间路径。文化传播从“多人、多链条、低速变异”转向“少数模型、高保真、低成本复制”。

更复杂的一层是声望偏置与二阶线索。Jiménez 与 Mesoudi 总结的声望偏置研究指出，当我们难以直接判断一个人的能力时，会使用其他人对其关注和敬重程度作为间接线索[21]。在 AI 环境中，这种机制可能被转译为：某答案被大量引用、某模板出现在很多报告里、某模型被很多专业人士使用，于是“很多人都在这么做”反过来成为可信度线索。然而，如果这些人本来就依赖同一模型，那么所谓“很多独立主体的支持”其实是一份源头被复制了很多次的证据。

从众传递会进一步强化这种错觉。Denton 等人关于从众与反从众文化传递的模型显示，从众偏置能够深刻改变文化变体的频率动力学[22]。对 AI 输出而言，一种说法一旦占据高频位置，就更容易进入搜索结果、培训材料、组织范文与下一轮写作，从而提高被再次观察和复制的概率。此时，流行不再只是结果，也成为下一轮传播的原因。

本章据此提出“共识独立性折扣”：当我们观察到 m 个主体持有同一判断时，不能直接把 m 当成 m 份证据，而应追问它们背后的来源与生成路径有多少是独立的。如果十篇文章都直接或间接由同一基础模型和同一原始来源生成，它们的社会可见度可以是十倍，证据独立性却可能接近一倍。这种“可见共识膨胀”是 AI 时代文化传播中特别值得治理的机制。

相关错误放大回路：一次局部模型偏差如何长成“大家都这么认为”

三门学科碰撞到这里，可以构造出一条完整的动态回路。第一步，某高性能模型在绝大多数任务上显著提高用户表现，于是获得信任与使用规模；第二步，更多个体采用同一模型、同类检索源和相似提示模板，群体错误相关性上升；第三步，模型在某个少见问题上存在共同盲点，很多人独立操作却得到相近错误答案；第四步，这些答案被写入文章、代码注释、内部文档、社交媒体与数据库；第五步，后来者看到“多个来源都这样说”，把重复复制误认成独立支持；第六步，主流答案频率上升，异质路径因为成本更高、可见度更低而被进一步边缘化；第七步，下一轮 AI 检索与人类写作又从这一被放大的信息环境中取材，使原本局部的偏差获得路径依赖。

这条链条可以称为“相关错误放大回路”（Correlated Error Amplification Loop, CEAL）。它与经典的信息瀑布、回音室或错误传播并不完全相同。它的特殊性在于，第一步并不是低质量信息获胜，而是一个总体质量很高的系统因为成功而获得更大文化权重；正是高平均质量降低了人们保留替代路径的动机。风险因此不会首先以“大量明显错误”的形式出现，而可能以“绝大部分时间都很好用，偶尔大家一起错”的形式出现。

这也解释了为什么传统的错误率监控可能发现得太晚。只要模型在总体基准上持续进步，机构就会有理由继续集中采购与标准化；只要日常任务的平均质量持续提高，用户就会更愿意把独立检索和手工推导视为低效率。于是，认知异源性像一种看不见的储备逐步被消耗。直到遭遇一个恰好落在共同盲点上的高后果问题，系统才发现自己虽然有很多人、很多文档、很多“复核”，却缺少真正独立的第二条路。

因此，AI 时代的系统性风险不应只用“模型会不会幻觉”描述。更深的风险是：一个原本局部可纠正的错误，通过成功偏置、复制便利与共识线索，转化为跨主体的共同失效。模型风险与社会传播结构在这里发生了二阶耦合。

第三重碰撞：生态学——为什么“功能都很好”仍然可能是一片认知单一林?

生态学提供了一个比“多样性越多越好”更精确的概念：响应多样性（response diversity）。Elmqvist 等人把它定义为承担相同生态功能的不同物种对环境变化作出不同反应的程度，并指出这种差异对于生态系统在扰动后的更新与重组至关重要[16]。关键不在于物种名称不同，而在于当温度、干旱、病原或资源条件改变时，它们是否会同时失败。

这一概念迁移到群体认知非常有力量。一个组织需要很多人完成相同功能，例如审查合同、判断研究证据、评估投资方案、检测代码安全。功能冗余本来是好事：一个人失误，另一个人可以接住。但如果所有人都使用同一模型、同一数据库和同一推理模板，那么他们在日常条件下看起来效率极高、输出也整齐一致，却可能拥有极低的响应多样性。一旦共同模型在某一类条件上失效，所有“冗余”会同时失灵。

本章因此提出“认知响应多样性”（Cognitive Response Diversity, CRD）：在群体成员执行相同认知功能时，面对事实扰动、反例、边界条件变化、数据异常或模型失效，能够产生不同检测信号、不同解释与不同纠错动作的程度。CRD 不是要求每个人坚持不同意见，而是要求系统保留不同的响应函数。两个人平时完全可以都支持同一个结论，但当关键证据被撤掉时，一个人会从统计稳健性上发现问题，另一个人会从法理边界上发现问题；这就是高响应多样性。

生态学还提示我们区分“功能冗余”和“响应冗余”。十个完全相同的备份可以提高应对随机单点故障的能力，却不能抵御共同病原；十个不同品种承担相似功能，才更可能在环境突变时至少保留部分功能。认知系统同样如此：同一个 AI 答案由十个人分别润色，增加的是生产冗余；由不同模型、不同资料源、不同专业方法对同一承重结论进行复核，增加的才是响应冗余。

这也揭示“认知单一化”的真正含义。它不是所有人说同一句话，而是系统的关键功能越来越依赖少数共同的生成机制。当多样的职业训练、地方经验、手工技能与独立资料源逐渐退出日常使用，系统可能仍然保留人口学上的多样、表达风格上的多样，却在关键失效机制上变成单一作物。

新概念一：认知单一化悖论（CMP）——更高平均质量如何与更低群体韧性同时成立

基于上述碰撞，本章定义“认知单一化悖论”（Cognitive Monoculture Paradox, CMP）：当一个高性能认知工具提高多数个体的平均表现，同时显著提高不同个体之间的错误相关性、压缩独立路径并强化同源社会传播时，群体可以在平均绩效上进步，却在发现和修正低频共同错误的能力上下降。

这个悖论至少包含四个彼此可分离的变量。A 表示个体平均准确度；ρ 表示关键错误相关性；D 表示认知响应多样性；R 表示群体在共同失效后的恢复能力。AI 可能使 A 上升，同时使 ρ 上升、D 下降；而 R 取决于群体是否仍保留未被共同模型吞并的替代路径。只报告 A，会系统性高估这类系统的安全性。

因此，一个更完整的群体认知评估不应问“AI 辅助后平均得分提高多少”，还应问：困难项目上的错误重合度是多少？不同人是否从真正独立的资料源得到相同结论？当移除共同模型或共同数据时，结论是否仍然成立？是否存在少数人能够在主流答案错时稳定发出异议信号？错误一旦形成共识，系统需要多大成本才能恢复？这些问题分别对应 ρ、证据独立性、CRD 与 R。

CMP 也解释了为什么“模型越强，问题自然会消失”并不充分。模型更强当然可以继续降低 p，这是第一层收益；但如果整个社会越来越集中于少数强模型，那么 ρ 可能同时上升。两条趋势谁占主导，取决于任务类型与风险结构。对于大量低后果日常任务，p 下降的收益可能远大于相关性代价；对于少数高后果、非常规、分布外问题，相关性代价可能成为决定性因素。

所以，本章并不主张为了多样性拒绝更强 AI，也不主张人人都必须从头独立完成任务。真正需要的是把 AI 的“平均能力红利”与群体的“异源性储备”同时纳入设计，使系统在常规环境下享受规模效率，又在罕见扰动到来时仍拥有逃离共同错误的路径。

新概念二：认知响应多样性（CRD）——群体需要的是“同功能、异响应”

CRD 可以进一步拆成四类响应。第一是证据响应多样性：面对同一事实争议，不同路径会调用不同来源，例如原始数据、法规原文、现场记录、独立数据库，而不是都检索同一二手摘要。第二是方法响应多样性：有人通过计算复核，有人通过反例测试，有人通过历史基准，有人通过因果识别。第三是异常响应多样性：同一异常出现时，有的路径检查数据质量，有的路径检查模型假设，有的路径检查制度约束。第四是价值响应多样性：在事实相同的情况下，不同角色能够暴露不同利益与风险权重，防止模型默认价值排序被误当成唯一答案。

一个高 CRD 群体并不一定在最终表决时分裂。恰恰相反，它可以比低 CRD 群体更快形成可靠共识，因为各条独立路径在关键点上已经完成交叉验证。真正要避免的是“先由同一模型形成答案，再让所有人围绕它做微调”。一旦共同锚点先出现，后续判断很容易变成相关修正，而不是独立推理。

由此可以把 AI 辅助流程分为两类。第一类是“先汇聚后分工”：AI 先给全体一个标准答案，成员再分别编辑、检查和补充。这类流程效率高，但极易造成锚定与共因失效。第二类是“先独立后汇聚”：不同成员或代理在看见彼此答案之前，分别从不同来源/方法形成判断，最后才聚合。后一类成本更高，却能显著保护独立性。高风险任务应优先采用第二类。

Navajas 等人的研究表明，把大型群体组织成若干彼此独立的小组先讨论，再聚合小组共识，可以在保留局部社会学习好处的同时获得高质量集体判断[33]。更广泛的集体智能研究也显示，网络结构、信息流与独立性并非越密越好，而会与任务复杂度发生交互[6][7][28]。CRD 因此不只是“人员构成”属性，也是流程与网络结构属性。

三学科同构表：从模型集成、文化传播到生态韧性

为了避免跨学科类比停留在修辞层面，可以把三门学科中的对应结构明确列出。三者都区分“单体表现”与“系统表现”，都承认相关性会削弱冗余价值，也都说明过度集中虽然提高短期协调效率，却可能降低面对新扰动时的恢复空间。

共同错误从哪里来：三个层级的共因失效

相关性不是一个抽象数字，它总要通过某种共同原因产生。第一层是“模型共因”：不同个体直接使用同一基础模型，或者使用虽然品牌不同但共享相似架构、训练数据与对齐方法的模型。此时共同错误可能来自知识截止、训练分布缺口、相似推理偏好或同类安全策略。Kim 等人的结果说明，同提供商、同架构与模型规模都与错误相关性有关，但即使控制这些因素，更准确模型之间仍可能保持较高相关[10]。因此，表面上的产品多样不等于失效机制多样。

第二层是“证据共因”：即使人和模型不同，只要所有判断最终引用同一个错误原始来源，群体仍会共同失败。现实中这类共因尤其隐蔽，因为十篇文章可能分别由不同作者、不同模型写成，语言风格完全不同，但它们的关键数字都来自同一篇错误报告。若只看作者数量或文本差异，会把一个证据节点错误计为十个独立节点。共识谱系审计因此必须追溯到事实与数据的源头，而不能停在“用了几个模型”。

第三层是“文化共因”：不同人甚至可以拥有不同证据，却因为先接受同一个问题框架而共同遗漏某类替代解释。生成式 AI 非常擅长给出完整的问题表述、分类框架和标准论证结构，这能显著提高组织效率，却也可能让群体在尚未意识到“问题还可以怎样被问”之前就共享同一认知坐标系。此时共同错误不表现为某个事实错，而表现为共同没有提出某个关键问题。

这三个层级可以形成嵌套：同一模型提供共同框架，共同框架引导大家检索同一类资料，同一类资料又强化最初框架。于是，共因失效并不一定能够通过“再增加一个人”解决；如果新加入的人仍进入同一个模型—资料—框架闭环，他只会提高可见共识，不会提高有效独立性。

因此，评估群体异源性时至少要问三个“不同”：底层生成机制是否不同？承重证据是否不同？问题框架与检验方法是否不同？只要三者全部同源，人数再多也更像一个

放大的单体系统。反之，哪怕只有两三条路径，只要它们在这三个层级中至少有一个关键层真正独立，就可能对共同错误形成高价值屏障。

群体认知韧性不能压成一个平均分：四指标向量

为了避免用单一准确率遮蔽系统结构，本章建议把群体认知状态表示为一个四维向量 G = (A, N_eff, CRD, T_c)。A 是平均个体准确度，表示常规任务中的能力水平；N_eff 是有效独立判断数，表示群体中真正可被视为不同证据路径的数量；CRD 是认知响应多样性，表示在条件变化与共同盲点出现时，群体能否产生不同报警和纠错动作；T_c 是纠错时延，表示共同错误一旦进入系统，从第一个异常信号出现到主流判断被修正需要多长时间。

这四个指标之间不存在简单的单调关系。一个高度标准化的 AI 团队可能拥有很高 A、很低 T_c 于普通可测试错误，却同时拥有低 N_eff 和低 CRD；另一个跨专业团队可能平均速度稍慢，却在分布外问题上拥有更高 CRD。真正的系统设计不是追求四项全部最大化，而是根据任务后果寻找组合。对于低风险任务，A 与速度最重要；对于高风险任务，N_eff、CRD 与共同错误的 T_c 具有更高权重。

T_c 尤其值得强调，因为“能不能最终纠正”与“能不能在造成现实后果前纠正”并不等价。一个错误如果六个月后被学术界修正，却已经在此期间进入成千上万篇 AI 辅助文本，其传播成本可能巨大；一个软件安全错误如果上线五分钟后就由独立监控捕捉，系统韧性仍然较高。生态学中的韧性同样不仅关心扰动后是否恢复，也关心功能在扰动期间如何维持。

由此，群体认知评测可以从“谁答得最准”升级为“谁在什么情况下会一起错、谁能先发现、系统多久能恢复”。这比文本多样性指标更接近真正的公共风险，也使认知单一化成为可实验、可比较的系统属性。

一个十人团队的思想实验：为什么平均错误更少，重大共同错误仍可能更难发现

考虑两个都由十人组成的分析团队。团队 H 主要依靠不同人的专业训练完成任务，成员平均错误率 20%，错误之间相关性较低；团队 A 全部使用一个高性能 AI，成员平均错误率降为 10%，但错误高度相关。对 100 个普通问题，团队 A 显然更舒服：多数成员更快、更准，重复劳动更少。仅看平均成绩，我们完全应该选择 AI 辅助。

现在加入第 101 个问题：它属于训练分布边缘，关键事实又恰好在共同模型中被系统性误解。团队 H 虽然平均能力较低，却可能因为一名成员来自不同专业、另一名成员查了不同数据库而出现两三个异议；团队 A 的十名成员则都得到同样流畅答案。由于过

去 100 题中 AI 表现优异，团队 A 对一致性反而拥有更高主观信心。这个时刻，平均绩效红利与共同错误风险同时存在。

若第 101 题只是低后果知识题，问题不大；若它决定一次不可逆投资、生产安全设置或法律行动，共同错误的后果会突然占据主导。因此，认知单一化风险本质上与“错误后果分布”耦合：平均错误率下降主要改善日常损失，而相关性上升主要影响尾部联合损失。用平均准确率评价高后果系统，就像只用平常产量评价一片单一作物，却忽略同一种病害可能让整片农田同时失收。

最合理的策略并不是让团队 A 退回全部手工工作，而是在第 101 类高风险问题上引入一条低相关路径：例如一名成员不看 AI 结论，直接核对原始数据；或者调用不同方法做反事实测试。只要这一条路径能够在共同模型出错时产生异信号，十人团队就重新获得了关键的响应多样性。异源性预算的意义正是在此：用很少的额外成本购买一份与主导系统不共因的保险。

从“共同错误”到“共同承诺”：风险真正跃迁发生在哪里?

共同错误本身并不自动构成系统灾难。一个群体可以在草稿阶段共同犯错，只要错误尚未进入不可逆行动，仍然有充足机会被外部反馈纠正。真正的风险跃迁发生在“共同认知”变成“共同承诺”的那一刻：报告被正式发布、代码被部署、资金被划拨、合同被签署、政策被执行、训练数据被纳入下一轮系统。此时，错误从信息状态跨越为现实状态。

因此，相关错误治理必须与承诺点绑定。低 CRD 并不意味着每个草稿都要暂停；它意味着在高后果承诺点之前，系统必须补上一条足够独立的路径。换言之，群体可以共享 AI 完成 90%的生产过程，但不能让共享 AI 同时成为最后一道裁决机制。最终承诺越不可逆，越要确保至少存在一个不与主导路径共因的证据来源或责任主体。

这一设计也能避免“为了独立而独立”的成本膨胀。许多早期想法、探索性分析和内部备忘录可以允许高度同源，因为它们仍处在可回滚阶段；只有当结论准备穿过现实边界时，独立性要求才显著上升。这样，认知响应多样性就不再是一种全天候的昂贵冗余，而是一种在关键承诺点被激活的韧性机制。

进一步说，AI 时代的群体治理需要区分三种状态：生成态、验证态与承诺态。生成态追求速度和覆盖；验证态追求失效暴露与证据去相关；承诺态要求责任主体确认残余共同错误风险已经可接受。若三种状态被混在一起，团队很容易把“十个人都看过”误当成验证完成。真正关键的不是看过多少人，而是在进入承诺态之前，是否已经引入足以打破共同错误的独立信号。

由此还可以提出“独立性保存率”这一未来指标：比较一个任务从最初个体判断进入 AI 协作、多人讨论、文档汇总再到最终承诺的全过程，最初存在的独立证据路径有多少在中途被合并、覆盖或遗忘。一个系统即使起点拥有多元团队，如果所有分歧都在第一轮 AI 总结时被压成统一表述，最终仍可能表现为低独立性。保护 CRD 因此不仅是增加新路径，还包括防止已有的独立信号在汇聚过程中被过早抹平。

为什么“多模型”仍不必然等于“多路径”：从品牌多样性到功能异源性

一个看似直接的解决办法，是让不同成员分别调用不同的生成式 AI 模型，再把答案相互比较。但“使用多个模型”只能构成候选的异源性，并不能自动构成有效独立性。不同模型可能共享大量公开互联网语料、相似的人类反馈规范、相近的检索入口与主流知识框架；即使参数架构和提供商不同，它们在某一具体问题上的失效原因仍可能高度重合。因此，模型数量是一种结构性代理变量，而不是独立性的最终判据。

真正需要测量的是“功能异源性”：当主导路径发生某类错误时，另一条路径是否有较高概率给出不同报警。两个模型若在常规问题上措辞差异很大，却在关键反例、边界条件和缺失证据上总是一起失败，它们对群体韧性的贡献就很有限；相反，一个 AI 模型与一名直接读取原始实验记录的人，即使最终结论经常相同，只要二者的证据获取和失效机制不同，就可能形成高价值的独立组合。

这也意味着，未来不应只用“答案相似度”评估认知多样性，而应进行失效扰动测试：人为改变证据来源、加入反例、删除主流线索、改变问题表述或制造分布外条件，观察不同路径的错误是否仍同步发生。若在这些扰动下交叉错误率持续很高，则说明所谓多模型系统仍然接近认知单一化；若某些路径能稳定地在其他路径失效时产生异常信号，它们才真正提高 CRD。

此外，人类自身也不能被浪漫化为天然独立源。十名专家若受同一教材体系、同一行业规范和同一机构激励塑造，同样可能共享盲点。因此，本章所说的“异源”从来不是人机二分，也不是品牌二分，而是对证据、方法、训练历史和责任结构的共因审计。最有韧性的组合可能是人+AI，也可能是不同专业的人、不同工具链的 AI，或现场测量与模型推断的交叉。

由此可以提出一个更严格的设计原则：不要购买“多答案”，而要购买“低相关失效”。群体需要的不是十个看起来不同的说法，而是在关键错误发生时仍能至少有一条路径拒绝随主流一起失败。这个原则把模型选择从功能排行榜问题转化为系统可靠性问题，也使认知响应多样性具有了可操作的工程含义。

表面共识与真实共识：AI 时代为什么必须给“多数”打独立性折扣？

在传统群体决策中，多数意见之所以具有信息价值，一部分原因在于不同个体拥有不同私有信号。若 100 个人分别观察世界后有 80 个人得出同一结论，这个 80%通常比一个人的判断更有分量。但如果 100 个人先看了同一个 AI 答案，然后 80 个人表示赞同，80%并不具有同样的信息含量。两种情形的“票数”相同，生成过程完全不同。

Lorenz 等人的经典实验表明，社会影响可以降低群体判断的多样性，却不一定改善集体误差，并可能使参与者对收敛后的答案更加有信心[4]。AI 让这一问题具有新的强度：一个模型不是普通同伴意见，而是一种极其流畅、覆盖广、可即时解释的高权威锚点。它能够在群体成员形成第一判断之前就提供一个完整叙事，从而改变后续意见的相关结构。

因此，在研究综述、政策咨询、投资委员会或医疗会诊中，仅统计“有多少人同意”越来越不够。更重要的是记录每个判断的证据谱系：哪些成员在看 AI 答案前已有独立判断？哪些成员调用了不同模型？哪些结论基于共同数据库？哪些异议真正来自独立来源？这可以称为“共识谱系审计”。

本章提出一个原则：共识强度应按证据独立性折扣，而不是按人数原值计价。 当多个判断共享同一来源、模型或推理模板时，应把它们视为一个相关簇；真正增加集体信心的是来自不同簇的一致，而不是同一簇内部的重复。这个原则与上一篇“最小充分验证路径”中的独立第二路径可以自然衔接：对高后果结论，至少需要一个不属于主导认知簇的验证路径。

为什么 AI 会使异议更难出现：从“异议成本”到“反常识税”

共同模型不仅提高主流答案的生成效率，也可能提高异议的相对成本。过去，大家都要花三小时查资料，一个异议者多花一小时寻找反例并不算太离群；当 AI 把主流答案压缩到三分钟后，仍坚持从原始资料重建问题的人，成本可能高出十倍。随着组织越来越追求速度，这种人很容易被评价为低效、固执或“不会用 AI”。于是，异源性不是被禁止，而是在成本结构中自然退出。

这种现象可以称为“反常识税”：当主流认知路径由高性能 AI 大幅降价，任何拒绝共同锚点、保留独立路径的人都要支付额外时间、技能与社会解释成本。若组织没有为这种成本设置预算，异议最终只会保留在那些个人动机极强的人身上，难以成为稳定系统能力。

文化进化论中的成功偏置进一步放大这种成本差。主流 AI 工作流不断产生可见成功案例，少数路径则只有在主流出错时才显示价值；但低频重大错误本来就很少出现，

因此少数路径的贡献在日常绩效指标中经常不可见。组织会倾向于削减“平时看起来没有产出”的冗余，就像把备用系统视为闲置成本。

生态学的视角恰好提醒我们：韧性资源的价值往往只在扰动到来时显现。一个耐旱物种在正常降水年份可能并不高产，却在极端干旱中维持系统功能；一个坚持使用原始数据和不同模型的审查者，在 99 次普通任务中可能只是慢一点，却可能在第 100 次共同模型失误时成为唯一的纠偏入口。

因此，保护认知异源性不能只靠“欢迎不同意见”的文化口号，还需要在绩效制度中承认冗余价值：为独立核验预留时间，为少数报告设置正式入口，让“发现主流错误”成为可计量贡献，而不是把所有人都按平均产出速度评价。

什么时候趋同是好事，什么时候趋同变成危险？

并非所有认知趋同都值得担心。科学进步本来就包含对更优理论的收敛，工程标准化可以显著降低接口错误，统一法律文本能减少歧义，成熟软件库也比每个人重复造轮子更可靠。问题不是“大家相同”，而是“为什么相同”和“相同之后是否还有逃生路径”。

本章给出四个判断条件。第一，证据可验证性：如果结论可以被低成本、客观测试，例如程序是否通过测试、计算是否重算一致，那么趋同风险较低；如果结论涉及不可观察因果、未来预测或稀缺信息，相关错误更难暴露。第二，错误后果：低后果任务可以用效率优先，高后果任务需要更高 CRD。第三，环境稳定性：在重复、稳定、分布内任务中，统一最优流程通常有效；在快速变化、分布外或对抗环境中，异响应更重要。第四，可逆性：错误可快速回滚时可以更集中；不可逆决策则需要更多独立路径。

由此可以形成一个“趋同—异源性调节原则”：常规、可验证、低后果、可逆的任务允许高度 AI 标准化；新颖、难验证、高后果、不可逆的任务必须提高异源性预算。不是给所有任务强制多模型、多专家，而是把 CRD 当作一种按风险配置的资源。

这也是生态学与工程思想的共同点：系统并不需要在所有部件上无限多样，而要在共同失效最危险的关键功能上保留足够独立冗余。认知治理同样应把多样性资源集中在承重判断、关键预测、事实不确定区和价值冲突区。

四类 AI 协作场景：认知单一化风险如何具体出现

14.1 研究与学术写作。若大量研究者用同一模型做文献综述、生成研究问题和解释结果，平均写作结构可能显著改善，但“哪些文献先被看到、哪些解释最自然、哪些反例值得追”的入口也更可能趋同。最危险的不是句式相似，而是同一篇关键论文被错误

概括后，被不同作者从 AI 输出中反复引用，最终形成虚假的二级共识。对策不是禁止 AI 写综述，而是要求关键结论至少有一条原始文献路径、一个对立检索式和一个非同源数据库。

14.2 编程与软件工程。高性能代码模型可以让更多人写出可运行代码，但如果团队大量依赖同一模型生成权限控制、异常处理或数据库操作，常见漏洞模式可能跨项目复制。十名工程师分别让同一模型“审查安全问题”也不等于十次独立安全审计。高后果模块应保留不同静态分析器、不同测试策略、人工威胁建模或独立安全团队，目的就是降低共同失效。

14.3 法律与政策分析。法律任务高度依赖法域、时间、事实前提与例外。模型若对一个最新修法理解错误，多个分析师共享同一模型后可能同步给出措辞完美但适用错误的意见。随后这些意见在组织内部互相引用，会制造“我们多个部门都得出一致结论”的外观。这里最重要的是共识谱系：一致究竟来自多次独立法源核对，还是来自同一个 AI 摘要。

14.4 商业与战略决策。AI 可以快速生成 SWOT、市场规模、竞争分析和情景预测，减少团队的信息整理成本；但如果所有团队在会议前都看到同一份 AI 总结，独立判断在开会前已经被锚定。真正的“十人决策”可能退化为“一份 AI 先验+十个相关修订”。对于重大决策，可以要求关键参与者先提交盲独立判断，再开放 AI 综合；或者把团队分为使用不同数据/模型的两组，最终再对冲差异。

实证证据已经出现到什么程度：我们知道什么，又不能过度推出什么？

本章的理论并非建立在“AI 一定让所有人变得相似”的假设上。当前经验研究呈现的是条件性结果。Doshi 与 Hauser 发现，生成式 AI 可以提高个体创造性评价，同时降低群体故事之间的多样性[8]；Hosanagar 与 Ahn 的实验研究也观察到 AI 参与程度提高时，聚合层面的内容多样性可能下降，而保留人类在早期创意任务中的参与可以缓解这种趋势[34]。另一方面，Ashkinaze 等人的动态实验发现，高强度暴露于 AI 创意在其任务设计中反而提高了集体创意多样性[14]。这说明“AI→单一化”不是机械定律，而取决于模型采样、任务、交互设计和信息网络。

这种不一致反而强化了本章的观点：我们不应只测最终文本相似度，而应测“失效相关性”和“响应多样性”。两个故事可以语义不同，却在事实问题上共享同一错误；两个法律意见可以措辞相似，却分别经过独立法源验证，因此并不存在同等风险。表面内容多样性只能回答“看起来有多不同”，无法直接回答“遇到共同扰动会不会一起错”。

Kim 等人对 LLM 错误相关性的研究为共同失效提供了更直接证据[10]，但它仍主要测量模型之间，而不是“人+AI”群体。未来真正关键的实验应比较四种群体：纯人独立判断、共享同一 AI、使用不同 AI、使用同一 AI 但强制独立资料源；同时记录平均准确率、错误重合度、少数正确者比例、共识形成速度与错误纠正时间。只有这样，才能直接检验 CMP。

同时，Burton 等人关于 LLM 与集体智能的综述指出，LLM 可能同时改变信息获取、协调、偏差与群体多样性，影响方向取决于具体制度设计[9]。因此本章的定位是“机制理论+可检验预测”，而不是把尚未充分测量的社会长期效应写成既成事实。

六条可检验命题：如何把“认知单一化”从隐喻变成研究对象

命题一：平均绩效—相关性分离命题。 在 AI 辅助条件下，个体平均准确率可以上升，同时个体错误指示之间的相关系数 ρ 上升；两者在经验上应被分开估计。若只测平均分，会漏掉系统性共同失效的变化。

命题二：有效独立数下降命题。 在参与人数 n 固定时，随着 ρ 上升，N_eff 下降；当 N_eff 显著下降时，多数投票、重复复核和多代理一致性所带来的置信增益应按相关性折扣。

命题三：成功偏置加速命题。 AI 在常规任务上带来的可见绩效优势越大，个体越倾向复制其工具和工作流；如果模型生态高度集中，这种复制会比低性能工具更快压缩来源与方法异源性。

命题四：共识伪独立命题。 当 AI 生成内容被大量复制进入社会信息环境后，后来者会把同源重复误判为多源支持；若向判断者披露信息谱系，主流答案的可信度评价将发生系统变化。

命题五：响应多样性—尾部韧性命题。 在常规任务上，低 CRD 群体可能与高 CRD 群体表现相当甚至更高效；在分布外、规则突变或共同模型盲点任务上，高 CRD 群体应具有更高的至少一人正确概率、更短的纠错时间和更低的错误传播范围。

命题六：独立优先命题。 “先独立判断后看 AI/他人答案”的流程，相比“先看统一 AI 答案后分别复核”，应产生更低的错误相关性和更高的异议发现率，即使最终平均准确率差异不大。

治理框架一：给共识做“独立性会计”，而不是只数赞成票

第一个治理工具是“共识独立性账本”。对高风险结论，每个支持判断至少记录三个标签：主要模型/工具、主要证据源、主要方法。若多个判断共享同一三元组，就把

它们视为一个认知簇，而不是多份完全独立证据。组织不必为所有日常任务做这件事，但在重大投资、科研结论、法律意见、安全部署等场景中，这一成本是可接受的。

第二个工具是“独立性折扣后的共识强度”。例如委员会十人中九人支持某方案，但其中八人都基于同一个 AI 生成的市场分析，只有一人使用独立数据源，那么不能简单写“9/10 专家一致”。更准确的表述应该说明共识的来源结构。这样做并非贬低 AI，而是恢复证据计数的基本诚实：复制不能被当成重复观察。

第三个工具是对 AI 多代理系统应用同样标准。多个代理只要共享同一基础模型、上下文与检索源，就可能具有强共因失效。多代理讨论可以提高搜索和推理覆盖，但“几个代理都同意”不能自动获得与几个独立专家同等的置信增益。至少一个代理需要在数据、模型家族、工具或验证方法上真正异构。

治理框架二：异源性预算——不是人人都独立重做，而是在关键处买一条不同的路

保护认知异源性的最大现实障碍是成本。如果要求每个人在使用 AI 后都完全独立重做一次任务，效率红利会被抵消。因此本章提出“异源性预算”（Heterogeneity Budget）：组织不追求全流程人人异构，而是在共同失效后果最高的节点，预先分配少量资源维持独立路径。

异源性预算可以有四种支出方式。第一，模型预算：关键判断至少使用一个不同模型家族或不依赖 LLM 的工具；第二，资料源预算：承重事实至少从一个独立原始源核对；第三，方法预算：核心结论至少用一种不同于生成路径的方法验证，例如 AI 做定性综述，人用数据复算；第四，人员预算：保留至少一个不先看主流答案的审查者。

这种设计与“最小充分验证”思想一致：不是让所有地方都多样，而是找到结论翻转集与共同故障点，在那里买独立性。对低后果内容，单模型高效生成完全合理；对关键节点，则必须把一部分节省下来的认知劳动重新投资到失效去相关上。

可以把异源性预算理解为 AI 效率红利的一部分保险费。一个组织若因为 AI 把原来十小时的任务降到两小时，未必要把节省的八小时全部转化为更多产量；在高风险任务上拿出其中一小时建立独立路径，仍然拥有巨大净效率增益，却显著提高系统韧性。

异源性预算与本书第八章的最小充分验证共用一条原则，此处需要说明它们的分工，以免读者以为是同一件事重复了两遍。

两章共同承认：同源的重复不构成独立的证据。差别在使用方向。第八章面对的是一份已经生成的输出，问题是"用什么样的验证动作，以最低成本压缩掉那些会翻转结论的风险"；本章面对的是一群将要形成共识的判断者，问题是"如何在共同失效发生之

前，保住至少一条不会一起错的路径"。前者是事后的风险压缩，后者是事前的多样性保全。

这个分工有一个实践后果：第八章的验证可以在结论产生之后补做，本章的异源性不能。一旦所有人都已经看过同一份 AI 生成的分析，异源性就无法通过事后增加检查来恢复——这正是本章"先独立，后汇聚"之所以是顺序问题而非努力问题的原因。因此在高后果任务上，本章的措施必须先于第八章的措施部署；两者的次序不可互换。

治理框架三：保护少数路径——让“平时看起来多余”的人和方法在关键时刻仍然存在

生态系统中的稀有功能与响应类型在平常时期容易被低估，认知系统亦然。组织如果只按平均效率优化，很可能逐步淘汰手工检索、独立推导、小众数据库、非主流模型和不同专业视角。短期看，这是标准化；长期看，却可能把所有纠错能力集中到同一基础设施。

因此，高风险组织需要“少数路径保护”。这不是保护任何特定反对意见，而是保护至少一条不依赖主导路径的能力。例如：研究团队保留能直接读原始文献的人，软件团队保留独立安全工具链，法务团队保留权威法源检索，投资团队保留不看市场共识先写反方备忘录的角色。

少数路径必须被制度化，否则它会在日常绩效压力下自然消失。可以采用轮值红队、盲独立预判、双轨检索、异构工具强制项、定期“断 AI 演练”等方式，使替代路径保持可用。所谓断 AI 演练，不是反技术，而是测试：当共同模型不可用或明显失效时，团队是否仍有能力恢复关键判断。

Shirado 与 Christakis 的网络实验甚至发现，在某些协调任务中，少量行为上带有噪声的自主代理反而能帮助人群摆脱局部僵局、提升全局协调[30]；Ueshima 等人的研究也显示，简单自主代理可以增强人类群体的创造性语义发现[29]。这些结果共同提示：系统中的“非一致性”并不总是成本，适度设计的异质扰动可以成为探索与纠偏资源。

治理框架四：“先独立，后汇聚”——改变信息顺序，比要求大家勇敢反对更有效

群体认知单一化有一个高度可操作的控制变量：信息出现的顺序。如果 AI 标准答案在最开始就展示给所有人，后续判断很难真正独立；如果先要求成员形成初始判断，再开放 AI 与群体信息，至少可以保存一轮未被共同锚点污染的信号。

因此，对高风险任务可以采用三阶段流程。第一阶段“盲独立”：关键参与者分别形成简短初判，并记录最关键的证据和不确定点；第二阶段“AI 扩展”：使用 AI 补充资料、挑战假设、生成反例和备选方案；第三阶段“结构化汇聚”：比较各路径分歧，优先调查那些由不同独立路径产生的冲突，而不是简单投票。

这一流程看似比直接让 AI 先写答案慢，却并不需要每个人完成完整报告。独立阶段可以只要求回答几个承重问题，例如“最可能使本结论失效的三项事实是什么”“你当前首选方案及置信度是多少”。只要这些初始信号在共同锚定之前被保存，就能显著提高后续共识的可解释性。

对多人+多 AI 系统亦可如此。不同代理先在隔离上下文中独立推理，使用差异化工具与资料源，再进入辩论；而不是所有代理共享同一长上下文反复互相引用。独立先行并不能消除共同训练语料带来的相关性，但能减少由直接社会影响和共同锚点造成的额外相关。

风险分级：哪些任务最需要认知响应多样性?

CRD 不应成为所有 AI 使用场景的统一合规负担。本章提出四级风险：L0 为创意探索与可随时撤销的低后果任务，重点追求效率，不要求额外异源性；L1 为普通知识工作，关键事实做来源抽查即可；L2 为重要专业判断，需要至少一条独立资料或方法路径；L3 为高后果、不可逆或系统传播任务，需要结构化异源性预算、共识谱系审计和独立第二路径。

风险升级信号包括：结论将直接触发资金、法律权利、医疗处置、生产系统变更；错误会被自动传播给大量下游用户；问题高度新颖或处于模型知识边界；多个参与者都使用同一 AI；搜索结果高度同质；结论缺少可快速实验验证；或者主流答案的支持主要来自互相引用的二手资料。

相反，如果任务能够被快速客观测试，例如一段本地脚本是否正确解析文件，哪怕所有人都用同一 AI 也未必需要很高 CRD，因为外部执行结果本身就是强独立反馈。认知异源性是为“没有便宜真值反馈”的场景准备的昂贵资源。

对 AI 产品设计的启示：不要只优化“给出更好的一个答案”，还要优化“保留不同的可检验路径”

当前许多 AI 产品的默认目标是给用户一个更完整、更一致、更自信的答案。从单用户体验看，这非常合理；从群体韧性看，却可能产生过度中心化。未来产品可以增

加“异源模式”：在高风险问题上，不只是给三种措辞，而是明确调用不同来源、不同方法或不同模型角色，展示哪些部分共享证据、哪些是真正独立得到。

第二，产品可以显示“证据谱系”。当一个回答引用多个网页时，应尽量帮助用户识别这些网页是否都转载同一个原始来源；当多个代理都同意时，应显示它们是否共享同一基础模型与检索结果。这样，用户才能区别“多次重复”与“多源一致”。

第三，产品可以支持“盲独立回答”。在团队场景中，系统先私下收集成员初判，再统一展示 AI 建议和同伴意见，降低早期锚定。第四，系统可以在检测到高模型集中度时提示“当前结论主要依赖单一认知簇”，并建议添加独立资料源或第二方法。

第五，对模型评测也应增加“生态指标”。除了单模型准确率、幻觉率和基准得分，可以测量不同模型之间的错误相关矩阵、模型簇结构、共同盲点以及多模型组合的有效独立数。一个稍弱但错误模式高度互补的模型，在生态系统中可能比又一个高度同质的强模型更有公共价值。

对教育的启示：AI 时代真正不可外包的，不只是“会不会独立做题”，而是“能不能生成异源判断”

如果教育仍把“独立完成所有认知劳动”作为理解的唯一证明，它会错过 AI 带来的真实能力扩展；但如果教育只训练学生会向 AI 提问和整合答案，又可能让一代人的推理路径高度同质。更合适的目标，是培养学生在使用 AI 的同时保留生成独立信号的能力。

例如，同一个问题可以要求学生先给出自己的最小判断，再让 AI 回答，随后比较差异；要求学生寻找一个足以推翻 AI 结论的反例；要求两组学生使用不同资料源或不同模型，再分析为什么会收敛或分歧；要求标注“哪些判断来自 AI，哪些来自原始证据，哪些来自个人经验”。这些训练不是怀旧式拒绝 AI，而是在培养 CRD。

真正重要的能力因此从“我能不能不借助工具完成一切”转向“当所有人都借助同一强工具时，我还能不能看见它没有看见的东西”。这与 AI 时代的理解标准自然连接：能够识别适用边界、发现关键缺口、改变前提后重组解释，本质上也是一种认知响应多样性。

对公共知识与平台治理的启示：防止“同源复制”伪装成“社会共识”

公共信息环境中的一个新挑战，是内容数量与独立来源数量逐渐脱钩。生成式 AI 可以让同一源头在短时间内产生大量改写版本；如果平台、搜索或读者只按网页数量、提及次数和文本表面差异理解“共识”，就可能把复制强度误当成证据强度。

因此，未来的信息质量治理需要更多“来源聚类”而非简单计数。对事实性议题，可以尝试识别多篇内容是否追溯到同一原始数据、同一新闻稿、同一论文或同一 AI 生成链。对社会共识的展示，也应尽量区分“独立机构分别验证”与“多个站点转述同一来源”。

这并非 AI 独有的问题，传统新闻转载和信息瀑布早已存在；AI 的变化在于复制成本更低、改写能力更强，表面文本差异更大，因此“同源性”更难肉眼识别。共识谱系由此从学术规范问题上升为公共认知基础设施问题。

理论边界：多样性本身也会失败，异源性不是新的万能答案

第一，过度多样会降低协调效率。Baumann、Czaplicka 与 Rahwan 的模型研究指出，多样性对集体学习的效果取决于任务复杂度与网络结构；在简单任务中，多样性甚至可能降低表现，而在复杂任务和特定网络条件下才显示优势[7]。因此，本章反对把“多样”道德化。CRD 是一种风险资源，不是越高越好。

第二，独立路径也可能共同受现实世界同一个偏差影响。两个模型来自不同公司，不代表训练语料真正独立；两个数据库也可能都转录同一个错误源。所谓异源性必须追溯到足以改变失效机制的层级，而不是换皮。

第三，少数意见不天然正确。文化进化中的反从众同样可能维持有害或低质量变体[22]。保护少数路径的目的不是赋予少数结论更高真理地位，而是保留一种“在多数共同失效时还有别的传感器”的结构能力。少数路径也必须接受证据和结果检验。

第四，AI 也可以主动增加多样性，而不是只减少多样性。Wan 与 Kalman 通过设置不同 AI“人格/角色”生成差异化创意，在其实验中恢复了群体创作的多样性[13]；Ashkinaze 等人的动态实验也显示，特定条件下 AI 暴露能够增加创意多样性[14]。这意味着真正的问题是设计：单一默认模型会不会成为唯一认知入口，以及系统是否有意识地制造独立的探索路径。

第五，本章的 N_eff 公式采用等相关近似，只用于概念说明，现实群体的错误相关矩阵往往高度异质。更严谨的测量应使用完整协方差结构、聚类相关性或任务级共同错误图谱。不能把一个简化指标误当成精确社会测量。

未来研究议程：从“文本相似度”走向“共同失效科学”

第一，应建立人—AI 群体的错误相关基准。现有评测主要比较单模型准确率，未来需要报告“哪些题大家一起错”，并区分同模型、跨模型、人+模型、不同资料源组合。错误相关矩阵可能成为比排行榜更重要的生态指标。

第二，应开发 CRD 测量。可以设计一组受控扰动：改变关键事实、撤掉某一证据、引入分布外案例、替换法规版本、制造数据异常，观察不同判断者是否产生不同的报警与纠错动作。CRD 应测“响应函数差异”，而不是最终答案词向量距离。

第三，应研究共识谱系对信任的影响。把“十个独立来源支持”与“十个同源 AI 改写支持”随机展示给参与者，看披露来源依赖后其置信度如何变化，可以直接检验共识伪独立效应。

第四，应研究异源性预算的最优分配。在给定时间与成本下，究竟把预算用于第二模型、原始资料、人工专家、红队还是随机抽样，哪一种最能降低高后果共同错误？这可以与上一篇 MSVP 的风险压缩率研究结合，形成 AI 验证的资源配置科学。

第五，应研究长期文化反馈。短期实验通常只观察一次 AI 辅助，而真正的文化进化发生在多轮传递中。需要让参与者的 AI 辅助产出成为下一轮参与者的输入，观察变异、错误与共识如何跨代传播，并比较有无谱系标记、反从众角色和异构模型时的动力学差异。

第六，应研究“能力提升与异源性衰减”的临界点。如果更强模型持续降低 p，什么时候它的能力红利足以抵消 ρ 上升？又在什么任务上 ρ 的尾部风险会占主导？这一问题需要任务风险分布、错误后果和相关结构共同进入模型，而不能只比较平均分。

结论：AI 时代要保护的不是“大家意见不同”，而是“大家不能只剩一种失败方式”

生成式 AI 带来的一个巨大文明收益，是把高质量认知能力以前所未有的低成本分发给更多人。它可以提升普通人的写作、研究、编程和分析水平，也可以缩小个体之间的技能差距。本章并不把这种趋同理解为退步。真正需要警惕的是另一种更隐蔽的趋同：不同个体虽然都变得更强，却越来越依赖同一个基础模型、同一批信息源和相似的推理模板，于是剩余错误从“各错各的”转变为“偶尔一起错”。

统计学习理论告诉我们，平均准确率与错误相关性是两个不同变量；文化进化论告诉我们，高性能工具恰恰会因为成功而被更快复制，并把频率转化为可信线索；生态学告诉我们，承担相同功能的多个高性能单元，如果对扰动缺少响应多样性，仍然可能组成一个脆弱系统。三者碰撞后，认知单一化悖论就不再只是“AI 会让文字千篇一律”的审美担忧，而成为一个可形式化、可测量、可治理的集体认知问题。

因此，AI 时代群体智慧的新公式不应是“更多人 + 更强 AI = 更聪明的群体”，而应增加一个被长期忽略的条件：这些人和 AI 之间必须保留足够的失效去相关性。十个

人的价值不只是十倍算力，而是十种可能发现不同错误的路径；一旦这些路径被压成同一个认知簇，人数仍在，群体的有效独立性却已经下降。

真正成熟的人机协作制度应同时追求两件事：让多数人借助最强工具快速提高平均表现，同时在关键节点制度性地保存少数异源路径。我们可以让 90%的日常工作高度标准化，但必须知道剩下 10%的独立性放在哪里；可以让 AI 先完成绝大多数搜索和生成，但必须保留一个不同资料源、不同模型、不同方法或不同责任主体，能够在主导路径失效时发出不一样的信号。

最终，群体认知韧性不是靠人人永远争论获得的，而是靠一种更精细的结构：平时可以快速收敛，关键时刻能够重新分叉；多数路径可以共享高性能 AI，系统仍保留不会与它同时失灵的替代路径；共识可以形成，但必须知道它究竟来自多少独立证据。AI 越强，这种能力越重要。因为一个低质量工具的错误通常容易被看见，而一个高质量、被广泛信任的工具所留下的少数共同错误，恰恰最可能穿过整个群体而不被发现。

理论贡献与跨学科意义

本章的第一项理论贡献，是把“AI 造成内容同质化”的讨论推进到“错误相关结构”层面。文本相似本身未必危险，真正决定群体纠错能力的是同一个错误是否会同时穿透多个判断者。由此，研究对象从内容表面多样性转向失效函数多样性。

第二项贡献，是通过 N_eff 把统计学习中的相关性问题翻译为群体认知可理解的指标：十个人不一定等于十份独立证据。这个概念为“多数意见”“多模型一致”“多人复核”提供了一个统一的独立性折扣语言。

第三项贡献，是从生态学引入并重构“响应多样性”，提出认知响应多样性 CRD。它比一般的观点多元更严格，因为它要求承担同一认知功能的主体，在面对环境变化和共同盲点时能够产生不同检测与纠错反应。

第四项贡献，是把文化进化论置于错误相关性的动态生成过程中。共享 AI 不是静态技术选择；高性能会通过成功偏置增加采用率，采用率又通过从众与声望线索转化为更强社会可信度，最终形成相关错误放大回路。这样，模型层的偏差与文化层的路径依赖被连接起来。

第五项贡献，是提出一组不以“拒绝 AI”为前提的治理工具：共识独立性账本、异源性预算、少数路径保护、先独立后汇聚以及风险分级。它们的共同原则不是让每个人重新独立完成全部任务，而是在共同失效最危险的地方保留一条不同的路。

本章附表

维度	低异源性表现	高异源性表现	主要治理问题
来源	多人引用同一摘要/数据库	至少一个原始或独立资料源	来源谱系是否独立
模型	同一基础模型多代理自洽	不同模型家族或非 LLM 工具	共同训练/架构盲点
方法	全部依赖语言推理	计算、实验、反例、现场证据并存	是否有不同失效敏感性
流程	先看统一答案再复核	先盲独立再汇聚	锚定与社会影响
人员	同专业、同激励、同角色	跨专业/红队/少数路径	异议成本与角色保护
传播	同源内容被重复计为共识	标注谱系并做独立性折扣	可见共识是否等于独立证据

统计学习理论	文化进化论	生态学	AI 时代的集体认知对应
单个模型准确度	个体学习成功率	单个物种/功能单元表现	个体 AI 辅助绩效
模型错误相关/协方差	共同来源与高保真复制	共同敏感性/共同失效	不同人的 AI 错误相关性
集成多样性	文化变异与替代路径	响应多样性	认知响应多样性（CRD）
集成收益被相关性吞噬	从众与成功偏置使路径锁定	单一化降低韧性	人数增加但有效独立判断数下降
异构模型/去相关	维持探索与反从众通道	功能冗余+异响应	异源性预算与独立第二路径
总体损失/尾部风险	错误的社会放大与路径依赖	扰动后的恢复能力	共同错误的发现、纠正与恢复速度

风险级别	典型任务	CRD 要求	建议机制
L0 低风险	创意草稿、可撤销日常文本	低	单模型即可，事后快速检查
L1 一般	普通研究整理、内部分析	中低	关键事实独立来源抽查
L2 重要	专业意见、重要数据分析、生产代码	中高	独立资料/方法路径 + 盲初判
L3 高风险	重大资金、法律权利、安全、医疗、不可逆部署	高	共识谱系审计 + 异源性预算 + 独立第二路径

第十章

同一条脊梁的九次换位

九章的共同判断、三对最危险的近亲、九个读数的三种形状，以及会让整本书失效的那几件事

九章为什么应当放进一本书

前面九章分别处理九个看上去互不相干的问题：能力归谁、能力怎么维持、帮助要不要撤、绩效为什么骗人、判断属于谁、理解算不算数、责任落在哪里、验证做到什么程度算够、群体为什么会一起错。它们各自动用的三门学科几乎没有重叠——认知神经科学与核设施运维之间、解释学与审计学之间、生态心理学与统计学习理论之间，本来没有对话的必要。

把它们放进一本书，不是因为都谈生成式人工智能。共同的题材不构成一本书，只构成一个专辑。真正让九章成为一本书的，是它们在各自领域内被逼出的同一句判断：

能力、判断与责任并不是先失去，而是先失去可观测性；正常运行的成功本身销毁了证明它们仍在的证据，因此它们只能在代理被有意打断的断面上被重新判定。

这句话在九章里没有一次以相同的措辞出现。第一章把它写成"结果质量不能认证能力"，第二章写成"零事故不能区分备用能力可用与不可用"，第三章写成"辅助态的好表现不再透明地显示这个人本人能做到什么"，第四章写成"能力可能在绩效持续上升时变得不可见"，第五章写成"最后动作在人不等于判断由人形成"，第六章写成"流畅的迁移可能恰恰掩盖了不会撤回"，第七章写成"每个节点都有责任语言，没有节点能改写下一轮"，第八章写成"检查覆盖率不是充分性"，第九章写成"多数同意不等于多份独立证据"。

九种措辞、九个学科、九个对象，句子的骨架却是同一根：一个本来用来证明 X 还在的信号，在 AI 持续代行之后不再证明 X 还在，而它失效的方式恰恰是它看起来更好了。

本章的任务是把这根骨架显式地摆出来，说明九章之间是什么关系、哪几对最危险、九个读数其实是几种形状，以及在什么条件下整本书应当被放弃。

九次换位

同一条判断在九章里换了九次位置。换的是三样东西：被宣布为"没出事"的那个信号、断口应当开在哪里、以及断口上读什么。

章	被当作"还在"的证据	断口开在哪里	读数

一	任务结果的质量	对代理施加一条可判定的扰动	生差、裁差、改向、回写四环节是否由主体重新闭合
二	长期正常运行、无事故	在真实故障之前主动做一次接管证明	异常发现、情境重建、独立裁决、干预执行、恢复验证五门串联是否全通
三	辅助状态下的当前表现	周期性收回首判、差分、盲段与恢复回合	证据龄比：距上一次有效独立表现证据过去了多久
四	持续上升的绩效曲线	锚定、扰动、撤援、延迟迁移四层测量	可观察绩效对潜在能力的预测力是否下降
五	最终由人签字、点击、批准	扰动、反事实边界、重新发起三类测试	是否持有决定性理由、能否识别失效条件、能否主动重开判断
六	能解释、能举例、能迁移到新情境	表面相似而关键前提已破坏的边界破坏任务	首次边界修订由谁发起：主体自主撤回，还是外部提示之后
七	流程上有明确的责任人	事故前置审计与近失误保留	认知可达性、因果控制杠杆、改写权限三者是否落在同一主体身上
八	已经检查了很大比例的内容	用与原生成路径失效模式低相关的程序去攻击承重节点	残余未核错误是否还足以翻转结论
九	多数人独立得出了同一结论	分布外、规则突变与共同盲点任务	有效独立判断数 $N_{eff} = n / [1 + (n-1)\rho]$，以及至少一人正确的概率

这张表是全书的索引，也是全书的论证。读者可以从任何一行进入：每一行的第二列都是一个在自己领域里被广泛承认的良好信号，第三列都是一次有意制造的中断，第四列都是只有在中断之后才存在的量。

三列之间的关系不是并列而是递推。正因为第二列的信号在 AI 代行条件下失去了分辨力，第三列的中断才必须由人主动安排；也正因为中断是被安排的，第四列的读数才可能存在。 如果第二列的信号仍然有效，全书九章都没有必要。

三个共绕的动作

把九章的判断按动作分类，落在三个族上，而且九章同时落在三个族的交叉点。

第一个动作是遮蔽。 九章无一例外地描述了同一件事：成功的运行掩盖了证据。第二章的零事故掩盖了备用通道的潜伏失效，第四章的绩效曲线掩盖了潜在能力的变化，第六章的流畅解释掩盖了越界，第七章的无事故运行掩盖了责任结构的空洞，第九章的

表面共识掩盖了同源。值得注意的是遮蔽的方向：不是失败被掩盖，而是成功本身成了掩盖物。这与传统的隐瞒、舞弊、指标造假是不同的机制——没有人隐瞒任何东西，信息也没有被销毁，只是产生这类信息的场合不再出现了。

第二个动作是错置。 九章都指出了一次证据类型的偷换：把正确答案当成能力的证据（第一章），把无事故当成可用的证据（第二章），把当前表现当成本人会的证据（第三章），把产出当成能力的证据（第四章），把末端动作当成判断的证据（第五章），把迁移当成理解的证据（第六章），把签字当成责任的证据（第七章），把检查覆盖率当成充分性的证据（第八章），把多数同意当成独立性的证据（第九章）。九次错置有一个共同形状：被偷换掉的那一项，全都是"过程属于谁"，换上来的全都是"结果长什么样"。 AI 之所以让这种错置变得危险，正是因为它极大地提高了结果的质量，同时极大地改变了过程的归属，而制度只盯着前者。

第三个动作是断口。 九章给出的解法在形式上完全一致：不去问"它现在怎么样"，而去制造一个中断，问"它在中断处怎么样"。第一章的定向代理扰动、第二章的接管证明、第三章的盲段、第四章的撤援、第五章的重新发起测试、第六章的边界破坏、第七章的近失误、第八章的判别性检验、第九章的独立性折扣，都是同一个动作的九种工程形式。

三个动作合起来，就是全书的方法论：遮蔽解释了为什么看不见，错置解释了为什么以为看得见，断口是唯一能重新看见的地方。

还有一个次一级的动作族值得单列，因为它决定了这本书是不是只在描述一次性事故：自强化。第四章的绩能遮蔽循环、第七章的末端责任压缩—改写责任空位循环、第九章的相关错误放大回路，以及第二章那个"AI 越可靠，人获得自然接管练习的机会越少"的悖论，都说明遮蔽不是静态的。看不见会使决策者更倾向继续外包，继续外包又使更少的证据被产生。 这条正反馈是本书判断"这不是一次误判，而是一种会自己长大的结构"的依据，也是第十一章要处理的问题：一个会自己长大的结构，多半也会长出应付检查的办法。

三对最危险的近亲

一本由九章构成的书，最容易受到的攻击不是"这些说法不对"，而是"这九章其实是同一章"。本节主动把最危险的三对摆出来，并给出可判定的合并条件。凡满足合并条件的，本书应当合并相应章节，而不是辩解。

第三章与第四章是最可能被合并的一对。 两章都说绩效与能力之间的连接被 AI 削弱了：第三章称之为能力证据债，第四章称之为能力可见性塌缩。两者的形式陈述几乎

可以互相翻译——第四章的"可观察绩效对潜在能力的预测力下降"正是第三章"辅助态表现不再显示本人能做到什么"的计量版本。分界线只有一条：第三章不主张能力真的发生了变化，它只主张系统失去了判断依据；第四章则给出了一个能力本身与能力可见性同时下降的动力学模型。 前者是认识论主张，后者是因果主张。判定方法：设计一组实验，让辅助态绩效与独立能力的相关下降，同时用锚定任务证明独立能力并未下降。若这种情形在经验上根本不出现——即相关下降总是伴随能力下降——那么第三章的"债"就只是第四章"塌缩"的早期阶段，两章应当合并，本书降为十章。 本书事先声明：这是九章里最可能被合并的一处。

第一章与第五章是第二对。 两者都在处理归属：一个问能力归谁，一个问判断归谁；一个用四环节，一个用三节点。四环节的"生差、裁差、改向、回写"与三节点的"理由持有、失效识别、重新发起"在测试设计上高度重叠——两章都用扰动，都看主体能否主动发起。分界在于对象的性质：能力是可以长期不用而仍然存在的，判断则是每一次都要重新形成的。 因此第一章允许深度外包只要断面上能复闭环，第五章却要求在每一次被声称为"人做的决定"里都保有理由。合并条件：若在同一批受试者身上，能力归属测试与判断归属测试的通过与否高度一致（相关高于 0.8），则两者不是两个维度，第五章应并入第一章作为一节。

第二章与第三章是第三对。 两章都在做维护设计，一个叫最低接管维持剂量，一个叫能力校准回路；一个的循环是"拆门—微调用—待命暴露累计—风险触发证明—扰动接管—恢复回写"，另一个是"首判—协作—差分—盲段—恢复"。这两条循环的相似程度足以让读者怀疑是同一条。分界在目的：第二章的循环是为了让能力保持可用，第三章的循环是为了让能力保持可见。 前者若成功，人的接管可靠度提高；后者若成功，系统对人的能力状态的判断精度提高，而人的能力可能一点没变。合并条件：若按第三章的校准回路执行一年，接管可靠度的提升与按第二章的剂量向量执行没有差别，则维持与校准不是两件事，第二章应并入第三章。

另有两对次危险的，一并写明：第六章与第一章（理解与能力，第六章已在其末节给出分界：一个测代理坏掉时能否闭合，一个测代理还好用时能否停止）；第八章与第九章（验证与群体，第八章的"证据独立度"与第九章的"独立性折扣"是同一个量在个体任务与群体判断上的两次使用，若两者能被同一个形式量统一，两章的独立性将显著削弱）。

九个读数是三型，不是九型

第二节那张表的第四列列出了九个读数。若九个读数彼此独立，这本书就有九个互不相干的测量传统，读者无从下手。实际情况要简单得多：九个读数只有三种形状。

第一型是串联门型，占五章。第一章的四环节、第二章的五门、第五章的三节点、第六章的撤回与否、第七章的 ECR 三要素，都是若干个二值判定，且都明确规定不许取平均——五门中有一门不过，整条接管路径就不通；ECR 缺一，改写责任就空着。这一型的共同前提是：能力的失效是路径性的，不是程度性的。一条路上任何一处断了，整条路就不通，而把各段完好程度加权平均会得到一个漂亮但无意义的分数。

第二型是相关与折扣型，占三章。第三章的证据龄比、第四章的预测力下降、第九章的 N_eff 与 ρ，都是连续量，都表达"名义上的数量要按某种相关性打折"。第九章打的是判断之间的折扣，第三章打的是时间上的折扣，第四章打的是测量通道上的折扣。三者可以写成同一个形式：有效证据量 = 名义证据量 × (1 − 冗余度)。这一型的共同前提是：证据可以累加，但同源的证据不能重复计数。

第三型只有一章：成本效率型。第八章的 η(v) = ΔR(v)/c(v) 是全书唯一一个把验证动作按"单位成本压缩多少结论相关风险"排序的量。它之所以孤立，是因为只有第八章的对象是可以选择做多少的——其余八章的对象是"有没有"，第八章的对象是"做到哪里为止"。

这个归类有一个直接的实践后果，也是本书给组织的最短建议：不要为九章各造一套指标。串联门型的五章可以共用同一张门卡，只需替换门的名目；相关折扣型的三章可以共用同一个冗余度估计程序；成本效率型只在需要决定投入多少验证资源时启用。九章，三种量表，一套记录格式。

同时必须承认这个归类暴露的一处缺陷：九个读数没有一个是分布量。全部九个读数都以事件或个人为单位取值，没有任何一个回答"在一个一百人的组织里，合格的那些事件是不是集中在少数人身上"。一个团队完全可能九项全部合格，而合格全部来自三个人。本书对此没有答案，只能要求凡使用本书方法的组织同时报告参与率与事件集中度。这是全书自设的第一个盲区。

与最近的邻居划界

本书九章各自与自己领域的近邻做了划界。这里处理的是另一件事：与在同一时期、同一题域、用同一类方法工作的其他研究者划界。最近的一支来自对"AI 时代两极分化"的研究，其核心判断是：在人机耦合状态下，不同个体的当下产出差距会收敛，而一旦脱钩，能力差距反而扩大；由此提出耦合可逆性这一分化轴，并进一步描述一种没有能力承载的成功状态。

这一支与本书第四章距离最近，因为两者都在说"辅助态的表现掩盖了个体之间的真实差别"。分界有三处，都可判定：

第一，变量不同。那一支关心的是个体之间的差距如何变化，是一个分布量；本书第四章关心的是绩效对能力的指示能力如何变化，是一个信号量。两者可以正交：一个群体的耦合态差距可以完全不收敛，而每个人的绩效对其能力的指示力都在下降。

第二，断口的用途不同。那一支把脱钩当作分化的显现时刻，是一个描述性的分水岭；本书把断口当作必须被主动、周期性安排的测量装置，是一个规程。前者研究脱钩发生时会看到什么，后者主张不能等脱钩自然发生。

第三，本书恰好补上了那一支的一个空位。若耦合可逆性确实是一条分化轴，那么就需要一种在脱钩发生之前测量它的方法——本书第三章的证据龄比与第四章的四层测量正是这样的方法。反过来，那一支给本书提供了一个本书没有的东西：分布视角，正是上一节承认的那个盲区。

因此这两支不是竞争关系而是互补关系，且互补的方向是明确的：本书提供断口上的读数，那一支提供读数在人群中的分布。 若将来的研究表明，脱钩后的能力差距扩大完全可以由本书九个读数的差异预测，则本书应当承认那一支是本书的推论而非独立发现；若表明差距扩大主要来自本书读数之外的因素（例如资源、机会与既往积累），则本书应当承认自己只覆盖了分化机制的一半。

什么会让这本书失效

九章各自写了自己的证伪条件。本节给出书一级的三条撤回条款——不是某一章错，而是整本书的判断不成立。

第一条：如果正常运行的信号并没有失去分辨力。 本书全部论证的起点是"良好的日常绩效不再能证明能力、判断与责任仍在位"。这是一个经验命题，可以被推翻。推翻的方式是：在若干个 AI 深度介入的真实场景中，长期收集日常绩效指标与断口读数，若日常绩效对断口读数具有稳定且足够高的预测力（例如跨场景相关持续高于 0.8），那么断口就是多余的，本书的全部工程建议都应当撤回，只保留一句提醒。

第二条：如果断口读数本身不能预测真实失效。 本书要求组织付出成本去制造中断，其正当性完全依赖于断口读数能预测出事时会发生什么。如果在真实的 AI 失效事件中，事前断口读数良好的个体与读数糟糕的个体，在异常发现速度、恢复时间与损失规模上没有系统差别，那么本书就只是发明了一种昂贵而无效的考试。这一条最难检验，因为真实失效事件稀少；本书接受用高保真模拟作为替代，但同时承认这是一个减分的替代。

第三条：如果这套方法在被制度化之后，其最省事的达标方式恰好摧毁它要保护的东西。 这一条与前两条性质不同——它不问理论是否为真，而问理论为真时会不会自我

毁灭。如果组织采用本书方法之后，普遍出现的是"为通过断口测试而专门训练的应试能力"，而真实接管能力并未提高，那么本书即使在理论上正确，在实践上也应当被撤回。这一条由第十一章单独处理，因为它是本书自认最可能发生的一种失败。

三条撤回条款按可检验性排序：第一条最容易，第二条最重要，第三条最可能。

这本书没有回答的

最后诚实地列出四件事，它们都在本书的射程之外，写在这里以免读者把本书当成一套完整方案。

一，本书不知道断口应当多密。九章都能说明为什么需要中断，也都能说明中断处读什么，但没有一章能回答"多久一次"。第二章提出了最大静默期加事件触发的框架，第三章提出了刷新窗口，但两者的具体数值都没有经验依据，目前只能由后果等级与衰减速度粗略推导。

二，本书的读数全部以事件或个人为单位，没有分布量。上一节已经承认，此处重申，因为它同时影响九章。

三，本书没有处理权限问题。第七章指出改写责任需要认知可达性、因果控制杠杆与改写权限三者合一，但其余八章在设计断口测试时，默认了被测者拥有改向与干预的权限。现实中大量人员没有这个权限，把"制度不给权限"读成"个人没有能力"是本书方法最容易造成的伤害。凡实施本书方法的组织，须先分离这两者。

四，本书自身的证据由 AI 深度参与生成。九章的写作、检索、结构与语言都不同程度地使用了生成式系统。按本书自己的判据，这意味着本书的论证质量并不能自动证明作者拥有相应的理解与判断。本书的自我处置是：把每一章的承重命题、竞争解释与撤回条件写死在正文里，使读者可以绕过作者直接检验——这正是本书对自己开的那个断口。检验它的方式与检验其他任何一章相同：找出一处表面相似而关键前提已经改变的地方，看这本书有没有在外部指出之前主动撤回。第三章那次因公开数据不支持而把主张从"能力必然退化"收窄为"能力可观测性下降"，第一章那次因转场效应而把测试从"闭卷"收窄为"定向扰动"，第六章那次因反向证据而撤回"AI 削弱理解"的宽命题，是本书能提供的三处记录。它们不足以证明本书理解了自己在说什么，但至少给出了可供检验的痕迹。

第 十一 章

断口测试自己会被应付掉

一套用于揭露遮蔽的方法，为什么会长出新的遮蔽；四种可预测的应付路径、四条对策及其代价，以及本书愿意接受的降级方案

一个必须由本书自己说出口的问题

前十章的建议可以概括成一句话：不要相信运行良好，要周期性地制造中断，在中断处读数。这套办法的说服力来自它揭露了一种旧办法看不见的东西。但一套办法一旦被组织采纳、写进制度、接上考核，它就不再只是一种观察手段，而成了被观察者所处环境的一部分。观察一旦变成环境，被观察者就会开始适应它。

这不是对本书的外部批评，而是本书自己九章逻辑的直接推论。第四章说得最清楚：当一个指标被用来评价能力，人和组织就会朝这个指标优化，而优化的最省事路径往往不是提高被测量的东西，而是提高测量结果。第七章说得同样清楚：无事故运行会让组织把"制度看起来有效"当成"制度确实有效"。把这两条推论用在本书自己身上，结论无可回避：

断口测试一旦制度化，最省事的达标方式很可能恰恰摧毁它要保护的东西。

第十章把它列为全书三条撤回条款中"最可能发生"的一条。本章的任务是把这条模糊的担忧变成可预测的具体形态、可执行的对策、可观测的预警信号，以及在对策失效时本书愿意接受的降级方案。一个不肯写这一章的理论，等于把自己失败的责任预先推给了执行者。

四条应付路径

应付不会以一种笼统的"敷衍"形式出现。按本书九章的读数结构，它有四条各自具体、各自可预测的路径。

第一条是预告化。 断口的全部价值来自被中断者事先不知道中断会在哪里、以什么形式发生。一旦接管证明被排进季度日历、盲段被固定在每月第一个周五、撤援任务写进培训计划的第三周，被测者所测到的就不再是"长期待命之后重新进入控制"，而是"针对一次已知考试的短期准备"。第二章特意区分过这两者：接管的困难不在动作本身，而在冷启动；一次被预告的接管，恰恰把冷启动这一项去掉了。预告化最隐蔽之处在于，它使读数系统性变好——通过率上升，管理者看到的是能力提高。

第二条是题库化。 断口需要扰动，扰动需要被设计出来。设计成本使组织倾向于复用：同一批事实错误、同一组冲突答案、同一个被删掉的关键前提，第二年照旧使用。扰动一旦稳定，它就从"制造一次真实的越界"退化为"一道可以被认出的题"。第六章的边界破坏任务对此最敏感——它要求主体在表面相似而关键前提已改变时主动撤回，可如果这个"关键前提已改变"的信号本身变成了熟悉的题型标志，被测者撤回的依据就不是理解，而是认出了这是一道撤回题。这时读数记录的是应试敏感度，而第六章正是全书里最需要区分"会用"与"知道何时不用"的一章。

第三条是代考化，也是四条里最尖锐的一条。 当断口测试成为一项要通过的任务，人会用手边最有效的工具去准备它——而手边最有效的工具正是那个被测试要暂时切断的代理。研究者可以让模型帮他预演"如果这份综述里有一条引用被错误解释，我应该怎样发现"；工程师可以让模型生成一份故障注入的应对清单；学生可以让模型替他准备"当模型出错时我该怎么答"。用被测对象来通过对它的测试，在形式上是荒谬的，在实践中却是最自然的选择。它的危险不在于作弊——多数情况下当事人并不认为自己在作弊——而在于它精确地复现了本书第一章要排除的那种状态：结果看起来是主体产生的，闭环实际上仍由代理闭合。

第四条是记账化。 第三章的证据龄比是全书最容易被制度直接吸收的一个读数，因为它形式简单：距上一次有效独立表现证据过去了多久。简单意味着可以做成字段，字段意味着可以做成看板，看板意味着可以要求它变绿。于是"刷新证据"就变成"把这个字段刷新"，而一次为了刷新字段而安排的、时长最短、难度最低、恰好能被记为有效的独立回合，在账面上与一次真正暴露缺口的回合完全等价。第三章自己说过，清偿证据债不等于提升能力——记账化把这句话变成了制度性的常态：债一直在被偿还，能力从来没有被看见。

四条路径有一个共同结构，而这个结构正是本书第四章的循环换了个对象：指标持续变好，被指标指示的东西持续变得不可见。 本书写了十章去揭露这个循环，然后自己造出了它的第十次实例。

这不是执行不力

必须明确一点：上述四条不是"制度设计得不够仔细"或"执行者不够认真"的结果。它们是任何一种把观察接上后果的安排都会产生的。一个用于评价的量会被朝它优化，这是本书第四章引用组织学习与测量理论时就已经承认的机制；本书没有豁免权。

更具体地说，本书的方法有三处结构性弱点，使它比一般指标更容易被应付。

其一，断口是被安排的。与真实故障不同，断口由人设计、由制度触发，因此它的时间、形式与难度都在组织的控制之内，而组织同时也是被这个读数评价的对象。第二章讨论核设施监督试验时提到过一个类似困境：试验本身有成本和风险，于是存在缩减试验强度的压力，而缩减之后账面上的可用性反而更好看。

其二，本书的多数读数是二值门。第十章指出五章的读数是串联门型。二值门的好处是不许取平均，坏处是只要通过就没有区分度——它无法区分勉强通过与轻松通过，因此优化的目标非常明确：把每一道门都调到刚好能过。连续量至少还能显示分布变化，二值门不能。

其三，本书为了不让能力维护变成昂贵形式主义，主张最小化。第二章的最低剂量、第三章的最短证据事件、第八章的最小充分验证，都在往"少而准"的方向压。这个主张本身是对的，但它与应付共享同一句口号：都在寻找通过的最低成本。区别只在于一个要的是最低成本下的真实暴露，另一个要的是最低成本下的账面达标，而这两者在事前很难从制度文本上区分开。

承认这三处弱点，比声称本书方法足够严密要有用得多，因为对策必须针对具体弱点，而不是针对"态度"。

四条对策，以及每条的代价

以下四条对策与四条应付路径一一对应。每条都写明代价——不写代价的对策，本身就是本书第十章批评过的那种空转。

对策一：不可预告性。断口的触发应由随机窗口与事件触发共同决定，而不由固定日历决定；被测者可以知道"本季度会有一次"，但不应知道是哪一天、哪一个环节、哪一类扰动。第二章的"最大静默期加事件触发"框架已经给出了形式：日历只作后备，真正的触发来自模型版本变更、数据源变化、职责变化、长时间未遇异常等可见事件。

代价：组织摩擦与信任成本。不可预告意味着人可能在高压时段被抽中，意味着一次失败的读数会被怀疑是时机不巧。若组织不愿承担这个代价，就不应声称自己在做断口测试，而应诚实地说自己在做定期演练——两者的价值不同，不应混用同一套结论。

对策二：扰动要生成，不要枚举。扰动不应来自一份固定题库，而应从真实事故记录、近失误报告与当期系统变更中持续生成。第七章把近失误称为责任结构的传感器，此处它承担第二个功能：近失误是唯一一种既真实又不昂贵的扰动来源。一个组织如果每年发生的近失误足够多、记录足够细，就永远不缺新的扰动。

代价：需要维护一份不被考核的近失误库。这是最难的一条——近失误一旦与考核挂钩，上报量立刻下降，扰动来源随之枯竭。因此这份库必须与任何绩效评价彻底隔离，且隔离必须是可验证的，而不是承诺的。

对策三：断口测试期间禁止使用同一代理。这一条针对代考化。测试所用的准备资源、判定标准与被测能力，必须与被切断的那个代理在失效模式上低相关。第八章的"证据独立度"与第九章的"独立性折扣"在此处直接可用：如果被测者的准备、执行与自评全部经由同一个模型，那么这次测试的有效独立判断数是一，无论过程看起来多完整。

代价：成本与可行性。要求异源工具意味着组织要为断口测试单独准备一套不同来源的判定材料，而在某些领域根本不存在真正异源的第二条路径。此时正确的做法不是假装做到了，而是在记录中标注这次测试的独立度不足，并按第九章的折扣规则降低它的证据权重。

对策四：断口读数只用于诊断，不用于个体奖惩。这一条是四条里最重要的，因为前三条的效力都依赖于它。一旦读数与晋升、考评、留用挂钩，被测者优化读数的动力将大于任何技术性防护：日历会被打听出来，扰动会被交流出来，代理会被用来准备，字段会被刷绿。反过来，如果读数只进入"哪些能力节点正在变空"的诊断，被测者没有理由伪装，一次不通过反而是有用的信息。

代价：组织不容易接受一项要花钱、要占时间、却不能用于人事决定的测量。这一条的说服只能靠一个论证：读数一旦用于奖惩就会失真，失真之后连诊断价值也没有了；用于诊断的读数是真的，用于奖惩的读数两头落空。

一条不可让步的底线

上一节四条对策不是并列的。对策四是其余三条的前提。如果断口读数被用于个体奖惩，前三条对策会以可预测的方式失效：不可预告性会被组织内部的信息流动抵消，扰动生成会因近失误上报枯竭而退回题库，异源要求会因成本压力被形式化为一句声明。

因此本书给出一条不可让步的底线，任何声称采用本书方法的组织都应当明示：断口读数不进入个体的绩效、晋升与留用决定。它可以进入岗位配置的讨论（哪些节点需要更多支持）、进入系统设计的讨论（哪些环节的自动化程度需要调整）、进入培训资源的分配，但不进入对某个人的评价。

这条底线也划出了本书方法的适用边界。在无法保证这条底线的组织里，本书的建议不应被采纳——不是因为它没用，而是因为它在那种环境下会以看起来成功的方式失败，这比明显失败更糟。

怎样知道应付已经发生

对策会失效，底线会被侵蚀。因此需要一组不依赖当事人自述的信号，用来判断断口测试是否已经退化。以下五个信号都可以从记录本身读出，且都指向本节前述的具体路径。

信号一：通过率单调上升，而恢复时间不变。 如果历次断口测试的通过率持续提高，但真实事件中的异常发现速度与恢复时长没有同向改善，那么提高的是应试而非能力。这一条对应预告化与题库化。

信号二：首次边界修订全部发生在提示之后。 第六章的关键读数是撤回由谁发起。如果记录显示几乎所有的撤回都出现在测试者给出提示、追问或倒计时之后，而自主撤回的比例接近零，那么被测者掌握的是应答流程而非结构边界。

信号三：证据刷新集中在窗口末尾。 第三章的证据龄比若被记账化，刷新事件的时间分布会出现明显的尾部堆积——大量刷新恰好发生在窗口到期前的最后几天，且时长与难度显著低于窗口内其他时段的刷新。这是记账化最容易被抓住的痕迹。

信号四：不同人的断口记录趋于同质。 如果不同背景、不同资历的被测者在同类扰动上给出的判断路径、措辞与裁决依据高度相似，多半说明存在共享的准备材料或共同的代理。这一条正是第九章的方法用在本书自己身上：记录之间的相关性上升，意味着这批读数的有效独立数在下降。

信号五：同一扰动复用超过若干次后通过率跃升。 这是题库化最直接的证据，也最容易检验——只需在记录中标注每个扰动的复用次数，观察通过率与复用次数的关系是否出现台阶。

五个信号中，信号一最重要但最慢，信号三与信号五最快也最便宜。任何采用本书方法的组织，至少应当记录扰动的复用次数与刷新事件的时间分布——这是两项几乎零成本的自我监测，却能挡住四条应付路径中的两条。

如果应付不可避免

假设上述对策全部部署，五个信号仍然出现，本书应当怎么办？此处给出两级降级方案，而不是坚持原方案。

第一级降级：缩小到高后果节点。 放弃在全部能力节点上做断口，只保留那些"若 AI 在此出错、未来仍要求人发现、限制或恢复"且后果不可逆的少数节点。节点越少，扰动越可以做到真实且不重复，不可预告性越容易维持。这一级的代价是覆盖率——大量中低后果的能力将不再被观察，本书接受这个结果，因为一份小而真的读数优于一份全而假的读数。

第二级降级：只记录，不结论。 放弃"通过／不通过"的判定，只保留过程记录：谁在什么时间发现了异常、依据什么、下一步做了什么、结果如何。不打分意味着没有可优化的目标，应付的动力随之消失；代价是失去了可比性与聚合能力，读数无法进入任何管理看板。

如果连第二级也无法维持——即组织在没有分数的情况下不愿投入任何资源做中断——那么本书对该组织的建议只剩下一句，且这句话不需要任何制度支持：在把 AI 的结论用于不可逆的行动之前，先写下一句"如果这是错的，最先出问题的会是什么"，然后去看那一处。 这是本书全部方法压缩到零成本时剩下的东西。它比什么都不做要好，也比一套被应付掉的制度要好。

本章自己也会被应付

最后必须承认：本章同样可以被应付。一个组织完全可以把本章的五个信号做成一份自查表，每季度填写，全部打勾，然后宣布自己没有发生应付。这正是本书第七章描述的那种情形——形式责任越明确，实质责任越可能落空。

本书对此没有更高一层的解法，而且相信不存在这样的解法：任何一层监督都可以被下一层的形式化吞掉。本书能做的只有两件事。

第一，把判断权留在组织之外。本章的五个信号全部是从原始记录（扰动复用次数、刷新时间戳、撤回发起方、记录文本相似度、恢复时长）算出来的，不依赖任何自评。因此只要原始记录对外部审查者开放，应付就无法只靠自查表完成。本书对采用者的最后一项要求是保存并开放这些原始记录，而不是保存结论。

第二，承认这本书可能是错的，并说明它错在哪里最有价值。如果若干年后回看，本书方法在多数组织里演化成了又一套季度合规动作，而人的接管能力与判断归属并未改善，那么这不会是执行的失败，而是本书第十章第三条撤回条款的兑现。届时值得研究的问题也已经变了：不是"怎样测出人还在不在"，而是"为什么每一种用来揭露遮蔽的办法，最后都长成了新的遮蔽"。那个问题比本书大，本书只能把自己作为一个案例交给它。

结语　把第一眼交出去

十一章走到这里，可以把全书压成三句话。

第一句：一直没出事，不能区分两种状态。它可以说明主运行通道在已经发生过的任务分布里没有暴露严重失败，不能说明人的备用路径是否仍可启动。这句话在航空、核电、组织治理与教育里同时成立，而生成式人工智能把它从少数高风险行业的专门问题，变成了几乎所有知识工作的日常处境。

第二句：失去的次序是反的。能力、判断与责任不是先消失、然后被发现，而是先变得不可观测，然后才可能不存在。销毁证据的不是失败，是成功——没有人隐瞒什么，只是产生这类证据的场合不再出现了。而这个过程会自我强化：越看不见，越不知道值不值得去看；越不去看，越看不见。

第三句：唯一的观测位置是断口。常态运行既然已经失去分辨力，读数就只能来自一次有意制造的中断。九章把这个中断做成了九种工程形式，第十章证明这九种形式只有三种量表，第十一章说明它们会怎样被应付掉。

这本书没有做成的事

本书没有提出任何提高人类能力的办法。它自始至终只处理"看得见"这一侧：诊断出问题不等于已经治疗，清偿证据债不等于提升能力。把本书当作训练方案来用，一定会失望。

本书也没有解决它自己指出的那个最深的困难。断口需要成本，付成本需要有人先相信"可能已经不行了"，而让人相信这件事的材料，恰恰是还没有被生产出来的那些证据。这本书能给出全部的方法，唯独给不出第一眼。

第一眼只能由人交出来。它可能来自一次真实的事故，来自一份外部审计，来自监管的强制，也可能只来自某个人在某个晚上忽然想知道：这件事我们上一次真正看见，是什么时候。前三种通常来得太晚，第四种最便宜，也最不可靠——它依赖的是一种正在被系统性削弱的东西，也就是一个人对自己是否还看得见的不安。

一句可以马上做的话

如果这本书的全部制度建议都无法在你所处的位置上落地，还剩下一件不需要任何人同意的事：

在把人工智能的结论用于一次不可逆的行动之前，先写下一句"如果这是错的，最先出问题的会是什么"，然后去看那一处。

这是全书十九万字压缩到零成本之后剩下的东西。它比什么都不做要好，也比一套被应付掉的制度要好。

最后一件

本书用十一章说明，一个只呈现良好读数、从不暴露自己在哪里会失败的系统，那份良好读数本身就应当被怀疑。这条要求首先适用于本书自己。因此全书写明了三条撤回条款、三对可能被合并的章节、四件本书没有回答的事，以及三处因公开数据不利而被迫收窄的主张。

如果若干年后回看，这套方法在多数组织里演化成了又一套季度合规动作，而人的接管能力与判断归属并未改善，那么这不会是执行的失败，而是本书第十章第三条撤回条款的兑现。届时值得研究的问题也已经变了：不是"怎样测出人还在不在"，而是"为什么每一种用来揭露遮蔽的办法，最后都长成了新的遮蔽"。

那个问题比这本书大。这本书只能把自己作为一个案例交给它。

参考文献

本表由全书各章文献合并去重而成，西文在前、中文在后，各按著者字顺排列。各章正文中的引注编号为该章原编号，与本表序号不对应；查证时请按著者与年份检索。

1. Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaïd, M., & Pentland, A. (2020). Adaptive Social Networks Promote the Wisdom of Crowds. Proceedings of the National Academy of Sciences, 117(21), 11379–11386.

2. Andrada, G., & Carter, J. A. (2025). Mind-Technology Problems for Know-How Anti-Intellectualism. Social Epistemology. https://doi.org/10.1080/02691728.2025.2463059

3. Ashkinaze, J., Mendelsohn, J., Qiwei, L., Budak, C., & Gilbert, E. (2024/2025). How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment. arXiv:2401.13481.

4. Bainbridge L. Ironies of automation[J]. Automatica, 1983, 19(6): 775-779. DOI:10.1016/0005-1098(83)90046-8.

5. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779.

6. Bandura, A. (1999). Moral Disengagement in the Perpetration of Inhumanities. Personality and Social Psychology Review, 3(3), 193–209.

7. Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of Moral Disengagement in the Exercise of Moral Agency. Journal of Personality and Social Psychology, 71(2), 364–374.

8. Barkoczi, D., & Galesic, M. (2016). Social Learning Strategies Modify the Effect of Network Structure on Group Performance. Nature Communications, 7, 13109.

9. Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

10. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. DOI: 10.1073/pnas.2422633122. （另见 2025 年作者更正：e2518204122）

11. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2422633122

12. Baumann, F., Czaplicka, A., & Rahwan, I. (2024). Network Structure Shapes the Impact of Diversity in Collective Learning. Scientific Reports, 14, 2491.

13. Bickel C S, Cross J M, Bamman M M. Exercise dosing to retain resistance training adaptations in young and older adults[J]. Medicine & Science in Sports & Exercise, 2011, 43(7): 1177-1187. DOI:10.1249/MSS.0b013e318207c15d.

14. Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417–444.

15. Bleher, H., Braun, M., & others. (2022). Attributions of Responsibility in the Use of AI-driven Clinical Decision Support Systems. Frontiers / related empirical literature.

16. Blythe, T., & Associates. (1998). The Teaching for Understanding Guide. San Francisco: Jossey-Bass; Project Zero, Harvard Graduate School of Education.

17. Bovens, M., Schillemans, T., & Goodin, R. E. (eds.). (2014). The Oxford Handbook of Public Accountability. Oxford University Press. 相关章节讨论多手问题与组织责任。

18. Boyd, R., & Richerson, P. J. (1985). Culture and the Evolutionary Process. University of Chicago Press.

19. Boyd, R., & Richerson, P. J. (2005). The Origin and Evolution of Cultures. Oxford University Press.

20. Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24, 61–100. https://doi.org/10.3102/0091732X024001061

21. Braverman H. Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century[M]. New York: Monthly Review Press, 1974.

22. Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), Handbook of Child Psychology, Vol. 3. Wiley.

23. Brown, G., Wyatt, J. L., & Tiňo, P. (2005). Managing Diversity in Regression Ensembles. Journal of Machine Learning Research, 6, 1621–1650.

24. Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at Work. Quarterly Journal of Economics, 140(2), 889–942.

25. Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., et al. (2024). How Large Language Models Can Reshape Collective Intelligence. Nature Human Behaviour, 8(9), 1643–1655.

26. Butler, O. (2025). Algorithmic Decision-Making, Delegation and the Modern Machinery of Government. Oxford Journal of Legal Studies, 45(3), 727–754.

27. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21.

28. Cananea, G. della. (2024). Administrative Procedure Acts in Europe: An Emerging Common Core? American Journal of Comparative Law, 72(2), 324–376.

29. Cane, P. (2002). Responsibility in Law and Morality. Hart Publishing.

30. Cao, S., Liu, A., & Huang, C.-M. (2024). Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. arXiv:2401.05612.

31. Caosun, M., & Aral, S. (2026). The Augmentation Trap: AI Productivity and the Cost of Cognitive Offloading. arXiv:2604.03501.

32. Casner S M, Geven R W, Recker M P, Schooler J W. The retention of manual flying skills in the automated cockpit[J]. Human Factors, 2014, 56(8): 1506-1516. DOI:10.1177/0018720814535628.

33. Caspar, E. A., Christensen, J. F., Cleeremans, A., & Haggard, P. (2016). Coercion Changes the Sense of Agency in the Human Brain. Current Biology, 26(5), 585–592.

34. Centola, D. (2022). The Network Science of Collective Intelligence. Trends in Cognitive Sciences, 26(11), 923–941.

35. Centola, D., & Baronchelli, A. (2015). The Spontaneous Emergence of Conventions: An Experimental Study of Cultural Evolution. Proceedings of the National Academy of Sciences, 112(7), 1989–1994.

36. Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

37. Chirikov, I., Smirnov, I., & Kizilcec, R. F. (2026). Generative AI use and misuse call for assessment reform in higher education. Science, 392, 818–820. https://doi.org/10.1126/science.aec5115

38. Clark, A. (2008). Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford University Press.

39. Clark, A., & Chalmers, D. (1998). The Extended Mind. Analysis, 58(1), 7–19.

40. Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), Knowing, Learning, and Instruction. Lawrence Erlbaum Associates.

41. Collins, H. M., & Evans, R. (2002). The third wave of science studies: Studies of expertise and experience. Social Studies of Science, 32(2), 235–296.

42. Collins, H., & Evans, R. (2007). Rethinking Expertise. Chicago: University of Chicago Press.

43. Constantinescu, M., et al. (2025). Responsibility Gaps, LLMs & Organisations: Many Agents, Distributed Accountability and Business Ethics. AI and Ethics / PMC.

44. Contractor, Z., & Reyes, G. (2026). Experimental Evidence on the Learning Impact of Generative AI. arXiv:2607.08849. Preprint.

45. Crossan, M. M., Lane, H. W., & White, R. E. (1999). An Organizational Learning Framework: From Intuition to Institution. Academy of Management Review, 24(3), 522–537.

46. Crozier, M., & Friedberg, E. (1980). Actors and Systems: The Politics of Collective Action. University of Chicago Press.

47. Darley, J. M., & Latané, B. (1968). Bystander Intervention in Emergencies: Diffusion of Responsibility. Journal of Personality and Social Psychology, 8(4), 377–383.

48. Dekker, S. (2014). The Field Guide to Understanding Human Error (3rd ed.). CRC Press.

49. Dell' Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper 24-013.

50. Denton, K. K., Ram, Y., Liberman, U., & Feldman, M. W. (2020). Cultural Evolution of Conformity and Anticonformity. Proceedings of the National Academy of Sciences, 117, 13603–13614.

51. Doshi, A. R., & Hauser, O. P. (2024). Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. Science Advances, 10(28), eadn5290.

52. Doshi-Velez, F., et al. (2017). Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center / arXiv.

53. Dunbar, K. (1993). Concept discovery in a scientific domain. Cognitive Science, 17(3), 397–434.

54. Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, 5, 40–60.

55. Elmqvist, T., Folke, C., Nyström, M., Peterson, G., Bengtsson, J., Walker, B., & Norberg, J. (2003). Response Diversity, Ecosystem Change, and Resilience. Frontiers in Ecology and the Environment, 1(9), 488–494.

56. Endsley M R, Kiris E O. The out-of-the-loop performance problem and level of control in automation[J]. Human Factors, 1995, 37(2): 381-394.

57. Endsley M R. Toward a theory of situation awareness in dynamic systems[J]. Human Factors, 1995, 37(1): 32-64.

58. Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.

59. Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. Human Factors, 37(2), 381–394. https://doi.org/10.1518/001872095779064555

60. European Union. (2024). Directive (EU) 2024/2853 on liability for defective products.

61. European Union. (2024). Directive (EU) 2024/2853, Product Liability Directive, modernised rules for software and defective products.

62. European Union. (2024). Regulation (EU) 2024/1689, Article 14 Human Oversight and Article 25 Responsibilities along the AI Value Chain.

63. European Union. (2024, consolidated text 2026). Regulation (EU) 2024/1689 (Artificial Intelligence Act), especially Articles 14, 25–26.

64. European Union. (2024/2026 consolidated text). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, Article 14: Human oversight.

65. Faas, C., Uth, R., Sterz, S., Langer, M., & Feit, A. M. (2026). Don’t blame me: How Intelligent Support Affects Moral Responsibility in Human Oversight. arXiv preprint.

66. Federal Aviation Administration. AC 61-98E: Currency Requirements and Guidance for the Flight Review and Instrument Proficiency Check[S]. Washington, DC: FAA, 2024.

67. Ferdman, A. (2025). AI deskilling is a structural problem. AI & Society. DOI: 10.1007/s00146-025-02686-z.

68. Fleisher, W. (2025). Responsibility and Accountability in an Algorithmic Society. Philosophy & Technology.

69. Fleishman E A, Parker J F. Factors in the retention and relearning of perceptual-motor skill[J]. Journal of Experimental Psychology, 1962, 64(3): 215-226.

70. Gadamer, H.-G. (2004). Truth and Method (2nd rev. ed., J. Weinsheimer & D. G. Marshall, Trans.). London/New York: Continuum. (Original work published 1960).

71. Gettier, E. L. (1963). Is Justified True Belief Knowledge? Analysis, 23(6), 121–123. https://doi.org/10.1093/analys/23.6.121

72. Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin.

73. Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

74. Goldman, A. I. (1979). What Is Justified Belief? In G. Pappas (Ed.), Justification and Knowledge: New Studies in Epistemology (pp. 1–25). D. Reidel.

75. Greeno, J. G. (1998). The situativity of knowing, learning, and research. American Psychologist, 53(1), 5–26.

76. Guo, Z., Wu, Y., Hartline, J., & Hullman, J. (2024). A Decision Theoretic Framework for Measuring AI Reliance. arXiv:2401.15356.

77. Haggard, P. (2017). Sense of Agency in the Human Brain. Nature Reviews Neuroscience, 18, 196–207.

78. Hanna, G. (1990). Some pedagogical aspects of proof. Interchange, 21(1), 6–13.

79. Hardwig, J. (1985). Epistemic Dependence. The Journal of Philosophy, 82(7), 335–349.

80. Harlow, C., & Rawlings, R. (2009). Law and Administration (3rd ed.). Cambridge University Press.

81. Hart, H. L. A., & Honoré, T. (1985). Causation in the Law (2nd ed.). Oxford University Press.

82. Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), Child Development and Education in Japan (pp. 262–272). New York: W. H. Freeman.

83. Heft, H. (2001). Ecological Psychology in Context: James Gibson, Roger Barker, and the Legacy of William James' s Radical Empiricism. Lawrence Erlbaum Associates.

84. Heidegger, M. (1962). Being and Time (J. Macquarrie & E. Robinson, Trans.). Oxford: Blackwell. (Original work published 1927).

85. Henrich, J. (2004). Demography and Cultural Evolution: How Adaptive Cultural Processes Can Produce Maladaptive Losses—The Tasmanian Case. American Antiquity, 69(2), 197–214.

86. Henrich, J., & Gil-White, F. J. (2001). The Evolution of Prestige: Freely Conferred Deference as a Mechanism for Enhancing the Benefits of Cultural Transmission. Evolution and Human Behavior, 22(3), 165–196.

87. Hepburn, J. (2012). The Duty to Give Reasons for Administrative Decisions in International Law. International & Comparative Law Quarterly, 61(3), 641–663.

88. Hickson R C, Rosenkoetter M A. Reduced training frequencies and maintenance of increased aerobic power[J]. Medicine & Science in Sports & Exercise, 1981, 13(1): 13-16.

89. Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction, 7(2), 174–196.

90. Hong, L., & Page, S. E. (2004). Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. Proceedings of the National Academy of Sciences, 101(46), 16385–16389.

91. Horvitz, E. (1999). Principles of mixed-initiative user interfaces. Proceedings of CHI 1999, 159–166.

92. Hosanagar, K., & Ahn, D. (2024). Designing Human and Generative AI Collaboration. arXiv:2412.14199.

93. Huesmann, K., et al. (2024). Perception-action coupling in anticipation research: a classification and its application to racket sports. Frontiers in Psychology, 15, 1396873.

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

95. International Atomic Energy Agency. Maintenance, Testing, Surveillance and Inspection in Nuclear Power Plants[S]. IAEA Safety Standards Series No. SSG-74. Vienna: IAEA, 2022.

96. International Atomic Energy Agency. Risk Based Optimization of Technical Specifications for Operation of Nuclear Power Plants[R]. IAEA-TECDOC-729. Vienna: IAEA, 1993.

97. Jiménez, Á. V., & Mesoudi, A. (2019). Prestige-Biased Social Learning: Current Evidence and Outstanding Questions. Humanities and Social Sciences Communications, 5, 20.

98. Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theoretical Issues in Ergonomics Science, 5(2), 113–153.

99. Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

100. Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263–291.

101. Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1–73.

102. Kapur, M. (2008). Productive failure. Cognition and Instruction, 26(3), 379–424.

103. Karpicke, J. D., & Roediger, H. L. III. (2008). The critical importance of retrieval for learning. Science, 319(5865), 966–968.

104. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15, 17458. DOI: 10.1038/s41598-025-97652-6.

105. Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated Errors in Large Language Models. arXiv:2506.07962.

106. Kirsh, D. (2010). Thinking with external representations. AI & Society, 25, 441–454.

107. Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior.

108. Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. Educational Psychology Review, 19(3), 239–264.

109. Krenn, M., Pollice, R., Guo, S. Y., et al. (2022). On scientific understanding with artificial intelligence. Nature Reviews Physics, 4, 761–769. https://doi.org/10.1038/s42254-022-00518-3

110. Krogh, A., & Vedelsby, J. (1995). Neural Network Ensembles, Cross Validation, and Active Learning. Advances in Neural Information Processing Systems 7, 231–238.

111. Kuhn, D. (2005). Education for Thinking. Cambridge, MA: Harvard University Press.

112. Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. Quarterly Journal of Economics, 112(2), 443–477.

113. Lakatos, I. (1976). Proofs and Refutations: The Logic of Mathematical Discovery. Cambridge: Cambridge University Press.

114. Lee H P, Sarkar A, Tankelevitch L, Drosos I, Rintel S, Banks R, Wilson N. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers[C]//Proceedings of the CHI Conference on Human Factors in Computing Systems. 2025.

115. Lee, H.-P. (H.), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. Proceedings of CHI 2025, Article 1121, 1–22. https://doi.org/10.1145/3706598.3713778

116. Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. Proceedings of CHI 2025. DOI: 10.1145/3706598.3713778.

117. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.

118. Legaspi, R., et al. (2024). The Sense of Agency in Human–AI Interactions. Knowledge-Based Systems.

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

120. Levitt, B., & March, J. G. (1988). Organizational Learning. Annual Review of Sociology, 14, 319–340.

121. Liu, G., Christian, B., Dumbalska, T., Bakker, M. A., & Dubey, R. (2026). AI Assistance Reduces Persistence and Hurts Independent Performance. arXiv:2604.04721.

122. Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The History and Philosophy of Ecological Psychology. Frontiers in Psychology, 9, 2228.

123. Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How Social Influence Can Undermine the Wisdom of Crowd Effect. Proceedings of the National Academy of Sciences, 108(22), 9020–9025.

124. Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., & Ma, X. (2025). Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. Proceedings of CHI 2025.

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

126. March, J. G., & Simon, H. A. (1958). Organizations. Wiley.

127. Mashaw, J. L. (2018). Reasoned Administration and Democratic Legitimacy: How Administrative Law Supports Democratic Government. Cambridge University Press.

128. Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. Ethics and Information Technology, 6(3), 175–183.

129. McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate Neural Systems Value Immediate and Delayed Monetary Rewards. Science, 306(5695), 503–507.

130. Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). Macmillan.

131. Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Educational Researcher, 23(2), 13–23.

132. Morgan, T. J. H., & Feldman, M. W. (2025). Human Culture Is Uniquely Open-Ended Rather Than Uniquely Cumulative. Nature Human Behaviour, 9, 28–42.

133. NIST AI Resource Center. (2023–2026). AI RMF Core and Playbook: Govern, Map, Measure, Manage.

134. NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1.

135. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

136. National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1.

137. Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated Knowledge from a Small Number of Debates Outperforms the Wisdom of Large Crowds. Nature Human Behaviour, 2, 126–132.

138. Nisioti, E., Risi, S., Momennejad, I., Oudeyer, P.-Y., & Moulin-Frier, C. (2024). Collective Innovation in Groups of Large Language Models. arXiv:2407.05377.

139. Nissenbaum, H. (1996). Accountability in a Computerized Society. Science and Engineering Ethics, 2(1), 25–42.

140. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192.

141. Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse[J]. Human Factors, 1997, 39(2): 230-253.

142. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.

143. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A, 30(3), 286–297.

144. Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590.

145. Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Basic Books.

146. Polanyi, M. (1958). Personal Knowledge: Towards a Post-Critical Philosophy. Chicago: University of Chicago Press.

147. Raees, M., Khan, V.-J., Lykourentzou, I., & Papangelis, K. (2026). Do People Appropriately Rely on AI-Advice? An Analytical Review of HCI Research on Human-AI Decision-Making. Proceedings of CHI 2026.

148. Raghavan, M. (2024). Competition and Diversity in Generative AI. arXiv:2412.08610.

149. Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. Journal of Experimental Education, 70(4), 293–315.

150. Ricoeur, P. (1973). The model of the text: Meaningful action considered as a text. New Literary History, 5(1), 91–117. https://doi.org/10.2307/468410

151. Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

152. Ryle, G. (1949). The Concept of Mind. London: Hutchinson.

153. Salatino, A., Prével, A., Caspar, E. A., & Lo Bue, S. (2025). Influence of AI behavior on human moral decisions, agency, and responsibility. Scientific Reports, 15.

154. Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science, 311(5762), 854–856.

155. Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. Educational Psychologist, 24(2), 113–142.

156. Savage, L. J. (1954). The Foundations of Statistics. John Wiley & Sons.

157. Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of IUI 2023, 410–422.

158. Schemmer, M., et al. (2023). Towards Effective Human-AI Decision-Making: The Role of Human Learning in Appropriate Reliance on AI Advice. ICIS 2023 Proceedings.

159. Schlonsak, R., Jetter, H.-C., & Matthies, D. J. C. (2026). The Cost of Convenience: How Delegation and Decision Horizons Erode Human Agency when overusing LLM Systems. CHI EA ’ 26. DOI: 10.1145/3772363.3798760.

160. Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. Proceedings of CHI 2024.

161. Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. Cognition and Instruction, 16(4), 475–522.

162. Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22(2), 129–184.

163. Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. P. Mestre (Ed.), Transfer of Learning from a Modern Multidisciplinary Perspective. Information Age Publishing.

164. Shen, J. H., & Tamkin, A. (2026). How AI Impacts Skill Formation. arXiv:2601.20245.

165. Shen, J. H., & Tamkin, A. (2026). How AI Impacts Skill Formation. arXiv:2601.20245v2.

166. Sheridan, T. B., Verplank, W. L., & Brooks, T. L. (1978). Human/Computer Control of Undersea Teleoperators. Proceedings of the 14th Annual Conference on Manual Control / NASA Technical Reports Server, Document 19790007441.

167. Shirado, H., & Christakis, N. A. (2017). Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments. Nature, 545, 370–374.

168. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI Models Collapse When Trained on Recursively Generated Data. Nature, 631, 755–759.

169. Simon, H. A. (1947/1997). Administrative Behavior. Free Press.

170. Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of Experimental Psychology: Human Learning and Memory, 4(6), 592–604.

171. Snook, S. A. (2000). Friendly Fire: The Accidental Shootdown of U.S. Black Hawks over Northern Iraq. Princeton University Press.

172. Soderstrom, N. C., & Bjork, R. A. (2015). Learning Versus Performance: An Integrative Review. Perspectives on Psychological Science, 10(2), 176–199.

173. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. Science, 333(6043), 776–778. https://doi.org/10.1126/science.1207745

174. Stapleton, J. (2008). Choosing What We Mean by "Causation" in the Law. Missouri Law Review, 73.

175. Steiner, M. (1978). Mathematical explanation. Philosophical Studies, 34(2), 135–151.

176. Taylor, H. L., Talleur, D. A., Bradshaw, G. L., Emanuel, T. W. Jr., Hulin, C. L., Rantanen, E. M., & Lendrum, L. (2003). Effectiveness of Personal Computers to Meet Recency of Experience Requirements. Federal Aviation Administration technical report 03/03.

177. Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. American Political Science Review, 74(4), 905–916.

178. Thurston, W. P. (1994). On proof and progress in mathematics. Bulletin of the American Mathematical Society, 30(2), 161–177. https://doi.org/10.1090/S0273-0979-1994-00502-6

179. Ueda, K., et al. (2021). Influence of levels of automation on the sense of agency during continuous action. Scientific Reports, 11. PMID: 33510395.

180. Ueshima, A., Jones, M. I., & Christakis, N. A. (2024). Simple Autonomous Agents Can Enhance Creative Semantic Discovery by Human Groups. Nature Communications, 15, 5212.

181. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–38.

182. Vaughan, D. (1996). The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press.

183. Vered, M., et al. (2023). The effects of explanations on automation bias. Artificial Intelligence, 322, 103952.

184. Wan, Y., & Kalman, Y. M. (2025). Using Generative AI Personas Increases Collective Diversity in Human Ideation. arXiv:2504.13868.

185. Westbrook, A., Kester, D., & Braver, T. S. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. PLoS ONE, 8(7), e68210.

186. Whitehead, A. N. (1929). The Aims of Education and Other Essays. New York: Macmillan.

187. Wiggins, G., & McTighe, J. (2005). Understanding by Design (2nd ed.). Alexandria, VA: ASCD.

188. Wiles, E., Krayer, L., Abbadi, M., Awasthi, U., Kennedy, R., Mishkin, P., Sack, D., & Candelon, F. (2024). GenAI as an Exoskeleton: Experimental Evidence on Knowledge Workers Using GenAI on New Skills. SSRN 4944588.

189. Wilkenfeld, D. A. (2013). Understanding as representation manipulability. Synthese, 190, 997–1016. https://doi.org/10.1007/s11229-011-0055-x

190. Williams, R. (2022). Rethinking Administrative Law for Algorithmic Decision Making. Oxford Journal of Legal Studies, 42(2), 468–494.

191. Wise, S. L. (2019). Controlling Construct-Irrelevant Factors Through Computer-Based Testing. Assessment in Education: Principles, Policy & Practice.

192. Wittgenstein, L. (1953). Philosophical Investigations. Oxford: Blackwell.

193. Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An Internal Model for Sensorimotor Integration. Science, 269(5232), 1880–1882. https://doi.org/10.1126/science.7569931

194. Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100.

195. Wood, D., Mu, T., Webb, A. M., Reeve, H. W. J., Luján, M., & Brown, G. (2023). A Unified Theory of Diversity in Ensemble Learning. Journal of Machine Learning Research, 24(359), 1–49.

196. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science, 330(6004), 686–688.

197. Wu, S., Liu, Y., Ruan, M., Chen, S., & Xie, X. (2025). Human–Generative AI Collaboration Enhances Task Performance but Undermines Human’s Intrinsic Motivation. Scientific Reports, 15, 15105.

198. Wu, S., Liu, Y., Ruan, M., Chen, S., et al. (2025). Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. Scientific Reports, 15, 15105. https://doi.org/10.1038/s41598-025-98385-2

199. Wu, S., Liu, Y., Ruan, M., et al. (2025). Human-generative AI collaboration enhances task performance but undermines human’s intrinsic motivation. Scientific Reports, 15, 15105. https://doi.org/10.1038/s41598-025-98385-2

200. Yan, L., Greiff, S., Lodge, J. M., & Gaševic,́ D. (2025). Distinguishing Performance Gains from Learning When Using Generative AI. Nature Reviews Psychology, 4, 435–436.

201. Zöller, N., et al. (2025). Human–AI Collectives Most Accurately Diagnose Clinical Vignettes. Proceedings of the National Academy of Sciences, 122(24), e2426153122.

202. de Regt, H. W. (2017). Understanding Scientific Understanding. Oxford: Oxford University Press.

203. de Villiers, M. (1990). The role and function of proof in mathematics. Pythagoras, 24, 17–24.

204. van de Poel, I. (2011/2012). The Problem of Many Hands: Climate Change as an Example. Science and Engineering Ethics, 17/18.

205. von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton University Press.

206. 孔凡鹤. 正确答案不等于能力：生成式 AI 时代“不可外包能力”的断代理复闭环判据[R]. , 2026.

207. 孔凡鹤（2026）.《消化性裁决》. 未刊稿。

208. 孔凡鹤（2026）.《退出迟滞：无明显症状阶段成瘾风险的动态识别与正当干预》.. 未刊稿。

209. 本章为理论与方法论文，不声称“结构辖域判据”已经获得直接实证确认。文中对生成式 AI 学习效应的讨论依据截至 2026 年 8 月 11 日可核验的公开论文与预印本；其中 Contractor 与 Reyes（2026）明确按预印本处理。公开实验主要用于压力测试“AI 辅助表现能否直接等同于人的理解”这一较弱命题，尚未直接检验本章提出的边界破坏任务。

210. ［10］National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press, 2019.

211. ［11］Popper K. The Logic of Scientific Discovery. London: Hutchinson, 1959.

212. ［12］Mayo D G. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge: Cambridge University Press, 2018.

213. ［13］Huang J, Chen J, Ji Z, et al. Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations (ICLR), 2024.

214. ［14］Manakul P, Liusie A, Gales M. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP, 2023: 9004-9017.

215. ［15］Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. Nature, 2024, 630: 625-630.

216. ［16］Dhuliawala S, Komeili M, Xu J, et al. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495, 2023.

217. ［17］Wei J, Yang C, Song X, et al. Long-form Factuality in Large Language Models. arXiv:2403.18802, 2024.

218. ［18］Min S, Krishna K, Lyu X, et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP, 2023.

219. ［19］Gou Z, Shao Z, Gong Y, et al. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. ICLR, 2024.

220. ［1］International Auditing and Assurance Standards Board (IAASB). International Standard on Auditing 200: Overall Objectives of the Independent

Auditor and the Conduct of an Audit in Accordance with International Standards on Auditing. In: Handbook of International Quality Management, Auditing, Review, Other Assurance, and Related Services Pronouncements.

221. ［20］Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 2025.

222. ［21］NIST. Developing Cyber-Resilient Systems: A Systems Security Engineering Approach, NIST SP 800-160 Vol. 2 Rev. 1. 2021.

223. ［22］Ioannidis J P A. Why Most Published Research Findings Are False. PLoS Medicine, 2005, 2(8): e124.

224. ［23］Gundersen O E. The Fundamental Principles of Reproducibility. Philosophical Transactions of the Royal Society A, 2021, 379: 20200210.

225. ［24］Devezer B, Nardin L G, Baumgaertner B, Buzbas E O. Scientific Discovery in a Model-Centric Framework: Reproducibility, Innovation, and Epistemic Diversity. PLoS ONE, 2019, 14(5): e0216125.

226. ［25］NASA. NASA-STD-8729.1A: NASA Reliability and Maintainability (R&M) Standard for Spaceflight and Support Systems. 2017.

227. ［26］NASA. NASA-HDBK-8709.22: Safety and Mission Assurance Acronyms, Abbreviations, and Definitions. 2018.

228. ［2］IAASB. International Standard on Auditing 320: Materiality in Planning and Performing an Audit.

229. ［3］IAASB. International Standard on Auditing 530: Audit Sampling.

230. ［4］IAASB. ISA 315 (Revised 2019): Identifying and Assessing the Risks of Material Misstatement. 2019.

231. ［5］NASA. NASA Systems Engineering Handbook, NASA/SP-2016-6105 Rev 2. Washington, DC: National Aeronautics and Space Administration, 2016.

232. ［6］NASA Goddard Space Flight Center. GSFC-HDBK-8004: Failure Modes, Effects, and Criticality Analysis (FMECA) Handbook. 2024.

233. ［7］NASA Software Engineering Handbook. SWE-141: Software Independent Verification and Validation (IV&V).

234. ［8］National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. 2023.

235. ［9］NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. 2024.

后记

这本书的九章是分开写成的，写的时候并不知道它们会变成一本书。

真正让我决定把它们合起来的，是第五篇写完之后的一次困惑：我发现自己在用不同的词说同一句话，而且每一次都以为那是一个新发现。这种情形通常有两种解释——要么是同一条判断在不同领域反复成立，要么是我把一个模板套到了九个地方。这两种解释在文本上很难区分，因为它们产生的文字看起来一模一样。

第十章就是为了区分它们而写的。它没有辩解九章不重复，而是反过来做：把最可能被合并的三对摆出来，写明在什么经验条件下本书应当合并、应当降为十章。写完之后我对这本书的信心反而增加了，因为如果那九章真的是同一章，第十章会是最先塌掉的地方，而它没有塌。

第十一章的写作过程不太一样。那一章是在我意识到一件不舒服的事之后才动笔的：我用十章批评了一种"指标越好看、被指标指示的东西越不可见"的机制，然后正在造出这个机制的第十次实例。承认这一点比想象中难，因为它意味着这本书最有可能的下场，不是被反驳，而是被采纳之后做成一份季度表格。我把这一章放进正文而不是附录，是想让它不容易被跳过。

书里几处最硬的判断，来自数据把我原来的话打掉之后。第三章那句"人工智能可以在不损害平均成绩的情况下，降低当前表现作为能力信号的保真度"，最初的版本要响亮得多，也要错得多。我保留了修改的痕迹，写在前言里。做研究久了会知道，一句被数据修剪过的话，读起来往往不如原来的漂亮，但它是唯一能立得住的那一句。

要感谢的人不少。促成这本书的三个场景各自有它们的当事人：那位说不出学生会了什么的老师、那位说"下次很可能还会发生"的工程师，以及航空与核电文献里几十年前就把这个问题写清楚的那些研究者。他们并不知道自己参与了这本书。

最后想说的一句，与本书主题有关，也与写作本身有关。这本书的写作过程深度使用了生成式系统，因此我比多数作者更没有资格声称，这些文字证明了我理解自己在说什么。我唯一能做的处置，是把每一章的承重命题、竞争解释与撤回条件写死在正文里，让检验绕过我进行。至于我自己有没有在被指出之前主动撤回过——书里留下了三处记录，判断权在读者。

孔凡鹤

二〇二六年八月

一直没出事，
不能区分两种状态。

航空用它证明机队适航，核电用它证明备用系统可用，学校用它证明教学没有问题，企业用它证明治理健全。这句话之所以长期有效，靠的是一个从未明说的前提：产生结果的过程与承担能力的主体是同一个。生成式人工智能抽掉了这个前提。

本书由九项跨学科研究与两章合论组成，得到同一条判断：能力、判断与责任并不是先失去，而是先失去可观测性；正常运行的成功本身销毁了证明它们仍在的证据。九个领域的良好信号——结果质量、零事故、上升的绩效曲线、末端签字、能迁移、有责任人、检查覆盖率、多数同意——全都失去了分辨力，而且失效的方式是同一种：它们不是变差了，是变得更好看了。

书中给出九个可执行的读数、三种量表、一套记录格式，以及三条会让本书自己失效的撤回条款。最后一章处理它自认最可能应验的失败：一套用于揭露遮蔽的方法，为什么会长出新的遮蔽。

孔凡鹤

照护与人机协作研究者。另著《照护的智慧》（专著第 72 号）。

研究关注长期代行条件下，能力、判断与责任的归属与可观测性。