溶解性边界

AI进入大学课堂后真正需警惕的，不是一切认知外包、也不是人机边界流动本身，而是学习者在尚未形成稳定判断时把问题界定、方案生成、证据筛选连续交给系统、却仍被要求对产出负责。本文将"溶解性边界"界定为一种可判定的风险：控制权、可追溯性与责任彼此失配。

作者 陈晓艳 · SDE 学员 · 发表于2026年7月25日 · 全球文库校订版

生成式人工智能进入大学课堂后，真正需要警惕的并非一切认知外包，也不是人机边界本身变得流动，而是学习者在尚未形成稳定专业判断时，把问题界定、方案生成、证据筛选与表达整合连续交给系统，却仍被要求对最终产出负责。本文将“溶解性边界”界定为一种可判定的教育风险：关键认知环节的控制权、过程可追溯性与结果责任彼此失配，以致学习者无法重建判断链、校准信心或在无 AI 条件下迁移能力。文章区分有益的边界开放与有害的责任失配，并提出四项判据：决策权不清、理由链断裂、能力归因失准、责任不可履行。基于此，本文设计“先行立场—显式授权—双轨比较—口头辩护—无辅助迁移”的教学流程及可证伪指标。核心主张是：大学无需守护一个封闭、纯粹的个体心智，而应保证学生对关键判断拥有可解释的否决权，并具备与其责任相称的理解、追溯和迁移能力。

一、问题不在边界流动，而在控制、理解与责任脱节

大学关于 AI 的讨论常被压缩为“能不能用”和“用到什么程度”。前者容易滑向禁令，后者容易滑向文本占比。两种治理方式都把主体性想象成一块固定领地：只要某些环节仍由学生完成，主体就被保住；只要 AI 越过红线，主体就受损。这个空间隐喻过于粗糙。写作、研究和专业判断从来依赖书籍、教师、同伴、软件与制度，边界开放本身不是病理。

更准确的问题是：谁决定问题怎样被界定？谁能够否决一个看似合理的方案？关键理由能否被重建？当结果造成错误时，学生是否拥有足以承担责任的理解与控制？生成式 AI 的特殊性，不是它第一次把认知放到头脑之外，而是它能一次性承担构架、生成、筛选和表述，并以高度流畅的界面压缩这些环节之间的差异。学生看似全程参与，却可能只在系统已经划定的选择空间内确认。

因此，本文保留“溶解性边界”这个命名，但改变其理论负荷。它不指“自我与工具的界线消失”，也不声称 AI 必然侵蚀主体；它专指一种控制—责任失配：学习者仍被视为成果作者和责任主体，却无法说明关键判断如何形成、哪些选择由自己作出、何时应当推翻系统建议。被溶解的不是一个形而上的自我，而是可问责判断所需要的决策结构。

二、理论文章的证据边界：从情境发现问题，而非用故事替代证明

（一）复合情境的认识功能

本文所使用的文学课与编程课情境，是根据大学课堂中常见的人机协作现象构造的复合情境。它们的作用不是报告某所学校的事实，也不是证明某种风险已经普遍发生，而是把一个容易被成品质量遮蔽的问题呈现出来：学生可能交出结构完整、语言流畅、技术上合格的作品，却无法说明决定作品方向的关键选择如何形成。

复合情境能够帮助理论辨认机制，却不能提供发生率或因果效应。因而，本文从这些情境中提出的是可检验问题：当 AI 介入问题界定、候选方案生成和结果整合时，学习者的理由重建、信心校准与无辅助迁移是否发生变化？只有比较实验、过程材料和延迟测试，才能判断这种变化是否存在以及由什么条件触发。

（二）概念论证的三类证据

第一类是哲学与社会理论资源。异化、劳动过程、技术德性和延展心灵分别揭示活动与主体的关系、构想与执行的分离、技术中介中的实践智慧，以及认知边界的开放性。它们为问题提供坐标，却不能直接证明生成式 AI 对学生造成了某种后果。

第二类是认知与学习科学证据。认知卸载与合意困难研究能够说明外部工具、任务负荷和练习条件如何影响学习，但从一般工具推到生成式 AI，仍需要分析其适配性、生成性、黑箱性和互动连续性。第三类是教育现场证据，包括作业版本、对话记录、口头辩护、变式任务和延迟迁移。只有三类证据相互约束，主体性讨论才不会停留在宏大判断。

三、理论重建：主体性不是独占生产，而是可问责的判断能力

（一）异化并不预设一个历史上完整的起点

异化批判并不依赖一段可验证的“前异化黄金时代”；它可以分析社会关系如何使人的活动、产品和能力以异己力量出现。学生主体尚在形成，也不会自动使异化框架失效。AI 教育中的特殊问题在于，学习者可能尚未形成稳定的专业判断，关键构架活动就已被外部系统承担。因此，需要讨论的不只是既有能力被剥夺，还包括能力形成所需的练习、反馈与责任结构是否仍然存在。

（二）认知卸载不是失败，失配才是风险

Risko 与 Gilbert 将认知卸载概括为通过外部行动改变任务的信息加工要求、降低认知负荷。记笔记、使用计算器、检索数据库都属于卸载。卸载可以释放工作记忆，让学习者处理更高层的问题；把它与能力退化直接画等号是不成立的。风险取决于卸载对象、学习阶段、监督能力和迁移要求。专家把常规计算交给工具，可能提升判断；新手在尚未理解问题结构时外包建模，则可能失去形成内部模型的机会。

因此，关键不是“亲历亲为”本身，而是功能匹配：外包后剩余的监督能力是否足以发现错误？学习者是否知道工具何时不可靠？任务的最终责任是否超过其实际控制？当责任高于控制、控制高于理解，或自信高于可迁移能力时，才出现值得教育干预的失配。

（三）延展心灵反驳了封闭边界预设

Clark 与 Chalmers 的延展心灵论表明，外部资源在满足特定耦合条件时可以构成认知过程的一部分。延展认知恰恰质疑内外界线的先验地位，因此，AI 协作是否有害，不能由边界是否清楚单独决定。工具进入认知系统，既可能扩大行动能力，也可能在监督能力不足时放大错误；判断其性质，需要考察耦合方式、可纠正性与责任结构。

更稳健的判据是规范性的而非空间性的：外部系统是否稳定可用、其局限是否被理解、学习者能否在关键节点修正或拒绝、责任是否与实际决定权相匹配。由此，有益的“边界渗透”与有害的“边界溶解”可以被区分：前者扩大可行动能力，后者制造作者身份、控制能力和问责要求之间的裂缝。

（四）学科共同体可能防御，也可能重构

Becher 的“学术部落与领地”提供了理解学科文化的经典资源。教师制定 AI 规范，可能服务于公平、版权、隐私、学习目标和评价有效性，也可能在某些条件下以程序治理替代课程目标重审。当规范讨论持续增加，而“本学科究竟要求学生形成什么不可让渡的判断”长期缺席时，技术治理就可能成为学科回避自我重构的方式。这个命题需要教师访谈、政策文本变化和替代解释比较来检验。

四、核心模型：四种失配构成溶解性边界

溶解性边界不是一种神秘的主体状态，而是四种可以观察的失配同时累积。任何单项出现都不足以诊断；只有它们形成闭环，并导致迁移或问责失败，概念才有解释价值。

四种失配的因果顺序并非固定。系统先提供框架，可能造成决策权失配；学生因结果顺利而高估能力，形成归因失配；评价制度又把成果全部归给学生，产生责任承载失配；为了维持作者身份，学生可能事后编造理由，使真正的判断链更加不可见。反过来，口头辩护、版本比较和独立迁移测试能够打断这一循环。

五、复合情境：同一份优秀作业的两种解释

设想一名文学课学生先阅读《狂人日记》，再请 AI 生成若干解释路径，最后选择“注视”作为切入点并完成论文。答辩时，她不能解释为何该角度优于“吃人”主题或日记体形式。这个情境至少有两种解释。弱解释是她准备不足、表达紧张或尚未掌握叙事学；强解释才是关键判断被 AI 预构架，导致她无法重建理由链。仅凭一次答不上来，不能直接选择强解释。

判别需要追加测试：让她脱离原对话提出两个新框架；给出一条反例，看她是否能修正中心论点；要求她标注 AI 输出中哪一步改变了方向；一周后在另一篇文本上重复任务但不使用 AI。如果她能够完成这些任务，最初卡顿只是局部表现；如果她持续只能修改系统给出的选项、信心明显高于独立迁移成绩，控制—责任失配的解释才获得支持。

编程课同理。“代码能跑但说不清复杂度”可能来自口头表达、代码理解或 AI 代为生成的不同组合。应比较代码追踪、错误定位、变式任务与无辅助重写，而不是把“不能从头写出完整代码”设成唯一主体性标准。真实的软件工程本就依赖库、文档和协作；教育要测的是学生能否在风险相称的层级上理解和负责。

六、最强反驳：为什么不能把一切直觉判断都视为悬浮

成熟专家经常“先知道该怎么做，后说出理由”。直觉并不等于运气，它可能是长期训练压缩出的模式识别。不能即时言说不等于无法迁移，“直觉操作者”的成功也不能一概归为稳定环境中的偶然。可言说性只是判断能力的一个指标，不是全部。过度要求学生把每一步显性化，还可能破坏流畅技能，并奖励善于事后合理化的人。

所以，本文不要求所有判断都能被完整语言化，而要求证据三角互证：其一，行为上能在新情境中迁移；其二，校准上知道何时不确定并寻求复核；其三，责任上能在关键节点定位错误并修正。若学生说不出学术化理由，却能稳定迁移、准确校准并有效纠偏，就不应被判定为主体性受损。反之，漂亮的过程说明也不能替代实际能力。

另一个反驳来自分布式认知：未来的重要能力可能是组织多个人与模型，而非独自完成所有环节。本文接受这一点。策展型主体不需要垄断内容生产，但必须拥有任务级控制：能设定目标、比较方案、改变分工、识别系统性盲点，并承担与决定权相称的责任。主体性从“我亲手写了多少”转向“哪些决定必须经我授权，我凭什么授权”。

七、教学设计：把边界规范改造成判断基础设施

1. 先行立场。在调用 AI 前，学生用短文本写下初步问题、证据标准和最大不确定性。目的不是禁止改变，而是为改变建立可比较基线。

2. 显式授权。把任务拆成问题界定、方案生成、证据核验、取舍、表达五类环节；学生说明哪些交给 AI、哪些保留否决权以及风险理由。

3. 双轨比较。至少对一个关键节点同时保留自主方案与 AI 方案，比较各自证据、盲点和失败条件，防止首个完整框架垄断选择空间。

4. 口头辩护。教师追问关键转折、反例和替代前提，但不把口才等同于理解；允许学生借助图、代码执行或文本证据完成辩护。

5. 无辅助迁移。在低风险、短时的变式任务中撤去 AI，测量内部模型与迁移能力；结果用于反馈和再分配授权，而非单纯惩罚。

这套流程不适用于所有任务。格式转换、语言校对和低风险信息整理没有必要承受完整审计成本。它应集中于高杠杆节点：问题定义会改变结论、错误后果较大、学生正处于基础能力形成期、或课程目标明确要求独立迁移。教学设计的智慧不在于增加最多的步骤，而在于把摩擦放在真正塑造判断的地方。

八、研究议程与证伪条件

概念若要超过修辞，必须转化为竞争性假设。研究可采用交叉实验：同一学生分别完成无 AI、仅检索、AI 生成候选、AI 全程协作四种任务；记录任务质量、关键决策来源、信心、延迟迁移与纠错表现。再操纵先行立场、来源提示、模型解释和口头辩护，以判断哪些装置真正降低失配。

还需警惕测量反身性：要求学生写 AI 使用说明，可能只培养更娴熟的合规叙事；过程日志也可以被事后加工。因此，证据必须结合行为任务、版本记录、延迟测试与访谈，而不能只看自述。不同学科的“关键判断”也不同，文学解释、数学证明、临床判断和程序设计不应共享一套简单阈值。

九、结论：大学要保护的是否决权与责任能力

生成式 AI 不会因为进入思维流程就自动取消主体性。主体从来在语言、工具和共同体中形成。真正危险的是一种制度性错配：系统深度参与决定，学习者却看不见参与发生在哪里；学习者无法重建理由，却被视为充分理解；产出被全部归于学生，责任也被完整压回学生。此时，边界问题才从协作方式上升为教育问题。

“溶解性边界”的理论价值，不在于宣称发现一种前所未有的本体论危机，而在于把模糊焦虑转写成可检查的关系：决策权、理由链、能力归因与责任承载是否匹配。边界可以开放，认知可以分布，作者身份也可以协商；但关键判断必须保留可解释的否决权，学习者必须拥有与其责任相称的理解、校准和迁移能力。

因此，大学 AI 教育不应以“完全独立”作为怀旧目标，也不应以“高效协作”作为免责口号。更成熟的目标是培养可问责的协同主体：知道何时授权、凭什么采纳、何时撤回信任，以及在系统失误时如何重新接管。主体性不是不借助他者，而是在借助他者之后仍能对方向作出自己的判断。

最近邻加固：与"分布式认知"的判决性划界。 "溶解性边界"最贴身的近邻是哈钦斯（Hutchins，1995）的分布式认知——认知过程正

当地分布于人、工具与环境所构成的系统之中，而非局限于单个头脑。表面上，"人机认知边界变得流动"几乎就是分布式认知的一个说法，本文却主张真正要警惕的不是边界流动本身。两者的关键差别在于责任与控制是否仍然对齐。分布式认知描述的是认知在系统中的健康分布——飞机驾驶舱里认知分布于机组与仪表，但每个环节的控制权、可追溯性与责任始终对齐，谁负责什么是清楚的。而溶解性边界界定的恰恰是一种可判定的失配状态：关键认知环节的控制权、过程可追溯性与结果责任彼此脱钩，学习者被要求对最终产出负责，却无法重建判断链、校准信心或在无 AI 条件下迁移能力。判决性的分离线在于责任是否仍追踪控制：分布式认知预测，只要认知有效分布、系统整体运转良好即可，无需苛求单点独立;而本文预测相反——认知分布得再"高效",一旦出现决策权不清、理由链断裂、能力归因失准、责任不可履行这四项失配,系统就落入溶解性边界,因为教育情境里学习者必须最终独自承担责任,而分布式认知的健康分布恰恰不要求任何单个节点能独立重建全过程。分布式认知无法区分健康的边界开放与有害的责任失配——这个可判定的失配判据,是本文多出的承重件。

参考文献

[1] 马克思. 1844年经济学哲学手稿[M]//马克思恩格斯文集：第1卷. 北京：人民出版社，2009.

[2] Braverman, H. Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century[M]. New York: Monthly Review Press, 1974.

[3] Vallor, S. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting[M]. New York: Oxford University Press, 2016.

[4] Clark, A., & Chalmers, D. The extended mind[J]. Analysis, 1998, 58(1): 7–19. DOI: 10.1093/analys/58.1.7.

[5] Becher, T. Academic Tribes and Territories: Intellectual Enquiry and the Cultures of Disciplines[M]. Milton Keynes: SRHE & Open University Press, 1989.

[6] Trowler, P., Saunders, M., & Bamber, V. Tribes and Territories in the 21st Century: Rethinking the Significance of Disciplines in Higher Education[M]. London: Routledge, 2012.

[7] Bjork, E. L., & Bjork, R. A. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning[C]//Gernsbacher, M. A., Pew, R. W., Hough, L. M., & Pomerantz, J. R. (eds.). Psychology and the Real World. New York: Worth Publishers, 2011: 56–64.

[8] Risko, E. F., & Gilbert, S. J. Cognitive offloading[J]. Trends in Cognitive Sciences, 2016, 20(9): 676–688. DOI: 10.1016/j.tics.2016.07.002.

[9] Selwyn, N. Digital downsides: Exploring university students' negative engagements with digital technology[J]. Teaching in Higher Education, 2016, 21(8): 1006–1021. DOI: 10.1080/13562517.2016.1213229.

这一篇把「人机边界」的讨论从空间隐喻里拉了出来：需要警惕的不是一切认知外包，也不是边界本身变得流动，而是控制、理解与责任三者脱节——学习者在尚未形成稳定专业判断时，把问题界定、方案生成、证据筛选与表达整合连续交给系统，却仍被要求为最终产出负责。

「溶解性边界」由此被界定为一种可判定的教育风险：关键认知环节的控制权、过程可追溯性与结果责任彼此失配，以致学习者无法重建判断链、校准信心或在无 AI 条件下迁移能力。作者还明确借延展心灵反驳了封闭边界预设，并指出认知卸载不是失败、失配才是风险——这两句把本文与保守派的「守住边界」区分开了。

课程与作业设计　评估风险时用三问：控制权在谁、过程可否追溯、责任由谁承担。三者一致即安全，失配即风险，与使用比例无关。

学术规范　把「AI 使用占比」这类指标换成可追溯性要求（保留提示、候选与取舍理由），前者不可测且不相关，后者可查。

学生自查　撤掉工具后能否重建这条判断链？不能，就是本文所说的失配，而不是效率高。

管理与工程团队　同一判据适用于人机协作的生产流程：谁控制、谁能复现、谁签字，三者必须对齐。