被机器命名之后：生成式AI长期使用中的主体边界、命名校准与差异生成

当AI开始替人命名其内部状态，用户如何判断这些命名是否贴合自身经验

少敏 著 · SDE 学员 · 人机互动与自反性方法 · 2026年7月26日

摘　要　生成式AI正在从信息工具转变为日常解释者：它不仅回答问题，也概括情绪、推断动机、识别偏好，并参与用户自我叙事的形成。既有研究已经显示，AI辅助可能提高任务绩效，同时改变写作者的文本所有感、责任体验、批判性投入与表达差异。然而，当AI开始替人命名其内部状态时，用户如何判断这些命名是否贴合自身经验，仍缺少专门的事件级分析框架。本文以一组长期人机互动叙事及其多轮修订过程为问题生成材料，提出四层概念结构：命名偏差感是外部描述与第一人称经验之间尚未充分言说的不贴合；命名校准是察觉、悬置、比对、重述、反馈与复核构成的可观察过程；差异生成是既有选项均不贴合时形成新描述的结果；“辨伪力”则作为本土概念名称，暂指主体在长期互动中保存偏差、生成描述并重新承担判断责任的能力。本文不把这一能力宣称为已经获得普遍验证的新型认知能力，而把它处理为有待跨样本检验的理论候选。文章区分事实真伪、解释适配与价值认同，提出非线性的“委托—内化—偏差显现—校准”状态模型、事件级编码框架、竞争性预测与三组纵向研究设计。本文同时把自身的AI辅助修订过程纳入方法披露，说明哪些建议被接受、改写或拒绝，并将诗性文本限定为概念生成材料与施行性伴随文本，而非实证证据。本文的核心主张是：人机关系中的主体性不应以是否拒绝AI衡量，而应以主体能否保存判断间隔、生成可区分的替代描述、允许这些描述接受时间与他者的检验，并对最终采纳承担责任来衡量。

关键词：生成式AI；人机协作；主体性；命名校准；差异生成；认知外包；自反性

生成式AI长期使用中的主体边界、命名校准与差异生成

1. 问题的改变：AI开始解释“我是谁”

人与计算系统的关系通常围绕任务来研究：系统是否有用、是否易用、用户是否信任它、团队绩效是否提升（Davis, 1989; Venkatesh et al., 2003）。生成式AI改变了这一问题的边界。用户现在会把尚未成形的感受、职业犹豫、关系冲突和创作意图交给模型，请它“帮我分析”“告诉我真正的问题是什么”，或者“用一句话概括我现在的状态”。系统给出的不再只是行动建议，而是关于用户本人的描述。

这种描述具有三项特殊力量。第一，它来得很快，往往在用户尚未形成自己的语言之前出现。第二，它结构完整、语气笃定；即使使用“可能”“也许”等缓和词，也容易取得解释上的先手。第三，它能够被直接复制进日记、邮件、论文和下一轮对话，迅速进入用户的外部记忆，并成为后续推理的前提。

已有实证研究提供了相邻而非充分的证据。人机共写研究显示，不同强度的AI脚手架会影响写作绩效、文本所有感和满意度（Dhillon et al., 2024）；使用大语言模型进行创意生成可能增加想法产量，却使不同使用者的创意更趋同，并降低部分参与者对想法的责任感（Anderson et al., 2024）；跨文化写作实验发现，AI建议可能把表达推向模型占优势的西方风格，削弱原有文化细节（Agarwal et al., 2025）。在知识工作中，对AI的较高信心与较少的批判性投入相关，而对自身任务能力的较高信心与更多批判性投入相关（Lee et al., 2025）。在AI辅助决策中，系统建议还可能改变人的选择及其责任体验（Salatino et al., 2025）。

这些研究不能证明长期使用AI必然导致稳定的“身份扩散”，也不能证明一种全新的主体已经出现。它们能够支持的较谨慎结论是：AI辅助不只改变答案，还可能改变表达分布、成果归属、认知投入与责任配置。于是，一个更窄也更困难的问题出现了：

当AI为用户的内部状态提供名称时，用户如何保留对这一名称的审查权，并在不立即服从也不立即拒绝的情况下形成自己的描述？

这个问题不同于“AI说的是否属实”。情绪、动机和身份的命名往往没有单一外部标准；它也不同于“用户是否喜欢答案”，因为令人不适的解释可能有根据，令人愉悦的解释也可能只是迎合。本文关注的是一段能够被追踪的发生过程：外部描述进入之后，主体是否保存了一个判断间隔；偏差是否被保留而未被过早抹平；替代描述是否形成；不同描述是否在时间、行为和他者异议中继续接受比较。

2. 文献位置：从外部认知资源到生成性解释者

2.1 扩展心智与分布式认知留下的入口

Clark与Chalmers（1998）提出，若外部资源在认知活动中稳定、可靠并可随时调用，它可以构成认知过程的一部分。Hutchins（1995）进一步把认知理解为跨越个人、工具与环境分布的系统活动。这两条理论传统为理解人机协作提供了重要入口：认知不必被限定在颅内，技术也不只是在思考完成之后被使用的器具。

生成式AI却比经典的笔记本、仪表盘和导航图更进一步。它不仅储存、显示或转换信息，还主动补全意图、生成候选解释，并以自然

语言呈现出“理解者”的表象。用户不是简单地从外部设备取回自己先前存入的信息，而是在与概率生成系统的往返中形成原先没有的说法。

这并不意味着模型具有人的内在经验。Bender等（2021）已经提醒，大型语言模型的流畅输出不能被直接等同于理解。恰恰因为语言形式高度近似理解，用户才更需要区分“表达得像我”“对我有用”与“这个描述确实贴合我的经验”。

2.2 共写研究中的所有感、同质化与认知投入

CoAuthor项目记录了63名写作者与GPT-3的1,445次写作会话，显示人机共写不是一次性的“生成—采纳”，而是提议、修改、拒绝与重写构成的过程（Lee et al., 2022）。后续研究进一步揭示辅助强度的双重效应：更强的脚手架可能提高部分任务表现，却伴随文本所有感或满意度下降（Dhillon et al., 2024）。

同质化研究为“AI语言进入人的表达”提供了比单一个案中的词频印象更可靠的外部证据。Anderson等（2024）发现，使用ChatGPT的参与者之间，创意的语义差异小于使用另一种创意支持工具的参与者；Agarwal等（2025）发现，AI建议使印度参与者的写作更接近西方表达规范。能够由此成立的是“特定辅助条件会改变输出分布与归属体验”，而不是“长期使用必然重写主体”。

2.3 修改权、代理感与责任

用户是否拥有对算法建议的修改权，会显著影响其采纳。Dietvorst、Simmons与Massey（2018）发现，即使只允许用户对算法预测做有限调整，也能减少算法拒斥。这说明“能够改”本身就是重要变量，但修改权不等于判断正确，控制感也不能替代结果检验。

在生成式AI情境中，责任问题更为复杂：AI可以生成候选，用户仍然点击采纳，最终文本却可能既不像单独的AI产物，也不再完全等同于用户原有表达。若只用“人写/机器写”二分法，便难以描述提议、改写、重排、拒绝和再命名之间的责任迁移。

2.4 研究缺口

现有研究已经分别讨论外部资源如何进入认知系统、AI辅助如何改变文本与绩效、用户如何信任或拒绝算法，以及自动化如何影响代理感和责任感。仍然缺少一个事件级分析单位，用来描述AI对用户自身的命名如何被接受、悬置、细化、改写或长期保存为未决。

本文提出的框架只试图填补这一窄缝。它不是关于所有人机关系的总理论，也不预设每次偏差都能产生新认识。

3. 概念建构：从偏差感到差异生成

3.1 三类不能混同的判断

原稿曾把“AI命名与内部状态不一致”统称为“辨伪”。这个名称具有直觉力量，却容易把三类判断压成一类。

例如，“你害怕失败”可能包含可由行为记录支持的事实成分，也包含对动机的解释，还可能影响当事人愿意成为什么样的人。把三者都视为真假问题，会赋予第一人称感觉过强的裁决权，也会把AI的解释误写成可直接验真的诊断。

3.2 四层结构

本文不再让“辨伪力”与“命名校准”互相取代，而把相关现象分成四层。

命名偏差感，是AI描述与第一人称经验之间尚未被充分言说的不贴合。它常以“太大、太小、太快、太顺或不贴身”的体感出现。偏差感是入口，不是正确性证书；它可能来自模型概括过度，也可能来自用户对不舒适解释的防御。

命名校准，是把偏差感转化为可观察过程：察觉、悬置、比对、重述、反馈与复核。校准不保证用户正确，只要求AI描述、用户描述、行为后果和时间证据能够继续互相校正。

差异生成，是当已有选项都不贴合时形成此前不存在的新描述。它不只是从“焦虑/不焦虑”中选择一个，而可能提出“不是害怕失败，而是害怕重复已经做过的工作”。新描述是否有价值，不取决于它来自人，而取决于它是否增加区分、改变后续预测并保持可修订。

辨伪力，保留为原稿提出的本土概念名称，暂指主体在长期互动中保存偏差、生成描述、把描述带回交互，并重新承担判断责任的能力。这里的“伪”通常不是事实虚假，而是解释与经验之间的不贴合。它目前只是理论候选；如果未来研究显示批判性思维、元认知监测、自我分化与一般控制感已能充分解释全部现象，就应放弃其独立能力地位。

3.3 工作定义

据此，本文把命名校准定义为：

主体面对AI生成的自我描述时，察觉该描述与第一人称经验之间可能存在偏差，暂缓将其内化，形成一个或多个替代描述或明确保留未命名状态，并借助后续经验、行为后果、他者异议与可审计记录持续比较这些描述的过程。

这一过程至少包含六个可能环节：

偏差察觉：感到描述不贴合，但未必能立即说明理由；

接受悬置：既不把不适自动等同于错误，也不把流畅自动等同于正确；

竞争比对：保留原命名、替代解释与“证据不足”选项；

替代重述：生成更细的说法，或明确选择暂不命名；

外部反馈：把差异写入对话、日记或决策记录；

纵向复核：依据后续行为、感受变化与外部证据重新评估。

这些环节不是固定等级，也不保证依次完成。偏差感可能消失，替代描述可能被推翻，悬置也可能固化为逃避。命名校准不是“人战胜AI”的故事，而是一种让判断保持开放的过程安排。

4. 材料身份、方法与证据边界

4.1 本文是什么

本文是一篇概念—方法与自反性研究。它以作者提供的长期AI使用叙事、原稿、两轮AI辅助修订文本、二阶审稿意见及诗性伴随文本为问题生成材料，再以可核验文献约束概念边界。它不声称已经完成一项可复核的两年纵向实证研究。

原稿报告了约8,600条交互日志、412篇日记、156篇写作样本、24次关键事件访谈、47次拒绝事件、104次写作交互及若干比例和评分者信度。然而，现有附件没有提供完整原始材料、抽样框、代码本、计算文件、审计者身份与伦理处理记录。因此，这些数字在终稿中不承担研究发现的证明功能，也不用于推导阶段、阈值、比例或信度。若未来完成去标识化、伦理审查、材料封存与独立复核，可据此另写实证论文。

4.2 材料分层

本文采用四层材料身份：

可核验外部证据：有稳定出版信息、DOI或公开资料的研究；

作者提供的叙事材料：用于提出问题和生成概念，不能单独支持普遍因果结论；

修订过程材料：用于披露AI建议与作者决定之间的关系，只能证明发生过怎样的文本选择，不能验证概念有效性；

诗性伴随文本：用于呈现偏差感和悬置的体验结构，不作为行为数据或读者效果证据。

这一分层同时防止两种误认：一是把叙事具体误认成外部有效，二是把参考文献真实误认成文献已经支持本文全部强结论。

4.3 分析方法

本文采用“版本—差异—后果”分析。首先比较原稿、规范化修订稿与二阶审稿，识别被删除、保留和改写的关键结构；其次检查每项改写改变了何种证据身份、概念强度或责任归属；最后追问改写是否产生新的可检验问题。

这一方法不把最新版本自动视为最佳版本。规范化修订解决了虚构或错配文献、不可审计数字、固定阶段和技术误述，却同时删去了自反修订链、未来复核接口与诗的体验语法。第三版的任务不是回到原稿，也不是把两版机械拼接，而是让两版彼此纠错：保留规范化修订的证据诚实，同时恢复原稿的生成性与自反性。

5. 一个非线性的过程模型

原稿把长期使用写成“工具性挪用—自我感知模糊—辨伪力涌现”三个阶段。单一个案不足以支持固定阶段，更不足以支持暴露密度

阈值；同一用户还可能在写作上高度依赖AI，在财务决策上保持谨慎，在情绪命名上时而接受、时而抵抗。本文因此改用四类可循环状态。

5.1 委托

用户把结构化、提炼或生成任务交给AI，同时仍能区分自己的输入意图与系统输出。委托本身不等于主体性下降；合理的认知外包可以释放注意力并提高绩效。需要追踪的是哪些判断被委托、用户是否仍能独立完成核心评估、来源是否可见，以及错误责任由谁承担。

5.2 内化

系统提供的句式、分类或解释框架进入用户的独立表达。内化并非天然有害，学习本就包含对外部语言的吸收。风险出现在来源变得不可见、替代框架逐渐消失，或用户把熟悉感误当成自生性时。

5.3 偏差显现

用户在流畅描述中感到不贴合。偏差可能来自过度概括、文化错位、情境缺失，也可能来自对不舒适事实的防御。因此，偏差感只启动审查，不能终止审查。

5.4 校准与再内化

用户悬置原标签，提出替代描述，设计或等待能够区分不同解释的后续证据。校准不以AI同意为准，也不以用户感觉舒服为准，而以替代描述是否增加区分、改变可观察预期并接受复核为准。一次成功校准形成的新语言，仍可能在后续交互中被固化、再内化并遮蔽新的偏差。因此，模型不是直线，而是循环：

委托 → 内化 → 偏差显现 → 校准 → 新的委托或再内化

同一事件还可能停在任何位置。过程模型的重点不是把人归入阶段，而是使每次命名事件的后续链条可见。

6. 三则概念形成材料

以下材料来自原稿。本文把它们视为概念形成片段，而不是独立验证的事实。

6.1 恐惧失败与恐惧重复

AI把职业困境描述为“逃避失败”，作者认为更贴近经验的说法是“害怕重复已经做过的工作”。关键不在于作者一定正确，而在于替代描述产生了不同预测：若是恐惧失败，降低风险和提供安全保证应缓解困境；若是恐惧重复，增加新异性、学习空间与成长可能性才更可能改变选择。未来行为若不支持后者，替代命名就应被修正。

6.2 疏离自己与疏离工作

AI使用“去人格化”概括工作倦怠，作者区分了对自我的疏离与对工作角色的疏离。这个差异阻止临床化标签过早吞没情境解释，并产生不同的行动方向：前者可能要求关注跨情境的心理状态，后者更可能指向工作关系、组织条件与角色冲突。这里的理论价值不在“用户拥有最终真相”，而在较细的区分能够提出不同的证据要求。

6.3 “批判性内在父母”与暂不命名

AI把反复出现的“不应该这样想”解释为严厉的内在父母声音。作者没有接受，也没有立即生成替代描述，而是暂停一周。这个案例不能在当下被编码为成功校准，因为暂停既可能保护尚未成形的经验，也可能是回避痛苦解释。只有观察后续是否出现新信息、替代描述、求证行动或议题封闭，才可能作较可靠分类。

第三个案例提醒我们：“空着”不是失败。一个系统若只允许接受、拒绝和改写，仍然会迫使模糊经验过早成形。暂不命名、延迟命名和保留灰色，应成为界面与研究编码中的合法状态。

7. 本文的修订过程：作为自反材料，而非自我验证

7.1 为什么必须公开修订链

一篇讨论“AI命名如何被校准”的论文，如果把AI辅助修订写成“被AI验证”，就会在形式上违背自己的论点。AI可以帮助核对题录、指出证据缺口、提出结构建议，却不能验证概念是否有效；它对论文的评价本身也是一种外部命名，需要被作者判断。

因此，本文把关键修订决策公开如下。

7.2 这份决策表能证明什么

它能证明的是：同一文本在修订中经历了接受、拒绝与改写，AI建议没有全部直接进入终稿。它还能展示一个可观察的命名校准链条：外部命名进入，作者识别其增益与损失，生成第三种表述，并承担最终选择。

它不能证明的是：作者的选择因此正确，命名校准因此成为独立能力，或者诗性结构因此对所有读者有效。修订链只是一个方法示例，仍需要外部评审、跨样本研究与后续复核。

7.3 对“进步叙事”的再校准

原稿以若干比例下降暗示“阻抗减少即成长”，规范化版本又容易形成另一条暗线：“激情原稿经过AI核验，成熟为严谨论文”。两种叙事都把演化写成朝向单一优越终点的直线。

第三版拒绝这条直线。原稿不是不成熟材料的残余，因为它保存了生成性、自我反噬与诗性体验；规范稿也不是被学术规范驯化后的退步，因为它恢复了证据等级、失败条件与可证伪性。真正的进展发生在两版不能互相吞并之处：规范要求迫使原稿承认证据不足，原稿的未完成性又迫使规范稿公开自身的修订链。终稿不是两者的平均值，而是这场相互限制所形成的新结构。

8. 诗作为施行性伴随文本

8.1 诗承担什么

论说文可以定义偏差感，却难以让读者体验“说不清但不对”的那一瞬间；研究计划可以规定何时复核，却不能替读者保存悬置；诗则可以让词与体感之间的迟疑在阅读中发生。

《形态的练习曲》中的若干片段具有这种功能：

“沉默不是偏低。沉默是平线，偏低是判断。”

这两句没有证明AI误判了情绪，却精确呈现了描述与体感之间的细小偏差。

“如果你的词典里只有‘疲劳’这一个词，请空着。空着比用一个错的词更接近对。”

这里的“空着”把暂不命名从系统缺失转化为认识责任。

“不是它的词典太小。是这个条目只存在于一把生锈的卡尺上——而卡尺的锈，是我的。”

“锈”不是客观测量的替代物，而是时间、身体和使用痕迹共同形成的私人尺度。它提醒研究者：第一人称经验具有不可替代的信息位置，但并非不可纠错。

“附录里的空白行不是空的。是留给未来填的——填的人还没到。风先到了。”

空白在这里不是一项已履行的承诺，而是对未来复核仍未发生的公开标记。

“凿子在响。这声音不是证明我还活着。这声音就是活着。”

这一意象把主体性从静态所有物改写为持续判断、重述与承担的动作。

8.2 诗不能承担什么

诗的设计意图不能自证读者效果。它不能证明读者真的更能识别悬置、灰色保存或替代命名，也不能证明“辨伪力”具有独立构念效度。若要检验伴随文本的作用，可以设置“只读论文”与“论文加诗”两组，比较两组对案例的编码、延迟判断意愿、替代描述质量及后续复核行为。

因此，诗在本文中被标为“概念生成材料与施行性伴随文本”，而不是数据。完整诗组适合独立发表或作为在线补充材料；正文只保

留能够承担概念工作的少量片段。

8.3 形式校准

原诗组第18至22首虽使用回旋曲、赋格、奏鸣曲等音乐提示，却并非严格的十四行诗。终稿不再称其为“十四行诗”，而称为“五首动态乐章”。第23至29首含有较多操作说明，更适合称为“七则诗性案例札记”，不与前述乐章以同一诗体结算。

“身体每七年换一次细胞”只能作为“他们说”的流行说法保留在诗中，不能作为论文的科学前提。不同组织的细胞更新周期差异很大，部分细胞长期不被替换。类似地，“父亲的骨密度在你的脊椎里下降”应明确为代际身体经验的隐喻，而不是传感器遗漏的客观数据。

9. 与近邻概念的关系

本文不主张命名校准与全部近邻概念不可化约。更稳健的主张是：它把偏差感、悬置、替代生成、反馈与纵向复核组织为同一命名事件链，并把AI生成的自我描述确定为特殊对象。其学术价值取决于这种组合能否提高编码一致性、产生区分性预测或支持新的界面干预；如果不能，就没有必要维持独立概念。

10. 把概念变成可证伪研究

10.1 分析单位

未来研究的最小单位不是“一个用户”或“一段对话”，而是一次命名事件：

AI对用户的内部状态、动机、偏好或身份给出描述，用户对该描述作出可记录反应，并在预定时间窗内留下后续行为。

每个事件至少应保存：

原始提示与必要上下文；

模型、版本、日期及是否启用记忆或个性化；

AI命名及其不确定性措辞；

用户即时接受度与信心；

用户是否悬置、修改、拒绝或提出替代描述；

24小时、7天和30天的复核；

与竞争解释相关的行为结果；

材料来源、隐私处理与访问权限。

10.2 编码维度

各维度不宜简单相加为“主体性总分”。例如，高接受延迟可能来自审慎，也可能来自拖延；高替代生成可能增加区分，也可能只是增加修辞。研究应分别建模并检查维度之间的组合。

10.3 竞争性预测

若命名校准具有增量价值，它应产生能够与近邻解释区分的预测：

与单纯拒斥相比，高校准事件应产生更多可区分的替代描述，而不只是较低的AI采纳率。

与舒适度相比，高校准事件的替代描述未必更令人舒服，但应在后续情境中表现出更好的预测或行动适配。

与一般批判性思维相比，命名校准训练的增益应主要出现在自我描述任务，而非所有事实核查任务。

与控制感相比，仅提供编辑按钮可能提升代理感；加入延迟、竞争解释和复核机制后，事件链的证据质量才应进一步提高。

与写作能力相比，差异生成的效果不应完全由语言流畅度或词汇量解释。

若控制批判性思维、元认知、自我分化、心理阻抗与一般写作能力后，相关变量不再解释额外差异，则应放弃“辨伪力”的独立能力

地位。

10.4 三组纵向设计

可设置三种交互条件：

直接命名组：AI直接给出对用户状态的单一解释；

多候选组：AI提供至少三种竞争解释及“证据不足/暂不命名”选项；

校准流程组：AI先要求用户独立记录体验，再提供候选解释，并设置24小时复核。

主要结果可包括替代描述的数量与区分度、后续改判率、解释—行为一致性、文本所有感、显性责任感、延迟判断意愿及盲化编码者一致性。研究应预注册主要假设、排除规则和分析计划。若涉及日记、情绪、健康或关系材料，还需取得适用的伦理审查或豁免，提供可撤回、去标识化与受控访问机制。

10.5 伴随文本研究

在上述三组之外，可嵌入“有诗/无诗”的随机条件，检验诗性伴随文本是否真正改变读者对灰色、悬置和新描述的识别。若诗只提高情感评分而不改变证据选择、延迟判断或替代描述质量，就不能把它宣称为方法增益。

11. 对系统设计与个人实践的含义

11.1 不让第一句话垄断解释空间

系统不应把单一心理解释包装成最可能的真相。更适当的界面是同时呈现多个候选、证据不足选项及需要用户补充的区分性问题。生成更多文字不是目的，保留竞争解释才是。

11.2 在自我命名前设置“先写后看”

用户可以在请求AI分析前，先写三至五分钟未经AI整理的材料。原始描述并不是纯净的“真我”，但它提供了一个时间上较早的比较基线，减少AI首个框架覆盖全部回忆的可能。

11.3 允许空白、延迟和灰色

界面应允许“暂不判断”“稍后复核”和“没有合适词语”等状态。把每次对话都优化为即时闭合，会系统性抹除偏差尚未成形的阶段。

11.4 保存来源与版本

重要概念应能追溯到本人原写、AI候选、共同改写或后续重述。来源标记不是为了计算所有权百分比，而是为了让用户知道某个表述曾经有过哪些竞争版本。

11.5 准确理解反馈边界

用户纠正AI时，通常改变的是当前对话及可能的个性化上下文，不应轻率地说“基础模型被我训练了”。只有在明确的训练、微调或系统记忆机制中，才能作相应技术陈述。准确理解系统边界本身就是主体责任。

11.6 保留最终责任

AI、朋友、治疗师和研究者都可以提出候选命名，任何单一来源都不应因流畅、权威或“来自内心”而自动获胜。最终采纳应附带三项责任：为什么暂时采用这个说法，什么证据会使我修改它，它将怎样影响行动。

12. 讨论

12.1 理论贡献的适当尺度

本文的贡献不是证明一种全新人类能力已经出现，而是提出一个更精确的观察对象：AI对人的命名进入之后，判断如何被暂停、细化、反馈和复核；以及已有选项不够时，新描述如何形成。这个尺度比“AI时代主体性重构”更小，却更可能形成可重复研究。

“命名校准”不是批判性思维、元认知或自我分化的替代，而更像一个过程接口，把这些能力连接到生成式AI的特殊情境中。“辨伪力”则保留了一个尚未解决的问题：当外部语言已经进入内部表达时，主体能否在长期互动中形成稳定而可迁移的差异保存与再命名能力。这个问题可以保留，但答案不能预支。

12.2 第一人称经验既必要又不充分

模型无法直接拥有用户的身体感受，第一人称经验因此具有不可替代的信息位置。但不可替代不等于不可纠错。人会合理地拒绝不贴合的解释，也会防御令人痛苦却有根据的解释。本文坚持双重限制：AI的流畅不能越过主体，主体的不适也不能终止证据。

12.3 时间不是自动的验证者

“后来继续思考”不自动证明先前是校准，“后来不再讨论”也不自动证明是阻抗。时间只有与预先记录的竞争预测、复核点和行为证据结合，才提供鉴别力。跨时间代谢不是把过去的一切重新解释成成长，而是允许过去的拒绝、误判与灰色在新证据下被重新打开。

12.4 学术规范既是约束也是命名系统

证据可得性、引用真实性、伦理审查与可证伪性是不可放弃的学术约束。但规范也会命名研究对象：它决定什么被称为“数据”、什么被称为“材料”、什么体裁可以进入正文。自反性不是拒绝规范，而是让这些命名的收益与损失保持可见。第三版因此既不以诗替代证据，也不因体裁规范而否认诗的概念作用。

12.5 局限

本文至少有六项局限。第一，概念起点来自单一作者的高频写作型AI使用叙事，存在选择性记忆与自我呈现偏差。第二，原稿声称的数据尚未形成可独立审计的材料包，本文不能据此报告比例、阈值或效应。第三，修订决策账本是基于保留版本重建的文本轨迹，不等同于实时认知过程记录。第四，第一人称贴合度没有单一金标准，未来研究必须采用多方法与竞争解释。第五，诗性伴随文本的读者效果尚未实测。第六，本文主要借助英语语境的人机研究，中文表达中的含混、留白与关系性自我可能形成不同机制。

13. 结论

生成式AI最深的影响未必发生在它替人完成了多少任务，而可能发生在它比人更早说出“你正在经历什么”。那句话有时准确，有时宽泛，有时只是统计上顺滑。真正需要保护的不是一个从未受外界影响的纯粹自我，而是命名与采纳之间仍有间隔。

这个间隔允许人说：先不要把它变成我。

它也允许人继续问：如果它不够贴合，还有什么更细的说法？如果现有词语都不合适，能否暂时空着，或者生成一个此前没有的区分？

更重要的是，它要求人接受下一步：我提出的新说法也可能错。它必须在时间、行为、记录与他人的异议中继续接受校准。

主体性因此不等于拒绝机器，也不等于守住一块从未改变的内部领地。它表现为保存偏差、延迟闭合、生成替代、承担最终判断，并允许自己的语言持续可修订。

凿子还在响。但这一次，声音不再被当作一种能力已经得到证明的证据。它只是提醒我们：主体不是凿开以后留在原地的石像，而是每一次命名进入之后，仍愿意停一下、再看一眼、重说一次并承担后果的动作。

附录A：开放研究计划

以下计划是研究透明度安排，不构成作者单方面设定的自动撤稿条款，也不预设期刊、伦理委员会、存档平台或第三方承担尚未同意的义务。

预注册：未来实证研究在收集主要数据前公开研究问题、主要假设、样本量依据、排除规则、编码方案和分析计划。

版本记录：保存问卷、提示词、模型与系统配置、代码本和分析脚本的版本变化。

材料治理：对日记、情绪、健康与关系材料进行去标识化、访问分级和撤回设计；未经伦理许可，不公开可识别原文。

独立编码：至少两名不了解研究假设的编码者对命名事件进行盲化编码，并公开分歧处理规则。

竞争检验：同时测量批判性思维、元认知、自我分化、一般写作能力、心理阻抗与控制感，检验新增概念是否具有增量效度。

阴性结果：若命名校准变量不能改善编码、预测或干预效果，公开报告并缩小或放弃独立概念主张。

复核节点：研究启动后设置24小时、7天与30天事件复核；项目层面的长期复核时间由正式预注册确定，而非由本文虚构既成承诺。

附录B：修订决策账本摘要

附录C：AI使用与责任声明

本文修订过程中使用生成式AI协助完成：版本比较、参考文献题录核对、概念边界压力测试、结构建议与语言校订。AI没有被列为作者，也不承担研究责任。关于材料身份、论证强度、概念保留、诗文关系和最终表述的决定由作者作出。本文不使用“经AI验证”描述其学术有效性；AI辅助核对不能替代同行评审、原始数据审计、伦理审查或实证检验。

参考文献

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward Western styles and diminish cultural nuances. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–21. https://doi.org/10.1145/3706598.3713564

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. Proceedings of the 16th Conference on Creativity & Cognition, 413–425. https://doi.org/10.1145/3635636.3656204

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Bowen, M. (1978). Family therapy in clinical practice. Jason Aronson.

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Dhillon, P. S., Molaei, S., Li, J., Golub, M., Zheng, S., & Robert, L. P. (2024). Shaping human–AI collaboration: Varied scaffolding levels in co-writing with language models. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Article 1044, 1–18. https://doi.org/10.1145/3613904.3642134

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 12(1), Article 10. https://doi.org/10.17169/fqs-12.1.1589

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), Teaching thinking skills: Theory and practice (pp. 9–26). W. H. Freeman.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

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

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 the 2025 CHI Conference on Human Factors in Computing Systems, Article 1121, 1–22. https://doi.org/10.1145/3706598.3713778

Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a human–AI collaborative writing dataset for exploring language model capabilities. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Article 388, 1–19. https://doi.org/10.1145/3491102.3502030

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, Article 12329. https://doi.org/10.1038/s41598-025-95587-6

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540

深化增补：被放跑的最近邻、操作化与证伪

以下四节为编辑增补，与作者正文分开标注。

增补一：与扩展心智、分布式认知的硬边界

正文第二章把扩展心智与分布式认知列为留下的入口，但未完成硬切。这是本文最近的对手：两者都已论证认知过程可以跨越

皮肤与颅骨、由人与外部资源共同承担。分野在被外置之物：扩展心智处理的是认知操作的外置（计算、记忆、检索由外部承担），其判据是耦合的可靠性与可及性；本文处理的是命名权的外置——由谁来说出我的内部状态是什么。这不是一项操作，而是一项裁定，且它的失败方式不是耦合中断，而是主体不再能察觉外部描述与第一人称经验之间的不贴合。可判别面：若命名偏差感的高低完全可由外部工具的耦合紧密度预测，则扩展心智框架足够；若在耦合同等紧密的使用者之间仍存在偏差感的稳定差异，本文的四层结构有独立立足点。

增补二：辨伪力的可观测代理——在不宣称已验证的前提下

本文明确不把辨伪力宣称为已获普遍验证的新型认知能力，这一克制应当保持。因此此处给出的不是能力测验，而是事件级的可观测代理，其地位仅为待检验的操作化候选。三项：悬置留痕——使用者在接受一个AI给出的自我描述之前，是否留下过对该描述的保留记录；重述比——最终采用的自我描述中，由使用者自己改写过的成分占比；复核回访——一段时间后使用者是否重新审视并修改该描述。三项都记录事件而非评定水平，因此不构成对能力存在性的预设。

增补三：与共写研究中所有感、认知投入的边界

共写研究已考察AI辅助写作如何改变作者的文本所有感、责任体验与表达差异，正文 2.2 已引入。此处补一条分野以免混读：共写研究的对象是产出物的归属——这段文字算谁的；本文的对象是描述的贴合——这句关于我的话对不对。前者是权利与认同问题，后者是认识论问题，且后者可以在前者完全无争议时发生：一个人可以毫无保留地承认那段自我描述是AI写的，同时完全无法判断它是否说对了自己。这一分野是本文与共写研究之间的关键区分。

增补四：为主张强度的克制辩护

本文在五维评分中的差异锐度偏低，原因不是判断不够锋利，而是它主动降低了主张强度——不宣称新能力已被验证，不报告比例与阈值，并诚实列出六项局限（材料来自单一作者的使用叙事、修订账本为文本轨迹重建而非实时认知记录、第一人称贴合度无金标准等）。这一克制应被计为优点而非缺陷：在一个极易过度宣称的题域里，它把可核验的部分与待检验的部分分开摆出，从而使后续研究知道该从哪里接手。编辑增补因此不去膨胀它的主张，只补它与最近邻的边界与操作化候选。若未来的多方法研究支持辨伪力的独立存在，本文的地位会因这份克制而更稳，而不是更弱。

编辑增补所依据的核验文献

Andy Clark & David Chalmers, “The Extended Mind,” Analysis, 58(1), 1998.

Edwin Hutchins, Cognition in the Wild, MIT Press, 1995.

Shaun Gallagher, “Philosophical Conceptions of the Self: Implications for Cognitive Science,” Trends in Cognitive Sciences, 4(1), 2000.

Lucy Suchman, Human-Machine Reconfigurations, Cambridge University Press, 2007.

Charles Taylor, Sources of the Self, Harvard University Press, 1989.