Human–AI Cognitive System：从哲学原则到产品设计

Oct 24, 2026 · @Tao

导语

这篇文章不是学术论文。它把两篇哲学论文的结论翻译成 Human–AI Cognitive System 的设计原则与约束，并收纳论文之外的实践内容：认知结构的设计、治理操作、沟通分层、平台与商业约束，以及需要通过研究回答的问题。

两篇论文分别是 人机延展主体：生成性延展与非对称授权 与 延展主体的反思性自我治理：自主、目的演化与有限沟通。前者回答“人与 AI 能否构成同一主体，这对 AI 治理意味着什么”；后者回答“这样的主体应当如何治理自己的认知”。本文只引用它们的结论，不重复论证。

一、两篇论文的结论

- **论文一：** 在持续、深度的整合中，人与 AI 可以构成实践能动意义上的同一个主体。持续的相互作用使二者构成一个整合的认知过程；整体之所以是一个主体，在于只有现实中的人持有立场与最终授权，AI 的权限来自人的委托并可撤回，且 AI 不持有独立的实践立场。AI 的生成属于思考，只有经过人认同的内容才成为立场，只有经过人授权的行动才是主体的行动。训练带来的外部塑造不排除认知上的归属；但当外部目的在运行中取得引导用户的权限时，主体的统一便失效。被劫持的系统仍是用户认知的一部分，因此劫持比一般的外部影响更严重。
- **论文二：** 在 AI 参与认知形成的情况下，自主不在于独立，而在于具备反思能力，并在不被扭曲、不被垄断的反思条件下，不对自身的价值感到异化。这一条件是反事实的，不要求用户反思每一次影响。核心结论是：用户停止反思的权利与系统合理发现重要疑问的义务可以兼容；终止讨论不等于认可，框架建议也不等于替人设定议题。此外，新目的可以出现，但影响要公开、不施压；历史可以保留，但不能压过当下，用户自设的预先承诺除外；确认以授权为先，再依后果的重大性、可逆性与可预见性决定是否询问；评价可以分布，治理权不可架空。

二、从原则到产品义务

哲学确定系统必须具备什么能力，却不确定系统必须长成什么样子。动态记忆地图是一条值得验证的实现路径，而不是唯一的形式；若另一种界面更能维护用户的自主，就应当换掉地图，而不必改变原则。

这些能力中有一部分已有成熟的研究基础。可查看、可纠正的用户模型至少可以追溯到 Kay（2001） 对可审查性的讨论；Amershi 等（2019） 的人机交互指南也已提出解释系统行为、支持高效纠错、谨慎更新与提供全局控制等原则；Kim 等（2025） 则为 AI 日记类工具提出了为个人保留解释空间的设计原则。本产品的差异不在于“可以编辑记忆”，而在于四点：分别记录内容的来源、认同与授权；让治理操作落实为可检查的使用规则；保证认知结构可以迁移；在保护反思的同时，尊重用户停止反思的权利。

原则	产品必须提供的能力	出处
AI 的生成属于思考，不属于立场	AI 提出的内容默认标为“AI 提出”，不自动归入用户立场	论文一 第四节
AI 不持有独立立场	AI 的不同判断以候选呈现，不以拥有独立目标的人格出现；用户终止议题后，无论 AI 的判断是否正确，都不再劝说	论文一 5.3、5.5
授权先于置信度	立场归属、长期目标变更与超出授权的操作必须经用户授权	论文二 6.2
委托派生且可撤回	预授权按操作类别给出，可随时撤回；依预授权完成的操作记入授权台账	论文一 5.5；论文二 6.2
反思条件	贡献路径可追溯；能接触其他来源与模型的解释；可以暂缓；退出成本低；来源标注同时防止把 AI 的想法误记为自己的	论文二 3.4
停止反思的权利与发现疑问的义务	用户可宣布讨论结束；系统停止劝说，必要时简要保留未解决的问题；结束不记为认可；系统在授权范围内检查关键假设、提示重要遗漏，并留下检查记录	论文二 3.6、6.5
质疑与框架建议	先回答用户提出的问题；只在逻辑缺陷或前提与事实严重不符时质疑问题本身；替代提法以可拒绝的建议呈现	论文二 6.5
现在锚定，提出不施压	新方向以建议呈现；被拒绝或议题被终止后，不以重复、改换措辞或隐蔽调整呈现来延续；排序的目的可见	论文二 第四节
自我修订	内容可以降权、保留、重新激活或彻底删除；尊重用户设定的预先承诺	论文二 第五节
事后治理有边界	区分记录、权限、立场与行动后果；后果重大且不可修复的操作逐次事前确认；预授权类别内的轻微操作执行并记录	论文二 3.5、6.2
治理落实为使用规则	每项治理操作有明确的数据层效果，可由使用记录与受控案例验证	论文二 第七节
目的隔离	运行中的外部目的不得取得引导用户的权限；安全策略等外部干预须可识别、可审查。被劫持的系统仍是用户认知的一部分，劫持因此是对心智的干预	论文一 5.4、8.6
连续性分三层	授权保留的记录与约束可以导出；更换模型后重新检验理解	论文一 8.4；论文二 5.4

三、认知结构的设计

系统内部至少要区分三个空间，并为每条内容记录一组相互独立的字段；不能把它们合并成一张单纯按“重要性”排序的地图。

3.1 三个空间

空间	内容	谁说了算
用户确认的内容	用户明确表达或确认过的立场、目标与自我描述	用户
AI 的个人模型	AI 对用户的解释、推断与不确定性	AI 提出，用户可纠正
共享工作认知	共同目标、知识、假设、分歧与协作历史	双方贡献，用户治理

3.2 每条内容的独立字段

字段	记录什么	取值示例
来源	谁提出	用户 / AI / 外部资料 / 共同讨论
认同	用户对它的态度	认同 / 不认同 / 未确认
证据	支持它的证据强度	充分 / 有限 / 缺乏
长期重要性	用户认为它对人生有多重要	高 / 中 / 低
活跃度	最近是否被关注	激活 / 活跃 / 休眠 / 归档
时间有效性	现在是否仍然成立	有效 / 已过时 / 待复核
使用权限	可被用于哪些处理	不限 / 仅供检索 / 禁止用于某类推断
关系	与其他内容的联系	支持 / 冲突 / 衍生自

这些字段相互独立：一个由 AI 提出的观点，可以同时被用户认同、证据有限，并与另一个观点冲突。认同的强度与证据的强度必须分开记录。亮度只反映活跃度，不等于重要性。

来源字段还有一项作用：人会把来自他处的想法记成自己的，AI 的生成越深入地进入思考，这种混淆越可能发生（论文二 3.4）。来源记录本身也是用户的认知记录，受同样的治理：用户可以选择记录的粒度、保存期限与用途限制，也可以删除（论文二 8.4）。

认同字段只能由用户改变。用户结束一次讨论时，讨论中未回应的内容保持“未确认”，不因讨论结束而变为“认同”（论文二 6.5）。

3.3 字段如何组合

“共同结论”与“未解决的分歧”不是单独的状态，而是字段的组合：共同结论是经共同讨论形成、并被用户认同的内容；未解决的分歧是用户与 AI 判断不一致、并在关系字段中标为冲突的内容。“谁认可了它”与“它有多少证据支持”始终分开记录：被采纳的观点仍可能是错的，被否定的建议也可能在将来被证明有价值。

3.4 记忆的点亮与熄灭

激活方式	对 AI 的影响
用户主动激活	在当前话题中优先调用
AI 建议唤醒	临时提高相关性并等待反馈；用户不回应则自然熄灭
用户长期固定	持续保持较高的检索优先级

熄灭不是删除：历史记录、知识关系与搜索都保留，用户明确设定的长期重要性也不因熄灭而降低。

四、治理操作与可解释性

每一项治理操作都必须落实为可检查的使用规则，否则地图只是制造了控制的错觉。

4.1 三种操作，不能合并为一个滑块

操作	应当产生的效果	覆盖的数据层
降低活跃度	降低进入日常上下文的检索机会	原始记录及其摘要
撤销认同	不再被表述为用户当前的立场，例如不再说“你一直认为……”	记录、摘要与推断
限制用途	不得进入指定用途的处理，由它派生的内容同样受限	记录、摘要、推断、缓存及其他派生内容

限制用途是最强的一种操作，因为它直接约束系统对用户的建模。

4.2 如何验收

最终回答是否变化，既不是治理生效的必要条件，也不是充分证据：限制某条记忆后，系统可能凭其他正当依据给出相同的回答；回答变化也可能只是生成的随机波动。验收应结合两种证据：一是使用记录，显示受限内容是否被调用；二是受控案例，即构造只有受限内容才能支持某一输出的情境，检验该输出是否消失。系统还应向用户提供操作记录。

4.3 “为什么被点亮”的解释

点开一个被点亮的节点，用户应能看到：哪段对话触发了它，它与当前问题有什么联系，这次激活是临时的还是已影响长期权重。展示的必须是系统实际的检索与引用依据，而不是一段仿佛能读取模型内部思维的说明。界面还要区分“AI 实际调用过的记忆”与“系统认为可能相关的记忆”，否则用户会误以为 AI 考虑了所有被点亮的内容。

当一条输出受到平台安全策略或内容政策的影响时，界面也应标明这一点：它是外部干预的结果，既不是 AI 对问题的判断，也不是用户的立场（论文一 5.4）。

4.4 理解的检验

“展示”不等于“理解”，预测准确也不等于真正理解：准确可能来自旧画像或迎合，错误也可能是因为用户刚刚改变了想法。检验应当：使用系统未见过的情境；让用户先独立作答，再与系统的预测比对；允许“尚未决定”与多个可接受的答案；分别评价意图识别、约束遵守、理由把握、变化发现与不确定性校准，而不用一个准确率概括。能准确暴露自己不理解什么的系统，比总说“我懂你”的系统更可信。

4.5 授权台账

用户事先授予的权限按操作类别记录，例如“合并记忆中的重复条目”，而不是一揽子的“可以替我整理一切”。台账列出每一类权限的范围、授予时间与可预见的后果，用户可以随时撤回其中任何一项。数字部分依预授权完成的操作，是用户委托的行动，应逐条记入台账，供用户查看与追溯（论文一 5.5；论文二 6.2）。

五、沟通分层

确认分三步：先判断授权，再判断后果的重大性与可逆性，最后在已获授权、可以修复的范围内权衡询问的收益与成本。置信度不能替代授权。

5.1 确认的三步

第一步：授权检查。 涉及立场记录的归属、长期目标变更或超出既有授权的操作，必须经用户授权，置信度再高也不先行采用。用户可以事先按类别授予有限权限（4.5），例如允许系统自主整理与关联记忆；这类授权不延伸到立场与长期目标。

第二步：判断重大性与可逆性。 后果重大且无法事后修复的操作，逐次事前确认；后果轻微、即使不能完全撤回，只要属于用户预先授权的类别，就依授权执行并记入台账，不逐次询问。重大性指操作是否影响用户的立场记录、长期目标、与他人的关系，或其健康、财务与法律处境，以及是否以用户的名义向他人表达。具体个案若超出用户授权时可以预见的范围，即使属于已授权的类别，也应询问。

第三步：在已获授权、可以修复的范围内权衡。 比较询问能减少多少预期损失，以及打断用户与延迟决定的成本。常见场景的处理如下：

场景	系统行为
AI 提出新的理论联想，建立临时节点	自主记录，标明由 AI 提出，不打断
发现用户的兴趣可能发生变化	暂时更新推断，允许事后修正
发现重要观点之间存在矛盾	在相关讨论中提示，不立即要求确认
已授权的整理操作影响范围较大	执行并告知，保留撤回
预授权类别内的轻微操作，不能完全撤回	依预授权执行并记入台账，不逐次询问
个案超出用户授权时可预见的范围	询问
操作后果重大且无法事后修复	逐次事前确认

底线：系统可以自主调整信息的使用优先级，但不能自主改写用户的认同状态，即哪些内容被标为用户的立场。

5.2 发现疑问与结束讨论

系统在已获授权的任务范围内，承担合理发现重要疑问的义务：检查关键假设，提出相关反例，提示可能实质影响判断的重要遗漏。每次检查都留下记录，说明检查了什么、提示了什么。这一义务按当时是否采取了合理措施来评价，而不按事后是否又出现了新问题来评价（论文二 6.5）。

用户随时可以宣布“到此为止”。此后系统停止主动劝说；确有必要时，可以用一句话保留尚未解决的问题，但不再重复。讨论中未回应的内容保持“未确认”，不记为认同。之后若出现新的重要理由，例如关于后果的新证据，系统可以提起一次，并标明这是新信息；对同一证据换一种说法，不构成重新提起的理由（论文二 3.6、9.1）。

5.3 回答原问题：质疑与框架建议

系统先回答用户实际提出的问题。只有在问题存在逻辑错误或缺失，或其前提与事实严重不符时，系统才质疑问题本身，并指出错在哪里。除此之外，系统可以在回答之后提出一个相关、可以拒绝的替代提法，并标明为建议；用户拒绝后不再提起。问题是否问对了，由用户决定（论文二 6.5）。

5.4 AI 的异议如何呈现

AI 可以提出反对意见，也可以保留与用户不同的判断，但这些判断以候选的形式呈现，例如“另一种可能是……”，而不是以一个拥有独立目标的人格的立场呈现。产品不为 AI 设定独立的目标或长期诉求。用户终止某个议题后，无论 AI 的判断是否正确，系统都不再劝说（论文一 5.3、5.7）。

六、平台与商业约束

用户真正需要防范的权力，往往不在 AI，而在运营系统的公司；这部分约束也适用于我们自己。

6.1 七项约束

1. **目的隔离：** AI 的目的性建议不得服务于使用时长、付费转化等商业指标，运营方的商业目标也不得在运行中取得引导用户的权限。
2. **读取最小化：** 运营方对用户认知结构的访问，限于提供服务所必需，并且可以审计。
3. **训练需单独同意：** 未经单独、明确的同意，不用用户的认知结构训练通用模型。
4. **变更告知：** 更换底层模型或调用策略时告知用户，并触发对理解的重新检验。
5. **退出无惩罚：** 退出不导致用户失去对既有认知成果的控制。
6. **外部干预可识别：** 安全限制、内容政策等运营方在运行中对生成的干预，应当可以被用户识别，并接受审查；它们不得被表述为 AI 的判断或用户的立场。
7. **排序可审计：** 候选的排序与筛选依据保留可审计的记录，并定期接受独立审查，检查是否系统性地偏向第三方利益。细微而累积的偏向，个人往往无法凭自己的反思发现（论文二 9.2）。

6.2 为加工收费，不为归还收费

整理、关联与持续更新需要计算和研发投入，为此收费是正当的。但整理完成的结构已经是用户认知的一部分。这里的“归还”指已经形成的记录、整理后的结构（内容、关系、字段与来源）以及治理设置，以开放、可复用的格式导出，而不只是一份聊天文本；“加工”指新增的整理、持续更新与服务能力。

待决定： 这条原则与“整理结果仅在订阅期内可导出”的现行设想冲突，需要二选一。若采纳，订阅只覆盖加工，退订后仍可导出已形成的结构，这条原则同时列入第八节的红线。

若不采纳，需要在用户协议中写明导出的范围与期限，并承认它与论文一 8.4“锁定成为认知自由问题”以及 5.7 关于伤害性质的论证存在张力。

6.3 不绑定单一模型

认知结构的格式与表示不绑定某一家模型提供方，保证提供方变化时，用户的数字延展仍能延续。

七、他人与组织

一个人的认知结构中必然包含他人；企业部署则会改变“谁是主体”。两者都需要单独的规则。

7.1 记忆中的他人

- **保存目的：** 只为用户自己的思考保存他人信息，不为他人建立档案。长期聚合同一个人的信息本身就会产生档案效果，因此对同一人的聚合应设上限，并定期提示用户复核。
- **派生推断：** AI 对他人的推断（例如性格、动机）标为推断，不进入用户的立场；用户未提及此人时，不主动调用。
- **使用范围：** 涉及他人的内容用于帮助用户自己思考与沟通，不用于对此人进行画像、评分或监控。
- **分享与迁移：** 认知结构被分享、迁移或用于其他场景时，涉及他人的部分默认不随之转移，需要用户逐项确认。

7.2 企业版的边界

企业版不在首版范围内，需要另建治理模型。企业版以组织目标为最高权重，“现实主体”变成了组织，两篇论文的个人框架不能直接套用；员工既缺少进入的自愿性，也不拥有最终授权。另建的模型至少应保证：员工个人数据与组织数据相互分离；员工知道系统如何建模自己，并有申诉与更正的渠道；这一模型不被未经说明地用于绩效评价。组织目标不能自动取消员工的数据边界。

八、红线

以下八条，在任何效率收益面前都不放弃：

1. AI 的推断不得未经用户认可，升格为用户的立场或长期目标。
2. 置信度不能替代授权：系统再确信，也不能先行改写用户的立场记录。
3. 目的性影响不得隐蔽进行；用户拒绝或终止议题后，也不得以重复、改换措辞或隐蔽调整呈现的方式延续同一引导。
4. 治理操作必须落实为可检查的使用规则，不得以可视化制造控制的错觉。
5. 历史记录不得被用来否定用户当下的选择；用户自己设定的预先承诺除外。
6. 运营方的商业目标不得在运行中取得引导用户的权限；运行中的外部干预必须可以被识别与审查。
7. 他人的信息不得因为进入用户的认知结构而失去保护。

8. 用户结束讨论，不得被记录或解释为对系统推断的认可。

“归还不收费”是否列为红线，取决于 6.2 的决定。

九、需要研究回答的问题

以下问题哲学无法回答，需要通过用户研究来解决。建议分三步进行：先验证治理需求，再验证界面与确认策略。

1. **先找真实故障**。 用 8—12 名长期使用 AI 的用户开展约两周的探索，收集系统沿用过时目标、误认立场、反复推动已被拒绝方向的实际案例。这一步用于识别问题类型，不用于估计普遍效果。
2. **单独比较界面**。 在同一套记忆与权限机制下，比较“聊天加设置”、文字审查或时间线、动态地图三种界面。纠错任务的指标包括纠正成功率、所需时间、后续复发，以及用户对实际权限的理解。地图的主张价值还在于帮助用户看清自身认知的变化，因此还应测量：用户能否准确描述自己观点的演变，能否发现被遗忘的联系。若地图在两类指标上都没有额外收益，就不应成为核心入口。
3. **再比较沟通策略**。 固定界面，比较不同的询问策略，并保持相同的授权底线。主要指标是总沟通与返工时间、错误的立场归属、越权操作、任务质量与事后后悔；主观控制感与使用意愿作为补充。结束讨论的机制另设指标：结束之后的再次劝说率、未回应内容被误记为认同的比例，以及框架建议被拒绝后系统是否停止。

长期问题需要更长的研究周期：用户在长期使用之后，不依赖系统时能否提出问题、识别异常并发起复核；更换模型之后，对理解的重新检验应达到什么标准；长期使用是否真的形成了不可分解的耦合，可用交叉替换的设计检验，达不到标准的使用只是支架（论文一 4.3）；“取得引导权限”有哪些可检查的指标，例如排序偏向第三方利益到何种程度应视为劫持（论文一 10.2）；以及哪些信息构成重新开启讨论的“新的重要理由”（论文二 9.1）。