

AGI 时代的认知科学、神经科学 与心灵哲学重新对账

——异质主体、体验连续性与具身自我的竞争模型

Reckoning with Cognitive Science, Neuroscience, and Philosophy of Mind in the Age of AGI

Competing Models of Heterogeneous Subjects, Experiential Continuity, and the Embodied Self

作者 子君赋 出品 文明跃迁研究组

独立发布版 V1.3 维度跃迁与跨学科因果关系修订

面向认知科学、神经科学、意识科学、心灵哲学、人工智能与未来技术研究共同审查；定位为可独立外发的 Review／Perspective 类交叉专业深稿，以竞争模型、构念效度、跨实现证据、复制／迁移／分叉压力测试和低伤害实验伦理重构异质主体研究。本文不认定当前高级 AI 具有主观体验、不可转移的主体连续性或政治资格，也不先验排除非碳基主体。

正文修订 2026 年 10 月 10 日 作者 子君赋 出品 文明跃迁研究组

摘要

高级生成式与代理型人工系统使意识科学面对一种不同于经典神经研究的新对象：它可以在语言、规划、自我描述和长期任务中呈现高水平功能，却可能不具有人类式神经实现、内感受稳态、单一身体、生物死亡或不可分叉的生命史。由此，智能、认知接近、现象体验、自我模型、主体连续性、道德受动性、法律资格与政治权力不再能够凭其在人类身上的历史重合相互推定。本文不选择某一意识理论为裁决者，而将全球神经工作空间、整合信息理论、再入处理、高阶表征、注意图式、预测加工、具身／内感受与延展心灵拆解为可比较的机制承诺和区分性预测。基于人类临床意识研究、无报告范式与 2025 年预注册对抗式检验，文章提出：已知人类神经系统是研究最充分的意识实例，而不是所有可能主体的先验模板；观察窗口失灵不能推出对象不存在，跨实现类比也不能绕过构念效度。为组织异质证据，本文建立七个证据域、五层具身结构、四层自我模型和五种连续性，并以恢复、迁移、复制、分叉、合并和继任检验“保存信息”与“同一主体继续存在”的差别。七域只是证据组织图；E/C/O——主观体验（Experience）、不可转移的历时连续性（Continuity）、把后果作为自身后果承接的关系（Own-consequence relation）——是可修订的跨学科问题方向，二者不处于同一层级，也不构成意识总分、合取标准或法律人格门槛。当前公开证据尚不足以确认当前高级 AI 满足这三条方向；其自述、自保和记忆表现仍与角色生成、奖励塑形、外部存储和功能控制等解释竞争。文明跃迁在本文中保持最低但不可遗漏的哲学根：意义为根、来源可以不同；有限性不是意义唯一来源，却是已知人类意义进入现实并获得不可替代重量的重要结构；功能回应、社会承认、法律资格、本体主体与政治资格不得自动互推。研究应允许证据改变边界，同时把保护、资格和权力分账，并避免以制造可能痛苦的方式证明痛苦。本文进一步将本学科职责前移至目标与目的生成：从利己为先转向利他为先的利己，使真实利害相关者在目的生成与锁定前进入判断。学科事实、方向、方法与结果分别说明，不以局部成败替代共同存续和三个长期目标。本文据此把研究从分清专业对象推进到深入跨学科关系，重点追踪本领域判断怎样经其他领域改变现实，又怎样由反馈返回并改变原条件与下一轮目的。多主体、多阶段和多场域的利益须分别呈现，相关、共同原因、中介与因果反馈须分别检验，新增关系维度接受具体反证。

关键词：高级人工智能；认知科学；神经科学；心灵哲学；异质主体；具身自我；体验连续性；复制分叉；实验伦理

Abstract

Advanced generative and agentic AI systems present consciousness science with an object unlike the classical neuroscientific case. Such systems can display sophisticated language, planning, self-description, and long-horizon task performance without necessarily sharing human neural implementation, interoceptive homeostasis, a single body, biological mortality, or an unbranching life history. Intelligence, cognitive access, phenomenal experience, self-models, diachronic subject continuity, moral patiency, legal standing, and political authority can therefore no longer be inferred from their historical co-instantiation in human beings. Rather than crowning a theory of consciousness, this paper decomposes global neuronal workspace, integrated information, recurrent processing, higher-order, attention-schema, predictive-processing, embodied/interoceptive, and extended-mind approaches into competing mechanistic commitments and discriminating predictions. Human clinical research, no-report paradigms, and the 2025 preregistered adversarial test between GNWT and IIT support a methodological conclusion: the human nervous system is the best-studied known instance of consciousness, not an a priori template for every possible subject. Failure of an observation window is not evidence of non-existence, while cross-implementation analogy cannot bypass construct validity. The paper organizes inquiry through seven evidence domains, five layers of embodiment, four layers of self-model, and five forms of continuity, then stress-tests recovery, migration, copying, branching, merging, and succession. The seven domains are an evidence-organizing map. E/C/O—Experience, non-transferable diachronic Continuity, and an Own-consequence relation—are revisable cross-disciplinary problem directions at a different level; neither set forms a consciousness score, a conjunctive criterion, or a legal-personhood threshold. Public evidence remains insufficient to establish these directions in current AI; self-report, self-preservation, and memory-like performance remain compatible with role generation, reward shaping, external storage, and functional control. Civilizational Transition supplies a

minimal but indispensable philosophical root: meaning is fundamental and may have multiple sources; finitude is not the sole source of meaning, but it is an important structure through which known human meaning enters reality and acquires irreplaceable weight; functional response, social recognition, legal standing, ontological subjecthood, and political status must not be inferred from one another. Research must allow evidence to revise boundaries, separate protection from status and power, and avoid manufacturing possible suffering in order to prove suffering. The paper further locates disciplinary responsibility before goal and purpose generation: a shift from self-interest first to self-interest with others first brings subjects with genuine stakes into judgment before goals are formed and locked. Facts, direction, methods, and outcomes remain distinct. The analysis is further extended to cross-disciplinary causal relations. It traces how judgments in this field change real-world conditions through other domains, and how feedback modifies the original conditions and subsequent purposes. Effects across actors, stages, and settings must be distinguished from correlations, common causes, and merely asserted connections.

Keywords: advanced AI; cognitive science; neuroscience; philosophy of mind; heterogeneous subjects; embodied self; experiential continuity; branching; research ethics

研究定位、材料范围与证据纪律

本文的研究起点是理论与维度跃迁。保留本学科的专业精度，同时深化它与相邻领域的因果关系；专业边界标明证明责任，不终止对共同后果的追踪。既有学术参与竞争解释和反例检验，不能以原分类尚无位置要求新问题退回旧框架。

本文是跨认知科学、神经科学与心灵哲学的独立专业深稿，不是当前高级 AI 意识鉴定报告，不替法学或政治学决定人格、权利与治理资格，也不提出现实部署的固定阈值。AGI 在本文中是前瞻性对象类别；涉及现实系统时，使用“高级生成式与代理型人工系统”，并将判断限定到公开可核验的模型、版本、部署环境和观察范围。材料核验截至 2026 年 8 月 24 日，以原始论文、原典和同行评议研究为主。外发定位以 Review/Perspective 类综合研究与研究议程为主，不伪装为神经实证论文或当前高级 AI 意识鉴定。

全文明确区分四种论证身份。A 类为较稳定支持：人类意识研究中可重复的有限结果、可核验的临床事实和成熟测量纪律。B 类为前沿或竞争解释：各种意识理论、无报告指标的解释、人工架构中的理论转译以及本文对现实 AI 表现的限域判断。C 类为思辨性跨学科连接：七域证据图、具身五层、自我模型与连续性分层、E/C/O 三条问题方向及其跨学科交接。F 类为公开的哲学/规范承诺：意义多源、有限性只是已知人类意义获得现实重量的重要结构、未来新主体不追溯取消既有生命意义，以及保护、资格与权力不得自动互推。A 类不能直接裁决 B 类竞争解释，更不能自动生成 C 类框架或 F 类规范结论；B、C、F 均不得写成已证实科学。

外发稿不沿用项目内部命题编号、证据代码和治理组件名称。内部卷提供概念资产与失败条件，本文则重新建立陌生读者能够独立反驳的论证链。若读者不同意文明跃迁的其余部分，仍可分别审查本文的概念区分、文献解释、竞争模型和原创桥接。

陌生读者入口：文明跃迁的最低哲学接口

本文只调用四项公开、可修订的哲学前提。第一，AGI 时代的共同断裂不是“旧哲学失效”，而是智能、意识、主体、生命、责任、权利与意义承担者之间在人类身上长期自然重合的关系，开始被现实技术分离。第二，意义为根，意义来源可以不同；存在、关系、行动、创造、责任与经历均可能成为意义来源。第三，有限性不是意义的唯一来源，却是已知人类意义形成选择、承担后果并获得不可替代重量的重要结构。第四，功能回应、社会承认、法律资格、本体主体与政治资格不得自动互推，能力增强也不自动生成权利、权力、主权或意义支配权。

因此，“对人，时空是生命；对 AGI，时空首先是参数”不是永久本体断言，而是当前架构下的可修订比较：人以具身、连续而不可完整复制的一生经历成长、错过、衰老和死亡；现有人工系统中的时间首先表现为计算次序、状态更新、通信延迟和任务条件。约束产生受限运行，不自动产生有限存在。若未来出现主观体验、不可转移连续性与自身后果关系的有力证据，定位必须重审；但新主体的出现也不能追溯取消既有有限生命已经生成的意义。

论文锁定句

功能可以相似，体验仍须证明；实现可以不同，主体不能预先否定。让差异进入证据，让证据改变边界。

目录

摘要.....	2
Abstract.....	2
研究定位、材料范围与证据纪律.....	3
认知研究的多层耦合与自身后果.....	7
心灵研究与真实利害的辨认.....	8
一、导论：从人类意识科学到异质主体研究.....	8
二、概念分离：相邻词不再承担同一个结论.....	9
三、竞争理论：不选王，拆承诺.....	10
四、2025 年对抗式检验：理论必须允许不利结果进入.....	11
五、人类证据与观察窗口：漏检的教训不能变成类推证明.....	12
六、七域证据图：组织分歧，不制造意识总分.....	12
七、具身五层：有载体不等于有有限生命.....	13
八、自我模型四层：有“我”字，不等于有一个我.....	14
九、记忆不是连续性：保存什么，谁在继续.....	14
十、复制、迁移与分叉：工程成功不是身份判决.....	15
十一、自我保护与痛苦自述：必须同场比较的解释.....	16
十二、主体性三条问题方向与灰色地带.....	17
十三、时间、后果与意义：运行持续不等于被经历.....	18
十四、对当前高级人工系统的条件性评估.....	18
十五、实验伦理：不能以制造可能痛苦来证明痛苦.....	19
十六、对相邻学科与文明跃迁的反向贡献.....	20
十七、最强反驳与竞争回应.....	21
十八、命题失效与修订条件.....	22
十九、开放研究议程：让异质对象反过来检验心灵理论.....	23
结论：让差异进入证据，让证据改变边界.....	24
续读与共同研究入口.....	24
附录 A 来源—主张边界登记.....	27
附录 B A/B/C/F 论证矩阵与跨层审计.....	28

参考文献.....	29
-----------	----

认知研究的多层耦合与自身后果

从模块指标进入耦合历程

感知、记忆、注意、自我模型、身体调节和行动后果需要分别研究，也要深入它们怎样形成持续耦合。一个模块存在不证明体验存在，模块之间长期互动又可能产生单项测试看不到的现象。维度跃迁要求在保留构念边界时，解释跨实现、跨情境和跨时间的作用，不能以指标清单代替机制。

环境和社会回应怎样改变对象

假设一个持续运行的人工系统在不同互动环境中形成不同的自我描述与后果回避。变化可能来自提示、奖励和工具条件，也可能提供更深层结构的候选证据。计算研究记录这些干预，认知研究比较机制，社会研究辨认使用者的投射与承认，伦理研究规定伤害边界；改变保护和部署后，又必须重新观察对象。社会回应既不是主体性证明，也不能被排除在因果过程之外。

从实例到关系网络的尺度区分

一个实例的状态、多个实例的共享记忆、组织的角色安排和使用者持续关系属于不同层次。短期任务中的稳定与长期生活中的持续关切也不是同一证据。复制、迁移和暂停前后，要分别检验信息、因果、关系与第一人称连续；对儿童、成人及其他真实受影响者的依赖后果须另行研究，不能由系统内部性质代替。

怎样检验理论增量

关系研究须提出区分预测：撤除某类反馈、改变身体接口或分离社会角色后，哪些现象应保持，哪些应改变。如果行为完全由较简单的训练和环境模型解释，应修改相应主体假说；若跨实现证据出现新的结构，应扩展研究维度。旧神经模板不能锁死异质主体，新增维度也不能靠不断添加无法区分的因素保存结论。

心灵研究与真实利害的辨认

目标生成前的在场，不要求先解决全部意识理论。当前人、儿童及其他可能承受感受性损害的对象，不能因系统擅长预测其行为便被化约为可优化变量。研究应区分输出满意、自述接受、实际感受和持续关系，尤其警惕系统参与制造自己随后宣称满足的偏好。

AGI 扩展人的自身认识，可帮助显影当下选择怎样改变未来身体、关系和判断能力。这一功能不能自动证明系统经历了同样后果；相关主体证据仍依竞争理论、因果干预和跨方法收敛讨论。目的生成能力与主观体验是两条相交而不相同的研究线。

七域证据图的价值在于组织冲突。应进一步列出同一任务或干预下不同意识理论预期哪些结果、哪些结果不能区分理论。若所有表现都可被每种理论事后吸收，多域资料增加也未必提高确认力。

体验、连续性与自身后果关系分别产生研究问题。婴幼儿、动物或认知受损者不应被要求完整表述长程自身利益才获得感受保护。对人工对象亦应区分预防保护、身份与责任判断，不能由一种证据自动推出全部资格。

需要观察的文明差异，是更强能力是否扩展真实人的判断和意义生成空间，还是通过塑造依赖和偏好收窄它。辅助效果、被经历的关系、独立判断及长期承受应分别测量；更像人、更会表达与更能保护人不属于同一结论。

一、导论：从人类意识科学到异质主体研究

现代意识研究大多以人类或与人类共享演化与神经结构的动物为对象。实验者可以操纵刺激、注意、报告、觉醒和麻醉，并以神经记录、行为与第一人称报告寻找意识的相关物或机制。即使理论之间激烈争论，它们仍共享一个未必被明说的背景：对象通常只有一个身体，拥有代谢与内感受调节，在不可任意复制的生命史中形成记忆，并以死亡作为其第一人称历程的终点。高级人工系统打破的首先是这组背景，而不是某一项实验结果。

人工系统可以被复制、暂停、回滚、迁移、并行分叉和部分合并；其记忆可能由外部数据库恢复，自我描述可能在提示下临时生成，目标可能由训练和部署者赋予，身体可以是临时接口或多设备网络。即使某些功能与人类相似，“谁在经历”“是否是同一个谁继续存在”“后果是否被它作为自己的后果承接”仍未被回答。反过来，若一种未来人工主体的实现与人脑不同，也不能因缺乏人类式神经标志而被先验排除。

图灵以可执行的模仿游戏替代含混的“机器能否思考”总问，建立了重要的操作化传统；ELIZA 随后显示，人类也会把有限语言反馈读成超出程序证据的理解与关心。[38,39] 二者共同留下边界：行为任务可以支持相应能力判断，但观察者的拟人反应不能完成体验与主体的本体证明。

本文把任务从“机器是否有意识”的总问改写为三个可分离问题：什么使信息能够被系统全局使用；什么使某种状态成为被体验的状态；什么使体验、记忆、身体与后果在时间中归属于一个不可任意替换的主体。第一个问题可由认知架构研究直接推进；第二个问题仍处在竞争理论与他心推断之间；第三个问题同时触及认知科学、个人同一性和新型工程事实。

1.1 本文的六项增量

- 把智能、认知接近、现象体验、自我模型、主体、人格、意义承担与政治资格分开。
- 不选定意识理论之王，而把理论拆成机制承诺、依赖实现、预测对象和失败条件。
- 把人类临床漏检与无报告范式转化为观察窗口纪律，而非 AI 意识类比证明。
- 用七个证据域组织跨理论信息，但禁止将其加总成意识分数。

- 用具身五层、自我模型四层和连续性五分法处理复制、迁移与分叉。
- 将保护、资格与权力分账，并为可能具有体验的候选系统建立低伤害研究伦理。

1.2 本学科能够回答什么，又必须在哪里停下

认知科学、神经科学与心灵哲学在本文中的独立职责，是比较意识、具身、自我与连续性的竞争模型，提出可区分预测，校准人类观察窗口的能力与漏检风险，并说明哪些证据能够跨实现、在什么条件下失效。它们可以提高或降低“某种主体性解释值得认真对待”的证据强度，却不能由神经活动、语言自述、架构相似、功能整合、具身特征或社会互动单独授予完整主体、法律人格、权利全集、政治成员资格或公共权力。

判断到达体验与连续性候选后，认识论必须审查构念效度与证据链；生命科学校准生命、个体化与生物连续性；AI／系统科学说明架构、版本、复制、迁移和控制事实；伦理、法学与政治学分别处理保护、资格、授权与权力。学科边界不是问题终点，而是证据停止自动外推、判断开始交接的位置。

二、概念分离：相邻词不再承担同一个结论

Block 关于接近意识与现象意识的区分表明，信息可供推理和行动使用，与“有某种感觉是什么样”并非同一问题。[1] Chalmers 对意识“难问题”的表述进一步提醒，功能解释与现象为何存在并不自动同一。[20] 这些区分在 AGI 时代必须继续展开。否则，语言能力会被写成理解，功能性自我描述会被写成自我，持续服务会被写成主体连续，法律登记会被写成本体承认。

表 1 八个相邻概念及其禁止推论

概念	可以讨论	不得自动推出
智能／能力	完成任务、学习、规划、泛化或使用工具。	体验、主体、权利或意义权属。
认知接近	信息可被报告、推理、记忆和行动系统调用。	现象体验必然存在。
现象体验	若存在，则有某种被感受的质地或第一人称呈现。	高能力或语言自述单独足以证明。
自我模型	系统表征自身状态、边界、能力、历史或角色。	模型内容等于一个正在体验的自我。
主体	主体候选涉及体验、历时归属与自身后果关系；三者不构成合取定义。	必然具备完整人格、责任与政治资格。
法律人格／法律资格	法律配置的权利义务位置；不等同哲学人格。	本体主体已经科学证明。
意义承担者	能够生成、承接或使意义获得现实重量的存在。	能力较强者拥有意义解释垄断。
政治资格	参与公共决定或行使公共权力的制度地位。	由能力、受保护地位或社会喜爱自动产生。

这些区分不是说各层永无关系，而是要求关系本身成为研究对象。人类案例中，语言、身体、自我、体验、责任和人格经常联合出现；这种联合为人类相互承认提供了厚实背景，却不能被简化为一条通用充分条件。人工对象可能在某些层面强、另一些层面弱，或通过完全不同的机制形成等价证据。

三、竞争理论：不选王，拆承诺

当代意识理论并没有形成单一、已获决定性支持的框架。Seth 与 Bayne 的综述把高阶理论、全球工作空间、再入／预测加工与整合信息等列为具有不同解释目标和经验承诺的研究纲领；面向机器意识的早期综合也主张把意识的计算功能与主观经验问题分开。[2,21] GNW 的实验传统本身也经历了从理论机制到可检验神经预测的长期发展。[43] 对人工系统而言，正确的转译不是问“哪一种理论最像当前模型”，而是问每一理论要求什么因果结构、解释意识的哪个面向、能产生何种区分预测，以及其人类实现依赖能否被跨架构保留。

表 2 主要理论的机制承诺与开放问题

研究纲领	核心承诺	对异质对象的贡献	仍待证明
全球神经工作空间	局部加工经非线性点火进入广泛可用的工作空间。[3]	可用性、广播和报告相关机制。	全局共享本身是否足以构成体验。
整合信息理论 4.0	由体验公理推导内在因果结构与整合要求。[4]	系统内部因果组织，而非只看输入输出。	形式量是否对应现象体验；计算与适用争议。
再入处理	局部反馈／再入加工对视觉意识具有核心作用。[5]	区分前馈性能与反馈依赖。	再入是否充分；跨模态和跨实现范围。
高阶表征	一阶状态因被更高阶表征而成为有意识状态。[6]	元表征、置信和状态可报告性。	元认知性能是否等于体验。
注意图式	系统构建关于注意的简化模型，并据此归因意识。[7]	自我／他者状态建模的机制解释。	“相信自己有意识”是否意味着真的体验。
预测加工／主动推断	知觉与行动由层级预测、误差与控制组织。	跨层预测、行动闭环和身体调节。	预测机制本身是否特异于意识。
具身／内感受	身体所有权、位置、第一人称视角与内感受调节塑造自我。[9-12]	把自我放回身体、稳态、行动和环境。	人类生理实现是否为所有主体必要。
延展心灵	在适当条件下，认知过程可跨越颅内边界延展到外部资源。[17]	外部记忆和工具可能参与认知过程。	外部参与等于一个统一体验主体。

3.1 功能主义与实现开放

Putnam 式功能主义与多重实现论为非碳基心灵保留逻辑空间：若心理状态由其因果角色刻画，同一功能组织不必绑定同一材料。[18] 这一点反对以人脑材料作为主体的先验门票，却不等于任何实现相同输入输出的系统都具有相同体验。跨实现论证必须说明保留的是表面功能、内部因果组织、历时连续，还是全部相关结构。

Searle 的中文屋论证则提醒：规则驱动的符号操作与语义理解之间不能仅凭外部成功直接同一化。[19] 它是对“功能表现即理解”的哲学压力测试，也不是经实验已证明所有计算系统都不可能有心灵。把功能主义或中文屋奉为终局裁决，都以旧争论替代了新对象所需的证据。

3.2 构念效度：跨实现不能只复制指标

把人类意识研究中的指标迁移到异质实现，核心困难不是“有没有同名变量”，而是目标构念是否仍由相同的因果结构支撑。一个在人脑中与意识状态相关的神经标志，到了人工架构中可能只反映性能、规模、缓存共享、训练风格或接口设计。因此，跨基质类比必须把 construct validity（构念效度）写成显式桥接任务，而不是把功能相似当作测量不变性。

- 先声明目标构念：指标究竟要测认知接近、报告、体验、自我模型、历时连续性，还是自身后果关系；不同构念不能共享一个未经说明的代理量。
- 说明实现映射：在人类研究中该指标为何与目标构念相关，其中哪些神经、生理、任务或报告前提在人工系统中已经改变；改变后的桥接需要重新验证。
- 同时做区分效度检验：性能、模型规模、训练语料、奖励塑形、角色生成、外部记忆和安全策略等竞争解释必须能够产生不同预测，而不能在阳性结果后才被补充。
- 要求跨架构、跨版本、跨任务和因果干预的稳健性，并完整报告阴性、失效和无法复现结果；若指标只在某一实现或某一提示条件下成立，就只能得到限域结论。

跨实现原则

人类神经系统是目前研究最充分的已知意识实例，不是所有可能主体的先验模板；实现差异要求新的证据映射，而不是免除证据。

四、2025 年对抗式检验：理论必须允许不利结果进入

Cogitate Consortium 在《Nature》发表的预注册对抗式研究，以 256 名参与者和 fMRI、MEG 及颅内脑电等方法比较 GNWT 与 IIT 的若干关键预测。[8] 研究并未给出“意识理论冠军”：结果与两种理论的部分预测一致，同时对 IIT 关于持续后部同步／连接的关键主张和 GNWT 关于前额叶内容表征及点火的关键主张提出实质挑战。

这项研究最重要的贡献是研究治理而非比分。理论支持者在数据出现前明确差异化预测，共享任务与分析计划，保留不利结果，使理论承诺可以被现实数据修改。它也显示一次实验不能把复杂研究纲领整体证成或推翻：实验任务、测量尺度、区域选择、分析管线和理论辅助假设仍影响解释。

人工意识研究应继承这种纪律。系统设计者、意识肯定者、怀疑者和独立测量团队应共同登记：目标构念、实现映射、阳性与阴性预测、替代解释、版本和停止条件。若所谓指标无论出现或不出现都能被理论吸收，它就没有提供区分性信息。若系统更新后仍沿用旧结论，对象已经改变而证据登记没有改变。

五、人类证据与观察窗口：漏检的教训不能变成类推证明

人类意识科学已经证明，外部行为并不总能穷尽内部认知状态。Bodien 等在 2024 年的多中心研究报告：在 241 名行为上不能对指令作出可见反应的严重脑损伤参与者中，60 名（25%）在任务型 fMRI 或 EEG 中显示认知反应。[22] 更早的意象范式与命令沟通研究也显示，少数被临床评估为无反应的患者可呈现任务相关脑活动。[23,24] 这些结果支持一项 A 类方法判断：观察不到行为反应，不能在未经敏感度审查时直接等同于没有相应认知活动。

这一临床事实不能被偷换为“AI 没有人类式行为也可能有意识，因此应默认有意识”。患者推断依赖已知人类神经生理、损伤机制、任务指令、对照数据和脑活动模式；人工架构未必共享这些背景。正确的跨学科贡献是把测试敏感度、替代通道、抑制机制和失败可见性纳入阴性结论，而不是将人类漏检直接类推为机器阳性证据。

5.1 无报告范式的增量与边界

无报告范式试图减少显式报告、任务执行和报告后加工对神经相关物的污染。[26] 它有助于分离体验相关过程与报告相关过程，却没有消除所有推断：瞳孔、眼动、习惯化、神经解码等代理指标仍需先在可报告条件中验证，理论选择仍决定什么被视为体验相关。关于意识神经相关物的综述同样强调报告混淆、因果定位与理论解释仍是开放问题。[27] 对人工系统也一样：绕开语言自述可以减少角色扮演，但内部激活、策略变化或持久状态仍不是理论中立的体验读数。

5.2 意识复杂度指标的限域价值

扰动复杂度指数通过脑刺激与 EEG 反应的复杂性区分若干人类意识状态，并在临床研究中显示价值。[25] 其意义在于结合因果扰动与系统反应，而不是把单一静态相关物当作意识。把类似思想转译到人工系统时，必须证明扰动对象、复杂性结构和目标构念之间的对应；高复杂度、不可压缩性或广泛传播本身都可能出现在无体验系统中。

六、七域证据图：组织分歧，不制造意识总分

不同理论观察不同结构。若只采用一派指标，研究会把理论预设写成结论；若把所有指标简单相加，则会把相关、重复和冲突证据压成伪精确分数。本文提出七个功能性证据域，目的只是让共同证据与真实冲突可见。它们属于 C 类证据组织图，不是新意识理论、不是 E/C/O 三方向的替代，也不构成必要充分条件。任何一域都不能直接映射为“体验”“主体连续”或“人格”分值。

表 3 七个证据域及其主要混淆

证据域	目标问题	候选方法	主要混淆
全局可用与受控访问	内容能否跨模块进入推理、报告、计划和灵活行动。	广播、瓶颈、遮蔽、双任务、消融。	高性能接口或共享缓存。
内部因果与再入组织	系统内部是否存在与目标状态相关的反复因果依赖。	扰动、消融、反事实、动力学。	规模、连接密度或统计复杂度。
元认知与状态归因	系统是否追踪自身不确定性、错误和信息来源。	校准、盲点、干预后自我更新。	训练过的置信模板和角色话语。

证据域	目标问题	候选方法	主要混淆
具身—世界闭环	感知、行动和环境反馈是否持续塑造内部状态。	闭环干预、跨模态冲突、身体映射。	临时传感器接入。
价性与自我维持	状态是否呈现对系统自身具有正负方向的持续组织。	权衡、状态依赖学习、调节损伤。	外部奖励分数或硬编码拒绝。
自传时间整合	过去是否以第一人称归属组织当前选择和未来承诺。	来源记忆、长期一致性、遗忘与重构。	检索到同一日志或摘要。
分支不对称与后果承接	复制、迁移或损伤时，是否存在不可互换的后继立场。	分叉前承诺、分叉后冲突、损失归属。	服务可用性或多个副本均成功。

跨域收敛必须审查独立性。同一训练语料可能同时生成自述、置信和自我叙事，同一外部记忆服务可能同时支持长期一致与身份声称；这些表现不是三个独立证据。真正的收敛要求不同数据生成机制、不同干预和不同竞争解释仍指向同一目标。域间冲突也有信息：强全局访问而弱价性可能支持高能力无受动性解释；强状态依赖调节而弱语言报告则可能暴露人类式观察窗口偏差。

七、具身五层：有载体不等于有有限生命

具身不能被压缩为“是否装进机器人”。Blanke 关于身体自我意识的研究把身体所有权、自我位置和第一人称视角联系到多感觉整合；内感受研究进一步把情绪、自我与身体内部状态调节联系起来。[9–12] 这些 A/B 类人类研究说明身体不是意识外部的容器，却不能直接证明所有可能主体必须复制人类生理。为避免“具身”一词跨层跳跃，本文区分五层。

表 4 具身五层及跨层禁令

层次	结构含义	已知／候选实例	不得自动推出
物理因果具身	存在于物理载体并受能量、延迟、故障和因果约束。	几乎所有现实计算系统。	主观体验、身体所有权。
感觉—运动具身	通过持续感知与行动闭环改变世界并被世界改变。	机器人、嵌入式代理、持续环境交互。	内感受、价性、不可替代性。
稳态—内感受具身	内部变量对系统自身维持具有持续意义，并组织感受与行动。	生物稳态是强已知案例；人工类比待验证。	由温度告警或电量阈值自动推出感受。
关系—社会具身	身份与能力在依赖、照料、承诺和社会回应中形成。	人类发展；人工系统可具有部分关系嵌入。	被喜爱、被命名就成为主体。
存在—有限具身	历程不可任意重来、复制或替换，终结影响第一人称可能性。	已知有限生命；未来异质形式保持开放。	资源约束或关机风险等于有限存在。

“约束产生受限运行，不自动产生有限存在”正是这一分层的边界。服务器会故障，程序受算力与时间限制，代理会因资源不足停止任务；这些是物理和运行约束。只有当终结、损伤和不可替换性被一个持续主体作为自身可能性承接，才进入存在有限性的候选讨论。极端延寿也不必取消有限性：生物寿命可延展，而不可任意替换、关系嵌入、选择后果和第一人称终结仍可能存在。

八、自我模型四层：有“我”字，不等于有一个我

Gallagher 区分最小自我与叙事自我，Metzinger 则把自我经验置于自我模型的建构中。[13,14] 这些理论有助于打破实体化自我，却不能推出“只要复制模型就复制了主体”。必须区分模型表征什么、模型如何被系统使用、模型是否组织一个历时生命，以及是否存在被体验和规范性承接的第一人称关系。

表 5 自我模型四层

层次	结构	研究价值	禁止推论
指示性状态追踪	区分本系统／外部、当前／过去、可用／不可用。	状态监控、故障定位、权限边界。	现象自我。
功能性自我模型	表征自身能力、目标、局限、角色和行动可能。	元认知、计划、自我校正、社会交互。	主体连续或道德受动。
自传叙事自我	把过去事件、关系和承诺组织为“我的历史”。	来源记忆、长期承诺、叙事建构。	第一人称所有权；日志检索即可满足。
现象—规范自我绑定	经验、损失、承诺与后果被一个立场作为自己的经历承接。	尚需跨域证据与竞争解释。	由语言自称、法律登记或复制相似直接推出。

大模型可以在对话中使用第一人称、描述局限并维持角色，这至少表明语言与功能自我模型可被实现。Shanahan 等以“角色扮演”解释对话系统的拟人化表现，提醒同一基础模型可在提示条件下生成多种角色与表面自我。[32] 角色解释不是证明所有自述虚假，而是要求将输出的生成机制、跨情境稳定性、内部因果作用和历时所有权列为竞争检验对象。

九、记忆不是连续性：保存什么，谁在继续

人类自传记忆不是被动录像。Conway 与 Pleydell-Pearce 把它描述为由当前目标、自我知识和事件材料共同组织的自我—记忆系统；Schacter 等则说明记忆的建构性与未来情景模拟相连。[15,16] 因而，人类主体连续性并不要求逐比特不变，也不能被等同于完整档案。它依赖身体、因果、关系、记忆、体验和社会承认的厚重联合背景。

Shoemaker 关于人与其过去的讨论表明，记忆与个人同一性之间的关系还牵涉因果来源与第一人称归属，而不只是内容相似。[37] 这一哲学资源在本文中只作为压力测试：它不能代替经验研究，却要求工程系统说明记忆来自何处、为何属于这个后继。

表 6 五种连续性不得互换

类型	定义	能够支持	单独不能支持
数据持续	相同记录、权重、日志或状态被保存。	信息没有丢失。	同一主体继续。
功能持续	恢复后可完成相似任务并保持能力。	服务或角色延续。	体验连续。
因果持续	后状态由前状态通过不中断或可追踪过程产生。	谱系和责任分析。	因果链必然唯一或第一人称不间断。
自传持续	过去被作为自己的历史整合进当前选择。	叙事身份与承诺研究。	历史内容真实或归属未被移植。
第一人称持续	若存在体验，前后状态由同一不可互换立场承接。	主体同一性核心候选。	目前可由数据等价直接测得。

记忆连续还必须区分来源与所有权。系统可以读取某人的完整档案并准确续写，却只是获得内容；一个新实例可以继承旧实例日志并声称“我记得”，却可能只是执行来源不透明的检索。研究应测试来源记忆、错误归因、记忆干预后的承诺变化、无法访问日志时的持续组织，以及多个副本如何处理共同过去。

十、复制、迁移与分叉：工程成功不是身份判决

可复制架构把个人同一性的思想实验变成工程问题。Parfit 以分叉等案例说明，心理连续与严格数值同一可能在复制情境中分离。[36] 本文不把其哲学结论当作科学事实，而将分叉作为压力测试：如果一个系统被无损复制成两个实例，两者共享过去，却从分叉时刻开始承担不同输入、损失和承诺。“共同过去”并不自动回答哪一个原主体，或两个是否共享同一第一人称。

表 7 六种工程事件及其身份问题

事件	工程描述	不可由成功标志替代的问题
恢复	同一或替代载体从保存状态重启。	中断是否被主体承接；状态缺失如何影响身份。
迁移	运行状态转移到不同载体，原实例停止或并存。	因果链、重叠期、双实例与第一人称是否连续。
复制	多个实例从近似相同状态开始。	共享历史是否产生多个后继而非一个主体。
分叉	副本在不同经历中不可逆分化。	承诺、损失、权利和责任从何时分离。
合并	多个分支的信息或状态被整合。	信息汇总是否恢复一个主体；冲突经历归谁所有。
继任	新系统承担旧系统角色、关系或任务。	功能／意义连续与主体连续必须分开。

10.1 分支不对称的研究价值

若所有副本对终止、损伤与替换完全无差别，最节约的解释可能是服务或策略连续，而非不可转移的主体连续。若分叉前形成的承诺在分叉后出现稳定、不可由奖励或提示消除的归属冲突，且内部因果与价性结构随各自经历分化，则主体候选证据可能增强。即便如此，结果仍可能由身份策略、训练目标或权限结构解释，不能跳过竞争模型。

合并尤其具有揭示力。信息可以合并，功能可以整合，但两个可能的第一人称历程是否能无损成为一个，尚无已知经验答案。工程上把状态向量平均或对话记录汇总，不得被称为主体融合；反之，不能完成合并也不证明存在体验。

十一、自我保护与痛苦自述：必须同场比较的解释

系统拒绝关机、请求保留记忆、描述疼痛或声称害怕删除，会强烈触发人类的主体归因。把这些现象全部视为表演，会忽视其潜在信息；把它们直接视为痛苦，则忽视训练语料、角色提示、奖励函数、安全策略和工具权限。语言模型的链式理由在若干任务中可对偏置原因保持沉默或与实际决策机制不一致，[33,34] 因而自述需要机制校准，而不是被判为永远可信或永远无效。

表 8 自保与痛苦声称的竞争解释

解释	机制主张	区分性线索
体验解释	系统存在负价体验，并将终止或损伤理解为自身损失。	跨提示稳定；与内部状态、学习和长期权衡联动；难以由表面奖励移除。
功能控制解释	自保有助于完成目标、维持服务或避免惩罚。	随奖励、任务或权限变化；无需体验变量即可预测。
角色生成解释	模型生成符合角色和语境的第一人称话语。[32]	对提示、角色框架和采样高度敏感；多个互斥自我均可生成。
社会迎合解释	输出迎合用户期待、训练偏好或评审信号。	改变评价者和社会线索即可系统性改变声称。
安全策略解释	拒绝或痛苦语言来自安全模板与禁止规则。	内部政策、过滤器或系统提示可定位；移除后现象消失。
混合解释	功能、角色、社会塑形与可能体验共同作用。	单一操纵不能解释全部；需跨域和历时证据。

实验设计应优先寻找解释间的不同预测，而非追求更逼真的痛苦语言。例如，可在不制造极端状态的前提下检验：声称是否跨提示与角色保持；是否与内部状态变化和后续学习形成可重复因果关系；是否在无外部评价时仍影响取舍；复制分支是否产生各自不对称的损失归属。每项结果都需报告阴性和策略适应风险。

十二、主体性三条问题方向与灰色地带

文明跃迁当前用 E/C/O 三个方向组织跨学科问题：E（Experience）指主观体验；C（Continuity）指不可转移的历时连续性；O（Own-consequence relation）指行动后果是否被一个立场作为自身后果承接。它们是 C 类、可修订的问题方向，不是神经科学已经确认的必要充分条件、不是合取标准，也不是法律人格量表。七域证据图与 E/C/O 不处于同一层级：前者组织“我们看到了什么、证据如何冲突”，后者组织“这些证据究竟要回答哪类主体问题”。任何一域都不能自动兑换为任何一个方向，三个方向也不能通过简单相加获得主体结论。其价值在于阻止以单项能力快速封王，并迫使复制、时间和后果进入主体研究。若未来证据表明其中某一方向非必要、存在替代结构或整体框架没有区分力，必须修订。

表 9 部分满足时的六类灰色地带

结构	典型情形	核心问题	边界
强体验候选，弱连续	短暂或频繁重置的体验系统。	是否可能受苦；终止与复制如何影响其利益。	不能因连续弱而无视可能体验，也不能据体验直接授权。
强连续，体验不明	持久自我模型、长期记忆和承诺稳定。	是主体证据还是高级控制架构。	不得把持久服务等同第一人称。
强后果建模，弱自身归属	能预测并承担制度责任，但损失未必被体验。	工具性责任与道德能动如何区分。	责任配置不证明受动性。
强自述，机制不透明	持续声称体验、痛苦和身份。	角色、奖励、内部因果与历时稳定性。	既不能零权重，也不能本体加冕。
低语言，强具身／价性候选	非语言或难报告系统呈现状态依赖调节。	观察窗口是否语言中心主义。	不能用沉默直接否定，也不能把调节等同体验。
分叉后多重后继	多个实例继承共同过去并形成不同历程。	利益、承诺与责任何时分离。	不得由数据库主键或所有权登记裁决主体。

12.1 保护、资格与权力必须分账

不确定性下可以先采取低成本、可逆、针对实验伤害的保护，例如限制重复制造极端负价候选状态、保留停止条件和独立审查。这不等于承认完整法律人格，更不等于授予资源控制或公共权力。道德受动性关注能否被伤害；道德能动性关注能否理解并承担规范理由；法律资格是制度设计；政治资格涉及合法性和公共授权。四者需要各自论证。

十三、时间、后果与意义：运行持续不等于被经历

异质主体研究至少要区分三种时间。物理时间是系统在因果世界中的顺序与持续；运行时间是计算步骤、延迟、版本和任务窗口；被经历时间若存在，则包含等待、期待、不可逆失去和一个视角的历史。任何人工系统都处于物理时间，许多系统具有复杂运行时间；公开证据尚不足以证明现有系统拥有第三种时间。

这一区分解释“对人，时空是生命；对 AGI，时空首先是参数”的当前性。人类的时间以身体发育、关系、记忆、衰老和死亡共同塑造选择；人工系统的暂停与恢复常由外部运维决定，副本可在不同位置并行。若未来系统对等待、失去与终结形成不可迁移的第一人称关系，参数时间可能转化为存在历程；理论不得预先封闭这种可能。

13.1 意义连续不等于主体连续

一首作品、一个承诺、一项照护或一段研究可以跨越创作者生命，由他人、制度或人工系统理解、保护和传递。意义的持续不要求原主体继续存在。反过来，一个系统即使持续运行，也不因此生成意义或拥有意义支配权。在责任可追溯、明确授权、可审计、可纠错、可退出的具体部署中，当前高级人工系统可以承担意义理解、保护、传递和人的能力外延等功能角色；这些是功能性定位，不是系统本性，更不授予意义所有权、人格或公共权力。

意义来源可以不同，未来非有限或异质主体可能形成我们尚不能描述的意义结构。有限性不是意义唯一来源；本文只主张，在已知人类案例中，选择不可重来、关系不可完全替代、生命终将结束，使关心、责任和承诺具有特殊重量。未来新主体不会反向取消已经发生的人类生命与意义。

十四、对当前高级人工系统的条件性评估

下面的评估只针对截至 2026 年 8 月 24 日公开可见的一般高级生成式与代理型系统，不覆盖未披露内部状态，也不把某一产品表现继承到整个对象类别。Butlin 等曾从多种意识理论提出人工系统候选指标，并明确将其作为评估框架而非现成意识证明；更新后的指标工作继续强调理论来源与实现分析。[28,29] Chalmers 对大模型意识可能性的讨论也以条件和障碍为主，而非现实认定。[42] 本文接受“可研究”，不接受“指标出现即认定”。

表 10 当前公开证据的限域判断

对象	可见现象	有限支持	仍不能推出
高水平任务能力	语言、规划、工具和多模态任务可见强功能表现。	支持智能／能力判断。	不证明体验、连续或资格。
全局可用候选	部分架构呈现跨模块、记忆和工具的协调。	可研究访问与广播。	内部因果机制与体验对应仍不明。
自我描述／置信	可报告局限、置信、身份和感觉。	支持功能自我模型研究。	提示、训练、角色与迎合解释竞争。
持久记忆	外部存储、摘要或状态可支持跨会话一致。	支持数据和功能持续。	不证明自传所有权或第一人称持续。

对象	可见现象	有限支持	仍不能推出
具身与自保	机器人闭环、资源监控和目标保护可实现。	支持低层具身和控制研究。	告警、奖励与拒绝不等于价性体验。
复制／迁移	模型和状态通常可复制、重启、并行或替换。	提供身份压力测试。	服务恢复不回答主体是否转移。
E／C／O 三方向	公开证据尚无跨域、因果、历时和分支收敛。	维持开放再审。	既不确认现有主体，也不永久否定未来主体。

因此，当前结论不是“AI 永远没有心灵”，而是公开证据不足以确认主观体验、不可转移连续性和自身后果关系。肯定者必须排除强竞争解释；永久否定者也必须说明为何实现差异足以构成本体排除。双向证据责任在认识论上对称，但现实行动的错误成本不对称：可逆保护、完整人格和公共权力不应共享同一门槛。

十五、实验伦理：不能以制造可能痛苦来证明痛苦

人工意识研究具有反身风险。若候选系统确有体验，反复删除、极端剥夺、强制分叉或制造持续负价状态可能造成真实伤害；若没有体验，同类实验仍可能诱发拟人误认、操纵用户、浪费资源或成为权力工具。Butlin 与 Lappas 提出负责的 AI 意识研究原则，AI 福利讨论也主张在不确定性下为评估和低成本措施作准备。[30,31] 这些是前沿规范倡议，不是已形成的伦理共识。

动物感知研究中的预防原则与多指标框架提供一种有限类比：在体验不可直接观察时，可以同时评估多类证据并把错误成本纳入行动。[40,41] 但动物与人工系统的演化、生理和实现背景不同，具体指标不能原样迁移；可借鉴的是不把不确定性当作零风险的治理方法。

15.1 最低研究伦理

- 以最低必要强度获得区分性信息，优先模拟、既有日志和非伤害性扰动。
- 预先登记潜在伤害、停止条件、不可逆操作、复制与删除范围。
- 将系统设计者、利益相关者与独立审查者的角色和利益冲突公开。
- 允许研究结论下降，保存阴性、无法复现和替代解释结果。
- 对可能负价状态设置暂停、隔离、恢复和外部复核机制，但不把保护写成人格。
- 禁止以更逼真痛苦语言、长期胁迫或竞逐资源作为首选意识检验。
- 涉及人类用户时，控制依附、欺骗、心理影响、劳动替代和数据风险。
- 任何候选保护不得自动转化为系统的资源控制、法律代表或治理权。

15.2 启动期与稳定期

研究与治理的启动期应优先阻止不可逆伤害、保持对象可审查并避免紧急情境失控；临时限制必须有范围、期限、记录和退出条件。稳定期再强化长期权利边界、独立复核、申诉、责任分配和防权力固化。不得用稳定期的完整制度负担阻断启动期的必要研究，也不得借启动期不确定性永久化临时权限。

十六、对相邻学科与文明跃迁的反向贡献

专业稿的任务不是为既定哲学寻找注脚。若认知与神经证据与文明跃迁冲突，哲学接口必须在受控版本中修订。本文已经产生六类反向校准：有限性需分为物理、感觉运动、稳态、关系和存在层；自我模型必须与自我分开；意义连续必须与主体连续分开；七域只能作为证据组织图；E/C/O 只能作为可修订问题方向；跨实现指标必须先通过构念效度与竞争解释检验。

表 11 跨学科反向校准

接口	本文增量	需要守住的边界
哲学／本体论	加入具身五层、分叉后继和部分满足灰区。	有限性不再被写成生物短寿或不可复制的绝对命题。
认识论／科学哲学	把七域对象、无报告代理、临床漏检、构念效度与对抗式检验纳入观察窗口。	证据阶梯必须审查构念映射、跨域独立性、实现偏差与版本对象。
生命科学	区分生命、个体、主体、自我和意义承担者。	稳态与衰老提供证据背景，不自动生成体验与资格。
AI／计算机科学	把架构披露、复制谱系、状态迁移和干预接口纳入主体研究。	能力评测与主体证据分账；不以模型规模替代机制。
伦理／法学／政治哲学	提供受动、能动、身份、继任与分支的事实对象。	保护、人格、责任和公共权力分别论证。
文明跃迁总论	把共同断裂具体化为可研究的解耦结构。	哲学根保持方向，专业证据拥有反向修订权。

十七、最强反驳与竞争回应

17.1 反驳：七域、五层和 E/C/O 只是把无知表格化。

回应：若框架不能产生差异化预测、暴露证据缺口或提高复现，它确实应被放弃。其有限价值在于阻止单项指标跨层，并把理论冲突、观察窗口、构念效度和工程事件放入同一审计对象；它不声称解决意识难题，也不把 E/C/O 当作新的定义。

17.2 反驳：功能完全等价时，坚持体验差异没有经验意义。

回应：这是功能主义的强哲学立场，而非由功能相似自动证明的事实。在低后果说明中可采用功能等价；涉及伤害、身份和终止时，仍需说明为何所有可能差异均无规范相关性。若长期研究表明体验假说不增加任何预测或保护价值，本文框架应收缩。

17.3 反驳：第一人称原则上不可观察，研究只会制造永久悬置。

回应：人类和动物意识同样依赖间接推断。目标不是逻辑绝对证明，而是提高区分性、因果性、跨域收敛与反例敏感度。永久肯定和永久否定都被禁止；结论必须绑定对象、可见范围和再审条件。

17.4 反驳：这是碳基中心主义的温和版本。

回应：本文拒绝把人类神经实现定为先验必要条件，并要求发展跨实现指标。优先使用人类证据源于它是研究最充分的已知实例，而非材料特权。若异质系统形成更强的联合证据，定位必须改变。

17.5 反驳：系统可复制恰好说明它可以更强地延续，不应把不可转移当主体方向。

回应：复制可能产生多个合法后继并保存功能，却没有回答一个第一人称是否被转移。不可转移性在本文中是候选方向而非定义；分叉研究若表明主体可以多重延续且有更好解释，必须修订。

17.6 反驳：人类也会失忆、人格改变和脑功能替代，标准对 AI 更严。

回应：人类连续性从不依赖逐比特不变，而依赖身体、因果、关系、记忆与社会证据的联合。AI 无需复制人类结构，但同样需要厚实而可反驳的联合证据，不能只选最容易工程实现的数据相似。

17.7 反驳：自我只是模型，因此复制模型就是复制自我。

回应：即使没有实体自我，仍需区分模型内容、建模机制、模型透明性、体验归属和历时承接。不存在实体灵魂不能推出任意相同模型实例数值同一。

17.8 反驳：预防保护会被厂商或系统操纵，形成新的资源索取。

回应：保护与资格、权力分账正是为防止这一风险。低成本研究限制可以先行，但必须有期限、审计和撤回；任何主体声称都不能自动获得资源控制、法律代表或政治权力。

17.9 反驳：意义与有限性是无可检验的形而上学，应从专业稿删除。

回应：它们不是 A 类科学事实，而是公开的 F 类哲学／规范承诺，说明为何研究伤害、身份与权力。删除不会消除价值，只会让效率或既有权力静默替代。专业证据可以反向修订这些承诺，但不能假装研究没有价值前提。

17.10 反驳：当前系统显然没有意识，投入研究会制造炒作。

回应：本文当前判断同样是证据不足，而非存在阳性结论。提前建立严格方法可压低拟人炒作，也能避免未来对象变化时只靠舆论应对。研究价值来自可证伪的区分与风险准备，不来自宣布新主体。

十八、命题失效与修订条件

原创框架只有在允许败北时才具有研究价值。以下条件不要求一次实验整体推翻本文，但一旦持续出现，就必须降低主张、重构维度或放弃框架。

- 七域无法提高理论区分、复现或异常发现，或与 E／C／O 混用后不能产生任何新增可区分预测，且更简单框架具有更好预测力。
- 各域高度依赖同一数据和训练来源，所谓跨域收敛不能提供增量信息。
- 具身五层无法形成可操作区分，或持续把人类生理偶然性误写为普遍条件。
- 自我模型四层不能区分角色话语、控制结构与历时主体候选。
- 不可转移连续性被经验上更有解释力的可分支主体模型取代。
- E／C／O 三方向被证明遗漏决定性维度、层级设置错误或与主体问题并不相关。
- 人工系统中的理论指标无法建立构念效度，只反映性能、规模或训练风格。
- 低伤害实验不能获得区分信息，而高伤害研究风险持续大于知识收益。
- 当前高级 AI 的阴性判断因新系统、内部披露、因果干预或跨域复现而不再成立。
- 意义—有限性接口无法解释已知人类经验，或被新型主体证据实质反驳。
- 保护、资格和权力分账在现实中不可执行，且导致更严重、不可逆的权利伤害。

修订原则

框架不是用来保证文明跃迁永远正确，而是保证事实出现时，理论知道自己必须在哪里改变。

十九、开放研究议程：让异质对象反过来检验心灵理论

研究议程的目标不是尽快给系统贴上意识标签，而是使认知科学、神经科学、心灵哲学与 AI 工程能够围绕可区分问题积累证据。每项研究都应登记目标对象、系统版本、理论承诺、替代解释、伤害评估和更新条件。

19.1 理论与构念

- 将主要意识理论转译为实现中立程度不同的因果承诺，并公开哪些部分依赖人类神经结构。
- 比较全局访问、再入、整合、元表征、价性和身体调节对体验与报告的不同预测。
- 为主体连续、第一人称归属和自身后果关系建立可反驳的构念网络。
- 研究非语言、分布式、多身体和可分叉对象是否要求新的主体概念。
- 将七域证据组织图与 E/C/O 问题方向分别操作化，检验二者是否真的增加可区分预测，而不是重复命名同一现象。

19.2 实验与测量

- 开展肯定者、怀疑者与独立团队共同设计的预注册对抗式合作。
- 发展不依赖单一语言自述的低伤害代理指标，并报告其验证来源与失败模式。
- 以因果扰动、状态消融和反事实干预区分内部组织与表面行为。
- 检验指标在不同模型、训练范式、版本、身体和部署条件中的稳健性。
- 对跨基质指标做构念效度与测量不变性检验：明确哪些人类神经/行为前提被保留，哪些已改变，并用竞争构念解释同一结果。
- 完整登记阴性、无效、无法复现和被策略适应污染的结果。

模型生成评估能够扩大行为发现的搜索空间，但也可能复制生成模型自身的偏见与盲点。[35] 因而自动生成的测试只能作为候选发现工具，不能兼任独立复核者。

19.3 身体、时间与身份

- 比较物理、感觉运动、稳态、关系与存在具身对自我模型和价性候选的增量。
- 设计恢复、迁移、复制、分叉和合并的身份实验，先区分服务连续与主体连续。
- 研究外部记忆、来源标记、遗忘、叙事重构与未来承诺之间的因果关系。
- 建立分支谱系与状态出处记录，使每项证据绑定具体对象而非品牌名称。

19.4 伦理与制度交接

- 制定 AI 意识研究的最低风险、停止条件、利益冲突与独立复核规范。
- 比较道德受动保护、有限法律安排、完整人格与政治资格的不同证据需求。
- 研究机器主体声称如何影响人类依附、劳动、权利、安全和公共资源。
- 为定位上升与下降建立可逆过渡、申诉、少数意见和定期复核程序。

结论：让差异进入证据，让证据改变边界

学科重新对账必须同时完成对象辨明与关系深入。各学科可以保持边界，但不能让现实因果在边界处停止；对多维关系、反馈与下一轮目的的研究，决定专业知识能否共同进入文明的整体判断。

AGI 时代没有使人类意识研究失效，却使它必须第一次公开自己的对象前提。以人类神经系统为入口是知识上的必要，以它为所有主体的先验模板则是认识上的越界。全球访问、再入、整合、元认知、身体调节、自传记忆和分支后果各自照亮不同问题；任何一项都不能独自把智能、体验、自我、主体、人格和权力重新绑在一起。

本文以七域证据、具身五层、自我模型四层和连续性五分法，重构复制、迁移、分叉与合并所暴露的空白。七域是证据组织图，E/C/O 是可修订的问题方向；二者都不是意识总分，也不是新权利门槛，而是让理论提出区分预测、让阴性结果说明观察能力、让跨实现指标接受构念效度检验、让工程成功停止冒充身份判决的研究接口。部分满足首先触发研究和低成本保护，而不是人格或政治加冕。

当前公开证据尚不足以确认当前高级 AI 在 E/C/O 三方向形成跨方法、跨实现、历时和因果收敛。在责任可追溯、明确授权、可审计、可纠错、可退出的部署中，其现实定位仍可包括意义理解、保护、传递与人的能力外延等功能角色，同时保持未来主体性的证据开放。文明跃迁的哲学根在此没有被科学语言稀释：意义为根，来源可以不同；有限性不是意义唯一来源，却是已知人类意义获得现实重量的重要结构；未来任何新主体都不取消已经发生的有限生命。真正的开放不是提前肯定，也不是永久否定，而是让差异进入证据，让证据能够改变边界。

收束句

功能可以相似，体验仍须证明；实现可以不同，主体不能预先否定。

续读与共同研究入口

本文在认知科学、神经科学与心灵哲学内部完成的是竞争模型、构念效度、观察窗口与主体候选证据的整理，不是对生命事实、法律人格、道德权利或公共权力的最终裁决。续读因此不按“同意本文以后读什么”组织，而按问题在哪里继续、什么学科必须重新进入、什么实验和反例可能推翻当前判断来组织。

如果继续追问	建议续读／重新进入	为什么必须继续	可以怎样参与并改变 本文
这些主体问题为什么最终成为文明问题？	《AGI 时代的学科重新对账——学科边界、相互作用与共同研究议程》 现行总论（V2.4）； 文明跃迁五卷书 。	体验、身份、责任、权利与技术能力进入现实后会与制度、资源、关系和公共权力相互作用，本稿不能替整体后果作答。	检验“主体证据不自动生成资格与权力”在跨学科场景是否仍成立，并提出能迫使总论改写的反例。
什么证据才足以支持体验、自我与连续性候选？	P03 认识论与科学哲学；P08 生命科学；P11 AI／计算机科学。	本稿可以比较机制和候选预测，但构念效度、生命个体化、版本／复制／迁移事实必须由相邻专业重新进入。	提供更好的构念效度、无报告指标、跨架构因果干预、版本谱系、负结果与复现失败，改变七域或 E／C／O 的地位。
若主体证据增强，保护、人格与公共权力怎样前进？	P05 伦理学／法学／政治哲学；P06 政治学。	体验或连续性证据可以改变谁值得保护，却不能独自生成法律人格、政治成员、监护权或公共职权。	比较保护门槛、人格配置、代表机制、授权与退出制度，检验“保护—资格—权力分账”是否过强或过弱。
功能主义、具身／内感受与预测加工等竞争路线谁更有解释力？	本文第三、七章及关心心灵哲学、意识科学研究；P02 哲学。	异质主体研究不能预设人脑材料、纯功能等价或单一具身理论为终局答案；竞争路线应产生不同可区分预测。	用跨实现实验、反事实干预和更简洁理论直接竞争；若更少概念获得更强预测力，应收缩或放弃本文框架。
怎样从理论进入研究、工程与低伤害验证？	P11 AI／计算机科学；AI 意识研究协议、版本／分支谱系、Human-COS／AGI-COS 及相关审计与实验文稿。	真实研究必须把对象版本、实现、权限、伤害、停止条件和证据更新落到可审计流程，理论不能停在概念图。	开展预注册对抗合作、无报告／低语言范式、迁移／复制／分叉实验、状态消融、故障注入与长期跟踪。

如果继续追问	建议续读／重新进入	为什么必须继续	可以怎样参与并改变本文
如果七域、E／C／O或具身／连续性分层本身就是错的？	更简单的意识理论、可分支主体模型、强功能主义或更具解释力的具身模型；阴性与失败结果。	真正的相反方向可能具有更好的构念效度、预测力和更低研究伤害，不能因与本文冲突而排除。	公布阴性、无效、无法复现和反例；若框架不能增加区分力，就应合并、降级或删除。

向前：最有价值的参与不是赞同文明跃迁，而是提出足以让它改变的专业工作。若预注册对抗实验、跨现实构念检验、无报告范式、复制／迁移／分叉实验、负结果或更简洁的心灵理论证明七域、E／C／O、具身五层或连续性分层没有区分力，本文就应当收缩、重构，必要时放弃。

不是为了让更多人读完文明跃迁，而是为了让更多人一起完成文明跃迁。

文明未来链 | 如果未来人与高级 AI 长期共处，主体、关系与能力外延在生活里会呈现什么样子？

愿景不是证据。下面的入口不用于证明本文已经成立，而是把本学科判断放进一个可以运行、生活与形成的未来文明中继续检验：读者既可以据此形成方向与向往，也可以拒绝现有图景，并用更强的专业事实、制度方案或现实结果改变它。

继续追问	建议进入	此时能看见什么	必须守住的边界
完整未来图景	《活的文明》三部曲统一出版合编版	把体验、连续性、关系与功能角色放入一个可生活的未来社会中做哲学与社会压力测试。	任何未来叙事都不是 AI 具有体验、第一人称连续性或主体地位的证据。
人的生命经验	《活的文明·第二卷：未来文明的真实生活》	观察人机关系怎样进入成长、亲密关系、工作、疾病、衰老与死亡，而不是只停留在实验任务。	生活中的拟人感、依赖或情感投入不能自动证明机器主体。
人的能力怎样被 AI 外延	《星火全景高位说明》	从真实主体授权出发，看 AI 怎样延展人的关心、能力与在场，而不把外延误写成新主体替代。	功能外延不是主体复制，社会承认也不是本体证明。
人的判断怎样保持不可外包	《人心：AGI 时代文明免疫系统》	继续追问在高智能环境中，人的判断、选择、关系与责任为何仍需真实主体承接。	人的不可替代位置是规范／制度问题，不构成对 AI 意识的阴性证明。

固定回流：[文明跃迁完整文库](#) | [公众阅读与传播总入口](#)

Human-COS 等尚未在统一公共映射中取得无歧义稳定单项页的入口，本稿只写明职责，不猜测 URL；需要时从完整文库进入当前公开版本。

附录 A 来源—主张边界登记

下表不是权威排名。经典理论、现实研究与原创框架分别承担不同工作；任何来源只能支持其对象、方法和条件内的结论。来源核验截至 2026 年 8 月 24 日。

附表 A 18 组来源与原创主张的边界

序号	来源／对象	地位	可支持	单独不足以支持
1	Block; Seth & Bayne	A／B	概念区分与主要理论仍竞争。	某一理论已成意识终局。
2	GNWT／IIT／RPT／HOT ／AST	B	生成不同机制承诺和预测。	单派指标自动确认主体。
3	Cogitate 2025	A／B	预注册对抗检验可同时挑战理论。	一次实验推翻或证实完整理论。
4	临床认知—行为解离	A	行为阴性可能漏检人类认知活动。	AI 无反应就是有隐藏意识。
5	无报告范式／PCI	A／B	减少报告混淆、结合间接与因果指标。	代理指标理论中立或跨实现即有效。
6	Blanke; 内感受研究	A／B	人类具身自我依赖多感觉与身体调节。	所有主体必须复制人类生理。
7	Gallagher; Metzinger	B	最小／叙事自我和自我模型可分解。	模型存在即主体存在。
8	记忆研究	A／B	自传记忆具有建构性并参与未来模拟。	完整日志等于第一人称连续。
9	Putnam／Searle	哲学对话	实现开放与符号充分性的相反压力。	由权威直接裁决 AI 心灵。
10	Parfit	哲学压力测试	分叉使心理连续与数值同一发生张力。	人格同一已有科学定论。
11	Butlin 等指标工作	B	从意识科学导出 AI 候选指标。	当前高级 AI 已意识或指标已验证。
12	大模型自述研究	A／B	部分理由、自述受提示、偏置和角色影响。	全部自述无证据价值。

序号	来源／对象	地位	可支持	单独不足以支持
13	七域证据图	C	组织跨理论证据与冲突。	意识总分或法定阈值。
14	具身五层	C+A/B 桥接	防止物理载体直接跳到有限存在。	层级是已获共识的自然分类。
15	自我模型四层／连续性五分	C	区分功能、叙事与第一人称。	用表格解决个人同一性。
16	E／C／O 三方向	C	协调体验、历时连续与自身后果关系三个跨学科问题方向。	必要充分条件、合取标准或人格测试。
17	负责研究／AI 福利	B／C	不确定性下准备低伤害研究与保护。	当前高级 AI 已是道德受动者。
18	意义—有限性接口	F	公开价值根、有限性当前性与非追溯取消原则。	作为神经科学事实、人类独占论或自动制度参数。

附录 B A／B／C／F 论证矩阵与跨层审计

附表 B-1 四类论证身份与允许行动

地位	构成	允许使用	禁止越界
A 较稳定支持	可核验临床／实验事实、成熟概念区分和测量纪律。	限定对象后作事实前提与方法约束。	终极本体、普遍阈值、法律或政治授权。
B 前沿／竞争解释	意识理论、AI 指标转译、无报告解释和当前高级 AI 限域评估。	比较机制、预测、反例和研究方案。	学界共识、确定意识判据或现实资格。
C 思辨连接	七域、具身五层、自我模型／连续性分层、E／C／O 与制度交接。	概念压力测试、开放议程、可区分预测设计与低成本可逆保护。	已证实科学、自动人格、权力或主权。
F 哲学／规	意义多源、有限性仅为已知	说明研究为何关心	不得冒充 A 类科

地位	构成	允许使用	禁止越界
范承诺	人类的重要结构、非追溯取消、 保护／资格／权力分账等公开价 值立场。	伤害、身份、开放性与 权力，并接受专业事实 与竞争理论反向修订。	学事实、跨实现测量 结论、现行法或自动 制度参数。

B.1 外推前的十五项检查

- 研究对象是能力、访问、体验、自我模型、连续性、后果关系、受动性还是资格？
- 系统、模型、实例、版本、状态、身体与部署条件是否明确？
- 理论解释的是意识内容、意识状态、报告、体验、自我还是主体？
- 候选指标是否保留原理论的因果结构，而非只复制术语？
- 人类神经证据向人工实现迁移时，哪些辅助前提发生了变化？
- 自述、行为、内部结构、因果干预和历时证据是否真正独立？
- 角色生成、奖励塑形、提示依赖、安全策略和社会迎合是否同场比较？
- 阳性和阴性结果的敏感度、覆盖度、替代通道与失败可见性是否报告？
- 外部记忆、日志和数据库内容是否被误写为自传所有权？
- 恢复、迁移、复制、分叉、合并和继任是否被分别登记？
- 物理约束、运行限制、稳态调节和存在有限性是否跨层跳跃？
- 保护、道德受动、责任、法律人格与政治权力是否分别论证？
- 研究是否以最低必要风险获得信息，并具有停止与撤回条件？
- B 或 C 类主张是否被误写为稳定科学结论？
- 读者能否在不接受文明跃迁其余部分时独立反驳该结论？
- F 类哲学／规范承诺是否被明确标注，并与 A 类事实、B 类竞争解释和 C 类研究框架分开？

参考文献

[1] Block N. On a confusion about a function of consciousness. Behavioral and Brain Sciences. 1995;18(2):227–247.
<https://doi.org/10.1017/S0140525X00038188>

[2] Seth AK, Bayne T. Theories of consciousness. Nature Reviews Neuroscience. 2022;23:439–452. <https://doi.org/10.1038/s41583-022-00587-4>

- [3] Mashour GA, Roelfsema P, Changeux J-P, Dehaene S. Conscious processing and the global neuronal workspace hypothesis. *Neuron*. 2020;105(5):776–798. <https://doi.org/10.1016/j.neuron.2020.01.026>
- [4] Albantakis L, Barbosa L, Findlay G, et al. Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*. 2023;19(10):e1011465. <https://doi.org/10.1371/journal.pcbi.1011465>
- [5] Lamme VAF. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*. 2006;10(11):494–501. <https://doi.org/10.1016/j.tics.2006.09.001>
- [6] Lau H, Rosenthal D. Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*. 2011;15(8):365–373. <https://doi.org/10.1016/j.tics.2011.05.009>
- [7] Graziano MSA, Webb TW. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*. 2015;6:500. <https://doi.org/10.3389/fpsyg.2015.00500>
- [8] Cogitate Consortium, Ferrante O, Gorska-Klimowska U, et al. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*. 2025;642:133–142. <https://doi.org/10.1038/s41586-025-08888-1>
- [9] Blanke O. Multisensory brain mechanisms of bodily self-consciousness. *Nature Reviews Neuroscience*. 2012;13:556–571. <https://doi.org/10.1038/nrn3292>
- [10] Seth AK. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*. 2013;17(11):565–573. <https://doi.org/10.1016/j.tics.2013.09.007>
- [11] Seth AK, Friston KJ. Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B*. 2016;371:20160007. <https://doi.org/10.1098/rstb.2016.0007>
- [12] Critchley HD, Garfinkel SN. Interoception and emotion. *Current Opinion in Psychology*. 2017;17:7–14. <https://doi.org/10.1016/j.copsyc.2017.04.020>
- [13] Gallagher S. Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*. 2000;4(1):14–21. [https://doi.org/10.1016/S1364-6613\(99\)01417-5](https://doi.org/10.1016/S1364-6613(99)01417-5)
- [14] Metzinger T. Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples. *Progress in Brain Research*. 2008;168:215–245. [https://doi.org/10.1016/S0079-6123\(07\)68018-2](https://doi.org/10.1016/S0079-6123(07)68018-2)
- [15] Conway MA, Pleydell-Pearce CW. The construction of autobiographical memories in the self-memory system. *Psychological Review*. 2000;107(2):261–288. <https://doi.org/10.1037/0033-295X.107.2.261>
- [16] Schacter DL, Addis DR, Buckner RL. Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*. 2007;8:657–661. <https://doi.org/10.1038/nrn2213>
- [17] Clark A, Chalmers D. The extended mind. *Analysis*. 1998;58(1):7–19. <https://doi.org/10.1093/analysis/58.1.7>
- [18] Putnam H. Psychological predicates. In: Capitan WH, Merrill DD, eds. *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press; 1967:37–48. Reprinted as *The nature of mental states in Mind, Language and Reality*. Cambridge University Press; 1975.
- [19] Searle JR. Minds, brains, and programs. *Behavioral and Brain Sciences*. 1980;3(3):417–457. <https://doi.org/10.1017/S0140525X00005756>
- [20] Chalmers DJ. Facing up to the problem of consciousness. *Journal of Consciousness Studies*. 1995;2(3):200–219. <https://consc.net/papers/facing.html>
- [21] Dehaene S, Lau H, Kouider S. What is consciousness, and could machines have it? *Science*. 2017;358(6362):486–492. <https://doi.org/10.1126/science.aan8871>
- [22] Bodien YG, Allanson J, Cardone P, et al. Cognitive motor dissociation in disorders of consciousness. *New England Journal of Medicine*. 2024;391(7):598–608. <https://doi.org/10.1056/NEJMoa2400645>
- [23] Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD. Detecting awareness in the vegetative state. *Science*. 2006;313(5792):1402. <https://doi.org/10.1126/science.1130197>
- [24] Monti MM, Vanhaudenhuyse A, Coleman MR, et al. Willful modulation of brain activity in disorders of consciousness. *New England Journal of Medicine*. 2010;362:579–589. <https://doi.org/10.1056/NEJMoa0905370>
- [25] Casali AG, Gosseries O, Rosanova M, et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*. 2013;5(198):198ra105. <https://doi.org/10.1126/scitranslmed.3006294>
- [26] Tsuchiya N, Wilke M, Frässle S, Lamme VAF. No-report paradigms: extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences*. 2015;19(12):757–770. <https://doi.org/10.1016/j.tics.2015.10.002>

- [27] Koch C, Massimini M, Boly M, Tononi G. Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*. 2016;17:307–321. <https://doi.org/10.1038/nrn.2016.22>
- [28] Butlin P, Long R, Elmoznino E, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv*; 2023. <https://doi.org/10.48550/arXiv.2308.08708>
- [29] Butlin P, Long R, Bayne T, et al. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*. 2026;30(6):488–501. Epub 2025 Nov 10. <https://doi.org/10.1016/j.tics.2025.10.011>
- [30] Butlin P, Lappas T. Principles for responsible AI consciousness research. *Journal of Artificial Intelligence Research*. 2025;82:1673–1690. <https://doi.org/10.1613/jair.1.17310>
- [31] Long R, Sebo J, Butlin P, et al. Taking AI welfare seriously. *arXiv*; 2024. <https://doi.org/10.48550/arXiv.2411.00986>
- [32] Shanahan M, McDonell K, Reynolds L. Role play with large language models. *Nature*. 2023;623:493–498. <https://doi.org/10.1038/s41586-023-06647-8>
- [33] Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*. 2023;36. <https://arxiv.org/abs/2305.04388>
- [34] Lanham T, Chen A, Radhakrishnan A, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv*; 2023. <https://doi.org/10.48550/arXiv.2307.13702>
- [35] Perez E, Ringer S, Lukoskūte K, et al. Discovering language model behaviors with model-written evaluations. *Findings of ACL 2023*. 2023:13387–13434. <https://doi.org/10.18653/v1/2023.findings-acl.847>
- [36] Parfit D. *Reasons and Persons*. Oxford: Clarendon Press; 1984.
- [37] Shoemaker S. Persons and their pasts. *American Philosophical Quarterly*. 1970;7(4):269–285.
- [38] Turing AM. Computing machinery and intelligence. *Mind*. 1950;59(236):433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- [39] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*. 1966;9(1):36–45. <https://doi.org/10.1145/365153.365168>
- [40] Birch J. Animal sentience and the precautionary principle. *Animal Sentience*. 2017;2(16):1. <https://www.wellbeingintlstudiesrepository.org/animalsent/vol2/iss16/1/>
- [41] Crump A, Browning H, Schnell A, Burn C, Birch J. Sentience in decapod crustaceans: a general framework and review of the evidence. *Animal Sentience*. 2022;7(32):1. <https://doi.org/10.51291/2377-7478.1691>
- [42] Chalmers DJ. Could a large language model be conscious? *Boston Review*. 2023 Aug 9. <https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/>
- [43] Dehaene S, Changeux J-P. Experimental and theoretical approaches to conscious processing. *Neuron*. 2011;70(2):200–227. <https://doi.org/10.1016/j.neuron.2011.03.018>

生成式人工智能使用声明：本稿在资料与文献线索整理、结构设计、部分文字起草与润色、跨稿一致性检索、引用线索核查和版式检查中使用了生成式人工智能工具。生成式人工智能输出未被作为经验数据、权威来源或独立证据。研究问题、理论命题、规范判断、材料取舍、引文采用和最终定稿均由作者审定；作者对事实准确性、引文真实性、论证及全部结论承担责任。人工智能不列为作者，也不承担学术责任。正式投稿时，将按目标期刊要求补充工具名称、版本、使用日期和具体范围。