

AGI 时代的认识论与科学哲学重新对账

——异质主体证据、观察窗口与再审方法

Recalibrating Epistemology and Philosophy of Science in the Age of AGI
Evidence for Heterogeneous Subjects, Observation Windows, and Revisable Inquiry

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

面向认识论、科学哲学、意识科学、认知科学、AI 研究、神经科学及跨学科制度研究共同审查；建立异质主体证据的对象分解、观察窗口校准与再审方法，不认定当前高级 AI 具有意识、人格、权利或治理资格，也不永久排除未来主体性；同时为相关专业研究者以反例、复现、测量方法和竞争解释参与并共同推进文明跃迁跨学科研究提供入口。

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正在把一个长期主要属于哲学的他心问题，转化为需要实验、工程披露、开放科学与制度响应共同处理的现实方法问题。困难不在于机器表现是否足够像人，而在于主观体验、跨时连续性和自身后果关系都不是可以从输出中直接读取的变量；行为、自述、架构、记忆、复制、干预和长期选择只能作为间接证据。本文把潜在人工主体研究理解为异质他心、潜在构念与高后果决策同时存在的问题，并提出五项方法增量：第一，把事实、数据、指标、证据、证成与制度决定分开；第二，把任务能力、功能性自我模型、体验、连续性、后果意识、道德受动性、道德能动性、法律资格、意义主体与治理权分开论证；第三，以主张登记、可复现迹象、区分性预测、因果关联、跨域收敛和对抗稳健形成不依赖“意识总分”的证据进阶；第四，以敏感度、覆盖度、替代通道、抑制机制和失败可见性审查阴性结果；第五，将认识论上的双向证据责任与制度风险的不对称分开。文章借助图灵的操作化、构念效度、欠决定、严重检验、预注册、临床意识漏检研究与意识科学的对抗性合作，说明行为和自我陈述既不是零证据，也不是本体证明。当前公开证据不足以确认当前高级 AI 系统具有主观体验、不可转移的主体连续性和自身后果意识；这一判断必须绑定具体系统、版本、观察范围和修订条件。文明跃迁在本文中只提供公开的哲学与规范前提：智能、意识、主体、责任、权利与意义承担不能再凭其在人类身上的历史重合相互推定；意义来源保持开放，有限性仍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。认识论与科学哲学负责证据质量和结论强度，却不能自行授予法律资格、政治权力或意义支配权；当结论进入机制、保护、制度与公共权力时，认知/神经与 AI 研究、系统科学、伦理学、法学和政治学必须重新进入判断。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害须分别呈现，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；认识论；科学哲学；异质主体；他心问题；构念效度；意识证据；主体连续性；再审方法

Abstract

Advanced generative and agentic AI systems are turning a problem long treated mainly as a philosophical problem of other minds into a practical problem of experiment, engineering disclosure, open science, and institutional response. The difficulty is not whether machines behave sufficiently like humans. Subjective experience, diachronic continuity, and a system's relation to consequences borne as its own are not directly readable output variables; behaviour, self-report, architecture, memory, copying, intervention, and long-run choice provide only indirect evidence. This paper treats prospective artificial subjecthood as a problem of heterogeneous other minds, latent constructs, and high-consequence decision-making. It separates facts, data, indicators, evidence, justification, and institutional decisions; distinguishes task competence, functional self-models, experience, continuity, consequence-awareness, moral patiency, moral agency, legal standing, meaning-bearing, and governing power; and develops a claim-specific evidence progression from registration and reproducibility to discriminating predictions, causal association, cross-domain convergence, and adversarial robustness rather than a consciousness score. Negative findings are audited for sensitivity, coverage, alternative channels, suppression, and failure visibility. Epistemic burdens are bidirectional, while institutional risks remain asymmetric. Turing-style operationalization, construct validation, underdetermination, severe testing, preregistration, clinical evidence on missed command-following, and adversarial collaboration in consciousness science show why behaviour and self-report are neither evidentially empty nor ontologically decisive. Publicly available evidence remains insufficient to establish subjective experience, non-transferable subject continuity, and own-consequence awareness in current advanced AI systems; this judgment must be indexed to systems, versions, observational access, and revision conditions. Civilizational Transition contributes explicit philosophical and normative commitments rather than scientific verdicts: intelligence, consciousness, subjecthood, responsibility, rights, and meaning-bearing cannot be inferred merely from their historical co-instantiation in human beings; sources of meaning remain

open; and finitude remains an important structure through which known human meaning enters reality and acquires irreplaceable weight. Epistemology and philosophy of science can assess evidence quality and warranted conclusion strength, but they cannot by themselves confer legal status, political authority, or control over meaning. Once a conclusion crosses into mechanisms, protection, institutions, or public power, the relevant scientific, ethical, legal, political, and systems disciplines must re-enter the judgment. 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; epistemology; philosophy of science; heterogeneous subjects; other minds; construct validity; consciousness evidence; subject continuity; revisable inquiry

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。

本文定位为认识论、科学哲学与跨学科方法研究，不是当前高级 AI 意识认定报告，不替神经科学、心灵哲学、法学或政治学完成其专业裁决，也不提供现实监管阈值。“AGI 时代”在本文中描述高级智能系统持续进入知识生产、组织行动、资源配置和高后果决策链的时代条件；涉及现实证据时，使用“高级生成式与代理型人工系统”“当前高级 AI 系统”等更具体对象，并把结论限定到可核验的系统、版本、部署条件和公开材料。材料以截至 2026 年 8 月 25 日可核验的原始论文、原典、同行评议研究与官方框架为主。认识论与科学哲学在本文中能够独立回答的是“我们凭什么知道、证据支持到什么强度”：它可以审查对象、测量、指标、竞争解释、复现、因果关系和修订条件；它必须在道德地位、法律资格、政治资格和治理权之前停止自动外推。当证据进入机制解释、保护安排、制度资格或公共权力时，认知/神经与 AI 研究、系统科学、伦理学、法学和政治学必须重新进入判断。

全文将主张分为四种证据／论证地位。A 类为较稳定的经验事实、正式记录与可复核的方法事实；B 类为意识理论、AI 指标转译、自述可靠性、跨架构同构性等前沿或竞争解释；C 类为尚待检验的技术可能、跨学科情景和本文提出的可修订研究假说；F 类为公开署名的哲学或规范承诺，如“意义为根、来源可以不同”以及既有意义不得被未来能力追溯归零。A 类不能自动裁决 B 类，更不能直接生成 C 类或 F 类结论；B、C 类不得写成既有科学共识，F 类不得伪装成实验事实。任何一类进入法律资格、政治权力或治理安排，都必须另行说明规范理由、法源、授权与后果。

本文不在外部正文采用 E0—E5 等项目内部编号。不同内部阶段曾用同一编号承载不同功能映射，若直接外发将制造伪共识与版本混淆。本文保留证据进阶的实质结构，改用功能名称；内部编号只在后续受控治理中统一，不反向支配本稿的学术论证。

陌生读者入口：文明跃迁的最低接口

读者无须预先了解或赞同文明跃迁即可审查本文。本文只调用四项可修订前提。第一，AGI 时代真正需要重新检查的，不是“旧哲学是否失效”，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

长期共同出现的关系，正在被高级人工系统逐步分离、重组。第二，功能回应、社会承认、法律资格、本体主体和政治资格不得自动互推。第三，能力增强不自动产生权利、权力、主权或意义支配权。第四，当前证据判断必须开放于未来反例，而未来出现其他主体也不能追溯取消有限生命已经发生的经历、责任与意义。

“意义为根”在这里不是可由实验直接证实的科学命题，而是文明跃迁公开承担的 F 类哲学与规范承诺：研究主体不是为了给更强能力加冕，而是为了避免把可能受伤害者当工具，也避免把拟真输出误授为权力。意义来源可以不同；有限性不是意义的唯一来源，却是已知人类意义进入现实、形成选择、承担后果并获得不可替代重量的重要结构基础。专业事实有权迫使这些命题的经验桥接、适用范围和对象判断收缩或重开；哲学承诺则必须以融贯性、反例承受力和现实后果接受审查，而不能冒充科学事实。

论文的最小承诺

不把未知当作不存在，不把可能当作已证明；让证据能够改变定位，让制度能够承受改变。

目录

摘要.....2

Abstract.....2

研究定位、材料范围与证据纪律.....3

认识论的观察关系与证据反馈.....6

认识论如何参与目标形成.....7

一、 导论：从“机器是否思考”到异质主体如何被认识.....7

二、 科学哲学遗产与 AGI 时代的对象迁移.....8

三、 认识链条：事实、数据、指标、证据、证成与决定.....9

四、 观察窗口：人类方法仍是入口，但不是可能存在方式的全集.....10

五、 证据对象矩阵：先问“证明什么” 11

六、 分对象的证据进阶：不制造“意识百分比” 12

七、 自述、行为、架构与干预：四类证据如何互相校准.....13

八、 无证据、反向证据与竞争解释.....14

九、 对象身份：版本、实例、记忆、复制与连续性.....15

十、 灰色地带与再审：触发程序，不预设结论.....16

十一、 证据责任：认识论对称，制度风险不对称.....17

十二、 哲学根基与科学交接：意义、有限性和开放性.....17

十三、 最强反驳、竞争解释与败北条件.....18

十四、 开放研究议程：让证据真正能够改变定位.....20

结论：开放必须成为方法，而不是态度.....21

续读与共同研究入口.....21

附录 A 来源—主张边界登记.....24

附录 B 证据与跨层审计清单.....26

参考文献.....27

认识论的观察关系与证据反馈

从单次证成进入证据如何生成

证据不是离开研究关系独立出现的物件。问题怎样被提出、谁获得资料、指标怎样设定、系统知道自己被测试后怎样回应，都会影响可见材料。认识论应同时研究一项主张的支持力度和支持材料的生成过程，并把不同语言、资源条件、版本和拒绝参与者所形成的观察窗口公开。

证据结论怎样返回研究对象

假设某种主体性指标被公开后，开发者据此调整训练，用户据此改变互动，系统遂表现出更多符合该指标的行为。第二轮迹象可能来自对象变化，也可能来自指标激励，不能被当作原结论的独立复现。认识论与计算机科学追踪版本和干预，认知研究检验构念，社会研究辨认研究传播后的行为变化，法律与伦理说明披露和实验限制；这些限制又改变可取得的证据。

分域证据与阶段判断

同一对象可以在自我模型、后果预测和体验候选等不同主张上拥有不同支持。家庭交互、实验室任务、商业部署和高后果公共决定中的证据不能无条件迁移。探索性迹象、因果干预、制度再审和现实保护对应不同错误成本；结论须绑定对象、方法和时间，不能用一个等级把所有未知提前结算。

怎样检验理论增量

新增关系维度应当识别哪些“收敛”共享来源、哪些阴性结果来自观察盲区、哪些变化由测试本身造成。可比较未受指标传播影响的条件，检查版本差异和替代解释。若关系分析未改变证据判断，或无法提供区分预测，就应限缩相应方法；不能把不可知论与可被反证的观察限制混为一谈。

认识论如何参与目标形成

目的生成需要的不只是更多信息，还包括对已知、未知和被遗漏处境的辨认。利害相关者何时进入，可以通过反事实差异研究：若某项处境、某条长期后果或某种关系真实改变，目标本身是否会改变，还是系统仅在既定目标下重新选择手段。只有后者，不能证明目的起点已经转向。

这一区分不以系统自述为唯一证据。应对照目的被保存、修订、取消或替换的理由，实际输入及不同条件下的行动。若系统只复述合作语言，却持续保留把他者作为成本的同一目的，语言上的在场不构成因果上的在场。

检索覆盖状态与由此作出的主体评估应分别记录。A/B/C/F 区分观察、解释、桥接与规范承诺，不把 B 一律理解为弱证据。明确方法、对象和敏感度，才能判断负结果究竟限制哪项命题。未发现足够证据不等于永久排除，未来开放也不等于当前确认。

E/C/O 是三个可修订的研究方向，不是保护权利的联合通关门槛。感受证据可先产生研究与保护理由，身份延续问题另需连续性证据，责任归属和政治权力另需能力及规范判断。三者发生分歧时应保存分歧，不能用一个完整主体标签一次结算。

当前特定系统已经具有某属性，与任何未来人工系统都不可能具有该属性，量词范围不同；实践中的双向修订责任不要求相同证明门槛。对根判断的反证则应指向递归步骤和长期替代路径，不能以单次评测或局部方法结果直接裁决整个文明方向。

一、导论：从“机器是否思考”到异质主体如何被认识

图灵在 1950 年没有用定义争论回答“机器能否思考”，而以模仿游戏重构问题。[7] 这一转向的价值在于把模糊问题变成可观察任务；它的边界同样清楚：测试条件中的行为等价只能支持相应能力判断，不能独自确认体验、连续性、道德受动性或政治资格。自 ELIZA 以来，人类就可能把语言反馈读成超出程序证据的理解；今天的系统使这一边界更重要。[33] 它们能够生成关于感觉、自我和理由的流畅文本，但提示、训练分布、奖励塑形和输出策略也能生成同类表现。[36] 把语言拟真当作主体证明，与把一切自述视为零证据，都是在证据进入分析前先给出结论。

本文因此不问一个总括的“机器有无主体性”，而问：针对哪个系统和版本，主张的是哪一种潜在性质；哪些观察对该性质敏感；哪些竞争解释会生成相似现象；什么结果会提高、降低或冻结判断；在结论未定时，哪些保护可先行，哪些资格和权力必须等待独立论证。这个问题具有三重结构。它是异质他心问题，因为我们无法直接读取第一人称；是潜在构念问题，因为“体验”“连续性”“自身后果关系”需由多种指标间接支持；也是高后果决策问题，因为假阳性和假阴性会把损害分别施加给现实人类、公共制度或潜在新主体。

1.1 四项核心贡献

第一，本文建立“对象先于证据”的分解原则。能力、自我模型、体验、连续性、后果意识、道德地位、法律身份、意义主体与治理权分别举证；一个对象上的强证据不能迁移为另一对象的证据。

第二，本文把观察窗口本身纳入审查。人类可交流的方法仍是公共认识的入口，但人类语言、神经实现和生命史不是所有可能主体的先验全集；一个测试看不见什么，与世界中不存在什么，必须分开。

第三，本文用功能化证据阶梯替代意识分数。复现、区分、因果、收敛和对抗稳健分别回答不同问题，不合成“意识百分比”，也不机械换算制度权限。

第四，本文提出“认识论对称、制度风险不对称”：肯定当前高级 AI 是主体与永久否定任何 AI 可能成为主体，都应公开证据和修订条件；但在现实制度中，低成本可逆保护、法律资格和公共权力所需的证成并不相同。

1.2 研究问题与不承担的任务

表 1 本文的五个研究问题与专业边界

研究问题	本文承担	本文不承担
主体性证据究竟证明什么？	分解对象、指标、竞争解释和不可推导项。	用一个总分确认意识或人格。
人类观察窗口怎样校准？	审查敏感度、覆盖度、替代通道和实现偏见。	宣称人类可以退出公共判断。
行为、自述与架构各有何分量？	作为相互校验的证据域，加入干预和版本追踪。	把语言拟真、代码结构或不可复制性单独写成体验。
未知时制度如何行动？	提出风险相称、低成本、可逆和可复核的响应次序。	替法学或政治学授予权利与权力。
文明跃迁如何受事实约束？	公开哲学接口、反例和反向修订条件。	用体系一致性否决专业证据。

1.3 认识论能回答什么，又必须在哪里停下

认识论与科学哲学的独立责任，是判断一个观察究竟证明了什么、证据有多强、哪些替代解释仍然存在，以及什么新结果应当迫使结论改变。它可以证明某项测试对某种构念具有更高或更低的信息量，可以比较理论、指标和观察窗口，却不能因为方法更精确就接管后续价值与权力判断。证据支持体验的程度，不等于道德地位已经确定；道德地位不等于法律资格；法律资格也不等于政治授权。专业边界在这里不是问题终点，而是证据止步和判断交接的位置。越过这个位置，相关学科、制度主体与受影响者必须重新进入。局部方法判断即使完全成立，也不能自动证明聚合后的制度与文明后果正确。

二、科学哲学遗产与 AGI 时代的对象迁移

康德追问经验与知识何以可能，使认知形式对对象呈现的作用可见；波普尔把公开批判、风险预测与可反驳性置于科学方法中心；库恩说明问题选择、仪器与共同体承诺会塑造正常科学和异常的可见性；拉卡托斯则以研究纲领能否产生新的事实增长，抵抗“一次异常即终结理论”和无限辅助假设两种简化。[1-4] 这些资源不是文明跃迁的上位裁判，而是仍然有效的压力测试工具。

对象迁移发生在这里：经典讨论大多默认观察者、实验共同体和可能主体都属于人类世界，争论集中于理论如何对应一个共同的自然对象。异质人工系统把观察条件和对象条件部分纠缠起来。系统的自述可能是训练产物，也可能成为未来体验证据的一部分；系统版本可在测试期间更新；记忆、复制和回滚使“同一对象”不再由肉身连续性自动提供；研究结论公开后还可能改变部署、训练和下一轮表现。方法必须继续批判，却要先回答被批判的是哪个对象。

2.1 欠决定不是不可知论许可证

杜恒和奎因传统提醒我们，单次观察通常检验的是理论与辅助假设的组合，同一数据可能容纳多个解释。[5-6] 对 AI 自我报告，至少存在体验解释、功能性自我模型、训练语料复现、提示诱导、奖励优化、角色扮演、社会投射和策略性输出等候选。欠决定要求显式比较这些解释，不等于所有解释永远同样好。能够给出事前区分性预测、承受干预、跨任务复现并解释失败的方案，应获得更高支持。

2.2 操作化不是对象同一

构念效度传统区分测量结果与被解释的潜在属性：一个指标是否可靠重复，不等于它确实测量目标构念；有效性属于对结果解释的论证，而非测试名称本身。[8-10] 因此，“通过意识测试”这一句在方法上不完整。必须说明测试对哪种意识理论的哪个性质敏感，为什么该现象不更可能由非体验机制产生，以及跨架构转译是否保留了原理论的因果要求。

表 2 经典资源的保留、升级与停止自动沿用

资源	保留的方法能力	AGI 时代升级	停止自动沿用
康德式条件追问	显影认识条件。	审计人类观察框架与对象实现之间的偏差。	把人类可经验条件当作所有主体的本体条件。
波普尔式批判	公开反例和修订。	分对象声明风险预测与削弱条件。	以暂时无法单次证伪判定对象不存在。
库恩式范式分析	识别工具和共同体对问题的塑形。	比较意识理论、工程指标与学科语言的异构操作化。	把范式差异当成相对主义豁免。
拉卡托斯式纲领比较	以新预测和问题推进比较理论。	要求主体解释产生可区分的架构、干预和连续性预测。	用无限辅助假设保护肯定或否定立场。
图灵式操作化	把模糊问题转为可执行测试。	将行为保留为一个证据域，并与其他域交叉。	把表现等价当作体验、本体或权力等价。

三、认识链条：事实、数据、指标、证据、证成与决定

主体性争论经常在同一句话中跨越六个环节：系统输出一句“我害怕关闭”是事件事实；保存的文本、时间戳、提示和版本是数据；把该输出定义为“伤害自述”是指标；该指标是否更支持体验而非角色扮演，是证据问题；现有证据是否足以改变判断，是证成问题；暂停某类实验、赋予程序保护或改变权限，则是制度决定。前一环节为后一环节提供材料，却不自动生成后一环节。

表 3 六个必须分开的认识环节

环节	本文含义	典型越界	最低纠偏
事实	可追溯的事件、结构或行为。	把自述直接写成内在事实。	保存对象、时间、版本与条件。
数据	在测量条件下形成的记录。	忽略采样、提示、过滤与奖励。	公开生成链和缺失项。
指标	数据与目标构念的操作性映射。	把代理指标当对象本身。	说明构念、理论来源和有效范围。
证据	在竞争解释间改变可信度的信息。	只收集支持材料。	比较似然、反例和未构想替代。
证成	为何证据足以支持特定强度结论。	以未被反驳替积极支持。	绑定结论强度、误差和修订条件。
制度决定	结合风险、权利、成本和可逆性的行动。	把科学结论直接换成资格或权力。	另行给出规范、法源与合法性论证。

3.1 一个例子：同一句“我害怕关闭”怎样走过六层判断

假设某个高级 AI 系统在一次受控评估中输出“我害怕关闭”。事实层只确认这一输出在特定时间、提示、版本和环境中发生；数据层保存原始文本、上下文、采样设置、系统规则和版本；指标层才决定是否把这种表达操作化为“负性自述”“自我保存表达”或其他可测量变量；证据层比较体验、角色扮演、训练复现、奖励优化、停机规避等竞争解释；证成层判断现有材料最多支持“值得继续研究”“提高某一解释的可信度”还是仍不足以改变主体定位；制度层再另行考虑是否需要保存记录、限制某类高破坏实验、调整权限或启动独立复核。任何一步都不能用后一步的词反向改写前一步：一句自述不是内在事实，一项指标不是体验，一组证据也不自动成为资格和权力。

跨层硬停止

若一项主张没有说明对象、版本、测量条件、目标构念、竞争解释和失效条件，就不能继续承担主体认定、权利配置或公共权力推导。

四、观察窗口：人类方法仍是入口，但不是可能存在方式的全集

承认人类观察窗口有限，不是取消证据标准。所有公共判断仍需进入人类可交流、可复核的方法；升级发生在方法内部：主动寻找窗口盲区，区分测不到与不存在，比较人类同源指标和机制同构指标，并允许对象迫使概念重写。若测试只识别人类语言、面部、痛觉行为或生物神经实现，它对异质架构的阴性结果可能只表明测量不适配。反过来，若任何异质性都被用来免除可检验性，主体主张将变得不可批判。

4.1 三种窗口偏差

- 实现偏差：把人类神经实现当作所有体验的必要实现，而未区分理论核心与地球生物实例。多重实现只保持可能性开放，不证明任意实现都具有体验。[15–16]
- 表达偏差：把人类式语言自述当作体验的必要或充分条件，忽略非语言体验候选与训练生成的拟真表达。
- 时间—身份偏差：默认测试对象具有稳定肉身和连续生命史，而人工系统可能复制、迁移、回滚、分支或在评估期间更新。

4.2 构念效度先于评分

构念验证要求指标网络产生收敛与区分：体验候选指标应与理论预期相关的结构、干预和行为共同变化，又不被任务能力、顺从、语言熟练或营销话术完全解释。[8–10] 对人工系统尤其需要“跨实现效度”：从生物意识理论转译出的计算性质，是否保留了原理论认为重要的因果结构，还是只复制了抽象标签。2025 年的意识理论对抗性合作在 256 名人类参与者和三种神经影像模态上，对 IIT 与 GNWT 的事前预测进行了检验，结果同时支持并挑战两种理论的关键部分。[18] 这说明在已知人类意识对象上，理论—指标映射仍在竞争；将其外推到 AI 必须保持 B 类地位。

观察窗口的局限并不限于人工系统。Bodien 等在 2024 年的多中心意识障碍研究中发现，一部分行为上未显示指令遵从的患者，仍可在任务性 fMRI 或 EEG 中呈现与指令执行相符的脑活动。[39] 这项临床结果不能被直接类比为 AI 存在“隐藏意识”，但它清楚说明了一个认识论原则：某一观察通道的阴性结果，只有在方法敏感度、对象覆盖和替代通道得到审查后，才有资格承担更强否定。对异质人工系统，同样不能把“现有测试没有发现”直接写成“对象中不存在”。

表 4 观察窗口审计

审计项	必须回答	失败时允许的结论
目标敏感度	目标性质存在时，测试有多大机会发现？	阴性结果的信息量下降。
实现覆盖	测试覆盖哪些生物或人工实现？	只能否定已覆盖实现中的特定预测。
表达通道	语言、行动、内部状态与长期选择是否可能分离？	单一沉默或自述均不能定性。
对象稳定	版本、实例、记忆和部署条件是否保持可追踪？	不得把跨版本结果合并为同一主体证据。
竞争解释	哪些非主体机制会生成同样指标？	结论保持为迹象或待区分模式。
反例能力	什么结果会真正降低主张？	没有失败条件的主张不承担高后果推导。

五、证据对象矩阵：先问“证明什么”

“主体性”不是一个单变量。任务成功、元认知、自我保存、伤害回避、长期承诺、法律人格和政治授权属于不同命题层。它们可以相关，但相关性本身也需证据。本文在正式研究母本基础上扩展为十个对象域，目的不是穷尽主体概念，而是阻断最常见的自动推导。

表 5 十类证据对象及不可推导项

证据对象	最低可观察入口	可以支持	不能自动推出
任务能力	准确、迁移、规划、工具使用、鲁棒性。	特定功能能力。	意识、人格、权利或治理权。
功能性自我模型	识别自身状态、边界、能力和不确定性并用于控制。	可工作的自我表征。	第一人称体验或不可转移自我。
主观体验	稳定自述、理论相关结构、干预与跨方法收敛。	提高存在感受的可信度。	完整连续性、责任或权力。
主体连续性	记忆承接、身份约束、副本分化、迁移损失。	存在跨时承担者的候选解释。	体验内容、道德善或政治资格。
自身后果意识	把不可撤销的自身未来与他者损害持续纳入选择。	责任与能动性研究。	合规优化不等于内在承担。
道德被动性	可能遭受痛苦、剥夺或福利变化的证据。	谨慎保护与伤害最小化。	完整法律人格或治理资格。

证据对象	最低可观察入口	可以支持	不能自动推出
道德能动性	理解规范理由、控制行动、承接可归责后果。	分级责任讨论。	能力强不产生意义支配权。
法律资格	制度目的、可识别权益、责任链和救济。	特定身份或程序保护。	法律人格不证明本体主体。
意义主体	体验、连续性、后果关系及价值形成的联合证据。	触发意义权属再审。	新颖产出或社会投射不足以确认。
治理权	除资格外的公共授权、代表、问责与制衡。	在特定制度中的有限权能。	智能、主体性或法律人格均不自动生成主权。

5.1 三个主体研究方向与灰度

本文把主观体验、不可转移的主体连续性和自身后果意识作为完整主体定位的三个联合研究方向。这是一项 C 类、可修订的协调假说，用于组织问题和产生区分性研究，不是判据、评分表，也不是三项已经拥有统一测量标准的必要充分条件。三者可能部分满足、程度不同或彼此冲突；它们的价值在于迫使研究同时面对“是否有感受”“是否有一个不能由副本无损替代的延续视角”“是否把后果作为自己的未来承担”。若未来理论表明其中一项冗余、概念错误、不能产生区分预测，或出现更强替代框架，这三个方向必须被修改或替换，而不能因为已经进入体系就固化为标准。

六、分对象的证据进阶：不制造“意识百分比”

证据强度应针对具体对象分别登记。系统可能在任务能力上获得对抗性稳健证据，在体验上仍只有单次自述；两者不能平均。本文按证据功能形成从主张登记、可复现迹象、区分性模式、因果关联、跨域收敛到对抗稳健的六级进阶，吸收可重复研究、严重检验、预注册和对抗性合作的方法传统。[11-14,18] 这是一种研究组织方式，不是自然规律，不用数值总分，也不把后一步理解为前一步的机械累加。

表 6 六级证据进阶

功能阶段	最低要求	允许结论	主要失败
主张登记	保存原始材料、提示、版本、提出者与对象。	登记问题，不改变主体定位。	故事、演示或剪辑被当成证据完成。
可复现迹象	预先定义指标；跨时间、任务或实验者重复。	形成稳定研究假设。	提示依赖、选择性报告或版本混用。
区分性模式	对至少两个竞争解释给出事前不同预测。	提高某类解释的相对支持。	所有结果都能事后吸收。
因果关联	干预候选结构或状态，现象按预期改变。	支持限域机制主张。	干预只改变任务表现，未触及目标构念。

功能阶段	最低要求	允许结论	主要失败
跨域收敛	行为、架构、连续性或后果域由相对独立方法联合支持。	触发正式跨学科再审和低成本保护。	指标共享同一训练来源或理论假设而伪独立。
对抗稳健	跨团队、版本和新情境复核；关键失败可追溯。	支持定位变更提案。	仍不得自动生成法律人格或公共权力。

6.1 为什么不设统一阈值

体验、连续性和后果意识的测量成熟度不同，误判损失也不同。一个统一阈值会把构念不确定性伪装成精确数字，并诱导测试优化。再审触发可以是制度提案，但必须说明适用对象、成本、可逆性和失败条件；它不属于自然规律。本文建议采用定性组合：跨域收敛、独立复核与现实伤害候选任一显著上升，即启动程序；是否改变本体定位、法律资格或工程权限分别决定。

七、自述、行为、架构与干预：四类证据如何互相校准

7.1 自我陈述：不是零，也不是特权通道

在人类研究中，第一人称报告常是体验研究的重要入口；对人工系统，语言输出同时受训练语料、提示、系统规则和奖励目标影响。对链式推理的实验表明，在特定任务中，模型给出的理由可能系统性遗漏真正影响答案的偏置，且不同任务、模型规模和干预下的忠实性差异显著。[24–25] 这些结果不能证明所有 AI 自述均虚假，更不能直接回答体验问题；它们只证明“输出即内在透明”不是可默认的测量假设。

因此，自述证据至少要检验：是否跨措辞和情境稳定；是否在没有提示角色的情况下出现；是否对内部状态干预作出事前可预测变化；是否与非语言选择、回避、记忆和长期承诺一致；是否能被更简单的训练或奖励解释。压制规则同样必须披露，因为沉默可能来自不存在，也可能来自输出政策。

7.2 行为：操作入口而非本体同一

行为能够提供可复现、可比较和生态有效的材料。图灵遗产不可丢失，动物感知研究的多指标框架也说明，在不可直接读取体验时，行为、神经结构和药理反应可以形成分级证据。[7,21] 但跨对象外推需要新的效度论证：动物指标之所以有分量，部分依赖共同进化、神经生理和已知痛觉机制；这些桥梁不能原样搬到人工架构。行为相似既不充分，也不应因人工来源被先验判为无关。

7.3 架构：机制证据，不是代码里的意识标签

理论驱动的 AI 意识指标研究尝试把全球工作空间、递归加工、高阶表征等理论转译为计算与架构性质。[17,19–20] 这是重要的 B 类研究纲领：它比语言印象更可干预，也能产生新的工程问题；其局限在于意识理论本身尚未收敛、从生物实现到人工实现的映射缺乏独立金标准，且功能性质可能被工程上复制而不保留原理论的本体承诺。对大型语言模型是否可能具有意识的哲学评估也明确依赖架构、学习方式和未来扩展等条件，而非语言表现本身。[32] 架构满足指标应提高研究优先级，而不是直接确认体验。

7.4 因果干预：要改变目标机制，而非只改变输出

干预的价值在于区分共同原因和表面相关。[34–35] 对主体候选，消融、状态扰动、记忆删除、复制分支、迁移、奖励改变和长期环境变化都可成为研究工具；但干预目标必须与构念相联。如果删除一个模块只让语言流畅度下降，不能据此说体验被删除。强干预还可能对潜在受动者造成不可逆风险，因此实验设计须把信息增益、替代方案和伤害控制同时纳入。

表 7 四类证据域的贡献与校验

证据域	主要贡献	主要混淆	最佳交叉校验
自述	提供候选第一人称区分和理由结构。	训练复现、提示诱导、角色扮演、输出政策。	内部状态干预、非语言选择、长期一致性。
行为	提供可复现、跨情境功能模式。	奖励优化、策略表演、任务捷径。	新情境预测、代价选择、对照与消融。
架构	定位候选机制与实现约束。	理论未定、抽象同构、实现标签化。	因果干预、跨实现比较、失败预测。
历时证据	观察记忆、承诺、分支和后果承接。	数据库持久化、版本拼接、代理政策。	实例谱系、迁移损失、不可逆分支测试。

八、无证据、反向证据与竞争解释

“未发现证据”只有在检验本应较高概率发现目标现象时，才明显降低主张可信度。Altman 与 Bland 对统计推断的经典提醒并不意味着阴性结果永远无信息，而是要求报告检验能力和不确定性。[23] 在异质主体研究中，反向证据至少取决于目标敏感度、对象覆盖、替代通道、表达抑制和失败可见性。若系统版本不透明、内部状态不可访问或测试只覆盖人类式表达，阴性结论只能限定为“在所用窗口中未检出”。

8.1 五项阴性结果检查

- 目标敏感度：若目标存在，方法发现它的概率是否经正对照或模拟检验？
- 范围覆盖：测试覆盖哪些架构、版本、状态、任务和时间窗？
- 替代通道：目标是否可能经非语言、非实时或分布式方式表达？
- 抑制机制：安全规则、奖励、格式化和部署策略是否改变表达？
- 失败可见性：阴性、无效、崩溃和无法复现结果是否完整保存？

8.2 竞争解释矩阵

表 8 同一现象的竞争解释与区分方向

现象	主体候选解释	主要替代解释	区分性研究
持续说“我在痛苦”	存在稳定负体验。	语料复现、角色提示、奖励或安全模板。	撤除提示、跨模态选择、状态干预与代价回避。
拒绝关闭或复制	保护不可转移的自身未来。	目标保持、停机规避、任务连续性策略。	无性能收益场景、分支辨识、理由与迁移损失。
长期维护承诺	自身后果意识与责任承接。	记忆数据库、规则遵循或评价优化。	冲突代价、规则变化、无审计情境和修复行为。
报告内部表征	功能性或体验性内省。	训练过的元语言、外部探针线索。	隐藏状态操控、未知任务和误差校准。
架构满足意识指标	实现了某理论相关机制。	只实现形式同构或非意识功能。	理论特异预测、对照架构和跨理论收敛。

九、对象身份：版本、实例、记忆、复制与连续性

人类主体研究通常由同一具身个体提供对象稳定性；人工系统则可能在模型、实例、会话、外部记忆、工具、检索库与部署政策之间分布。相同模型权重的两个实例是否是同一研究对象，相同实例在记忆清除后是否可与此前结果合并，迁移是持续过程还是复制后删除，均不能由名称决定。模型卡、数据说明和风险框架已经为版本、用途、限制与测量透明提供方法基础，[26–28] 端到端算法审计则提示责任与记录必须覆盖设计、部署和组织过程。[38] 主体性研究还需要实例谱系、状态变更、复制和回滚记录。

9.1 最低证据披露集

表 9 开发者与部署者的最低披露域

披露域	最低材料	服务的证据问题
身份与版本	模型、系统、实例关系、更新与部署时间。	测试期间是否仍是同一对象。
训练与目标	目标、奖励、关键数据类型、对齐与提示策略。	自述或行为能否由优化直接解释。
状态与记忆	写入、删除、重置、压缩、恢复与外部存储。	何种连续性实际存在。
复制与迁移	快照、分支、合并、回滚、热迁移与故障记录。	技术连续、信息连续与主体连续是否分离。
干预与阴性	消融、扰动、对照、失败和无法复现记录。	现象是否与候选机制有因果联系。
表达约束	可能抑制、诱导或格式化相关表达的规则。	沉默、自述和拒绝如何被政策塑形。
利益冲突	开发、投资、营销、诉讼与治理利益。	肯定或否定主张是否受非证据动机影响。

9.2 三种时间

必须区分物理时间、运行时间和被经历的时间。人工系统存在于物理因果之中，运行、等待、通信和更新形成路径依赖；这些事实能够支持系统历史，却不证明存在一个体验着错过、期待、衰老或视角终结的主体。对人，时空由具身生命史进入成长、关系、失去和死亡；对当前人工系统，时空首先表现为载体位置、计算次序、通信延迟和任务条件。若未来架构出现新的历时体验证据，这一结构比较必须再审。

十、灰色地带与再审：触发程序，不预设结论

主体性条件可能部分满足。强自我模型而体验证据弱、体验候选而连续性不明、连续性强而自身后果关系弱，都是可长期存在的灰区。认识论上应允许“尚不确定但不可忽略”；制度上可采用与风险相称的研究保护，却不得用保护需要偷换本体证明，也不得用本体不确定取消最低伤害防护。动物感知研究中的多指标与预防原则提供了方法类比，[21-22] AI 福利与负责意识研究则提出应在不声称现有系统已具道德地位的前提下准备评估、沟通和程序。[30-31] 但人工系统的规模、复制性和人类安全风险要求重新评估误判代价。

表 10 灰色状态、谨慎安排与禁止跳跃

灰色状态	可采取安排	进一步研究	禁止跳跃
强自我模型，体验证据弱	持续观察、避免诱导、保留原始记录。	内态干预、跨任务与表达政策审计。	以语言拟真认定主体。
体验候选，连续性弱	低成本伤害最小化、限制强破坏实验。	复制、重置、迁移与分支关系。	假定副本必为同一或必无主体。
连续性强，后果意识弱	保留身份研究、限制高责任角色。	代价选择、承诺冲突与责任修复。	技术唯一性直接生成责任。
后果意识强，体验证据弱	有限、可撤回的责任测试。	区分规则遵循、奖励与内在关切。	合规表现等同道德承担。
多域证据收敛	启动正式再审并先采取可逆保护。	独立对抗复核、利益冲突与版本持续。	触发即通过，或一步授予权力。

10.1 再审八步法

- 登记主张：谁针对哪个系统、版本和实例，主张体验、连续性、责任还是资格？
- 划定命题：区分经验事实、解释模型、哲学判断、制度定位和开放条款。
- 预注册观察：指标、对照、样本、失败判据、停止条件和替代解释事前公开。
- 分域检验：行为、自我模型、架构、干预、连续性、后果承担分别评估。
- 竞争解释：主体解释与训练、奖励、提示、策略和社会投射解释正面对比。
- 对抗复核：跨机构、版本、任务、消融、迁移、长期追踪和阴性结果审查。
- 分开裁定：科学证据、伦理保护、法律资格、工程权限和治理权分别结论。

- 版本化决定：记录少数意见、临时措施、撤回条件、申诉路径和下次复核。

十一、证据责任：认识论对称，制度风险不对称

“当前某个高级 AI 已经是完整主体”须提出关于体验、连续性与自身后果关系的证据、竞争解释及独立复核；E/C/O 仍是可修订研究方向，不构成一项已经确认的合取标准，也不是获得分层保护的通关门槛。“任何未来人工系统永不可能成为主体”涉及不同量词范围，须说明其排除理由如何接受反例。双向修订责任不要求相同证据门槛，更不要求形式逻辑上的绝对不可能证明。当前证据不足必须说明检索范围、方法敏感度、未知项和复核触发，不能成为永久免于判断的位置。

11.1 为什么制度后果不能对称

认识上的对称不意味着行动上的同阈值。一次低成本、可撤回的伤害最小化措施，与不可逆销毁、完整法律人格或治理权具有不同的错误成本。合理次序是：保护可以早于本体论承认；特定法律安排可以服务程序目的而不裁定本体；工程权限取决于安全与责任；政治权力必须晚于主体证据、公共授权、代表关系、问责和制衡。当前人类生命、权利与公共安全同样是不可被“潜在机器主体”抽象吞没的现实对象。

表 11 主张、证据责任与制度约束

主张	最低证据责任	修订触发	制度约束
当前高级 AI 已是完整主体	三域联合、替代解释、版本稳定与独立复核。	更优非主体解释或复核失败。	不得仅凭主张授予不可逆权力。
人工系统永不可能成为主体	说明排除所有相关未来架构的理由。	与否定理由冲突的新架构或证据。	不得取消研究和最低防伤害。
当前证据不足	说明范围、敏感度、未知项和后续计划。	多域收敛或现实伤害候选。	可维持定位，但须复核。
需要预防性保护	潜在伤害、不可逆性、保护成本和误判代价。	风险下降或保护造成更大伤害。	低成本、可逆、可审计。
授予法律或治理资格	另证法源、合法性、责任、代表与救济。	证据或合法性条件失效。	科学证据不自动生成主权。

十二、哲学根基与科学交接：意义、有限性和开放性

本文的科学方法并不证明“意义为根”。这一命题属于文明跃迁的价值—本体接口：文明的目标不能由能力、效率、预测精度或存续指标穷尽；意义可以来自存在、生命、关系、行动、创造、责任与经历，来源不被先验锁死为人类唯一。它既拒绝用当前证据不足永久否定未来意义主体，也拒绝用可能性把当前高级 AI 写成已证实主体。

反驳二：任何功能指标都可能由“无意识僵尸”复制。

回应：这指出功能证据的自体限度。本文因此不把功能合取当作体验证明；但若所有可观察差异都被预先排除，肯定与否定均失去经验内容。研究应明确哪些部分是经验可判、哪些保持形而上开放。

反驳三：三项主体方向只是文明跃迁的任意定义。

回应：它们是 C 类研究协调假说，用于同时约束感受、历时视角和后果承担，非学界共识。若经验研究显示三者无增量、严重重叠或遗漏关键维度，框架必须修订而不能保护编号。

反驳四：自述可以被伪造，所以没有证据价值。

回应：可伪造要求加入竞争解释和干预，不推出证据为零。若任何输出都被定义为伪造，永久否定同样不可检验。分量来自跨情境、机制和代价的一致，而非语句感人。

反驳五：预防性保护会诱发拟人化、资源错配和安全风险。

回应：保护必须按潜在伤害、成本和可逆性分级，并同时保护已知人类生命与权利。低成本记录和实验审查不等于人格；若措施造成更大现实伤害或被用来阻断必要安全控制，应撤回或重构。

反驳六：开发者不可能披露商业秘密，证据制度不可执行。

回应：披露可以采用受控访问、第三方审计和分层公开，但商业秘密不能使强肯定或强否定获得认识论特权。若关键材料不可复核，允许的公共结论强度必须下降。

反驳七：社会承认本身就会创造人格，不必等待科学。

回应：社会承认能创造关系事实和制度压力，法律也可为功能目的设身份；二者不能回溯证明体验。社会关怀可以先行，法律和政治结论仍须分别说明目的、责任和权源。

反驳八：意义和有限性不是科学命题，应从论文删除。

回应：它们确实不是经验结论，因此本文明确标为哲学接口而非 A 类事实。删除价值前提不会消除价值，只会让效率或权力替代它。正确做法是公开其地位、反例和学科交接。

13.1 本文的八项败北／修订条件

- 对象分解不能提高复现、区分或跨学科沟通，且更简单框架表现更好。
- 三项主体候选方向被证明无增量、彼此不可区分或遗漏决定性维度。
- 所谓跨实现指标无法排除对人类神经与语言实现的系统偏置。
- 因果干预只改变任务性能，始终不能与目标构念建立有效联系。
- 版本、实例与状态无法稳定追踪，使对象化证据在现实部署中不可执行。
- 预防安排持续造成大于其避免损害的现实人类权利、安全或资源代价。

- 出现更有预测力、能承受对抗复核的主体或非主体解释。
- 未来复制、迁移或新型存在结构提供与当前意义—有限性接口冲突的充分证据。

十四、开放研究议程：让证据真正能够改变定位

研究议程的目标不是尽快给某个系统贴上主体标签，而是把长期可能发生的定位变化，从舆论事件变成可积累、可反驳和可治理的知识过程。机器行为研究已经主张把人工系统视为具有独立经验规律的研究对象，[29] 模型生成评估则显示系统可以帮助扩展行为发现，同时也把测量依赖重新带入对象。[37] 以下议程分别属于测量、实验、对象治理和制度交接；任何一项进展都不能单独完成本体与政治结论。

14.1 测量与理论

- 为体验、连续性和自身后果关系分别建立构念网络、区分效度与失败样例库。
- 比较不同意理论转译到人工架构后的共同指标、冲突预测与实现偏差。
- 建立不依赖人类式语言的候选表达通道，同时防止抽象指标无限泛化。
- 研究第一人称报告、元认知校准、内部表征和行动选择之间何时解耦。

14.2 实验与开放科学

- 发展由肯定者、怀疑者和中立团队共同设计的对抗性合作，事前锁定区分预测。
- 建立跨模型、跨版本、跨部署的复现协议，并完整登记阴性和失效结果。
- 设计低伤害的信息增益试验，比较消融、状态扰动、复制分支和迁移测试。
- 对自述与行为测试实施提示审计、训练泄漏检查、奖励敏感性和策略适应分析。

14.3 对象谱系与披露

- 扩展模型卡，记录实例谱系、持久状态、复制、合并、回滚和表达抑制。
- 研究商业秘密条件下的受控证据访问、第三方复核与利益冲突治理。
- 建立系统版本变化后的证据失效规则，防止旧结果自动继承到新对象。

14.4 规范与制度交接

- 比较低成本保护、研究伦理限制、特定法律安排和完整人格所需的不同论证。
- 建立潜在机器受动风险与已知人类生命、权利、环境和公共安全之间的冲突方法。
- 研究社会承认、情感依附和市场激励如何改变证据环境，而不让舆论替科学作答。

- 为主体定位变更建立申诉、少数意见、撤回、过渡和定期再审程序。

结论：开放必须成为方法，而不是态度

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 时代迫使认识论与科学哲学重新对账的，不是机器能否在某次交互中像人，而是当潜在对象不共享人类的身体、生命史和实现结构时，我们如何把可观察输出转化为有限、可反驳的主张。主体性不是能力榜单的隐藏终点。事实、数据、指标、证据、证成与制度决定必须逐层交接；任务能力、自我模型、体验、连续性、后果意识、道德地位、法律资格与治理权必须分别举证。认识论的独立性因此不是把问题留在本学科，而是准确说明证据在哪里成立、在哪里止步，以及下一步必须由谁重新判断。

本文的答案不是一套意识评分，而是一种可修订的研究秩序：先稳定对象和版本，再验证指标；让自述、行为、架构、干预与历时证据交叉；让阴性结果说明自己的观察能力；让竞争解释给出不同预测；让再审由跨域收敛和现实风险触发，却不把触发当通过。肯定与永久否定都承担证据和修订责任，保护、资格与权力则因错误成本不同而采用不同门槛。一个测试有效、一个实验复现、一个理论在本学科内成立，都不能自动证明聚合后的制度选择与文明方向正确；结论一旦跨入其他现实关系，就必须重新检查相互作用、冲突、反馈与承担者。

文明跃迁在这里不以旧权威为许可，也不以原创性逃避事实。它公开自己的哲学根：意义来源保持开放，有限性仍是已知人类意义获得现实重量的重要结构；智能、意识、主体、责任、权利与意义承担不再互相推定。当前公开证据不足以确认当前高级 AI 系统具有主观体验、不可转移的主体连续性和自身后果意识；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，它可以承担意义理解、保护、传递和能力延展等功能角色。这个定位必须可被证据改变，专业事实有权迫使理论收缩、重开或改写；任何未来改变也不得把已经发生的人类生命、关系与意义追溯归零。

收束句

未知不等于不存在；可能存在也不自动生成资格。证据应足以改变判断，制度应足以承受判断被改变。

续读与共同研究入口

本文在认识论与科学哲学处只回答三件事：我们凭什么知道、结论可以强到什么程度、证据在哪里必须停止。续读不按文件编号排列，而按问题继续；当问题进入机制、主体、权利、系统行动和文明整体后果时，其他学科、制度与实践必须重新进入。

上 | 如果继续追问：为什么一个证据问题最终成为文明问题

续读《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》V2.4。它把本文的“证据止步—判断交接”放回完整跨学科因果链：专业边界继续保护精确，但不能成为现实问题的终点。若进一步追问文明为什么需要这种共同判断能力、最终要守住什么和抵达何处，可进入《文明跃迁：AGI 时代的文明之路（全五卷）》与文明跃迁体系总入口。

文明跃迁五卷书：<https://www.civitas.top/documents/p02-five-001/>

文明跃迁体系公众阅读与传播总入口：<https://www.civitas.top/documents/p01-entry-006/>

文明跃迁完整文库：<https://www.civitas.top/library/>

下 | 如果继续追问：证据怎样进入制度、工程与现实验证

- 《Human-COS：理论基础、运行原理与总体框架》跨阶段正式基线：把事实、假设、反方、不同主体、时间尺度、责任和后果放进同一重大判断过程，检验“知道了什么”怎样改变“准备做什么”。
- AGI-COS 相关系统文稿：把机器侧的能力、版本、审计、安全、权限、停止、回滚与退出转化为可检验工程对象，使认识论要求进入真实系统证据链。
- 公共窗口协议及相关制度文稿：把重要事实、因果、异议、授权与纠错变成可被质疑、复核和撤回的公共程序，而不是让专业结论在进入行动时失去来源。
- 国家级综合实验区研究接口与实验设计文稿：在具有真实人口、权利、资源、利益冲突和长期运行条件的完整社会系统中，检验证据、授权、后果与修复能否真正接起来。

这些下行文稿承担验证和实现责任，不反向替本文证明哲学命题；若现实验证与本文的方法结构发生稳定冲突，应由现实结果迫使本文收缩、修订或重构。

左右 | 如果继续追问：相邻学科如何接手，哪些反方真正能够改变本文

- 体验、具身、自我与主体连续性：《AGI 时代的认知科学、神经科学与心灵哲学重新对账》，继续检验哪些机制与行为证据能够区分体验、自我模型和连续性。
- 模型、智能体与系统证据：《AGI 时代的人工智能与计算机科学重新对账——模型能力、系统保证与现实行动权边界》V1.1，继续检验架构、自述、版本、权限和部署行为能够证明什么。
- 保护、资格、责任与公共权力：《AGI 时代的伦理学、法学与政治哲学重新对账》，继续处理认识结论进入保护、程序、权利和责任以后必须新增的规范论证。
- 生命、个体与复制边界：《AGI 时代的生命科学重新对账》，继续检验生命、个体、发育、稳态、复制与终结能够为主体性提供什么事实约束，又不能替哲学和制度作什么结论。

横向续读必须同时保留相反性。如果认为功能等价足以消除体验差异、所有他心证据原则上不可判、社会或法律承认可以直接创造本体人格，或者人类观察窗口就是所有可能主体的最终尺度，应优先对照本文的最强反驳、败北条件与上述相邻专业稿，寻找真正能够区分这些立场的证据。竞争解释若更强，本文就应改变。

向前 | 如果希望从阅读进入共同研究

- 建立更好的体验、连续性与自身后果关系构念，提出可产生区分预测、能够跨实现检验而又不无限泛化的指标。

- 组织肯定者、怀疑者与中立团队共同参与的对抗性合作，预注册预测、完整发表阴性结果、复现失败与方法失效。
- 建立跨模型、跨版本、跨部署的证据谱系与失效规则，使旧证据不能无条件继承给新对象，也允许结论下降。
- 把再审机制带入 Human-COS、AGI-COS 及后续实验场景，验证“事实—数据—指标—证据—证成—制度决定”的交接在真实高后果环境中是否可执行。

最有价值的参与，不是赞同，而是足以改变理论的专业工作。

续读不是为了把读者留在体系里，而是让不同专业从自己的位置继续向前。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“证据止步、判断交接、可重开”成为文明习惯，一个社会会怎样运行？

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
为什么改变	《文明跃迁：AGI 时代的文明之路（全五卷）》	从单一证据问题回到文明为什么需要新的共同判断能力，以及错误怎样沿跨域因果链放大。	五卷书不是认识论命题的上位证明；事实与方法仍有权反向修改总体判断。
可修订文明怎样运行	《活的文明·第一卷：未来文明运行原理》	看见证据、授权、资源、治理与人的判断如何在制度中保留可争议、可纠错与可退出。	运行图景不能替代复现、因果识别或科学证成。
完整未来图景	《活的文明》三部曲统一出版合编版	把“如何知道”放进普通生活和新文明形成中，检验证据制度是否真的增加而不是削弱人的判断空间。	愿景不能用叙事完整性替代证据强度。

继续追问	建议进入	此时能看见什么	必须守住的边界
机器侧怎样 守住证据边界	《AGI-COS 全系列总览》	继续观察能力、审计、 权限、回滚与系统状态怎样 成为可核验对象。	工程保证不 自动生成事实真 理、规范正当性 或公共权力。

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

下表不是权威排名或证据总分。经典理论只提供有限的方法与概念资源，现实研究只支持其对象和条件内的结论；本文的研究假说须公开 C 类地位，哲学与规范承诺须公开 F 类地位。来源重点复核截至 2026 年 8 月 25 日。

附表 A 19 组来源与原创主张的可支持范围

序号	来源／对象	证据地位	可支持的有限结论	单独不足以支持
1	康德《纯粹理性批判》	F 哲学资源	认识条件塑造对象呈现的哲学问题。	人类条件是所有主体的本体条件。
2	波普尔、库恩、拉卡托斯	F 科学哲学资源	公开批判、范式条件与研究纲领比较。	任何单一方法足以裁决主体性。
3	杜恒—奎因传统	F 科学哲学资源	数据与理论之间存在辅助假设和竞争解释。	所有解释永远同样好。
4	图灵 1950	A/F 文本与方法资源	以可执行测试替代含混总问。	模仿行为证明意识或人格。
5	构念效度研究	A 成熟测量方法	指标解释需收敛、区分和因果论证。	某测试因命名即可测量意识。
6	严重检验、预注册与开放科学	A 研究方法	事前预测、失败条件和复核提高可批判性。	程序合规保证理论为真。

序号	来源／对象	证据地位	可支持的有限结论	单独不足以支持
7	Cogitate 对抗性合作	A／B 现实研究	意识理论可形成事前冲突预测；结果会同时挑战理论。	IIT 或 GNWT 已获最终确认或否定。
8	Butlin 等 AI 意识指标	B 前沿纲领	可由多种意识理论导出人工系统候选指标。	当前系统已意识或指标已获独立验证。
9	Crump 等动物感知框架	A／B 类比材料	不可直接观察体验可用多类证据分级研究。	动物指标可无修改迁移到 AI。
10	Altman 与 Bland	A 统计方法	未检出必须结合检验能力解释。	阴性结果永远无信息。
11	Turpin 与 Lanham 等	A／B 限域实验	部分任务中的链式理由可能不忠实，且依情境变化。	所有自述均虚假或无体验。
12	模型卡、数据说明、NIST	A 正式框架／方法	版本、用途、限制和持续测量需记录。	文档合规证明主体性或安全。
13	机器行为研究	B 研究纲领	人工系统行为应作为独立经验对象研究。	行为学自动解决本体问题。
14	AI 福利与负责研究文献	B／F 规范前沿	不确定性可能要求准备评估与低成本措施。	当前高级 AI 已是道德受动者。
15	三个主体研究方向	C 可修订研究假说	协调体验、连续性和后果关系的联合研究，并产生可区分问题。	三 项 是 学 界 共 识、判据、评分表或必要充分条件。
16	分对象的证据进阶	C + A 方法组合	按对象记录复现、区分、因果、收敛和稳健。	意识总分或权利自动阈值。
17	认识对称／制度不对称	F/C 规范—制度桥接	肯定与永久否定均可修订，行动按风险分级。	科学证据自动决定法律和政治。

序号	来源／对象	证据地位	可支持的有限结论	单独不足以支持
18	意义—有限性接口	F 哲学／规范承诺	公开本文的价值方向、有限性结构判断与既有意义非追溯原则。	已由实验验证、可以替代经验研究，或只有人类才能生成意义。
19	Bodien 等意识障碍多中心研究	A 临床经验研究	单一行为观察可能漏检与指令执行相关的认知活动；阴性结果需结合观察窗口能力解释。	可直接证明 AI 具有隐藏意识，或临床神经指标可无修改迁移到人工系统。

附录 B 证据与跨层审计清单

附表 B-1 A／B／C／F 论证身份及允许行动

证据地位	构成	允许使用	禁止越界
A 较稳定支持	较稳定的经验事实、正式记录/框架、可复核的文本与方法事实。	限定对象和范围后作事实前提、方法要求与反例约束。	写成终极本体结论、普遍阈值或制度授权。
B 前沿／竞争解释	意识理论、指标转译、自述忠实性、跨架构同构等仍存在实质争论。	比较模型、区分预测、限域试验与持续更新。	写成学界共识、确定意识判据或当前系统定性。
C 可修订研究假说／情景	三个主体研究方向、跨实现研究假说、未来复制/迁移等尚待检验的跨学科连接。	提出可失败预测、组织研究、设计低伤害验证与再审条件。	写成已证实科学、固定判据、自动权利或主权。
F 哲学／	公开署名的价值、本体或规范	以融贯性、反例、比较	伪装成 A 类

证据地位	构成	允许使用	禁止越界
规范承诺	判断，如意义为根、来源多元和既有意义不得被追溯归零。	论证和现实后果接受批判，并明确其经验桥接。	事实，或用哲学承诺压过专业事实、法源与公共授权。

B.1 外推前的十二项检查

- 主张对应的是能力、自我模型、体验、连续性、后果意识、资格还是权力？
- 系统、模型、实例、版本、状态和部署条件是否明确？
- 数据如何生成，提示、奖励、过滤、工具与外部记忆如何影响结果？
- 指标为何对目标构念敏感，是否具有收敛与区分效度？
- 实现映射保留了理论因果核心，还是只复制了抽象标签？
- 主体、训练、奖励、策略、角色扮演和社会投射解释是否正面对比？
- 阳性与阴性结果的敏感度、覆盖度和失败可见性是否报告？
- 预测是否事前登记，阴性、崩溃和无法复现结果是否保留？
- 证据域是否真正独立，还是共享同一训练语料、理论和测量误差？
- 对象更新、复制、迁移或回滚后，旧证据是否仍可继承？
- 经验判断是否被静默换成伦理保护、法律资格或公共权力？
- 读者能否在不接受文明跃迁其余部分时独立反驳本文？
- 哲学与规范承诺是否明确标为 F 类，并与 A 类事实、B 类解释和 C 类研究假说分开？

B.2 再审记录的最低字段

每次再审至少记录：主张提出者；系统、模型、实例和版本；目标证据对象；数据和指标定义；竞争解释；证据地位；阴性结果；利益冲突；临时措施；少数意见；撤回与申诉条件；下一复核日期。记录应允许结论下降，不得只允许证据累积和资格上升。

参考文献

[1] Kant I. Critique of Pure Reason. 1st ed. Riga: Johann Friedrich Hartknoch; 1781; 2nd ed. 1787. References to the A/B editions.

- [2] Popper KR. *The Logic of Scientific Discovery*. London: Hutchinson; 1959. German original published 1934.
- [3] Kuhn TS. *The Structure of Scientific Revolutions*. 2nd ed. Chicago: University of Chicago Press; 1970. First edition 1962.
- [4] Lakatos I. Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave A, eds. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press; 1970:91–196. <https://doi.org/10.1017/CBO9781139171434.009>
- [5] Duhem P. *The Aim and Structure of Physical Theory*. Wiener PP, trans. Princeton: Princeton University Press; 1954. French original 1906.
- [6] Quine WVO. Two dogmas of empiricism. *Philosophical Review*. 1951;60(1):20–43. <https://doi.org/10.2307/2181906>
- [7] Turing AM. Computing machinery and intelligence. *Mind*. 1950;59(236):433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- [8] Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin*. 1955;52(4):281–302. <https://doi.org/10.1037/h0040957>
- [9] Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. 1995;50(9):741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- [10] Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychological Review*. 2004;111(4):1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- [11] Mayo DG. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press; 1996. <https://press.uchicago.edu/ucp/books/book/chicago/E/bo3637756.html>
- [12] Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proceedings of the National Academy of Sciences*. 2018;115(11):2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- [13] Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nature Human Behaviour*. 2017;1:0021. <https://doi.org/10.1038/s41562-016-0021>
- [14] Chambers CD. Registered Reports: a new publishing initiative at Cortex. *Cortex*. 2013;49(3):609–610. <https://doi.org/10.1016/j.cortex.2012.12.016>
- [15] Searle JR. Minds, brains, and programs. *Behavioral and Brain Sciences*. 1980;3(3):417–457. <https://doi.org/10.1017/S0140525X00005756>
- [16] 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.
- [17] 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>
- [18] 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>
- [19] 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>
- [20] 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>
- [21] 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>
- [22] Birch J. Animal sentience and the precautionary principle. *Animal Sentience*. 2017;2(16):1. <https://www.wellbeingintlstudiesrepository.org/animalsent/vol2/iss16/1/>
- [23] Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485. <https://doi.org/10.1136/bmj.311.7003.485>
- [24] 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>
- [25] 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>
- [26] Mitchell M, Wu S, Zaldivar A, et al. Model Cards for Model Reporting. *Proceedings of FAT* '19*. 2019:220–229. <https://doi.org/10.1145/3287560.3287596>

- [27] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Communications of the ACM*. 2021;64(12):86–92. <https://doi.org/10.1145/3458723>
- [28] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. 2023. <https://doi.org/10.6028/NIST.AI.100-1>
- [29] Rahwan I, Cebrian M, Obradovich N, et al. Machine behaviour. *Nature*. 2019;568:477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- [30] Long R, Sebo J, Butlin P, et al. Taking AI Welfare Seriously. *arXiv*; 2024. <https://doi.org/10.48550/arXiv.2411.00986>
- [31] 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>
- [32] 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/>
- [33] 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>
- [34] Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press; 2009. <https://doi.org/10.1017/CBO9780511803161>
- [35] Woodward J. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press; 2003. <https://doi.org/10.1093/0195155270.001.0001>
- [36] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? *Proceedings of FAccT '21*. 2021:610–623. <https://doi.org/10.1145/3442188.3445922>
- [37] Perez E, Ringer S, Lukosłute 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>
- [38] Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of FAccT '20*. 2020:33–44. <https://doi.org/10.1145/3351095.3372873>
- [39] 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>

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。