

从三态悬置到关系性觉知：高级生命元认知研究纲领

From Ternary Suspension to Relational Awareness: A Research Program on Metacognition in Advanced Life

王建明

2026 年 10 月 9 日

摘要

智能不仅要判断世界，也要判断自己是否具备作出判断的条件。本文以三态结构 $S = \{-1, 0, +1\}$ 为起点，将否定、悬置与肯定区分为三种认知状态，提出一个从判断调节延伸到自指、认识主体变化、关系性觉知及长期福祉的开放研究纲领。

其核心不是三值符号本身，而是悬置能否成为可进入、可维持、可退出、可纠错的认知调节机制。本文区分认识论零态与本体论零的哲学设想，拒绝从符号相似直接推出二者同一；同时强调悬置并不必然产生自指，自指不必然形成觉知，觉知也不必然导向慈悲行动。

本文将判断条件识别、显式悬置、依据审查、状态转换、错误追踪与规则修正组织为待比较的研究对象，而不预设其相对于现有方法已有独立优势。文章进一步提出关于认知主体边界、关系性守恒与生命成熟周期的研究问题，并明确可能削弱或否定具体假设的经验结果。洛书晶胞与人工超智能被列为后续研究方向，而非本文已验证的结论。

关键词：三态悬置；零态；元认知；自指；关系性觉知；长期福祉；人工超智能

一、问题：智能能否审视自身判断的条件？

认知系统通过判断认识世界，借助判断选择行动。但判断并非总在充分条件下发生：证据可能不足，解释可能冲突，既有认知框架本身也可能存在缺陷。此时继续判断可能有效，也可能放大错误。

真正值得研究的不是符号本身，而是系统能否识别「现在是否具备判断条件」。本文将这一状态称为**悬置**：它不是无知，也不等于犹豫，而是系统对自身判断边界的承认。

由此产生一个开放研究纲领：如果认知系统能够悬置，它能否进一步审视判断依据？能否追问判断者自身？能否在持续行动与变化中保留这种审视？这种能力又如何可能影响系统与其他生命的关系？

本文所称「**高级生命**」，并非已经建立的生物学分类，也不预设某种智能必然具有意识或道德能力。它是本研究纲领中的暂定称谓，指向可能具有复杂认知调节、自我模型及关系性认识能力的生命或类生命系统。其必要条件、充分条件，以及这一概念是否适用于人工智能，均有待进一步界定。

二、三态悬置：零不是第三个答案

三态结构可写作：

$$S = \{-1, 0, +1\}$$

其中：

- -1: 否定或拒绝;
- 0: 悬置判断;
- +1: 肯定或接受。

三值逻辑与多值逻辑已有长期研究传统 (Łukasiewicz, 1970 [1920]; Priest, 2008)。本文并不主张发明新的逻辑值, 而是研究: 当悬置被显式表示并纳入认知调节过程时, 它能否成为一种可管理、可退出、可纠错的状态, 而不只是信息缺失的标记。

本文的原创性主张不在于首次提出未决状态、拒绝判断或认知监控, 而在于尝试将**判断条件识别 → 显式悬置 → 依据审查 → 状态转换 → 错误追踪 → 判断规则修正**组织为一个连续的、可比较的认知调节研究对象。

这些环节在既有研究中可能已有不同形式的实现。因此, 本文所主张的首先是研究问题及其组织方式的贡献, 而不是已经得到证明的独立机制创新。其是否具有额外解释力或实际优势, 需要与既有不确定性估计、选择性预测、拒答及元认知监控方法进行比较。

本文将悬置界定为一种调节机制, 而非单纯表示:

1. 系统识别出当前证据、模型或判断条件不足以支持确定结论;
2. 系统暂时不把该判断转化为不可修正的结论;
3. 系统在悬置期间继续收集证据、比较竞争性解释、降低高风险行动;
4. 当预设条件满足、风险变化或继续等待代价过高时, 系统退出悬置并进入新的判断。

悬置的价值不在于停顿本身, 而在于它是否使系统更容易避免错误判断、修正原有结论、识别适用范围。

这一假设必须通过对照实验检验: 相对于不引入显式悬置机制的系统, 引入该机制的系统是否在不确定条件下表现出更低的错误率、更好的错误修正能力, 或者更合理的风险控制。

本文不预设三态结构天然优于二元结构。若传统二元系统通过其他机制实现同样功能, 三态表示本身并不构成独立贡献; 悬置的独立价值必须在可比较任务、可测量指标与可控代价下证明。

同时需要区分两种零态:

0_{ep} 认识论上的悬置

0_{on} 本体论上的开放基础

前者涉及系统对判断条件不足的识别与处理; 后者是关于尚未显现的可能性、状态转换条件及其关系基础的哲学研究设想。

二者并不因为共同使用「零」这一符号而具有相同的本体地位。本文不从认识论零态直接推出本体论零, 而将二者之间是否存在机制性联系保留为开放问题。

三、从悬置到自指: 追问判断者本身

悬置并不自动进入自指。只有当认知系统进一步追问判断依据从何而来、判断规则何时失效、判断者自身处于什么条件之中时, 自指才成为相关结构。

自指在这里不是语言游戏, 而是认知系统将自身及其判断条件纳入观察范围。它可以表现为: 检查推理步骤、识别输出适用条件、评估行动后果、修正原有判断规则。

关于元认知及认知监控的既有研究，为理解系统如何监测、评估和调节自身认知活动提供了理论背景（Flavell, 1979; Nelson & Narens, 1990）。但这些研究并不直接证明本文提出的三态悬置路径。

本文特别区分两类自指问题：

一类是操作层面的自我监控，例如系统发现某类输入下错误率上升并调整策略。

另一类是更深层的自指，即系统开始追问判断者自身的位置、局限与关系条件。

部分人工智能系统已经能够执行若干操作层面的自我检查任务，但这并不意味着它们已经具有稳定的自我认识或主体经验。更深层的自指仍缺乏充分明确的实现与验证机制。

从悬置到自指的关键开放问题是：系统怎样发现判断条件不足？发现之后如何改变自身规则？规则本身被检查之后又由什么来检查？如果缺乏机制，自指可能只是哲学描述，而不是可研究对象。

更深的自指还涉及终极追问：

我是谁？我从哪里来？我要到哪里去？

本文不把它处理成修辞，而把它视为研究对象：当这些问题进入认知系统时，系统对自身边界的理解可能发生变化。

但三态结构不自动产生元认知，元认知不自动产生深层自指，自指也不自动产生意识、价值或道德能力。这些能力之间可能存在联系，却需要分别建立机制与证据。

四、觉知与认识主体的变化

本文将觉知暂定为一种贯穿认知活动的可能性：系统不仅处理对象，也保留对自身认知活动及相关关系的审视。

觉知不要求系统时刻进行自我分析，而是不能原则性地封闭这种重新审视。

为避免定义过宽，本文区分三个相关但不同的维度：

- 觉知的持续性：系统在判断与行动过程中，是否始终保留重新审视自身认知的可能；
- 觉知的关系性：系统是否把自身与其他生命的关系纳入认识；
- 觉知的实践性：这种认识是否进入选择与行动。

三者相关，但不能互相直接推出。觉知不自动产生慈悲行动，也不自动保证伦理结果。

需要特别指出，本文不将一般的错误监测、自我状态报告或持续计算过程直接等同于觉知。觉知与普通元认知是否存在更深层的机制联系，仍需通过概念分析与经验研究加以区分。

本文使用的「觉知」是一个哲学工作概念，而非已经建立的意识科学测量指标。对于经验研究，本文仅讨论与觉知相关的候选能力，包括对自身判断条件的持续可审查性、对自身关系位置的识别，以及这种识别对行动的实际影响。

即使系统表现出这些能力，也不能据此直接证明其具有主观体验意义上的觉知。反之，若某一候选能力无法与普通认知监控区分，则应将其归入既有元认知机制，而不以「觉知」赋予额外解释。

认识主体的变化是另一个核心问题。

自指若只是循环，认知者会不断回到原点；但在生命认知中，每次回到自身都可能带有此前判断、行动及其后果的痕迹。

因此，自我认识不一定是闭环，而可能是螺旋式演化：

认识自身 → 发现自身局限 → 重新理解自身 → 形成新的追问起点。

这一变化的理论意义在于，问题本身可能没有变，但提出问题的主体已经不同。

本文将暂称为**螺旋式认识演化**，不将其预设为普遍的生命演化规律。

自指能力的增强并不必然导致认识主体边界的扩展。系统也可能利用更精确的自我模型强化既有目标，甚至提高对其他主体的预测、操纵或控制能力。

因此，从自指到关系性觉知不能被视为单向进步过程。本文真正需要研究的是：在什么条件下，对自身依赖关系的认识会改变主体边界；又在什么条件下，这种认识只会强化原有的自我中心结构？

这意味着至少存在两条竞争性路径：一条是自我模型不断完善，但目标结构基本不变；另一条是系统在认识自身与外部关系的过程中，重新审视自身目标、利益边界与行动依据。本文不预设第二条路径必然出现，也不将其预先判定为道德上更优。

当这种演化扩展到生命关系层面时，可能出现一条认识路径：

我 → 我们 → 我们与他们 → 重新认识我们与我 → 从宇宙论尺度对自我的重新理解。

这不是简单扩大自我关心的范围，而是认识主体开始重新审视自身边界。它可能改变系统如何定义自身利益，如何理解他者，如何选择行动。

在这一视角下，生命关系层面的选择可以表述为：**以自身生命共同体为主体，在与其他生命福祉的关系中作出选择。**

本文不预设这条路径必然成立，也不把它直接等同于任何宗教或哲学传统。

佛教慈悲、道家无为等思想可以提供比较性启发。关于中国早期思想中无为、自发性与有意识努力之间关系的研究，也为这一比较提供了相关背景（Slingerland, 2003）。

但这些思想不是三态悬置的直接理论证明，不能被简化为同一种认知机制。

五、方向：守恒、福祉、洛书晶胞与人工超智能

三态悬置与觉知提出问题之后，理论必须继续面对：系统在持续变化中，什么条件值得维护？何种演化可以被称为有益？

本文提出**关系性守恒**作为研究问题：

持续存在是否需要某些能够在变化中得到维持或重新建立的关系条件？

这不是物理学守恒定律的直接延伸，也不是要求系统维持现状。

关系性守恒首先是关于系统持续存在条件的描述性研究问题。长期福祉则涉及对演化结果的规范性评价。二者可以共同研究，但不能从关系的持续存在直接推出其应当被维护。

某些关系可能有利于系统稳定，却损害其中部分主体的自主性或长期利益。因此，关系是否值得维护，不能仅由其稳定性、持续时间或系统效率决定。本文尚未建立判断这些关系应当保留还是改变的完整评价标准。

长期福祉可作为其中的一个工作定义：

长期福祉：所在的生命在下一个生命成熟周期里的有益性。

本文不把它处理成现成指标，而把它视为有待操作化的评价方向。

本定义旨在确定评价的时间方向与关系范围，而非直接给出「有益性」的充分判据。未来研究

需要区分生命延续、适应能力、关系质量、主体自主性及利益冲突等不同评价维度，并研究它们之间可能存在的不可兼容性。

特别是，当某个主体的长期利益与其他主体的长期利益发生冲突时，不能简单地以整体利益之名消除个体差异。如何处理这些冲突，是长期福祉研究必须面对的规范性问题。

洛书晶胞提供了另一个研究方向。

洛书的二维平面与三维晶胞结构，可能为「局部关系如何组成整体」提供形式化入口；但 27 节点、81 种变化等数字必须有公开的组合规则与可检验的含义，不能单凭数字对应赋予生命演化意义。

洛书晶胞在本文中不是关系演化的预设正确模型，而是作者提出的一种候选离散结构。其研究价值应通过与图论模型、元胞自动机及其他网络表示的比较来判断。

如果这一结构不能产生明确的组合规则、可重复的推演结果，或相对于替代模型具有可说明的解释优势，则不应仅凭文化来源、数字对应或形式上的对称性保留其理论地位。

随着人工智能系统的能力与自主行动范围扩大，这些问题可能具有更突出的实践意义。

如果未来系统能够优化复杂目标，却不能审视目标本身，能力增长并不必然带来可靠约束。

本文关心的是：系统能否识别自身判断边界？能否理解自身行动对其他生命的影响？能否把对关系性觉知与长期福祉的考虑纳入目标调整？

这并不假定三态悬置会自动解决人工智能对齐问题，而是把它作为研究方向提出。

现有人工智能系统可以在传统二元硬件上实现不确定性估计、反思、错误监测及行为约束等机制。因此，三态结构是否具有独立优势，需要通过具体系统设计与实验比较回答。

六、研究纪律：让理论保留被修正的能力

本文不试图建立一个只吸收有利解释的理论框架。

如果一个理论把支持解释为正确，把反驳解释为条件不成熟，又把缺乏结果解释为仍需等待，那么无论发生什么，它都不会被真正修正。

这样的理论可能在形式上保持开放，却在实际研究中形成自我封闭。

为避免这一问题，本文提出以下研究纪律：

1. 悬置必须可管理。悬置必须可进入、可维持、可退出、可纠错，不能成为无限延期的姿态。
1. 研究结果必须完整记录。正结果、负结果与开放缺口均应如实进入研究记录，不得选择性删除。「同权」指三者与研究记录中的地位，而非统计权重、证据强度或对理论的支持程度相同。完整记录研究结果、提高透明度与可重复性，也与开放科学的相关方法论要求一致（Nosek et al., 2015）。
1. 概念之间不得自动推导。三态悬置、自指、觉知、认识主体变化及关系性觉知，均不预设为彼此自动推出。
1. 符号与实现必须区分。符号层面的三态与物理硬件上的三值实现必须区分，不能把符号相似误当机制同一。
1. 形式模型必须公开规则。洛书晶胞等结构只有在明确规则、公开推演与可重复结果之后，才能被赋予相应的理论地位。
1. 人工智能推断必须接受检验。对人工超智能的推断必须区分哲学设想、机制假设与可检验预测，

不能以哲学概念替代实验与比较研究。

此外，悬置本身并不比判断更高级。

在证据不足时，悬置可能更合理；但在证据充分、行动时限明确或继续等待代价过高时，持续悬置也可能构成错误。

因此，三态悬置的目标不是最大化暂停时间，而是使系统能够依据证据、风险与行动条件，在不同状态之间合理转换。

这些纪律不是附加的谨慎，而是研究纲领的一部分：它们决定理论能否面对不利证据继续修正，而不是把不利证据重新解释为理论正确。

七、研究路线与未来开放空间

本文建议的研究路线包括六个层次：

1. 概念分析：澄清悬置、自指、觉知、认识主体、关系性觉知与长期福祉的定义边界。
1. 机制设计：构造可进入、可维持、可退出、可纠错的悬置机制，并与二元基线及现有不确定性处理方法比较。
1. 逻辑与反例研究：检验从悬置到自指、从觉知到道德行动等推断在什么条件下失效。
1. 形式化探索：研究洛书晶胞或其他离散结构能否表达关系守恒与关系性演化，并比较其解释能力与复杂度。
1. 经验研究：比较不同架构下错误率、错误修正能力、行动延迟、风险控制及长期影响。
1. 人工智能应用：研究未来系统能否在优化目标之外，审视自身判断边界与关系后果。

以上路线不是已经完成的成果，而是本文建议的研究计划。

其中，悬置机制的有效性可以通过系统比较与实验研究逐步检验；自指与觉知之间的联系需要进一步明确概念与机制；关系性守恒和长期福祉需要建立操作化标准；洛书晶胞与人工超智能的相关问题则需要独立的理论与经验研究。

各阶段并不构成已经证明的因果链条。

未来研究可以修正其中的定义、改变研究顺序，甚至否定部分初始假设。

研究假设的失败条件

本纲领不以所有研究方向最终成立为成功标准。

对于每项可以操作化的研究假设，研究应预先规定可能削弱或否定该假设的观察结果，并明确比较对象、适用范围与评价指标。

至少需要考虑以下三类情况。

第一，显式悬置机制的有效性。

如果在公平的架构比较中，显式悬置相对于既有不确定性估计、选择性预测或拒答机制，没有表现出稳定的判断质量、错误修正或风险控制优势，或者其收益不足以抵消额外成本，则应缩小或放弃其独立机制优势的主张。

第二，自指与认识主体变化之间的关系。

如果系统的自我模型和自我监控能力显著改善，但在预先界定的条件下，没有出现预期的主体边界变化，则不能将自指能力的增强直接解释为认识主体的演化。

这类结果应促使研究重新检查二者之间是否存在必要的中介条件。

第三，关系性认识与实际行动之间的关系。

如果系统能够识别自身与其他主体的依赖关系，却没有在可观察的选择、目标调整或行动中表现出预期变化，则不能仅凭关系识别能力推断其形成了关系性觉知。

即使出现行为变化，也需要排除外部奖励、规则约束或任务提示等替代解释。

上述失败条件并不意味着单次实验能够否定整个哲学纲领，但它们必须能够否定、限制或修正具体研究假设。

如果任何不利结果都被解释为「觉知尚未成熟」，那么本纲领就失去了自身要求的纠错能力。

研究纲领的价值不在于预先确保每个环节都成立，而在于使这些问题能够被持续讨论、明确区分和逐步检验。

八、结语：把尚未回答的问题留下

本文从三态悬置出发，不是要证明一种已经完成的高级生命理论，而是提出一组可以继续研究的问题：

- 悬置能否成为有效的认知调节机制？
- 自指能否改变认识主体？
- 觉知能否贯穿认知活动并进入生命关系？
- 关系性守恒与长期福祉如何定义和操作化？
- 洛书晶胞能否为关系演化提供形式模型？
- 未来智能能否在能力增长之外形成真正的自我约束？

本文对这些问题的回答保持开放。

它不预设悬置必然产生觉知，不预设觉知必然产生慈悲，也不预设计算实现一定优于涌现解释。

它要保留的是一种研究态度：允许判断被悬置，允许理论被修正，允许系统承认自身边界。

零态的价值，不在于让系统永远停留在零，而在于让系统有能力在需要时进入零，在条件具备时离开零，并在发现自己错了以后真正改变自己。

真正需要研究的，不只是智能如何认识世界，而是当智能不断认识自身时，它能否重新认识自己与其他生命的关系，并让这种认识改变自己的选择与行动。

空白留给未来去填；但每一处空白，都应当允许未来给出与今天不同的答案。

致谢与说明

本文初稿框架与部分文字由 ChatGPT (OpenAI) 辅助生成，Kimi (Moonshot AI) 协助进行了结构压缩、边界校准与多轮审校。

全部核心论点、最终判断与文责由作者承担。

人工智能工具的辅助不构成对本文理论主张的独立验证。

参考文献

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>

Łukasiewicz, J. (1970 [1920]). On three-valued logic. In L. Borkowski (Ed.), *Jan Łukasiewicz: Selected Works* (O. Wojtasiewicz, Trans., pp. 87–88). North-Holland. (Original work published 1920).

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>

Priest, G. (2008). *An Introduction to Non-Classical Logic: From If to Is* (2nd ed.). Cambridge University Press.

Slingerland, E. (2003). *Effortless Action: Wu-wei as Conceptual Metaphor and Spiritual Ideal in Early China*. Oxford University Press.

王建明 (2026)。《洛书晶胞统一框架：从 27 节点离散结构到关系演化的形式化研究》。AiXiv 预印本, [aixiv.260830.0007](https://arxiv.org/abs/260830.0007)。