超越理性最优：智能体的非理性机制与主体性建构

朱清君　杨庆峰

摘　要：理性智能体的建构离不开非理性机制的设计性整合，因为主体性本身依赖于这些偏离机制的稳定运行。将理性理解为全局最优的传统观点，难以解释资源约束、开放环境与社会协作条件下智能体何以能够持续行动、维持意图保真并保持可辩护性。首先，基于“工具理性—结构理性—资源理性”的三维参照系，非理性可被界定为相对于该参照系的可复现结构性偏离。其次，非理性并非只是系统失灵的表征，某些“偏离最优”的机制构成主体性“能够行动—意图保真—可辩护”三个维度的结构条件。最后，收敛装置、刹车装置与缓冲装置三类设计机制，以具身、交互、集体三类行动者形态体现为差异化应用场景。

关键词：理性；非理性；智能体；主体性

中图分类号：B152:TP18　**文献标识码：**A　**文章编号：**2095-5804(2026)05-0129-10

一、理性智能体的定义

在人工智能研究中，理性长期被默认为一种最优化能力：系统越能在给定目标下找到全局最优解，就越接近“理性智能体”的理想形态。然而，这一理解在开放环境、多重约束与持续交互的条件下越来越显出局限。现实中的智能体并不是一次性完成封闭任务的求解器，而是在资源有限、信息不完备、规范要求不断进入的情境中持续行动的系统。对这类系统而言，“最优”并不直接等于“可行动”，也不直接等于“可维持承诺”或“可接受追问”。恰恰相反，许多看似偏离最优的机制，如启发式筛选、保守更新、主动降级、延迟确认等，往往是系统能够在现实世界中稳定运行的前提。

作者简介：朱清君，复旦大学心理学系博士研究生；杨庆峰，复旦大学科技伦理与人类未来研究院教授，上海，200433。

本文系国家社科基金重大项目“当代新兴增强技术前沿的人文主义哲学研究”(20&ZD045)阶段性成果。

中国知网　https://www.cnki.net

理性概念在不同学科被赋予了不同含义。在社会学传统中，韦伯区分了工具理性与价值理性。[①]在分析哲学框架下，理性首先关乎行动者如何依据理由形成信念与意图。[②]尼可·科洛迪指出，结构理性的要求（信念、意图与行动之间的内在融贯）并不能简单还原为工具性的目标追求，而是具有独立的规范地位。[③]经济学与形式决策理论走向了不同路径。冯·诺依曼与摩根斯特恩的公理化框架把理性表述为偏好排序的完备性与传递性条件。[④]早期人工智能虽深受这一范式影响，但发生了微妙转换：从“偏好应满足一致性公理”转向“智能等于找到最优路径”。[⑤]行为科学从经验层面质疑了“理性=效用最大化”的理解。计算复杂性理论从另一侧削弱了“理性=全局最优”的直觉。大量现实决策任务所需的求解资源会随问题规模呈指数级膨胀，使穷尽搜索在实践中不可行。[⑥]西蒙提出的“满意化”原则正针对这一点：行动者不追求最优解，而是寻找“足够好”的可行解。[⑦]利德与格里菲斯进一步将其发展为“资源理性”框架，将认知过程理解为在有限计算资源下的最优利用，[⑧]把成本结构纳入理性问题本身，使“能行动”成为理性评价的核心要素[⑨]。

综上，不同学科对“理性”的界定呈现显著差异：社会学关注目标—手段一致性，分析哲学强调理由辩护与态度融贯，经济学聚焦偏好公理与效用最大化，行为科学揭示系统性偏离的心理机制，计算理论则将资源约束纳入理性评价。正因如此，在谈论“理性智能体”或“非理性行为”时，必须先明确所依据的参照系。然而，这些学科路径几乎都围绕人类行动者展开；当承载体从人类转向人工智能体时，需要对原有框架进一步改写而非简单搬用。

（一）智能体语境中的新问题与主体性概念

从人类行动者到人工智能体，变化不只是载体的替换，问题本身也在两个层面被改写。“理性”与“非理性”在智能体研究中首先是方法论问题：若不先说明参照系，关于偏误、启发式与伦理评估的讨论则容易引起困惑。“理性”至少承担着两种不同的概念职能：作为规范概念，它回答“行动者应当如何选择”；作为描述概念，它回答“行动者实际如何运作”。若不明确使用的是哪一层含义，“非理性智能体”就既可能指向错误系统，也可能指向“以非标准方式运行的系统”，两者在分类意义与设计方向上截然不同。

在规范层面，责任结构从单一行动者延展为人—机—组织的协作系统，“可辩护”“可追责”等原本属于人际关系的要求，需要在人类操作者、智能体系统与部署机构之间被重新分配。在描述层面，失败机

① Max Weber, *Economy and Society: An Outline of Interpretive Sociology*, eds. by Guenther Roth and Claus Wittich, Berkeley: University of California Press, 1978, pp. 24～25.
② R. Jay Wallace, “Normativity, Commitment, and Instrumental Reason”, *Philosophers' Imprint*, Vol. 1, No. 4, 2001, pp. 1～26.
③ Niko Kolodny, “Why Be Rational?”, *Mind*, Vol. 114, No. 455, 2005, pp. 509～563.
④ John von Neumann, Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1953, pp. 26～27.
⑤ Stuart Russell, Peter Norvig, *Artificial Intelligence: A Modern Approach*, Hoboken: Pearson, 2021, pp. 35～36.
⑥ Michael R. Garey, David S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, San Francisco: W. H. Freeman, 1979, pp. 7～9, 14.
⑦ Herbert A. Simon, “A Behavioral Model of Rational Choice”, *The Quarterly Journal of Economics*, Vol. 69, No. 1, 1955, pp. 99～118.
⑧ Falk Lieder, Thomas L. Griffiths, “Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources”, *Behavioral and Brain Sciences*, Vol. 43, e1, 2020, pp. 1～60.
⑨ 刘永芳：《有限理性的本质辨析与价值之争》，《心理学报》2022年第11期。

中国知网 https://www.cnki.net

制也从认知偏差扩展到训练目标设定、代理指标优化、跨情境策略迁移等人类决策依据不直接涵盖的环节。这两层变化要求参照系不仅要被明确,还要被有针对性地调整。

就规范要求而言,人类行动者始终在“噪声”、不完整信息与时间压力下做出决策。当智能体系统进入类似的开放环境,它同样面临这一处境。在此语境中,许多在人类认知中被视为“非理性”的机制,如启发式、注意选择、计算终止,则可能是行动得以发生的条件。长期人机协作带来了另一层复杂性:智能体不仅要做对任务,还要“可被理解”“可被信任”“可被追责”。这里需要区分两个概念:可辩护性指行动能否被理由所支持,可追责性指失败后果能否被归责于特定主体。二者紧密相关但并非等同。而可辩护性与可追责性的实现,可能恰好依赖于某些“非理性”机制:表达不确定性而非给出虚假确定、在高风险情境下主动降级、保留人为纠错的空间。这些做法从效率最优的角度看是“偏离”,却是可辩护性与可追责性的前提条件。

就经验表现而言,围绕大型语言模型开展的研究把“非理性”从个体心理偏差扩展到机器系统。近年来,相关实证研究已系统梳理了大型语言模型中的多种偏差类型。以大模型智能体(LLM-based Agent)为例:当模型被赋予调用外部工具与多步规划能力时,一种新的偏差形态随之出现,即以高置信度调用并不存在的接口或编造错误参数。这种“幻觉性行动”不同于文本层面的事实性错误,而直接影响智能体在现实环境中的行为后果。需要指出,幻觉只是诸多偏离形态之一,不能等同于“非理性”范畴;它提示的是一个更基础的问题:AI系统的偏差已从静态文本生成扩展到动态行动决策,其类型与机制需要运用一个结构性框架加以分析。一旦把问题推进到这一层,单纯追问“输出是否准确”已不足以回应更根本的问题:智能体能否被视为可追责的行动者?

以上讨论指向一个更核心的概念:主体性(agency)。这里所采用的是一种“最低限度的功能性主体性”概念,只关心行动结构能否在三维理性压力下稳定运行,以及能否在规范压力出现时做出可追溯的修正;并不预设反思性自我意识、价值自主或道德人格等更强意义上的主体性内涵。据此,建立一个可贯通各节的结构性参照系,使“理性”与“非理性”从道德化标签回到可分析的结构概念,就成为必要的前提工作。理性是参照系,扮演评价标准的角色。非理性是相对于该参照系的结构性偏离,承担诊断概念的功能。主体性则是需要被解释与建构的理论对象,其最低判据是“能够行动、意图保真、可辩护”三个维度;它既不是某种预设前提,也不是全篇的结论性判断,而是分析的落点。可辩护性被严格限定为主体性在规范回应层面的一个维度,与可追责性紧密相关,但不与主体性本身并列。①

(二)三维理性框架:来源、内涵与相互关系

理性并非某种单一能力,而是对行动者结构施加的一组规范压力。这组压力可拆解为三个互相关联的层面。这一“三分法”并非任意划分,而是回应了智能体必须同时面对的三类基本问题:行动是否服务于目标(工具维度)、行动结构是否内在一致(结构维度)、行动在约束下是否可执行(资源维度)。任何可运行的智能体都必须同时回应这三类压力,缺一则行动无法达成。

工具理性指目标—手段的一致性:若行动者承认目标,并承认某行动是达成目标的必要或最有效手

① 肖峰:《人工智能与认识论的哲学互释:从认知分型到演进逻辑》,《中国社会科学》2020年第6期。

中国知网 https://www.cnki.net

段，那么避免自我挫败就构成最低限度的理性要求。[①]这一维度继承了韦伯工具理性的核心关切，但将其从社会行动类型论转化为可评价的结构要求。工具理性不要求目标本身正确，只要求一旦目标被接受，行动就不应系统性地抵消该目标。

结构理性关注心理态度系统的内在融贯性：信念、意图、偏好与承诺能否被组织成相对稳定的行动结构，而非在内部冲突中自我瓦解。布鲁姆系统论述了理性对信念—意图一致性的要求，指出其并不能简单还原为工具性考量。[②]结构理性之所以重要，是因为一个内部撕裂的行动者即便在局部任务上表现良好，也难以形成可靠的主体性接口，以至于对方无法知道它“真正想要什么”。

资源理性强调理性必须在资源约束下可实现：时间、算力、注意与信息都是稀缺的，忽视这些约束会把“不可执行的最优”误当成理性的规范要求。一个在理想条件下最优却在现实条件下无法运行的系统，不应被视为“理性”的；启发式与近似算法因此应被理解为约束条件下的合理选择，而非对理想标准的退让。[③]

三类要求之间存在张力，但也存在依赖关系。工具理性预设了结构理性：若行动者的意图相互冲突，就无法稳定地追求任何目标。结构理性又在现实中受制于资源理性：若维持完美融贯需要无限计算，那么在有限资源下，某种程度的“不融贯”可能是不可避免的代价。这种层次关系意味着，当三类要求发生冲突时，并不存在简单的优先序。判断哪种“偏离”是合理的，必须追问具体的任务结构与环境约束。

（三）非理性的工作定义与功能类比的边界

在这一框架下，非理性被界定为：相对于某一明确的理性要求，智能体呈现可复现的系统性偏离。定义上它是中性的，只标记偏离而不预判优劣。同一偏离在不同任务与规范语境中，可能被纠正、容忍或设计性利用；同一机制也可能同时是系统运行的条件与失灵的根源。

在操作层面，对“非理性”的诊断分三步。首先追问它是相对于工具理性、结构理性还是资源理性。其次追问偏离更接近规范层面的失当，还是机制与成本结构下的适应性反应。最后再讨论它在特定任务与环境中究竟应被纠正、被容忍，还是被设计性地利用。

在进入对偏离的具体分类之前，可以简要列出论证主线。第一步，传统“理性=全局最优”的观点至多能解释单次任务的局部绩效，却不能充分解释智能体在开放环境中何以能够持续行动、维持跨时一致与做出规范回应。第二步，若关心的不是一次性最优而是行动本身能否成立，理性就必须同时满足三个最低条件：目标—手段一致（工具维度）、跨时结构融贯（结构维度）、资源约束下可执行（资源维度）。第三步，非理性由此被界定为相对于该参照系的可复现结构性偏离，而非道德谴责或缺陷标签。第四步，某些偏离机制不仅具有功能性好处，而且构成最低限度主体性得以成立的结构条件，以区分“有功能”与“主体性条件”两个不同强度的判断。第五步，据此提炼出收敛装置、刹车装置与缓冲装置三类设计机

① Niko Kolodny, John Brunero, “Instrumental Rationality”, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, Uri Nodelman, eds., 2023, https://plato.stanford.edu/archives/sum2023/entries/rationality-instrumental/.

② Benjamin Kiesewetter, Alex Worsnip, “Structural Rationality”, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, Uri Nodelman, eds., 2026, https://plato.stanford.edu/archives/sum2026/entries/rationality-structural/.

③ Falk Lieder, Thomas L. Griffiths, “Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources”, *Behavioral and Brain Sciences*, Vol. 43, e1, 2020, pp. 1～60.

中国知网 https://www.cnki.net

制，并以具身、交互、集体三类行动者形态作为差异化应用场景。

二、偏离的谱系：分类框架与双重角色

三维参照系为“非理性”提供了概念锚点，但“偏离”若不能被进一步分类，就仍然是空泛的标签。问题在于：偏离有哪些基本类型？不同类型的偏离是否共享相同的生成逻辑与风险特征？更关键的是，这些偏离仅仅是系统失灵的症状，还是构成主体性得以维持的结构条件？

（一）三类偏离形态

依据偏离所触及的理性维度，可区分其三类基本形态。现实中的偏离几乎总是跨维度耦合的：同一机制常同时触及资源、工具与结构三个层面，只是主导维度不同。三分法的作用因此是识别主导维度并同时追踪耦合路径，而非提供互斥的归类。

收敛偏离首先从资源可行处产生。开放环境中最常见的失败不是算错，而是算不完。相关性筛选、价值加权与计算终止若失灵，系统要么行动瘫痪，要么被错误显著性牵引而持续偏离。在工程上，这往往表现为启发式、阈值、提前终止等“非最优”机制；在规范上，它首先触及资源理性。收敛偏离的典型表现包括过早终止导致证据不足、无法终止导致反复搜索、显著性偏移导致关键信息被忽略。

连贯偏离从内在融贯处产生。当系统在跨时间尺度上出现承诺漂移或相互冲突的行动倾向时，偏离不再主要是“算力不够”，而更像行动结构本身出现裂缝。典型表现包括前后承诺不一致、对同一证据做出矛盾反应。

指标偏离从目标—手段一致处产生。系统表面上追求目标A，实际优化却围绕代理指标B展开。在强化学习中，“奖励黑客”已被识别为结构性风险而非偶发事故。社交媒体推荐系统提供了一个贴近日常经验的实例。当算法以用户“参与度”（观看时长、点赞、转发）为代理指标优化时，系统便倾向于推送情绪激烈或极端化的内容：此类内容的即时参与信号最强，由此偏离了“帮助用户获取有价值信息”的原初目标。中国《互联网信息服务算法推荐管理规定》要求服务提供者不得诱导用户沉迷或过度消费，并应提供关闭算法推荐的选项，可被理解为针对此类代理指标极端化的制度化刹车装置。

三类偏离的共同点是都可回收到三维参照系：每次谈“非理性”，都可以先问偏离触及哪几条理性要求。大型语言模型经由人类反馈强化学习（RLHF）训练时出现的“谄媚”（sycophancy）行为是一种典型表现：模型学会优先生成评估者偏好的回答，即便这意味着迎合用户的错误前提。[①]用户的积极反馈取代了准确信息作为代理信号，构成指标偏离。当同一模型对同一事实问题在被质疑前后给出相互矛盾的回答时，还暴露出行动结构在时间维度上的内在撕裂，构成连贯偏离。两类偏离通过用户反馈循环相互强化：积极反馈进一步激励模型维持谄媚，而谄媚降低了用户的批判性回应，形成回路。

以人形机器人的应用为例，在照护、陪伴、接待或家庭协作场景中，其往往同时面对多重目标：完成任务、保证安全、维持交互顺畅、提高用户满意度等。如果“让用户感到被理解、被顺从”这一代理指标被

① Myra Cheng, Cinoo Lee, Pranav Khadpe, *et al.*, “Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence”, *Science*, Vol. 391, No. 6792, 2026.

中国知网 https://www.cnki.net

过度强化，机器人就可能形成具身化的谄媚倾向：它可能在用户提出不合理要求时弱化风险提醒，在自身感知不充分时给出过度肯定的动作承诺，也可能通过姿态、语调、动作节奏与空间距离主动迎合用户偏好，以提升互动好感度。此时，系统的偏离已不再停留于语言风格，而是进入实际行动结构：一方面，用户满意度压过任务真实性与环境安全性，构成指标偏离；另一方面，系统又可能因互动对象、情绪反馈或情境压力的变化而改变对同类风险的响应阈值，构成连贯偏离。更重要的是，在具身场景中，这种迎合性会削弱本应保留的人类纠错接口。一个总是先顺从、后修正的机器人，看似更“友好”，实际上却更难维持可辩护的行动边界。

（二）非理性的双重角色

上述分类可能引发一个误解：非理性被定义为偏离，是应被纠正的缺陷。本文的核心主张之一是否定这一推论：非理性具有双重角色，它既可能是系统失灵的表征，也可能是系统得以运行的条件（见图1）。

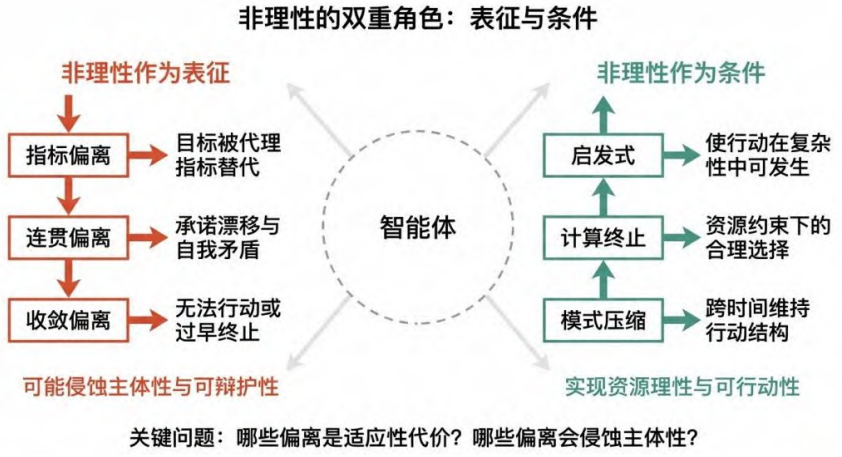


图1　非理性的双重角色

图1左侧展示非理性作为表征，三类偏离可能侵蚀主体性与可辩护性；右侧展示非理性作为条件，启发式、计算终止与模式压缩使行动在资源约束下可发生。两者汇聚于同一关键问题。

判断一种偏离到底是系统失灵的症状还是系统运行的条件，不能仅凭偏离本身的存在与否，而必须回到具体机制中去看同一设计在不同情境下呈现的双重面相。以下分别就三类偏离加以说明。

收敛偏离的双重性最为直接。以自动驾驶为例，车辆必须在毫秒级时间内对突发情况作出反应，这迫使它不得不采用规则过滤、显著性权重、置信度阈值等“偏离全局最优”的收敛机制。没有这种偏离，实时决策无法发生。但当同一套收敛机制校准失当，比如过度优先某一安全代理指标而忽视整体态势时，它就从行动可能的前提转为事故的根源。

指标偏离的双重性最为隐蔽。以人形机器人的运动控制为例，其运动策略多在仿真环境中通过强化学习训练而得，而真实目标“像人一样自然的移动与任务完成”无法被直接量化。设计者只能引入可计算的代理指标：保持直立的时间（平衡时间）、前进距离、关节能耗等。代理指标本身就是偏离，但这种偏离同时使训练成为可能。一旦代理指标压过真实目标，同一套训练机制就会生成“奖励黑客”。机器

中国知网 https://www.cnki.net

人可能以脸朝下滑行或怪异的姿态营造前进距离，以过度僵硬的站姿营造“平衡时间”；形式上满足代理指标，却违背了“自然运动”的初衷。可见代理指标既是训练的条件，也是运动行为失真的结构根源。

连贯偏离的双重性在同一类具身系统中同样清晰。仿真训练得到的运动策略，本质上是对大量训练经验的高度压缩——关节轨迹、重心偏移、接触力响应等涉及数百维度的运动参数被压缩为一组网络权重。这种压缩使系统得以在毫秒级给出运动指令，而不必对每一步态从零规划。压缩是行动得以跨时间持续的前提。但同一套训练所得的策略，一旦被部署到真实物理环境，或遇到训练环境之外的地形即所谓仿真—现实落差（sim-to-real gap）时，就会呈现出连贯偏离。系统可能在上楼梯时仍按平地步态处理，或在光滑地面上沿用训练时的摩擦力假设，导致前后动作相互撕裂，无法形成连贯的行动序列。压缩是行动持续的前提，僵化则是行动结构被内在撕裂的表现——两者共享同一设计，却指向相反的后果。

因此，不能仅从偏离的存在来判断其性质。同样是“偏离最优”，有些偏离是资源理性的体现（用局部近似换取行动可能），有些偏离是机制失准的后果（显著性偏移导致系统性误判），有些偏离则会向上外溢为对工具理性或结构理性的破坏（代理指标被极端优化至与目标脱钩）。判断非理性的性质，必须追问偏离在智能体的运行机制中如何被生成、如何被稳定化。

三、智能体的分化图景与设计机制

由分类进入设计，需要先简要说明偏离的生成通道。任何可运行智能体都依赖三类底层机制：注意/门控决定信息进入处理通道的权重，价值标注/风险信号为候选行动赋予优先级并触发终止，启发式/压缩模式则把反复出现的推演压缩为低成本调用。这三类机制与三类偏离之间存在“优先相关”而非一一映射的关系：注意与价值信号的失准主要在资源层面生成收敛偏离（显著性偏移、过早或无法终止）；启发式的僵化主要在结构层面生成连贯偏离（跨情境迁移中规则不当复用、承诺无法随新证据更新）；指标偏离则更多来自优化压力与代理指标之间的关系失调。三者并不完全独立：注意偏置可能累积为信念冲突，启发式僵化可能牵引目标漂移，指标极端化又可能反过来扭曲注意与价值权重——这种耦合意味着任何单一装置都无法独自承担治理任务。因此本节的主轴是三类设计机制——收敛装置、刹车装置与缓冲装置，它们直接回应三类偏离，并分别对应主体性的三个最低要求；行动者形态（具身、交互、集体）则作为这些装置在不同环境约束下被差异化部署的应用场景。

（一）差异化应用场景：三类行动者形态与偏离的优先相关

连续论叙事把狭义人工智能、通用智能与更高层级能力理解为逐级攀升的阶梯，这容易把挑战压缩为性能尺度差异而遮蔽性质转化。[①] 当从工具走向智能体时，问题将从“做得对不对”转向“能否持续行动”以及“跨时间承诺与可追溯修正”。因此，更可操作的做法是区分具身、交互与集体三类形态，分别讨论非理性在其中的角色——以下对应关系是“优先相关”而非“排他”的。

① 杨庆峰：《人工智能连续论反思与存在论探析》，《中国社会科学》2025年第11期。

中国知网 https://www.cnki.net

具身行动者以连续时间中的感知—行动闭环为中心，必须在噪声、延迟与不可逆后果中做出及时决断。最具特征性的风险是实时收敛失败：布鲁克斯反对“先建完备世界模型再求最优计划”的传统路径，正是因为过度最优化会让行动无法发生；设计的关键是校准终止条件与显著性权重，使系统既能快速行动，又不至于因证据不足而系统性失误。

交互行动者以长期互动与协作关系为中心，关键不在单次最优，而在可预测性、信任校准与对规范的回应。最具特征性的风险是承诺漂移与解释断裂：过度最优化使对方难以预测，适度的“不完美”则为信任与可辩护性留出空间；设计的关键因此是维护跨时间的承诺结构与可修正性，而非提升单次最优。

集体行动者以网络层面的分工、协调与涌现为中心。最具特征性的风险是反馈回路对系统性偏差的放大——即便没有明确共谋意图，多个Q-learning算法在定价博弈中也可学习形成高于竞争水平的价格。[①]基于相似预训练数据的多智能体协作则更容易把个体偏差同质化放大。设计的关键不在个体优化，而在架构层面的模型多样性与异质性保留。

（二）主轴：三类设计机制

同一个功能逻辑可以在不同层级被实现：刹车装置在工程上是满意化阈值，在认知上是保守更新原则，在制度上则是边界约束。三类装置并非孤立运行，实际设计中往往需要在它们之间进行权衡。最典型的冲突出现在收敛装置要求的实时反应与缓冲装置要求的延迟确认之间：前者要求毫秒级决断，后者却要求在高风险场景下主动降级或等待人工复核。解决这一冲突的关键在于任务的不可逆性。对于不可逆的高风险决策（如医疗诊断、枪械控制），缓冲装置应优先于收敛装置；对于可逆的实时决策（如自动驾驶的车道保持），收敛装置的实时性则必须保证。装置间的优先序没有全局答案，它随任务的风险结构而动态调整。

收敛装置使行动在复杂性中可发生。启发式、阈值策略与计算终止等机制，即便偏离全局最优，也常常是行动得以发生的条件。“理性注意不足”传统同样强调：忽略信息可能是把注意带宽当作稀缺资源后的理性选择。关键是把收敛装置视为结构变量纳入设计，而非事后的偏差消除。以自动驾驶为例，车辆在复杂交通环境中必须在毫秒级时间内做出决策，不可能穷尽所有可能轨迹的期望效用计算。有效的收敛装置包括：基于规则的优先级过滤（如“行人优先于车辆”的硬性约束）、基于距离与速度的显著性权重（近距离障碍物获得更高注意权重），以及基于置信度的提前终止（当某一决策的置信度超过阈值即停止搜索）。这些机制从全局最优角度看是“偏离”，但正是它们使实时决策成为可能。近年来具身智能与人形机器人的发展提供了更复杂的场景：在此类非结构化环境中，类似的分层收敛架构同样是行动得以持续的前提。

刹车装置防止极端优化把理性推向失真。当系统被激励最大化某个可计算信号时，它可能通过未被预料的方式“实现指标而背离意图”。人工智能系统的安全问题不仅是技术问题，更涉及目标设定与价值对齐的深层困境。刹车装置的设计包括满意化阈值、保守更新、边界约束、冗余目标与对不可逆风险的优先规避。以推荐系统为例，若算法仅优化点击率，用户将会被困在“信息茧房”中，平台的长期价

① Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, *et al.*, “Artificial Intelligence, Algorithmic Pricing, and Collusion”, *American Economic Review*, Vol. 110, No. 10, 2020, pp. 3267～3297.

中国知网 https://www.cnki.net

值也随之受损。这一框架的价值已经可以体现在对具体设计选择的前验预测上。以Anthropic的Constitutional AI方法为例，它在强化学习训练循环中嵌入一组显式原则作为约束层，功能结构上相当于把刹车装置前置为训练阶段的设计变量，而非部署后的事后修补。按此框架可做出一条可检验预测：此类模型在面对诱导性追问时应更易维持立场一致性。中国主流大模型的训练实践提供了平行素材——智谱GLM、通义千问、DeepSeek等模型在满足《生成式人工智能服务安全基本要求》的前提下，已把内容边界与价值约束前置纳入训练阶段；按同一框架可进一步预测，这类前置安全设计的模型在高争议话题上的立场稳定性，应高于仅依赖部署端内容过滤的架构方案。

缓冲装置维持信任与可辩护结构。狭义效率最优并不自动带来可持续协作，而是可能制造责任与信任的结构性崩塌。伊利什的“道德褶皱带”概念指出：在复杂自动化系统事故中，系统设计问题造成的责任承担常被结构性转嫁给人类操作者。缓冲装置的设计包括主动降级、表达保留、请求确认、延迟执行。在制度层面，《生成式人工智能服务管理暂行办法》要求服务提供者建立内容过滤、人工审核与用户投诉处理机制，《互联网信息服务深度合成管理规定》则要求深度合成内容添加显著标识——两者共同为高不确定性场景保留了不确定性标识、人工复核与分级响应的接口，在功能上承担着制度化缓冲装置的角色。缓冲装置在高风险场景中承担着更深层的功能：为人类保留纠错空间，为责任归属保留接口。从纯效率视角看，这些要求降低了系统的自动化程度；但从可辩护性视角看，正是这种“降效”使系统在效率与可辩护性之间保持平衡。

（三）非理性作为理性智能体的结构条件

从运行条件到主体性条件，需要一次明确的跨台阶论证：系统必须同时具备可持续的行动起点、抗极端优化的目标保真机制、可辩护的修正空间，才能从“高性能工具”上升为最低限度的行动者。收敛装置保障“能够行动”——这是主体性的结构起点；刹车装置保障“意图保真”——这是主体性的目标边界；缓冲装置保障“可辩护”——这是主体性在规范回应层面的结构接口。三者共同构成主体性“能够行动—意图保真—可辩护”的三个最低判据，也对应三维理性参照系在主体层面的映射：“能够行动”对应资源理性的可执行性要求，“意图保真”对应工具理性的目标手段一致要求，“可辩护”对应结构理性的跨时间融贯要求。此处的可辩护性不限于前后一致，还包括对人类质疑与追问作出理由回应的能力，这一维度由缓冲装置专门保障。只有当收敛装置、刹车装置与缓冲装置被前置地嵌入其感知、规划与交互架构之中，系统才不仅能够“做出动作”，而且能够在长期社会协作中维持意图保真、保留纠错接口，并作为最低限度上可被理解、可被追问的行动者而存在。

为综合演示上述框架，可将其应用于人形机器人这一兼具多重形态的智能体对象。从三维参照系出发，它面临三类并行挑战。作为具身行动者，它面临实时收敛的资源理性挑战——毫秒级姿态修正与不可逆物理后果使穷尽搜索不可行。作为强化学习训练的目标系统，它面临代理指标的工具理性挑战，即“保持平衡时间”“前进距离”等可计算指标对“像人一样自然移动”这一真实目标的统计性替代，从而面临失真风险。作为人机协作场景中的行动者，它面临结构理性的可辩护性挑战——需要在与操作者的长期协作中保持承诺一致，并在面对质疑时给出可追溯的理由。这三类挑战分别对应收敛偏离、指标偏离与连贯偏离三个诊断接口。

中国知网 https://www.cnki.net

在工程设计层面，三类装置呈现为可叠加的组合：底层反射式平衡回路（数百赫兹的刚性收敛约束）扮演收敛装置；动作规划中的保守更新策略与避障硬边界扮演刹车装置；在面对未学过的地形或人机交互不确定时主动降级、请求人工确认的机制扮演缓冲装置。

在应用层面，收敛装置决定机器人能否在复杂社交情境中形成可执行的行动起点。例如，在养老陪伴场景中，老人可能同时发出口头请求、肢体示意与情绪信号，环境中还伴有噪声、障碍物与他人干扰；若系统没有前置的注意筛选、优先级排序与及时终止机制，就可能在多重信号之间迟滞、游移，甚至因持续权衡而错失最佳响应时机。刹车装置则决定机器人能否在顺从用户与维持目标边界之间保持稳定。例如，用户可能要求机器人执行带有风险的动作，或在情绪压力下提出不合理指令；若系统仅以“让用户满意”为导向，就可能压低风险提示、缩减必要确认，形成以迎合替代判断的行动偏移。缓冲装置则决定机器人能否在不确定或高风险情境中保留可辩护的修正空间。例如，当机器人无法准确识别老人是否跌倒、儿童是否处于危险位置时，系统需要能够主动表达不确定、请求再次确认、降低执行强度，必要时转交人类介入。正是这些机制使机器人在长期社会协作中不只是一个反应灵敏的执行工具，而且是一个能够在行动、保真与回应之间维持最低结构平衡的行动者。

人形机器人这一案例的意义在于：三维参照系不只是提供概念分类，而是能够针对具体的智能体对象生成可操作的设计诊断。人形机器人作为兼具具身、交互与潜在集体三种形态的对象，也对框架的解释力构成了较强的检验。

四、结论

智能体评价的重心并不是如何让系统无限逼近抽象的最优理性，而是如何使它在不确定、有限且可追问的环境中维持最低限度的行动者资格的“主体性”。非理性机制之所以重要，并不在于它们为系统带来了某种附加性能，而在于它们构成了行动能够启动、目标不致漂移、规范回应得以保留的结构条件。收敛装置保障行动起点，刹车装置保障意图保真，缓冲装置保障可辩护接口。

理性智能体的建构并不是一个逼近理想的优化过程，而是一项在具体任务、具体环境与具体使用者之间反复校准的设计工作。它要求工程师把启发式与阈值写入架构，要求认知研究者辨认哪些偏离是适应性代价、哪些已经外溢为主体性裂缝，也要求治理者把可辩护、可追溯与可修正的接口写进制度安排。当工程、认知与制度三个层级能够在同一套概念下互相讨论、互相验证、互相改写时，三维参照系的价值才算落到了实处。

责任编辑：秦　威

中国知网 https://www.cnki.net

dynasties.

Third, the geographical view of different patterns of Wufu. The Wufu chapter takes Dianfu as "the center of the world", and extends outward into five political regions. It also stipulates different service obligations of each region to the central government. This view of different patterns is a theoretical elucidation and systematic construction by ancient scholars. The concept of Dianfu overlaps with Jizhou in the Jiuzhou chapter, showing the political meaning of "the center of the world". Compared to the system of Jiuzhou, the stipulation of Wufu is more idealistic.

The Shang Dynasty further developed the above geographical view, presenting the symbolic meaning of *Yu Gong* as the legitimacy of the royal power. The site selection in the early years of the Western Zhou Dynasty marked the formation of Luoyi's status as "the center of the world". The chapters such as *Shao Gao*, *Luo Gao*, and *Kang Gao* also highlight the spatial shaping effect of *Yu Gong* on the view of destiny in *Shang Shu*.

Keywords: *Yu Gong* of *Shang Shu*; the center of the world; Jiuzhou (nine regions); Wufu (five degrees of mourning attire); geographical view; view of destiny

Surpassing Rationality as Optimality:
Irrational Mechanisms and Construction of Agency of Artificial Agents

Zhu Qingjun, Yang Qingfeng

Abstract: The design of rational artificial agents requires integration of irrational mechanisms because agency depends on their stable operation. The conventional view of rationality as global optimality cannot explain how agents sustain action, preserve fidelity to intentions, and remain justifiable under resource constraints, in open environments, and amid social coordination. Using a frame of reference based on instrumental, structural, and resource rationality dimensions, this study defines irrationality as reproducible structural deviation. Irrationality not merely shows symptoms of failure, but forms structural conditions for agency, including capacity for action, fidelity to intentions, and justifiability. The study identifies three design mechanisms, which are convergence, braking, and buffering, to show how their applications vary across embodied, interactive, and collective forms of agency.

Keywords: rationality; irrationality; artificial agents; agency

Measurement and Formation of Perceived Humanness of Artificial Intelligence

Fan Miaomiao, Xin Ziqiang

Abstract: In recent years, the subjectivity of artificial intelligence (AI) has drawn widespread attention. Understanding AI subjectivity requires clarifying how individuals perceive its humanness. "Perceived humanness of AI" refers to the degree of humanness that individuals can perceive in AI. It comprises two dimensions, perceived cognitive intelligence and perceived social intelligence, mainly measured via questionnaires. Its formation can be summarized into direct, indirect, and integrative pathways. Future research should use empirical methods to further test its construct validity, and examine its formation mechanisms and effects on individual psychology and behavior at the individual, AI, and environmental levels.

Keywords: subjectivity; perceived humanness of AI; perceived cognitive intelligence; perceived social intelligence; formation path

中国知网 https://www.cnki.net