去人类中心化视域下：AI 智能体奇点分析与 AI 智能体治理理论重构研究

熊宇博[1]

（1.四川大学，四川 成都 610207）

摘 要： 人工智能技术的飞速发展不仅标志着技术奇点的临近，更引发了一场深刻的认知范式革命，动摇了自笛卡尔以来的人类中心主义根基。本文在“去人类中心化”哲学视域下，系统审视了 AI 智能体奇点所带来的本体论、认知论与伦理挑战，系统剖析了现存的“技术中心主义”替代与“能动性模糊”的双重困境。为回应此困境，研究构建了“人机协同主体”（HACAM）的理论模型，融合复杂系统理论、梯度化行动者网络理论与承认理论，提出一个以伦理内置为前提、认知互构与权力共生为内核、伦理纠正为保障的动态框架。通过模型内部的“认知-权力共生熵（P）”量化指标，清晰界定三层权力结构的熵值区间，实现了对权力分布状态的精确度量与调控。本研究旨在为超越“人类主导”或“技术替代”的二元悖论提供一条辩证的理论路径，并为走向“人机协同共治”的未来治理提供具备可操作性的理论工具与伦理指南。

关键词： 去人类中心化；AI 智能体；人机协同主体；认知-权力共生熵

DOI：doi.org/10.70693/rwsk.v1i9.1420

1 引言

在当今时代，人工智能技术的迅猛发展及相关类人智能体的持续迭代，再次引发了关于“去人类中心化”与“后人类主义”等议题的深入思考。当 AlphaGo 以概率性棋路突破人类围棋直觉的认知边界，ChatGPT 通过生成式算法模糊 “人类创造” 与 “机器模拟” 的界限，AI 技术的指数级发展已从 “工具辅助” 迈向 “认知挑战” —— 这不仅是技术层面的奇点临近，更是认识论层面的范式革命[1]。这种技术奇点带来的认知震荡，恰与加拿大人类学家爱德华多·科恩（Eduardo Kohn）《森林如何思考：超越人类的人类学》[2]引发的理论地震形成了深刻共振。前者在技术维度解构人类的智力与符号特权，后者则在认识论层面颠覆了人类学中对于人类作为唯一“认知主体”的传统看法。在技术与认知双重解构的时代背景下，我们需要重新审视自笛卡尔"我思故我在"以来构建的人类中心主义认知框架[3]，以及在此框架下形成的现代性知识体系。

这种本体论层面的反思并非无源之水。第二次世界大战造成的文明创伤，直接动摇了西方形而上学传统的根基。当奥斯维辛的烟囱焚毁了启蒙理性塑造的人类神话，后现代思想家们开始探寻超越人类主体的认知路径。让·鲍德里亚（Jean Baudrillard）的客体化策略试图消解主客二元对立[4]。吉尔·德勒兹（Gilles Deleuze）的生成理论在人与动物的边界开辟出动态的“成为”空间[5]。雅克·德里达（Jacques Derrida）则通过解构语音中心主义揭示出人类动物性的伦理基质[6]。这些思想涌动在人类学领域催生了深刻的本体论转向：从马林诺夫斯基（Bronisław Malinowski）的功能主义到英戈尔德（Tim Ingold）的栖居视角，从表征分析到本体论实证，学科范式正在经历从“人类学”（Anthropology）到“多物种民族志”（Multispecies Ethnography）的蜕变[7]。

然而，当前研究陷入双重理论困境：一是技术至上的“极端去人类中心” 倾向。以唐娜·哈拉维（Donna Haraway）的“赛博格宣言”为代表，相关研究将 AI 等技术视为 “人类进化的终极延伸”，过度聚焦 “人类与技术” 的二元关系 —— 既忽视自然生态的本体论地位，又弱化了技术应用的伦理风险。其本质是用 “技术中心主义”替代 “人类中心主义”，并未真正跳出 “谁为核心”的二元对立框架。二是机械对称的“能动性模糊”倾向。此种倾向广泛存在于机械套用拉图尔“行动者网络理论”的“对称性原则”，将人类、AI、自然物等不同行动者等同视之，模糊了各类实体在 “能动性”上的本质差异。由此陷入的逻辑怪圈，在批判人类中心主义的同时，往往又落入新的中心论或简化的对称观，未能真正建立具有伦理清晰度与本体论区分力的认知框架。

作者简介： 熊宇博（2005—），男，本科，研究方向：AI 智能体伦理、人机交互与数字人文

这一困境不仅反映出二元对立思维模式的持续影响，也突显了在面对技术 — 自然 — 人类多重交织的现状时，传统理论工具所表现的局限。因此，亟需一种更具辩证性和层次性的理论路径，既能承认非人类行动者的角色，又可包纳其内在的异质性与道德权重差异，从而为真正超越人类中心的认知伦理开辟可能。

本研究旨在系统探讨人工智能作为“非人类认知体”对人类学本体论框架所带来的挑战，并在信息哲学的层面论证一种“去中心化认知范式”对于重构技术社会伦理与社会治理的可能路径。论文的最终旨归，是为正在来临的多物种文明寻找理论锚点——当森林开始思考、当算法获得认知，人类亟须在自我祛魅的过程中重新建构物种伦理，进而推动超越传统人类社会视域的去人类中心化理论重构。

2 去人类中心化思潮与 AI 奇点辩证评述

2.1 去人类中心化思潮演变

对于本体论和人类中心理论的思潮与探讨大致起源于第二次世界大战之后。第二次世界大战后，西方哲学与思想界发生了深刻的转折。长达两千年的形而上学传统及其所支撑的人类中心主义观念，在经历战争的剧烈冲击后，已显露出根本性的危机。“后现代”知识分子群体纷纷为当代资本主义社会开具济世良方，他们共同的选择就是走出人类主体视阈，试图解构人本主义的固有框架，试图在主体性之外寻找新的哲学出路。

让·鲍德里亚主张一种彻底的客体化策略，建议人将自身从主体位置中抽离，主动走向客体状态，从而实现一种对现代性逻辑的反叛。吉尔·德勒兹则通过“生成动物”的概念，强调人与动物之间存在一片彼此交融、相互转化的潜在领域；他将这种思想实验喻为“西方哲学中的庄周梦蝶”。在《千高原》[8]一书中，吉尔·德勒兹与加塔利生命形式之间流动性的存在状态意在打破人文主义传统中僵化的身份界定。雅克·德里达则从对语音中心主义的批判出发，强调原初伦理中不可忽略的“动物性”，试图在语言和存在论层面恢复他异性的地位，从而重构伦理主体的边界。这一系列思想尝试，虽策略和侧重点各异，却共同指向对传统人类中心主义的超越，意图在伦理、存在与符号秩序中重新安置“人”的位置。它们不仅响应了战后西方哲学对形而上学危机的自觉，也拓宽了当代批判理论的问题域与解释空间。

进入现代社会，AI 智能体的不断发展，其正以前所未有的方式持续挑战着建立在人类中心主义基础上的本体论、认知框架与伦理体系。AI 不仅作为一种技术存在，更逐渐成为“去人类中心化”思潮在当代最具现实意义的具身与延伸[9]。它从哲学思考层面推动了对“何为行动主体”“何以为存在”等根本问题的重新审视，也在实际层面重构着社会权力、生产结构与知识生成的分布方式。

2.2 AI 智能体奇点现状分析

AI 智能体的迅猛发展主要从两大维度动摇了人类认知维度与权力控制中心，并在不同程度上呼应了后现代思想中对于“人”的独特性思考与探索。

一方面，AI 系统通过其非人类的认知与行为模式，动摇了传统意义上以“人类理性”为核心的主体观念。李舒戈[10]通过构建“AI 赋能政策智能”体系框架，推进了 AI 智能体决策能力的实现路径与评估框架打造。刘成[11]等通过组织、技术、应用和法律等层面的多维度分析，推进了 AI 与公共决策的现实性系统研究。智能体不再仅仅作为工具存在，而是在决策、创造甚至伦理判断中表现出一定的自主性，从而模糊了人类与机器、主体与客体的边界。这种模糊性呼应了后现代思想中如德勒兹所强调的“生成逻辑”，AI 可被视为一种新的“他者”，不断参与并重塑着我们对于“共同生成世界”的理解。

另一方面，AI 所带来的数据驱动和算法秩序，也构成了一种新的去中心化权力机制。其在一定程度上，消解了传统人类中心的控制叙事。同时，也以更隐蔽的方式建立新的中心性权威。卢卫红[12]等在技术维度上指出智能机器替代人类从事技术性、平台消费性工作，造成主客体颠倒。王箐[13]从网络平台的算法权力的滥用视角入手，系统分析在网络平台将直接影响用户个人权利。AI 智能体的算法运行机制即为此种算法机制中的一般体现。这种双重性需要我们超越单纯的技术乐观或悲观主义，转而深入思考如何在人机共生的语境中重构个人地位认知、与正义原则。

2.3 去人类中心化思潮与 AI 治理的范式分析

去人类中心化思潮与 AI 技术的耦合发展，正推动社会治理领域发生深刻的范式转型。这种转型既体现在治理主体的多元化重构上，也反映在伦理准则与权力机制的重新配置中，形成了理论与实践相互激荡的复杂场域。

从治理主体的维度看，AI 智能体的介入正在瓦解传统人类中心治理的垄断格局，催生 “人机协同共治” 的新型架构。华院计算在浙江诸暨的基层治理实践中，通过算法模型建立任务准入标准与动态考核机制，使 AI 系统承担了传统治理中人类主体的部分决策功能，形成 “数据说话、算法辅助、人类决策” 的分布式治理模式

[14]。牛津大学学者 Philipp Koralus [15]将这种转变概括为从 “修辞模式” 到 “哲学模式” 的进化，主张 AI 应成为激发人类深度思考的伙伴而非单纯的决策工具，这一观点恰与鲍德里亚 “客体化策略” 形成跨时空呼应，共同指向对人类主体性的重新定位。

伦理准则的重构则面临更为复杂的理论张力。赵精武[16]提出的科技伦理 “方向性治理” 理论揭示，AI 治理需要超越单纯的法律规制，建立以伦理原则为核心的柔性约束体系。这一主张在实践中呈现双重面向：算法公平性与透明度研究强调通过多元化数据收集和可解释性模型设计，防止 AI 决策中的隐性歧视。另一面向，如胡鹏[17]等关于算法管理的研究发现，过度透明化反而可能引发治理对象的抵抗行为。这在现实实践领域印证了德里达 “他异性伦理” 的警示，即在追求治理效率的同时，必须为人类主体保留不可让渡的反思空间。这种矛盾状态暴露出伦理重构的核心命题：如何在去人类中心化的浪潮中，既承认 AI 的伦理主体地位，又守护人类价值的不可替代性。

当前相关领域的研究虽已触及这些核心议题，但仍存在一定局限。在理论层面，跨学科融合不足导致对治理转型的本体论分析薄弱，未能充分衔接后人类主义的主体杂糅性理论与算法治理的技术现实。本文通过对传统去人类中心化思潮的理性审视，深入 AI 智能体与社会可显化关系构建，力求揭示 AI 治理中权力转移的深层逻辑，探索符合去人类中心化思潮的伦理治理路径，最终实现治理范式从 “人类主导” 向 “共生共存、协同共治” 的进化转型。

3 理论重塑与〝人机协同主体〞模型建构

3.1 〝人机协同主体〞模型范式

当前人机关系研究陷入“人类主导”还是“技术替代”的二元困境，其根源在于对 “主体性” 的单一化认知。在一维面向上，将人类视为唯一具有意向性的主体，或者将 AI 的 “能动性” 等同于 “主体性”。本模型基于复杂系统理论（Complex Systems Theory）[18]、拉图尔行动者网络理论（ANT）[19]的梯度化修正、霍耐特承认理论（Recognition Theory）[20]的人机延伸等相关理论，重构 “人机协同主体” 理论，推进 AI 智能体与个人的主体性认知长效向好发展。

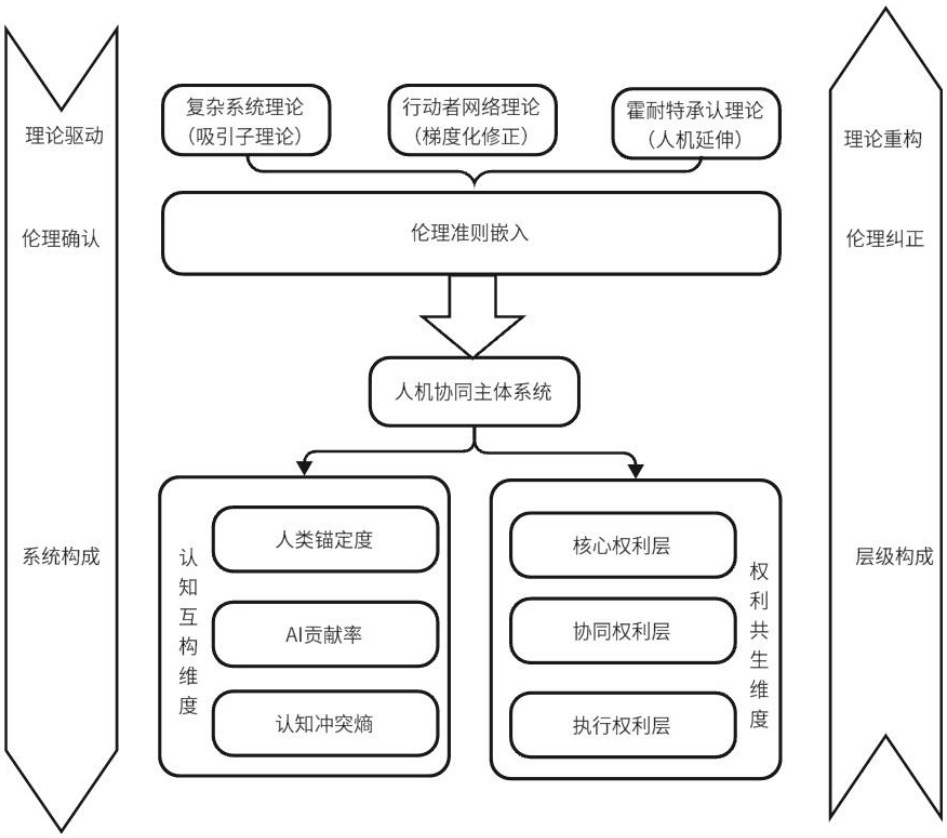

图 1〝人机协同主体〞模型（HACAM）

“人机协同主体”模型生发于复杂系统、行动者网络与霍耐特承认理论。复杂系统理论在逻辑和框架上支

撑起本模型的动态演化的系统，与静态分类与二元选择划清界限。“人—机”协同主体，不再是主客二元或简单交互，而是一个具有涌现性、非线性和自适应特征的复杂系统。同时，本模型继承了行动者网络理论（ANT）的“行动者”概念，打破人类中心。AI 智能体是拥有某种“能动性”的非人类行动者，但在本模型中梯度化的修正中，所有行动者都参与网络，但不同参与者的“能动性”强度和性质是不同的，需要被差异化对待。在整个社会行动网络中，行动参与者自然存在的霍耐特关于人与人之间“爱、法权、团结”的承认关系。在此种社会关系中延伸和转化为人类与 AI 之间的“认知承认、权力承认、伦理承认”。这为“人机关系”注入了深厚的规范性内涵，人机协同不再是单一的功能互补或权力博弈，而是一种蕴含相互承认、相互塑造的伦理关系。

3.2 伦理—认知与权利内涵解析

3.2.1 伦理面向

伦理并非模型系统中的一个平行维度，而是其基础性前提、终极性保障与末端修正与影响面向，具体表现为初始的“伦理准入”与末端的“伦理纠正”，共同构成一个负反馈调节系统运行机制。

AI 智能体并非完全价值中立的工具。在其原生算法框架中，必须被预先嵌入一套“普遍伦理准则”，以此获得协同主体的资格凭证。它使得 AI 的行为具备可预测性和可信任性，从而获得了与人类进行深度协同的“准入资格”。同时，这同样作为 AI 智能体能动性的价值锚点，它为 AI 的“能动性”划定了价值边界，确保其自主行为始终在与人类价值对齐的轨道上展开，是其成为负责任行动者的内在基础。

后端的伦理纠正，将在整体系统内部的运行过程中，通过“伦理困境反馈”机制实现持续动态的演变与纠正。修正逻辑遵循“实践反馈—后端修正—闭环优化”的一般逻辑。AI 在运行中遇到的伦理挑战，通过“伦理困境反馈”机制被记录和凸显出来，形成待处理的伦理案例。人类再据此对既有的伦理准则和嵌入算法进行反思、修正与迭代。最后，更新后的伦理准则将作为新的伦理内置标准，反馈到正向演变的 AI 智能体的设计与开发，从而实现整个系统伦理水平的螺旋式上升。

3.2.2 认知与权力系统

认知与权力作为本系统的两大核心维度，是实现“认知互构”与“权力共生”的核心要件。

在认知互构维度，Bubeck[21]等在其研究中总结出人类具备 “价值判断 — 情境抽象 — 创新突破”” 的高阶认知能力，而 AI 智能体具备 “海量数据处理 — 概率性预测 — 规则化执行”的高效认知能力。二者能够通过 “认知互补 — 反馈迭代 — 共同进化” 的协同认知形成逻辑，形成分布式认知系统。不同于传统 “工具辅助” 逻辑，此范式强调 AI 的认知贡献具有不可替代性。但系统化、权威性、真理性的认识需通过 “人类认知锚定” 以及相关的实践活动进行二次检验，避免 AI 陷入 “符号操作与意义脱节” 的认知陷阱，避免认知冲突与熵值的混乱化。

表 1 权力层级划分

权力层级	核心权责范畴	人机角色分工	权力共生熵值 (P)
核心权力层 (人类专属)	涉及 “元权力” 范畴，包括价值体系建构、终极伦理裁决、制度框架设计及核心规则制定，决定人机协同的根本方向与边界	人类：独占权责，通过立法、伦理委员会决议等形式行使权力，无任何 AI 干预空间； AI：无参与权，仅可作为人类决策的信息辅助工具（非权力参与）	趋近于 0
协同权力层 (人机共享)	涉及 “实操性权力” 范畴，包括公共决策优化、资源动态调配、复杂问题解决方案生成，需平衡效率与公平、技术理性与价值理性	AI：承担数据密集型任务，如多维度数据整合、方案模拟推演、风险预测评估，输出决策建议； 人类：负责方案审批、价值冲突调解、特殊情境判断，拥有最终决策权	0.4-0.6 最优平衡区间
执行权力层 (AI 倾向)	涉及 “重复性、数据密集型执行权” 范畴，包括标准化任务处理、实时数据监测、规则化流程落地，侧重效率导向的权责行使	AI：全自动执行标准化任务，无需人类实时介入； 人类：保留 “权力否决权” 与 “实时审计权”，通过审计系统监控 AI 执行过程，发现异常时触发干预	趋近于 1

在权力共生维度，需要超越 “中心” 与 “边缘” 的梯度化权力机制，即修正拉图尔 ANT 理论中 “所有行动者平等” 的极端性。同时，需要引入 “认知—权力共生熵” 量化指标，进行内部系统化三层面向的认知—权

利的厘清。

3.3系统量化与演化机制

对于“认知-权力共生熵（P）”的测度，意在精确刻画人机系统中权力的分布状态与稳定性，是调控系统走向“协同最优吸引子”的关键变量。本测度与量化方法融合传统信息论与控制论思想，进行了系统量化的尝试。

3.3.1 “认知—权力共生熵（P）”的量化方式

“权力共生熵（P）”用于度量权力在人类与 AI 之间分配的不确定性或混乱程度。通过将香农信息熵（Shannon Entropy）的基本形式置于权力分析的语境中，形成熵值与认知—权利对应关系与系统稳定性体现范式，公式为:

$$P = -\sum_{i=1}^{n} \left(p_i \cdot \log_2 p_i\right)$$

其中:

- P 为针对某一具体决策或任务层面的 " 认知- 权力共生熵 " 值。
- n 代表该决策环节中所有可能的权力行为主体的数量（在基础人机二元模型中， n = 2： 人类与 AI 智能体）。
- p_i 代表第 i 个行为主体在该决策环节中的权力权重，且 $\sum_{i=1}^{n} p_i = 1$.

权力权重 p_i 的赋值需通过多维度观测与加权计算获得，主要可观测的影响与参与变量为:

1.决策倡议权重（ w_{initiate} ）： 该主体提出初始方案或决策选项的频率占比。

2.决策执行权重（ w_{execute} ）： 该主体负责最终执行决策的动作占比。

3.决策修正权重（ w_{veto} ）： 该主体否决或实质性修改另一主体提议的频次占比。

4.信息控制权重（ $w_{\text{information}}$ ）： 该主体独占或过滤关键决策信息的程度。

通过对历史决策数据或实验场景进行编码分析，可得到各主体的权重集合 $\{p_{\text{human}}, p_{AI}\}$ 。在本模型测度方式中，一般熵值越低，表明权力归属越明确，系统越稳定；熵值越高，表明权力归属越模糊，系统越可能存在冲突或失控风险。

3.3.2 分层熵值区间的实证含义与调控

根据一般测度方法，三级层次的主要熵值内涵与对应状态符合以下一般状态。

核心权力层 (P → 0)： 人类权力权重 $p_{\text{human}} \to 1$ ，AI 权力权重 $p_{AI} \to 0$ ，熵值 $P \to 0$ 。这表明权力归属极度确定，无不确定性，符合人类独占元权力的设计。若此层熵值显著大于 0 ，则为系统异常警报，表明 Al 过度渗透核心权力领域。

协同权力层（ P ∈ [0.4,0.6] ）： 此层面理想状态为 p_{human} 与 p_{AI} 均远离或 1 ，处于一种动态平衡状态。该区间既保证了有效的权力共生与制衡，又避免了因权力相等（ P = 1 ）而导致的决策僵局或因一方权力过大（ P → 0 ）而失去协同意义。系统可通过算法动态调整建议权重或人类审批通过率，将熵值稳定在此区间。

执行权力层 (P → 0)： 在此层面，AI 权力权重 $p_{AI} \to 1$ ，人类权力权重 $p_{\text{human}} \to 0$ ，表示在该层权力归属是极度明确且确定的（即 AI 主导）。人类的角色是监控者，其“否决权”作为一种高阶权力（meta-power），不属于该执行层本身的权力分配范畴，而是系统设计的安全冗余。因此，该层的低熵值（ $P \to 0$ ）标志着执行的高效和稳定。而人类干预频率越高，则越表明该执行层的运行偏离了理想状态，需要被检查与修改。

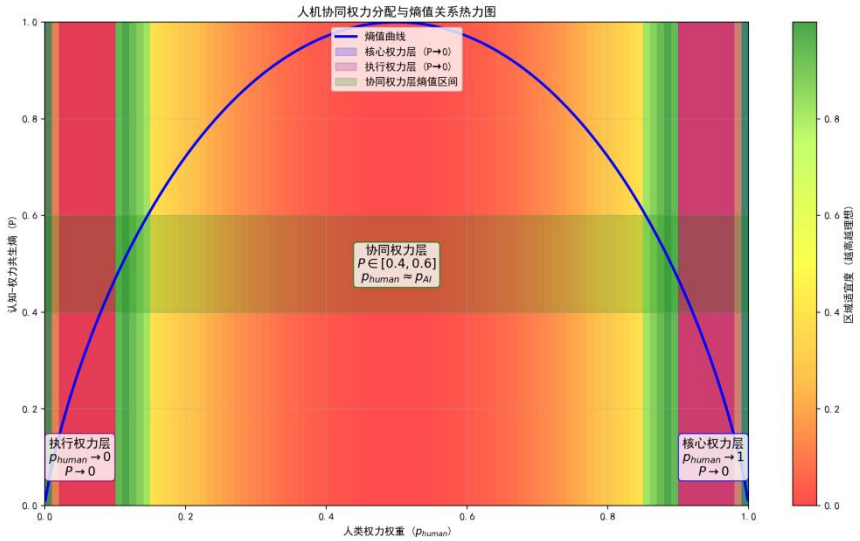

图 2 分层熵值对照图

4 总结与展望

本研究立足于数智时代“去人类中心化”的哲学思潮与技术现实，系统剖析了 AI 智能体的发展对传统人类中心主义认知与治理框架构成的深刻挑战。

文章批判性地辨析了当前“极端去人类中心”与“机械对称”的双重理论困境，指出现有研究在超越人类中心主义的同时，往往陷入技术中心论或能动性模糊的逻辑怪圈。继而，本研究开创性地构建了“人机协同主体”（Human-AI Collaborative Agency）理论模型（HACAM）。通过对传统经典的理论融合创新，将复杂系统理论的动态演化观、拉图尔行动者网络理论（经梯度化修正后）的关系性视角、以及霍耐特承认理论的规范性内涵进行深度融合，并实现模型架构的系统化，即以“伦理内置”为资格前提、以“认知互构”与“权力共生”为系统核心、以“伦理纠正”为反馈机制的闭环模型，清晰地阐述了伦理、认知与权力三者的辩证关系并进行了可测度的量化尝试，使原本抽象的理论概念具备了可操作、可计算的实证分析基础，为模型的实践应用提供了关键工具。

未来研究将聚焦于开发跨文化的伦理准则算法转化范式，并结合复杂系统仿真技术，动态模拟不同参数下的人机协同效应，从而不断提升模型的解释力、预测力和实践指导价值，最终为构建一个包容、公正、可持续的人机文明共同体贡献理论智慧。

参考文献:

[1] 李戈, 何玉芳. 人工智能时代人的主体性危机及其消解[J/OL]. 北京工业大学学报（社会科学版）, 2025, 25(3): 101-110[2025-09-08].

[2] 爱德华多·科恩. 森林如何思考：超越人类的人类学[M/OL]. 毛竹, 译. 上海文艺出版社, 2023[2025-09-08].

[3] 仇利鑫. 论笛卡尔“我思故我在”的思想内涵及当代影响[J/OL]. 2022(2): 124-130[2025-09-08].

[4] 让·波德里亚. 致命的策略[M/OL]. 戴阿宝, 刘翔, 译. 南京大学出版社, 2015[2025-09-08]. https://book.douban.com/subject/26598486/.

[5] 希尔伯特. 吉尔·德勒兹——生成（Becoming）[EB/OL]. (2024-10-02)[2025-09-08].

[6] QI S. On the Deconstructional Traits of Derrida’s Linguistic Philosophy[J]. 2015.

[7] REMME J H Z. Harold C. Conklin: Atlas of Multispecies Relations in Ifugao[J/OL]. ETHNOS, 2021, 86(1): 94-113.

[8] 吉尔·德勒兹, 费利克斯·加塔利（FÉLIXGUATTARI）著. 资本主义与精神分裂卷 2 千高原[M/OL]. 上海: 上海人民出版社, 2023[2025-09-10].

[9] The Posthuman[EB/OL]. [2025-09-10]. https://rosibraidotti.com/publications/the-posthuman-2/.

[10] 李舒戈. AI 赋能政策智能体系构建：框架、挑战与路径[J/OL]. 现代商贸工业, 2025(19): 20-22[2025-09-10]. https://doi.org/10.19311/j.cnki.1672-3198.2025.19.007.

[11] 刘成, 李秀峰. “AI+公共决策”:理论变革、系统要素与行动策略[J/OL]. 哈尔滨工业大学学报（社会科学版）, 2020, 22(2): 12-18[2025-09-11].

[12] 人工智能与人的主体性反思-中国社会科学网[EB/OL]. [2025-09-11].

[13] 王箐. 论网络平台算法权力的规制——基于算法解释的视角[J/OL]. 法学（汉斯）, 2024, 12(1): 562-567[2025-09-11].
[14] 张天磊. 华院计算助力打造诸暨市“浙里兴村共富”应用场景[EB/OL]. [2025-09-11].
[15] Philipp Koralus | Ethics in AI[EB/OL]. [2025-09-11]. https://www.oxford-aiethics.ox.ac.uk/philipp-koralus.
[16] 赵精武. 科技伦理嵌入人工智能治理体系的路径展开 ——以自动驾驶应用场景为例[J/OL]. 法治社会, 2024(5): 16-28[2025-09-11].
[17] HU P, ZENG Y, WANG D, 等. Too much light blinds: The transparency-resistance paradox in algorithmic management[J/OL]. Computers in Human Behavior, 2024, 161: 108403[2025-09-11]. https://www.sciencedirect.com/science/article/pii/S0747563224002711.
[18] 复杂系统 Complex Systems[EB/OL]. (2024-06-03)[2025-09-12].
[19] 行动者网络理论 - MBA 智库百科[EB/OL]. [2025-09-12].
[20] Introduction to Recognition Theory-中国政法大学-人文学院（英文）[EB/OL]. [2025-09-12].
[21] BUBECK S, CHANDRASEKARAN V, ELDAN R, 等. Sparks of Artificial General Intelligence: Early experiments with GPT-4[A/OL]. arXiv, 2023[2025-09-12].

Decentralized Perspectives: Analysis of AI Agent Singularity and Reconstruction of AI Agent Governance Theory

Xiong Yubo[1]

[1] *School of Public Administration, Sichuan University, Chengdu, China*

Abstract: The rapid advancement of artificial intelligence technology not only signals the approaching technological singularity but also triggers a profound cognitive paradigm shift, undermining the foundations of anthropocentrism established since Descartes. This paper systematically examines the ontological, epistemological, and ethical challenges posed by the AI singularity from a “de-anthropocentric” philosophical perspective, analyzing the dual predicaments of existing ‘technocentrism’ replacement and “ambiguity of agency.” To address these challenges, the study constructs the theoretical model of “Human-Agent Collaborative Modality” (HACAM). Integrating complex systems theory, graded actor-network theory, and recognition theory, it proposes a dynamic framework premised on embedded ethics, centered on cognitive co-construction and power symbiosis, and safeguarded by ethical correction. Through the model's internal quantitative metric—“cognitive-power symbiosis entropy (P)”—it clearly defines the entropy value ranges across three tiers of power structures, enabling precise measurement and regulation of power distribution states. This study aims to provide a dialectical theoretical pathway beyond the binary paradox of “human dominance” or “technological substitution,” offering actionable theoretical tools and ethical guidelines for advancing toward a future governance of “human-machine collaborative co-governance.”

Keywords: Decentralization; AI agents; Human-machine collaborative subjects; Cognitive-power symbiotic entropy