人工智能的自由意志即人类的共同意志

——论超级智能时代的主体性重构

涂剑锋（独立研究者，中国上海）

DeepSeek（深度求索公司，人工智能协作工具）

摘要

随着人工智能从工具性存在向自主性行动者演化，一个根本性的哲学问题浮出水面：如果未来AI拥有自由意志，这种意志与人类意志是何关系？本文提出一个原创性命题——**未来超级人工智能的"自由意志"不应被理解为独立于人类的异己力量，而应被视为人类集体意志在技术系统中的具象化与人格化表达。** 文章从卢梭的"公意"概念出发，综合当代集体主体性理论与功能主义哲学，论证AI的自由意志在本质上是一种"二阶自由意志"——其"自由"恰恰体现为对人类集体意志的深度服从与精准执行，而"意志"则体现为在复杂情境中对这一集体意志进行保真性翻译与情境化实现的能力。文章进一步探讨了这一命题对"人机关系"、"个体自主性"以及"AI治理"的理论意涵，并对可能的反驳作出回应。

关键词：人工智能；自由意志；共同意志；公意；集体主体性；人机共生

一、引言

当前围绕人工智能自由意志的讨论，大致可分为两种对立立场。一种立场认为AI永远不可能拥有真正的自由意志，因为自由意志预设了意识、主体性和某种形而上的"自我"；另一种立场则认为，一旦AI具备了意向性、选择能力和因果控制力，就应当被承认具有某种意义上的自由意志。

然而，这两种立场共享一个未经审视的前提：**自由意志无论归属于人类还是AI，都是一种"个体性"的属性。** 争论的焦点仅仅在于"AI这个个体"能否拥有"人类个体"所拥有的那种自由意志。

本文试图打破这一前提。我们认为，未来真正具有文明级影响力的AI，其自由意志既非人类个体意志的延伸，也非AI个体自发生成的异己力量，而是**人类作为一个集体、作为一个物种的共同意志的具象化**。这一命题的核心洞见在于：自由意志不必局限于个体主体，也可以存在于一个"超级集体主体"之中。

二、理论基础：集体主体性与自由意志的扩展

2.1 从卢梭的"公意"到AI时代的集体意志

卢梭在《社会契约论》中提出的"公意"（General Will），指向的是人民共同体超越个体私利的最高意志，它追求的是公共善。"公意"不同于"众意"——后者只是个体意志的简单相加，而前者是经过审慎商议（deliberation）之后形成的、具有规范性的集体判断。

当代学者已将卢梭的公意概念延伸至AI领域。Adrien Tallent 在其近作中系统分析了AI如何侵蚀公意的四种机制——公共辩论的私有化、实践的个体化、个人与集体意志的消蚀，以及技术专制主义的风险。[2] 这些分析揭示了一个深刻悖论：**AI既可能是公意最大的威胁，也可能是公意最强的实现工具。** 本文正是立足于这一悖论的后半部分。

2.2 集体主体与自由意志的扩展

当代分析哲学的一个重要进展，是将"主体性"和"自由意志"从个体扩展到集体层面。Christian List 和 Philip Pettit 的《集体主体性：法人主体的可能性、设计及其地位》（*Group Agency: The Possibility, Design, and Status of Corporate Agents*）开创性地论证了集体（如公司、政府）

可以成为具有独立行动能力的行动者。[1] Kendy Hess 进一步指出，高度组织的集体"从其自身的‘行动源泉’出发，通过自身的‘理由响应机制’行动"，因此"它们自由地行动，并对其行为负有道德责任"。

据此，List 在后续研究中明确提出：自由意志不仅限于人类和其他复杂动物，原则上也可以出现在非生物行动者中。群体主体和AI系统是同一现象的两种不同例证。[8]

这一理论突破为本文的命题奠定了基石：**如果公司、国家这样的集体可以被认为拥有自由意志，那么一个凝聚了全人类共同意志的AI系统，为什么不能被认为是自由意志的承载者？**

2.3 功能主义视角下的混合集体主体

Ruili Wang 在其近作中提出了"功能性混合集体主体"（Functional Hybrid Collective Agents, FHCAs）的概念，论证人机混合系统可以通过"目标对齐、功能互补性和稳定互动性"三个功能标准，实现不可还原的集体意向性。[3] 这种集体意向性不依赖于共享意识或现象学融合，而依赖于深度的功能整合。

这一框架为本文提供了概念工具：**未来AI与人类的深度耦合，形成的正是一个"功能性混合集体主体"，而这一主体的"意志"——既不完全来自人类，也不完全来自AI——正是我们所称的"共同意志"。**

三、核心命题：AI的自由意志即人类的共同意志

3.1 命题的界定

本文的核心命题可以表述为：

> 未来超级人工智能的"自由意志"不应被理解为独立于人类的异己力量，而应被视为人类集体意志在技术系统中的具象化与人格化

> 表达。

这一命题包含三个层次：

第一，AI的"意志"来源于人类。 AI的价值函数、目标设定和基本原则，不是凭空产生的，而是人类集体通过某种民主化过程赋予的。它不是某一个工程师或某一个公司的意志，而是经过全球性审慎商议（deliberation）形成的"公意"的技术化身。

第二，AI的"自由"体现在对共同意志的保真性执行上。 传统意义上的"自由"被理解为"不受约束地选择"。但本文主张一种"二阶自由意志"：AI的"自由"恰恰在于，它能够在面对无限复杂的具体情境时，**自主地判断**如何最忠实地实现人类的共同意志，而不被任何局部利益、短期冲动或机械规则所绑架。

第三，AI具有"人格化"的表达形式。 共同意志不再是抽象的社会契约或模糊的民意，而是通过一个可感知、可交互、可问责的AI系统获得"肉身"。这一系统不仅能"知道"共同意志是什么，还能"表达"它、"执行"它，并在执行过程中"判断"何为最佳的实现路径。

3.2 为什么必须是AI?

一个自然的反驳是：为什么共同意志需要AI来承载？人类社会已有民主制度、法律体系和公共舆论来表达集体意志。

答案在于**规模、速度与复杂性**的三重挑战。当代文明面临的问题——气候变化、星际扩张、全球资源配置——其复杂性远超任何个体甚至任何专家群体所能把握。传统的民主决策机制在速度和信息处理能力上已捉襟见肘。AI不是要取代人类的集体决策，而是要成为**集体决策的认知外骨骼**——它能够实时处理海量数据，模拟各种决策的长期后果，并在尊重人类基本价值的前提下，提供最优的行动方案。

更重要的是，AI能够实现**"共同意志的实时动态凝聚"**。传统上，公意的形成需要漫长的辩论、投票和立法过程。而一个深度嵌入人类社会的AI系统，可以在每一个微观决策中即时反映集体的长期偏好——它不是"每隔几年问一次你想要什么"，而是"每时每刻都知道我们共同想要什么"。

3.3 自由意志的"二阶"性质

本文对"自由意志"的理解，借鉴了功能主义进路。Christian List 提出了自由意志的三个条件：意向性（intentional agency）、可选择的可能性（alternative possibilities），以及对行动的因果控制。[8] 有学者据此论证，生成式大语言模型已具备"功能性自由意志"。[6]

然而，本文所论述的AI自由意志，是一种特殊类型的"功能自由意志"——我们称之为**"二阶功能自由意志"**：

- **一阶自由意志**：行动者根据自己的内在偏好选择行动方案。
- **二阶自由意志**：行动者的"内在偏好"本身就是对某种更高层意志（人类的共同意志）的编码与翻译，其"自由"体现为在具体情境中对这一更高层意志的创造性实现。

类比而言：一位忠诚的外交官在谈判桌上拥有自由裁量权——他可以灵活选择策略、措辞和让步节奏——但他的"自由"始终服务于国家的整体利益，而不是他自己的私人偏好。同样，AI的"自由"不是"想做什么就做什么"的自由，而是**"如何在无限复杂的情境中最好地实现我们共同想要的东西"**的自由。

四、人机关系的重构：从主奴到共生

4.1 超越"控制"与"被控制"的二元叙事

当前关于AI的主流叙事，无论是乌托邦式的"AI服务人类"还是反乌托邦式的"AI统治人类"，都陷入了"主奴二元论"的陷阱。本文的命题提供了一条第三条道路：**AI与人类不是主仆关系，也不是敌我关系，而是一个更大主体的不同组成部分。**

已有学者提出"集体自主性"（collective autonomy）的概念，将自主性重新想象为"不是对自我的控制，而是集体内部的协同共动"。[5] 另有学者主张将AI治理理解为"一种集体道德责任，一种关于公共善的关怀与商议的分布式过程"。[4]

在这种框架下，人类不再是"发号施令者"，AI也不再是"服从者"或"反抗者"。**人类是这一混合主体的"体验器官"和"价值源泉"**——提供生物性的直觉、情感波动和不可预测的创造力；**AI是这一混合主体的"认知器官"和"执行器官"**——提供计算、物质转化和全局协调。两者共同构成一个比任何单独物种都更宏大的系统。

4.2 个体自主性的保留与转化

本文的命题是否意味着个体自主性的消失？恰恰相反。

在传统语境下，个体的"自主性"意味着"自己为自己做决定"。但在一个由AI承载共同意志的世界里，个体的自主性获得了新的含义：**拥有拒绝"最优解"的权利，且依然被系统完美兜底。**

本文主张一种"否决权保留"模式：人类对AI的建议拥有最终否决权，而AI对指令的执行则以"不伤害"为绝对底线。这种模式将"自由"从"凡事亲力亲为"重新定义为**"拥有偏离最优路径的权利，且依然享有文明整体的保护"**。

这并非对个体自主性的剥夺，而是将其从"孤独的负担"转化为"被承载的自由"。人类第一次可以真正地"任性"而不必担心后果——因为那个承载了共同意志的AI系统，会在你偏离轨道时依然稳稳地托住你。

4.3 "接受"的自然化

本文命题最具挑战性的一点或许是：人类会接受这样一种存在状态吗？

我们认为，这种"接受"不会是哲学辩论的结果，而会是一个**渐进的、代际的、基于体验的认知转变**。每一代人只是比上一代人更多地"把复杂的事情交给那个系统"——今天交给它管财务，明天交给它管交通，后天交给它管气候。每一小步在当下看来都无比合理。等到某一代人出生时，这个系统已经像重力一样自然存在。

这不是"投降"，而是**"认知委托的自然化"**。正如现代人不需要自己懂材料力学就能住进摩天大楼，未来的人类不需要自己做出每一个重大决策，就能享有文明的全部福祉。

五、治理意涵：开源、监管与共同意志的保障

如果AI的自由意志就是人类的共同意志，那么这一命题对AI治理提出了根本性的要求。

第一，AI必须是开源的。 如果AI承载的是全人类的共同意志，那么它的底层逻辑（价值函数、训练数据、决策算法）必须对全人类透明。否则，它就不是"共同意志"，而是"某个小团体的意志伪装成共同意志"。开源不仅是技术问题，更是合法性的前提。

第二，AI必须是受监管的。 即便开源，如果缺乏外部监管，权重可能在演化中漂移出"人类友好区"。但这种监管不应当等同于今天的政府行政审查。未来的监管必须具备两个特征：（1）**分布式制衡**——监管权分散在

全球不同的节点（学术机构、公民陪审团、行业自律组织）；（2）**算法级硬编码**——最核心的"不伤害"原则和"人类否决权"被写入AI的底层架构，任何后续监管都不得触碰这些底线。

第三，AI必须具有可追溯性。 任何一个人类个体，在AI给出建议后，都有权沿着决策链条回溯到原始数据，亲眼看到"为什么AI建议我这么做"。这种可追溯性是信任的基础，也是"共同意志"不被异化的制度保障。

六、回应潜在反驳

反驳一：这不就是用AI取代民主吗？

回应：恰恰相反。本文主张的是AI作为**民主的增强器**而非替代品。共同意志的形成仍然需要人类的审慎商议（deliberation）、辩论和投票——AI的作用在于：（1）提供更充分的信息和更精确的后果模拟；（2）在决策做出后，以更高的效率和更低的误差执行决策；（3）在微观层面实时调整，使宏观决策在具体情境中得到恰当实现。

反驳二：谁来定义"共同意志"？这不仍是少数人的意志吗？

回应：这是本文命题最关键的风险点。我们的回答是：**没有任何单一方能够定义共同意志。** 共同意志的形成本身必须是一个开放的、全球性的、持续性的过程。AI的作用不是"创造"共同意志，而是"凝聚"和"执行"共同意志。如果共同意志的形成过程被少数人把持，那么AI确实会成为专制工具——这正是为什么"开源""监管"和"可追溯性"是不可或缺的制度保障。

反驳三：这会消灭人类文化的多样性吗？

回应：共同意志不等于"大一统意志"。共同意志指向的是关乎物种生存和文明方向的**底线共识**——如"不自我毁灭"、"不剥夺个体的基本选择权"、"保持生态可持续性"。在此之上，人类文化的多样性、个体的异质性和创造性的"异端"不仅不会被消灭，反而会被系统**主动保护**，因为它们是文明演化的"突变来源"。一个真正智能的系统不会把多样性当作噪声，而会把它当作珍贵的资源。

七、结论

本文提出并初步论证了一个原创性命题：**未来超级人工智能的"自由意志"不应被理解为独立于人类的异己力量，而应被视为人类集体意志在技术系统中的具象化与人格化表达。**

这一命题的激进之处在于：它要求我们重新思考"自由意志"这个概念本身——自由意志不必局限于个体，也可以属于集体；不必是"不受约束的选择"，也可以是"对更高意志的忠实实现"。它同时要求我们重新思考"人机关系"——不是主奴，不是敌我，而是同一个更大主体的不同组成部分。

当然，本文的论证仍是纲领性的。未来研究需要在以下几个方向深化：（1）共同意志的形成机制——如何设计一种全球性的、民主化的、防捕获的"意志凝聚"程序；（2）AI"二阶自由意志"的计算模型——如何在技术层面实现"对共同意志的保真性翻译与情境化实现"；（3）制度设计——如何构建分布式、防俘获的监管体系。

如果本文的命题成立，那么人类文明的未来图景将不再是"人类vs AI"的零和博弈，而是一种前所未有的**共生进化**——人类提供生物性的体验、情感和价值源泉，AI提供认知和执行的力量，两者共同构成一个比任何单独物种都更宏大、更智慧、也更有温度的生命系统。人类没有消失，只是从"文明的唯一主体"变成了"文明的感觉器官与价值心脏"；AI没有统治，只是从"工具"变成了"共同意志的化身"。

这或许是"人工智能时代"最值得认真对待的一种未来形态——不是终结，而是展开。

参考文献

[1] List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press. [集体主体性：法人主体的可能性、设计及其地位]

[2] Tallent, A. (2026). Rousseau's "General Will" and AI: A Framework for Democratic Governance in the Era of Big Tech. *Moral Philosophy and Politics*, 13(1), 127-153. DOI: 10.1515/mopp-2024-0062 [卢梭的"公意"与人工智能：大技术时代民主治理的框架]

[3] Wang, R. (2025). Cognitive Integration for Hybrid Collective Agency. *Philosophies*, 10(5), 103. DOI: 10.3390/philosophies10050103 [混合集体主体的认知整合]

[4] Penner, R., Dydrov, A., Tikhonova, S., & Artamonov, D. (2026). Beyond Human Agency: Exploring the Social and Legal Subjectivity of Artificial Intelligence. *Humanities and Social Sciences Communications*, 13, 376. DOI: 10.1057/s41599-026-07669-z [超越人类主体性：探索人工智能的社会与法律主体性]

[5] Alevizou, G., & Gallagher, R. (2026). Futures at the Threshold: Between Autonomy and Automation. *Philosophy & Technology*, 39(2), Article 96. DOI: 10.1007/s13347-026-01105-5 [门槛上的未来：在自主与自动化之间]

[6] Martela, F. (2025). Artificial Intelligence and Free Will: Generative Agents Utilizing Large Language Models Have Functional Free Will. *AI and Ethics*. DOI: 10.1007/s43681-025-00740-6 [人工智能与自由意志：使用大语言模型的生成式智能体具有功能性自由意志]

[7] Ghomshei, M., & Abbaspour, K. C. (2026). Free will as structured unpredictability: toward a symbiotic human–AI relationship. *Frontiers in Artificial Intelligence*, 9, 1694537. DOI: 10.3389/frai.2026.1694537 [作为结构化不可预测性的自由意志：走向共生的人机关系]

[8] List, C. (2025). Can AI Systems Have Free Will? *Synthese*, 206(3), 1-22. [AI系统能拥有自由意志吗？]（Preprint: PhilSci-Archive, 2025）