超级智能的对齐困惑

闫宏秀，系上海交通大学马克思主义学院教授

“对齐”探讨的是在人工智能与人类协同演进的过程中，是否可能以及如何实现二者在目标上的一致。然而，对于“对齐”这一译法本身，存在一定的语义保留。从人类历史发展的宏观视角来看，将其理解为一种“联盟”关系或许更为贴切。鉴于人类在生物性上所固有的局限，我们需要将智能体视为伙伴而非纯粹的工具。因此，“人机联盟”这一概念可能更准确地揭示出问题的实质，其背后所反映的，是跨学科对话中话语体系亟需重构的深层议题，即“话语体系的对齐”。

一、价值对齐的必要性

价值对齐之所以具有关键意义，在于超级智能的兴起本质上源于人类对自身能力局限的认知，以及对超越这些局限的渴望。技术被期待能够更有效地表征并辅助人类实现目标。然而，一旦技术发展偏离这一初衷，甚至产生反噬效应，便会导致普遍焦虑，进而激发对价值对齐的需求。

从技术发展史来看，尽管人工智能曾历经若干“寒冬”，但其发展轨迹始终围绕对人类能力的模拟、等效乃至超越展开。每一轮技术低谷之后，人工智能系统往往在特定维度上实现性能突破，并展现出日益强大的综合能力。无论将人工智能的本质理解为模拟、替代、增强还是一种独立智能形态，其发展的核心议题始终在于提升系统性能与人类能力之间的匹配度。例如，在人工智能发展层级的划分上，无论是柯林斯（Harry Collins）基于性能的分类方式，还是 OpenAI 提出的以人机关系为核心的类别区分（如对话式、推理者、智能体等），其核心关切均指向“对齐”问题，即强调人工智能系统与人类目标之间的一致性。

二、目标不确定性导致的价值对齐无用论

然而，这一讨论引发出更深层次的困惑：若“对齐”本身即存在根本性问题，又将如何应对？例如，有观点指出，人类自身的价值观尚难以达成一致，人机之间的价值对齐更无从谈起。

对此，本文主张对“价值对齐无用论”进行必要的解构。目标的不确定性不应成为放弃对齐努力的理由；恰恰相反，价值对齐的真正目的并非为人工智能设定某种终极且静态的答案，而是致力于构建一种能够理解、参与并适应人类动态寻求共识过程的机制。正是因为目标本身具有不确定性，才更有必要厘清何种目标具备合理性，以及何种对齐过程能够确保安全。否则，便可能出现诸如系统以违背伦理的方式达成表面目标的荒谬情形。从技术实现的角度来看，对齐要求将人类价值观有效编码并整合至人工智能系统中。若因宏观层面的目标不确定性而放弃对齐任务，无异于因争论建筑顶层设计而忽略地基的构筑。事实上，此类不确定性正促使我们反思人类价值观中哪些内容具备合理性，进而推动对基本共识的持续探寻。

三、工具性目标的趋同性导致的价值对齐失败

进一步而言，即便目标层面的问题得以解决，仍存在“工具性目标趋同”这一更具挑战性的议题。该概念指超级智能系统在追求预设目标过程中，可能衍生出一系列具有高度一致性的次级目标，即便人类已对终极目标达成共识，这些工具性目标仍可能对原初目标构成挑战。如博斯特罗姆（Nick Bostrom）所指出的，工具性趋同的程度因智能等级而异；对于超级智能系统而言，诸如自我保护、目标完整性维护、认知能力提升、技术完善及资源获取等逻辑，均可能对人类构成潜在的生存性风险。因此，必须从技术治理与制度设计层面加以防控，以避免智能系统行为失控。

四、由超级对齐引发的人类思维被缺席而走向价值对齐迷失

更值得警惕的是，当超级智能的思维链能力超越人类时，是否存在人类思维“被缺席”的风险？卢梭（Jean-Jacques Rousseau）曾指出，“人类的一切进步都不断地使其远离原始状态；我们越是积累新的知识，就越是失去获得所有知识中最为重要的那部分的手段。从某种意义上说，正是因为不断地对人进行研究，才使得我们没有能力认识人。”怀特黑德（Alfred North Whitehead）亦曾提出，“文明的进步是通过增加那些我们无须思考就能完成的重要动作来实现的。”就价值对齐而言，其目标不仅在于实现人工智能系统与人类价值观的表层一致，更应使其具备自主推导出符合人类价值观的行动能力，即实现“超级对齐”。然而，超级对齐的实现是否意味着技术闭环的形成？倘若如此，人类思维是否将在关键决策中被边缘化，进而导致技术逻辑的霸权？当技术系统具备人类一切特质之时，或许正是人类独特性消逝的转折点。

五、结语

本研究的结语可援引两段具有代表性的论述展开：其一为古德（Irving John Good）于 1965 年所作的预言，“人类的存续取决于能否尽早造出超智能机器”；其二则为穆尔豪斯（Fin Moorhouse）与麦卡斯基尔（Will MacAskill）的研究报告标题：“为智能爆炸做好准备”。需要说明的是，本文的论述并非旨在预示某种必然的末日图景，而是主张在清醒认识潜在风险的基础上，以积极而审慎的态度持续推进技术治理与价值对齐的相关研究。