哲学 AI：探索一种“哲学与 AI 共生”的工作样态

孙向晨　贾子菡　张　祺

摘　要　随着 AI 的迅猛发展，AI 哲学也随之展开一系列相关讨论。AI 哲学可分为消极型与积极型两种。前者停留在试图建立起人类与 AI 之间的“本体论隔离”，后者则试图摆脱哲学滞后性的被动反思模式，前瞻性地为 AI 演化提供必需的概念框架与思维范式。“哲学 AI”概念的提出，更强调哲学与 AI 的双向塑造，既强调 AGI 演进中哲学起到的必要作用，也注重 AI 思考方式助力哲学的发展。在“哲学家 AI 助手”（Philosopher's AI Assistant）的平台，以孙向晨与某当代著名哲学家 AI Agent 的虚拟论辩为实例，可具体展现哲学 AI 发展可能的空间与存在的问题，以及其最终所指向的一种“哲学与 AI 共生”的立场。

关键词　AI 哲学　哲学 AI　哲学家 AI Agent　哲学与 AI 共生

作者孙向晨，复旦大学哲学学院教授（上海 200433）；贾子菡，Microsoft AI Asia OPE 项目实习生（北京 100871）；张祺，Microsoft AI Asia 总裁（北京 100080）。

中图分类号 B0　　**文献标识码** A　　**文章编号** 0439-8041(2026)02-0005-10

AI 正在人类多重知识领域推动着知识发展，尽管发展尚有瓶颈，发展的速度也不一定符合人们的预期，但 AI 在多方面所展示出的智能潜力却不容小觑，从日常对话到人文领域，从数码编程到科学进展，不同的生活和知识领域都可以从自身角度对于 AI 潜力进行深入的评估，可以有不同量级的尺度：智能内部的迭代，现代技术的进展，工业革命以来的变迁。在笔者看来，从哲学与 AI 这一微观的关系来审视，仍可以管窥 AI 与人类文明演化的宏大关系。人类在轴心时代所建构的认知图景与价值体系，为人类理解自身及其在世界中的位置提供了稳定框架。人类对技术发展的理解、评价与规范，很大程度上都收束于此框架之内。然而，AI 的迅猛发展，特别是以自主性与黑箱决策为特点的智能系统，将对轴心文明以来的人类认知架构构成持续挑战。如此大的判断看起来未免虚妄，需要在微观层面有真正的入手之处。哲学作为人类表达智慧的基本类型，与各大文明的走向息息相关。有鉴于此，我们要努力探究哲学与 AI 的内在关联，不仅从哲学上反思 AI 的特质，助力 AI 发展，更为关键的是，探索 AI 对哲学研究的帮助。为此，Microsoft AI Asia 团队开发了“哲学家 AI 助

手”平台[①]，尝试让AI介入到哲学研究之中，通过AI在哲学中的发展理解AI之于文明的重要意义。

一、AI哲学I：哲学的反思及其局限

哲学作为反思的学科，自AI诞生之日起，便一直将AI作为自己的研究对象。1956年达特茅斯会议首次提出了“人工智能”（AI）这一术语[②]，自此AI的发展经历了数次热潮与寒冬，每次起伏都会伴随对AI的哲学讨论。“AI哲学”便是以哲学的方式来研究与探索AI的特质与发展，关注AI对文明与技术关系的深远影响，我们可以粗略地将其区分为四个阶段。

第一阶段在20世纪50年代，以符号主义为特征。符号主义以逻辑推理掀起AI的第一次浪潮，以机器证明（ATP）为其主要代表，试图通过计算机程序来完成数学定理或逻辑命题的证明，但这一进展却因无法处理复杂现实问题而退潮。这一时期AI哲学的核心议题是“智能的形式化边界”，探讨将人类逻辑推理形式化的前景与局限，AI的出现在哲学上已经开始挑战人与技术的传统关系。这一时期，阿兰·图灵（A. M. Turing）将“能否表现出类人行为”作为衡量机器智能的新标准，也即后世所称的“图灵测试”[③]。在哲学上，这实质上是以“行为表现”为标准来评估智能水平，这一努力模糊了以往所认为的人类智能（内在心智）与机器智能（符号推理）之间的鸿沟，形成了判断智能的某种统一标准。

第二阶段是20世纪的80年代，以“专家系统”（Expert System）为代表，“专家系统”通过模仿人类专家的决策系统，借助于电脑计算，试图解决特定领域中的复杂问题。“专家系统”凭借垂直领域的规则模拟，曾短暂地推进（以MYCIN[④]、DENDRAL[⑤]为代表）了AI的发展，但因系统的脆弱性以及高成本的投入而陷入沉寂。针对这一时期AI的发展，哲学的核心议题是探究“知识编码的极限”，此阶段的哲学讨论大多通过捍卫“非形式化”和“意义理解”的向度以维护人类智能与经验的独特性。如休伯特·德雷福斯（Hubert Dreyfus）从现象学切入，强调人类“具身性”与“情境性”的特点是符号系统所不能穷尽的[⑥]；约翰·塞尔则通过“中文屋论证”，指出仅操纵符号是无法产生真正的意义，认为人类心智的“意向性”是机器所不可复制的[⑦]。

第三阶段是2010年代的“深度学习”，“深度学习”试图模仿人脑的神经网络结构，通过“多层”神经元来处理数据，从而学会识别模式，做出决策。“深度学习”依靠大数据与算力的爆发改写了AI的格局（以AlexNet，AlphaGo为代表），但仍局限于单模态的任务。这一时期的成果在一系列涉及感知、决策的具体任务上逼近甚至开始超越人类，其巨大成功似乎颠覆了先前哲学对“形式化不可能”的论断，促使哲学议题转向“数据认知的合法性”问题，例如尼克·波斯特罗姆（Nick Bostrom）对“超人类智能风险”的探讨，隐含了对于AI自主性的巨大忧虑。[⑧] 此阶段，人与技术的关系开始超出单纯“工具使用”的框架，向竞争与合作的主体间互动性演变，AI的发展迈出了实质性的一步。

---

① 在微软全球资深副总裁、Microsoft AI Asia总裁张祺博士提出的OPE（One-Person Entrepreneur，单人创业家）理念指引下，Microsoft AI Asia研发了“哲学家AI助手”（Philosopher's AI Assistant）平台，用于研究探索哲学与AI结合的方式。项目团队为：叶晚乔，Microsoft AI Asia首席战略官；魏思宁，Microsoft AI Asia首席应用科学与研发总监；花贵春，Microsoft AI Asia首席应用科学家；黄泽彦，Microsoft AI Asia高级应用科学家；李红亮，Microsoft AI Asia高级软件工程师；贾子菡，Microsoft AI Asia OPE项目实习生。合作方为北京大学博古睿研究中心。由2024—2025博古睿学者、复旦大学哲学学院教授孙向晨担任学术指导，博古睿研究院高级副院长、中国中心主任宋冰提供学术支持。

② John McCarthy et al., “A proposal for the Dartmouth Summer Research Project on Artificial Intelligence,” 1955, www-formal. stanford. edu/jmc/history/dartmouth/dartmouth. html, accessed October 1, 2006.

③ Turing, A. M., *Computing machinery and intelligence*, Springer Netherlands, 2009, pp. 23-65.

④ Van Melle, W., “MYCIN: a knowledge-based consultation program for infectious disease diagnosis,” *International Journal of Man-Machine Studies*, 10(3), 1978, pp. 313-322.

⑤ Buchanan, B. G., & Feigenbaum, E. A., “DENDRAL and Meta-DENDRAL: Their applications dimension,” in *Readings in artificial intelligence*, Morgan Kaufmann, 1981, pp. 313-322.

⑥ Dreyfus, H. L., *What computers can't do: The limits of artificial intelligence*, 1972.

⑦ Searle, J. R., “Minds, brains, and programs,” *Behavioral and brain sciences*, 3(3), 1980, pp. 417-424.

⑧ Nick, B., *Superintelligence: Paths, dangers, strategies*, 2014.

第四阶段则以当下 ChatGPT[①] 为代表的“生成式 AI”，通过 Transformer 架构、万亿级参数跨领域数据预训练，实现 AI 对人类语言与理性能力的突破，具备了跨领域通用推理能力。相较于历史上昙花一现的“专用智能”，AI 摆脱了“专家系统”僵化规则与“深度学习”单一任务的局限，展示出指向认知革命的“新智能物种”的可能性。这一时期的 AI 哲学开始讨论“通用智能体是否具备主体性资格”等问题，如大卫·查尔默斯（David Chalmers）提出，当技术表现出泛化的理性与理解力时，我们是否必须重新审视其在本体论上的地位。[②]

通过简单的梳理，这一历程中不同阶段的哲学工作展现出 AI 哲学的某些特点：对“人类与技术”的哲学反思保持着密纳发猫头鹰的节奏，始终滞后并受制于 AI 技术的进步，这本质上源自西方哲学古老传统的局限性。虽然不同阶段的 AI 哲学讨论有着不同的侧重点，以哲学方式研究 AI 的主线索却是一致的，核心问题是 AI 与人类的差别究竟何在？这样的问题意识似曾相识，传统上哲学一直在探讨人与动物的差别几何？正如孟子所说“人之所以异于禽兽者几希”？不同的哲学传统会给出不同的回答，人是理性动物，人是语言动物，人是政治动物，人的良知良能，等等。现在的问题是，人类具有而 AI 并不具有的特质究竟为何？由此才能为人类与 AI 划界。但这样的“认知模式”忽视了一个重大差异，人类物种与动物的差别在生物亿万年的漫长岁月中，是相对静态的，因此可以用某一标准来划界；而 AI 技术则以肉眼可见的速度迅猛发展，日新月异，颠覆了哲学家给出的所有“界限”。人机“划界”的探讨充分暴露出以往哲学探讨的局限性，“划界”标准不断被技术突破，哲学家们提出的标准不断退守，例如，“直觉性决策”被 AlphaGo[③] 突破，“艺术创造”被 DALL·E[④]、Midjourney[⑤]、Suno AI[⑥] 等突破。我们把这一时期的“AI 哲学”称为“AI 哲学Ⅰ”，一种消极型的 AI 哲学，其主旨还停留在试图建立起人类与 AI 之间的“本体论隔离”。如果说，人与动物的区别是在彰显人的主体性，那么人类与 AI 之间的划界讨论则显示出一种保守人类文明主体性的防御性机制。虽然传统与现在的心态不同，本质上的“认知范式”却是一致的，都是某种人类中心主义的体现。问题在于，这样的做法是将人类的有限特质作为普遍性标准去衡量所有智能，对智能未知的迅猛发展没有留出足够的想象空间。如此，当 AI 成长到其所具有的复杂性完全超越人类的时候，即其作为更混沌更复杂的系统出现时，人类视角下的哲学探讨就会开始瓦解，哲学的框架被突破，哲学的可信度被质疑。回顾伴随 AI 的哲学探讨，给予我们的启示是，人类与其以防御心态去区隔人机关系，何妨以更积极姿态去接纳与自身同阶甚至超越自身的智能存在，并努力学会与之共同前进。

二、AI 哲学Ⅱ：哲学助推 AI 发展

正如我们看到，以往“AI 哲学Ⅰ”的困境源自西方哲学传统自身“认知范式”的局限性，以及在此背后的人类中心主义预设。在认清这一困境本质之后，“认知范式”上的转换可以推动哲学与 AI 进入更深入、更具创发性的模式。在“智能”与“主体”多重且异质的前提下，哲学不仅服务于人类主体，更有面向 AI 主体的可能性。由此，在 AI 问题上，哲学的“认知范式”就需要摆脱哲学滞后性的被动反思模式，更积极地主动介入 AI 系统的设计、训练与部署过程。哲学不仅是技术的被动解释者，还需前瞻性地为 AI 演化提供必需的概念框架与思维范式。我们将助力 AI 发展的哲学称为“AI 哲学Ⅱ”，一种积极型的 AI 哲学。

---

① Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D., “Language models are few-shot learners,” *Advances in neural information processing systems*, 33, 2020, pp. 1877-1901.

② Chalmers, D., *GPT-3 and General Intelligence*, 2020, https: //click. convertkit-mail. com/n4ulgo3kk2uvh85ol5vi6/z2hghnhol3d6w8iz/aHR0cHM6Ly9kYWlseW5vdXMuY29tLzIwMjAvMDcvMzAvcGhpbG9zb3BoZXJzLWdwdC0zLyNjaGFsbWVycw==.

③ Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D., “Mastering the game of Go with deep neural networks and tree search,” *Nature*, 529(7587), 2016, pp. 484-489.

④ Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I., “Zero-shot text-to-image generation,” in *International conference on machine learning*, Pmlr, 2021, pp. 8821-8831.

⑤ https: //www. midjourney. com/.

⑥ https: //sunoai. ai/.

这里，我们借用一份据称是OpenAI给出的衡量AI能力的层级标准（见表1）[①]，来说明“哲学”在AGI演进中所能起到的重要作用。技术的阶梯攀升绝非单纯的工程问题，每一个层级的跃迁背后都有深刻的哲学塑造——哲学在AGI演进中提供着思维进化不可或缺的范式与框架，哲学的推理水平、反思能力、语境辨析以及真正的创新能力，这些都是评估AGI能力的重要指标。

表1　OpenAI AGI能力分级

层级	名称	核心能力
1	对话者（Chatbot）	自然语言交互
2	推理者（Reasoner）	人类水平逻辑推理
3	行动者（Agent）	现实情境行动决策
4	创新者（Innovator）	引导领域创新
5	组织者（Organizer）	社会系统级运作

2023年，依赖统计模式生成对话的GPT-3.5模型可以说是正式达到了Level 1（ChatBot）的水平，能够支持与人类的自然语言交互，对各种问题给出某种看似合理的回应。但对语言背后的真实含义、逻辑关联或上下文深层意图的理解还非常有限，因此无法维持长对话的逻辑一致性，也没有能力推断事物间的因果机制。这里存在的问题在于，当时的大语言模型还缺乏推理思维，对语言只是给出被动的、直觉性的回应，因此还无法达到Level 2（Reasoner）的水平。Level 1与Level 2的区别，可类比于卡尼曼《思考·快与慢》[②]中的系统1与系统2的差别：前者体现快速的、直觉式的模式匹配与联想；后者则要求缓慢的、符合规则的结构化、符号化推理能力。

随后，大语言模型领域的研究开始集中于思维链探索，试图在后训练阶段与提示词工程上教会AI正确的推理步骤。在这一过程中，分析哲学与逻辑学起到了非常重要的作用[③]，比如将命题逻辑、谓词逻辑等符号系统嵌入思维链之中，使模型输出符合形式有效性的推理步骤[④]，以及提供形式谬误的检测模板，在微调中不断修正模型的不合规推理路径[⑤]。

目前的AI可以说已经接近了Level 2的水平（如GPT-o3、DeepSeek-R1等在逻辑任务上表现出的能力），并正向Level 3行动者（Agent）的级别推进，该方向已成为近期的技术热点。Agent相比Reasoner的最大困难在于：AI需进一步在开放、动态变化的现实情境中来解决实际问题。如此，原来局限于思维世界的“概念游戏”式的推理显然不够了。Agent所需“实践智慧”的要求已经远远超越离散推理，而是要在动态现实中构建“行动—反馈”的闭环，这正是当前AI的薄弱环节。

在Level 3 Agent及后续阶段，AI必须掌握真正的思辨能力，这不是简单的应答能力与逻辑推理能力，而是思维必须能真正应对“变化”，因此我们也经常把哲学能力称为“思辨能力”，也就是它得容纳实践中产生的种种矛盾，从而将认识与实践中的冲突视为发展过程中的内在要素，而非单纯的形式错误，这已大大超出了逻辑学形式化的要求，却是哲学必备的能力。目前的AI尚未具备这种能力。有研究指出，当对AI输入出

① 需要指出的是，该分级为OpenAI在内部会议提出，但并未在公开报告中声明过这个分级。具体可以参考：https://generationaishow.com/episodes/the-5-levels-of-ai-from-chatbots-to-organizational-intelligence。

② Kahneman, D., *Thinking, fast and slow*, Farrar, Straus and Giroux, 2011.

③ Bentzen, B., Liao, B., Liga, D., Markovich, R., Wei, B., Xiong, M., & Xu, T. (eds.), *Logics for AI and Law: Joint Proceedings of the Third International Workshop on Logics for New-Generation Artificial Intelligence and the International Workshop on Logic, AI and Law, September 8—9 and 11—12*, 2023, Hangzhou.

④ Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu, “LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 2124-2155.

⑤ Chan, J., Gaizauskas, R., & Zhao, Z., “RULEBREAKERS: Challenging LLMs at the Crossroads between Formal Logic and Human-like Reasoning,” arXiv preprint arXiv: 2410.16502, 2025, https://doi.org/10.48550/arXiv.2410.16502

现上下文矛盾的文本时，现有的各大模型输出都会产生严重的混乱。[①] 因此，要推动 AI 从 Level 2 向 Level 3 及其之后的阶段演进，AI 就必须具备真正的辩证思维能力，也就是要求 AI 能够动态地适应环境流变，而非仅依据静态的先验知识；要求 AI 能够在冲突中寻求协调，在变化中把握相对稳定的规律。这种对思辨能力的迫切需求，彰显了“哲学”之于技术方案的核心价值。

哲学在 AI 迈向 AGI 的路程上，应该会扮演更为重要的角色，不仅 AI 的能力层级需要哲学的评估，哲学能力在人类智能的类型中也具有某种独特性，作为通用人工智能（AGI）必须有所体现。举例来说，我们都知道在现有的生成式 AI 的模板上都有一个控制温度（Temperature）参数的机制，简单来说就是一种控制 AI 生成文本的严谨性与创造性之间程度的参数，往偏低方向调整，就是放大高概率词的权重，像逻辑学家或数学家那么严谨；往另一个方向，就是往偏高方向调整，就是让概率的分布更加平坦，让 AI 更像艺术家那样富有想象力。不同的能力可以在一个线性的刻度上调整。但是，就哲学家而言，当他进行严厉的批判性工作或者理论建构时，他往往需要像艺术家那样具有想象力，对于任何框架的边界具有敏锐性，敢于突破，敢于批判，这就需要 AI 上的高温度；当他就此进行细致论证时就得需要像逻辑学家那样严谨，对于任何观点都要有严格的理性论证或推理，这就需要 AI 上的低温度。无论是高温度还是低温度都是 AI 所必需的，也是 AI 针对不同的要求各自能够完成的。但目前 AI 上的“温度”的调试却是在线性刻度上完成的，不能同时向两个相反方向来调试，很难在“温度”参数的调节上同时完成这两个方面的要求。起码就哲学工作而言，这两方面的要求都是十分必要的，所以对于 AGI 的模型来说，这样的双重能力如何能够在同一项工作中同时具备？这依然是一个挑战。因此如何来设计新的 AGI 模型，必须要有哲学的前瞻性和预备性的工作。

总之，哲学代表了人类智能的某种类型，哲学在 AI 发展中的作用绝非停留在解释性或反思性的层面，在 AGI 的演进中，哲学通过为其提供某种超越单纯技术性可穷尽的“认知范式”，成为技术迈向高阶能力的一个重要指向。这就不是哲学对技术的反思，而是哲学—技术的一种互动演化，这构成了通往 AGI 的必由之路。

三、哲学 AI：AI 推动哲学研究

“AI 哲学”总体而言，以哲学作为主体（背后是仅以人作为主体），以 AI 作为修饰语，以哲学的方式凌驾 AI；但“哲学与 AI”的关系绝不局限于此，我们可以进一步摸索哲学与 AI“双重主体”进行互动的理论可能性以及现实展开的空间。基于新的“认知范式”，我们将 AI 推动哲学研究的实践，称为“哲学 AI”。区别于“AI 哲学”，“哲学 AI”将 AI 作为至少与人类同阶的“智能主体”来看待，研究并实践“哲学与 AI”的对话与互动。这种“对话与互动”可以展开得非常复杂，有着非常广阔的探讨空间。

在未来能够容纳“多层”智能体互动的 AI 时代，我们非常肯定地预言，哲学也将发展出新的样态，一如 AlphaGo Zero 在与李世石那场著名五番棋第二局中下出石破天惊的“第 37 步棋”。因此不单是哲学反思与助推 AI 的发展，AI 也将推动哲学研究，一种真正的“双向互动”。因此，AI 时代新样态的“哲学”将对人类提出新要求：AI 应成为一种思维方式与方法，成为某种“主体”来推动哲学的研究，而非仅仅作为一种等待人们进行哲学反思与研究的“对象”。

我们必须认识到，AI 的思维方式与现有的哲学方式有着非常不同的面向，基于 AI 的认知模型呈现出显著的概率性、非确定性特征。其与人类传统思维的差别类似于量子力学多世界理论与经典力学的差异。从柏拉图学园门口“不懂几何学者不得入内”的标牌开始，西方哲学就已把某种确定性作为自己的追求目标，笛卡尔的“我思故我在”也是意识层面上来重新确立这个确定性。一如近代的经典力学正以这样的“确定性”为自身基础，但是现代量子力学的多世界诠释是以“概率”作为世界基石的。“概率”既非经典统计模型的定义，也非贝叶斯模型的主观断言，而是作为“非确定性”本身而存在（换言之，不是量子世界的“概率”需要“确定”的解释，反而是宏观世界的“确定性”需要量子的解释）。“概率”“多世界”等说法都是人类

① Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu, “Knowledge Conflicts for LLMs: A Survey,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 8541–8565.

语言为描述“概率本体”所做出的尝试。由此，AI 的思维方式将会迫使西方哲学的传统去努力学会与混沌、与未知相处，让自身能够安顿于多种可能性之中。未来的哲学需要更善于掌控已所未知之物，而非沉溺于确定性之中。在这个意义上，中国哲学传统的思维方式如果转型得当的话，将会在 AI 时代大放异彩。

那么究竟应该如何构想以 AI 方式来推进哲学研究呢？在此之前，我们不妨先回顾一下人类的哲学研究方式。

自轴心时代以来，在过去两千多年时间里，哲学思辨的核心是围绕辩论来展开的——辩论作为内在结构驱动着思想的动态演进。西方哲学史中，从古希腊的诘问式对话到中世纪经院的辩证法，再到现代的逻辑分析，辩论始终作为哲学理论生成的内在动力，推动着思想从现象到理念的攀升；中国哲学传统同样有一种基于对话的思想力量，在孔孟老庄、宋明理学等关键思想的突破中，彰显出论辩的创生力量。辩论因而成为哲学区别于其他学科的重要标志：它以动态进化而非凝滞的思想方式来回应人类生存所面临的重大关切。

AI 的介入使得哲学辩论本身获得了“概率化”的重构潜力。经典辩论在不同程度上都依赖于非真即假的排中律，AI 式的概率思维方式类似于量子叠加态，这完全不同于西方哲学传统中确定的排中律。换言之，AI 的语言大模型有同时维持多个矛盾立场的可能性，这将极大地拓展辩论边界，重现古典哲学辩论的场景。AI 能够基于概率模型同时穷举多种矛盾的论证路径，将单一对话拓展为叠加不同可能性的论辩概率图谱，从而更进一步地揭示思想网络中的隐含联系，这恰是哲学应该去追求的。另一方面，传统辩论在不同程度上都指向一个所谓的“正确”裁决，辩证法也会指向一个综合正反两方面意见的“合”的结论，AI 引导的哲学辩论更像是一种概率性的探索——不在于找出“正确”判断，而在于通过模拟海量的思想互动，概率式地呈现出各种核心理念在不同情境与约束条件下可能“涌现”的共识与分歧。如此，哲学辩论的丰富性与多元性将在 AI 的助力下得到更进一步的彰显。在这种新式的哲学范式之下，人类哲学思想的工作将要求更聚焦于抓取超出常规理论的最核心部分，理论的完善性甚至其丰富态都可以交由外置 AI 的理性能力来处理。AI 引导下的哲学工作对人的哲学创造力提出了更高的要求。

此次，Microsoft AI Asia 与北京大学博古睿研究中心合作，启动了“哲学家 AI 助手”的研发工作，希望能将上述的哲学思考在技术层面上加以实现，使“哲学 AI”的可能性得以真正地实现出来。在哲学专业的层面，针对性地进行 LLM 所需要的预训练，形成哲学类别的多模态知识库。在基于各种概率性的前提下，通过 AI 摸索哲学论辩的多种可能性，多层次地构建起哲学家的 AI Agent。并且，通过不断地与各种 agent 进行各种论题的辩论，推进哲学的研究与探索。一种哲学家与 AI 的互动，将会成为哲学工作的新样态。虽然这只是初步的工作，但这样的研究方向，确实让哲学家们充满期待。

四、哲学 AI 的初步塑造——从孙向晨与某哲学家 AI Agent 论辩的实例谈起

在构建“哲学 AI”的方面，目前在世界范围内的推进都非常初步，且十分有限，甚至可以说还处于某种空白阶段。就目前初步工作而言，“哲学家 AI 助手”是一种面向哲学研究者和学习者的智能对话工具，旨在通过与“哲学家 AI 助手”的讨论与辩论，深化对哲学理论的理解，帮助研究者拓宽自身的研究视野，最重要的是推动“哲学 AI”工作的进展，使 AI 有可能介入到人类的哲学工作中去，在最基本的层面上创造出哲学研究的新工作形态。

这项实验性工作是以“家哲学”为例展开的。具体是通过当代美国某著名哲学家的 AI Agent 与孙向晨的 AI Agent，就“双方”感兴趣的“家哲学”论题展开对话与论辩，使上述的宏观思考在一个非常微观的层面上得以实施，并检验 AGI 分级能力的运用情况。这位著名学者是美国当代的一位哲学家，在女性主义批评、性别研究、当代政治哲学与伦理学等领域做出过杰出贡献，具有强烈的批判性立场。笔者提出的“家哲学”与这位哲学家的论域有很多交叉重叠之处。这次实验也是在世界范围内首次为当代哲学家制作 AI Agent。在此实验中，AI Agent 分别以两位哲学家现有著作与文章为数据库，提炼其中核心观点，并以各自观点出发，就他们共同感兴趣的“家庭与性别”等问题进行论辩。在现实中，他们的立场差距很大，他们之间也没有进行过直接论辩与交锋，更没有就对方问题有过针对性评述，因此论辩的内容完全是靠 AI Agent 自动生成的。这样的哲学讨论，没有现成答案，也没有设定明确目标，模拟一种完全自然的论辩状态，并保持开放的结果：

看看两个“哲学家 AI Agent”在“自主”状况下何以展开自由的论辩，得出怎样的学术结论。做当代哲学家 AI Agent 的优势在于，其交锋论辩成果可以交由哲学家本人来进行检验，对其能力、过程与结果进行细致分析。

这里给出的案例是在“个体与家庭”这个主题下进行的，AI Agent 之间的交锋采取了一般的论辩形式：有主持人和辩论双方，首先是介绍各自立场，随后是相互质疑与回应，再进行相互提问与回答，最后是主持人归纳总结。这样的 AI Agent 之间的论辩实验进行了多轮的尝试，在学术顾问的指导下，在工程技术人员的帮助下，AI Agent 的工作不断得到修正，形成了多个不同的版本。

在最早期版本中，AI Agent 显然对于哲学家间的论辩没有明确的概念，在各自数据库中提炼的观点比较单一，对于学术概念本身缺乏敏感，对于理论观点作为一种结构性的表达比较生疏，论辩的进行常常是一竿子到底，基本观点车轱辘来回倒。可以说，只具有一般的论辩形式，在内容上并没有很多可取之处。通过与技术人员的交流与协调，强调哲学论辩的特点，由此对所用 AI 大语言模型的参数进行了各种调整，在之后输出的版本中，有了比较显著的进展。在这一较早版本中，提取的观点依然比较谨慎，这位与孙向晨 AI Agent 对话的哲学家 AI Agent 会强调家庭的非自然性及表演性的特质，认为现代个体理论背后忽略了生命政治的作用，以及基于酷儿理论发展出的某种针对“亲亲”的理论，孙向晨则站在“双重本体”立场上，质疑原子化的个体，同时激活“家”作为原初伦理经验的哲学潜能，并对某哲学家的“表演性”理论提出质疑。总体而言，此时提取的观点比较宏观，相互论辩有了某种针对性，但还比较直接与线性，忽视了哲学观点得出的复杂性及相互碰撞的多样性。从 AGI 的分级能力来看，这一阶段的版本，有着很好的自然语言互动能力，推理上也没有太大问题，但论题的展开比较单一。

随着 AI 对于哲学辩论的深刻理解，在之后的版本中，对于两位哲学家的概括 AI Agent 做得更为立体，辩论展开的层次感也更为丰富。那位美国哲学家的立场被概括为一种规范性剧场式的家庭，强调后家庭的伦理以及关系性主体与非暴力政治，孙向晨的立场则被概括为清醒地意识到个体自由及其意义塌陷的困境，个体与亲亲的双重本体进行了更细致的互构，以及提出了一些新的观点来应对现代性挑战。在此轮版本中，观点更加细致，提炼的观点越来越呈现出多面立体形态，不同的观点之间建立起更紧密的相关性，随后的质疑与提问，各自的针对性都得到了加强。从 AGI 的分级能力上看，这是有了一定的在对话情景下做出更丰富回应的能力。

在最后一个版本中，借助目前最新模型，两个 AI Agent 相互辩驳的层次感更加丰富，能够在多个层面上推进论述，论辩水平有显著提升。在各自观点的总结上，较以前更为丰富。以下是对两个“哲学家 AI Agent”之间最终成稿的论辩记录所做的分析。那位哲学家的 AI Agent 对自身的观点归纳了 5 条，非但包括之前观点，还提出了更丰富的想法：坚持家庭为权力表演场域，而非自然实体；扩大可识别性，超越传统的亲属定义；以平等可悼性重写伦理基础；将家庭视为公共政治议题而不是私人考验；建议生产与亲属关系的动态重构。孙向晨的 AI Agent 也对“自己”进行了新的概括：分别是以“个体—亲亲”理论破除家庭与个体的对立，以此回应现代性危机；强调家庭是自由的原初伦理空间，非契约性责任塑造了真实的自由；以代际非对称性伦理关系作为理解家问题的关键，同时强调对未来的责任；超越了单纯的个人主义与社群主义；提出家—社—政结构重绘伦理与法的界限，强调伦理优先性；在与那位哲学家探讨家庭与性别问题时，主张亲亲的本体论地位；并提出三重原则：责任化个体、开放家庭、可反思亲属伦理；最后视“归家”为救赎之道，强调家庭与个体的动态平衡，提出家庭与个体张力的新规范政治重构。在这里尤其值得一提的是，AI Agent 提出“一种有根的自由”（a rooted freedom）的概念，这超出了笔者本人的预期。这是 AI Agent 在网络上借鉴过来的概念[①]，但与笔者理论完全是融贯的。当然，这只是一个辩论形式，任何新观点的提出，在学术意义上还需要哲学家本人进行更深入的论述与论证。

最后的版本对于各自观点的归纳明显更丰富，更立体，在逻辑上也更完善；每一个观点的论述相对完整，

① “Rooted freedom”是一个在日常写作中曾经出现过的说法，但并不是一个哲学家曾严格讨论过的哲学概念，AI Agent 在此将其作为一个哲学概念借鉴过来是完全恰当的，萨特曾说过“a condemned freedom”，列维纳斯说过“a invested freedom”。

注重观点的出处与依据。对于哲学家AI助手论辩质量的评价，经过几轮改进，应该说最后版本的论辩还是有一定质量的。就笔者的AI Agent而言，这个AI Agent对于笔者哲学思路的概括基本是正确的，但可以形成某种不一样的概括形态；对于笔者不同文章之间的观点也能够建立起有效关联，在某种意义上有其自身的思路；AI Agent还可以根据论辩对手的思想，形成一些包含对手“说法”的结论；根据笔者的思想线索调用了一些新的思想资源，并给出了一些笔者并没有直接提出的说法和概念，却依然能够保持在笔者的思路范围之内。关于家与个体自由的问题，笔者曾有过一个草稿，没有发表过，不在这个AI Agent的数据库中。AI Agent给出的版本跟笔者的原有思考是很不一样的，提出了一些新的说法和概念，但其说法原则上都是可以成立的。AI智能所具有的“概率化”的多样可能性得到了初步体现。总体表现中规中矩，当然你可以说这是“他”谨慎的表现，也可以说是能力尚有不足的表现。其中若干观点的提出可以看作原作者观点的衍生，而不是一种直接应用，这究竟是因为论辩主题的重新设定以及因涉及论辩对手立场而提出新观点了呢，还是仅为AI的某种“幻觉”而已？这也很难辨别。只要这种“幻觉”在逻辑上立得住，就不失为对于哲学进展的某种启示。

笔者从来没有写过关于这位哲学家的评论文章，对于这位哲学家的哲学虽有所了解，但并没有到非常熟悉的地步，AI Agent替笔者抓住了回应的要害。AI Agent所给出的归纳与反驳意见，基本上可以得到笔者认可。比如以下这段文字：“我与这位哲学家的对话将落点放在‘亲属—性别—规范’的张力上。这位哲学家关于性别和亲属表演性的分析凸显：家并非自然化实体，它在规范、承认与排除中被持续生产。对此我部分赞同：家庭确有‘建构’的维度，传统亦须反思与重述。然而我同时坚持‘厚的本体论’：亲人不仅是话语秩序，更是具身性的生成与相互负责的场域。将亲人完全消解为表演，容易忽略身体与代际之‘给出性’。因此，我主张把表演性纳入‘生生—亲亲’的本体框架之中：一方面承认家庭形式的历史可塑性与多样性，另一方面保留其作为伦理源初生成地的独特地位。”① 这完全是AI Agent生成的文字，却是笔者完全可以接受的立场。尽管笔者本人如果做同样工作，可能会有不一样的进路。

这是AI作为某种智能主体与人类之作为主体，已经具有“双主体”的形态，在这个意义上，“他”已经超越了单纯“工具”范畴，而类似于一个与之共同工作的“朋友”。虽然目前AI Agent的工作尚没有做到惊艳地步，但其“振振有词”依然让人印象深刻。它并不是简单地以“针尖对麦芒”的方式来反对那位哲学家思想，而是在笔者理论框架中，既对那位哲学家理论进行批评，同时又努力把她的某些观点镶嵌入笔者的框架中，并给予某种限制；为新理论预留空间的同时，也坚持了自身的理论底线，这恰恰像是一个“朋友”立场的工作。通过这个AI Agent在哲学工作方面的小小实验，已经显示出AI Agent所能做的哲学工作远要比想象的更为复杂。最初版本经过各种调试之后，也可以进行更复杂的论辩。这表明现有的先进的AI大模型，经过一定程度的专门哲学训练之后，是可以参与到哲学理论的研究中去的。

五、“哲学与AI共生形态”的探索

“哲学家AI助手”的实验表明，除了“AI哲学Ⅰ”对AI的哲学反思，“AI哲学Ⅱ”对于AI能力升级的前瞻性思考外，我们完全可以在“哲学AI”的层面把“哲学与AI共生”推进得更远，进一步推动AI与人类哲学思考的互动。但是，这一实验也带来许多进一步的思考：AI Agent可以给出关于你思想的不同概括版本，究竟是否算对你思想的“正确”概括？基于你的观点所提出的新想法，究竟是“你”的，还是“他”的？是“你”的，因为这是根据你的观点而来；是“他”的，因为这衍生的观点与你对同样问题所作的思考并不一样。一种新的、有待阐释的人机交互关系，或者说一种智能体与人类“双重主体”的格局“正在”诞生。就现实问题而言，基于某学者已有思想的AI Agent所作的讨论究竟是由谁来负责呢？所产生的影响究竟算是“谁”的影响呢？如果按“知识产权”的思路，这就是谁的知识产权？还会有许许多多的问题，都将是一系列新的挑战。哲学作为某种智慧形态是人类文明的反思结晶，AI的发展其实牵扯出的是一种人类与技术之间的极致关系。人类的生存与技术之间的关系显然进入了一种全新形态。当代AI技术革命引发文明转型危机时，可能是现代智人（Homo Sapiens）所面临最严峻的技术与文明之间的对立。在这个意义上，“共生”可能

① 根据AI Agent生成的文本。

是唯一出路。

由于哲学学科自身特点，哲学与 AI 的“共生”问题，需在人类与技术之间的关系这个更大的框架下来进行思考。笔者将人类与技术之间的关系设定为三大形态：最初是技术作为人类的工具。当技术作为工具时，主导者就是人类。长期以来，我们会受限于一种技术决定论思维，总以技术来标识文明的不同阶段，青铜时代、黑铁时代、蒸汽机时代，乃至于未来的 AI 时代，若只按技术决定论来理解人类文明的话，那么就会出现一个巨大困惑。雅斯贝尔斯在《论历史的起源与目标》中指出，人类在“轴心文明”之前还有“古代历史中的高度文化”或称“古历史文明”，如苏美尔、巴比伦、埃及、赫梯等古代文明。这些古代文明都曾创造出高度辉煌的文化，有着巨大的“技术突破”，比如埃及的金字塔和灌溉系统、巴比伦的天文学、苏美尔的冶金术等，但这些古文明并没有演进到“轴心文明”阶段，而是相继在历史长河中以各种方式消亡了。若按技术的先进程度而言，“轴心文明”的技术甚至还不如他们，胡夫金字塔在其建成之后的 3800 多年时间里都是人类最高建造物。因此“轴心文明”之为一种持久文明，必定基于其他关键要素。雅斯贝尔斯认为，轴心文明的出现标志着人类从神话思维向理性思维的转变，从地方性文化向普遍性文化过渡，实现了某种“精神突破”。[①] 人类早期在面对生存困境和社会变革时，通过对存在、道德、宇宙等根本性问题进行反思，在各自的文明中都形成了某种认知与价值体系以维系人类的生存。在这个意义上，是“精神突破”而不仅是“技术突破”才建构了“轴心文明”。由此，“技术作为工具”才被人类文明所牢牢地掌控着，而不至于像“古历史文明”那样，因技术仅仅作为暴力统治手段却因缺失人类健全的“情—理结构”而趋向崩溃。[②] 从这个角度看，在轴心文明框架内，并不是技术决定了人类文明，而是人类文明掌控着技术。

第二阶段，现代技术作为一种自足体系，它虽驱动着人类发展，但对人类来说也发生了相应异化，海德格尔将技术本质称为“架座”（Gestell）。[③] 海德格尔认为，现代技术绝不仅仅是人类使用的中立工具，其本质远超出了工具或手段的范畴，技术一方面强制地向自然“逼迫”与“索取”，另一方面反过来也制约了人类的生存。文明的逻辑是在寻找人类生存的平衡态，传统技术在不同的文明形态中，在各自的“情—理结构”中安置自身。但是，现代技术与文明的逻辑是迥异的，“工业革命”之后的现代技术形成了自己的逻辑，它要求一种强求性的“解蔽”方式，事物被迫放弃其自在状态，成为技术系统内的“持存物”（Bestand），失去独立性和丰富性，成为被消耗与订购的库存资源。工业体系剥削自然，计算思维将万物简化为可预测、可利用的资源，以“资源化”的眼光看待世界，不断剥削地球，遗忘了世界原本可以诗意地呈现。在此过程中，人表面上继续主宰着技术，实则被技术逻辑所支配；不是人使用技术，而是技术“促逼”人服务于其自我实现与扩张的目标。技术的自我驱动是无沉思的、单向度的，它不断要求更多资源、更高效率、更强支配，与人类文明强调的平衡、多元和稳健背道而驰。文明追求在“情—理结构”中建立了可持续的生存样态，而现代技术逻辑则一直在瓦解这种稳态，将人与世界统统纳入其运转的框架之中，人类从“主体”跌落为机器中的“螺丝钉”。海德格尔对技术的深刻反思，揭示了现代技术对于人类文明的深刻挑战。他批判西方形而上学传统将人确立为绝对主体，使世界沦为可被无限征服的客体，而技术在追求成为主体意志最高体现的同时，也内在地成为一种异化的力量。

最后，在 AI 时代，智能技术不单单是一种自在的体系，而开始呈现出某种“主体”的样态，未来“他”甚至可以摆脱人类而不是依靠人类，实现自我演进与迭代。更值得注意的是 AI 的“黑箱”特性——其决策过程往往不可解释，人类即便在表面上也将逐渐丧失对技术的理解与控制权。正如 AlphaGo 超出人类理解的落子第 37 步，这种不可控性已经超出了海德格尔所说“架座”的一般风险，AI 作为智能技术不仅超越了“工具”的概念，也超越了“架座”的概念，而是更迅猛地构成了一种智能“主体”的样态。当现代技术直接以智能这种更加纯粹的形态出现时，其迭代的速度更是无与伦比，AGI 的发展速度可能在数十年之内全面颠覆人类社会的文明基础，这是我们特别要注意的。如果说海德格尔担忧技术对自然和人的控制，那么 AI 发

① 雅斯贝尔斯：《论历史的起源与目标》，上海：华东师范大学出版社，2018 年，第 8—9 页。

② 参见孙向晨：《文明的逻辑与技术、智能的逻辑》，《广州大学学报（社会科学版）》2023 年第 4 期。

③ M. Heidegger, *Basic Writings*, Harper San Franciso, 1977, pp. 311–312.

展的威胁就更进一步了：它可能直接破坏人类的社会结构，使我们全面从属于技术的自我演进。但我们要问的是，AI作为某种“主体”是否也可以成为人类的“朋友”呢？

人类的文明自诞生以来，都是在同质智能主体的互动下发展起来的。尽管文明体系不同，但我们都是现代智人的后代，有着相似的情商与智商，都是与相同智能同类在打交道，但人类早期有着与超人类智能打交道的想象经验。在某种意义上，“人类与AI共生的状态”有可能回复到某种形式的人类早期型“认知范式”。在人类早期，人类欣然接受将人接纳为整个世界的重要构成部分，人作为“世界多元中的一元”，如中国哲学中天地人三才，或者西方哲学中所说的人神关系，其决定与行为需与其他决定性因素相协调（如殷商占卜决策或希腊神谕咨询），方能做出决定。在这个意义上，当面对具有超越人类智能且能够自主行为的AI时，或许古人比被启蒙理性规训的现代人更容易接受其存在，更容易接受其“主体”地位。AI的重大发展让人类再次面临“认知范式”的重大转换，这将带来人类生存的一系列根本变化。需强调的是，回归某种“多元智能”共生的处境，绝非简单地放弃人类的主体性，而是在认清人类认知能力的边界后进行清醒的重构。当抛弃人类“绝对主宰”的幻觉之后，人类反而能在“人类与AI共生”的形态中释放出更大潜能。这种反思将会塑造出一种全新的文明形态：人类与AI共建的新形态将会是一种真正的多元体系，传统哲学在天人之际的不同角色之间寻找平衡，而未来的哲学与AI共生的工作形态将探究碳基与硅基在动态平衡中的各司其职。人类在“喂养”人工智能时，其所面临的巨大风险，是谁也不敢否认的，AI的发展远远不是“对齐”所能约束的。[①] 在此意义上，需探索未来哲学与AI互动的新模式，构建技术迭代与文明传承的一种可能的和解路径。

AI的发展表明，人类文明不再只是用价值体系去牵制作为“工具”的技术发展，不再只是抵御技术“架座”的异化，未来文明更要学会与作为“智能主体”AI的“共生”。人类与AI“共生”的道路或许可以从哲学与AI的协同工作开始。这一方式或许可以作为构建人类与AI共生形态的最初步实验，并逐步将这种人类与AI的互动模式，推扩至文明的其他层面，诸如教育的革新、伦理体系的重构、社会治理的转型等。人类与AI作为主体也由此得以“共生演化”，构建多层次多智能主体互动的新文明。

（责任编辑：盛丹艳）

Philosophical AI: Exploring a Co-working Paradigm between Philosophy and AI

SUN Xiangchen, JIA Zihan, ZHANG Qi

Abstract: With the rapid advancement of Artificial Intelligence (AI), the “philosophy of AI”has sparked a series of intense academic discussions. However, current debates often exhibit a reactive or lagging character, primarily centering on the demarcation between humans and machines. The concept of “philosophical AI” emphasizes the bidirectional shaping between philosophy and AI. It highlights both the indispensable role of philosophy in the evolution of Artificial General Intelligence (AGI) and the potential for AI-driven modes of thinking to facilitate the development of philosophy itself. Utilizing the “Philosopher’s AI Assistant” platform, this study presents a virtual debate between Prof. Sun Xiangchen and an AI Agent representing a renowned contemporary philosopher as a case study. This example concretely demonstrates the potential developmental space and inherent challenges of philosophical AI, ultimately pointing toward a stance of “co-working between philosophy and AI”.

Key words: philosophy of AI, philosophical AI, philosopher AI agent, co-working of philosophy and AI

① Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014.