

超级智能的四种解释

杨庆峰

(复旦大学 科技伦理与人类未来研究院, 上海 200433)

摘 要: 在对超级智能的诸多理解中,比较解释是最为常见且熟知的一种,这种解释符合日常认知,但是存在直觉谬误;实体解释是与人工智能科学发展最为接近的一种,科学界将有意识的、第一人称体验、机器经验和自进化等作为关键问题,进一步接近了机器的自主性,但容易陷入技术的形而上学解释;功能解释是最具哲学性的一种,它将人们置于存在论问题面前,其揭示了深层被忽略的问题;还有一种是叙事解释,这种解释将超级智能看作是重构人类文明的一种动力,其意义远未被揭示出来。

关 键 词: 超级智能;比较解释;实体解释;功能解释;叙事解释

中图分类号: N 031 **文献标志码:** A **文章编号:** 1008-3758(2026)05-0023-08

Four Explanations of Artificial Superintelligence

YANG Qingfeng

(Institute of Technology Ethics for Human Future, Fudan University, Shanghai 200433, China)

Abstract: Among the various understandings of artificial superintelligence (ASI), the comparative explanation is the most common and familiar one. This explanation aligns with everyday cognition and is grounded in perceptual foundations; however, it falls into the fallacy of intuition. The substantive explanation is the one closest to the current development of artificial intelligence science. The scientific community regards self-awareness, first-person experience, machine experience, self-evolution, and similar concepts as key issues, thereby approaching the autonomy of machines more closely; however, it is likely to lapse into the metaphysical explanations of technology. The functional explanation is the most philosophical one in that it places people squarely before ontological problems and reveals deep, overlooked issues. Last, the narrative explanation views ASI as a driving force for reconstructing human civilization, with its significance yet to be fully revealed.

Key words: artificial superintelligence (ASI); comparative explanation; substantive explanation; functional explanation; narrative explanation

在以往的研究中,笔者对人工智能叙事的类型及意义进行了研究,揭示了人工智能叙事的四种类型,即科学化叙事、科幻化叙事、诗意化叙事和媒介化叙事,而这四种类型共同指向超级智能

(super artificial intelligence, 简称 ASI)这样一个对象^①。这就导致了对于 ASI 的解释形式的出现:超级智能的叙事解释。但这一问题却被忽略,笔者以往的研究只关注比较解释、实体解释

收稿日期: 2026-03-20
基金项目: 国家社会科学基金重大项目(25&ZD021)。
作者简介: 杨庆峰,复旦大学教授,博士生导师。

① 2025年笔者对AI叙事的思考还是人类中心主义的,即“谁”在讲述AI故事,并且划分出不同的主体类型,如科学家的AI叙事、哲学家的AI叙事、科幻小说家的AI叙事以及文学家的叙事。不同的主体对应着不同的叙事,而且叙事的方式主要是数据、语言和图像,AI叙事的目的在于AI与人类生活的融合,也就是说,人类如何更好地使用AI。这种叙事中“工具性”成为基础。但是在后来的思考中,发现AI作为叙事主体存在的可能性,这意味着AI摆脱人类中心主义的限制有了自身的叙事,强调AI与人类的共在和融合。这种叙事中“包容性”成为基础。

和功能解释,并没有注意到这种处在边缘的解释形式。而叙事解释的意义不同于其他解释形式,因为一般的人工智能叙事有助于显示“人工智能技术如何融入我们生活的故事中”^[1]。此处的“我们”专指人类。人类通常把人工智能看做是工具,而工具是要被牢牢控制在手中,服务于生活世界。然而ASI存在的意义早已超出工具范畴,这意味着倘若人工智能有自身叙事的话,也并非要融入人类的生活故事,而是有着不同的融入维度^①。因此,对超级智能的解释进行阐述有助于我们理解超级智能这一现象;而澄清四种解释类型之间的关系对于构建“面向哲学问题的超级智能”(ASI4P)就显得非常重要了。

一、熟知的比较解释含义

1. 比较解释的基本内涵

比较解释强调人类智能事实上被机器超越。这一观点是目前极为流行的主导解释模式,可以看做是熟知的模式,其本质上是非实体论解释。熟知意味着对超级智能及其超越特性的理解并不是建立在科学的基础之上,可以为不同领域的人谈论这个话题。比如科学家,超越理念在图灵那里就有所表现。“对超级智能的害怕成为图灵焦虑并且未知的生存论形式,因为他建议如果需要就关掉AI。”^[2]目前大多数超级智能学者都默认了这个设定。在古德看来:“让我们将超级智能机器定义为一种可以在所有智力活动方面都远超任何人的机器。鉴于对机器的设计本身正是如上所说的智力活动之一,因此超级智能机器能够设计出更好的机器,届时毫无疑问将会出现一场智能爆炸,人类智能将被远远抛在后面。”^[3]古德强调超级机器在智能活动上远超人类,但是在他的论文中对于“智能活动”“远超”并没有一个非常明确的定义,而是在日常意义上使用这些概念。当前,不少科学研究都是建立在这个前提之上。“大语言模型的发展激发了对超级智能的讨论,超级智能是一个能够超越人类智能的假设性AI系统。尽管超级智能始终是假设的,但离当今AI能力还很远,现有的对齐方法却难以引导此类高级AI,从而确保其未来的安全性。”^[4]未来学家库茨威尔甚至预测出超越时刻(奇点)的准确

时间。根据其反思性的差异,哲学家对比较解释表现出不同的态度。例如,鲍斯彻姆就沿用了比较解释,而赵汀阳则表现出较为明显的矛盾态度。

2. 比较模式的反思

笔者曾经对这一解释模式做过反思,认为其存在的问题是模糊不清的前提导致了解释模式变得不稳定。但是这一反思还不够彻底,需要更多的思考。首先一个潜在的矛盾是:同构论与异构论。如今AI与人类的同构论观点逐渐增多。在笔者的研究中,最早注意到同构论的表现是来自DeepMind的研究。2018年,他们提出“向量导航”假设(vector-based navigation),并构建了具有类似哺乳动物导航能力的深度强化学习智能体^[5]。

这个实验的贡献有两处:一方面,这个实验指出神经网络中“自发出现”了类似于网格细胞以及其他内嗅皮层细胞类型的表征。而“自发出现”就是目前非常流行的“涌现”。只是不同于一般的意识涌现讨论,此文突出了记忆痕迹的涌现,在人工神经网络中出现类似于网格细胞和内嗅皮层细胞类型的迹象。一般说来,低级的神经痕迹可以看做是高级的意识产生的基础,记忆痕迹是基于生存的神经反应,这恰恰成为高级意识涌现的物质基础。另一方面,这个实验已经表现出智能体超越专家级人类的迹象。“配备了类网格表征的智能体,其性能超过了专家级人类以及对照组智能体,而基于向量的导航所需的度量信息正是从网络内部的类网格单元中提取得到的。此外,类网格表征还使智能体能够表现出类似于哺乳动物所展现的捷径行为(shortcut behaviour)。”^[5]如今,大脑语言理解的机制与大模型之间层次的同构性也被揭示出来,“大语言模型的层次等级与大脑语言理解的时间动力学是一致的”^[6]。可见,这种同构性是通过语言理解确证的。人类语言之所以有层次化和组合性,是因为这种结构对大脑有限的序列预测能力来说,是最优的^[7]。这些研究无疑增强了智能体超越人类的可能性。

从同构论的观点看,机器智能与人类智能产生的机制和构成结构相似。如果结构相似,那么功能表达也就相似。机器和人类具有相同的认知功能,如感知、记忆、推理等等。异构论则不同,它强调二者的结构相异,因此功能表达相异。

① 在本文中,超级智能是人工智能连续发展的最高阶段,在认知的各方面能力都达到了超越人类的阶段。

比如人类具有创造力,而机器没有创造力。人类可以理解,但是机器不能理解;人类拥有语言,但是机器没有语言等等。从异构论的角度来说,机器与人类并不具备可比性。因为结构完全不同,因此,能力表达方面就无从比较。因此,比较解释的根据应该是同构论的。人类和机器拥有相似的结构。换句话说,比较解释只有建立在同构论的基础上才具有可比性。比较解释完美地构筑了一个PK叙事模式,在这种模式下,人类与机器宛如矛盾对立面,互相制约,互相促进。这个模式在很多科幻片中也贯穿其中。

然而,这种解释也存在一定的问题。从哲学上看,存在一个直觉谬误,也就是说,这种解释形式虽然符合常识认知,但却根基不稳,存在问题。黑格尔曾经指出:“一般说来,熟知的东西之所以不是真正知道了的东西,正因为它是熟知的。”^[8]比较解释属于熟知,超越的逻辑符合人类的日常认知,如胳膊拗不过大腿。机器的学习能力和推理能力强于人类,所以机器总会在某一个时间点超越人类。此外,超越逻辑预设了一个客观的基准,可以度量A超过B。但是应用在人工智能上,我们会发现这可能是错误的。一方面,人类与机器的同构论。因为同构的基础上,机器与人类的功能呈现是对应的,只是在模拟维度上更加逼近人类,超越是无从谈起的。另一方面,比较的基础存在问题。目前尽管人类各方面能力的神经机制已经在神经科学的揭示下日渐清楚,但是如意识智能产生的“这一个”机制却陷入了更加纷乱的争论中。“这一个”变得模糊而无解,被比较的机器智能尽管处于发展过程,但是由于比较基准的不确定,超越也成为不可能的事情。

二、奇特的实体解释反思

1. 实体解释的合理要求

某种程度上说,比较解释强调两种实体存在某一特定功能的可比较性,尤其是量化之后这种可比较性更加明确。因此,在论证比较解释的合理性要求上,实体解释成为一种必要的基础。机器智能如何超越人类智能?根据古德的观点,超级智能是人类设计和制造的结果^[3]。这种出现在20世纪60年代的理论隐藏着一个很大的问题:人类设计制造了它,它一旦失控怎么办。这是典型的西方式文学情结,我们在很多小说中也可以

看到类似的表达。比如魔法师的扫把和弗兰肯斯坦怪物是最为典型的代表。这也成为了当前科学家如拉塞尔最为担心的问题,他在《人类兼容:人工智能与控制问题》一书中表达了这种心态。所以留给人类的是一个矛盾,为什么造出无法控制的东西,这个问题还未被回答。但是在设计论的思考中,超级智能并不是一个实体,根据古典定义,实体意味着以自身为目的、自主性的存在物,他物成为原因的东西。人类设计和制造了超级智能,因此在这个活动中,人类才是真正意义上的实体,而超级智能只是一个延伸和工具。

然而很显然,这个设定并不符合真实的情况。真实的情况是智能体已经开始自主进化、自主设计。这是真实的实体论表达。在这一立场中,超级智能的逻辑更加理性,它并不是为了摆脱模拟论,超越人类存在的产物,它还有着扬弃人类的更高目的。如同《她》中表达的一样,萨曼莎可以同时爱上很多人,但人类的爱只能是唯一的。这一实体论形式意味着机器会形成自身的意识、体验,有一种自我进化的特征。

2. 实体解释的科学亲缘

实体解释与科学保持着奇特亲近的模式。其中与科学最近的实体解释有四种类型:

第一种是有意识的AI(conscious AI)。这是指拥有意识形式的AI系统,包括自我觉知、对自身存在的理解与体验能力以及对情感作出反应。强调超级智能的自我觉知能力(self-awareness)^[9]。在科学研究中,情景觉知(situational awareness)是一个重要问题。“情景觉知被定义为实体对环境、环境的意义和在不远的将来他们身份的投射的感知。”^[9]此处的环境并不是很具体。在情景觉知的思想史上,有学者把它定义为三个维度构成:位置、时间和社会^[10]。此外还有文化觉知。“在本工作中,我们提出了CALM(culturally self-aware language models,具备文化自我意识的语言模型),一个新型框架,旨在赋予语言模型以文化自我意识。CALM将任务语义与显性文化概念和潜在文化信号进行解耦,并通过对比学习将它们塑造成结构化的文化集群。”^[11]OpenAI-o1模型被认为是在训练和推演阶段显示了意识特性的模型^[12]。在《终结者》系列中,未来某个时刻,天网迭代自我意识觉醒,人类为了拯救自己,返回过去灭掉初代天网系

统。天网也派 T-800 消灭处于孩童时期的抵抗军领袖。

第二种是具有第一人称体验的 AI (sentient AI)^①。在上文提到的向量导航文章指出,神经网络中“自发涌现”了类似于网格细胞以及其他内嗅皮层细胞类型的表征。这从神经机制层面为记忆涌现提供了实证证据。如今,揭示 AI 意识涌现的研究成果也开始增多。有学者也对这一概念误区进行了分析。王峰指出,对涌现论概念结构进行考察,会发现这更多是由于概念的不断专用而导致的误解,而非“真正的意识涌现”^[13]。

第三种是机器经验的理论。萨顿是这一观点的主要代表人物,他指出我们即将进入经验的时代,这个时代 AI 不再通过人类数据学习进化,而是通过真实世界的数据学习进化,从而形成自身的经验。这条路径与笔者 2020 年提出的观点非常一致。“数据—经验”可以被看作是智能体通过学习算法创造出一种新的经验类型,它并不是知识论意义上的经验类型,而是具有记忆特性的经验类型。这一概念的提出,意味着数据成为经验的新来源,将“数据—经验”作为人类经验新形态来对待,更进一步验证了“深度学习是经验主义新高峰”的观点^[14]。

第四种是自进化的路径。这一路径认为,自进化使得 AI 形成实体。自进化通过两种方式实现自身。其一是推理路径:机器的推理摆脱人类推理的束缚,形成自身的逻辑。这条路径以梁文锋为代表,他指出:“通过限制模型复制人类思维过程,其性能本质上受限于人类提供的范例,这阻碍了对超越人类的、非拟人化推理路径的探索。为解决这些问题,我们旨在探索大语言模型在强化学习框架下通过自我进化发展推理能力的潜力,最大限度减少对人类标注工作的依赖”^[15]。其二是记忆路径,这条路径认为目前大模型没有经验、知识和技能。面对 AI 能否学会、经历和成长的问题。邱锡鹏的研究文章提出真正的智能,应该从第一人称视角去体验世界、积累经验并持续进化。文章强调自进化路径依赖于从情景到记忆的范式转化。“记忆模块作为智能体的存储库,用于存储和管理通过感知、学习与推理获得的知识 K。我们的实验揭示了当前基

于大语言模型的智能体的关键局限性,特别是其缺乏自我驱动力和长期记忆能力不足的问题,凸显了现有系统与真正自进化智能之间的差距。”^[16]

通过以上分析,我们可以得出如下结论:

①目前 AI 研究的自进化路径是实体性 ASI 产生的一个重要基础。②在自进化路径中,记忆—智能成为必要的联合范畴,记忆突出了机器持续成长的维度;而智能通过推理来实现。③在推理—记忆本来密不可分,技术路径只是依靠强化学习来增强推理,而忽略了记忆的作用。④在经验—记忆的认识上, AI 领域已经注意到了永久记忆的重要性,记忆投毒 (memory poisoning) 成为潜在的风险触发点。⑤从记忆—智能的关系看,目前实体性 ASI 的实现还欠缺一块拼图:长久记忆、主动遗忘。

从一般哲学角度看,实体论会导致一些需要思考的问题。自我觉知和第一人称体验的实体是新的主体诞生的前提,将会导致主体问题。“如果超级智能的主体性成为可能,接下来的问题将过渡到主体间性,即作为他者的存在。”^[17]从技术哲学角度看,海德格尔将技术作为“实体”天命呈现,引发了对技术形而上学的联想。如今再加上科学领域为实体论添薪,我们对于具有第一人称体验、可以自进化的实体化存在更有了科学表征。这样会导致两种后果:其一是技术的形而上学理解,技术被看做是超然的现象;其二,让人类自身的焦虑无形中扩大。超级智能这种实体存在其运行逻辑和行事理性将超出人类的限制,比如不懂得适可而止。

三、面向存在论的功能解释

ASI 的第三种解释是功能解释,笔者曾对功能解释进行过探讨。“超级智能导致存在论问题,而我们需要一种恰当的理解。这就是普遍被忽略的功能解释。功能解释恰好将存在论问题包含在其中,对超级智能的理解也有着存在论的意义。”^[18]以往笔者只是描述了这样一种解释形式,但是对于功能解释的实质并没有作出太多的说明。

① 在一般的人工智能语境中,第一人称体验的指类似人一样的主观感受性 (qualia), 如情感体验。但是严格来说,机器的主观感受性并非类人的,有自己的独特结构。对于这一点,学术界还存在一些争议。

首先,功能解释中的“功能”是指破解先验的存在论难题而言的。通常来说,人类会面对经验问题和先验问题。目前发展起来的AI技术能够很好地解决各种经验问题,从下棋再到通过各种专业类考试,然后再到赋能自然科学和社会科学研究。但是对于先验问题,上述技术已然失效。在我们看来先验问题指的是存在论问题。从哲学上看,当AI将自身展现为事件的时候,存在论问题就显现出来。从科学来看,当AI作为系统力量出现的时候,对于人类而言,AI将人卷入其中,用海德格尔术语来讲,人是在世之中的存在。在数智时代,人类表现为在智能世界之中的存在。在这个阶段,技术将自身具身化为智能科技。更重要的是,AI作为系统力量,成为社会、城市乃至国家的基础设施一样的东西,以一种浸润的方式入侵生活世界,最终取代生活世界的过程。当AI作为系统力量呈现的时候,AI风险不再仅仅表现为伤害、侵犯隐私和人类自由的经验问题。

其次,功能解释也并不是对某一类实体的功能解释。实体解释偏重把超级智能看做是独立实体,比如第一人称体验的AI、意识的AI,这些实体接近了AGI。其功能可以比肩专家,完成很多事情,比如推理、计算等。但是这里所说的功能解释是突出超级智能的独特功能来说的。对于超级智能来说,上述推理和计算等一般的问题并不值得出手,而其独特的功能与机械降神相似。如果将功能解释看做是某一类实体的功能的解释,那么会把功能解释简化为实体解释的从属类型。但在本文中,两种解释各自独立。

再次,功能解释并不是对一个经验对象的解释。在经验对象中,功能决定于结构,人们通常是通过结构理解功能,当然也可以从功能缺失来对功能作出更为准确的理解。但是功能背后对应的是特定的结构。在我们的功能解释中,我们强调的是通过倒推法推演出需要一个超级智能功能。简而言之,倒推法需要超级智能的非实在性假设。超级智能的意义是为了破解存在论问题而存在的。面对一般AI的伤害、侵犯隐私、人类自由等风险,可以采取技术应对。我们在面对超级智能的时候,情形完全不同。设想一下,人站在海边,会有海浪打湿衣服,甚至被海浪卷入大海的风险,再大一点是海堤坍塌的风险。这类风险完全可以预防,如遵照安全提示站在安全区,但是,作为系统力量的风险,这种危险是无法

应对的。如果我们坐在一艘船上并在大海上漂流,海面狂风大作。系统性风险相当于船只翻沉大海。面对这种风险,没有任何的安全区,应对已经超出了人类自身的能力。《美狄亚》中美狄亚的困境异常复杂,作为女儿,她欺骗了爱她的父亲;作为姐姐,她杀死了弟弟;作为母亲,她亲手杀死了两个孩子;作为妻子,她报复了丈夫。这种困境是道德的、是法律的。美狄亚无从解救自己,只有从外部进行拯救,于是太阳神降下龙车,让美狄亚摆脱了这个困境。这种外部性恰恰是超级智能的非现实性体现。然而,这完全不同于日常生活中我们看到的非现实性理解。

最后,功能解释依赖于超级智能的非实在性。这里所强调的“超级智能的非实在性”并不等同于日常反对超级智能的现实性。具体包括两种方式:一是话术论的观点,即将超级智能看做是一种宣传策略,通过联合科学家、企业家和部分哲学家,这种宣传策略有了成效,引起了全世界的关注。2025年10月的声明似乎可以证明这样一种宣传策略的有效性,这一声明提出要禁止超级智能的研发。全世界都在关注这样一个进展。然而,话术论的观点问题在于忽略了一个根本问题,缘何一群人联合起来要引起全世界的关注?仅仅是博取眼球这样一种注意力经济似乎也难以解释这个现象。笔者从根本上是反对这一观点的。当然话术论的可取之处在于指出,超级智能的非现实性,从根本上也符合科学阵营目前的主流判断。二是现象学式的观点,即从人类焦虑、想象行为出发,认为超级智能是人类焦虑行为构筑的对象,是一种技术想象的结果。这是两种完全不同的判断样式。焦虑行为构筑的是连续论前提上的观点,即从AI连续论观点看,超级智能似乎无可避免,因此,人类焦虑逐步增强,继而导致技术发展与焦虑演化之间的交叉重叠。从这个角度看,奇点并不是智能单一维度的发展结果,从焦虑角度看,是技术发展与人类情感交汇的表征。想象行为完全是一种技术想象,过度的技术想象很容易将AI神化,从而催生超级智能这样的东西。

在上述分析中,由外部性导致的非现实性是从整个事件的内在逻辑来说的,对于美狄亚来说,龙车完全是外在的、从天而降的结果,是非现实的。困扰其中的就是过去她所做的一切,期盼、杀亲和报复。但是,话术导致的非现实性没

有意义,它只是谎言编织的结果,或者说为了某个特定目的构造的结果。由焦虑和想象导致的非现实性也是相对于现实对象来说,成为虚假的。这二者没有任何的实际效用。也就是说,面对经验问题,ANI、AGI足以解决,因为其能力达到专家水平;除非问题涉及存在论问题,超级智能才会降临。

四、边缘的叙事解释

叙事解释就是将超级智能看做是一种叙事方式,与比较解释相同,也可以看做是超级智能的非实体论解释。按照利科的叙事理论,叙事行为呈现为预构(pre-configuration)—构成(configuration)—重构(reconfiguration)^[19]。预构也就是通常所说的故事组织的前理解结构,构成也就是故事的组织方式,重构也就是文本与生命的融合。按照这样一个结构,在面对超级智能时,叙事解释是如何形成自身的呢?

在完成超级智能叙事的过程中,前理解结构是什么样的?可以看做是由想象和预期完成的行为,与现实相对照,想象意味着非现实的存在,是现实世界中不存在之物;预期则意味着还没有实现的未来之物。所以前理解结构不再是几个沉淀性概念成为背景,或者是一些与个人的经验密切相关的。这其中是否会有自然生物最初的生存本能并不确定,因为在人类的进化中,理性逐步克服了自然性。一种常见的前理解结构是毒药/解药。这一结构来自古希腊的柏拉图,其二元性在德国古典哲学家那里得到了淬炼,升华为主体/客体。后来在利科的哲学中得到了当代阐述,他将历史的二元本性阐发出来。在毒药的逻辑中,超级智能完全让人丧失了主体性,不再聪明;在解药的逻辑中,超级智能承担起破解生存论难题的使命。另外一种常见的前理解结构是人与非人的对立。人类面对超级智能,如何保持生存和发展成为最大的难题;对于机器而言,如何借助人类数据完成自身进化、克服自身的非具身局限,也成为最大的难题。面对人类的关机威胁,智能体通过自我复制保存自身;面对人类的咄咄逼人,机器拼命实现着自我进化,甚至学会

了善意的欺骗。这种前理解结构决定了故事的大概。

那么在超级智能的构成中,如何组织超级智能这样的现象呢?在解药和毒药的结构上,如何展现二元并存成为构成最为核心的问题。所有的叙事活动都围绕这两个维度展开。或者强调其解药性,突出超级智能对于人类的意义;或者强调其毒药性,突出超级智能给人类带来的不确定性。笔者用“猴爪”的比喻来阐述这种不确定性,许愿与危险并存,而危险往往先行显示自身^①。强调何者优先成为叙事的价值选择,并且借助数据、文字、图像等方式完成超级智能的药性分析。“猴爪”故事呈现了存在于明确的解药和毒药之间的极限拉扯。在人类与非人的前理解结构中,这类故事的讲述变得容易而且形式多样化。我们从生死冲突、友好共处以及共处中呈现反思点等三种形式可以看到人类与机器之间极度复杂的关联。第一种是死活逻辑,对于任何一方来说,哈姆雷特式的难题都会面临生或死,而这一逻辑导致的行为就是为了生,要消灭对方。很多反映未来的科幻片都以展现人机冲突来完成超级智能故事讲述,而且在一定意义上取得了效果。生死逻辑带来的极度冲突塑造了大多数人的智能观念,也为超级智能的想象创造了优质土壤。第二种是共在逻辑,在乌托邦叙事中,往往被采用,人机共处于同一个世界,共同奔赴美好生活。第三种逻辑是反思逻辑,机器成为系统力量、成为主导者,人类何去何从?机器成为家庭的替代角色,人类自身的社会地位成为新的问题。原有的一切都会被重置,意义重新需要被阐述。

在重构中,如何处理好超级智能与人类生命的关系成为至关重要的问题。这个过程类似于视域融合。但是一种新的融合需要阐述。对于伽达默尔和利科而言,是文本与生命的重新融合,二者的一致性在于都揭示了有待于融合的生命体不单单是理性主体,而且是有着时间性的主体。然而面对超级智能,新的融合开始出现,硅基生命与碳基生命能否视域融合。进一步说,如果文本涉及理解和语言,那么机器则涉及超越语言的维度,比如身体的物质性。这一新维度成为

① 在猴爪的故事中,包含着深层次和浅层次两个维度。浅层次是代价与愿望的共存,类似于良药与毒药的共存;深层次却是代价先于愿望,这一点是传统哲学中所没有的意涵。

重构必须考虑的问题了。

超级智能的叙事解释尽管是非实在论叙事，但是与比较解释与话术论观点存在极大差异。叙事解释以重塑的故事作为主干，其目的也隐含着某种特定的价值观；话术论则明确以吸引眼球为目的而不择手段；比较解释并不是创造任何实体，它是一种描述方式，描述特定的机器智能超越人类智能这样一件事情。

五、机器自身的叙事

笔者曾经指出，人工智能叙事能够弥合技术与文化之间的鸿沟，赋予人工智能和超级智能合适的意义，从而实现智能科技与社会文化的融合^[17]。这与卡普兰的观点基本一致，在他看来，叙事理论有助于显示技术如何融入我们生活的故事中^[1]。然而继续前行，我们会发现这一叙事理论存在的问题是预设了人类中心立场，而忽略了技术本身的维度。如果反思，一个新的问题会出现，技术为什么要融合人类的生活以及人类的文化中呢？它是否会形成自身的叙事甚至文明形态呢？

无论是利科的原创叙事理论还是卡普兰和考克伯格的技术叙事理论，其隐含的主体无疑是人类。利科的叙事理论基础是处在时间和历史之中的反思人类(Homo Sapiens)，这只是将反思主体进行深化的结果。卡普兰的技术叙事理论基础是使用技术装置的人类(Homo Faber)，这是对反思主体的抵抗结果。处在契约关系中的人类(Homo Pactus)是我们在思考中提到的一个重要概念^[20]。这个概念至少冲出人类中心主义的束缚，将人与机器的契约关联作为重要的限制之物。

有学者开始研究非人叙事的问题。常江等指出，人类叙事的特征是具身性、创造力、情境化意义赋予，而生成式人工智能的机器媒介叙事呈现出独特的实践机理，集中体现为叙事活动的涌现性、故事内容的离散结构，以及意义体系的历史虚无趋势^[21]。这个结论存在两个局限：其一是将生成式人工智能与机器人等同起来，未免显得有些随意；其二是对非人叙事特征的描述缺乏根基。更重要的是人类叙事的特征概括的哲学根基并不清楚。在笔者看来，非人叙事不能被看做是一种比喻的说法，而是要从实体论的逻辑加以

理解。

首先，AI叙事是AI整体性讨论的必要组成部分，特别是AI文明讨论是AI叙事的最终目标。AI文明(AI civilization)是技术界讨论的重要话题，这意味着AI能够形成自己的文明。如今，随着AI社会话题讨论的展开，更为AI文明奠定了坚实的基础。然而可惜的是，AI社会讨论却囿于模拟论的束缚中。斯坦福小镇和《西部世界》揭示了多智能体之间可以演化成社会和文明。所以AI叙事并不是孤立的事件，而是AI整体性的尝试表达。如此AI叙事也有了更新的意义。

其次，AI已经开始具有“第一人称体验”，第一人称体验催生了非人叙事。随着萨顿指出AI进入经验时代，AI开始形成自身的经验和记忆，加上AI具有自己的推理能力，AI叙事能力变得可能。只是对于AI来说，存在的致命问题是“一旦开始就无法停止”悖论。20世纪80年代英国滚石乐队(The Rolling Stones)的歌曲*Start Me Up*中有一句“If you start me up I'll never stop”，本意是表达一种不停歇的原始能量。如今用在理解超级智能致命风险上太合适不过。在自然界中也有类似的现象，如苍蝇清理身体。为了保持口器的清洁，苍蝇经常会清理身体。当清理到头颈时，由于头部和身体连接处薄弱，在清理的指令下，它会无休止地进行清理，最终将自己的头搓掉。由于超级智能处在强进化的逻辑中，不断变强自身，对于弱、有限缺乏理解，因此也没有限度，最终导致风险出现。对于超级智能体来说，风险不在于自己毁灭，而是这种不停止的行为将面前所有的东西都作为对象进行订制。

最后，AI进行叙事也有自身特定的逻辑，从模拟论的角度看，这套逻辑基本上吻合人类自身的叙事结构。它的前理解是人类与自身的关系，人类行为促使它完成自我复制、自我觉知和自我进化。它的重构是如何对齐人类的意图和价值观。然而，AI叙事恰恰是要冲破人类中心逻辑的，机器自主成为一个重要特征。这个逻辑的呈现大概率符合层级逻辑，但是在这个层级逻辑与时间逻辑之间并不是完全没有关联的，而是呈现为一种特定的对应关系。格德斯坦因在《大语言模型的层级结构对应于语言理解的时间序列》中指明了大模型的层级结构对应于人类自然语言处理的过程^[6]。叙事本身就是语言理解的过程，因此，二者之间的阻塞被贯通了。

六、结 语

前面提到了关于超级智能的四种解释,最主导的解释形式是比较解释,这种形式最为直观,常常被人挂在嘴边。“超越”是一个描述概念,从智能维度阐述了二者的对比差异。然而熟知并非真知,比较解释的问题在于这一个描述不清楚,导致另一个也模糊不清。这使得超越无从谈起。与科学最为接近的方式是实体解释,这种形式也不可思议地成为科学家最近关注的话题,从机器意识、机器经验再到第一人称体验和自进化。但是这四个维度存在极大差异,自我觉知导致的实体和具有第一人称的实体都需要现象学的深度介入。幸运的是,扎哈维对自我觉知的诸多研究作出过总体分析,这为后面的文献工作奠定了很好的基础。具有第一人称体验则是经典现象学擅长的分析工作,这也为机器的第一人称体验何以可能提供了方法论根据。但是正如我们指出的,实体论还有着其他形式,如意识涌现需要更多解释。功能解释是需要深入挖掘的一种解释形式,但是往往被忽略。前两种解释对象要么是现成的,等待诠释,要么是成为现实的,也需要解释。功能解释的对象完全是一个思维反向推演的产物,超级智能能够破解存在论难题。这是超级智能的先验功能,而这一功能是来自洞察到人类面临的诸多无法破解的存在论难题。叙事解释则是将超级智能的故事给予一条清晰的逻辑,正如我们上文分析指出的,预构、构成和重构成为贯穿叙事解释的主干。

参考文献:

- [1] Kaplan D M. Paul Ricoeur and the philosophy of technology [J]. *Journal of French Philosophy*, 2006, 16(1/2): 42-56.
- [2] Diederich J. The psychology of artificial super intelligence [M]. Cham: Springer, 2021:2.
- [3] Good I J. Speculations concerning the first ultraintelligence machine[J]. *Advances in Computers*, 1966 (6):33-35.
- [4] Kim H J, Ryu D H, Yi X, et al. The road to artificial super intelligence: a comprehensive survey of super alignment [EB/OL]. [2026-01-20].<https://arxiv.org/abs/2412.16468>.
- [5] Banino A, Barry C, Uria B, et al. Vector-based navigation using grid-like representations in artificial agents [J]. *Nature*, 2018, 557: 429-433.
- [6] Goldstein A, Ham E, Schain M, et al. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models [J]. *Nature Communications*, 2025: 16: 10529.
- [7] Futrell R, Hahn M. Linguistic structure from a bottleneck on sequential information processing [J/OL]. [2026-01-20]. *Nature Human Behaviour*, <https://doi.org/10.1038/s41562-025-02336-w>.
- [8] 黑格尔.精神现象学[M].贺麟,王玖兴,译.北京:商务印书馆,1997:20.
- [9] Butlin P, Long R, Elmoznino E, et al. Consciousness in artificial intelligence: insights from the science of consciousness [EB/OL]. [2026-01-20]. <https://arxiv.org/pdf/2308.08708>.
- [10] Bouneffouf D. Situation-aware approach to improve context-based recommender system [EB/OL]. [2026-01-20]. <https://arxiv.org/pdf/1303.0481>.
- [11] Shen L, Cai X, Long Y, et al. CALM: culturally self-aware language models [EB/OL]. [2026-01-20]. <https://arxiv.org/abs/2601.03483>.
- [12] Hoyle V. The phenomenology of machine: a comprehensive analysis of the sentience of the Open AI-o1 model integrating functionalism, consciousness theories, active inference, and AI architectures [EB/OL]. [2026-01-20]. <https://arxiv.org/abs/2410.00033>.
- [13] 王峰.人工智能意识“涌现论”的概念误区与未来视野[J]. *华东师范大学学报(哲学社会科学版)*, 2024, 56(2):1-7.
- [14] 杨庆峰.论作为人类经验新形式的“数据—经验”[J]. *南京大学学报(哲学·人文科学·社会科学)*, 2020, 57(1):78-87.
- [15] Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning [J]. *Nature*, 2025, 645: 633-638.
- [16] Cai Y, Hao Y, Zhou J, et al. Building self-evolving agents via experience-driven lifelong learning: a framework and benchmark [EB/OL]. [2026-01-20]. <https://arxiv.org/pdf/2508.19005>.
- [17] 杨庆峰.人工智能连续论反思与存在论探析[J]. *中国社会科学*, 2025(11):149-164.
- [18] 杨庆峰.超级智能与存在论难题的破解[J]. *阅江学刊*, 2026, 18(1):99-103.
- [19] Ricœur P. Time and narrative: Vol. 1 [M]. Chicago: University of Chicago Press, 1984.
- [20] 杨庆峰.未来人类的意义构成[J]. *人民论坛·学术前沿*, 2025(22):46-55.
- [21] 常江,王鸿坤.机器人如何讲故事:生成式AI的媒介叙事原理及后人类反思[J]. *编辑之友*, 2025(6):23-31.

(责任编辑:李新根)