中州学刊

人工智能伦理研究中的拟人化之弊及其哲学反思

阮　凯

摘　要：在当前人工智能的人文社科研究中，拟人化叙事愈发普遍。拟人化指将人类特征尤其是心智与社会特征主观地投射或赋予非人对象，其根源在于人类难以完全超越人类中心主义视角。越来越多的学者以拟人化方式思考人工智能伦理问题，在一定程度上削弱了人工智能伦理研究的严谨性、客观性和可信度。这要求研究者深入探索人工智能伦理的研究规范问题：拟人化的思维和叙事方式是否适合人工智能伦理研究，在人工智能伦理研究中拟人化的界限在哪里。维特根斯坦哲学有助于全面地审视和克服人工智能伦理研究中的拟人化之弊：在语言使用层面，拟人化修辞与严谨务实的学术研究有实质性冲突；在思维方式层面，维特根斯坦提供了一套从语词和句子到语言活动和生活形式的整体主义哲学，提醒我们要谨慎对待将人类的道德概念和道德判断简单移植到人工智能的做法。探索人工智能拟人化的认知基础及其可能危害，能为发展人工智能伦理的研究规范提供理论启示。

关键词：人工智能伦理；人工智能拟人化；人工智能道德地位；维特根斯坦；整体主义思维

中图分类号：B82；TP18　　**文献标识码：**A　　**文章编号：**1003-0751(2026)06-0114-08

当前，全球人工智能(AI)技术正在飞速发展，各种大模型也在不断更新迭代，为我国的科技发展提供了历史性机遇；同时，人工智能正在赋能人类社会的很多重要领域，如人工智能驱动的科学(AI4S)成为科学发展的前沿领域，人工智能与医学的融合发展有力地促进了医学研究的创新。随着人工智能的影响日益广泛深入，人工智能伦理治理研究迫在眉睫，并成为当前国际和国内科技伦理的重要内容。中国政府在《关于加强科技伦理治理的意见》中强调，要“制定生命科学、医学、人工智能等重点领域的科技伦理规范、指南等……重点加强生命科学、医学、人工智能等领域的科技伦理立法研究”[1]。技术的发展以及国家和人民的关切，使得人工智能伦理研究一跃成为科技哲学、应用伦理等领域的重要研究方向。然而，人工智能伦理研究虽聚焦规范问题，但其自身的研究规范尚未真正形成，这使得它在研究的方法、方式、形态、目标等方面呈现一定的混乱态势。例如，一篇充满想象和拟人化思维的论文与一篇严谨务实、有科学和事实依据、有哲学思辨内容的论文都被不加区分地认为是人工智能伦理研究成果。在笔者看来，科技伦理研究者有责任探究人工智能伦理的研究规范问题，而本文则主要以拟人化为焦点，来以点带面地探索人工智能伦理研究的规范问题。

一、人类对待人工智能的拟人化倾向及其根源

拟人化(Anthropomorphism，又译拟人论)是把非人的事物人格化，把人类的特质(如思维、能力、情感、身份及社会特征等)主观赋予和投射到其他事物上。拟人化具体表现为拟人化修辞、拟人化思维和拟人化行为等方面。心理学家和行为科学家尼古拉斯·埃普利定义拟人化时更强调了行为要素，

收稿日期：2025-08-29

基金项目：国家社会科学基金重大项目“数字智能技术与哲学发展及知识生产范式变革研究”(24&ZD320)。

作者简介：阮凯，男，中山大学哲学系(珠海)副教授(广东珠海　519082)。

他认为,“拟人化描述了将非人类主体的真实或想象行为赋予类人的特征、动机、意图或情感的倾向”[2]864。动物伦理研究者也对拟人化问题进行了长期深入的研究,他们认为人的拟人化实践是指:“将类似人类的心理状态赋予非人类,或者将表示人类(通常被认为是人类独有的)现象状态的术语应用于非人类生物……这一定义也与弗朗斯·德·瓦尔(Frans de Waal)的定义相似:‘将类似人类的特征和体验(错误地)归因于其他物种。’”[3]这些定义虽然各有侧重,但也有共同指向,即将独属人类的特征、体验、心理状态主观地赋予非人类实体,用相对应的概念去言说非人类实体。

在人类的社会历史和日常生活中,常有三类拟人化的现象:一是将动物拟人化。动物与人类有着紧密的亲缘关系,人类生活中处处都有动物的身影。同时,一些智力水平较高的动物能展现与人类相似的社会活动、心理活动和情感活动,因此动物是人类常见的拟人化对象。二是将生活中常见的无生命的存在物拟人化。如李白诗云“我寄愁心与明月,随风直到夜郎西”,就是将明月拟人化,赋予月亮感受人类忧愁的特征和传达人类情感的能力。三是将宗教中的神明拟人化。如古希腊神话中的诸神都具有与人相似的形象,与人一样有七情六欲,他们身上的神性也是人性的投射和夸大。Anthropomorphism 这个词有时被翻译为神人同形同性论,也表明宗教中的拟人化现象体现为将人的“形象”与“人性”赋予神明。在前两种拟人化现象中,对象是一种客观存在,人类更多地是以一种拟人化的方式去认识、理解和对待这些事物;而在后一种拟人化现象中,从唯物论视角看,人类还以一种拟人化的方式创造了信仰对象和诸神诸仙。

拟人化在文学中是一种修辞手法,但从更深层面来看,将事物拟人化是人类的一种较为普遍的思维方式和行为倾向,在东西方不同文化中也的确普遍存在拟人化现象。我们不禁要问,人类的拟人化行为为何具有跨越文化的普遍性?要清楚、深入地回答这一问题并非易事,一种哲学式的宏观回答是,人类在看待和理解外界事物时,不可避免地带有人类中心主义视角。人类中心主义视角是人类生存的一种重要方式,无论人如何努力去除人类中心主义视角,都无法完全摆脱这种思维和视角,原因无他,只因我们是人类。当代哲学家托马斯·内格尔生动地指出了人类对世界的客观性认识和主观性理解并存,原因是:“我们是一个大的世界中的小的生物,我们对这个世界的理解非常偏颇,而且事物呈现给我们的方式,既取决于世界,也取决于我们的结构。”[4]人类中心主义的视角也正是我们人类的结构所形成的视角。

拟人化代表了人类视角,人类无法完全摆脱拟人化思维;然而,人类拟人化的活动并非时刻出现,人类只是在某些特定的情境下进行拟人化活动、使用拟人化思维。尼古拉斯·埃普利的研究向我们揭示了拟人化的心理学根源,同时也解释了拟人化的间歇性。他提出的理论认为:“拟人化有三个决定因素:可获取并适用的人类中心主义知识(被引出的主体知识,elicited agent knowledge)、解释和理解其他主体行为的动机(效能动机,effectance motivation),以及对社交接触和归属感的渴望(社交动机,sociality motivation)。”[2]我们可以将这三个决定因素分为两类:认知因素和动机因素。从认知因素看,人类习惯于用已有的知识来解释未知的知识,相较于关于其他动物的知识,人类关于自身的知识是较早获得并且更丰富的,同时人类个体也更容易和直接地把握自己内心活动和心理状态,因此,人们在试图解释动物的心理活动时,往往会借用人在相似情景下的心理活动来帮助理解。比如,一个孩童在看到鸟儿困在鸟笼中,他可能会想着这个鸟儿是不是为自己失去自由而感到深深的难过,是不是在等待亲人的解救,它着急的叫声是不是在请求我打开鸟笼。从动机因素看,效能动机和社交动机驱动人们不断地探索、理解和把握周遭环境,和环境中的同类和其他事物进行互动,产生社会性联系,拟人化也有助于人类探索和理解周遭,和环境中的非人类主体进行互动。

尼古拉斯·埃普利进一步解释了人类拟人化行为出现的条件:“当人类中心主义知识易于获取且适用时,当人们有动机成为有效的社会主体时,以及当缺乏与其他人类的社会联系感时,他们更有可能进行拟人化。”[2]由此可见,拟人化并不是时刻发生的,而只是在特定条件下发生。知识匮乏和与人相似,更易引起拟人化。就知识匮乏而言,孩童较成年人更易做拟人化行为,古人较今人更易做拟人化行为。知识匮乏是一个关键因素,如果人关于认知对象的知识相对匮乏,这时用关于自己和同类的知识来解释该对象也在情理之中。就与人相似而言,动物较植物更易成为人的拟人化对象,布娃娃和玩偶较钱包更易成为儿童的拟人化对象。当人们缺乏与他人的社会联系时,他们更愿意选择一个与人相似

的事物来补偿这种社会联系的缺失。

在人工智能(下文简称AI)时代,大语言模型和智能机器等各种形态的AI成为人类新的拟人化对象,我们可以从知识匮乏和与人相似来解释为何人类容易将AI拟人化。关于AI的知识匮乏有两个层面:一是个体知识匮乏。由于AI还是比较艰深的技术,一些人甚至是人文社科研究者在不了解当下AI技术的情况下大谈AI威胁论,认为AI将"统治"和"奴役"人类。二是集体知识匮乏。由于AI的黑箱特征以及大模型愈加复杂的内部结构和参数,AI是否有意识等问题仍存在很大争议。换言之,人类并未完全认识AI这个人造物,这也为将AI拟人化提供了可能。就与人相似而言,随着AI技术的进步,即AI在语言对话、推理认知、知识储备、身体形象乃至情感表达等方面愈发与人类相似,人类有将AI拟人化的倾向和冲动也在情理之中。慕尼黑大学的斯文·尼霍尔姆也指出:"大多数人倾向于将机器人拟人化,人类有一种自发地将人类的特质赋予机器人的倾向,并且在与机器人互动时也会认为它们好像有类似人的特质(humanlike qualities)。这是人机交互伦理中需要始终牢记的第二个关键考量。"[5]4尼霍尔姆的观点代表了越来越多的学者已经意识到将AI拟人化带来的学术问题,但有一个重要问题却容易被忽视:学者们也有拟人化的倾向,在伦理研究中将AI拟人化会导致灾难性后果,如果AI伦理研究缺乏科学事实依据,充斥着拟人化的想象与假设,那么我们将难以依赖这些研究成果来有效规范和治理AI的发展。在此意义上,研究AI伦理研究中的拟人化之弊及其解决方案就成为当务之急。

需要说明的是,"将AI拟人化"与"让AI类人化"既有联系,也有根本的区分。二者的联系是:将AI拟人化所体现的想象和希望也构成了让AI类人化的动力和目标。二者的区分则更为值得重视:将AI拟人化是指人类主观赋予和想象AI具有人类特质,实际上AI本不真正具有这样的特质;而让AI类人化是指通过科学研究和技术实践努力,让AI真实地模仿人,让AI实际上具有一定的类人特质①。例如,(1)模仿人的形象与行为,研发者在设计AI时试图让其形象与行为接近人类。(2)模仿人的认知与思维,从符号主义AI到联结主义AI都体现了对人类认知和思维不同层面的模仿。(3)模仿人的情绪与情感,现有的大模型、情绪识别和情感计算技术也在不断探索如何让AI更好地识别和模仿人的情绪与情感。澄清这两个概念的区分,也将有助于在下文更清楚、更深入地讨论AI伦理研究中的拟人化之弊,不至于让读者误以为本文在讨论AI类人化的问题。

二、拟人化之弊:以人工智能道德地位为例

在对AI的人文研究中,拟人化的思维和修辞容易产生极端化的思想。某一些人文学者长期从事人的研究,当他们思考AI时,很容易把人的方方面面想象成AI的能力特征,进而倾向于夸大AI伦理风险。例如,耶路撒冷希伯来大学历史系教授尤瓦尔·诺亚·赫拉利(Yuval Noah Harari)就以拟人化的方式倡导AI威胁论。他写道:"人类现在还忙着创造像人工智能(AI)这样的新技术,但这些技术有可能逃脱人类的掌控,反过来奴役或消灭人类。"[6]V"人工智能如果要夺取权力,最简单的方法不是逃出制造科学怪人的实验室,而是赶快去讨好偏执的提比略……只要一不小心,人工智能就会夺取权力。"[6]311从赫拉利的著作中我们看到,人工智能威胁论往往与拟人化紧密相关。他所说的"逃脱……掌控、奴役或消灭人类、夺取权力、讨好"等,本质上是一种拟人化的思维和修辞,这引导读者把AI想象成与人类处于对立阵营并拥有类人特征的敌人。反之,克制随意的拟人化冲动,也有助于避免AI威胁论的过度泛滥。

为了让研究更聚焦,本文主要讨论AI伦理研究中的拟人化之弊,并且以"AI是否具有道德地位(moral status)"这个代表性问题为切入口来展开讨论。同时,AI是否具有道德地位的问题与支持AI具有道德地位的立场紧密相关,如果学者们不认为该问题是一个真问题,自然也不会坚定地持有该立场。维森特·穆勒概括了国际学术界支持AI具有道德地位②的两类主要理由:关系转向(relational turn)的理由和行为主义(behaviorism)的理由[7]。这一概括也有助于我们针对性地反思,为何将AI拟人化会导致虚构的问题与经不起推敲的观点。

(一)以关系转向论证AI道德地位为何是拟人化

穆勒总结了关系转向论证的主要观点和代表人物,即AI和机器人是否具备道德地位,不应取决于其本质特性,而应取决于人类与其的互动关系。如果人类选择把它们视为道德行动主体,就能赋予其道德地位。大卫·甘克尔、马克·科克尔伯格等都

是该观点的代表人物[7]。科克尔伯格较早地从关系视角来思考 AI、机器人及动植物的道德地位问题。他提出了关系转向的代表性思路:“我们善待我们的猫,并不是因为我们对猫进行了道德推理,而是因为我们已经与它有了一种社会关系……如果我们给我们的狗起一个人名,那么——与我们食用的无名动物相比——我们已经赋予了它一种特殊的道德地位,这种地位是独立于它的客观属性的。采用这种关系的、批判的、非教条主义的方法,我们可以论证说,同样地,人工智能的地位将由人类赋予,并且将取决于它们将如何嵌入我们的社会生活、语言和人类文化中。”[8]

穆勒对这类论证提出严厉批评:“‘关系转向’实际上是一种道德地位的相对主义解释(relativist account of moral status),这带来了许多问题:无法判断方式的对错,无法评判观点的优劣,无法实现道德的进步。”[7]因此,我们应谨慎对待关系转向论证,因为过度强调关系特别是人类主观界定的关系,容易滑向相对主义,相对主义会导致谁也说服不了谁的境地。如一些人认为,我和 AI 的关系很亲密,我想赋予 AI 道德地位,这种“我认为我们之间有什么样的关系”和“我想赋予 AI 什么样的地位”,并不能说服那些坚定地认为“我和 AI 没有这样的关系”和“AI 没有道德地位”的人。如果你与玩具公仔或虚拟偶像有某种情感联络,你认为你和它们之间有某种关系,这是否意味着玩具公仔、虚拟偶像也拥有道德地位呢?这一推论难以成立,因为个体性的情感关系不能被直接等同于普遍性的道德判断。

笔者较为赞同穆勒的批评,在该批评的基础上笔者还认为,科克尔伯格的关系转向论证存在两方面问题:一是在社会生活中,具体的关系往往是个体与个体之间的,个体之间关系如何并不能回答一个公共性的问题。例如,个人如何善良地对待他养的猫,甚至拟人化地把它视为家庭成员,赋予它家庭地位,并不能为人类应该如何对待猫树立一个公共的标准,同样,AI 的道德地位等问题也是如此。二是赋予 AI 道德地位是将 AI 拟人化的表现。其理由是:人的道德特质、道德地位、权利义务等都是人类道德生活的重要组成部分,这些概念也是人根据道德生活的需要创造出来的,道德地位、道德行动主体等概念都是用来形容人的,因此主观地赋予 AI 道德地位也是一种拟人化的行为。此外,出于关系的原因而进行主观赋予,与前文提及的出于社交动机而进行主观赋予没有实质差别,“拟人化的主观赋予”与“真正和实质上拥有”才有着根本不同。某些伦理研究者如果以拟人的方式主观赋予 AI 以某种道德特征和地位,这就类似于诗人以拟人的方式主观赋予日月山川等人的情感和思想,这些拟人化可以用来表达个人情感和态度。但我们仍要谨记,人以拟人化的方式去想象和赋予 AI 类人的能力、特征乃至道德地位,与人真正拥有这些能力、特征和道德地位有根本的不同。

(二)以行为主义论证 AI 的道德地位也是拟人化

穆勒认为该论证以约翰·丹纳赫等人为代表,从行为主义角度论证 AI 的道德地位的理由是:“(1)如果在行为表现上,一个机器人大致等同于一个被普遍认为具有重要道德地位的实体(entities),那么赋予该机器人同样的道德地位是恰当的。(2)人们普遍认为,机器人在行为表现上可以大致等同于其他具有重要道德地位的实体。(3)因此,赋予机器人重要的道德地位是正确和恰当的。”[7]相信对 AI 技术与伦理学研究有一定了解的人,都会觉得如此论证过于简单、粗糙和无理,自然不太会认同这种行为主义论证。笔者对此的意见是:这一论证的大前提与小前提都存在很大问题。就小前提而言,什么是“行为表现大致等同”呢?这种模糊表达有损论证的严谨性。另外,至少还有很多人认为机器的行为和人类的行为还是有海量和本质差异的,丹纳赫为何就能轻易得出小前提中的共识呢?就大前提而言,大前提中的观点也是很多人不能接受的,即便机器在行为表现上与人类似、等效或者等同,这与机器具有道德地位是两件不同的事情,至少前者推不出后者。穆勒也主要是对该论证的大前提展开批评,他指出这种行为主义论证忽视了行为背后的主观状态,这可能导致人们错误地赋予机器人不应有的道德地位。

令人惊异的是,还有不少学者表达了与丹纳赫类似的主张,如科克尔伯格此时又成了丹纳赫的同路人,他写道:“人类有理由将虚拟的道德行动主体性和道德责任(virtual moral agency and moral responsibility)赋予那些与自身相似的非人类存在,并且在这种相似性的范围内,根据这一信念采取行动。”[9]在笔者看来,这些所谓的论证仅仅是基于拟人化思维的畅想,并不能充分证明 AI 具有各种类型的道德地位、道德行动主体性或道德责任。因为拟人化思维常常基于联想和想象:某物与人类有某方面的相似或相同,进而联想它能与人类有其他方面的相似或相同。如果仅凭 AI 在行为表现上与人类相似,就

主张AI具有人类的特征、地位、权利等,这样不负责任的研究方式和态度无疑严重损害了AI伦理研究的公信力。

综上所述,基于关系转向和行为主义等理由去论证AI具有道德地位,都和拟人化有着紧密的联系。由此我们得出结论:一是未来人们还会基于其他理由去主张AI具有道德地位、道德权利等观点,针对这些理由,我们首先也要审查它们是否是拟人化的;二是在拟人化的论证之外,目前缺少令人信服的理由来支持AI具有道德地位的观点;三是学术问题和学术观点总是相伴而生,主张AI具有道德地位的学者也把"AI是否具有道德地位"这个问题凸显出来。因此,对观点的批判也要延伸到对问题的批判,这类问题也可以视作人类是由拟人化思维构造出来的,所以西方学术界才会在这个被构造的问题上争论不休,找不到合适的答案。从维特根斯坦哲学看,消解虚构的假问题也是一项重要的哲学工作,对一个有批判意识的伦理学家而言,并非被学者们构想和热议的AI伦理问题都是真问题,推进AI伦理研究的先行工作是审查AI伦理问题本身。只有去伪存真,才能继续推进AI伦理研究;否则,将只能在错误的道路上越走越远。

值得特别关注的是,由拟人化思维出发倡导AI具有道德地位,这一现象并非个例,而只是代表性案例,它让我们理解在AI伦理研究中拟人化思维的重要弊端。国际上的研究逐渐认识到这一点,如哲学学者阿德里亚娜·普拉卡尼认为:"拟人化通过给本不具备人类特质的人工智能系统赋予人性化特征,从而夸大其能力与性能表现……拟人化(anthropomorphism)也被证明会扭曲人类对AI的道德判断,比如对其道德品质和地位的判断,以及对其责任和信任的判断。"[10]哲学学者、神经伦理专家阿琳·萨莱斯等人也提出担忧:"通过人类心理特征的视角来理解AI,可能会将AI降低为人类思维的复制品,这种做法会导致人类对AI带来的伦理问题的分析存在缺陷并最终受限。"[11]要言之,当人们把人类的道德特征赋予AI,将人类的道德词汇用于AI时,都是以拟人化的方式进行AI伦理研究,这也最终导致关于AI的某些人文研究尤其是伦理研究呈现出系统性错误。为了预防拟人化给AI伦理研究带来的系统性错误,我们需要在该领域慎用拟人化,而采纳更加严谨科学的研究方式。在这方面,维特根斯坦哲学能给予我们积极的理论启示。

三、谨慎使用拟人化:维特根斯坦式诊疗

如前所述,在AI伦理研究中拟人化让学者们不断地制造诸多缺乏根据的观点和问题,甚至严重损害了AI伦理和应用伦理的学术声誉,让旁观者误以为AI伦理研究仅仅是一些拟人化想象而已。因此,在认识到AI伦理研究中的拟人化之弊后,我们需要在相关研究中谨慎对待拟人化。拟人化思维与拟人化修辞是相辅相成的,都是拟人化在不同层面的表现。后期维特根斯坦哲学常被视为思考语言误用的诊疗型哲学,为我们提供了深刻的语言观和思维方式。故笔者认为,挖掘维特根斯坦的哲学思想资源,可以帮助我们更深刻地认识到,为何不应在AI伦理研究中倡导拟人化,在拟人化思维之外我们还能用哪些思维方式去研究AI伦理这门严肃的学问。要言之,维特根斯坦的后期思想有助于在方法论和研究范式层面,为AI伦理的研究规范建构提供新的启示。这些启示包括两方面内容。

(一)语用观的启示:为何AI伦理研究中拟人化不宜过多

后期维特根斯坦哲学的一个核心贡献是使对语言意义的讨论转向对语言用法的讨论:"一个词在语言中的用法便是其意义。语法描述语词在语言中的用法。"[12]与关切语用密切相关,他尤为关切因语言的误用引起的哲学疾病,并主张通过纠正语言的误用而完成哲学治疗,因此他提出了一个核心主张,"哲学家诊治一个问题;就像诊治一种疾病"[13]106。从该视域看,AI伦理研究中的拟人化语言和修辞经常是对语言的误用,由这种语言误用生成的哲学观点和哲学问题应被视为一种哲学疾病。其根本原因在于,每种修辞都有其适用范围,每个概念都有其通常用法(维特根斯坦称之为语法),这些都是使用语言的界限。拟人化修辞在一些文学创作和抒情场景是恰当的,可以增加表达的形象性,引起情感共鸣,但是在科学研究、AI伦理研究等严肃的学术研究中,这种修辞方式和思维方式并不十分适用,如果强行使用,就容易导致修辞学层面的误用,不仅会削弱研究的科学性、论证的严密性,也会带来各种哲学问题和哲学疾病。这就像降压药适用于高血压患者,但并不适合血压正常甚至血压偏低的人服用,如果误用可能会导致严重的低血压、头晕、晕厥甚至休克。因此,我们需要区分合理的拟人化与有害的拟人化。

虽然 AI 研究中的拟人化常是一种思维陷阱,引导人们走向谬误,但研究者习惯从一个极端走向另一个极端。如有研究者主张:“拟人化也是一种谬误,这一点经常被忽视。当一个人假设或做出毫无根据的推论,认为一个非人类的实体具有人类的品质时,就会出现这种谬论。这涉及将人类的特征赋予非人类事物,比如:‘我的车在生我的气。’”[10]在笔者看来,这种批评只有结合语境才更准确更全面。维特根斯坦后期哲学非常强调要在语境中理解某个语句或表达,如“我听到‘我害怕’这句话。我问道:‘在何种语境中你这样说?这是一声叹息,是一种坦白,一种自我观察,还是……?”“问题在于:‘它处于哪种语境中?’”[14]语境的重要性让我们认识到:拟人化(修辞、思维等)本身并没有对错,我们不能说人类的一种重要修辞方式和思维方式是错的,而要说它们用在哪些领域和语境中是合理、恰当的,用在哪些领域和语境中是不恰当乃至错误的。在文学和诗歌创作中,没有充分理由地认为一个非人类的实体具有人类的品质,反而可能会被赞誉为有想象力,而在 AI 伦理研究的一些严肃论证中,拟人化的表达和思维则容易将人引向谬误。

让我们进一步思考,拟人化为何适合一些语境而不适合另一些语境呢?笔者认为,一个可行的建议是:当语境不需要充分的论证和扎实的理由时,该语境不排斥拟人化;反之则不适合拟人化。因为拟人化强调主观赋予和主观想象,这会损害学术研究中论证的质量。例如,诗歌是人类的重要表达方式,其中有很多拟人化和想象的内容,诗人并不需要过多解释这种拟人化和想象的表述是否有现实对应物;而在 AI 伦理这类学术研究中,每个重要判断和理论观点则都需要充分的论证和扎实的理由。该建议也有助于使 AI 伦理学成为更具客观性、令人信服的学术研究,而非研究者自说自话、各自想象的科幻型研究。目前 AI 伦理研究备受关注,参与研究和讨论的人越来越多,几乎任何人都可以就 AI 伦理议题随意说上几句,这使得缺乏事实依据和充分论证、充满想象的 AI 伦理研究也越来越多。在笔者看来,维护 AI 伦理研究的声誉,不让该研究等同于科幻想象和个人意见彼此争执的场所,一个重要的工作就是树立 AI 伦理研究的规范,对于合理的拟人化和有害的拟人化的讨论,启发我们要树立谨慎使用拟人化的规范,在 AI 伦理核心问题的确立、关键观点的论证上尤其要慎用拟人化。谨慎使用拟人化并不等同于完全拒绝拟人化。原因在于:人终究无法完全摆脱拟人化思维,同时,完全拒绝拟人化也会导致人们在讨论 AI 时用其他形式的拟化来替代拟人化。例如,动物伦理研究中批判的拟人化思路也能对 AI 伦理研究提供有益启示:“避免拟人化只会产生其他形式的拟化,如机械化,与其回避拟人化,不如将其视为一种应当批判使用的沟通策略。”[15]

(二)整体主义思维方式的启示

维特根斯坦后期哲学不仅是一种语言哲学,同时也蕴含了深刻的整体主义思维方式,为我们批评和克服 AI 伦理研究中的拟人化提供了可靠的思维模式。在 AI 伦理研究中,如果慎用拟人化的表达和思维,我们会面临用什么表达方式和思维方式的问题,维特根斯坦哲学也能为我们开启新思路。他的语用论、语境论、语言游戏和生活形式等思想为我们提供了一个整体主义立场:不要孤立地去看待一个单词、一句话,而要从具体语境和语言游戏中去理解语词或句子。单个语词和句子都不是孤立的,而是语言系统的有机组成部分,在这个语言系统中,语词之间、句子之间都互相联系、互为支撑,而语言系统又与人的话语实践不可分割,进而与人的生活形式和现实世界紧密相关。维特根斯坦在《哲学研究》等著作中建构了语词—句子—语言活动—生活形式的整体性关联,如“理解一个句子就是说:理解一种语言。理解一种语言就是说:掌握一种技术”[13]93,“想象一种语言就叫做想象一种生活形式”[13]11,“‘语言游戏’这个用语这里是要强调:用语言来说话是某种行为举止的一部分,或某种生活形式的一部分”[13]15。他主张从语言系统和话语实践去理解句子和语词,又从人的生活世界和更广泛的实践活动去理解语言,为我们树立了整体主义语言观和思维方式的典范。

其他哲学家也在不同层面呼应了维特根斯坦的见解。如康德在其伦理学体系中就深刻地指出了重要道德观念之间是互相联系、互相依存的。他提醒我们:“自由当然是道德法则的 ratio essendi[存在根据],但道德法则却是自由的 ratio cognoscenti[认识根据]。因为如果不是在我们的理性中早就清楚地想到了道德法则,我们就绝不会认为自己有理由去假定像自由这样的东西。但如果没有自由,在我们里面也就根本找不到道德法则。”[16]自由与道德法则之间相辅相成,其他道德概念之间亦是如此。恩格斯则强调了语言和劳动实践的整体性联系,他说:“语言是从劳动中并和劳动一起产生出来的,这个解释是唯一正确的。”[17]劳动的发展促进人与人之

间更紧密的相互协作,这进一步促使他们进行更多的语言交流。

从维特根斯坦整体主义的哲学视角看,道德概念之间、道德命题之间都是有机联系的,它们并非孤立存在,而是互相联系、环环相扣,共同构成了一个有机的伦理道德系统;与这些概念、命题相对应,人类的道德能力、道德特质也是有机联系、互相依赖的。当我们希望用某个道德概念、命题来言说AI,或将某个道德特质赋予AI时,就必须反问,针对其他相关联的道德概念、道德命题和道德特质,我们是否能同时这样做呢?如不能,我们就必须谨慎对待甚至放弃这样的尝试。例如,当我们主张AI具有道德地位时,这会进一步要求我们思考它是道德行动者还是道德受动者,承认它是道德行动者,又要求我们承认它具备理性、自由意志和意图意向等;承认它具备理性与自由意志等心智特征,又要求进一步承认它具有其他特征。承认或赋予AI道德地位看似简单,但如果我们真的这样做了,会带来难以承受的理论负担。因此,从一种整体主义的语言观和思维方式出发,我们不能以拟人化的方式将道德地位赋予AI,否则,我们必须将其他相关联的道德观念、心智能力以及人类特征赋予它们,并充分证明这种赋予的合理性。赋予AI以行动主体性也会存在类似的问题,按照斯文·尼霍尔姆对"行动主体性"的总结,人有行动主体性意味着人拥有一种复杂能力,这种能力是人"采取行动、做出决定、对如何行动进行推理、与其他行动主体互动、制定如何行动的计划、评估过去的行为、为我们的行为负责等的复杂能力"[5]31。按照该要求,即便AI能做一些特定的自主行动、对如何行动进行推理,只要人类无法确认AI能为自己的行为负责,就不能认为AI具备行动主体性。

维特根斯坦进一步启发我们,道德地位、道德行动主体、责任与义务等诸道德概念也是与人的社会实践深度绑定的,没有人的社会实践和伦理生活,也就不会有相应的道德概念和伦理思想,如果忽略人的社会实践特别是伦理实践,就很容易导致伦理研究走向死胡同。例如,在西方学术界,"关于道德地位的讨论通常认为道德地位取决于属性,如在动物伦理中,雷根和辛格之间的辩论就是关于'什么样的属性是与道德有关的'"[18]。这类问题之所以没有终极答案,是因为道德不仅与人的特征属性相关,更与人的社会实践紧密相关。总之,人的社会实践、人的特征属性与人的道德概念紧密绑定在一起,这导致道德地位等概念是人类道德领域的专属概念,当一些研究者将这些道德概念随意用于AI时,这种应用只是体现了拟人化的思维方式,无论论证多么精巧,AI也很难具有真正的道德地位。以上讨论也表明,拟人化思维并不是很适合AI伦理研究,而整体主义思维更应该成为AI伦理这类严谨、科学的学术研究应该采取的思维框架和研究规范。

结 语

拟人化有其生物演化和社会文化基础,但人们如果不分场景地使用拟人化修辞和思维,会使得拟人化逐渐被滥用。AI伦理中的拟人化导致了诸多不合理的理论问题和理论观点,AI的道德地位问题就是代表性的案例。反思AI的道德地位问题,不仅要求我们思考支持AI具有道德地位的理由有哪些不合理之处,更要求我们反思和批评隐藏在这些具体问题之下的拟人化修辞和思维。在维特根斯坦的后期哲学视野下,我们可以更为清楚地认识到:一是AI伦理研究中的拟人化修辞常常是对语言的误用,这会导致一系列可疑的哲学伦理问题和观点被生产出来,处理这些以拟人化方式构造的问题和观点,要求我们从思维方式上去质疑其合理性,甚至只有解构了虚构的伪问题,真正的AI伦理问题才能得到更多的关注和研究。二是我们要根据具体的情境来区分合理的拟人化和有害的拟人化。AI伦理研究中的拟人化之所以经常带来危害,核心原因是这项学术研究需要充分的论证和扎实的理由,将AI拟人化反而有损这项研究的科学性。三是在AI伦理研究中,整体主义思维有助于抵御拟人化思维的惯性,当我们深刻认识到人类的道德概念与道德能力、道德实践的紧密联系时,就会明白试图证明AI具有道德地位的关系转向论证和行为主义论证是多么肤浅。从维特根斯坦哲学审视出发反思AI伦理研究中的拟人化现象,可以看到,经典哲学思想与AI伦理研究的结合具有重要的理论价值,既有助于思考和诊疗前沿问题,也有助于让经典哲学思想在AI时代焕发生命力。

反思AI伦理研究中的拟人化之弊,也为我们回答一个大问题提供了新视角,这就是:AI伦理研究应该更关切AI的伦理处境还是人的伦理处境;应该抽象地、拟人化地讨论AI道德地位、道德行动主体、权利义务等问题,还是以人为中心,关心AI对人类的影响以及人在制造、使用AI时带来的伦理与治理

问题。从拟人化视角看,前一类问题的形成与回答常常深受拟人化思维的影响,相关研究甚至逐渐成为某些学者建构的学术乐园。为了避免拟人化思维将 AI 伦理研究引入困境,AI 伦理研究应更关注 AI 技术对人类的社会生活、生存处境和未来发展带来的实质性影响,应该以人而非机器为本位,关切技术如何影响人以及人如何更规范地使用技术。只有这样,我们才能克制拟人化的冲动,在没有明确事实证据的情况下不将人的道德特质赋予 AI,以免制造出一系列伪问题并浪费宝贵的研究资源。通过剖析将 AI 拟人化导致的伪问题和混乱观点,我们进一步认识到:AI 伦理虽然主要研究关于 AI 各方面的规范问题,但该研究本身也应有其规范,如谨慎使用拟人化应当成为 AI 伦理研究中的重要规范,这包括在易引发误解的问题上要对使用拟人化做说明,在关键论证和结论上尽可能不使用拟人化等。以拟人化问题为起点,未来 AI 伦理的研究规范问题值得更深入的探索。

注释

①需要说明的是,有些学者在其研究中没有区分“将 AI 拟人化”与凭借各种新兴技术让 AI 类人化、让 AI 模仿人,这会导致他们在谈论“将 AI 拟人化”时包含两种不同的意思,容易产生思想上的混乱。笔者认为,前者更强调将人类特征主观赋予 AI,而后者是一个 AI 设计的技术实践问题,两者并非同一个问题。②维森特·穆勒对道德地位做了进一步界定,一个实体有道德地位意味着该实体是一个道德行动者或道德受动者。

参考文献

[1]中共中央办公厅　国务院办公厅印发《关于加强科技伦理治理的意见》[EB/OL].(2022-03-20)[2025-03-20].https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.

[2]Nicholas Epley, Adam Waytz, John T Cacioppo. On Seeing Human: A Three-factor Theory of Anthropomorphism[J]. Psychological Review,2007(114):864-886.

[3]Rebekah Humphreys. Animals, Ethics, and Language: The Philosophy of Meaningful Communication in the Lives of Animals[M]. London: Palgrave Macmillan, 2023:6.

[4]内格尔.本然的观点[M].贾可春,译.北京:中国人民大学出版社,2010:4.

[5]Sven Nyholm. Humans and Robots: Ethics, Agency, and Anthropomorphism[M]. London: Rowman & Littlefield, 2020.

[6]赫拉利.智人之上:从石器时代到 AI 时代的信息网络简史[M].林俊宏,译.北京:中信出版集团,2024.

[7]Vincent C Müller. Is It Time for Robot Rights? Moral Status in Artificial Entities[J]. Ethics and Information Technology, 2021 (23):579-587.

[8]Mark Coeckelbergh. AI Ethics[M]. Cambridge, MA: The MIT Press, 2020:59.

[9]Mark Coeckelbergh. Virtual Moral Agency, Virtual Moral Responsibility: on the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents[J]. AI & Society, 2009(24):181-189.

[10]Adriana Placani. Anthropomorphism in AI: Hype and Fallacy[J]. AI and Ethics, 2024 (4):691-698.

[11]Arleen Salles, Kathinka Evers, Michele Farisco. Anthropomorphism in AI[J]. AJOB Neuroscience, 2020 (11):88-95.

[12]维特根斯坦.哲学语法[M]//维特根斯坦文集:第 3 卷.韩林合,译.北京:商务印书馆,2019:28.

[13]维特根斯坦.哲学研究[M].陈嘉映,译.上海:上海人民出版社,2005.

[14]维特根斯坦.心理学哲学笔记[M]//维特根斯坦文集:第 7 卷.张励耕,译.北京:商务印书馆,2019:6-12.

[15]Fredrik Karlsson. Critical Anthropomorphism and Animal Ethics[J]. J Agric Environ Ethics, 2012(25):707-720.

[16]康德.康德著作全集:第 5 卷[M].李秋零,译.北京:中国人民大学出版社,2007:5.

[17]恩格斯.自然辩证法[M].北京:人民出版社, 2015:306.

[18]苏令银.当前国外机器人伦理研究综述[J].新疆师范大学学报(哲学社会科学版), 2019(1):105-122.

The Pitfalls of Anthropomorphism in AI Ethics Research and Its Philosophical Reflections

Ruan Kai

Abstract: Anthropomorphic narratives have grown increasingly prevalent in humanities and social science research on artificial intelligence. Anthropomorphism refers to the subjective projection or attribution of human characteristics—especially mental and social traits—onto non-human entities, rooted in humanity's inherent difficulty in transcending an anthropocentric perspective. A growing number of scholars analyze AI ethical issues through anthropomorphic thinking, which undermines the rigor, objectivity and credibility of relevant research to a certain extent. This necessitates in-depth exploration of normative standards for AI ethics research: whether anthropomorphic thinking and narratives are suitable for AI ethics research, and where the boundary of anthropomorphism lies in such research. Wittgenstein's philosophy provides a comprehensive framework to examine and overcome the pitfalls of anthropomorphism in AI ethics. At the linguistic level, anthropomorphic rhetoric substantially conflicts with rigorous, pragmatic academic research norms. At the thinking level, Wittgenstein's holistic philosophy reminds researchers to cautiously avoid mechanically transplanting human moral concepts and judgments onto artificial intelligence. Exploring the cognitive foundations and potential harms of anthropomorphizing AI can offer theoretical insights for establishing standardized norms for AI ethics research.

Key words: AI ethics; anthropomorphism of artificial intelligence; moral status of AI; Wittgenstein; holistic thinking

责任编辑:思　齐