机器主体道德意识能否产生:人工智能机器伦理治理的“内在进路”探析

谢静[1,2]

1. 黑龙江大学马克思主义学院,哈尔滨 150080
2. 黑龙江大学马克思主义理论博士后科研流动站,哈尔滨 150080

摘要 人工智能机器伦理治理的进路主要有“外在主义”和“内在主义”2种,“外在主义”强调以人类制定的伦理准则规约人工智能机器的发展,而“内在主义”则突出人工智能机器的自学习能力,以形成独立于人类的机器主体道德意识。通过介绍人工智能机器伦理与属人的伦理的本质区别,揭示了属人的伦理世界包含着的审美性意涵,即通往道德意识的美与美感无法被模拟与分层,形成道德意识的有机生命与审美判断力无法被还原,从而推理出人工智能机器的主体道德意识无法产生,得出在功利主义原则驱动下的人工智能机器伦理不能通往人类伦理意义世界的结论。人工智能的道德意向性是形式逻辑推演的结果,人工智能机器缺乏“具身”,不能从事感性实践活动并产生情感意识,其认知模式属于单一的还原论。因此,人工智能机器不能自主生成美善价值,故其治理的“内在进路”也是“外在主义”进路的延伸与扩展。

关键词 人工智能机器;内在进路;外在进路;美善价值

人工智能在为人类提供一种美好生活新形态的同时,又带来了诸多不确定性和未知风险,譬如数据茧房、算法歧视以及信息安全等问题。更引人深思的是人工智能与人类之间的伦理冲突问题,如失业、贫富差距、家庭伦理以及人类尊严等。西方马克思主义理论家齐泽克曾以“普罗米修斯的耻辱”来比喻人类与人工智能的关系。虽然人类创造了人工智能,但是随着科技的进步,人工智能算法体系日趋完善,甚至能够以自学习的方式产生主体意识,颠倒曾经的主客体之间关系,则有可能发展成为一种透明的、自律的意识实体,类似于自我运动的绝对精神,进而成为科技意识形态,与数字资

收稿日期:2023-03-08;修回日期:2024-04-01

基金项目:黑龙江省高校基本科研业务费黑龙江大学专项资金项目(2023-KYYWF-1535);黑龙江省哲学社会科学研究规划项目(青年项目)(23KSC119)

作者简介:谢静,讲师,研究方向为马克思主义基本原理,电子信箱:1014390789@qq.com

引用格式:谢静. 机器主体道德意识能否产生:人工智能机器伦理治理的“内在进路”探析[J]. 科技导报, 2024, 42(8): 120-128;
doi: 10.3981/j.issn.1000-7857.2023.03.00366

本的运行规律相互嵌入,操控人类社会,成为新的绝对大他者。其中暗示着一个核心问题,人工智能的伦理治理关键在于确定人工智能机器是否存在类似于人类认知的自我意识,使人工智能自身朝着伦理治理的“内在进路”健康发展。这涉及到的核心问题是机器意识能否和人的主体意识一样,根据社会实践自发地形成技术的道德意向性。总之,问题可以聚焦成一个,那就是人工智能能否产生机器主体道德意识,如果能的话,与人的智能之间的鸿沟[1]究竟有多远?

国内学者张卫将人工智能机器伦理的治理归纳为“外在主义”(externalism)和“内在主义”(internalism)2种进路,“其中,‘外在主义’进路主要关注的是人工智能的负面伦理价值,目的是用伦理来规约人工智能的发展;而‘内在主义’进路则主要关注的是人工智能的正面伦理价值,目标是用人工智能来解决伦理问题”[2]。具体而言,“外在主义”进路对人工智能伦理问题的治理,主要依靠立法方面的伦理调适,即出台法律法规对技术的发展进行监管,如中国的《新一代人工智能原则:发展负责任的人工智能》、美国的《人工智能原则:国防部应用人工智能伦理的若干建议》、欧盟的《可信任人工智能伦理指南》等。而“内在主义”进路则主要通过人工智能的自学习功能习得与本国相适应的法律法规和伦理规范,解决人类社会的伦理问题。与“外在主义”进路不同的是,“内在主义”进路强调人工智能的自主性和能动性,利用机器意识与认知能力借以产生机器本身的道德意向性,成为“后人类”新主体,即“完全的伦理能动者”[3]。其中关键的问题在于,人工智能能否成为独立于人类之外的新主体获得自我意识,形成一种机器伦理,进而使人类社会归顺于机器意识形态?笔者认为不能而且也不应该。

1　伦理概念之外的人工智能机器伦理

探寻人工智能机器伦理的产生与机器意识形成机制之前,有必要从“伦理”本身出发,来思考属人的伦理与机器伦理是否属于同质性的存在。抑或是本来诞生于人的伦理世界的机器伦理为何会成为像比尔·乔伊(Bill Joy)所担忧的那样,会“超出人类集体的价值观、伦理和道德”[4],机器即使有了意识,成为超强人工智能,能否自发生成属人的伦理范畴或意义世界?

1.1　属人的伦理

一般而言,伦理涉及社会成员在共同的政治生活中,在共同的“善”的驱使下形成具有一定普遍性的伦理规则和道德规范。实际上,日常生活中的伦理概念是建立在不同的利益需求基础上,按照共同政治生活的日常轨迹所形成的习惯性原则而确立的基本规范与秩序。伦理包含了2个向度:一是理想维度,可上升到形而上学,更倾向于政治生活中的理性追求和价值判断,隐含着向美的境界“趋近”与“引渡”以及对现实的超越的意涵;二是现实维度,更强调对现行制度的遵循,对现有政策层面的合理拥护,内隐着制约与规训的意蕴。而人工智能伦理最多能达到这个现实维度,更符合前文所提到的“外在进路”,更倾向于政治或制度本身对人工智能的外在规约与管束。

人工智能机器伦理与属人的伦理是否有本质区别?人工智能机器能否通往“彻底的合伦理”的生活?为了回答这些问题,需要回到亚里士多德的实践哲学中探讨“伦理”“道德”“政治”产生的原初语境。在亚里士多德的原初语境中,伦理活动隶属于道德活动,以寻求一种自我圆满的、自我规定的、自含意义的、目的在内的最高善为目的,而这种追求恰恰是目的本身,只不过是在一定的人际关系中去求得一种与他人关系的“完满”和美好的生活境界,是对整个美好生活之“善”和“美”的趋向过程,而这种趋向是既讲动机又讲效果的实践,这种动机就是“明智”。人在关系中追求“完满”的和谐,说明人是政治的动物[5]。而人生存的社会不只是由父母、妻子、儿女这种建立在血缘关系基础上的亲缘社会,还是和非亲缘的其他人所组成的城邦社会,因此政治的生活也是按照“自足”的、完满的“善”所导引而成的和谐生活。政治活动所追求的“善”又是较为具体的、借助于具体的善(德行)的原则,需要人在总体性的“善”的指引下,根据不同的、处于

变动中的情境进行适度选择,他称之为“中道”[5]。这种选择的能力,不只是从事政治活动所需要的,也是道德和伦理活动所需要的,这就是“明智”,是实践的理性,也称为实践智慧。“明智”是一种对整个美好生活的“善”的考虑和策划。因此,在伦理概念的原初哲学语境中,伦理就成了需要经由政治活动、借助于“明智”来实现对善好生活的选择的、具有超越属性的意义世界。回到亚里士多德关于伦理的原初语境中会发现,人工智能伦理的“外在进路”充其量能做到对现有政治制度或管理规范的逻辑展开、理论沿革、细节章法、法律适用的解读与遵循,仍困于“实然”领域,无法达到“伦理”或者是“德性”的境界。

1.2 无法通达属人的伦理的“外在进路”

回到亚里士多德的实践哲学中,探讨“伦理”“道德”“政治”产生的原初语境,会发现道德、伦理与政治不可分割。人工智能机器伦理的“外在进路”更倾向于城邦伦理政治生活中对现存的政治制度本身的遵守与服从,通过让渡个体的自由而换取城邦中所有人的自由。因为“外在进路”的机制是在给定的人类拟制的伦理框架下进行活动,而不能通过主体性的社会活动去“考虑”这个制度本身是否合理,更不能自发地、自觉地在由各个历史阶段所积淀下来的实践中改变、策划、修改与超越给定的框架。因此,达不到对“完满”的终极善的趋近与向往,也就是缺乏“明智”,而只有理智,失去了对实现完满的终极“至善”的超验之维以及对澄明的审美之境的憧憬。因此,这种进路下产生的人工智能机器伦理达不到亚里士多德意义上的伦理与道德,充其量是把伦理框架当作“至善”本身,却忽略了亚里士多德的 “至善”是模糊的,至真至纯的形式本身是谁都无法预料或企及的,实践活动只能是“爱”智慧的过程。

外在进路中的伦理框架是由确定不移的编码编制而成的,这个编码最终导向的是确定的、具体的、局部的、片面的“善”,不是“无用之用”,而是“小善”,是目的在外的,依赖于科学的“真”的认识论模式,而这也是亚氏所排除在道德、伦理活动之外的创制活动。这个外在的伦理框架所给出的伦理构成方式是自上而下的遵循,并不是亚里士多德所谈及的由“终极”意义上更合乎“德性”的趋近与生成,是单向度的类似于马尔库塞所说的“压抑性的道德”,不是向和谐人际关系与美好生活的、自下而上生成的“好”的道德,或多维性的、否定性的具有批判与建构意义上的“新”道德,最终导向了功利主义原则下的多种行为结果中的“最优解”,取缔了伦理的审美性意涵。

1.3 无法产生伦理意识的“内在进路”

人工智能伦理意识诞生的“内在进路”才是真正意义上的“伦理道德”,但属人的伦理所趋向的“应然”状态是人工智能机器无法企及的,而且也不能确定人工智能机器是否愿意“试图”或“筹谋”进入这个属人的意义世界。很难确定人工智能机器是否有超越自身功能属性的憧憬与期待,或者是康德意义上的“目的”,这需要探求机器主体意识的产生机制,方可略知一二。

机器意识属于弱人工智能阶段的产物,也就是说,从现有的科技论断和实验研究来看,人类还无法制造出完全具有自主意识的智能机器,机器的自主意识尚处于缺失状态。弱人工智能所遵循的是具体的、人类提前设定并植入于芯片中的法则信息,目的是完成某项特定的任务。但从科技的前瞻性来讲,人类绝不满足于使人工智能仅停留于此阶段,而是致力于研发出有意识的人工智能机器,使机器自发产生主体意识。机器意识是人工智能领域研究的一项专门课题,主要议题是如何让机器像人一样具有自主意识,这显然是一项重大的前沿科学技术工程研究,其中具有深刻的哲学意涵,无法仅从科学技术角度去探求。从哲学意义上来说,机器意识是在科技哲学理论基础上研究如何利用机器理解意识现象,由此推之,机器主体道德意识是研究如何使机器内生出道德选择进而做出道德行为并成为道德主体的哲学理论。近年来,人工智能领域内的机器意识研究中,科学家们往往将目光聚焦到如何在机器内复刻人脑,因此建立了一系列关于意识现象的模型。例如帕斯夸里和蒂默曼斯等的元认知意识模型,认为意识的产生在于大脑能否意识到自己已知的内容,类似于反思行为。那么如

何在机器中建立具有反思功能的大脑呢？关键在于实现低阶信息和高阶信息的交互,而这种信息的交流出自任务的需要,低阶信息是关于对世界的了解,而高阶信息则在于对这个世界进行表征[6]。如此一来,机器主体道德意识产生的关键就在于在任务的驱动下,对生活世界的信息进行低阶搜索,通过自下而上的控制流程,运用高阶知识,在伦理规则的框架内完成与低阶信息的往来互通。再如费科特发现神经元是复刻人脑的关键,只有依靠神经元方能激活大脑中的“感受质”,与此相关的数据才能产生出“感受质”来刺激神经元活动[7]。北村(T. K. Itamura)还注意到无意识层面在自我意识生成中的重要作用,虽然无意识认知无法用语言表述,但可以运用符号来表征无意识信息,这样能够区分机器人的意识层次结构,如意向性、预期、反馈体现和情感[8]。虽然这些人工智能机器的意识模型得到比较充分的论证,但目前为止仍然没有与人类完全一致的人工智能机器问世。因为思维的形成属于生物过程,仅依靠符号或信息进行表征也仅停留在模拟生物过程层面,即使通过生物酶的复刻,也无法形成具身性的肌体,因此生物过程的结果——“意识”也无法产生。而且意识并不仅是大脑这个物质器官或其内部局部神经物理机制的产物,即使将大脑复刻后植入人工智能机器内部,也不能形成“自我意识”。在马克思看来,“人类经历政治生活,但必定要超越政治才能实现人的最终解放。实现人的彻底的有道德的、合伦理的生活。”[9]也就是说,必须在属人的政治生活中,进行社会性交往活动,这是通往伦理这一意义世界的起点。

通过“内在进路”而形成的机器伦理意识无法产生,原因在于就算是给定了伦理原则与框架,机器仅凭借算法、程序,甚至是学习、模拟或者还原人类的神经元系统与复刻人脑,但是不能理解为什么人类能够在不同的情境下选择放弃、让渡个人权利甚至是个人生命的“心甘情愿”,以成全他人的幸福。人生命本能的直接升华与压抑性反升华,是复杂的过程,不只是依赖于感性“质料”,能够凭借一种与他人共在的善,甚至是纯粹的善的纯美指引而直接上升为情感意志的选择。这与亚氏的“明智”十分相似,却又有所不同,而这个不同恰在于亚氏所要抑制的合乎“欲望”的逻各斯,有时这种欲望超越了作为生物属性的力比多性生欲或死欲本身,而是在生欲与死欲交织的压抑性反升华中走向了非压抑的升华,从而诞生彻底合乎人性的“好”的道德。这恰恰是人工智能机器伦理意识无法复刻的。

2　人工智能机器伦理治理的“内在进路”为何行不通

即使在外赋予机器意识以伦理框架,但是机器却不能进行“明智”的选择或抉择,仅把感性的“质料”进行编码形成各种符号,但无法洞察符号中的隐喻、转喻的丰富的感性存在物。如《流浪地球2》中的男主角作为宇航员在获得通往地下城名额后的面试中,机器智能系统特别提示他不能使用隐喻、反问、暗示等方式回答,这说明即使是给流浪地球计划自主地设置难关的机器智能系统,也仅限于将被面试者及其家人的基本信息集中起来,以科学的求真认知模式为受试者提供“最优解”,而无法知晓受试者在面对面试问题时表象世界中的感性资料以及在回忆情境中的情绪与百转千回之后的意志抉择,无法恢复言语和语言中的否定性张力。这说明,无论是“外在进路”还是“内在进路”,如果人工智能机器意识仅达到“智能”水平,也就是通过转码、编码或复刻还原人的认知系统和人脑机能,仅能将想象力、幻想、冲动、情感限制在认知领域,即康德的纯粹理性批判层面,而无法将它们还原到判断力批判层面,进而无法进入实践理性批判层面。如此一来,无论是“外在进路”还是“内在进路”,机器伦理意识无法产生,即使产生也仅限于寻求“最优解”的功利主义道德领域,无法通达属人的伦理世界。

2.1　通往道德意识的美与美感无法被模拟与分层

根据对人工智能机器意识的最新研究进展,心灵哲学和意识的意向性成为热门的研究领域,将不同意识状态的各部分和各层次进行科学而严格的区分[10]是很多研究人工智能科学家需要应对的首要问题,但人们似乎习惯了将意识进行分层本身来

复刻人脑,把人的自我意识局限在生物学、神经学、认知科学、心理学、物理学等科学领域学科。实际上,自我意识属于哲学领域范畴[11]。如果仅在科学领域研究机器主体道德意识,一定是将人的认知能力和理性能力看作类似于康德所说的“先验统觉”,也就是变相承认了康德的知、情、意三分的先验意识结构。

然而,属人的伦理意识与美和美感有着密切的关联,康德就是在自由意义上是将美视作道德的象征[12]。在康德看来,美本身蕴含着超越形象、质料与感性认知的限制,从而进入无限理想之境的自由精神,而人只有在从事道德实践的过程中才能突破认识形式的局限,通达象征着无限自由领域的理想之维。康德看到了美之境界的无涉功利性,这就回到了亚里士多德的道德实践。美和美感象征着自由精神,是感性的人在感性的生存中产生了感性的需要,感性的需要不仅包括自然性的主体需要,而且还满足于社会性主体的需要,换言之,美善的道德是在首先满足有生命的个人的存在这一历史前提中产生的,而向往的是不被私有财产异化的“具有人的本质的这种全部丰富性的人”[13]的共产主义终极伦理关怀。而这种对“丰富的、全面而深刻的感觉的人作为这个社会的恒久的现实”[13]的向往恰恰是在来自“对攻击性和罪恶的超升”[14]以及对“在社会的范围内对充满生命的需求”[14]的爱欲冲动下产生的。对“感性、娱乐、安宁和美”[14]的社会的憧憬需要审美判断力方能实现“新感性”和“新理性”的统一,而审美判断力恰恰是不遵循既定的理性法则和知性法则的心灵、灵感、想象力等感性能力,需要在“具身”性的、整体性的混沌体验中将独创性的、未被觉知的、微小的、联结着无意识的感性因素的意向内容直接升华,通达道德的伦理之境。

也就是说,往往意识科学无法说明这种升华依靠的是什么形式。人脑只是意识产生的物质器官,意识是人脑的机能,但审美与道德的通路并非单纯依赖“智能”性的人脑产生,而是马克思所说的在有生命的个人生存中的通过劳动及其关系历史性地生成的“实践感”。这种实践感来源于纯然的赤子与婴儿状态的感性生存,是一种“混沌”的“绽出”,是单一的可分层的认知系统所无法理解的。因此,在美与美感作为“去远”化冲动的道德实践中,将意识进行内部分层,模拟人脑,将神经元系统加装在智能芯片当中,也无法让机器自发地或者“自愿自觉自由”地产生道德意识。

2.2 形成道德意识的有机生命与审美判断力无法被还原

加梅兹(David Gamez)认为机器意识的产生经历4个阶段,分别产生了MC1、MC2、MC3、MC4这4种类型机器意识[15]。MC1机器意识可以复制人类的有意识行为,复刻人的自主学习与模仿行为,但无法进入人的无意识领域。MC2是通过模拟人类的神经元系统,在意识和类似于想象、情感之间建立多重联系的机器意识,把人类复杂的、社会性的心理现象和生理现象还原为单一的物理性的生物行为,将想象力和判断力当作人的认知特征。MC3可以根据人的交往、劳动等需要,与其他机器实现机机交往从而进行数据交换,架构一个巴尔斯所说的“全局工作空间”。虽然加梅兹认识到人类意识产生的根源在社会性的实践活动,但将机器之间的交互行为当作无人身的理性自我繁衍的“一般智力”的产生行为,忽视了人的意义世界的建构是在社会化的对象性劳动创造中产生的。MC4是真实现象体验的机器,但这种“体感”是在人的体感基础上产生的,以人的身体感觉作为数据和信息的源泉。如科幻电视剧《三体》中的史强警官和冯森进入的称作“三体人的世界”的体感游戏,是通过游戏中虚拟的场景给予视觉与听觉的刺激,以此来激发他们的联觉系统,让自己的身体感受到冷热,而这种人类身体所产生的冷热的感觉又通过传感装置重新进入游戏系统成为新的数据信息,机器意识根据不断收集的新的信息再生出更丰富的体感场景,从而达到人机的交互。也就是说,离开了人的真实身体,机器意识赖以产生的信息数据无法产生与被收集,机器也仅仅是人的无机身体。即使MC1—MC4的产生貌似实现了弱人工智能到强人工智能的质性发展,但试图从“内在进路”入手的类人性的机器主体道德意识也无法产生。

我们可以从作为生命有机体的自然属性、关乎

美的感受力与判断力等角度来回答为何从“内在进路”生成的机器主体道德意识无法产生。在“善”的问题上,康德有2条思路,一是有用性,能够作用于某一目的就是“好的”,二是道德上“好的”行为。在“有用性”的世界里,人是“手段”,都是为了实现某一目的,都是为了实现善的法则的“手段”。但是在这个“善”的世界里,“至善”是人生的终极目的,人只能朝着这个目的运动但无法实现,这就是道德的世界,人始终对道德善怀有“憧憬”与“崇拜”之情,只有这样善的世界才能成为不压抑人的“人的世界”。而席勒把美与道德相联系,说明好的道德是既不压抑人的自然本能也不损害他人自由的行为的善。而人作为自然有机体的生命体,经营着自己的生命,这个过程就是“目的”本身。人的有机生命赋予了人以生物人性,那就是“性欲”,弗洛伊德将其称为“力比多”,深藏于人的无意识领域。但仅凭借性欲本能冲动的驱使,人无法成为“自为”的生物,而人的审美判断力能把来自身体本能的性欲转化为“爱欲”,真正无涉功利的纯美“至善”是要将爱欲本能所遵循的快乐原则经过“自我”的“异化”,并扬弃“本我”到“自我”阶段所接受的异化,才能进入“爱欲”的升华——“超我”,这时快乐原则才是真正超越性欲之“原罪”与“恶”的“新的现实原则”。这说明人的深层心理结构决定人的道德认知和道德生成。

机器不是有机生命体,缺失生命本能,更无法自主进入无意识结构深处,无法产生既不属于“悟性”也不属于“理性”的“判断力”,即使给定拟制好的伦理框架也无法形成“反省的能力”。“判断力”是“直感判断力”,与“判断”不同。“判断”是把“感觉”所接受的“现象”适用于“识别”规则的机能,机器智能所具有的能力属于此类别。而“判断力”则是“判定”适用于“规则”的方法是否得当的能力,是一种“直觉”,连通着想象力,作为边“自然”地“假定”“自由”边理解“自然”的能力,能够使人的力比多内化为“爱欲”,与对至善的追求的结合而成为“意志”的努力,从而内化为实现社会性的自由。“判断力”能将“想象力”和“识别”(悟性)调和不与任何特定的目的相结合,这一结合本身是生动的,是面向人的身体展开的,不能被还原为“智性”和“理性”的感觉,也不能被还原为纯粹的神经元系统内生的意识。即使是罗森塔尔·亚当斯也没有明确主张将不同的意识状态直接还原为不同的神经生理状态。因此,MC1和MC2机器意识无法实现有机的生命体由无意识的生物机能进入审美判断力的感性之维,将审美判断力所联通的想象力、情感等感性能力限制在认知领域,充其量达到“识别”与“判断”的程度。而“美,是靠‘想象力’与‘识别’(悟性)的调和的游动来决定的。这种心理状态满足了使认识作为可能的‘超越论’的条件。”[16]想象力能够将道德伦理中所牵涉的“理念”和“理想”这些抽象的概念,与获得“感动”和“刺激”的“质料”经由判断力这种“心灵的能力”转化为面向身体、符合人伦形态的形象关系,而这种形象关系也需要靠想象力或情感能力与视觉关系、听觉关系、语言的关系以及体态关系构成一个磁场,这个磁场就是“直感理念”,成为认识“善”“爱”“正义”这些“理性理念”的“隐蔽条件”,形成“审美共通感”的普遍法则,道德伦理的“普遍传达的可能性”方可建立。目前的人工智能机器伦理意识尚无法形成“直感理念”,且不具有生成它的有机生命,无法通过“无目的的合目的”的审美想象形成对“秩序”的“趣味”性的整理。

3　人工智能机器伦理意识的本质

人工智能机器伦理意识产生的“内在进路”研究离不开人工智能的道德意向性研究,而机器智能的道德意向性的展开与人性化的伦理世界的生成过程截然不同,这也揭示了人工智能机器伦理意识的本质。

3.1　形式逻辑推演的结果

人工智能的道德意向性并不是自主产生的,而是其设计、研发和维护者在其芯片中植入数字代码和算法公式,使人工智能利用科技前沿技术与传感交互系统模拟出与人类似的大脑智能,对外界数据和信息进行分析、收集以及处理、反馈。其自主学习能力不过是利用已经植入的算法公式对外部环境给出的数据信息进行逻辑推演出的结果,这充其

量是一种去除掉具身的、非神经性的条件反射,其推演的逻辑至多是形式逻辑,缺乏人类的感性思维能力。马克思说,人类历史的第一个前提是有生命的个体的存在,个体为了这个生命的存在而展开生殖、生产等各种活动,由此有了家庭活动、生产化活动、政治活动等。人为了维持肉体生命的存在,在实践活动中展开自己,然后展开自身的丰富性,即人能够“按照美的规律去生产”。只是由于人的本质的客观地展开的丰富性,主体的、人的感性才具有丰富性,如有音乐感的耳朵、能感受形式美的眼睛。那些能够成为人的享受的感觉,即确证自己是人的本质力量的感觉,才能一部分发展起来,一部分产生出来。因为不仅五官感觉,而且所谓精神感觉、实践感觉(意志、爱等),一句话,人的感觉、感觉的人性,都只是由于它的对象的存在,由于人化的自然界才产生出来的。五官感觉的形成是以往全部世界历史的总和。所以,人在与对象世界的客观性实践活动中产生自身的丰富性,产生美感与实践感觉,才能确证自身,获得自身的丰富性。而人工智能机器看上去拥有了人的感觉和情感意识,但它不具备实践的具身,虽然人工智能拥有传感器组件、智能控制和驱动效应器,但它将外界刺激都以数据信息的形式收集起来进行处理与编程,并不像人类那样基于实践将外部世界当作对象化的客体去认识与加以改造,因此它不能产生和人类一致的情感与情绪。因此,人工智能机器无法将实现自由作为自己的价值目标。

3.2 还原主义认知模式

人工智能机器与人类的道德意向性从认识论方面来看是完全不同的2种模式,其中人工智能机器的认知系统遵循的是还原主义原则,而人类智能的认知原则属于整体主义。马克思认为对于人的本质进行归纳要将人始终放在一定的社会历史环境和文化氛围中进行审视,因此人不只是生物学意义上的人,还是一定社会关系的总和,具有文化性格。故人类实践是在“感觉、思想、动机、意志”的共同加工下将外部世界变成“理想的意图”,进而形成“理想的力量”[17]。可见,马克思强调了人是在现实性的感性活动中获得整体性的认知的。而这种整体性认知的获得是不仅要依靠人的感觉器官,还不能脱离理性思维能力。因此,人的智能是在多种感官的整体协调与配合之下,获得对客观对象的属性的感性印象和感性片段,并在五感的互动基础上,“接受所有这些不同的感性印象,对它们进行加工,从而把它们综合为一个整体”[18],而这个整体并不仅仅是依靠人工智能的传感器组件和单一的类神经元结构所能完成的结果,而且感性与理性能力并不完全等同于人工智能的接受与处理信息的能力,同时还包含着各种偶然的可能性和多样化的创造力。换言之,正是因为人类的不完美和偶然犯错,人才具有独特性,许多发明创造正是在人偶然的一次错误中产生的。而人工智能通过传感器组件接收的外部刺激或活动图像往往被还原成为单一的数据信息,再通过语义系统和数字算法进行加工与推演,形成指令,再做出反应,这与人类通过感官协作与理性思维能动地改造世界的主体性创造活动是完全不同的。人类在满足生存需要以后,会逐步朝着自由而全面发展的人的“类本质”这一价值理想构造自身,并将自己当作“一切社会关系的总和”。这说明,人的认知水平是随着不同阶段与不同程度的社会交往与实践活动而逐步加深与提高的。但人工智能的认知水平的提高是通过设计者与研发者根据不同的人类实践需要,逐步修改并更新升级信息数据处理系统的功能来实现的,因此人工智能根据指令做出行动与人的创造生成性的实践活动和社会交往活动是完全不同的。

3.3 “一般智力”法则下的非身体的数据交换

伦理道德是属人的意义世界的内容之一,这离不开对道德的“理解”。这是一种不同于知识论的理解方式,其实质是主体将知觉经验纳入一个世界图景的实践活动,是一种更高阶的认知成就。而这个世界图景的构建离不开身体在“在世关系”中的建构作用。正如美善道德的人伦形态离不开“心地善良、纯洁、坚强、文静”等身体的表现,人类也需要在具身性的整体体验中不断完善与拓展对意义世界的“理解”。正如梅洛·庞蒂这一代存在主义思想家所言,我们在世界上存在的意义,是身体构筑起来的,而我们作为人类生存的意义,也正是在身体的基础上呈现的。主打高阶认知的MC3和MC4型人工智能机器意识的发明者恰恰也考虑这一点,强

调身体基础上的真实现象体验、实际交往与劳动基础上的人-机与机-机交互,但这种身体并不是如同米歇尔·亨利所说的那种“在原初本质上,属于生存范围”的、返还其“主体性本身的范围”的身体[19]。真正的身体是在现实的、动态历史性的交往实践中生成的,而不是存在于赛博空间或互联网中的“由数据和算法组成集合的对象”[20],这个集合的对象是在编程(object-oriented programming)中通过符号语言创立的包含状态、行为、标识的集合的对象,是虚体,而不是身体。虚体与虚体之间的交往不同于现实的感性的人的实践,它所依赖的是数据包之间的数据交换关系,不再以真实的、发展中的肉身为基础,所依赖的环境是数字和编码之中的数据流,试图把任何生命体和非生命体进行数据化,形成掏空现实的人的生存的、以算法规则取代人的内在价值尺度的“一般智力”(general intellect)。如此一来,有“使用实践力量的人”[21]将不复存在,基于真实的人的需要及以人作为活动主体的“实践的真理”将不复存在。而对美好未来的理想追求的应然之维离不开能进入人心的、成为实践理性的实践能力和实践需要,正如马克思曾经说过的“思想本身不能实现什么东西”[21]一样,“一般智力”本身只能在众多行为中筛选出最优解,但无法将人的价值倾向转化为具有实践冲力的意识形式,这种意识形式是在人格心理、意志情感与实践意图和内在动机模式融合后而形成的“实践感”,是“世界的准身体意图”,它“对那些虽非有意却依然是系统的、虽非按目的来安排和组织却依然带有回顾性和目的性‘选择’具有导向作用”[22],呈现为“能够使用理性和想象力改变现有物质”的“创造性性格结构”[23],表现出一种人之为人的超越性与创造性的道德意向性,从而实现对经济基础和现实生活的实践改造。

4　结论

人工智能不能成为“后人类”的机器道德主体,因为人工智能不能实现“人与世界之间实际的相互作用”即实践,它只是人类智能的异化形式,或者说是人类智能的理想状态和最高形式,它既不能受到感性自然和生理因素的制约,也不能有独立的实践目的,它的行动或反应无论多么形式多样,但终究不过是对外部信息进行接收加工处理后的直观反应或机械反应。因此,人工智能不能形成类似于人的“美感”,更不能自主地产生“美善”动机。从更深层次的意义上来说,人工智能治理的道德伦理的“内在主义”进路,仍然是在“外在主义”的伦理价值嵌套下的加工与整理的机制,无论它如何实现自学习风俗文化与法律法规,也是对外界现有的、由人类创造与提供的数据信息统筹后的结果。既然人工智能机器无法生成内在意识,是不是意味着在人工智能机器的伦理治理问题上人们就束手无策?实则不然,我们可以在人工智能机器伦理治理的“外在进路”上继续探索。其实,人工智能机器伦理问题产生的根源并不是技术本身,根源在于它的开发者和使用者。如果说外在伦理框架的植入,让人工智能机器进行自主判断与自我道德选择,这在现有的或可预期的技术范畴内是无法实现的,但我们可以对人工智能机器的开发者和使用者的伦理和文化素养继续做出规定,让人工智能机器所执行的任务本身符合“善好”的原则,即为人服务,为人类的美好生活造福的原则,制定相关的伦理规范引导人工智能机器的发展方向。我们不能让人工智能机器去套用属人的伦理道德规范与约束,是因为人工智能技术还有一定的局限性,这就需要继续在国际范围内,召开全球人工智能伦理议题的会议或讨论等,积极推动全球人工智能技术的和谐开发与运用,继续推进人工智能伦理治理的理论研究,加强人工智能机器研发设计与投产制造过程中的立法监督,这样能够在事故发生前做好预案,降低事故发生的频率,并能做好事故后的科学归责,降低事故损害。

参考文献(References)

[1] 陈肖东. 预见人工智能与人的智能之鸿沟——基于马克思认识论对认知神经科学的解构[J]. 理论界, 2021(9): 22-28.

[2] 张卫. 人工智能伦理的两种进路及其关系[J]. 云南社会科学, 2021(5): 21-27.

[3] Moor J H. The nature, importance, and difficulty of machine ethics[J]. IEEE Intelligent Systems, 2006, 21(4): 18-21.
[4] 庞洁. 机器意识研究的演进道路及其意向性分析[J]. 网络安全和信息化, 2023(1): 14-17.
[5] 亚里士多德. 尼各马可伦理学[M]. 廖申白, 译. 北京: 商务印书馆, 2003: 18-19, 52-57.
[6] Pasquali A, Timmermans B, Cleeremans A. Know thyself: Metacognitive networks and measures of consciousness[J]. Cognition, 2010, 117(2): 182-190.
[7] Fekete T, Edelman S. Towards a computational theory of experience[J]. Consciousness and Cognition, 2011, 20 (3): 807-827.
[8] Kitamura T, Tahara T, Asami I. How can a robot have consciousness?[J]. Advanced Robotics, 2000, 14(4): 263-275.
[9] 陶艳华. 马克思政治伦理思想研究[M]. 北京: 人民出版社, 2009: 5.
[10] 赵泽林. 内部区分: 意识与机器意识的新进展及其哲学前瞻[J]. 哲学动态, 2021(8): 118-127.
[11] 赵汀阳. 人工智能的自我意识何以可能[J]. 自然辩证法通讯, 2019(1): 1-8.
[12] 刘晓男. 康德审美道德论与浪漫主义思潮[J]. 学术交流, 2018(7): 37-42.
[13] 马克思. 1844年经济学哲学手稿[M]. 北京: 人民出版社, 2000: 88.
[14] 赫伯特·马尔库塞. 审美之维[M]. 李小兵, 译. 桂林: 广西师范大学出版社, 2001.
[15] Gamez D. Human and machine consciousness[M]. Cambridge: Open Book Publishers, 2018.
[16] 岩城见一. 感性论——为了被开放的经验理论[M]. 北京: 商务印书馆, 2008: 149.
[17] 马克思恩格斯选集(第4卷)[M]. 北京: 人民出版社, 1995: 936.
[18] 马克思恩格斯选集(第3卷)[M]. 北京: 人民出版社, 2019: 923.
[19] Henry M. Philosophy and phenomenology of the body [M]. The Hague: Springer, 1975: 7-8.
[20] 蓝江. 一般数据、虚体与数字资本——历史唯物主义视域下的数字资本主义批判[M]. 南京: 江苏人民出版社, 2022: 99.
[21] 马克思恩格斯全集(第2卷)[M]. 北京: 人民出版社, 1957: 152.
[22] 皮埃尔·布迪厄. 实践感[M]. 蒋梓骅, 译. 南京: 译林出版社, 2003: 101.
[23] 弗洛姆. 追寻自我[M]. 苏娜, 安定, 译. 延吉: 延边大学出版社, 1987: 100.

Whether the machine subject moral consciousness can be generated: The "internal approach" of the artificial intelligence machine ethical governance

XIE Jing[1,2]

1. School of Marxism, Heilongjiang University, Harbin 150080, China
2. Postdoctoral Research Station of Marxist Theory, Heilongjiang University, Harbin 150080, China

Abstract There are two approaches to the ethical governance of AI machines: "externalism" and "internalism". "Externalism" emphasizes the development of AI machines based on the ethical norms formulated by humans, while "internalism" highlights the self-learning ability of AI machines to form the moral consciousness of machine subjects independent of humans. By introducing the essential differences between artificial intelligence machine ethics and human ethics, this paper reveals the aesthetic implications contained in the human ethical world, that is, the beauty and aesthetic sense leading to moral consciousness cannot be simulated and stratified, and the organic life and aesthetic judgment forming moral consciousness cannot be restored. The subject moral consciousness of AI machine cannot be produced and the AI machine ethics driven by the principle of utilitarianism cannot lead to the conclusion of the meaning world of human ethics. The moral intentionality of artificial intelligence is the result of formal logic deduction, artificial intelligence machine lacks "embodiment", cannot engage in perceptual practice activities nor produce emotional consciousness, and its cognitive mode belongs to a single reductionism. Therefore, artificial intelligence machines cannot generate beauty value independently, so the "internal approach" of its governance is also an extension and expansion of the "externalism" approach.

Keywords artificial intelligence machine; internal approach; external approach; beauty value

(责任编辑 王丽娜)