

人工道德能动性的伦理证成：传统观点到开放立场

曲 蓉

摘 要：传统观点将道德能动性解释为人独有的道德行为能力，并提出了道德能动性的完备条件。传统观点并未直接否定人工道德自主体和人工道德能动性的可能性，但它固守人类中心主义立场，对道德能动性的完备条件缺乏充分反思。事实上，人类以及非人类存在物在缺乏某些条件的情况下，仍具备道德能动性。站在开放立场上，道德能动性可以宽泛地理解为实体积极施加有效道德影响的能力，要求实体具备由其内部状态“引发”的道德行为、内在认知能力、处于道德规范约束下等条件。道德能动性既可以是自然生成的，也可以是通过人工建构以软硬件形态呈现的无意识道德能动性，还可以是通过人机交互表现出来的集体道德能动性，包括授权的道德能动性、扩展的道德能动性和分布的道德能动性等不同形态。

关键词：人工智能；道德能动性；人工道德能动性；意识；意向性

中图分类号：B829 **文献标识码：**A **文章编号：**1003-0751(2025)06-0118-09

当前，人工智能正处于蓬勃发展期，不仅在文本、音频、图像合成等方面展示出强大的推理、判断和学习能力，而且在交通、医疗、金融、农业等越来越多的领域实现了产业化应用。人工智能可以在没有人类直接干预的情况下独立实施机器行为，承担起许多原来属于人类的职责任务，甚至在特定领域逐步取代人类。人工智能在相关技术道德事件中发挥了关键性作用；但它是否具备道德能动性，能否成为像人类一样的道德自主体仍存在广泛争议。按照传统观点，道德能动性是人类独有的重要属性，是人类价值和尊严的重要标志。人工智能不具备道德能动性，它只能作为技术工具协助人类发挥道德能动性。传统观点的问题在于：固守人类中心主义立场，试图构建道德能动性的完备条件，对道德能动性的复杂性和概念弹性缺乏足够反思。人工智能的高速迭代创新要求人文社会科学站在开放立场上重新思考道德能动性的条件，探讨人工道德能动性得以成立的底线要求和可能形态。

道德能动性是正确的道德责任归属的重要前提。“在我们知道能动性是否参与以及何种能动性参与了某些事件的发生之前，试图让任何自主体承担责任都是徒劳的。”^[1]人工道德能动性的提出意味着承认人工智能逐步发展而来的、对道德承受体施加有效道德影响的能力，并根据人与人工智能对技术道德事件的参与程度和实际作用来分配相应的道德责任。本文拟通过对人工道德能动性的伦理证成，阐明技术道德事件中的因果责任关系，从而为人工智能的道德责任归属问题提供必要的理论依据。

一、道德能动性的传统观点

传统观点从行为哲学出发，提出“能动性的标准概念和标准理论”。按照标准解释，道德能动性是人的道德行动能力，特指人“实施意向性行为的能力”^[2]。具备道德能动性的人是道德自主体。“能动性是内在于行动者之中的能力，而对于行动者

收稿日期：2025-04-08

基金项目：国家社会科学基金一般项目“人工智能的责任伦理研究”（19BZX109）。

作者简介：曲蓉，女，哲学博士，南京信息工程大学马克思主义学院教授（江苏南京 210044）。

最基本的理解是,行动者即对周围世界起作用的人。”^[3]传统观点站在人类价值立场上提出了道德能动性的完备条件,并据此对人工道德能动性进行了深入分析。

观点一:人类独特性

传统观点主张,能动性特别是道德能动性是人类心灵独有的功能,是人类意识和实践的一种特性,是人类区别于其他存在物的重要标志。“这一能动性是我们人类的一个核心特征,也是使我们每个人都成为独有的和有价值的个体的一个核心特征。它也是整体性评价人类价值的核心。重要的是,这个能动性包括我们的道德能动性。这是我们不能失去的。”^[4]按此观点,人工道德能动性的提法根本不成立。

斯派洛(Robert Sparrow)和博丁顿(Paula Bodington)代表了这一观点的两种不同思考路线,他们分别从人类个体和整体角度解释道德能动性的独特性。斯派洛基于个人主义观点,指出道德困境和道德决策都是个人的,个人的生活史深切地影响了他所做出的道德决策。不存在超出个人之上的道德权威,而试图在某个道德主张基础上教导人工智能道德的尝试是不可能实现的^[5]。博丁顿则认为道德能动性与“我们是谁”的身份认同直接相关,人工道德能动性的存在将损害人类独特的道德地位。与此同时,人类道德能动性极度脆弱,人类倾向于放弃自身的道德能动性或任由道德能动性受到侵蚀,承认人工道德能动性将加剧这一过程。“对于人类来说,我们的能动性至关重要,这也对我们是谁而言至关重要:我们的能动性可能会失败,我们的道德能动性可能会崩溃。”^[4]作为一种技术工具,人工智能的发展方向应增强而非削弱人类的道德能动性。当然,博丁顿并未直接否定人工道德能动性的可能性,而是试图告诫人们人工道德能动性对人类构成的潜在威胁。上述观点将道德能动性与人类整体或个体的身份认同和道德地位联系起来,强调人类道德行为的独特价值,从而对人工道德能动性持反对态度。

观点二:特定心理状态

传统观点认为人的某种特定心理状态是构成道德能动性的一个必备条件。大多数观点将这种心理状态理解为意向性,主张意向性构成行为的原因,在行为中发挥着不同于自然必然性的因果作用。“‘能动性’一词用在更狭义的意义,用来表示意向行为的表现。”^[2]还有观点认为人的全部意识(consciousness)都是形成道德能动性的心理条件。

因此,人工道德能动性取决于计算机能否构建人工意向状态或人工意识。

传统观点认为,“意向性是指心理状态指向外在对象的能力”^[6],它与自然因果效力不同,能够解释人类自主体与事件之间的因果关系,是道德或法律责任归属的前提条件。正如有学者所指出的,“在责任方面,这种区别是基本的,因为人类自主体可以被认为是道德和法律上对他们的行为负责(可以问责,有义务,值得赞扬/谴责),而人工物则不能”^[7]。传统观点内部对意向性的理解存在较大分歧。一种观点将意向性理解为信念/欲望组合,认为意向性是一种“关于其他事物的心理状态,如对X的欲求”^[8],信念即相信Y是实现X的必要手段,这一欲望和信念组合将引发一个实施Y的行为。另一种观点认为,信念/欲望组合停留在指向目标和行动的一阶欲望上,而对人的动机的反思性评价导致二阶欲望的形成。二阶欲望是认同一阶欲望及其效果,“拥有自由意志就是拥有二阶欲望”^[9]。如鲍尔斯(Thomas M. Powers)用慎思、实践理性和自由意志扩展了对意向性的理解。还有一种观点认为,意向性行为与有理由(reasons)行为之间有密切联系,意向性行为是出于理由的行为。“一个行为者是指出于理由而采取行动的人或事,这些理由可能受外部因素影响但却仍然是行为者的理由。”^[10]鲍尔斯进一步指出道德自主体应具备作为行动理由的内在意向状态。只有意向性成为其行为的理由时,一个人才能被视为自主体并具有道德能动性。他进而主张人类道德能动性的核心概念可以扩展至计算机道德的考虑之中,作为判断人工道德能动性的依据。

希玛(Kenneth Einar Himma)认为,构成道德能动性的必要且充分条件是人的两种能力:自由行为能力和道德认知能力。其中,自由行为能力意味着一个人的行为出于他自己的决定,而决定又离不开这个人的慎思能力,尽管并不是所有决定都是慎思的结果。道德认知能力是“知道对错的区分”^[8],更准确地说,是正确理解并应用道德原则进行道德推理的能力,而非许多人所理解的道德无误能力。这两种能力包括意向性都以意识为前提。据此,人工道德能动性是否可能取决于人工智能是否具有或产生意识,而按照目前的技术发展状况难以得出最终结论。

观点三:道德责任能力

传统观点认为问责或为行为负责的能力是道德

能动性的又一必备条件。“道德能动性意味着自主体可以对其行为负责,也就是说它可以被追究责任。”^[11]问责通常包括两类不同的责任实践:从后向来看,问责是使实体得其所应得;从前向来看,问责是使实体能够反思自身的选择和行为,不断发展自身的道德品质。希玛和阿桑罗(Peter M. Asaro)分别从上述两个向度进行了论证。希玛认为道德能动性是一个规范性概念,要求道德自主体在道德规范的约束下,为其所做出的行为负责,接受来自他人的赞扬、奖励、谴责、惩罚。而后者又要求自主体拥有“体验如骄傲和羞耻等意识状态”的能力。“从概念上,奖励或惩罚没有意识的东西是不可能的。”^[8]阿桑罗认为道德能动性离不开接受惩罚的能力。“道德能动性可能有多种定义,但它最终必须成为受惩罚的主体。”^[12]接受惩罚意味着道德自主体的道德品质能够得以发展和纠正。接受惩罚进一步要求道德自主体具有反思能力,能够将自己的潜在选择和行为与其他受到惩罚的选择和行为联系起来,使惩罚发挥其威慑作用。事实上,无论从前向抑或从后向,负责任能力的观点再次将问题聚焦于人工意向性或人工意识的构建问题上。这样看来,人工道德能动性的可能性似乎不再是一个哲学问题,而是一个技术问题。此外,还有观点指出为行为负责的能力需具备可以接受惩罚的身体,当然,后者是有机体特有的抑或是任何实体都具有的,同样存在争议。

二、对传统道德能动性的反思

从传统观点出发并不必然得出对人工道德能动性的否定性结论,相反,鲍尔斯等学者还指出构造道德机器大有希望,前提是计算机能否模拟或实现人工意识或人工意向性;但与此同时,传统观点固守人类价值立场,对道德能动性的完备条件和研究方法缺乏充分反思。

传统观点将道德能动性与人类的道德地位和生命价值联系起来,将人类与非人类实体在道德地位上根本对立起来,进而认为承认非人类实体的道德能动性将损害人类的道德地位。但是,将人类与技术人工物等非人类实体所属的群体类别作为区分道德能动性和道德地位的界标不仅面临诸多现实挑战,而且代表了一种落后的、带有歧视性的价值立场,即将群体的某些差异进行价值固化。正如人类历史上,那些因性别、肤色、国籍曾被排除在道德共

同体之外的群体最终获得了道德行为资格,其他非人类实体也不应因所属的群体而被直接否定其道德能动性和道德地位。更为重要的是,道德地位包括道德自主体地位和道德承受体地位,前者强调实体在道德活动中的主动地位,而后者强调实体在道德活动中的被动地位^[13]。从狭义上讲,道德地位特指后者,即实体的道德重要性或者说实体因其内在利益而具有道德上的重要价值。“如果一个实体具有道德地位,那么我们就不能随意对待它;我们在道德上有义务在我们的审议中重视其需要、利益或福祉。”^[13]承认人工道德能动性只是肯定技术人工物所具有的道德功能能力及其在道德活动中发挥的主动作用,它既不会威胁人类的内在道德价值,也不会损害人类的道德能动性。此外,人类道德能动性确有其脆弱性,但发展人类道德能动性的关键在于加强道德教育和道德修养,提高人类理性思维能力、道德选择能力和道德责任能力,而非错误性地减少其竞争对象。

传统观点提出道德能动性的完备条件,但这两个条件并非牢不可破。

一方面,特定心理状态是特指意向性还是包括意向性在内的全部意识,仍存在着争论。由于不同理论对意向性的内涵不断修正,意向性和道德能动性都成了具有高度弹性的概念。如果将意向性理解为信念/欲望组合的话,就无法将人的本能冲动与意向性行为区分开来。如果进而用自由意志来解释意向性,又可能陷入一种道德洁癖的误区。绝对命令是自由意志的法则,“绝对命令是道德能动性的构成性原则。为了采取行动,一个人必须以普遍规律为指导,这是一致性的行动原则。只要一个人的行为方式不那么一致,她就不是自主体”^[9]。该观点的问题是将偶然性行为或不道德行为排除在意向性行为之外。这显然不符合常识,因为我们的很多行为不是绝对前后一致或严格遵守道德原则的,但这些行为仍然是我们基于自身行为能力所做出的自主自愿的行为选择。意向性和与之相关的道德能动性概念被进一步修正:道德能动性离不开慎思能力,但并非所有道德能动性都出于慎思;道德能动性要求道德推理能力,但这又不同于道德无误^[8]。事实上,对意向性的修正,使意向性和道德能动性成为具有高度弹性的概念,可以在从宽泛至严格的不同维度上加以理解。从宽泛的意义上讲,我们可以将意向性理解为实体展现出一种特殊的功能能力,正如丹尼特通过“意向性立场”对能动性所做的界定,因

此人工意向性或人工道德能动性都具有可能性。

另一方面,传统观点对责任条件的理解不仅要面临正当性挑战,而且还存在循环论证的问题。在传统观点看来,“对某事负责任就是通过给予其行为应得的东西而对她的行为做出回应”^[8]。具体来说,超出道德义务的行为应受赞扬;违反道德义务的行为应受到指责或惩罚;履行道德义务的行为既不赞扬也不谴责。通过道德上的奖惩特别是惩罚,恢复行为所改变的社会秩序,从而“维持正义的平衡”。许多现代学者对此提出了质疑,认为惩罚作为“对过去错误的一种自然报复,这一点在有教养的社会中不应当再得到捍卫了”^{[14]59}。此外,传统观点既用道德责任能力解释和论证道德能动性的条件,同时又以道德能动性说明和解释道德责任的条件^[15]。正如前文所言,一个道德自主体要求其能为自己的行为负责;反过来讲,自主体为自身的行为负责,又要求他具有相应的道德行为能力和规范判断能力,即道德能动性。道德责任和道德能动性成为相互论证和诠释的问题。

现实地看,在道德实践中,相当一部分人类个体及某些非人类实体,在不满足完备条件的情况下仍被认为有能力实施道德行为,具体可分为五种类型。

一是缺乏充足理性能力。适龄儿童缺乏充足理性,他们对自己行为的价值仅有有限的认知,对道德规则的遵从仍处于道德他律阶段。“对于他们来说,因为无法看到相关的选择和后果,伦理行为的范围被大大缩小。”^{[16]52}儿童不是“道德共同体的无限制的成员”,但他们“正在逐渐成为有责任的行动者”^{[14]315}。一些国家将刑事责任年龄适度下调,从实践角度确认了适龄儿童具备道德行为能力和责任感。此外,工作犬等一部分动物表现出具有高度责任感的行为能力。有学者甚至主张,“能动性不需要包含自由和理性”^[11],道德能动性可以扩展至动物等非人类存在物。

二是缺乏意向性。道德能动性离不开人的意向性,但面对客观、不容忽视的义务或责任时,缺乏意向性不能成为否定个体道德行为能力和道德责任的托词和借口。俄狄浦斯神话是缺乏意向性的典型案例。赫费解释道,俄狄浦斯在完全无知无意中犯下弑父娶母的罪行,但无知无意的心理态度不能取消他的责任。“某些义务在任何情况下都不得受到蔑视,不依赖注意义务或过失问题。”^[17]弗洛里迪和桑德斯认为,俄狄浦斯没有恶的动机,相反,他竭尽善意的努力最终却做出了那个邪恶的行为,他对行

为的结果没有道德责任(moral responsibility),但他又是罪恶的唯一根源,需要为其行为负责任(accountability),他是没有责任的道德自主体^[18]。

三是缺乏意志力。意志是理性的实践能力,是个体按照理性法则行为的能力。成瘾症患者缺乏践履道德原则并将道德原则贯彻始终的意志力。他们有能力根据道德原则做出正确的道德判断,很多人常常为自己的错误行为感到懊悔、痛心疾首,甚至主动接受道德惩罚;但与此同时,他们又控制不住自己的错误行为。传统观点认为成瘾症患者受强烈欲望所驱使的行为属于其自觉自愿的选择。

四是缺乏道德感。道德感不是指“道德判断的能力”^[19],而是指“对善恶的好恶态度与反应”^[20]。个体为其行为负责要求他具备羞耻、愧疚、骄傲、自豪等道德感受能力。电影人物汉尼拔等反社会人格缺乏内在的道德感,他遵守道德规范的动力来自对免受道德惩罚的实际考量。汉尼拔还缺乏相应的反思能力,他不会因外部评价而重新审视自己的行为 and 动机。传统观点认为缺乏必要的道德感并不影响个体的道德能动性,他仍需为行为承担道德责任。

五是缺乏身体。身体是个体接受实质性惩罚的基础条件,也是负责任能力的重要前提。公司或法人组织缺乏可接受惩罚的身体,在司法实践中,对它们的惩罚不那么直观,或者针对公司或法人组织存在的根本原因——赚钱,或者针对其法律代理人进行惩罚。“公司可以拥有财产,签订合同,并对疏忽负责。”^[12]公司或法人组织拥有合法权利,可以实施法律行为,承担相应道德和法律责任,可以认为具有道德能动性。由此可见,在实践中所谓道德能动性的标准理论也是可疑的。

表1 不满足完备条件但具备道德能动性的主要类型

类型	一	二	三	四	五
缺乏条件	充足的理性能力	意向性	意志力	道德感	接受惩罚的身体
典型案例	适龄儿童、工作犬	俄狄浦斯	成瘾症患者	汉尼拔	公司、法人组织

传统观点运用人机类比方法,通过分析人工智能可否直接模拟人类道德能动性,或者分析人工物能否模拟诸如意向性、意识、负责任能力等构成要素,进而判断人工道德能动性的可能性^[21]。这种类比方法与人工智能发展的目标和理念是一致的,许多人工智能领域专家将模拟人类的思想或行动作为人工智能的建构方法。如著名的达特茅斯会议指

出人工智能的预期目标是“制造一台机器,该机器可以模拟学习或者智能的所有方面”^[22]。在伦理学领域,类比方法有助于利用已经发展起来的人类道德理论来阐释并解决人工智能引发的道德问题,这些问题具有一定相似性,如谁有能力实施道德行为、谁来承担道德责任等,相应的道德原理和方法也有相似之处,类比方法具备一定程度的适用性。但是,类比方法的局限性也不容忽视,它要么错误评估人工智能的功能能力进而否定人工道德能动性,要么陷入对人类道德能动性拙劣模仿的误区。约翰逊等学者就曾指出类比方法的本质是隐喻性的,过度夸大隐喻双方的相似性,忽略二者的根本差异性,将导致深刻的误解^[23]。事实上,由于现代科学对大脑的结构及功能的认识仍不足以构建大脑的数字模型或模拟人类意向性、意识状态,所以,通过类比人类道德能动性或其构成要素的研究方法很难得出真正可靠的结论。

三、道德能动性的开放立场

传统观点按照拥有完美理性能力和道德感受性的成年人形象提出了道德能动性的完备条件。“一个完备伦理自主体可以做出明确的道德判断,并且通常有能力合理地证明这些道德判断。”^[24]当然,传统观点在理论上或实践上面临的诸多困境,不仅要求我们审慎地对待传统观点及其结论,同时也要求我们站在开放立场重新思考道德能动性以及人工道德能动性的可能性。

站在开放立场上,道德能动性是一个包容性概念。“道德能动性的概念具有足够的弹性,允许不同种类的实体在道德意义上行动。”^[10]它包括道德的和不道德的、理性的和非理性的能动性^[9]。从这个意义上说,道德能动性将道德上有缺陷的个体,一部分动物、公司或法人组织等非人类实体均纳入其范畴内。道德能动性问题也由人的主观心智问题转变成客观的道德能力问题。也就是说,开放立场不再关注个体在实施道德行为时的特定心理状态,而重点关注人类个体或非人类实体所具有的道德功能能力,以及该能力对现实道德关系带来的实际影响。道德能动性是一种“功能主义现象”,即实体所表现出来的一种“功能主义特性”^[15]。图灵测试开启了功能主义先河,瓦拉赫和艾伦、鲍尔斯、汤肯斯(Tonkens)等学者也都表达了类似主张。根据功能主义,能动性可以宽泛地理解为实体“主动积极的

起动力作用的活动性能”^[25]。道德能动性是实体(不仅仅是人类)的道德行为能力,或者说实体积极施加有效道德影响的能力。

道德能动性不仅是一个描述性概念,而且还是一个规范性概念。道德能动性体现了道德自主体与道德承受体之间具有规范性的道德关系,是道德评价和道德责任归属的重要依据。据此,道德评价和道德责任归属不再依赖于实体的内部道德状态,而是依赖于实体的外部道德状态,即实体在道德关系中所处的地位以及实体对重要道德价值和具有重要道德价值的道德承受体的实际影响。例如,工作犬承担了由社会分配的特定的职责任务,它像人类个体一样尽职尽责,不论它是出于自由意志抑或本能驱动,人类都可以根据其行为表现来进行道德评价。人工智能也是如此。人工智能在与人类交互关系中不仅深刻地形塑人类自主体的言行,而且与人类形成了较为稳定、有道德意义的社会关系,那么,对人工智能自主实施的有道德意义的行为进行评价并要求其承担相应的道德责任就是可行的。“能动性和责任归属的审美现象学维度已经具有相当程度和足够的规范性力量来维系道德生活,不论‘内在现实’是什么。”^[11]开放立场上的道德能动性应该具备何种条件呢?

实体的内部状态“引发”了其道德行为表现。“引发”表明实体主动积极开启了道德意义上的因果关系链条,它成为道德承受体状态变化的特定原因。实体“引发”了某一道德行为表现,该行为及结果就可以归因于它,即便缺少任何心理状态或心理表征,它仍可以被视为道德行为者。就“引发”而言,道德自主体与道德承受体是相对而言的:道德自主体是道德行为的发出者,对道德承受体主动施加道德影响、发挥道德作用;而道德承受体是道德行为及后果的承受者,接受来自道德自主体的道德影响和作用。由内部状态“引发”表明,作为道德自主体的实体是其自身道德行为的第一推动力,当然这并不否定实体的内部状态是由其他实体或因素所引发或决定的。“基于行动是我们的能动性表达的这一观点。我们对我们的行动负道德责任是在仅当我们的行动显示我们是作为道德行动者的时候,即当作为我们的目的、承诺或价值表征的行动确实是由我们引起的时候。”^[26]

实体的内部状态不一定表现为意识,但应包含一定的认知能力。作为道德自主体的实体所实施的行为不是必然性的结果,而是基于对行为本身(包

括行为的主体、对象、目的、环境、条件等行为的构成要件)的认知所做出的自主选择。正是对行为的认知“引发”了实体的行为表现,与此同时,认知又给实体提供了进行选择的前提,使实体的行为表现具有道德意义。伦理是一种选择的品质。而缺乏对行为的充分认知,实体就看不到其他可能的选择,伦理行为的范围就会大大缩小,行为就不能说是真正自由的。这样看来,完全没有智能的实体不具备能动性。

作为道德自主体的实体受道德规范的约束,或者说,道德自主体与道德承受体的关系是道德规范性的。“道德自主体是负有多个责任或义务的实体,而道德承受体是要给予其至少一项责任或义务的实体。”^[8]按照道德规范的要求,道德自主体对道德承受体负有某种道德义务,他的行为受道德规范调节,同时按照道德规范进行评价和责任归属。作为事物之间相互影响、相互作用的因果能动性普遍存在,但在众多的因果条件中一个特定的因果条件被挑选出来作为该事件的原因,这种挑选不仅取决于经验性事实,也取决于道德规范。

四、人工道德能动性的可能形态

站在开放立场看,道德能动性既可以是自然生成的,也可以是通过人工建构以软硬件形态表现出来的,还可以是通过人机交互中生成的集体能动性表现出来的。后两者构成了人工道德能动性的可能形态。

1. 无意识的道德能动性

理性、情感、意志以及意向性均属于人的意识范畴。在人工智能领域,构建机器意识的研究一直在持续,同时也面临着巨大争议:我们对人类的意识和大脑的认识仍然非常粗浅,更不用说去用计算机思维来表征它们。如果不考虑意识的话,人工道德能动性只需考虑人工智能表现出来的功能能力即可。也就是说,如果人工智能表现出与人类自主体同样的道德行为能力和道德影响能力,那么它就具有道德能动性。鲍尔斯认为,人类与正确类型的计算机没有道德差异,但计算机的能动性属于“无意识道德”^[10]。

弗洛里迪和桑德斯详细阐述了“无意识道德”,认为实体可以在没有自由意志或其他精神状态情况下成为道德自主体。他们采用一种在抽象层次上描述、分析和讨论系统及其背景关系的抽象方法来解

释道德。“道德可以被认为是在确定所考虑的 LoA (抽象层次) 的界面中,对观察值定义的‘阈值’。”^[18]如果自主体的行为遵守了这个“阈值”,他在道德上就是善的;如果他不遵守这个“阈值”,他在道德上就是恶的。道德自主体的正确 LoA 特征可以概括为互动性、自主性、适应性和实施道德上合格的行为。其中,实施道德上合格行为是指实体“能够引发道德上善或恶”,而互动性、自主性、适应性表明了实体所具有的行为能力。艾伦等人提出了判断人工道德自主体的道德图灵测试及其替代版本,同时他们也认为考虑到人类不是纯善的,在测试中,成功的人工道德自主体可能是更道德的那一个^{[16]61}。

无意识的道德能动性不同于以往技术人工物的工具性功能。人类道德能动性的发挥离不开技术人工物的参与。例如,交通信号灯指导过往车辆和行人安全有序地通行,但交通信号灯是按照设定好的程序控制交通,有些交通信号灯能够根据环境变化自动调整控制参数,具有一定的自适应性,它没有认知的内部状态,也不能“引发”自身的道德行为表现。交通信号灯只是人类使用的工具,是影响人类能动性发挥的背景条件。人工智能具有自主行动能力,它可以在没有人类直接控制的情况下独立地实施机器行为,造成某种有意义的结果。例如,聊天机器人对外部信息做出积极回应,并根据自己在环境中学习的东西进行自主回答,还能通过学习不断改变自己的对话方式。聊天机器人的对话是它基于对聊天环境的认知自主引发的行为,是针对人类对话者的,并对后者带来积极或消极的影响。聊天机器人与人类对话者之间不再是一种纯粹的工具性关系,而是一种带有明显规范性的关系,都受道德规范的约束和指导。虽然聊天机器人是无意识的,没有任何主观上的目的、意图、情感或情绪,但它具有道德能动性。

显然,包括聊天机器人在内的人工智能有能力基于自身对外部环境信息的认知状态,根据社交准则或道德原则的要求,独立“引发”具有道德意义的行为,展现出不同程度的道德能动性。但与此同时,人工道德能动性与人类道德能动性存在着显著差异:一方面,人工智能的认知状态或来自符号推理过程,或来自对大数据的学习,本质上都是计算机的信息处理能力,而这也限定了人工智能进行道德选择的范围;另一方面,人工智能“引发”道德行为表现的功能能力是指其在与外部环境的交互作用中自主

学习特定的行为模式,并根据对外部环境信息的判断做出相应的决策和行动,并不意味着人工智能从根本上摆脱了人类的控制,成为独立的道德自主体。

由于人工智能距成为独立的人工道德自主体还有相当长的一段距离,在这一过程中,人工智能仅表现出有限的道德行为能力,其道德功能能力及道德选择的范围都是有限的,是准道德自主体。摩尔(James H. Moor)区分了不同层次的伦理自主体:伦理效果自主体(可从伦理效果上评估机器)、隐含式伦理自主体(解决机器的安全性和可靠性问题,避免不道德行为)、显现式伦理自主体(用机器描述道德并做出道德判断)和完备伦理自主体^[24]。在此基础上,瓦拉赫和艾伦根据自主性和对价值的敏感度两个维度,在操作性道德和可靠的道德自主体之间划出一个功能性道德区,以便容纳人工道德自主体的各种可能形态。他们的观点也可用于理解人工道德能动性:人工道德能动性是由人工智能在两个维度上定位决定的,而非简单的线性发展过程,人工道德能动性的进步与两个维度的进步密切相关。

此外,无意识的道德能动性并不代表人工智能获得了道德权利和道德地位。“作为主体的人及其实践活动是能动与受动的统一。”^[25]无论人类个体抑或公司、法人组织,都既受道德义务的约束,又受道德权利的保护,是道德自主体和道德承受体的统一。相较之下,人工智能缺乏意识和感受性,它虽然受道德义务的约束,却不受道德权利的保护。人工智能具备道德能动性,但没有道德受动性;是道德自主体,而不是道德承受体。人类对待人工智能的方式,在一定程度上受到它展现出的形象和功能能力的影响,如对人类表达关爱的聊天机器人,人类也会“投之以桃,报之以李”,表现出对机器的喜爱和关心;反过来,对满口污言秽语的聊天机器人,人类则会表现出憎恶和愤怒。但这种方式本质上仍是工具主义的。

2. 人机交互中的集体道德能动性

通常情况下,人工智能不是独立引发道德行为的,而是通过人机交互方式形成集体道德能动性,亦即团体道德能动性或共享道德能动性。利斯特(Christian List)明确主张集体能动性与人工智能系统具有关键的并行关系,集体能动性可以被视为人工智能系统的一个特例^[15]。他以集体能动性为参照来阐明人工智能的道德能动性和道德地位等问题。更为重要的是,人机在交互关系中构成混合行动体,并展现出不同于人类个体或单个机器的道德

行动能力。“人机混合道德能动者,顾名思义是由人与计算机、机器、人工智能等非人类能动者共同组成的行动主体,它属于一种多行动者系统。”^[27]根据人机关系的特殊性,人机交互中的集体道德能动性可分为三种类型。

一是授权的道德能动性。授权意味着将原属于人类的社会职责授权给人工智能,由人工智能代表人类自主体执行任务或实施有意义的道德行为。人工智能代表人类行事,这是人工智能发展的重要方向。“关于人工物能动性的论述似乎常常涉及那些为人类执行任务或代表人类行事的自主体”^[23],“机器人可以被视为代表他人行事的自主体”^[12]。例如,交通信号灯代替打手势的交通警察,洗碗机和扫地机器人代替清洁工人,自动驾驶汽车代替驾驶员,等等。与交通信号灯、扫地机器人不同,将驾驶任务委托给自动驾驶汽车,意味着允许它处理可能出现的各种路况信息,包括紧急情况下处理车辆与路人之间的利益冲突。人类向人工智能进行授权,给人工智能一定的行动自主权,人工智能在授权范围内相对独立“引发”道德行为。但授权并不意味着人工智能成为独立道德自主体:一方面,人工智能代理人类行事,行为后果的责任归属于授权人与人工智能的整体;另一方面,“人工物通过刻在其设计中的授权意图影响人类行为”^[23],人类的意图和道德能动性是通过开发—设计—测试—使用的纵向因果关系链条传递给人工智能并由后者展现出道德行为能力,人类制造者也参与了授权行为。当然,道德授权比较复杂,它不仅把人类的某个道德义务授权给人工智能,而且把对于人类的道德义务授权给人工智能。过多的道德授权可能造成人类主动放弃道德自主体地位,进而放弃对自身命运的掌握。因此,道德授权是有限度的。正如亚当(Alison Adam)的分析所言,道德授权应该增加人类的道德资本,提高人类的道德能力,而不是削弱它^[28]。

二是扩展的道德能动性。随着人工智能在多领域的广泛应用,很多道德事件是人工智能与人类使用者共同完成的,而不是由人类使用者或人工智能单独完成的。在此过程中,人机构成联合行为体共同“引发”了针对道德承受体的行为。人机联合行为体的“引发”行为来自他们对行为的集体认知状态,进而产生具有道德意义的行为后果。以寻人数据库为例。寻人数据库具备面部识别、地域推荐、多维度查询等功能,向人类用户提供带有倾向性的决策,与传统互联网工具已有本质不同。但寻人数据

库无法单独发挥作用,它需要在使用过程中将人类的理性与机器决策结合起来形成共同认知状态,这样才能发挥其道德影响。约翰逊将这种能动性称之为三元能动性。她认为处于因果链远端的人类用户和设计师发挥的意向能动性和因果效力,与处于因果链近端的人工智能的因果效力共同构成了集体能动性,她还预测未来处于因果链更加远端的人类用户仍将共同发挥集体能动性^[7]。汉森(Allan Hanson)认为由人机构成的联合行为体具有扩展能动性。“扩展能动性理论认为,行为事件(deed)是由学生、图书馆馆藏数据库以及使两者联系起来的自动化硬件和软件组成的联合实体共同完成的。这种超越个人能动性的扩展能动性其背后的基本理由是,如果一项行动只能在各种各样的人类和非人类参与者的共谋下完成,那么,实施该行为的主体或能动性不能局限于人类部分,而必须由他们所有人组成。”^[29]扩展能动性是人类用户与人工智能共同展现出的道德行动能力,相应的道德责任属于作为整体的扩展能动性,而不单独属于其中人类或非人类实体。

三是集体道德能动性还有一种特殊形式,即分布的道德能动性。随着人工智能技术的广泛应用,多主体系统的相互作用会使原本价值中立的行为转变为具有道德负载的行为。弗洛里迪对此进行了论述,他认为价值中立或在道德上微不足道的行为,通过适当的集聚或多主体系统的相互作用,如果突破一定的道德阈值,行为的性质就会发生变化,具有道德上善或恶的意义。例如,个体小数额捐款可能做不了什么,但借助慈善平台就能够兴建学校、医院及养老院等大型公益机构;相反,人们到公共场所活动是完全无害的,但人数超过极限时就可能发生踩踏事件。在集聚或多主体共同参与的情况下,行为结果的道德性质会使分布式行为的道德性质发生转变。道德阈值的边界为何,弗洛里迪认为并没有绝对的界限。这是因为分布式道德情境在道德上是弹性的。如果公共场所秩序良好,它可能容纳更多的人而不会出现踩踏事件;如果公共场所处于无序状态,它可能容纳的公众数量就会少很多。这说明公众进入公共场所的行为性质在道德上是否会变化,在一定程度上受基础性设施的环境容量的影响。在分布式道德中,人类或非人类实体确实“引发”了单个行为,但他们没有主动“引发”道德行为,是由系统的相互作用使行为形成集聚效应,才产生了具有道德意义的后果。在这里,道德能动性是通过社会

系统分布在人类和非人类实体包括人工智能之中的,“道德活动不仅仅依附于作为个体的人类或人工智能,而是可以在社会、组织和人工物与人类自主体网络中分布”^[28]。与此相应,道德责任也应在人类和非人类实体中进行按比例分配。

综上所述,站在开放立场来看,人工道德能动性具有不同表现形式,这反映了人工智能在道德事件中所发挥的功能作用带有差异性。特别是集体道德能动性,反映了人工智能时代人机交互关系的特殊性及人机共享的道德功能能力。一方面,它为人工智能道德事件进行道德责任归属提供了更加完备的理论依据;另一方面,它也对开发和人工智能提出明确要求,人工智能的开发和使用应与人类的价值观保持一致,应增加人类道德资本而不是削弱人类的道德能力和道德地位,应将人工智能作为社会技术系统整体上考虑其复杂的道德影响和社会价值。

参考文献

- [1] CIBRALIC B, MATTINGLY J. Machine agency and representation [J]. AI & Society, 2022. DOI:10.1007/s00146-022-01446-7.
- [2] SCHLOSSER M. Agency. The stanford encyclopedia of philosophy [EB/OL]. Edward N. Zalta (ed.), Winter 2019 Edition, (2019-10-28) [2023-11-08]. <https://plato.stanford.edu/archives/win2019/entries/agency/>.
- [3] 李磊,王国豫.深部脑刺激:同一性、能动性和责任[J].哲学动态, 2019(6):109-116.
- [4] BODDINGTON P. AI and moral thinking: How can we live well with machines to enhance our moral agency? [J]. AI and Ethics, 2021(1):109-111.
- [5] SPARROW R. Why machines cannot be moral [J]. AI & Society, 2021(36):685-693.
- [6] 王淑庆.人工道德能动性:从连续性进路到非连续性进路[J].湖北大学学报(哲学社会科学版),2021(9):17-24.
- [7] JOHNSON D G, VERDICCHIO M. AI, Agency and responsibility: The VW fraud case and beyond [J]. AI & Society, 2019(34):639-647.
- [8] HIMMA K E. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? [J]. Ethics and Information Technology, 2009(11):19-29.
- [9] MCDONALD F J. Agency and responsibility [J]. Journal of Value Inquiry, 2010(44):199-207.
- [10] POWERS T M. On the moral agency of computers [J]. Topoi, 2013(32):227-236.
- [11] COECKELBERGH M. 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.

- [12] ASARO P M. Robots and responsibility from a legal perspective [EB/OL]. (2007-01-20) [2023-11-08]. <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=78F466B9A56488D53D5CE75F6431A29B?doi=10.1.1.475.809&rep=rep1&type=pdf>.
- [13] 孟令宇.道德地位的双源假说及其意义[J].理论界, 2021(1): 45-51.
- [14] 徐向东.自由意志与道德责任[M].南京:江苏人民出版社,2006.
- [15] LIST C. Group agency and artificial intelligence[J]. Philosophy & Technology, 2021(34): 1213-1242.
- [16] 瓦拉赫,艾伦.道德机器:如何让机器人明辨是非[M].王小红,译.北京:北京大学出版社,2017.
- [17] 赫费.作为现代化之代价的道德:应用伦理学前沿问题研究[M].邓安庆,朱更生,译.上海:上海译文出版社,2005:17.
- [18] FLORIDI L, Sanders J W. On the morality of artificial agents[J]. Minds and Machine, 2004(14): 349-379.
- [19] 王荣麟.道德是天生的吗?——乔依斯“道德先天论”辨析[J].复旦学报(社会科学版), 2014(2): 119-127.
- [20] 高德胜.道德情感:本质、类别与意义[J].当代教育与文化, 2021(6): 1-10.
- [21] 王淑庆.人工道德能动性的三种反驳进路及其价值[J].哲学研究, 2021(4): 119-126.
- [22] 李德毅.人工智能导论[M].北京:中国科学技术出版社, 2018: 2.
- [23] JOHNSON D G, NOORMAN M. Artefactual agency and artefactual moral agency[C]// KROES P, Verbeek P. The Moral Status of Technical Artefacts. Springer Dordrecht, 2014: 143-158.
- [24] MOOR J. The nature, importance, and difficulty of machine ethics [J]. IEEE Intelligent Systems, 2006(4): 18-21.
- [25] 何祚榕.论能动性:与吴建国、崔绪治同志商榷[J].中国社会科学, 1981(2): 167-170.
- [26] 闫宏秀.数据时代的道德责任解析:从信任到结构[J].探索与争鸣, 2022(4): 37-46.
- [27] 邹桑.人机混合能动者的伦理意义、建构分析和启示[J].哲学分析, 2019(5): 41-52.
- [28] ADAM A. Delegating and distributing morality: Can we inscribe privacy protection in a machine? [J]. Ethics and Information Technology, 2005(7): 233-242.
- [29] HANSON A. Beyond the skin bag: On the moral responsibility of extended agencies [J]. Ethics and Information Technology, 2009(11): 91-99.

The Ethical Proof of Artificial Moral Initiative: From Traditional Perspectives to Open Positions

Qu Rong

Abstract: The traditional view interprets moral agency as a unique ability of human moral behavior and proposes complete conditions of moral agency. The traditional view does not directly deny the possibilities of the subjectivity and agency of artificial morality, but it adheres to the anthropocentric position and lacks sufficient reflection on the complete conditions for moral agency. In fact, both human and non-human beings still possess moral agency in the absence of certain conditions. From an open standpoint, moral agency can be broadly understood as the ability of an entity to actively exert effective moral influence, requiring the entity to possess moral behavior “triggered” by its internal state, internal cognitive ability, and be constrained by moral norms. Moral agency can be both naturally generated and unconsciously constructed through software and hardware, as well as collective moral agency manifested through human-computer interaction, including authorized moral agency, extended moral agency, and distributed moral agency.

Key words: artificial intelligence; moral agency; artificial moral agency; consciousness; intentionality

责任编辑:思 齐