【伦理与道德】

人工道德智能体的发展路向*

——基于道德判断问题的分析

余天放

摘　要:尽管人工道德智能体的发展受到了诸多批评,然而人们依然有一种强烈的直觉认为,应该让机器可以像人那样有道德地行动。莫尔对于人工道德智能体的分类就反映了这一看法,他相信一种"完全的伦理智能体"将是"明确的伦理智能体"的更高实现。但这一看法对于发展人工道德智能体是有误导性的,因为这两种智能体的设计并不相同,"完全的伦理智能体"的实现将更大程度上依赖于"机器意识"问题的解决,而不是在机器中植入道德规则的方案。支持这一观点的理由在于,人类道德判断能力包含有道德直觉和道德推理两种模式,而植入道德规则的方式只能够实现机器的道德推理,却不能够实现它们的道德直觉。因此,一种能够参与人类实践活动的道德实体才是道德智能体应被期待的发展路向。

关键词:人工道德智能体;道德判断;机器伦理;伦理智能体

中图分类号:B82　　**文献标识码**:A　　**文章编号**:1003-0751(2022)05-0087-09

伴随人工智能或自主式机器人在我们生活中的愈加介入,关于人工道德智能体(artificial moral agents, AMAs)的构想也得到更多的讨论,而这些讨论的目的在于回答这样两个问题:第一,一种具有自主行动能力的机器人应如何对待人类?第二,一种参与人类生活的机器人应如何有道德地行动?这两个问题分别针对两种不同的场景而被提出,前者仅将人工智能体(artificial agents)看作一类人造物,因此它们需要有道德地对待人类,或者至少对于人类而言是无害的;后者则将人工智能体看作一些类人(human-like)的存在物,它们需要像人类那样拥有道德,既包括与人类的关系,也包括与其他人工智能体的关系。对于第一个问题,学者们经常会考虑应用阿西莫夫的三定律来作为限制机器人行为的规则,即便这些规则被表明为不适当的①;对于第二个问题,学者们已提出了诸多在人工智能体中置入人类道德规范的建议,这些建议无疑推动了人工道德智能体的设计与发展,但同时却并没有使得人工道德智能体的发展路向问题得到真正的澄清。

据此,有关人工智能体能否有道德的讨论中出现了这样一种分歧:部分学者否定了一种赋予人工智能体以道德的构想,他们相信,由于机器人的道德责任归属是不同的,所以不应该期待一种能够实现人类道德的人工智能体;与之相反,另一些学者则持有更为积极的态度,他们不仅提供了发展人工道德智能体的理由,同时也已经构想了多种实现方案,包括"自上而下的""自下而上的"以及混合的方案等。针对这一分歧,我们将表明,在构想某种人工道德智能体时我们会被引向对于一种与人类道德能力相一致的智能体的期待,然而这个期待是误导性的。我们更应该期待一种作为道德实体的人工道德智能体,它们将参与我们的社会活动,但并非真正的道德

收稿日期:2022-01-22

*基金项目:国家社会科学基金后期资助项目"康德自我理论研究"(21FZXB066);扬州大学校教改课题"智能时代新德育模式研究"(YZUJX2021-C1)。

作者简介:余天放,男,扬州大学社会发展学院哲学系讲师,哲学博士(扬州　225002)。

主体。

一、人工道德智能体的分类

在对人工道德智能体所可能具有的形式进行构想时,学者们会使用一些分类法(taxonomy)来区分不同类别或不同等级的道德智能体,其中最常见的有这样两种:一是由莫尔(James. H. Moor)所提出的按照人工智能体实现道德的程度而进行的分类;二是由艾伦(Colin Allen)等人所提出的按照人工道德的设计方案而进行的区分。这两种分类法并不是相互排斥的,事实上一种综合了这两种方法的新方案能够更加全面地总结现有人工道德智能体的发展。[②]此外,分类法还具有的一个潜在作用则是为人们提供一个未来发展的预期目标,它通过展示一个等级次序而告诉我们应期待怎样的人工道德智能体。对此,我们指出,莫尔的分类法中对于这个最终目标的设定是有误的,它将我们错误地引向对于一种充分实现了人类道德的智能体的期待,并进而产生有关机器伦理的诸多争论。

在莫尔的分类法中,他将机器伦理的可能形式区分为"伦理作用智能体"(ethical impact agents)、"隐性的伦理智能体"(implicit ethical agents)、"明确的伦理智能体"(explicit ethical agents)以及"完全的伦理智能体"(full ethical agents)[③],这一区分是根据机器中所实现的伦理等级而做出的。第一等级的"伦理作用智能体"仅能够被动地产生一些伦理效果,例如,我们应用机器设备取代了一些高风险、大劳动量的工作,从而产生了一些好的伦理效果,此时这些机器即被看作一类"伦理作用智能体"。第二等级的"隐性伦理智能体"在第一等级的基础上实现了对于人类而言的安全性和可靠性,此时机器被编入了一定的代码,从而保证它在与人类的互动中遵守既定的规则。莫尔举例说银行的自动取款机以及飞机的自动巡航模式即扮演了这样的角色,它们将按照既定程序很好地完成自己的任务,从而避免发生与安全性相关的伦理问题。第三等级的"明确的伦理智能体"则能够更为清晰地展现一些伦理范畴,并且能够按照某种道德要求去进行决策或行动。在当前有关机器伦理的研发中,诸多计算模型或系统已朝向这一目标在进行设计,例如,MoralDM,Jeremy 以及 W. D. 等多个决策模型都被视为"明确的伦理智能体",它们能够根据被植入的道德算法而处理特定场景下的道德困境。[④]最后是第四等级的"完全的伦理智能体",它们实现了与人类的平均水平相似的道德判断(moral judgment)能力,同时也被要求具有意识、意向性以及自由意志等,从而成为一类完全意义上的道德主体(moral agents)[⑤]。很明显的是,这类"完全的伦理智能体"目前只是一种可设想的方案,它的出现将依赖于我们在多久的未来能够实现通用的人工智能。

此外,在艾伦等人的分类法中,人工道德智能体则按照在其中实现伦理价值的方式而被区分为"自上而下的"(top-down)、"自下而上的"(bottom-up)以及"混合的"(hybrid)三种类型。[⑥]"自上而下的"方式要求将一种或多种道德理论和原则作为机器人进行决策时所遵循的规则,此时它们只需要考虑如何最大化某种道德价值(义务论方案)或者计算如何使得最终结果最优(后果主义方案)即可。与之相对的是,"自下而上的"方式则并不要求在机器中预先植入某些道德理论和原则,而是提供让机器人在其中获得奖惩的环境,从而使得它们能够自主地产生对于道德价值的敏感性。"自上而下的"方式部分地模拟了我们人类的道德教育,即将那些已得到普遍同意的道德法则以知识的形式植入另一个主体当中。然而这一方法所面对的挑战是,不同的道德理论和原则之间可能存在冲突或者矛盾,以至于我们无法决定应选择哪种道德原则去遵守,此时就出现了某种道德困境或者道德不确定的情况。此外,对于机器人来说,"自上而下的"方式还可能存在运行上耗时过长的问题,因为在处理任何一个行动决策的问题时都需要进行大量的计算,而这在面对一些瞬间的特殊情况时被认为是不可能的。因此,为避免这些问题,"自下而上的"方案要求我们模拟人类进化过程中道德价值的产生方式,通过设计一些场景或平台而使得机器人可以自主地学习如何有道德地行动。然而这一方案也存在一些特定的困难,例如,让机器自主学习将无法保证它们的道德行为是必然的,以至于可能出现我们无法预测的非道德行为。并且这种"自下而上的"设计所需经历的时间是不可估的,最终的效果也是不确定的。据此,一种结合了"自上而下的"和"自下而上的"混合方案得到了部分设计者的青睐,例如,在一种名为 LIDA(Learning Intelligent Distribution Agent)的认知结构中就同时使用了以上两种设计方案,它通过区

分反应层和元认知层来实现这一点。在反应层,机器人将通过监测情绪反应而限制机器人可能存在的一些伤害行为;同时,在元认知层,又植入了一些类似康德式绝对命令的规则来保证机器人的基本行为。⑦

二、对发展人工道德智能体的批评

莫尔的分类法反映了直觉上人们对于发展人工道德智能体所拥有的这样一种想法:我们应该让机器也可以像人那样有道德地行动。然而这一想法也许在观念上有自相矛盾之处,就如同我们无法训练一只猎犬既能够自主地行动同时又符合人类的道德要求一般。因为对于猎犬这样的自然物而言,自主性和有道德似乎是很难共存的。关于这一看法,学者们对人工道德智能体的发展提出了诸多诘难,而这些诘难总体上可以被区分为这样两种态度:一是设计人工道德智能体时所需面对的技术挑战是无法克服的;二是让人工智能体拥有道德的观念是不适当的。

第一,如果一个机器可以拥有道德,那么我们需要考虑伦理规范如何从它的程序中产生出来,而这种方式已被艾伦等人归纳为上面的三种:“自上而下的”“自下而上的”以及“混合的”。然而这三种方式各自所面对的困难也是显著的,并且这些困难反映了一个更加基础的元伦理学问题:如何通过算法解决道德问题?⑧也就是说,如果我们把人类道德构想为一系列的指令或代码,那么此时会存在一个伦理规范向道德命题还原的问题,以至于部分的伦理学家可能会反对这样一种结合了道德实在论和道德认知主义的观点,而倾向于支持非实在论或者非认知主义的立场。或者他们至少质疑说,在人类中间没有形成关于什么是“善”的一致性看法之前,如何让机器人按照“善”去行动?此时,那种“自下而上的”方法似乎也无法绕开这一障碍,因为即便我们可以构造一种基于刺激—反应模式而行动的道德智能体,但我们依然很难说它就是一个真正意义上的道德主体,因为它可能并不明白自己为什么这么做,并且这样做与“善”有怎样的关系。此外,发展人工道德智能体还将面临的一个技术性挑战是,我们如何让机器人理解并识别某一道德场景,从而可以准确、迅速地调动特定的算法来处理当前问题?对于人类来说,这一能力似乎已发展为一种道德直觉,以至于我们在面对一些典型状况时会产生相应的道德情感来提示我们其中可能存在的道德价值。但对于机器人来说,我们又该如何通过设计而产生这种道德直觉,并且让它敏感于特定的价值呢?再退一步来说,即使以上的问题都得到了解决,但任何一个实现了人类道德推理(moral reasoning)能力的机器人在某个场景下所进行的运算都是巨大的,我们似乎很难要求这个智能体能够在极短的时间内完成对当下诸多可能性的运算,更不用说这些可能性能否完全体现预期的结果。因此,如果要实现机器伦理,我们至少需要克服这样几个技术上的困难:(1)在机器中实现的道德反映了人类的真实道德状况,而不是部分人的道德,或者部分理论所支持的道德;(2)机器人能够自主地做出道德判断,具有价值敏感性;(3)进行道德推理时,所需要的运行时间是受限的。

第二,发展人工道德智能体可能在观念上就是不可行的,因为这将有悖于它们只是一些机器的看法。与这一观点相一致的批评是多样的,例如,范·韦恩斯伯格(Aimee van Wynsberghe)等人指出,我们需要区分一个道德上可指责的情境和对一个道德角色的代表。⑨当动物被人类训练而用于一些治疗的场景时,这些动物对于人类安全的保证只是处于一个道德上可指责的情境当中,而不是真正代表了一个道德角色,就像一个成年人所具有的那样。因此,无论人工智能体未来的发展是怎样的,我们都只可能把它们作为一类达到某种目的的工具或手段,而不可能被作为目的本身。与之相类似,通肯斯(Ryan Tonkens)提出了在发展机器伦理时所难以避免的“伦理上不一致”的问题,这就是说,机器所被植入的伦理法则将无法与创造它们时所需遵守的伦理法则一致。⑩例如,在建造一个其行为符合康德式义务论的伦理机器人时,我们却不能够将它仅作为目的而不作为手段;或者在建造一个符合功利主义原则的机器人时,研发过程中所消耗的各种资源也明显地破坏了“满足最大多数人的最大幸福”原则。据此,从观念上而言,对机器伦理的发展至少需要解释这样两个问题:(1)我们应如何对人工智能体进行道德责任的归属,是归属于创造者还是归属于智能体本身?(2)在同一个伦理框架内如何实现创造人工智能体所遵循的道德规范与它们自身被植入的道德规范一致?

三、发展人工道德智能体的更多理由

尽管存在以上诸多反对发展人工道德智能体的理由，然而这些理由似乎并没有对将道德规范植入人工智能体的现实计划产生明显的阻碍，特别是近年来在自动驾驶、医疗保健、增强现实技术等领域中对智能体的应用，人们更加期待一种具有较高道德水平的智能体的出现。因此，针对这样一个现象我们需要考虑，仍然有哪些理由支持着我们发展人工道德智能体？这些理由是对于机器伦理的误解还是对以上挑战的回应？

第一，发展机器伦理的一个重要理由来自我们直觉上的这样一个想法：让机器可以像人一样有道德。据此，那种“自上而下的”设计方案成了研究者们的首选，因为它不仅相容于目前大多数智能机器设计中所运用的规则系统，在操作上是便利的、可行的；同时，也因为它与我们人类的道德教育模式一致，即将道德规则通过命题的方式传授给他人。然而在部分批评者看来，这种构想是没有必要的，我们可以在机器中编入一些指导它行动的执行程序，从而在面对一些复杂情况时可以根据个别原则来行动，例如“最低伤害”的原则，但这并不要求机器在真正意义上具有道德。并且上文也提及，我们能够承认机器将广泛地出现于一些道德上可指责的情境中，但这并不表示它们能够代替人而成为一个真正扮演道德角色的道德主体。针对这一批评，福摩沙（Paul Formosa）和瑞恩（Malcolm Ryan）回应说：“一些通过行动或不行动而会伤害人类的机器应该成为人工道德智能体，但另一些机器，例如一个傻傻的吐司机，则不应该。”⑪这一回应的论点在于，当一个机器拥有较高的自主行动能力时，那么它就应该被赋予一定的道德决策能力，例如在自动驾驶领域中，而这就成为一个“明确的伦理智能体”。此外，福摩沙和瑞恩还建议，我们可以根据机器自主行动能力的强弱来区分它们被归责的等级，例如，区分为在道德上完全、非完全以及部分有责任的主体，这只不过将推动我们形成一个关于道德责任的更复杂的看法。⑫

福摩沙和瑞恩虽然给出了一个有关人工智能体之责任归属问题的解决方案，但他们并没有真正回应批评者所隐含的这一质疑：为什么必须要让机器人具有道德？因为一种明显的可能是，我们可以始终让人类扮演那个最终道德责任承载者的角色，即便当机器人需要在瞬间做出决定时，我们也可以通过对责任偏好的预先设定来解决这一问题，例如，让智能体的使用者在利己偏好和利他偏好之间在先地进行程度选择。因此，一种能够为发展人工道德智能体进行更强辩护的观点需要对我们的直觉进行解释，即说明为什么我们会期待机器像人一样有道德。这种直觉同样也体现在莫尔的分类法中，他通过区分“明确的伦理智能体”和“完全的伦理智能体”来表明，我们似乎可以期待一种完全地实现了人类道德能力的人工智能体，只不过同时也需要实现我们对于人工智能体所期待的其他困难目标，例如意识、意向性、自由意志等。因此，让机器人具有道德的直觉性要求很大程度上来自我们对于通用人工智能的期待，我们似乎无法接受一个高度类似于人的智能体却是对道德无感知的。

第二，针对设计道德机器人时所面对的技术性困难，我们需要解释一个人工道德智能体如何能够在道德推理的运行上是有效的、即时的，同时又把握了一种真实的人类道德。在马尔提诺（Andreia Martinho）等人的构想中，他们将此问题看作一个如何在机器中既实现道德的异质性，同时又避免运行时间过长的问题。对此，他们通过引入离散选择分析（discrete choice analysis）的方法对人们的道德偏好和决策规则进行统计并编码，从而经验地反映出人们在道德推理中所存在的不确定性。⑬这一方案的优势在于，它以“自上而下的”方式实现了“自下而上的”目标，同时又不是一种简单的混合方案。一方面，该方案没有在先地依赖于某个道德理论或某部分人的道德偏好来对机器中运行的道德规则进行编码，而是通过对人们在具体场景下所做的道德选择进行分析，从而获得一个拥有潜在类别的离散选择模型，再通过将这一模型植入人工智能体的系统中来实现对人类真实道德状况的反映。另一方面，该方案也避免了传统经验性方法构造一个道德机器人的刺激—反应模型耗时过长的问题，并且最终所做出的道德推理也能够是瞬时的，因为它只需要在既有的规则系统中进行搜索即可完成。也许我们会继续质疑这种基于数据的分析能否真正反映人类的道德偏好，但至少马尔提诺等人的工作让我们看到了从技术上解决这一问题的可能性，并且指出了一种实现机器伦理的综合路径。

第三,还存在一些支持我们发展人工道德智能体的实用性理由,例如,防止机器人对人类的伤害,获得公众的信任以及通过机器伦理更好地理解人类道德等。[14]相对于以上基于直觉的要求以及对道德在技术上可实现性的辩护,这些实用性理由则显得更弱一些,我们或者可以选择其他的方式而达到同样的目的,或者可以仅仅在最低程度上设定机器不会伤害人类,而无须在机器中实现更加狭义的道德。因此,对于一个真正意义上的人工道德智能体来说,我们可能会拥有一些更高的期望,而这种期望就在于让机器人的道德满足一些内在条件,让它们可以像人类一样进行道德推理和决策。这事实上就成了支持者和反对者产生冲突的最后边界:是否有必要,并且是否可能让机器进行自主的道德判断?

这个问题的困难来自道德判断概念的模糊性。大多数研究者相信,人类之所以被看作是有道德的,不仅在于人类可以做出一些符合道德原则的行为,更重要的是人类具有一些内在的意向性状态,而这些状态才是产生出这些道德行为的原因。因此,在人工道德智能体的支持者看来,实现机器伦理的任务不仅在于让机器人的行为符合外在的道德标准,同时要求它们像人类一样具有这些内在状态,进而体现为一个道德智能体能够自主地进行道德推理,并产生道德决策的过程。[15]但在范·韦恩斯伯格所代表的反对者看来,“没有任何证据表明,让机器具有道德推理能力是不可避免的,而不管它们是否在一个道德显著的环境中工作”[16]。由此可以看出,这两派的争论最终集中于向智能体植入人类道德推理能力的问题上,支持者试图说明这种能力从人类向机器中的转移是如何可能的,但反对者却相信,这种转移或者是不可能的,或者是不必要的。此外,这一问题的复杂性还在于,一个实现了道德运算的“明确的伦理智能体”也可能不被认为是一个真正的道德主体,因为它所实现的也许不是人类所拥有的那种道德判断能力,它缺少区分对与错这一关键的能力。[17]基于这些困难,我们将进一步比较人类所具有的道德判断的方式与机器中所可能实现的方式,最终表明:由于道德判断方式上的差异性,我们并不应该期待一个与人类相类似的道德主体,相反,我们应该期待一个比人类更好的人工道德实体(entity)。

四、人类道德判断的方式与特征

在此,我们将给出一个关于人类道德判断的一般性描述,这一描述的目的并不在于形成一个在解释上具有较大前景的理论模型,而仅仅是通过对相关因素的考察来表明人类在做出某一道德决策时所可能受到的影响是怎样的,并且我们将在较大程度上依赖于近年来的一些经验性研究。

第一,意向(intention)与动机。解释道德判断的一个核心任务是说明我们如何通过一系列的思维活动而产生做某一行为的动机或意向,戴维森所提出的“欲望—信念”的行动理论模型可以作为我们考虑这一问题的出发点。戴维森认为,我们做某一行动的意向由两部分组成,分别是对某个行动的支持性态度(pro attitude),以及相信他的行动属于那一类的信念。[18]前者即我们对于做某一行动的欲望,后者则是对于实现该欲望之手段的信念。这种对于行动之意向性分析的理论模型此后得到了诸多学者的认同,进而人们相信,一个产生道德行为的意向或动机也需要获得这两方面的分析:某人做此事的欲望,以及相信此事能够实现他欲望的信念。但此时我们将面对的新困难是,该模型并未解释欲望的来源问题,我们可能是由于情感而产生了做某事的欲望,就像休谟所认为的那样;也可能是由于一系列道德原则而产生了做某事的理由,并且该理由导致我们最终选择某事。因此,在戴维森模型的基础上我们需要进一步说明的是,在进行道德判断时哪些因素将影响我们欲望的产生。

第二,道德原则。与机器伦理中的“自上而下的”方式相类似,人类道德的产生也存在这种“自上而下的”理解,即认为我们可以通过一系列道德原则的约束而产生做某事的动机或欲望。这一模式与道德认知主义者关于道德本质的看法相一致,他们相信,道德判断以一种认知的方式起作用,我们通过将一些具有普遍意义的道德原则应用于特定情境即可获知此时应怎样行动。而这些道德原则能够以命题的形式表达为一些具有真值的道德知识,同时具体的道德原则将由某个规范伦理学理论来确定,例如康德式的义务论或者密尔式的功利主义。因此,在道德推理时我们似乎进行了双层的操作,第一层决定选择哪条道德原则应用于此情境,第二层则在先地决定选择哪个规范性理论。需要指出的是,第二层的决定可能是隐秘的,或者是模糊的,因此我们并不总是意识到它。虽然这一考虑道德动机产生方式的观点近年遭到了道德非认知主义者的批评,然

而,我们依然可以为一种弱化的立场辩护说,我们所具的道德动机和意向至少部分是来自道德原则的,否则道德教育就是不可能的。

第三,道德情感。在哲学、心理学的文献中,道德情感被普遍地认为是一个与我们道德判断相关的重要因素,然而不清楚的是这种联系究竟是怎样的,或者说应该以怎样的方式去界定情感所产生的影响。例如,在形而上学的情感主义者那里,他们会认为从本质上相关于我们情感的是道德性质;而在认知的情感主义者那里,他们相信从本质上相关于我们情感的只是道德概念。前者承认一类道德性质的存在,并且这些道德性质由情感所决定;后者则可以不承认道德性质的存在,或者说可以是一个道德非实在论者,但却同时相信我们所使用的道德概念由情感所构成。⑲我们若搁置有关道德的本体论问题,那么根据现有的诸多经验性研究,至少我们能够肯定情感通常与道德判断相伴随,并且能够影响它们的获得⑳,例如,巴斯顿(C. Daniel Baston)的研究即解释了一种通过同情而产生利他动机的假设究竟是如何可能的㉑。这里需要强调的一个区分是,由于情感因素的介入,道德判断的概念并不同于道德推理,并且基于情感对道德行为的解释事实上将构成对传统道德推理学说的挑战,因为后者并不承认情感在道德判断中的作用,而只认同基于理性理由的道德动机。因此,休谟所提出的欲望和信念相分离的问题将始终出现在道德推理问题中,一种直面它的办法就是肯定意志薄弱现象的存在,从而为进一步分析欲望的结构留下空间。

第四,道德直觉。根据近年来一些针对道德直觉能力的研究,有更多的证据表明,我们的道德判断可能并非完全基于一个有理性的推理过程,而是来自一种综合了情感、社会规范以及人类认知特性等因素的直觉方式。格林(Joshua D. Greene)所提出的道德判断的“双加工”(dual-process)理论,便质疑了上述传统的道德推理学说。他指出,人脑具有两种处理信息的模式,可以类比于相机拍摄时所使用的“自动模式”和“手动模式”。这两种模式分别具有效率性和灵活性的优点,因此使得我们人类可以同时很好地应对一般情况和复杂情况。重要的是,格林认为我们的道德判断也因受到这两种处理模式的影响而分别呈现为一种自动的情感反应和受控的、有意识的推理,并且前者将导致义务论的判断,后者则产生后果主义的判断。㉒此外,海特(Jonathan Haidt)发现,当我们被问及做某一道德判断所依据的理由时经常会出现“道德失声的”(morally dumbfounded)现象,即我们并不能准确地描述出如此判断的原因是什么。㉓而这一现象也提示我们,一种基于明确理由的道德推理模式可能并不普遍;相反,我们更大程度上可能依赖于另一种没有明确理由,并且受情绪驱动的道德直觉能力。据此,一种结合了情感和推理能力的双加工理论也许是我们看待人类道德判断能力的更好选择。只不过格林同时也指出,非个人的道德判断以及个人道德判断中基于理由的考虑和直觉之间出现矛盾时,道德推理能力具有重要的作用。㉔

综上所述,人类的道德判断涉及道德推理与道德直觉两种能力,其中道德推理相关于某人所具有的信念、意向性以及道德原则等内容,而道德直觉则相关于某人的情感、情绪以及对于道德情境的识别等。并且海特、格林等人相信,在一般的道德场景中,我们只是使用我们的道德直觉处理相关道德问题,并且主要依赖我们的情感反应;而在一些复杂场景中,当我们的道德直觉无法准确给出相关道德判断时,则会诉诸另一种基于理由的道德推理模式。据此,上文所引入的莫尔分类法中,“明确的伦理智能体”和“完全的伦理智能体”之间存在着设计上的本质区别,前者无论是采用“自上而下的”还是“自下而上的”设计方案,它们都只能在最大程度上实现道德推理的模式,而非另一种以情感为基础的道德直觉模式。同理,当一些学者试图通过在机器中植入人类道德原则的方式而实现人类道德时,即便我们接受此时在机器内部也许会产生一些具有内容的意向状态,但这一模式仍然可能是非人类的,它只是实现了人类的道德推理能力,而非另一种依赖于情感、情绪的道德直觉模式。

五、从道德主体到道德实体

由于一种以推理规则为基础的道德实现方案并不同于人类真正的道德判断模式,因此莫尔的分类法事实上将我们引向了这样一种误区:“完全的伦理智能体”似乎是“明确的伦理智能体”的更高实现,前者是在后者的基础上伴随有意识、自由意志、意向性等。然而,根据我们在上文中所提供的理由,这两种智能体之间存在着根本的区别。“完全的伦

理智能体”被构想为对人类道德的完全复制,因此它至少需要同时实现道德直觉与道德推理两种能力;并且更为重要的是,当我们能够突破“机器意识”这一困难问题而实现了机器人自主的情感反应时,道德推理能力也将顺利地被实现。相反,“明确的伦理智能体”更大程度上依赖于将人类道德规则编入机器算法中,而这就使得机器人不需要那种为人类所具有的“双加工”模式,在面对任何道德情境时都只需要机器人按照既定规则运行即可。因此,从对于道德规则的遵守而言,“明确的伦理智能体”才是“完全的伦理智能体”的更高实现,后者将由于道德情感的存在而出现诸多道德直觉与道德推理不相一致的情况,此时一个近似于人的道德智能体也许会选择根据道德直觉来行动,从而产生一些非道德行为。

此外,即便我们不考虑道德直觉能力在机器中实现的特殊困难,道德推理能力在机器与人身上的体现也是不同的,其中人类的道德推理是“可废止的”(defeasible),我们在遵守某一种道德原则时可以存在许多例外。[25]例如,当我们遇到一个纳粹军官来盘问家中是否藏有犹太人时,便能够打破“不能够撒谎”的规定。在机器中若要实现这种“可废止性”,我们可能需要对规则进行“词典式排序”,或者增加一些具有语境敏感性的特殊规则。然而这两个方案似乎都不可行,前者将依然排除许多特殊情况,例如,我们若将对人类生命的保护置于一切其他规则之上,那么就无法实现“正当防卫”等情形;而后者则需要无数条针对特殊情境有效的规则,这对于计算机程序的编写和运行都是不可能的。

据此,我们提出,一种对于“类人”的人工道德智能体的直觉性期待需要得到调整,因为这并不符合当前人工智能设计的整体计划。从道德实践的角度而言,莫尔分类法中所区分的“完全的伦理智能体”并非一种对于道德规则的充分遵守,而只是一种对于人类道德判断方式的完全实现,并且这种实现在更大程度上依赖于我们对“机器意识”问题的解决,而不仅仅是让机器学会有道德地行动。这一看法也反映了道德和知识间的差异,当我们试图以构建专家系统的方式来设计一个道德智能体时,我们就不能指望这个智能体是在完全地模拟人类道德,因为一个明显的事实是:人类很难成为道德圣人,而机器却是可以的。此外,我们可以通过问答的方式去测试一台机器是否达到了人类智能的水平,但对于机器伦理而言,这种“道德图灵测试”的标准却是远远不足的。[26]我们既不能因为机器没有做某一道德行为就判断它的道德水平是低于人类的,也不能因为它更优地处理了某一个道德问题就判断它的道德水平是高于人类的。

基于这些批判性的主张,我们又应该对于人工道德智能体有怎样的期待呢?事实上,这一问题首先是一个规范性问题,而不是关于我们在技术上能够发展出怎样的人工智能体的事实性问题。或者说,由于人工智能体并不是某种先于人类或者独立于人类而存在的自然实体(natural entities),而是一类人造物(artifacts),因此我们需要提供一些发展它们的规范性标准,而非一个对它们未来之可能性的客观观察。而这些标准的提出和我们应在机器中植入怎样的道德规范是两个界限清晰的问题,前者相关于我们设计任何一种工具时所包含的道德责任,即使是一个“傻傻的吐司机”,我们也需要考虑它伤害到人类的潜在风险,只不过它的自主行动能力远远低于人工智能体而已;相反,在机器中植入道德规范却是一种试图实现真正道德主体的打算,在此过程中我们的目标是让机器成为人类社会中的一员,而非一个工具。

由于这两个问题中都包含了道德规范的成分,因此在讨论一种人工道德智能体的可能性时学者们常常将这两个问题相混淆,从而徘徊在对于人工智能体的工具性态度和主体性态度之间。摆脱这一混淆的一个有效办法就是区分道德实体(moral entities)和道德主体(moral agents)的概念。[27]前者用来表示那些参与人类社会活动,并将同时承担部分道德责任的实体,例如,主人饲养的宠物、科学家在实验室中培育的新生物以及人工智能体等。一方面,它们具有较高的自主性,能够自由地实现一些目标;另一方面,对于它们的道德归责也是受限的,需要考虑到它们的创造者、使用者、管理者的参与等。而道德主体是指那些既具有自由行动能力,同时又是一类非人造物的自然实体,主要包括人类。[28]据此,道德实体即便参与了人类的实践活动,但由于它们不是真正的道德主体,因此只负有部分的道德责任,甚至没有责任。

在此区分下,人工道德智能体的发展路向得以明确:我们应期待某种具有较高自主行动能力的道

德实体,而非某种道德主体。莫尔的分类法却模糊了这一区分,由于他将“完全的伦理智能体”看作“明确的伦理智能体”的更高实现,因此将我们引向了对一种与人类相等同的道德智能体的追求。然而,在上文中我们已表明,在机器中所可能实现的道德判断方式并不同于真正人类的道德判断方式,因此我们并不应该期待“明确的伦理智能体”在未来能够发展成为一种“完全的伦理智能体”,后者的实现更加依赖于对“机器意识”这些困难问题的解决,而非在机器中植入道德。事实上,这两种道德智能体之间的差别正如飞机和一个“仿鸟”(bird-like)飞行器之间的差别。当我们试图设计一架飞机时,我们并不必须完全地模拟鸟类的各方面能力,包括它们的飞行动作、对方向的感知,甚至它们的意识方式等,而仅需要部分地实现飞行的目标。相反,当我们需要设计一个能够完全地模拟鸟类飞行方式的飞行器时,仅满足空气动力学的要求则是明显不够的,我们需要在先地实现对于鸟类意识的模拟,而这被看作是与空气动力学的研究完全不相关的。

六、结论

通过对发展人工道德智能体的支持和反对意见的讨论,我们指出,尽管在机器中实现道德的尝试面对着诸多质疑,我们却始终持有这样一个强烈的直觉:机器应该像人一样有道德地行动。这一直觉反映在莫尔有关道德智能体的重要分类当中,当他将“完全的伦理智能体”看作“明确的伦理智能体”的一种更高的实现时,就暗示了我们在设计人工道德智能体时应朝向某种“类人”的智能体,因为只有后者才是对于我们人类道德能力的充分实现。然而当代有关道德判断的一些经验性研究却表明,人类在处理道德问题时依赖于一种“双加工”模式,其中道德直觉和道德推理都是道德判断所需要的,因此机器中根据规则而实现的道德判断能力并不同于人类,从而一种“明确的道德智能体”也不能够由于设计上复杂度的增加而达到“完全的伦理智能体”,后者更大程度上依赖于“机器意识”问题的解决。据此,莫尔的分类法事实上给我们带来了一些对于人工道德智能体的误解,他将我们引向对于“类人”智能体的期待,却忽视了这两类智能体在设计上的本质差别。

基于这些理由,我们提出,在发展人工道德智能体时并不应该期待一种与人类道德能力相一致的道德主体,因为后者的设计基础不在于道德规则的编码和植入,而在于如何让机器有意识。相反,我们更应该期待一种能够参与人类实践活动的道德实体,它们作为一类人造物而存在,同时也能够因道德的植入而规范地行动。并且更值得期待的是,由于人工智能体并不会受到情绪、疲劳程度、既有信念的影响,因此它们事实上能够成为一种比人类更好的道德决策者,例如,安德森(Michael Anderson)等人所设计的 MedEthEx,就是一种能够给人类提供医疗伦理方面建议的专家系统[29],而我们也不必因为它没有完全模拟人类道德而感到不满。

注释

①Anderson, Susan Leigh. Asimov's "Three Laws of Robotics" and Machine Metaethics. *AI and Society*, 2008, Vol.22, No.4, pp.477-478. ②④Cervantes, JoséAntonio, etc.. Artificial Moral Agents: A Survey of the Current Status, *Science and Engineering Ethics*, 2020, Vol.26, No.2, pp.503-504, pp.512-514. ③Moor, Jame H.. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 2006, Vol.21, No.4, pp.18-21.⑤本文涉及 artificial agents, artificial moral agents 和 moral agents 三种不同的主体或行动者,前两者因被强调为一种类人的智能机器人,故其中的 agent 一词被翻译为智能体;而 moral agents 更一般地表示那些具有自主意识和行动能力的人或者机器人,因此其中的 agent 一词被翻译为主体。⑥Allen, Colin, etc.. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 2005, Vol.7, No.3, pp.149-155.⑦Madl, Tamas & Franklin, Stan. Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture, *A Construction Manual for Robots´ Ethical Systems*. Springer, 2015, pp.137-140. ⑧Brundage, Miles. Limitations and Risks of Machine Ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 2014, Vol.26, No.3, p.356.⑨⑯Van Wynsberghe, Aimee, & Robbins, Scott. Critiquing the Reasons for Making Artificial Moral Agents, *Science and Engineering Ethics*, 2019, Vol.25, No.3, p.724, p.725. ⑩Tonkens, Ryan. A Challenge for Machine Ethics. *Minds and Machines*, 2009, Vol.19, No.3, pp.421-438.⑪⑫⑭Formosa, Paul & Ryan, Malcolm. Making Moral Machines: Why We Need Artificial Moral Agents. *AI & Society*, 2021, Vol.36, No.3, pp.839-851, pp.845-846, pp.846-850.⑬Martinho, Andreia, etc.. Computers Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence. *Minds and Machines*, 2021, Vol.31, No.2, pp.216-217.⑮Malle, Bertram F.. Integrating Robot Ethics and Machine Morality: the Study and Design of Moral Competence in Robots. *Ethics and Information Technology*, 2016, Vol.18, No.4, pp.248-249.⑰Himma, Kenneth Einar. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent?. *Ethics and Information Technology*, 2009,

Vol.11, No.1, p.23.⑱[美]唐纳德·戴维森:《行动、理由与原因》,牟博选编:《真理、意义与方法——戴维森哲学文选》,牟博译,商务印书馆,2008 年,第 387—392 页。⑲⑳Prinz, Jesse. *The Emotional Constructions of Morals*. Oxford University Press, 2007, pp.13-17, pp.19-29. ㉑Baston, C. Daniel. *Altruism in Humans*. Oxford University Press, 2011, pp.30-32. ㉒Greene, Joshua. Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics. *Ethics*, 2014, Vol.124, No.4, pp.695-699. ㉓Haidt, Jonathan. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 2001, Vol.108, No.4, p.817. ㉔Greene, Joshua. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 2001, Vol.293, No.5537, pp.2105-2108. ㉕Grzy, Jarek. *Some Technical Challenges in Designing an Artificial Moral Agent*. Artificial Intelligence and Soft Computing: 19th International Conference. Springer, 2020, pp.485-486. ㉖Arnold, Thomas, Scheutz, Matthias. Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems. *Ethics in Information Technology*, 2016, Vol.18, No.2, pp.103-115. ㉗Johnson, Deborah. Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*, 2006, Vol.8, No.4, pp.195-204. ㉘这里称道德主体主要包括人是因为,一些动物似乎也能够意识到同伴的"不道德"行为,例如对同伴偷取了自己的食物而表现出不满,所以在此并不排除动物中也存在有较高道德意识的道德主体。㉙Anderson, Michael & Anderson, Susan. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 2007, Vol.28, No.4, pp.15-26.

责任编辑:思　齐

On the Development Direction of Artificial Moral Agents

— An Analysis Based on the Issue of Moral Judgment

Yu Tianfang

Abstract: Although the development of Artificial Moral Agents (AMAs) has been criticized, humans still have a strong intuition that machines should be allowed to act morally like humans. This insight is reflected in Moor's taxonomy of AMAs, in which he classified "full ethical agents" as a higher realization of "explicit ethical agents". However, this view is misleading for the development of AMAs, because the design of the two agents is different. The realization of "explicit ethical agents" will depend more on the resolution of the issue of "machine Consciousness" rather than an implementation of moral codes into machines. The reason for this is that human moral judgment ability includes moral intuition and moral reasoning, and the way of embedding moral codes can only realize the moral reasoning of machines, but can not realize their moral intuition. Therefore, a moral entity that can participate in human practical activities is the development direction of moral agents.

Key words: artificial moral agent; moral judgment; machine ethics; ethical agents