doi: 10.15936/j.cnki.1008-3758.2020.03.003

人工智能体"有意不为"的伦理意蕴

王淑庆
(湖南师范大学 公共管理学院, 湖南 长沙 410081)

摘 要: 人工智能的发展使得人们期待人工智能体具有明辨道德是非的部分能力,特别是要求人工智能体能够主动地不做违背道德的行动。基于道德能动性的功能主义,说明人工道德主体在何种意义上是可能的。与人类有意不做某事类似,人工智能体在有意抑制或忽略做某事的时候,也应该出于道德理由考量。由此,人工智能体在"有意不为"上应当具备两种道德特性,即对可能伤害的强敏感性和对道德决策的弱自主性。最后,从控制论的角度设想如何将这两种特性嵌入到人工智能体中去,并论证对于自主人工智能体的伦理控制来说,"有意忽略"是更为困难的问题。

关 键 词: 人工道德智能体; 有意不为; 强敏感性; 弱自主性; 伦理控制
中图分类号: N 031 **文献标志码**: A **文章编号**: 1008-3758(2020)03-0014-07

Ethical Implications of Artificial Agents' "Intending Not to Do"

WANG Shu-qing
(College of Public Administration, Hunan Normal University, Changsha 410081, China)

Abstract: With the development of AI, artificial agents have been expected to own the partial ability to distinguish between "morality" and "non-morality", in particular not to perform immoral actions by themselves. Based on the functionalism of moral agency, it was discussed in what sense artificial moral agents are made possible. Similar to the mankind's negative actions, artificial agents should also be considered for some moral reasons when they intentionally refrain or ignore doing something. Therefore, artificial agents should possess two moral characteristics; namely, strong sensitivity to possible harms and weak autonomy in moral decision-making. Finally, from the perspective of cybernetics, the ways to embed these two characteristics into artificial agents are envisaged, and it is proved that "intentional negligence" is a more difficult problem for the ethical control of autonomous artificial agents.

Key words: artificial moral agent; intending not to do; strong sensibility; weak autonomy; ethical control

为避免人工智能体(artificial agent)对人类产生伤害,从伦理的角度对其进行探讨则成为了一个非常必要的研究主题[1]。人工智能伦理有三个基本问题[2]:一是如何把伦理规则植入到人工智能体中? 二是人工智能体设计者应该遵循什么伦理规范? 三是人们应该如何对待人工智能体? 其中,第一个问题可以看做人工智能体的道德行动问题。让人工智能体的行动符合道德,一种研究进路是构建人工道德智能体(artificial moral agents,简称 AMAs)。如何构建"道德上值得称

收稿日期: 2019-07-09
基金项目: 国家社会科学基金重大资助项目(17ZDA023)。
作者简介: 王淑庆(1986-),男,湖南耒阳人,湖南师范大学讲师,哲学博士,主要从事人工智能伦理、行动哲学研究。

赞的人工智能体”(morally praiseworthy artificial agents，简称 MPAAs)，近年受到了很多学者的关注[3-6]。笔者认为，MPAAs 具备明辨道德是非的能力，这在实践进路①上意味两层含义：一是它“知道”哪些行动符合道德；二是它“知道”哪些行动不符合道德。目前的研究主要集中在前者，但后者其实更能体现出 MPAAs 的自主选择能力②。比如，当一个人工智能体在与人类的交互中产生了“上瘾”[7]行动后，就需要它自主地进行抑制(refrain)，否则可能引发伦理风险；再如，当一个人工智能体在进行多项任务而面临突发状况时，应当主动地忽略(omit)某些行动以防止不良的道德后果。从根本上说，这些都是人工智能体的“否定性行动”(negative action)或“有意不为”的问题。

事实上，“抑制”和“忽略”这两类“否定性行动”在人类社会生活中具有重要意义[8]。那么，人工智能体的“抑制”和“忽略”能否具有道德属性以及应当具备何种道德特性？如何把它们嵌入到人工智能体中去？为探讨这两个问题，本文首先说明“有意不为”对于构建道德主体(moral agent)的意义，进而提出并论证“机器有意不为”应具有的两种道德特性，最后从自上而下的伦理嵌入思路探讨“机器有意不为”的伦理控制，以期为人工智能伦理嵌入提供一种新的研究视角。

一、“有意不为”的内涵及其对构建道德主体的意义

构建 MPAAs 面临的首要哲学问题是机器是否可能成为道德主体。目前对此问题已形成针锋相对的两大派别：支持者们宣称，只要机器能够“模仿”人类基于道德理由而实施行动的特性，并“主动”避免做伦理上有害的事情，机器就可以被看做某种道德主体[9]；而反对者们则主张，机器不可能拥有自我意识和意向性，因而机器不可能作为道德主体而存在[10]。笔者更倾向于支持论者，理由有二。首先，从理论上看，在一个共同体中，某存在物道德主体地位的确立，意味着它可以为自身的行动负责。随着人工智能体自主性程度越来越高，人机互动在理论上可能形成一个共同体。因此，如果人工智能体在相当高的自主性条件下，能够基于道德理由而实施行动，则可以把部分道德主体地位归属于它。其次，从实践上看，承认人工智能体有可能成为道德主体有利于尝试制造人工道德智能体，这其实也是不少人工智能实验室正在研究的课题。相反，如果极力反驳机器成为道德主体的可能性，则有可能进一步认为机器伦理也是不可能的，从而某种意义上会阻碍对机器行动的伦理控制。

当然，以上并不能得出人工智能体就是道德主体，而仅仅站在道德行为功能主义的立场上，说明人工智能体成为道德主体是可能的。在此前提下，才有必要考虑引言中所提到的“否定性行动”③对于构建 MPAAs 的意义。下面首先澄清“否定性行动”的内涵，并说明它对于构建 MPAAs 的积极意义。

在行动哲学中，“抑制”和“忽略”是两种最有代表性的“否定性行动”。对“抑制”和“忽略”的各种界定中，笔者认为沃尔顿(D. N. Walton)的界定最为基本。他给出了“行动者 A 抑制做某事”和“行动者 A 忽略做某事”两种界定。其中，“行动者 A 抑制做 a”，当且仅当[11]：

① 行动者 A 没有做 a 从而没有致使事态 q(q 是指成功实施动作 a 后的结果事态)；

② 行动者 A 执行了另一个动作 b，它与实施动作 a 后的结果事态 q 直接相关。

对于“行动者 A 忽略做某事”，他认为需要体现出义务性，即“行动者 A 忽略做 a”，当且仅当：

① 行动者 A 事实上没有做 a；

② 行动者 A 应该做 a。

由此可见，沃尔顿对“抑制”和“忽略”的界定可以概括为：前者是通过做其他的事而抑制做某件事，而后者就是指行动者没有做应该做的事。显然，他没有考虑心智状态在“否定性行动”中的影响。事实上，人类的“抑制”和“忽略”在很多时候是有意图的，否则很难算作行动。由于人工道德主体需要模仿人类行动的机制，因此笔者假定，

① 另一进路是理论进路，即考虑如何把义务论、功利主义等伦理理论嵌入到人工智能体中。

② 行动上的自主选择能力，被认为是人工道德智能体的前提。自主地选择不做某事，当然也是一种选择能力。事实上，选择做或不做都是智能体的能动性(agency)的发挥。

③ 尽管关于它们能否作为一个合法性的行动概念还存在很多争论，但“抑制”和“忽略”的确是现实生活中所需的一个概念，毕竟它们与伦理道德以及法律密切相关。

较高水平的人工智能体的“抑制”和“忽略”都是有意图的。因此,下文假定“有意不为”就是指“有意抑制”和“有意忽略”这两种“否定性行动”的统称。

人工智能体基于道德理由而行动包括两个方面:一是道德理由使得它主动实施某项具体的行动,并把这种行动坚持下去;二是道德理由使得它放弃可以做的某些具体行动,比如抑制正在造成危害的行动或者忽略本来将要去做的行动。可见,这两个方面都能展现能动性,但考虑到第二个方面需要更多的道德敏感性,以及需要更强的能动性,因而“有意不为”更能体现行动者的全部或部分道德主体地位。因此,有必要对人工智能体的“有意不为”进行更深入的研究,特别是探讨它应当具备的道德特性。

二、人工智能体在“有意不为”上应当具备的道德特性

虽然人工智能体的“有意不为”能够展现其能动性,但人与机器在“有意不为”上是不同的,这使得机器在明辨“非道德行动”上也具有特别之处,这至少体现在两个方面:一是不应做“明知”对人类有危害的行动,即尽量避免实施有伤害的行动;二是机器不应在能力范围内对人类正在或可能经历的伤害漠视不管,即尽量不出现“见死不救”的情况。下文从“机器不为”的角度对这两个方面进行探讨,着重论证人工智能体在“有意不为”方面应当具备两种道德特性:对可能伤害的强敏感性以及对道德决策的弱自主性。这两种道德特性分别对应于“抑制”和“忽略”,它们共同构成了人工智能体在“有意不为”上应当具有的最基本的道德特性。

1. 对可能伤害的强敏感性

在人工智能体的伤害问题上,人们达成了高度的一致,即都承认人工智能体不伤害人类是最起码的前提。如何才能使人工智能体不对人类造成伤害?在人类道德行动中,很多时候不去做某事是因为拥有某种情感,比如“害怕法律惩罚”就可以让人抑制做对他人有害的事。正如艾伦(C. Allen)所说:“在人类的行为中,情感毫无疑问起着重要作用。”[3]此外,由于“人工道德性在于构建对价值、伦理和法律等计算机或机器人的活动具有敏感性的AI系统”[12],而敏感性最为明显的就是对情感的敏感。

鉴于以上两点理由,从人工智能体“有意抑制做某事”的角度构建人工道德性,一种可设想的道路是让机器对某些可能的伤害具有强敏感性,即能够及时(有意地)抑制做正在产生较大伤害或可能有重大危害的事情。根据前文的分析,人工智能体抑制做某事可能针对两种情况:一是正在做的危害动作,二是上瘾动作。在这两种情况下让机器的行动体现出道德性,都有必要让人工智能体能够对可能的伤害产生强敏感性。

第一,对于“正在产生危害的动作”来说,人工智能体必须马上停止原动作。要做到这一点,必须让机器建立起对已有危害和进一步可能的危害的高度敏感性。这与人类的行动是非常不同的,因为人类个体发现自己的行动危害他人时并不会都进行抑制,甚至有时是有意对他人造成伤害。显然,人工智能体不应该“模仿”人类行动的这种特性。因此,人工智能体在理想的意义上应该被设计成“在道德上值得称赞的行动者”①,它将尽量模仿人类公认的最基本的道德情感品质(如不应当伤害无辜之人),而对可能伤害的强敏感性,正是最为初级的且直观上比较可行的模仿。

第二,对于机器的“上瘾动作”来说,人工智能体也应该进行适当地抑制。与人类的“上瘾”类似,机器的“上瘾”与正在发生的危害不同,它的不良后果一般不是立马就显现的。机器的“上瘾”行动至少在两方面可能对人类有害。一方面,长时间的动作造成资源的浪费,比如,机器的“上瘾”使得其作为行动者会重复一些不必要的动作,这在耗电与损耗机器材质方面显然是很浪费的。另一方面,机器长久“上瘾”可能引发伦理问题,比如,在人机互动的场合下,机器的“上瘾”动作可能会影响到人类对其的过分依赖。特别是人类本来就是一种容易上瘾的动物,如果得到人工智能体的主动配合,人类反倒会深受其害,并影响到个人与其他人的互动关系,从而造成不好影响。例如,对于有社交恐惧症的人来说,人工智能体上瘾地与其“促膝相谈”,很可能使得此类人更加孤僻。如果人工智能体对“上瘾”具有敏感性,那么它就可

① “道德上值得称赞”虽然是一个抽象的概念,但如果机器的行动不仅能够实现一些基本的道德品质,甚至还能超过多数人的品行,那么就可以说机器的行动在道德上值得称赞。艾伦甚至认为“道德上值得称赞”的机器是人工道德智能伦理的终极目标。

能抑制这种情况的发生。

因此，为使人工智能体能够主动地抑制有直接危害的行动，让它们能够对这些可能的危害具有高度的敏感性，就是一种比较可行的途径。显然，这种敏感性应当与人工智能体模型中的人工情感①对应起来。人工智能之父明斯基（M. Minsky）认为情感是“人类特殊的思维方式”[13]，而人工智能领域的著名学者博登（M. A. Boden）则认为：如果人工智能不能模拟情感，要实现强人工智能就好似做白日梦[14]。事实上，目前已有很多人工智能小组在研究人工情感。至于人工智能体如何才能对可能的伤害具有高度的敏感性，这是伦理嵌入的话题，笔者将在下文探讨这个问题。

2. 对道德决策的弱自主性

尽管人工智能体在可以预料的未来不太可能有意识，但并非不可以有“人工意图”②。如果把“人工意图”植入到机器中[15]，那么机器的动作就可以被看做某种“有意行动”。与“有意做某事”不同的是，人工智能体“有意忽略做某事”更能体现其自主性。如果人工智能体在道德决策上具有弱自主性，那么就能部分地体现人工智能体的道德主体地位。

人类个体一般需要为自己的行动负责，这种责任既可以是道德上的，也可以是法律上的，但是，由于个体能力与道德水平等方面的差异，人们的自主性责任是不相同的。自主性责任与事后责任是两回事情，前者基于自觉，而后者基于道德或法律评价。与人类不同，人工智能体“有意忽略做某事”是无法达到鲜明的自主性特征的。这是因为，在可以设想的未来，人工智能体的道德地位远不如人，它主要是为人类服务，而不是与人类公平地竞争或生活。因此，从伦理的意义上说，人工智能体的自主性仅仅是指道德推理上的。

以人工智能体“忽略做某事”为例，假设人们期待一个人工智能体 a 做某项行动，但是 a 在经过基于工具理性的推理（如效用最大化推理）后，产生了不做的人工意图，并进而导致没有去做此行动。在此情形中，a 的“忽略”就可以看做是行动自主性的体现，它的“忽略表现”就是一种典型的“有意不为”。它产生的伦理后果可能是不好的，甚至是人类无法接受的。解决此种问题的可能途径是降低人工智能体的道德自主性，使得在一些重要问题上，不让人工智能体单独决策，而求助于人的参与。正如珀德斯韦德克（F. Podschwadek）所论证的：完全自主的道德智能体一定蕴含着它会有意做非道德的行动，也就是完全自主的道德智能体在道德上面临着潜在的不可靠性[16]。显然，这与我们期待的构建道德上值得称赞的人工道德智能体是背道而驰的。

以上分析面临一种典型的反驳，即认为以上例子仅仅表明人工智能体中的算法让某些动作停止而已，并不能表明人工智能体具有自主性。比如，知璨·赫（P. Chisan Hew）就认为，在可以预见的技术条件下，人工智能体是不可能具有道德能动性或自主性的。在他看来，任何机器系统都是人设计的，它背后的设计者才具有真正的自主性。从伦理学上说，缺乏自主性的东西，不可能成为道德主体，因为它完全不具备承担道德责任的能力[17]。笔者认为，知璨·赫的反驳还是基于人类“自主性”的角度看待机器自主性和道德性的问题，而不是从机器能动性的角度探讨问题。如果人工智能体的“有意不为”具有伦理意义，那按照前面的分析，其在“有意忽略”上可能产生潜在的道德风险。例如，看护机器人看到病人突发心脏病却“漠视不管”③，这种“忽略”的后果可能是致命的。一种可以设想的方案是限制人工智能体的自主性，特别是让其在道德决策上具有弱自主性。

所谓道德决策上的弱自主性就是指非完全的自主，但又有一定的自主性。比如小孩子在一些重要选择上，一般要听从父母的建议或让父母参与决策，这就是弱自主的体现。对于人工智能体的“忽略”来说，道德决策的弱自主性至少有两个原因。第一，道德决策关系到人工智能体行动的后果，而“忽略”是本来可以避免的事情，如果人工智能体有意忽略做某些事情，这可能会有产生较大的道德风险。第二，按照珀德斯韦德克的观点，完全自主的道德智能体蕴含着它有主动作恶的可能性，这种可能性可以通过“有意忽略做某事”体

① 所谓人工情感，就是让机器的行动模拟人类情感的反应特性。例如，人类在后悔的时候会抑制做已经认定为错误的事情，如果机器在功能上也能做到这一点，则可以认为它具有“后悔”情感。

② 意图是指行动者欲望驱动下做某事的理由，若让机器也能基于意图而行动，则这样的构造就是人工意图在机器中的实现。事实上，智能机器领域已普遍采取了布莱特曼所提出的“信念－欲望－意图”模型。

③ 机器人可以通知病人的亲人或者拨打急救电话等，这些事项应该是未来的机器人可以做到的。

现出来。至少在“忽略”的层次上,构建弱自主性,就有利于防止人工智能体的一些“有意不作为”带来的道德后果与风险。

三、“有意不为”的伦理嵌入

既然人工智能体的“有意不为”应当具备两种道德特性,那么如何把这两种伦理特性嵌入到人工智能体中?在弱人工智能的意义上,人工智能伦理嵌入就是让机器尽量做人类道德上能接受的行动,并避免做人类道德上不能接受的行动。因而,人工智能伦理嵌入的本质是对机器行动的伦理控制。可见,伦理嵌入就与控制论有着天然的联系。这种伦理控制的关键是在人工智能中嵌入一个“伦理模块”①,以影响人工智能体在“有意做或不做某事”上的决策[18]。所以,有必要从控制论的角度探讨如何让人工智能体自动地明辨“有意不做”的伦理后果。

1.“抑制”的原理:基于反馈的闭环控制

从控制论的角度看,抑制正在做的事情属于“事后控制”,基于信息原理和反馈原理的现代控制思想和方法足以应付这种情况。所谓信息原理,就是指以信息接收与传递作为实现控制的基础方法[19]。就人工智能体的“抑制”来说,其信息是先前所做的行动以及环境的各方面数据。反馈方法在经典控制论中比信息方法更为根本,因为实际控制的自动化目标必须基于反馈原理。所谓反馈原理,就是控制系统把被控对象的信息输送出去,再通过监控器获得新的返回信息,然后把新返回来的信息作为输入信息的一部分,进而对信息的再输出产生影响的机制和过程,从而达到预定的控制目的。人工智能体抑制正在做的事情,当然不是凭空发生的,而是由于“伦理模块”感受到一些信息与人类的伦理准则相冲突,从而停止做某事(在停止的基础上可能还需要实施其他行动)。也就是说,如果外部的环境数据不满足一定的伦理要求,则人工智能体就可以通过调节行动来进行改变。

“事后控制”针对的是外部环境或场景的变化,而伦理原则必须事先设置在人工智能的伦理模块中。如前所述,人工智能体的灵活的“抑制”需要以“对可能伤害的强敏感性”为基础。这意味着需要把“人工情感”植入到伦理模块中,这在目前很难办到。然而,“强敏感性”并不需要把所有情感都植入到人工智能体中,只需要把一些比较敏感的情感植入到机器中。例如,如果把“人工害怕”情感植入机器中,当机器面临相关的场景产生“害怕”情感时,从而自动抑制正在做的行动——这可能是人工智能体情感的一种非常初步的形式②。

当然,基于伦理要求,人工智能体进行自我抑制在现实中没有这么简单。难点之一在于,在人工智能体的“伦理模块”中,各种伦理规则之间可能存在冲突。从难处说,电车难题之类的伦理困难对于人工智能体来说异常困难;从一般的意义上说,有些伦理规则并不是直接给定的,而是已给定的伦理规则的逻辑后承,那么人工智能体要准确理解就需要进行逻辑推理,而目前的道义逻辑是不太适应人工智能这种需要的。

2.“忽略”的原理:基于因果知识的开环控制

与“抑制”不同,人工智能体的“忽略”更多地需要事先知悉多数行动的伦理后果,从而“知道”哪些行动必须被禁止做,特别是那些人类不能承受的行动,比如人工智能体的伤人、杀人事件。此外,某些很难逆转的公共危害(如自动驾驶、公共服务机器人的伤人事件),等到其危害事件发生后,再通过信息反馈来实施控制,可能没有伦理控制效果或控制效果极差。此时,基于相关的知识和伦理原则事先控制是可以设想的唯一方法。

人工智能体忽略一项行动的过程本身就是事先控制的过程,因为它是由于“知悉”有些行动不符合伦理要求而主动不做的情况。因此,人工智能体的“忽略”不适用于基于信息反馈的控制,而需要基于因果知识的开环控制。这种开环控制不强调“反馈”在控制中的意义,而注重事先知道相关行动的结果,因此它是以相当多的“因果知识”为前提的。

一般来说,基于因果的开环的控制比基于反馈的控制要难得多,因此在人工智能体“有意不做

① 伦理模块用来专门调控机器行动的道德属性,它其实就是文献[1]中所说的“良芯(良心)”。关于伦理模块的具体论述,可参见文献[19]。

② 至于人工智能体的“害怕情感”的机制到底是怎么样的,目前还没有文献专门进行实验研究,这是因为人工情感目前还处于探索阶段,需要各方面(包括哲学、控制论、心理学、算法等)的进一步研究。

某事”的两种类型中，对“忽略”的伦理控制才是最有挑战性的。按照传统的基于逻辑的符号主义人工智能做法，这需要表示行动的前提与后果，但却面临框架问题①、后果问题等人工智能的典型困难。为了避免此困难，深度学习研究进路能够达到一定的效果。然而，在某种意义上说，深度学习是黑箱控制方法的另一种形式，而黑箱方法在伦理控制上有它的局限性。正如图灵奖得主珀尔(J. Pearl)所说的，基于黑箱方法的概率理论有其严重的理论局限，它不能推断干预和反思，而人类级别的智能更多地是基于因果推理[20]。珀尔就此认为，基于因果推理的控制在人工智能中是很有应用前景的：因果推理对于机器理解人类事务和人机交互，以及机器的道德决策，都具有根本性的意义[21]。根据类似的理由，张坤等人也认为机器学习中的黑箱方法有其局限，有必要探究数据中的因果关系，但传统的人为方法寻找因果关系，过程非常复杂，代价也非常昂贵，所以有必要基于大数据的深度学习，从观测数据中找出因果关系，进而实施自动控制[22]。与自然因果不同，道德上的因果推理对于深度学习方法的挑战更大，至少目前在道德决策上，基于逻辑方法更为可行一些。人工智能体需要在“忽略”上具有“对道德决策的弱自主性”，如何把这一点嵌入到“伦理模块”中，对于人工智能的伦理控制来说，自然是更为困难的问题。

四、结论与进一步的研究问题

根据以上分析与论证，可以得到两个基本结论：第一，人工智能体“有意不为”具有独特的道德特性，这种特性要求人工智能体主动地不实施一些行动；第二，人工智能伦理嵌入的本质是对人工智能体行动的伦理控制，就人工智能体的“有意不为”来说，“抑制”和“忽略”恰好适用于两种不同的控制思路。因此，若期待人工智能体在为人类服务或交往时体现出一定的道德水平，则“有意不为”的伦理嵌入研究在未来具有一定的价值。在基于自上而下的人工智能伦理嵌入研究中，“伦理模块”至少要有两部分内容：一是道德推理所使用的逻辑，比如心智逻辑、道义逻辑等；二是伦理方面的实质内容，最好是形式化的或算法化②，比如形式伦理研究。假如以上结论成立，就有必要对以下两个问题展开进一步研究。

第一，面向人工智能体的心智逻辑(mental logic)研究。所谓心智逻辑，就是指刻画了行动者知、情、意三方面心智状态的逻辑[23]。由于“有意不为”本身和意图等心智状态有关，同时还与行动(体现为意志状态)有关，甚至直接地是情感所引发的，所以知、情、意三方面的综合形式关系及其推理，对人工智能体主动地不做某事来说就是必要的。人类在进行道德决策时离不开道德推理，同理，人工智能体“有意不做某事”也需要道德推理。目前的心智逻辑几乎是分别地来刻画知、情、意三个方面，但人工智能体有意地不做某事，肯定不能只基于某一种心智逻辑，而是要综合进行知、情、意三个方面的推理，并作出道德决策(包括决定不做)。以人工智能体照看小孩为例，如果小孩正处在看似危险的环境中，而它却“忽略”可能的危险，则可将其看做无伦理推理能力的机器——在这样类似的情形中，最重要的也许不是根据命令而实施行动，而是需要基于心智逻辑和伦理规则进行推理，进而有意地做一些必要的保护行动以及忽略一些不必要的行动。

第二，面向人工智能体的形式伦理(formal ethics)研究。形式伦理是指把伦理原则用符号语言表达出来，以便于进行形式分析或推理。形式伦理对人工智能伦理嵌入具有一定的意义。首先，对人工智能来说，哪些伦理原则适合人工智能体，这是人工智能形式伦理的首要研究任务。在根斯勒(H. J. Gensler)看来，人类最基本的伦理原则有九条[24]，但它们显然在目前可以预见的阶段并不适合人工智能体。比如，“不应该做后悔的事”似乎就不适合人工智能体。其次，如何让机器“理解”形式伦理原则，则是人工智能形式伦理的另一个研究任务。不可否认，在弱人工智能阶段，机器不可能完全理解伦理原则，它们只“知道”这些伦理原则与对应行动的实施或忽略关系，以及这些伦理规则之间的逻辑关系。即使要达到这一步，也是比较困难的，因为一条原则对应的具体行动可能在数量上非常大，而且需要机器预测动作

① 一个动作完成之后，哪些状态不会发生改变——这就是所谓的框架问题。框架问题对于人类不是问题，但对于计算机或机器则是一个非常困难的问题。

② 不管是理论进路，还是实践进路，如果不能形式化或算法化，就无法被计算机或人工智能系统“理解”。

的后果。为了解决这些问题,仅仅依靠自上而下的逻辑方法肯定是做不到的,但把机器学习方法和逻辑方法结合,或许能够简化地解决一些问题。

参考文献:

[1] 李伦,孙保学. 给人工智能一颗"良芯(良心)"——人工智能伦理研究的四个维度[J]. 教学与研究, 2018(8):72-79.

[2] Asaro P. What Should We Want from a Robot Ethic?[M]// Capurro R, Nagenborg M. Ethics and Robotics. Amsterdam: IOS Press, 2009.

[3] Allen C, Varner G, Zinser J. Prolegomena to any Future Artificial Moral Agent[J]. Journal of Experimental and Theoretical Artificial Intelligence, 2000,12(3):251-261.

[4] Anderson M, Anderson S L. Machine Ethics: Creating an Ethical Intelligent Agent[J]. AI Mag, 2007,28(4):15-26.

[5] Etzionil A, Etzioni O. Incorporating Ethics into Artificial Intelligence[J]. The Journal of Ethics, 2017,21:403-418.

[6] 温德尔·瓦拉赫,科林·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.

[7] Bringsjord S, Thero D, Sundar N. Akratic Robots and the Computational Logic Thereof[C]// IEEE International Symposium on Ethics in Engineering, Science, and Technology. Chicago: IEEE, 2014:1-8.

[8] 王淑庆,程和祥. "否定性行动"的合法性之争[J]. 哲学动态, 2018(3):102-108.

[9] Floridi L, Sanders J W. On the Morality of Artificial Agents[J]. Minds and Machine, 2004,14:349-379.

[10] 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.

[11] Walton D N. Omissions and Other Negative Actions[J]. Metamedicine, 1980(1):305-324.

[12] Wallach W. Artificial Morality: Bounded Rationality, Bounded Morality and Emotions[M]// Smit I, Lasker G, Wallach W. Proceedings of the Intersymp 2004 Workshop on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. Baden-Baden: IIAS, 2004:1-6.

[13] 马文·明斯基. 情感机器[M]. 王文革,程玉婷,李小刚,译. 杭州:浙江人民出版社, 2016:9.

[14] 玛格丽特·博登. 人工智能的本质与未来[M]. 孙诗惠,译. 北京:中国人民大学出版社, 2017:69.

[15] 徐英瑾. 如何设计具有自主意图的人工智能体——一项基于安斯康"意图"哲学的跨学科探索[J]. 武汉大学学报(哲学社会科学版), 2018,71(6):79-92.

[16] Podschwadek F. Do Androids Dream of Normative Endorsement? On the Fallibility of Artificial Moral Agents[J]. Artificial Intelligent Law, 2017, 25: 325-339.

[17] Hew P C. Artificial Moral Agents are Infeasible with Foreseeable Technologies[J]. Ethics and Information Technology, 2014,16(3):197-206.

[18] Gips J. Towards the Ethical Robot[M]// Android Epistemology. Cambridge: MIT Press, 1991.

[19] 维纳. 控制论:或关于在动物和机器中控制和通讯的科学[M]. 2版. 郝季仁,译. 北京:科学出版社, 1963.

[20] Pearl J. Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution[C]// Proceeding of the 11th ACM International Conference on Web Search and Data Mining. Los Angeles: Association for Computing Machinery, 2017:1-8.

[21] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect[M]. New York: Basic Books, 2018.

[22] Zhang Kun, Schölkopf B, Spirtes P, et al. Learning Causality and Causality-related Learning: Some Recent Progress[J]. National Science Review, 2017(11):26-29.

[23] 潘天群. 意向性、心智模态与心智逻辑[J]. 浙江大学学报(人文社会科学版), 2010,40(6):125-133.

[24] Gensler H J. Formal Ethics[M]. New York: Routledge, 1996.

(责任编辑:李新根)