华南农业大学学报(社会科学版)
2014年第1期 JOURNAL OF SOUTH CHINA AGRICULTURAL UNIVERSITY No. 1 2014
(第13卷) (SOCIAL SCIENCE EDITION) (Vol. 13)

人工智能体的道德确立与伦理困境

王东浩

(南开大学 哲学院,天津 300071)

摘 要: 人工智能体道德的出现引发了人们对传统伦理体制的思考。随着技术不断革新,人工智能体的内涵也不断丰富,它的自由性和独立性日趋进步。人工智能体的安全性凸显,在设计和应用人工智能体的过程中,传统的研发进路已然落后。人们对人工智能体有无道德以及道德去向提出了质疑。人工智能体走向人机一体化可能更有助于人工智能体道德的确立,人机之间的互动与交流对于人工智能体道德所引发的伦理困境而言具有进步意义。人机之间的和谐相处可能是未来人工智能体道德迈出困境的一条路径。

关键词: 人工智能体; 道德冲突; 伦理困境; 人机关系

中图分类号: B82 **文献标识码:** A **文章编号:** 1672-0202(2014)01-0116-07

人工智能体已经充斥在我们生产、生活与学习的各个角落,它们给人类世界提供了极大的便利,但同时人工智能体尤其是机器人的广泛应用也冲击着传统的伦理和道德体制,从而进一步引发了人类对人工智能体道德存在与否的争论。人工智能体以其不可或缺的便利而独立存在,这已成为不容人类忽视的一个事实,那么理所当然,人工智能体所承载的伦理和道德规范也必然占据一席之地。当下的人工智能体尤其是机器人已有一定的意识,可以依据一定的程序行使其判断力,但这与人类的独立意识仍相差甚远。我们期待人工智能体可以发展成完全的道德主体,那么这个主体的道德内涵如何诠释呢?其道德责任又归属何处呢?人工智能体道德冲击着人类传统的伦理体制,它所表现出来的伦理困境正是人类所需面对的问题。尽管人们对于智能体道德这一概念已经广泛接受,但是也有很多人对此研究进路提出了反对意见,尤其是在设计机器人做出道德决定的能力方面。

一、人工智能体面对的伦理困境

计算机科学家 Rosalind Picard 指出"机器的自由化程度越高,就越需要道德标准"[1],因此,人工智能体的道德建构与发展就变得尤为必要。随着技术革新的加快,人工智能体在人类关系中的渗透已越来越深入,这种情况下就更需要开发智能体道德决策方面的能力。但人工智能体道德的提出对于人类传统的伦理体制无疑是一个挑战。基于此,处于道德两难境地的伦理学家虚构的"失控电车案例"[2]可能代表了人工智能体道德在目前所应承受的最大问题。

那么人工智能体道德"向善的一面"究竟是什么?众所周知,在软件运行中计算机病毒已经对人类造成了伤害。以信用卡支付系统(自动储存贸易系统)为例,它是自动化系统的一个典型,但

收稿日期: 2013-09-30

基金项目: 国家社科基金重点项目(11AZD056)

作者简介: 王东浩(1982—),男,河北衡水人,南开大学哲学院博士研究生,主要研究方向为科技哲学、逻辑学。E-mail: wangdhnk@163.com

由于它缺乏道德决策方面的能力,从而已经对我们的伦理生活产生了很大的影响。在当今社会,电脑信息处理方面的技术应用已经深入到我们生产生活的各个领域,包括在家庭中负责照顾老年人的服务型机器人,以及具有射击功能的军用机器人,它们缺乏伦理维度的考量,从而可能会对人类造成伤害。

在严格意义上的自动化系统中,其道德表现完全掌控在设计者或使用者的手里。随着技术的更新,机器变得更为复杂。这时,人工智能体就朝向"功能型道德体"机器人发展,这类机器人有能力对与道德挑战相关的情况进行评估和做出反应。而功能型机器的创造者同样也面临着许多限制,这从根本上归咎于当前技术的发展。然而,我们也应该认识到,至少在理论上,人工智能体最后可能会成为具有权利和义务的,能够与人类相媲美的真正意义上的道德主体。

我们是否想让计算机做出道德决定? 人工智能体自主做出道德决定是否有益? 人类对这一问题的担忧正是技术与人类文化相互影响的一个非常普遍的状况。基于传统技术哲学的人工智能,特别是人工智能体在现实生产生活诸领域的应用也引起了一些伦理方面的忧虑。例如,拟人化的机器人娃娃,机器人宠物、家庭用机器人、伙伴机器人、性玩具、甚至军用机器人,这些人造体在与人类的交互关系中逐步替代了传统的人际关系,这是否会造成一些问题。诸如,人工智能体面临的责任如何分配? 另外,从现在机器的发展趋势来看,人类的发展前景会受到机器的奴役,而技术进步所带来的利益与相关风险所引起的灾难更有待于评估。或者,当风险相对较低时,我们是否要考虑一个预防性的策略? 从历史的观点来看,技术哲学家的主要任务是批判,但是社会的发展却催生了哲学家的转向,也就是帮助把人类所具有的敏感性价值介绍进设计系统,从而进一步实现技术与哲学的交叉融合。

机器人真的能有道德吗? 人工智能体如何才能更进一步,缺乏人类意识和情感方面特质的机器人是否会成为道德主体? 包括哲学家在内的很多人相信,一个"纯粹"的机器不能称之为道德主体。在此,对这一问题我们保持中立的态度。然而,我们相信对于人工智能体的现实需求会促成一个更加务实的研究进路。我们接受一个完全成熟的道德主体(它依靠强有力的人工智能) 或即使是弱的人工智能体,但它必须具备通过图灵测试的能力——这一程序由 Alan Turing 设计[3],也就是用智能化的人类语言对机器进行等量的自由测试——可能会超越当前甚至未来的技术。目前来说,这一技术是需要时间和实践来验证的。然而,人工智能体发展得越迅速,我们就越能够把它置于"操作型道德体"与真正的道德体之间。我们把它标识为"功能型道德体"。我们相信基于人工智能和更多建立在人工神经网络与体验认知基础上的最新进路与传统的符号处理方法进路的结合,能够为"功能型道德体"的发展提供更多的技术支持。

二、人工智能体道德在设计中的伦理进路

人工智能体道德的设计通常有两条进路: 自上而下的理论体系与自下而上的理论体系。任何"自上而下"的伦理理论的实施都会带来计算和应用两方面的挑战。这一伦理体系的中心原则均以规则和义务为基础的伦理观念以及功利主义构造为背景条件。当设计机器人的相关标准时,Asimov 的三条规则已经渗透到我们的思维观念中,但即便如此,在智能体道德的程序设计方面,这些表面上明确的原则也无法应用到实践中去。高标准的规则,诸如黄金律则,康德的道义论,或对结果论的普遍要求,也会由于一些简单的计算而失效。然而,从诸多伦理标准中衍生出的伦理原则对于人工智能体道德的设计有可能起到导向的作用,或者这些原则本身就成为了人工智能体行动过程中的效果评估标准。

另一条关于人工智能体道德的设计和发展进路为“自下而上”的理论体系。它试图去仿效学习、启发和评价这样一个过程。像诸多伦理理论的形成和发展过程一样,人工智能体道德的理论来源也根植于实践,诸如人工生命的发展、机器人技术的应用以及众多智能体应用过程中所表现出来的不同道德能力。自下而上的进路也坚持这样的发展观念,即道德习惯是一种自我培养的模式,它主要是在与其它个体的合作和分享中实现的。然而,自下而上的理论体系同时也面临着挑战,即如何预防学习和进化中不好的行为习惯,并促使其向好的方面发展?

人工智能体道德的设计有其深刻的理论根源,这可以追溯到以美德为基础的亚里士多德的道德观念。它混合了自上而下和自下而上的两个理论体系,并对美德进行了清晰的描述(或者近似完美的描述),而建立在亚里士多德道德观念之上的个体实质上是自下而上的理论体系的产物。人工智能体道德的设计综合了这两条进路,但从根本上说,它应该以自下而上的理论体系为根本,这有助于解决人工智能体在道德设计过程中所面临的困难。试想如果把这一混合体系融入到计算程序中,那么具有类似于神经网络连接机制的机器人很有可能具备道德判断力[4]。

三、人工智能体道德研究的新进展

人工智能体道德的设计进路都强调了推理能力在伦理体制中的重要性。尤其是一些文学作品在对道德心理的描述时也多表现在理性的推理方面。相比之下,直观感性领域却很少涉及,但不可否认,诸如情感,情绪,语义理解和意识等也对人类道德决定的做出起到了重要作用。那么,对于人工智能体而言,哪方面因素才是必要的呢?并且,如果它是必要的,那么又如何在机器中实现。众所周知,关于情感计算,身体认知和机器意识等这些前沿的科技研究,它的目标是设计具有超越理性能力并具有潜在社会技能的计算机和机器人。从这一角度来讲,对于复杂的人机交互来说,感性因素有必要写入计算机运行程序中。然而,到目前为止,这一程序在机器人设计中还没有实现,也没有这样一个链接情感的机制,因此具备道德能力的兼具社会技能或身体认知的人工智能体还没有出现。

最近,出现了这样一个研究思路——人类认知的综合模型,它的目标是解释高阶认知功能,诸如思考和计算。对于高阶认知的计算程序来说,道德决定的做出在众多具有挑战性的任务中是可论证的,并且可以应用到人工认知领域比较复杂的计算模型中。迄今为止,关于人工智能体道德的介绍最为特别的是 Stan Franklin 的 LIDA 模型[5]。LIDA 同时提供了一系列计算工具和人类认知的模型,并衍生出具有解释能力的机械装置,即面对某一形势时,具有在从自下而上的进程中收集感性数据而做出决定和通过自上而下的进程做出决定这两者之间选择的能力。LIDA 模型也支持与人类情感的互动与融合,同时也会考虑到相关的伦理与道德因素,这进一步论证了人工智能体在未来某个时间可以做出道德决定的可能。

当然,计算机做出道德决定也包含了许多难以预测的风险,因此,它仍旧需要进行检测和控制。这样以来,在其安全性、发展方向和速度等方面就需要国家政策和社会监督机制发挥作用,并据此约束人工智能体朝向一种规范的进路发展。当然这并非是基于对未知的恐惧而阻止科学探究,而是为了在科学研究中摒恶扬善,从而激发和促进对人类有益的科学探索。另外,科幻小说和影视中对于人工智能体的描述,以及人工智能体在其中演绎出来的善恶表现对于我们防范真正的危险也是很有必要的。科学探究中难免会出现风险,因此对于机器人和它们的设计者来说,有必要在其权利和义务的关系之间进行制约和维护,赋予其权利的同时更要求他们履行为人类服务的

义务。这就需要在人工智能和人类智能之间寻求一个可能的“技术奇异点(technological singularity) ”[①],从而使人工智能体向具有道德意识的半机械人状态发展,而非超越人类,控制人类。

尽管在此我们强调的是人工智能体道德的发展前景,但是我们也相信它的实现是建立在技术进步的基础上[6]。人工智能体的道德是以作为道德主体和具有伦理特性的人类的一种自我反馈,而当前发展人工智能体道德的伦理体系存在一定的局限性,这也促使一些深层问题进一步突出。

四、人工智能体道德引发的争议和回应

人工智能体道德的出现,大概引发了以下五类争议: (1) 对于人工智能体道德而言,专家还没有就其研究范围与特征进行充分详当的论述; (2) 人工智能体的道德属性在计算机系统和机器人中不可能付诸实施; (3) 人们提出的关于发展人工智能体道德的进程太过于强调人的作用。机器人需要道德规则的约束,但不一定要完全和人类道德相一致; (4) 研究者工作的焦点集中在确保对人类有益的技术创新而非人工智能体的道德内涵; (5) 从人工智能体道德的现状分析,我们在使用机器人的过程中所遇到的危险是可以解决的,然而,许多危险却不能提前预防。换句话来说,人工智能体道德给人这样一个幻觉,就是存在一个技术修正,可以延缓或阻止有害系统的发展。

基于人工智能体所引发的一系列争议,我们从以下几个方面做出回应:

(一) 独立的人工智能体道德主体

人工智能体道德决定的做出有很多途径,分别体现在自上而下的理论推导规则、自下而上的聚合方法、不断学习积累的进程以及对超理性能力塑造等方面。那么人工智能体的道德属性如何划分和定义呢? 围绕这个问题有人提出“完全的道德主体所要求是”。这个空格需要具有美德以及道德社会共同特征的词语来填充。发问者认为专家可能忽略了这一问题,或者对这一问题缺少必要且充分的论证,从而以“同情心”或“良心”予以代替。

另外,还有这样一些批判,特别是来自于康德的观点认为我们谈论的与主体相关的道德是由于对潜在的不道德行为的选择而生成的。对于这一观点,我们认为人类道德的关键是在利己主义行为和对他人负责的行为之间进行斗争,甚至反对利己主义。通常人类道德生活包括两个方面,即自由选择自己的行为和接受因此而产生的责任。这一道德观念对于人类社会的运行是必要的,同时对于机器的发展也具有促进的作用。另外,我们认为机器的道德观念可以在哲学上描述,并应用到工程设计。

同时,我们认为把人工智能体看作是完全的道德主体是非常有趣的事情。在不同的社会形式中,人工智能体也会以一个完全的道德主体的形式而让人接受,甚至于机器人也会被看作是道德主体。但是,无论是人工智能体或者是机器人,它们在现实世界中以道德主体的形式出现还需要一定的时间。我们认为,人工智能体道德决定能力的做出会随着自动化系统的进步和优化而发展。但这个优化的过程是基于哪一个平台或策略却是一个未知数。完全道德主体是一个迷人的主题,鉴于现在或未来技术的发展,我们期望它能够成为一个更安全,更尊重道德价值观,越来越自主的机器人。

(二) 现有计算机平台的内在限制

从约翰·塞尔(John Searle) 的中文房间(Chinese Room) 的实验思想,到罗杰·彭罗斯(Roger Penrose) 提出的电脑中纯粹智能的可能性提议。有人据此得出人类的智力基本上是依靠超越计算

① 技术奇异点是一个根据技术发展史总结出的观点,认为未来将要发生一件不可避免的事件——技术发展将会在很短的时间内发生极大而接近于无限的进步。

机能力的量子力学原理。但不少理论家认为,现有的计算平台根本无法捕捉到人类智力和心理活动的基本特征。尽管如此,最近的一些批评人士仍旧认为,人类在直觉上理解数学概念的能力与纯粹智能体进行创造性工作的能力本质上是相同的。人类的理解力超越基于规则的系统不会引起争论,也就是它超越了所有基于规则和演算的系统对于我们而言都是正常的。但对于机器而言却未必如此,因为规则不能够直接复制到机器中——只有完全的道德主体是以规则为基础的。此外,即使我们在将来接受基于规则的系统(也就是机器能否拥有我们研究中的多种策略,比如自下而上的认知方法,或超越理性的能力等),人工智能主体也不会轻易就范,因为迄今还没有一个成功的推理和判断的系统模型。尽管人类是聪明的并具有创造性,但是在多重认知任务中,仍旧存在超越规则和演算系统的情况。事实上,有很多任务仍旧超越我们当前在计算机设计方面的能力,我们只是察觉到了这些问题,但对于敏感的计算机程序的开发我们现在仍然无能为力。尽管如此,基于人类固有的创造力,在将来也可能会找到协调的方法,从而能够复制人类的判断力,实现机器人向功能型道德的转向。

(三)人工智能体道德与人类道德并不完全一致

在对人工智能体道德的研究设计中,我们提到借助"自上而下"或"自下而上"等伦理体系去发展类人型道德。这样的结果如何暂且不论,我们先开展一下自我批评。也就是说我们的方法是否太专注于对机器人的道德建设。我们不妨转变一下思维模式,例如,Peter Danielson[7]提出一个非常合理的可能,即机器的不断进步与复杂化对于人工智能体道德而言,其表现形式的实现可能会更加困难,限制也可能会更多。在这一点上,我们同意 Danielson 的观点,虽然上面谈到了关于人工智能体美德方面的问题,但是我们应该认识到这与 Danielson 等类似的批评者所提出的问题还存在很大的差异。可贵的是,我们在此对这一问题的讨论可能会开启人工智能体道德主体设计方面所忽略的一些细节问题。因此,假设技术继续发展,机器人具有人类可能没有的感知能力、计算能力和机动能力,那么我们相信对于人工智能体道德主体进行适当的限制可能胜过对其一味的批判。

(四)技术革新的目标是创建友爱的人工智能体

人工智能体是由保守的计算机科学家建立,在其道德属性上他们仅限于基本的挑战而未超越某一阈值。相比之下,激进一些的人则相信在未来某个时间类人机器和超人系统将会出现,那么人类如何应对未来这些复杂的计算机和机器人系统呢?对此学术界已经有相关的理论和评价,并进一步认为这些将在未来二十至五十年间实现。从道德力的不同表现来分析,在不同的阈之间,机器人安全性的维护方面,我们有两个分组:机器伦理组织和奇点研究机构(singularitarians)(友善的人工智能),以后者为例,奇点研究机构的人工智能(Singularity Institute for Artificial Intelligence)(SIAI)其系统存在一定的危险性,但它们比人类要聪明。

基于数学模型的 SIAI 能够对人工智能在未来发展中所体现出来的利弊进行有效预测。Eliezer Yudkowsky[①]创建的 SIAI 其中一个项目是稳定人工智能系统先进形式的目标。如果预期模型和策略能够在目标体系内稳定发展,那么人工智能体作为道德主体的价值就是显而易见的。但是它们是否能够进一步发展,并且这一策略在人工智能应用之前,是否要超越其它的技术阈值?没有人会质疑机器在学习的过程中会促进人工智能其它技术的发展,包括道德决定的做出。但机器设计理论是不断革新与进步的,而其应用却具有推测性。因此在此我们强调不同组织之间的协调与合作,从而推动友善人工智能的发展。

① 埃利泽:什洛莫(Eliezer Yudkowsky)(出生于1979年9月)美国加州大学伯克利分校计算机科学和人工智能专家,倡导友好人工智能,在2000年共同创立了非营利性的奇异人工智能研究所(机器智能研究所)。

(五) 对于危险的人工智能系统幻想存在技术修正

在所有的批评家中,Deborah Johnson 对于我们建构人工智能体道德中技术上的缺陷反应最为强烈——在没有人类的直接控制下,人工产品假定可以自主做出道德控制的决定——而不是依赖嵌入其整个技术系统中人类的运作。Johnson 还认为,机器人所承载的能力应该考虑到潜在的危险。然而,机器人不是孤立的,人类在进行机器道德设计的过程中是以复杂而相互连接的机制为中心,并以整个社会运行系统为基础的设计理念。

同样,David Wood 和 Erik Hollnagel 认为机器人和它们的操作者是一个联合认知系统(JCSs)。如果对机器自主性进行孤立的关注则会扭曲 JCSs 系统中设计的价值。随着人工智能复杂性的强化,当一个 JCS 失败时,就会出现责备人类作为一个弱连接的事实,从而建议进一步增加机器设备自主性的解决方案。而且会出现这样一个错觉,即增加机器设备的自主性将允许设计者逃避人工智能体行为所产生的责任。机器在遭遇新的或惊奇的挑战时,往往会变得较为脆弱。这就需要人类操作者对于机器在新的形势下做出的行为具有预见性,并能有效协调机器人的行为使其发挥到最佳状态。然而,因为系统变得越来越复杂,人们对机器人行为的预测就变得很困难,从而潜在地增加了 JCSs 的不稳定性。所以,关注孤立的自主性会导致 JCSs 方向迷失。因此,Woods 与 Hollnagel 提议对于 JCSs 的设计更为关注的是它的弹性和协调性[8]。

对于这些批判,专家们感到非常遗憾。在设计机器人的过程中确实应该考虑一下人类在这个过程中所应担负的责任。机器人在从工厂制造到发展成自由漫步的人工智能体的过程中发展过快,这就使得人工智能体道德在设计的过程中无法与潜在系统的发展相协调。基于此,在人工智能体道德主体发展的未来,我们应该反思整个生产和使用它们的体系,从而建立一个良好的机器人应用机制。

我们希望那些忽略由自动化系统所产生危险的机器人专家把敏感的道德情感注入到机器人设计中去。如果对保障措施进行充分严格的设计和检查,那么相关危险就不会出现在实践过程中,所以,我们应该把从依赖自动化系统的注意力转移到社会系统中。

五、小　结

人工智能体道德的出现和发展进一步凸显了人类伦理的不完全性、以及人类对于道德完全性的追求。我们应该从人工智能体道德设计中反思人类伦理体系,并进一步完善它。人工智能体在近期不会也不可能发展成完全的道德主体。然而,我们对那些批评家所提出的人工智能体在道德属性上潜在的危险问题还是应该予以关注。

首先,人工智能体的设计,尤其是机器人的设计和应用必须遵守“机器人的三大定律”。以此为法则,在“不伤害”①的前提下[9],培养人工智能体的自我决策能力。

其次,人工智能体的道德问题涉及伦理和哲学等层面。在这些理论问题的实践应用中,人工智能体与人类和大自然之间的关系必须秉承“和谐”的原则,这有利于人工智能体深化服务质量,满足人类需求。

再次,人工智能体的道德属性根源于人类自身的道德认识。从根源上健全人类自身的伦理体系是必须而紧迫的,同时构筑有效的权力保障措施也是道德标准得以实现的必由路径。除此之外,对于人工智能体自身的发展而言,也需要一个与其相适应的伦理道德体系。

最后,人工智能体的发展必须有一个相对应的伦理道德体系。

---

① 甘绍平关于应用伦理学最核心的价值原则是“不伤害”原则的思想给了我们以较大的启发。

新技术的开拓和应用革新了人类的生产生活方式,提高了人类的文明标准。但技术上的进步并非人类生活的全部,更为重要的是人类精神文化方面的追求和满足。因此,人工智能体的发展乃至成熟,需要有完善的道德伦理体系予以支撑。这样不仅可以弥补人工智能体发展过程中所面临的诸多漏洞,对于人类整体的善恶标准也是一种维护和更新,从而有助于加速"良心的革命和伦理学范式的转换"[10]。在人机和谐与一体化的未来,聚焦于人工智能、人工生命或机器人的伦理问题研究,有助于和谐人类、和谐自然与社会的出现。

参考文献:

[1] PICARD, ROSALIND. Affective Computing [M]. Cambridge MA: MIT Press, 1997: 19.

[2] FOOT, PHILIPPA. The Problem of Abortion and the Doctrine of Double Effect [J]. Oxford Review, 1967, (5): 5 – 15.

[3] TURING, ALAN. Computing Machinery and Intelligence [J]. Mind, 1950, 59: 434 – 460.

[4] DEMOSS DAVID. Aristotle, Connectionism, and the Morally Excellent Brain [EB/OL]. (2010 – 05 – 18). http://www.bu.edu/wcp/Papers/Cogn/CognDemo.htm.

[5] FRANKLIN, STAN, BERNARD BAARS, UMA RAMAMURTHY, MATTHEW VENTURA. The Role of Consciousness in Memory [J]. Brains, Minds and Media, 2005, (1): 1 – 38.

[6] WALLACH, WENDELL. Robot Minds and Human Ethics: the Need for a Comprehensive Model of Moral Decision Making [J]. Ethics and Information Technology, 2010, 12(3): 243 – 250.

[7] DANIELSON, PETER. Can Robots Have a Conscience [J]. Nature, 2009: 457 – 540.

[8] WOODS, DAVID D, ERIK HOLLNAGEL. Joint Cognitive Systems: Patterns in Cognitive Systems Engineering [M]. Boca Raton, FL: Taylor and Francis Group, 2006: 23.

[9] 甘绍平. 应用伦理学前沿问题研究 [M]. 南昌: 江西人民出版社, 2002: 19.

[10] 卢 风. 应用伦理学——现代生活方式的哲学反思 [M]. 北京: 中央编译出版社, 2004: 504.

The Ethical Dilemmas Raised by Artificial Intelligence

WANG Dong-hao

(*Faculty of Philosophy, Nankai University, Tianjin* 300071, *China*)

Abstract: The emergence of artificial intelligence unlocked people's thoughts on traditional ethical ideas. With technology innovation, artificial intelligence has enriched its connotation; it's freer and more independent, which has protruded the problem of safety. While in the process of design and application, the traditional developing approaches have fell behind. People are beginning to question the morals and the whereabouts of the related ethical issues. Human-computer integration may be helpful in establishing artificial intelligence ethics and human-computer interaction and exchange may have a positive role in solving the ethical dilemmas raised by artificial intelligence. Harmonious coexistence between man and machine may be a path taking artificial intelligence out of moral dilemma in the future.

Key Words: artificial intelligence; moral conflict; ethical dilemma; human-computer relationship