·技术哲学·

道德计算是否可能
——对机器伦理的反思

陈 锐,孙庆春
(重庆大学 法学院,重庆 400045)

摘 要:机器伦理是解决人工智能伦理困境的新方案,其倡导者主张,道德是可计算的,通过计算不仅可确保机器合乎道德地行动,而且可更好地探求道德的本质。但通过对机器伦理诸假设的分析发现,道德计算在方法论上面临着严峻挑战,不仅无法确保机器的道德行动,而且会动摇机器伦理的必要性基础。更重要的是,道德计算主义必然将人类导向一种旁观者立场,这或许是一个满足机器伦理要求的道德解释,但对于人类来说,却是一个潜藏着系统性风险的道德陷阱。因此,基于道德计算主义的机器伦理并不是一个解决人工智能伦理困境的有效方案,我们仍需要将伦理规制的重心放在人类设计者身上。

关键词:人工智能;机器伦理;道德计算;方法论困境;道德陷阱
中图分类号:N031 文献标识码:A 文章编号:1674-7062(2020)04-0074-07

人工智能在给人类社会带来巨大福祉的同时也伴随着一定的风险,这将人类置于一个道德困境之中,即如何在发展技术以获取福利与抑制欲望以规避风险之间进行选择?机器伦理似乎表明,人类找到了一个两全的解决方案,即通过道德计算在机器上再现人类伦理。根据其倡导者的论述,机器伦理不仅可以确保机器的道德行动,而且可以更好地探求道德的本质。但是,经过深入考察发现,这种道德计算主义在工程实现上面临着严重的方法论困境,不仅无法确保人工智能积极的伦理效果,反而可能因扭曲了人们对道德的普遍理解而导致系统性的道德风险。因此,我们认为机器伦理并不是一个解决人工智能伦理困境的有效方案。为此,本文首先介绍了机器伦理及其关于道德计算的基本假设,然后分析了这些道德假设存在的方法论局限及其潜在的系统性道德风险,最后是对机器伦理进一步的批判性反思。

一 机器伦理的道德计算观

随着智能机器愈发"自主",人们开始担忧其是否会做出不道德的行为。在应对人工智能"恶"的一面时,伦理学首先提出的方案是传统的计算机伦理,它认为人工智能只是人类智慧的产物,对其规制的关键仍在于对人类设计者进行伦理约束。但是,由于担心更加自主的机器将使传统的计算机伦理在规范机器行动方面难有作为,因此一些学者提出了机器伦理这一新的解决思路,即存在"作恶"可能的机器本身应该具有自我约束的道德能力。

机器伦理主要涉及两方面的问题:一是如何使机器获得自主的道德能力;二是如何看待机器的道德地位[1]。虽然问题意识不同,但二者之间具有内在的因果关联,即关于机器的道德地位的主张往往建立在关于机器的道德能力的主张之上,因此这种具备道德能力的机器通常被称为"人工道德行动

【收稿日期】 2019-11-11
【基金项目】 中央高校基本科研业务费(2019CDSKXYFX0040)
【作者简介】 陈 锐(1968-),男,安徽安庆人,重庆大学法学院教授、博士生导师,研究方向为法哲学;
孙庆春(1988-),男,山东聊城人,重庆大学法学院博士研究生,研究方向为法哲学与人工智能伦理研究。

者”或“道德机器”。根据詹姆斯·摩尔提出的标准,道德机器可划分为隐性伦理体与显性伦理体,前者是指通过内部功能隐含地促进道德行为的机器,后者是指具备明确的道德原则并能依此做出可信的道德判断的机器[2]。基于此,迈克尔·安德森等认为道德机器设计的目标正是显性伦理体,它具有类似于人类运用道德准则进行道德推理的能力。这不仅使机器“能够通过诉诸伦理原则来解释为什么一个特定的行为是对的还是错的”,而且可“在新的情况下,甚至在新的领域中”确定合乎道德的行动[3]。由此可见,机器伦理的最终目标是,通过赋予机器道德能力以使其像人类一样自主地进行道德判断与行动。

为此,机器伦理的研究者提出了诸多实现这一目标的方法,如康德式方法、范畴理论方法以及表面义务的混合方法等。总体来说,这些设计方案可归为三种主要的研究进路。一是理论进路,其希望通过预设道德原则、规范的方式来规制机器的行为。其次是实践进路,即设计者希望人工智能在一个道德环境中进行自主探索与学习,从而在不断地道德实践中形成相对稳定的道德原则。最后是混合进路,它主张先赋予机器以必要的道德原则,然后依此指导机器自下而上的学习实践。

然而,无论采用哪种进路,道德机器的设计都将归诸计算主义,这是机器伦理总的方法论。计算主义的经典假设是,人类心灵是一台数字计算机,只要恰当地模仿,就可以在计算机上再现人类思维品质[4]。相应地,机器伦理在道德领域坚持了一种道德计算主义,即道德是可计算的,通过计算可以使机器再现人类伦理[5]。具体而言,这一主张包括了以下几个方面的假设:(1)道德现象可以还原为客观、抽象的自然事实。例如卢西亚诺·弗洛里迪等认为,对抗信息熵是一切道德行动者要遵循的一般道德法则,它揭示了信息熵与道德性之间具有一种负相关关系,即熵值越高表明越邪恶,熵值越低表明越道德[6]。(2)道德问题的相关特征解释项可以进行明确的表征。例如安德森等假设,伦理的相关特征提示了正确的行动方针,在一组给定的行为中,能够最小化或最大化某一伦理特征的行为便是正确的道德行为[3]。(3)道德判断与道德行动完全可以通过逻辑推理与审慎的思考过程来达成,不需要道德情感的参与。例如科林·艾伦认为,情感对于人工道德行动者来说并不重要,因为情绪等感性因素通常在决策中扮演负面的角色,因此没有情感影响的机器可以做出更好的道德决策[7]。(4)道德问题的解释不需要考虑主观精神因素,只需要关注客观行为及其后果。例如,安德森等明确表示意向性与自由意志并不是在道德困境中做出正确道德行为的必要条件[3]。

显然,机器伦理基于计算主义提出了一种形式化的道德假设,它排除了道德现象中的情感、意识、意义体验等,而关注客观的行为及其背后的计算要素。这种道德计算观不仅为机器获得道德能力提供了方法论上的支撑,而且作为一种道德解释,可以更好地探求道德的本质[3]。如果道德计算主义的上述假设为真,那么道德问题就能像莱布尼茨所期待的那样,计算一下就够了。但是,计算主义的道德假设能否成立?这些假设是否像机器伦理的倡导者所说的能更好地解释道德的本质?更重要的是,被赋予道德原则的机器能否确保其道德行为,从而解决人工智能的伦理困境?这都需要我们进行批判性的反思。

二 道德计算面临的方法论挑战

从当前研究来看,道德计算主要面临两方面的问题:一是如何确定一个理想的道德框架,二是如何将这一道德框架赋予机器。机器伦理在道德理论、背景表征、道德心理学以及主观能动性等方面的假设都是为了解决上述两个问题。但是,这些假设存在固有的方法论局限,难以从根本上满足机器伦理的要求。

(一)道德事实能否还原为自然事实?

机器伦理主张道德现象可还原为诸如客观行为、环境、后果、信息这样的自然事实要素,通过对这些要素的计算解释就可以指导机器做出理想的行动选择。但是,这必然会面临一个方法论挑战,即能否将道德事实还原为纯粹的自然要素?

这既是机器伦理面对的一个技术性问题,也是一个道德理论难题。一方面,将所有的道德事实还原为自然事实可能是一个难以完成的任务。首先要面对的问题是,我们无法确定将道德事实还原为哪一种自然事实。比如,我们能否把“善”还原为“快乐”,把恶还原为“痛苦”?这些还原性定义都会面临摩尔“开放性问题论证”的检验,它表明任何一种还原性定义都是不充分的。而且,这种还原会带来诸多伦理上的争议,从而难以达成共识。正如斯特金所说的,“要确信有这样一种(还原性的)定义,我们就必须要知道哪一种伦理学理论是正确的。但是,伦理学是一片充满极大争议的领域,我确信我们

还没有能力对此予以确认”[7]242。从某种意义上来说，争议性是伦理学或伦理问题的必然属性，这源于人们不同的道德直觉与价值观念。这正是机器伦理在道德理论框架上无法达成共识的深层原因。

另一方面，将所有道德事实还原为自然事实可能无法充分解释道德现象。尽管还原论认为一切道德性质与道德事实都可还原为自然性质与自然事实，但是我们无法对此进行确认。而且，考虑到对道德现象的解释时，我们反而不能确定道德性质与道德事实是多余的，因为这无法充分且合理的解释道德现象。以弗洛里迪的信息熵为例，我们无法从这样一个纯粹的数值中体会到行为的善恶，也无法用这样的物理描述来说明德性。如斯特金所言，“倘若存在任何连续的物理参量，那么世界就有无限的物理状态，然而任何语言，甚至包括理想的物理学语言，充其量只有有限的谓词”[7]240。所以，对于那些无法用物理语言来指代的部分，仍然需要其他的性质与语言来说明。同样地，我们也不能将这种道德性质与道德事实还原为生物学的或其他的性质与事实，正如我们不能因为细胞可以还原为更为基本的原子，而否认细胞的独立存在与属性一样。

因此，将道德性质与道德事实还原为自然性质与自然事实既面临着方法上的挑战，又面临着理论争议。这表明，机器伦理不能也不应将一切与道德事实相关的东西都解释为一种自然事实或几种自然事实的组合。

（二）道德背景能否被充分表征？

机器伦理的倡导者希望通过精心设计的算法来赋予机器正确行动的道德能力，包括在那些结构开放的道德情形中。但是，其所采取的表征主义技术进路决定了它们可能无法良好地实现机器伦理的最终目标，因为它们都无法回避现象学对表征主义明晰性假设的批判。

在机器伦理的设计中，为了理解道德问题必须将其结构化，或者为了识别道德情境必须对数据进行选择与解释。但是，我们如何把这个环境本身传送给机器呢？这正是现象学对表征主义发出的灵魂质问，也是表征主义的根本困难。这种困难表现在两个方面：一是机器是一个形式化的系统，它的任何行动都以形式化为前提，但并非现实世界中的任何活动都可以被形式化；二是机器的无身性使其孤立于环境之外，而无法把握作为认知与行动所必要的背景知识。其结果便是机器在道德能力上的严重缺陷：一是道德机器只能应对相关特征比较明确的道德情境，而无法应对复杂的情形，尤其是结构开放的领域；二是机器的无身性使其无法像人类那样通过整体感知来把握环境的全局与细节。因此，即使其可以模仿那些可明确表征的所谓高级理性功能，也难以驾驭那些低级功能。值得注意的是，随着联结主义的崛起，机器伦理的倡导者试图通过人工神经网络来建构具备学习能力的道德机器。但是，计算理论从符号主义向联结主义的转向并未摆脱表征主义的技术进路，因此并未使道德计算逃脱现象学的批判。

而且，机器的这种孤立性与形式性使其脱离了环境中的意义网络，也就无法在道德活动中把握真正的道德意义。道德是人类有意义的生活，人们总是发现自己处于一种意义空间之中，我们的道德判断与道德行动总是受到我们意义体验的影响。正如德雷福斯所指出的，环境一开始就是以人的需要和倾向组织起来的，而人的需要和倾向赋予事实以意义，并使事实成为事实。“当我们置身于世界之中的时候，我们是生活在意义世界之中的，那些嵌入这种上下文环境中的有意义的物体，不是存储在大脑或思维的世界模型，而是世界本身。”[8]

（三）道德理性是道德判断的充分条件吗？

道德计算的本质是逻辑，道德机器遵循逻辑推理来对道德情境做出判断并采取行动，而道德情感则被完全排除在外。正如安德森所言，在人类的道德行动中，道德情感往往是作为“冲昏头脑”的情绪而存在的，排除了道德情感的机器可以做出更好的道德决策。但事实可能并非如此。

在某种意义上，机器伦理的主张是道德理性主义在人工智能伦理领域的延续，它强调理性在道德中的根本性地位，同时极度贬低了道德情感。但是，在哲学上，同样存在着道德情感主义，它强调道德情感在道德判断与道德行动中的作用。例如，莎夫茨伯利认为道德上的善恶取决于情感上的善恶；在休谟看来，理性只是为情感提供了发挥作用的手段，只有情感才是引起行为发生的最终原因；此外，还有如赫起逊、哈奇森、斯密、弗洛伊德、斯洛特的阐述都突出了道德情感的重要性。而且，道德情感的作用得到越来越多现代心理学实验的证明。例如，格林等在电车困境与人行桥困境实验中，借助功能磁共振成像手段展开了道德决策中情绪作用的内部机制研究，证明了情绪在特定道德情境中发挥了主导决策与行动的作用[9]。海蒂也在其实验中得出结论，道德行为与道德情感的协变性大于与道德推理的协变

性，而且道德推理很少是道德判断的直接原因。[10]由此可见，道德情感并不像机器伦理的倡导者所认为的那样无用甚至只有消极作用。

那么，道德情感如何起作用？根据现代心理学的经验性证据，道德情感正是通过“移情”机制在道德行动中发挥作用的。例如，感受他人痛苦的“感同身受”的移情反应往往会促使人们去帮助他人摆脱痛苦或避免他人受到伤害。若一个人本身没有感知或体验别人情感的能力，便不可能把握别人的体验状态，也难以做出符合人类需要的道德决策。正是在这个意义上，一些道德理论家认为道德情感产生道德义务。需要说明的是，强调情感的作用并非意在否定理性的作用，而在于指出纯粹的道德推理并不是道德行动的充分条件。情感与理性共同作用于道德行动，缺乏任一机制都不可能使机器做出与人类相一致的道德决策。

（四）计算能否把握道德行动中的主观能动性？

机器伦理面临的技术局限性，使部分倡导者开始关注意识问题。瓦拉赫与艾伦认识到，在当前的技术条件下，让机器对具体规则的有效性做出连贯反应仍然是一个“遥远的梦”，如果机器像人类一样“有意识”，或许就可以解决道德计算所遭遇的前述难题[11]。那么，计算能使机器“有意识”吗？

虽然意识问题被称为世界之结，但计算主义者认为，意识并未超出可计算的范围，只要实现对人脑结构及神经元活动规律的充分的内部表征，创造人工意识并非不可能。但是，就目前来看，意识研究面临着两大问题：一是简单问题，二是困难问题。所谓简单问题是指理解有意识思维的信息处理方面的问题。之所以称其为“简单问题”，并不是因为它们容易解决，而是从功能主义框架内的计算和神经生物机制来看，解决这些问题是可能的。相比之下，困难问题是指意识的主观性问题，即与意识相关的主观体验问题。例如，我们能体验到颜色、红酒的味道、背部的疼痛等。与强调信息加工的功能意识相比，这种主观性体验通常被称为现象意识，对此尚没有令人信服的科学解释。所以，即使科学最终解释了所有的简单问题，也不能解释为什么这些机制或功能不是在没有主观意识的情况下发生的，而是伴随着主观经验。用约瑟夫·莱文的话说，在对意识的成功的计算性描述和伴随它的主观体验之间存在着一个“解释的鸿沟”[12]。

正是这种意识的主观性构成了对人工智能的挑战，同时也构成了对机器伦理的挑战。按照内格尔的观点，主观性是意识所具有的第一人称本体论的性质，与任何产生意识的机制的第三人称叙述存在根本区别，这种意识状态本身并不是感知对象，从而也就不能通过将感官引导到大脑来观察它们，也就无法通过客观的描述或表征来呈现。这种本体论上的主观性明显不同于认识论上的主观性，它是意识的根本特征，但其没有被任何熟悉的、最近设计的对精神的简化分析所捕捉[13]。詹姆斯·雷吉在总结近20年的意识研究时感慨道，“现有的人工意识方法都没有对现象意识进行令人信服的演示，甚至没有清楚地证明人工现象意识最终是可能的”[14]。

（五）小结

综上所述，由于道德计算在道德事实还原、道德背景知识的表征、道德行动的心理机制以及主观能动性方面都面临着方法论挑战，这使得机器的道德能力受到极大的限制，从而无法确保其在所有情境下做出正确的道德决策。或许，机器伦理的倡导者认为，有限理性和满意模型可以解决这一问题，因此只要注意到机器在知识的获得和处理方面是有限的就足够了，因为人类也不具有无限理性。但是，事实却是，这种基于计算的机器和人类在“性能”上存在巨大差距，无论是在精神上，还是在身体上。而且，就保持机器与人类的道德一致性这一目标来看，由于机器本身在意识、感受与情感等方面的固有缺陷，也无法确保其能做出与人类相一致的道德决策。

三　道德计算观的潜在危险

除了在方法论上的挑战之外，道德计算还面临着认识论上的挑战。根据倡导者的主张，道德计算观作为一种更好的道德解释为伦理转向提供了认识论基础。但是，这种“道德思维革命”本身存在着诸多潜在的伦理风险：一方面，道德计算观因贬抑了主观精神状态在道德实践中的决定性作用、扭曲了人们对道德的普遍理解，从而可能在道德评价与道德责任方面给人类社会带来系统性的道德风险；另一方面，道德计算观指导下的机器伦理可能会带来新的伦理问题，从而无法保证人工智能积极的社会效果。

（一）道德计算观导致道德评价的机械化

通过道德计算，机器伦理暗示其更好的阐释了道德的本质。因为它原则上坚持了一种更公正、更普遍的形式化道德解释，即道德不是关于意向性的，而是关于行为及其影响的，一切引起道德上善的或恶的影响的行为（行动者）都是道德行为（道德行动者）[5]。这种道德解释明显建立在计算与思维等价

的观念之上，使得机器伦理无论是在方法论问题上，还是在主体地位问题上都表现出一种形式化与抽象化倾向。

那么，这种方法阐释了道德真义吗？根据普遍理解，道德实践是主观精神状态下的意识行动，而非纯粹的行为。事实上，在道德生活中，人们总是关注行为背后的意图与动机。如梯利所言，我们把道德判断限定在有意识的人的行动上，希望这种行动有一个精神的或心理的基础，只有当行动是一个有意识的人的表现时，我们才能对它进行道德判断。如果行动并非出自行动者的意志，那我们就不对此进行评判。所以，道德判断与道德评价的对象正是人们有意识的行动。比如，刑事司法中定罪与量刑不仅要考虑嫌疑人客观的犯罪行为，也要评价其主观的认识、意志，甚至是动机。然而，在机器伦理的倡导者看来，这种民间心理学无法解释这样一种现象，即无意识行动也可能会造成道德上好的或坏的影响。无可否认这种现象的存在，但是同样需要注意的是，无论是在道德自治领域，还是在法律评价体系中，人们往往对这种无意识造成的道德错误表现出明显的宽容态度，例如心智不成熟的儿童与精神病患者往往不承担刑事责任。这表明，意识行动而非无意识行动才是道德生活的核心，那种抛开意识而谈论道德的观点可能不被人们所接受。

至于意识或主观精神状态的本质是什么，这正是所要解决的问题，而不能为了满足机器的“道德要求”而贬抑其在道德实践中的决定性作用。意识本身赋予人类以道德动力，我们的道德实践不需要某种外在的事物来给我们提供额外的驱动。但机器却不能，它不能脱离我们给其设定的道德意义而存在。作为一种人工制品，机器本身只是存在于人们事先编织好的道德意义空间之内，它不像任何自然事物那样有独立的目的。“在这种意义上来说，机器是什么或机器变成什么，完全取决于人类。”[15] 毫无疑问，人类的道德现象可以在特定抽象层次上进行特征化、形式化，但是这种特征化与形式化并不是道德生活本身。我们不能通过将道德定义在一个抽象化层面，来实现人类与机器的道德等价。这不仅扭曲了人们对道德的普遍理解，而且还可能导致人类的道德判断与道德评价走向机械化。

（二）道德计算观导致道德责任体系的混乱

与机器伦理密切相关的另一个问题是道德责任问题。在倡导者看来，道德行为足以解释道德的本质，从而为机器伦理的主体责任主张提供理论支撑，即可以做出某些在道德上有影响的行为的机器同样是道德行动者，进而可以或应当承担相应的道德责任[16]。但是，我们认为，这种“责任类比”不仅不能真正解决现实社会中的道德责任问题，反而可能导致作为人类社会基础的道德责任体系的混乱。

通常认为，伦理的核心在于责任。用列维纳斯的话来说，伦理关系首先是一种责任关系。正是基于此，机器伦理的倡导者认为如果能将道德责任归于机器，便反向论证了机器作为道德行动者之主张的合理性。那么，将责任归于机器如何可能？根据弗洛里迪的观点，将责任归于机器的可能性在于区分道德问责与道德责任，从而在前者的意义上去解决机器的责任问题。其中，道德问责意在表明“谁或什么是所讨论的道德行为的道德来源或原因”，而道德责任则是确定“该道德来源者是否也对该行为负有道德责任以及在多大程度上负有道德责任”[5]。在他们看来，这种区分不仅是可能的，而且是有意义的。因为，这将给人类社会带来实质性的好处，即当责任事故发生时，“我们可以停止寻找责任人的无限倒退”[5]。

但是，这种道德解释及其“实质好处”的背后却潜藏着系统性的风险。首先，机器可能根本无法获得成为合格的道德行动者所必要的责任能力，从而使归责成为空谈。因为责任能力通常要求行动者具备一定的条件，除了内在的意识、自由意志、感受之外，还应具备必要的物质条件，比如财产等，这些都是机器所不具备的。其次，将机器作为责任主体，即使只是所谓的道德问责的对象，也为人类设计者逃避责任提供了可能。正如乔安娜·布莱森所说的：“将我们设计的人工制品要执行的行为的责任分配给它，将意味着决绝否认我们对该设计的责任。”[17] 最后，过度强化机器的责任主体地位，可能降低人们作为负责任的道德行动者的意愿与能力，从而导致人们从道德行动者降低为道德病者，即“不再或不能对自主制定和追求自己的道德目标感兴趣”[18]。然而，我们有充分理由相信，人类有意义的私人生活与社会实践都是建立在负责任的道德行动者这一基础之上的。由此可见，将道德责任归于机器是不可能的。

（三）道德计算观可能带来新的伦理问题

机器伦理的初衷在于通过赋予机器伦理能力，来确保人工智能的部署产生积极的伦理效果。然而，由于道德计算观在意识、情感与意义体验等方面与人类道德行动者存在实质性差异，大规模部署机

器伦理可能根本无法确保其预期的积极效果。

首先，正如前文已经指出的，道德机器可能在一些道德情境下做出错误的道德决策。这种错误可能是因为道德规则的不周延而在一些不可预测的情况下做出不符合人类伦理的决策，也可能是因为在道德能力方面的缺陷而做出与人类不一致的道德决策，还有可能是因为恪守规则而造成的机械化决策。这些错误决策突出表现了道德机器在自主道德能力方面的不足。其次，机器伦理可能导致安全问题。人类道德环境本身并不纯粹，存在着恐怖主义、极端主义、暴力主义以及其他形式的不道德，这可能使机器在学习过程中培养出不符合人类道德价值观的品性。再次，机器伦理本身可能会带来歧视与偏见等伦理问题。现在机器伦理存在的一个极为突出的问题是伦理学家与工程师们无法就统一的伦理框架达成共识，因为现有的理论都不能令人满意。伦理的这种“未解决性质”将在机器伦理的设计中凸显出来，无论赋予机器哪种道德理论都可能被认为是符合一部分人的道德倾向，而与另一些人的道德直觉相冲突[19]。最后，机器伦理道德决策还可能被指责为过于重视眼前利益而对未来考虑不足，从而产生代际公平问题[20]。

上述问题并不是机器伦理可能产生的全部伦理问题，但这些问题足以说明通过道德计算解决人工智能的伦理困境是非常困难的。不可否认，其中某些问题可能会在伦理学家与工程师们的努力下得到优化，但我们不认为这些问题会通过计算得到彻底解决。因为这并不是机器的问题，而是人类自身的问题。只要我们还是有意识、有知觉、有感情、有身体的生物，我们就不可能像机器一样来解决道德问题。

（四）小结

为了满足机器的道德要求，道德计算观剔除了道德中的主观意识、情感因素以及意义体验等非计算要素。这种形式化的道德解释本身并不是一种更好的道德解释。因为“将道德与计算等价”的道德解释进一步为“机器与人类等价”的观点提供了认识论基础，为“将机器视为道德主体”的观点提供了理论支撑。当我们将这种道德计算观念放在道德的意义空间中进行考察时会发现，它扭曲了人们对道德的普遍理解。这不仅无法实现机器伦理在解决人工智能伦理困境方面的积极效果，而且还有可能带来系统性的道德风险。

四　结语

为了解决人工智能的伦理困境，机器伦理提出了道德计算的假设。不可否认，人类道德具有一种隐喻意义上的计算纬度，但道德本身不是通过计算而是通过体验来获得的。如果完全遵循计算主义法则来解释道德，这本身可能是不道德的，因为计算法则必然逻辑地将人类导向一个旁观者立场。也许，对于机器来说，这是一个倡导者所谓的更公正、更普遍的道德解释，但对人类来说，却是一个道德陷阱，它意味着更大的道德困境。此外，道德计算的方法论挑战也会动摇机器伦理的必要性基础——机器“作恶”，因为这种必要性假设只有在机器具备真正的自主能力时才能成立。也就是说，只有机器具备了能够自主地违背或破坏人类伦理的能力时，才有能力去遵守人类伦理。然而，就目前来看，机器并不具备这种自主性，它是什么以及做什么都未脱离人类的程序设定。这表明，当下存在的人工智能伦理问题的原因不在于机器，而在于人类自身。即使我们教会机器如何明辨是非，可能也无助于解决那些由人类引起的伦理问题，因为机器伦理的解决方案与人工智能伦理困境之间并不具备有效的因果关系。由此看来，建立在道德计算基础上的机器伦理难以实现其预期目标。因此，在机器并不具备自主“作恶”能力的情况下，应将伦理规制的重心放在人类设计者身上，而不是机器身上。

【参　考　文　献】

[1] 杜严勇. 机器伦理刍议 [J]. 科学技术哲学研究，2016，33(1)：96－101.

[2] MOOR J. The nature，importance，and difficulty of machine ethics [J]. IEEE intelligent systems，2006，21(4)：18－21.

[3] ANDERSON M，ANDERSON S. Machine ethics：creating an ethical intelligent agent [J]. AI magazine，2007，28(4)：15－25.

[4] TURING A. Computing machinery and intelligence [J]. Mind，1950，59(236)：433－460.

[5] FLORIDI L，SADERS J. On the morality of artificial agents [J]. Minds and machines，2004，14(3)：349－379.

[6] ALLEN C，VARNER G，ZINSER J. Prolegomena to any future artificial moral agent [J]. Journal of experimental &theoretical artificial intelligence，2000(12)：251－261.

[7] STURGEON N. Moral Explanations [C] // MCCORD G. Essays on moral realism. New York：Cornell University Press，1988.

[8] 德雷福斯. 计算机不能做什么：人工智能的极限 [M]. 宁春岩，译. 北京：三联书店，1986：274.

[9] GREENE J，SOMMERVILLE R，NYSTROM L E，et al.

An fMRI investigation of emotional engagement in moral judgment [J]. Science, 2001, 293: 2105 – 2108.
[10] HAIDT J. The emotional dog and its rational tail: a social intuitionist approach to moral judgment [J]. Psychological review, 2001, 108(4): 814 – 834.
[11] 瓦拉赫,艾伦. 道德机器: 如何让机器人明辨是非 [J]. 王小红,译. 北京: 北京大学出版社,2017: 73.
[12] LEVINE J. Materialism and qualia: the explanatory gap [J]. Pacific philosophical quarterly, 1983, 64 (10): 354 – 361.
[13] NAGEL T. What is it like to be a bat? [J]. The philosophical review, 1974, 83(4): 435 – 450.
[14] REGGIA J. The rise of machine consciousness: studying consciousness with computational models [J]. Neural networks, 2013, 44: 112 – 131.
[15] TOIVAKAINEN N. Machines and the face of ethics, ethics and information technology [J]. Ethics and information technology, 2016, 18(4): 269 – 282.
[16] FLORIDI L. Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions [J]. Philosophical transactions of the royal society A, 2016, 374(2083): 1 – 13.
[17] BRYSON J. Patiency is not a virtue: the design of intelligent systems and systems of ethics [J]. Ethics and information technology, 2018, 20(1): 15 – 26.
[18] DANAHER J. The rise of the robots and the crisis of moral patiency [J]. AI & society, 2019, 34(1): 129 – 136.
[19] BRUNDAGE M. Limitations and risks of machine ethics [J]. Journal of experimental & theoretical artificial intelligence, 2014, 26(3): 355 – 372.
[20] BAUM S. Social choice ethics in artificial intelligence [J]. AI & society, 2020, 35(1): 165 – 176.

Is Possible to Compute Morality

——*Reflection on Machine Ethics*

CHEN Rui, SUN Qing – chun
(*Chongqing University*, *Law School*, *Chongqing* 400045, *China*)

Abstract: Machine ethics is a new solution to the ethical dilemma of artificial intelligence. Its advocators claim that morality is computable, and that morality computation can not only ensure the ethical actions of machines, but also better explore the essence of morality. However, through the analysis of the assumptions of machine ethics, it is found that moral computationism faces severe challenges in methodology, not only failing to ensure the moral actions of machines, but also shaking the necessary foundation of machine ethics. More importantly, moral computationism will inevitably lead human beings to be a bystander, which may be a moral explanation to meet the ethical requirements of machines, but for human beings will be a moral trap with hidden systemic risks. Therefore, machine ethics based on moral computationism is not an effective solution to the ethical dilemma of artificial intelligence, and we still need to focus on human designers.

Key words: artificial intelligence; machine ethics; morality computation; methodology dilemma; moral trap

(责任编辑　魏屹东)