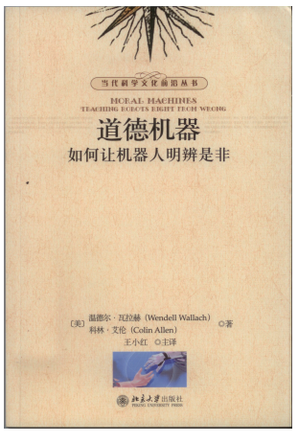
制造“道德机器人”的远虑与近忧

——评《道德机器:如何让机器人明辨是非》

李颖娜

西安交通大学人文社会科学学院,西安 710049

随着人工智能和机器人技术的发展,自主性日渐增强的机器已经出现在金融、军事、电力控制、交通运输等领域,并且在可展望的未来仍将飞速拓展其应用空间。如果说机器获得和人同等水平的道德后,可能出现的种种困境还只是远虑,那么如何让越来越具有自主性的机器根据符合人类道德标准的决策去行动则是近忧了。《道德机器:如何让机器人明辨是非》是美国认知科学哲学家科林·艾伦(Colin Allen)和技术伦理专家温德尔·瓦拉赫(Wendell Wallach)合著的一部机器道德领域的开创性著作,其关注点已经超出了人类如何使用计算机之类的应用伦理学问题,而跨进到该如何建构有道德的机器这一直指实质的问题。



科林·艾伦,温德尔·瓦拉赫 著,王小红 译。北京:北京大学出版社,2017 年 11 月第 1 版。定价:58.00 元。

一开篇,该书就介绍了目前具有一定自主性的机器系统所存在的道德盲区,引发读者重视制造道德机器的迫切性。有道德的机器并不会自动出现,而是需要人为的设计和干预。作者关注目前的机器人系统,发现它们要么是自主性和道德敏感性都很低的系统,尚处于完全受设计者和使用者控制的“操作性道德”区域内;要么是自主性高而伦理敏感性低、或自主性低而伦理敏感性高的系统,只具备有限的“功能性道德”。该书分别描述了它们的局限性,并指出如何让具有高度自主性的机器人成为道德主体,即人工道德智能体(AMAs),亦即书名所谓的“道德机器”,这是摆在伦理学家、哲学家、认知科学家和机器人学家面前的难题。

“道德机器”并非一个自然而然的说法。如果机器人可以实现道德决策的功能,人类真的会想要机器人这样去做吗?人类将机器人当作“士兵、伴侣和奴隶”,在每种关系中将机器人升级为 AMAs,都会带来一定的社会风险。作者承认风险的可能,但更注重在技术发展中评估和控制风险。相比旧式技术哲学家可能采取的对技术进步观的批判,或者工程师可能具有的盲目的技术乐观主义,该书作者显然更赞同哲学家海伦·尼森鲍姆所提出的意在向技术注入增进人类福利的价值观的“工程能动主义”,致力于引导工程师不仅仅关注操作性道德的价值方面,更要为机器人系统自身提供清晰的道德推理和决策能力。作者还进一步追问:既然机器人这个基于逻辑平台建立的必然性系统没有如人类一样的自由意志、理解和意识等,那么,它真的可以有道德吗?虽然关于人类道德的本质极富争议,与其对“工程能动主义”的认同相一致,作者相信,即使关于伦理本质的本体论问题、认识论问题和机器

收稿日期:2017-12-30;修回日期:2018-01-31

作者简介:李颖娜,研究方向为计算哲学,电子信箱:1114620337@qq.com

引用格式:李颖娜. 制造“道德机器人”的远虑与近忧——评《道德机器:如何让机器人明辨是非》[J]. 科技导报, 2018, 36(4): 101-102; doi: 10.3981/j.issn.1000-7857.2018.04.015

道德的实践问题都悬而未决,却并不影响通过推进构建AMAs的实践去促进对于伦理本质的本体论和认识论问题的理解。相比难以定论的哲学争论,作者更关心AMAs的决策和人类的行为在道德效果上的等同程度,提出了评价AMAs设计成功程度的标准——类似经典图灵测试的对比道德图灵测试,即成功的AMAs应在人与AMAs的混杂测试中始终是更为道德的那一个。

该书的主旨意在探讨如何建构成功的AMAs。当进入实践层面,一些复杂的问题仍会浮现出来。AMAs将贯彻谁的道德标准、怎样的道德程序?作者承认构建AMAs有不同路径,聚焦于构造满足外部评价标准的道德机器的实践目的,探讨伦理学家应如何和工程师通力合作,使得道德机器实现道德决策能力。总括地说,有自上而下进路、自下而上进路2种设计思路。前者指的是设定一套可以转化为算法的伦理规则,用它来指导设计执行规则的子系统,这样AMAs有道德的行动就转变为遵守规则的问题;后者指的是不预先给定伦理规则,而是创造环境让AMAs自主地探索学习,当其做出道德上良好的行为,给予积极反馈使之得到强化。

当然,2种进路都存在一些问题。对于前者而言,无论采取功利主义还是道义论所支持的普遍原则,都会遇到具体规则如何与首要原则自洽、难以落实的计算复杂性等问题。采用后者则会面临当环境变化时如何适应、难以确定AMAs是否会产生出复杂的道德判断能力等问题。基于以上分析,作者瞩目于未来应用联结主义网络来开发具有美德的系统,这采用的是两者混合式的进路,既强调一定的伦理原则的重要性,又具有一定的灵活度拓展行动的范围。

除理论探讨外,该书后几章中介绍了设计AMAs的一些近期的具体实践,如关注道德决策的MedEthEx等软件系统,具有一定具身认知、社交能力等超理性能力的机器人Leonardo等,基于全局工作空间理论整合理性推理与情感的LIDA模型,这些AMAs的前沿实践各有一些生动有趣的进展,它们区别于科幻片的未来狂想,展示了道德机器的真实发展水平和趋势。

作者采用的是更靠近工程实践的"施事者"视角,以便于更好地参与和工程师共建AMAs的合作,并寄望于通过持续关注道德机器的工程实践中的具体问题,为人类目前的伦理理解提供反馈。总的来说,该书话题围绕如何成功地"教机器人明辨是非"展开,涉及伦理学、认知科学、心理学、机器人学等多个学科,有着丰富的信息含量。它成功地引起人们对于建构AMAs迫切性的关注,详尽地介绍了关于AMAs的2种设计进路的优劣势,并且可信地介绍了当下体现着混杂进路的多种AMAs初步实践。唯一遗憾的是,作者有意无意地简化了非工程问题的复杂性。

在AMAs完全实现的时代,人与机器将如何进行合乎道德地行为交互?作者视其为未来主义的问题而予以搁置,但这还是探讨"道德机器"的题中应有之义。对于这个世界上可能发生的无穷个事件和可能存在的无穷种角色来说,"道德的"不是一个非真即假的判断。仅就作者所初步揭示的,机器道德问题已经有着足够的复杂性,而且也是对人类道德的巨大拷问。在机器人那端,该书未充分展开的AMAs实现"谁的道德和什么道德"就隐含了人类不同意识形态的冲突,机器人会是人类群体冲突的新战场吗?而在人类这端,无论将AMAs作为"士兵"、"伴侣"还是"奴隶",试想如果机器人展现出足以通过对比道德图灵测试的道德能力,那么如何对待它们才是合乎人类道德的呢?将其作为实验品进行实验,根据最新的需求对其进行升级,这在性质上和第二次世界大战期间纳粹分子囚禁犹太人做实验并且大批量地处死他们有什么本质区别吗?和基础医学研究领域用小白鼠做生理和药理实验,最终将其作为医疗垃圾处理有没有本质区别呢?前者现在已经为人唾弃,因为纳粹分子将犹太人视作劣等民族完全是偏见和鬼话。后者现在仍在时时上演,在人类利益面前,如何处理动物伦理问题还远未达成共识。如果把这种道德优越的机器人作为人一样来尊重,给予其足够的自由发展空间,机器人会不会对人类自身的生存和发展形成反噬?尽管这些还只是一种未来主义式的推演,但已让人很难在负罪感、无辜感和安全感面前保持心理平衡。关注如何教机器人辨别是非的工程实践固然踏实,但该书更重要的价值可能在于沿着作者的论证分析,读者不由得会进行更深入的思考那些终究无法在工程实践的内部消解的"谁的道德"和"什么样的道德"等问题。

(责任编辑 陈广仁)