人们对人工智能做道德决策的厌恶感之源及解决之道

The Origin of and Solution to People's Aversion to AI Making Moral Decisions

丁晓军 /DING Xiaojun[1]　喻丰 /YU Feng[2]　许丽颖 /XU Liying[3]

（1. 西安交通大学人文社会科学学院，陕西西安，710049；2. 武汉大学哲学学院，湖北武汉，430072；
3. 清华大学社会科学学院，北京，100875）
(1. School of Humanities and Social Sciences, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049; 2. School of Philosophy, Wuhan University, Wuhan, Hubei, 430072; 3. School of Social Sciences, Tsinghua University, Beijing, 100875)

摘　要：为了借助人工智能来应对当代社会的道德滞后问题，跨学科、跨文化的大量研究都在关注对人工道德智能体的设计、实现和发展。然而，一个更为重要的元问题应该是：人们究竟是否希望由人工智能做道德决策？来自道德心理学家和神经科学家的一些实证研究表明，答案是否定的，并且该厌恶感的来源可能在于人们认为人工智能缺乏完全的、完整的心智。而要想有效提高人们对人工智能做道德决策的接受度，一方面应将道德决策主体的合法范围延伸至人工智能，另一方面则应进一步对人工智能进行拟人化，提高人工智能被感知到的体验性（热情）和能动性（能力、专长性），以增强人类与人工智能之间的共情和信任，同时还应特别注意利用伊丽莎效应并避免恐怖谷效应。

关键词：人工道德智能体　道德决策　道德滞后问题　心智感知　拟人化

Abstract: In order to deal with the moral lag problem in the contemporary society with the help of AI, a large number of interdisciplinary and intercultural researchers are paying close attention to the design, implementation and development of artificial moral agents. However, a more important meta-question should be: Do people really want machine/robot/AI/computer to make moral decisions? Some empirical studies from moral psychologists and neuroscientists suggest that the answer may be "No", and this aversion is mediated by the perception that artificial intelligence lacks a complete mind. This paper argues that in order to effectively improve people's acceptance of AI making moral decisions, on the one hand, the legitimate circle of moral decision makers should be extended to artificial intelligence; on the other hand, we should further anthropomorphize AI and increase its perceived experience (warmth) and agency (competence, expertise), in order to enhance the mutual empathy and trust between humans and AI, while at the same time we should pay special attention to the utilization of the "Eliza Effect" and to the avoidance the "Uncanny Valley Effect".

Key Words: Artificial moral agent; Moral decision; Moral lag problem; Mind perception; Anthropomorphism

中图分类号：N031　　**文献标识码：**A　　DOI: 10.15994/j.1000−0763.2020.12.011

随着人工智能越来越多地从科幻走向现实、从高科技概念广告走进普通人的日常工作和生活，人们对于人工智能的依赖程度也在不断提高。虽然关于人工智能的许多哲学问题还在持续获得热

基金项目：教育部人文社会科学研究青年基金项目"理性行动的认知规范研究"（项目编号：19YJC720006）；国家社科基金青年项目"拟人化人工智能的道德责任归因研究"（项目编号：20CZX059）。
收稿日期：2019年7月19日
作者简介：丁晓军（1987−）女，江苏盐城人，西安交通大学人文社会科学学院副教授，研究方向为分析哲学、科技哲学、哲学咨询。Email: xiaojunding@xjtu.edu.cn
喻　丰（1985−）男，湖北武汉人，武汉大学哲学学院教授，研究方向为社会与文化心理学、道德心理学、积极心理学。Email: psychpedia@whu.edu.cn（通讯作者）
许丽颖（1993−）女，湖北襄阳人，清华大学社会科学学院博士后，研究方向为人工智能心理学。Email: liyingxu830@gmail.com

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

议、远未达成共识，但人工智能已然以大大超出我们预期的速度马不停蹄地融入人类世界，提高了人们的工作效率，提升了人们的生活品质，极大满足了人们对美好生活的期待与幻想。与此同时我们也看到，人工智能的发展也时不时地超出人类的控制和预期，给人们造成了不少麻烦，也由此滋生了许多道德困境与难题。因此，许多哲学家、心理学家、计算机科学家、认知科学家都在致力于设计、制造出能够像人类一样明辨是非、进行道德判断的人工道德智能体。温德尔·瓦拉赫（Wendell Wallach）和科林·艾伦（Colin Allen）就此提出三个值得探究的问题：人们需要人工道德智能体吗？人们是否希望由人工智能做道德决策？如果人们认为让人工智能做道德决策是必要的或不可避免的，那么工程师和哲学家应该如何设计这样的人工道德智能体？（[1]，p.9）这三个问题是紧密相关的，而本文将重点考察第二个问题，即：人们对人工智能做道德决策的喜恶情况如何？为何如此？如何进行应对？

一、人工智能道德决策的普遍性及必要性

时至今日，人工智能几乎已经全方面地渗透进人们日常生活的方方面面，成为在当代社会中人们所无法脱离、不容忽视的重要元素。得益于其卓越的计算能力、远超人类的理性、智能与战略部署能力，人工智能通常能够做出更加优化的决策，因此人工智能已经不同程度地被应用在危机管理、供应链配送、医学诊断、飞机航程线路计算、复杂库存管理、人类行为预测、象棋、围棋、智力问答节目等领域。[2]可以看出，人工智能不仅作为一种工具被用于替代人类完成一些高风险或高重复性的工作，也会被分派执行一些常人所无法胜任、需要较高心智能力和感官功能的精密作业。与此同时，随着5G时代的突然降临，对人工智能进行全面、深入地发展与应用的趋势又将会得到极大地加速。

有研究者甚至预计，当人工智能发展到一定阶段时，具有与人类相似的物理特征与行为特性的智能机器可能会在很多角色上对人类进行补充和替换，例如商店的维护和销售人员、酒店的接待员和内务人员、医院的护士和陪护，以及家中老人和孩子的同伴等。[3]目前，世界各主要大国都在人工智能、机器人等领域投入了尽可能多的关注和资源，对相关技术发明专利的抢夺与竞争也日趋白热化。在一些领域，人工智能系统已经或即将成熟到足以单独做出相应决策，有时候甚至是关涉人类荣辱与生死的道德决策。而随着相关科学技术的持续蓬勃发展，人工智能在人类生活中所扮演的角色将会越来越重要，人工智能做道德决策的势头也会越来越猛烈，随之而来的一个重要问题是：我们有必要让人工智能参与做出、甚至独立做出道德决策吗，尤其是当这些决策与人类的生死荣辱相关的时候？

人类世界所存在的道德滞后问题是人工智能道德决策之必要性的一个有力佐证。道德滞后问题的提出者米哈伊尔·克林切维奇（Michał Klincewicz）指出，一方面，正如亚里士多德在《尼各马可伦理学》中所言，如果儿童在其成长过程中没有受过很好的道德教育，那么他们在成年后几乎是不可能真正过上一种良善的道德生活的，而另一方面，即便人们在孩童时期获得了很好的道德教育，但是在日常生活中普遍存在于道德判断和广义理性之间的冲突也使得“人们很难做一个道德之人”。（[4]，pp.171–172）克林切维奇认为，时至今日，当代人想要过上一种道德生活的难度已经远高于亚里士多德那个年代，因为当代社会在科学技术上获得迅猛发展的同时，相关道德理论与实践却没有能够及时跟上，而人类从先祖那里继承而来的道德心理学并不能使我们做好足够准备以应对由信息技术、生物技术等的出现与发展而带来的道德困境。（[4]，p.172）如此种种问题导致了人类的道德水平一直都没有达到它们所能够达到或者所应该达到的程度，克林切维奇因而将这些问题概括为“道德滞后问题”。（[4]，p.172）如何来应对、解决道德滞后问题是大量道德哲学家目前所关注的焦点，其中一个重要思路就是利用强大的人工智能来补足人类智能在道德领域的先天及后天缺陷，增强人类的相关道德能力，帮助他们成为更道德的人。

杰森·博伦斯坦（Jason Borenstein）和罗恩·阿尔金（Ron Arkin）通过许多具体案例展示了人工智能的“道德助推者”（Moral Nudger）功能，即人工智能通过口头的或者非口头的提示来影响人类行为，以使得人类更有可能做出道德行为。（[5]，p.37）朱利安·萨乌列斯库（Julian Savulescu）和汉娜·马斯伦（Hannah Maslen）则探讨了利用道德人工智能（Moral AI）来实现对人类的道德增强（Moral Enhancement）的可能性。他们认为，面对世界全球化所带来的紧迫挑战，随着普适计算、环

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

境智能的发展，人工智能能够对人类的道德行为进行监测、提示和建议，从而帮助人类克服其道德心理学的一些先天固有局限；这种人工智能可以监测影响道德决策的物理因素和环境因素，识别并使人类意识到他们的偏见，进而根据人类的道德价值，就正确的行动过程向人类提供建议。（[6]，p.80）

对道德增强这个构想的一个共同反对意见是，由于尚不存在（并且在很长的时间内似乎也很难形成）一个关于正确行为的单一理论解释，人们在是非对错的道德判断问题上一直不能达成一致，因此所谓利用道德人工智能来增强人类道德的这种努力是注定要失败的。但是萨乌列斯库和马斯伦辩称，道德人工智能为人类量身定制，不仅能保持道德价值的多元性，还能通过促使人类的反思以及帮助人类克服其自然的心理局限性，从而增强人类的自主性；他们认为这也是在道德增强方面，道德人工智能超出其他形式的生物医学技术的一个重要优势。（[6]，p.79）

克林切维奇则指出道德人工智能的一个主要问题是其鲁棒性还不够强大，并不能切实在人类行为中起到规范作用，于是他提出了一个在他看来更有希望成功的方法，该方法依赖于一个人工道德推理引擎（artificial moral reasoning engine），此引擎向人类用户提供以一阶规范理论（如义务论或功利主义）为基础的道德论证，对推理敏感（reason–responsive）的人是能够被相应道德论证所说服的。克林切维奇认为人工道德推理引擎可以起到规范作用，也是一种更有前景、更有希望成功的道德增强途径，因为这样一个系统可以利用人们有时候对自动化技术的过度信任：当来自其他人类的类似论证可能无法说服某个人的时候，一个设计良好的道德推理系统却可能成功完成对此人的说服。[4]

二、人工智能道德决策的基本路径和规范

伦理学中已经有大量对人类做道德决策的研究，相关研究通常关注于人类在面临道德情境甚至困境的时候如何进行选择，而这些道德困境通常涉及人们自身与他人和群体在利益、价值方面的冲突，也会涉及不同道德规范或者理论立场之间的冲突，例如功利主义、义务论、美德论之间的冲突。[7] 当人工智能在特定道德情境甚至困境中成为道德决策者时，它们应该如何处理、解决上述多方面、多层次的冲突，这是道德机器研究者在进行相关软硬件设计制造时所需要着力克服的一个重要问题。对道德机器（人工道德智能体）的设计其核心正在于制定关于人工智能应如何做道德决策的规范，相关研究已有很多。瓦拉赫和艾伦将一些主流研究设计思路概括为自上而下路径、自下而上路径以及混合路径，接下来我们将概述他们对这三种路径的思考。

在最一般的意义上，人工道德智能体设计的自上而下路径需要有一套规范，这些规范又必须可以转化为一种算法；自上而下的伦理体系可能有多种来源，包括宗教来源、哲学来源和文学来源等等，例如道德金律（Golden Rule，“己所不欲，勿施于人”）、摩西十诫、后果论或功利主义伦理学、康德的道德律令、法律和职业守则以及阿西莫夫的机器人三定律等等。（[1]，p.84）也就是说，在自上而下路径下，人工道德智能体设计者必须选取一个具体的伦理理论，分析它的计算需求以指导设计能够执行该理论的算法和子系统；如果相关道德规范可以得到清晰地陈述，那么道德的行为就是遵守相关道德规范的行为，而没有遵守相关道德规范的行为就是不道德的行为。在这整个过程中，人工道德智能体所需要做的就只是去计算其行为是否为相关上层规范所允许。但是，正如我们在人类世界中所看到的，上述任何一条原则单独地来看都有可能会导致一个道德困境的出现，并且不同道德规范之间更是时常存在着冲突。目前看来，人类本身尚不能彻底有效地解决这样的问题。在这样的情况下，由人类所设计出来的人工道德智能体又能否回避或者解决该问题呢？此外，自上而下的伦理体系要求设计者为人工道德智能体所选取的伦理理论必须是表述清晰的、可计算的，那么人类世界中那些表达模糊的、意会默许性的道德规范又该如何有效地嵌入到计算机系统之中呢？（[1]，pp.83–98）

自下而上路径认为道德行为来自于学习和进化，而道德机器设计的重点便在于使人工道德智能体具备一些能力来动态地整合来自不同社会机制的输入，习得它们所处环境的道德规范。也就是说，正如人类在孩童时代所接受的道德教育一样，人工道德智能体在其发展阶段也需要通过经验、试错来获取关于其行为的道德可接受性或不可接受性的反馈。对于人类来说，这种反馈通常表现为对相关行为主体的奖励与惩罚、认可与不认可；但是人工道德智能体似乎并不能有意识地感受到快乐或痛苦，

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

应该如何给予它们以这样的反馈呢？另一方面，人工道德智能体作为学习机器，它的灵活性和适应性必须是十分突出的。然而，任何具有学习能力的系统都有可能学习错误的东西，它们甚至可能会撤销或覆盖设计者内置于其中的约束限制，进而给人类带来危害。此外，当语境、环境发生变化，当目标是多个，或者当可用的信息是混乱的或者不完整时，对于自下而上路径来说，提供一个清晰的行动方案就是一个十分困难的任务。最为重要的是，即使是在计算机系统的加速环境中，许多代人工智能体可以在几秒钟内进行变异和复制，但进化和学习可能仍然是非常缓慢的过程。并且，对于一个不断进化的人工道德智能体来说，决定其进行变异和复制的适应度标准是什么？什么是合适的目标？如何有效地为自组织系统定义这个目标？人们还不是十分清楚应该如何解决这些问题。（[1]，pp.106−118）

如前所述，自上而下路径和自下而上路径都面临着大量尚待解决的问题，二者中无论哪一个都不足以设计制造出令人满意的人工道德智能体，这提示人们可能需要考虑对两者进行某种混合。事实上，对自上而下、自下而上所进行的二分原本就有些简单化。工程师通常从对复杂任务进行自上而下的分析开始，以指导对组件进行自下而上的组装。自上而下路径似乎更安全、更有保障，但与此同时，该路径通常意味着很难满足的理想主义标准以及相关计算的复杂性。而自下而上路径的一个优势在于通过对组件的组装以实现一个目标。如果一个系统的组件被设计得很好，并且能够被正确地予以组装，那么人工道德智能体在应对由环境和社会语境所带来的挑战时其所面临的选择范围就将会扩大，而具有对该选择范围里的选项进行自上而下评估的能力的人工道德智能体将能够选择既符合其目标又符合可接受之社会规范的行动。（[1]，pp.114−115）为了说明自上而下路径和自下而上路径的混合交互方式，瓦拉赫和艾伦考虑尝试利用联结主义网络来开发一个具有良好品格或美德的计算机系统。（[1]，pp.171−187）

三、人们何以厌恶人工智能做道德决策

尽管揭示人工智能应如何做道德决策是十分重要的，但是调查一个更为基本的元问题似乎更加重要：人们究竟是否希望由人工智能做出道德决策？人工智能的发展将会对我们每一个个体以至人类社会全体都造成根本性影响与转变。而随着信息技术的日益发展，人们似乎已经不得不真正开始设计能够进行道德决策的计算机、机器人，例如，它们需要判断什么时候效率应该凌驾于隐私权之上，而什么时候隐私权又应该凌驾于效率之上。然而，人类究竟是否想要这样的技术？人们究竟是否愿意、是否喜欢由人工智能来做道德决策？（[8]，p.464）

答案是否定的。约哈南·比格曼（Yochanan Bigman）和库尔特·格雷（Kurt Gray）的一系列实验研究表明，人们不喜欢人工智能在自动驾驶、法律、医疗和军事方面做出与生死相关的道德决策。其研究1–4都是被试间设计。在研究1中，被试们更允许由人类驾驶员而不是由自主计算机程序来做出涉及生死的驾驶决策；在研究2中，被试们更允许由人类委员会而不是由超级计算机CompNet来做出假释决策；在研究3中，被试们更允许由琼斯医生而不是由基于统计的自主计算机系统HealthComp来做出最终会导致病人死亡的医疗决策；在研究4中，被试们更允许由琼斯上校而不是由基于统计的自主计算机系统CompNet来做出最终会导致无辜儿童死亡的军事决策。[2] 如何来解释这样的实验结果呢？

瓦拉赫和艾伦曾就“人们是否希望计算机做道德决策？”这个问题进行了初步探索。他们在技术哲学的语境下讨论了人们对人工智能做道德决策的厌恶感，认为这种厌恶可以在更一般的意义上被视为人们担忧、惧怕新技术对人类社会所可能造成的影响。例如，其中一个似乎尤为紧迫的关切就是人工道德智能体是否会导致人类取消对机器的责任。至于人类是否在将来会被机器奴役，这种担忧在他们看来在很大程度上还只是推测性的。因此，人工智能作为一种新兴技术其所可能带来的风险是人们厌恶由人工智能做道德决策的原因之一；而对技术风险进行评估所尚未解决的一个问题是：如何认真权衡一个新技术所提供的明显优势以及它所可能造成的灾难。（[1]，p.9）

除了对新技术的恐惧、担忧，比格曼和格雷的更多实验研究则进一步表明，这种厌恶感可能部分源于人们认为人工智能缺乏一个完全的、完整的心智。[2] 长久以来，人们对心智有着不同的感知维度，而希瑟·格雷（Heather M. Gray）等人则通过对2399份调查研究结果的统计分析得到了心智感

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

知的两个重要维度：能动性和体验性。[9]其中，能动性维度所包括的心智能力有7种：自我控制、道德、记忆、情绪识别、计划、交流和思考；体验性维度所包括的心智能力有11种：饥饿、恐惧、痛苦、愉悦、愤怒、欲望、个性、意识、骄傲、尴尬、欢乐。并且，这两个维度与亚里士多德关于道德能动者（moral agents，其行为在道德上可以是对的或是错的）和道德受动者（moral patients，别人可能会对他们做出一些在道德上或对或错的事情）的经典区分有关。心智感知的能动性维度与道德能动者挂钩，因此便也与责任挂钩，而心智感知的体验性维度与道德受动者挂钩，因此也就与权利和特权挂钩；这也表明了心智感知的能动性和体验性这两个维度实际上刻画了道德的不同方面。[9]

在比格曼和格雷的一系列实验研究中，被试要对人工智能决策者或人类决策者的12种不同心智能力进行打分，这些心智能力中的6个与能动性相关（“与他人交流”、“能够思考”、“计划自身行动”、“是智能的”、“有先见之明”和“能够对事情进行思考”），另外6个与体验性相关（“对痛苦敏感”、“体验到快乐”、“体验到恐惧”、“体验到同情”、“体验到共情”和“体验到内疚”）。分析显示，在整体心智方面，人类决策者被认为要高于人工智能决策者，心智感知（同时包括能动性和体验性二者）中介了人们对由人工智能做道德决策的厌恶感，并且这两种（能动性、体验性）间接效应都是显著的。[2]

进一步的研究还表明，人们对由人工智能做道德决策的厌恶感（以及通过心智感知所进行的中介）并不是因为人们认为人工智能会做出更糟糕的决定。即便决策结果是好的、积极的（研究5：“导弹袭击成功，杀死了恐怖分子，并且只对站在附近的一些平民造成轻微伤害”；研究6：“手术很成功，杰森活了下来并重新能够控制自己的身体”），人们也更倾向于由人类而不是由人工智能来做出相关道德决策。[2]

四、如何降低人们对人工智能做道德决策的厌恶感

在大致理清了人们厌恶人工智能做道德决策的现象及其可能原因之后，我们接下来自然而然要面临和解决的另一个问题就是，如果我们想要进一步有效推动人工智能在人类社会中的全面发展和深入应用，那么人工智能设计者、推广者、营销者和管理者等应该如何有效地降低人们对人工智能做道德决策的这种厌恶感呢？基于比格曼和格雷的后续实验研究，至少有以下三种可能的途径。

1. 将人工智能限制为咨询辅助类角色

研究7沿用了研究3和6中的医疗决策场景。在关于应该由谁来决定给不给杰森做手术这个问题上，被试共有三个选项：（1）HealthComp；（2）琼斯医生；（3）HealthComp辅助琼斯医生。最后，在100位被试中，4位选择了HealthComp，32位选择了琼斯医生，64位选择了HealthComp辅助琼斯医生。

这样的研究结果表明，通过将人工智能限制为咨询辅助类角色，也就是说，给予人类以最终决定权，可以降低人们对人工智能做道德决策的厌恶感。然而，即便如此，还是有相当一部分人（32%）选择了人类决策者本人（琼斯医生）。因此，这一途径看上去似乎并不能在很大程度上消除人们对人工智能做道德决策的厌恶感，我们还需要进一步考虑更多其他可能途径。

2. 提高人工智能被感知到的专长性

研究9包括被试内和被试间两种设计，仍然沿用了研究3、6中的医疗决策场景，被试必须选择应该由琼斯医生还是由HealthComp来决定给不给杰森做手术。被试内设计分为专长性相同组（琼斯医生和HealthComp都有75%的成功率）和人工智能专长性占优组（琼斯医生的成功率为75%，而HealthComp的成功率为95%）。卡方检验显示，在专长性相同组，只有7%的被试选择由HealthComp做决策，但是在人工智能专长性占优组，72.28%的被试会选择由HealthComp做决策。而在被试间设计下，方差分析显示，无论专长性水平如何，被试们都更允许由琼斯医生而不是由HealthComp来做决策。尤其值得注意的是，即便当HealthComp具有高专长性（成功率为95%）而琼斯医生只具有平均专长性（成功率为75%）时，被试还是更允许由琼斯医生而不是由HealthComp来做决策。也就是说，人们宁愿让胜率平平的人类医生而不是由胜率更高的专业人工智能来做相关决策，除非是像在被试内设计里那样，将HealthComp和琼斯医生在专长性上的差异通过成对比较变得特别显著之时，人们才可能会更多地倾向于由HealthComp来做相关决策。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

总之，一方面，提高人工智能被感知到的专长性可以降低人们对人工智能做道德决策的厌恶感，而另一方面，只有当人工智能在专长性上的优势特别突出时，此种效应才存在。因此，当人工智能在专长性上的优势还没有特别突出时，提高人工智能被感知到的专长性这一途径看上去似乎也无法在很大程度上消除人们对人工智能做道德决策的厌恶感，我们依然需要进一步考虑更多其他可能途径。

3. 提高人工智能被感知到的体验性

比格曼和格雷在探索人们厌恶人工智能做道德决策这个现象背后的可能原因时着重验证了心智感知的中介作用，并且进一步地，相关研究结果表明，人工智能被感知到的能动性要低于人类，而人工智能被感知到的体验性更是要显著低于人类。因此，一个自然而然的猜想便是，如果我们提高了人工智能被感知到的心智，那么人们对人工智能做道德决策的厌恶感就有可能就会降低。在上一小节的讨论其实就是在试图提高人工智能被感知到的能动性；当然，我们还可以进一步挖掘更多其他途径来提高人工智能被感知到的能动性。而研究8则是聚焦于试图提高人工智能被感知到的体验性。

研究8依然沿用了研究3、6中的医疗决策场景；但是与之前研究所不同的是，被试直接被告知将由HealthComp来决定给不给杰森做手术，并且随后会听取一段HealthComp的讲话录音。在低体验性组，HealthComp采用无情感的计算机语音，并将自己描述为没有情感；在高体验性组，HealthComp使用有情感、有表达力的声音，并将自己描述为具有体验情感的能力。方差分析显示，在高体验性组对由HealthComp做决策的可允许性和低体验性组对由HealthComp做决策的可允许性之间没有显著差异。而值得注意的是，当心智感知作为中介被纳入回归时，“被感知到的体验性”对“可允许性”的效应就变得显著，但是这种效应是负向的。

比格曼和格雷随后用“恐怖谷效应”（Uncanny Valley Effect）来解释为什么提高体验性反而可能会降低可允许性。恐怖谷效应由日本机器人专家森政弘（Masahiro Mori）于1970年提出：当机器人从样貌到行为都越来越像人的时候，人们对它们的喜爱程度也会不断提升，直到抵达一个山谷——“恐怖谷”；也就是说，森政弘假设，当一个类人机器人的样貌跟人类非常接近但是又没有达到极其逼真的程度的时候，人们对这个机器人的反应会在某个时刻突然由共情转变为厌恶、惊恐。（[10]，p.98）恐怖谷效应在机器人设计师中引起了不同的反应。人形机器人专家石黑浩（Hiroshi Ishiguru）认为，对于那些看起来更像人类的机器人，人们会更容易接受它们、与它们相处，而恐怖谷效应则是机器人设计师们需要克服的众多挑战之一。另一些机器人学家则认为，由恐怖谷效应所带来的启示是，最有效的机器人应该在具备一些类人特征的同时，又不会去假装自己是人类。瓦拉赫和艾伦便提出，机器人学家所采取的策略在很大程度上取决于他们的目标，而基于当前的技术，在外形和动作上与人类相似但有明显不同的机器人在促进人与机器人的互动方面要优于人形机器人。（[1]，p.44）

与此同时，当机器人被设计成外貌和行为都跟人类相像时，人们会倾向于将人类的特征归之于它们，该现象被称为“伊丽莎效应”（Eliza Effect）。该效应以模仿心理医生的计算机程序Eliza命名，指的是人工智能的拟人化趋势。[11] 伊丽莎程序由MIT计算机科学家约瑟夫·魏泽鲍姆（Joseph Weizenbaum）于1964–1966年间开发出来，其智能之处在于它能通过脚本理解简单的自然语言，并能产生类似人类的互动，使人类向其敞开心扉；人们甚至一度认为经由这种思路可以做出能够通过图灵测试的程序。魏泽鲍姆以萧伯纳的戏剧作品《卖花女》（*Pygmalion*）中主人公的名字来命名该程序，该剧作讲述的是一口伦敦英语腔的下层阶级卖花女伊丽莎·杜利特尔（Eliza Doolittle）被中产阶层语言学家亨利·希金斯（Henry Higgins）教授改造成优雅贵妇的故事 。与语言学家教伊丽莎如何打扮、说话以及举止得像个淑女一样，人工智能的开发人员使用设计特性使机器更像人，以引起积极的印象和反应。（[12]，pp.1–2）

总的来说，在伊丽莎效应下，提高人工智能被感知到的体验性可能会降低人们对人工智能做道德决策的厌恶感，而在恐怖谷效应下，提高人工智能被感知到的体验性反而可能会提高人们对人工智能做道德决策的厌恶感，因此人工智能的开发设计者就必须尽量发挥利用伊丽莎效应而避免发生恐怖谷效应。也就是说，在研究者对人工智能所进行的拟人化程度能够真正达到100%之前，较好的策略是在对人工智能的某些方面进行高度拟人化的同时仍然不忘为其保留一些异于人类的特征。这当中的张力、分寸和尺度是不好把握的，需要通过大量实验

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

研究来进行验证、调整和修正。于是，现在的问题又演变成：应如何对人工智能进行恰到好处的、不多不少的拟人化？

正如一个人的样貌、行为特征、被感知到的类人格特质等会影响他人对其接受度，被拟人化的人工智能亦是如此。在人机交互（Human -Machine/ Robot/ AI/ Computer -Interaction）的过程中，人类也会在某种程度上将人工智能（机器人）视为他人，并对其进行感知、评价。在社会心理学中，社会认知、社会判断的内容常常被概括为热情与能力两个基本维度，前者是对他人意图的反映，与关心、友善和社交等特质相关，即与一个人的社交能力、友好性和可信赖性相关，[13]后者是对他人能力的反映，与一个人的能力、智力和技能等特征有关。[14]类似地，人们对人工智能的社会认知和判断也可以从“热情”与“能力”这两个维度进行衡量，并且也可以通过提高人们对人工智能之“热情”的感知从而提高人们对其“体验性”的感知。相关研究是值得进一步深入的。

结　语

随着人工智能进一步融入我们的私人和社会生活之中，并在各个层面不同程度地担当起决策角色，我们越来越迫切地需要了解人们（作为个体、公众或消费者等）到底“是否”（以及“为什么”）喜欢由人工智能来做相关（道德）决策，以及“如何”能够通过一定的途径（智能设计、公共宣传等）来改变人们的态度。

在未来，相关研究还可以扩展到其他社会领域。或许通过在不同社会领域之间的比较，我们不仅可以进一步了解公众在这一重要问题上的偏好，以及如何对公众的这些偏好进行更新以应对快速的社会变革与技术发展。更为重要的是，相关厌恶感的存在并不意味着研究者应该停止探索、揭示如何设计道德机器，而是提醒我们深思人类到底想要“何种人工智能”来做“何种决策”。[15]

［参 考 文 献］

[1] Wallach, W., Allen, C. *Moral Machines: Teaching Robots Right from Wrong*[M]. New York: Oxford University Press, 2009.

[2] Bigman, Y. E., Gray, K. 'People Are Averse to Machines Making Moral Decisions'[J]. *Cognition*, 2018, 181: 21–34.

[3] Van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., Petersen, J. A. 'Domo Arigato Mr. Roboto: Emergence of Automated Social Presence in Organizational Frontlines and Customers' Service Experiences'[J]. *Journal of Service Research*, 2017, 20(1): 43–58.

[4] Klincewicz, M. 'Artificial Intelligence as a Means to Moral Enhancement'[J]. *Studies in Logic, Grammar and Rhetoric*, 2016, 48(1): 171–187.

[5] Borenstein, J., Arkin, R. 'Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being'[J]. *Science and Engineering Ethics*, 2016, 22(1): 31–46.

[6] Savulescu, J., Maslen, H. 'Moral Enhancement and Artificial Intelligence: Moral AI?'[A], Romport, J., Zackova, E., Kelemen, J. (Eds.) *Beyond Artificial Intelligence, Topics in Intelligent Engineering and Informatics 9*[C], Switzerland: Springer International Publishing, 2015, 77–95.

[7] Paxton, J. M., Ungar, L., Greene, J. D. 'Reflection and Reasoning in Moral Judgment'[J]. *Cognitive Science*, 2012, 36(1): 163–177.

[8] Wallach, W. 'Implementing Moral Decision Making Faculties in Computers and Robots'[J]. *AI & Society*, 2008, 22(4): 463–475.

[9] Gray, H. M., Gray, K., Wegner, D. M. 'Dimensions of Mind Perception'[J]. *Science*, 2007, 315(5812): 619.

[10] Mori, M. 'The Uncanny Valley [from the field]'[J]. *IEEE Robotics & Automation Magazine*, 2012, 19(2): 98–100.

[11] Ekbia, H. R. *Artificial Dreams: The Quest for Non-Biological Intelligence*[M]. New York: Cambridge University Press, 2008.

[12] Kim, S. Y., Schmitt, B. H., Thalmann, N. M. 'Eliza in the Uncanny Valley: Anthropomorphizing Consumer Robots Increases Their Perceived Warmth but Decreases Liking'[J]. *Marketing Letters*, 2019, 30: 1–12.

[13] Fiske, S. T., Cuddy, A. J. C., Glick. P. 'Universal Dimensions of Social Cognition: Warmth and Competence'[J]. *Trends in Cognitive Sciences*, 2007, 11(2): 77–83.

[14] Fiske, S. T., Cuddy, A. J. C., Glick. P., Xu, J. 'A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition'[J]. *Journal of Personality and Social Psychology*, 2002, 82(6): 878–902.

[15] 许丽颖、喻丰. 机器人接受度的影响因素 [J]. 科学通报，2020, 65（6）: 496–510.

［责任编辑　李斌　赵超］

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net