中国心理学前沿
2022 年 10 月第 4 卷第 10 期

人们对智能机器的道德决策期望研究概述

吴明证[1] 严梦瑶[1] 林 铭[1] 刘钇瑶[1] 孙晓玲[2]

1．浙江大学心理与行为科学系，杭州；
2．杭州师范大学心理系，杭州

摘 要 | 智能机器指能够在各类环境中自主地或交互地执行任务的人工智能系统，近年来，智能机器的道德决策受到越来越多研究者的关注。本文从人们对智能机器的道德决策期望、影响因素和心理机制三方面出发，系统介绍了智能机器的研究现状，并探讨了已有研究的不足及未来研究方向。

关键词 | 智能机器；道德决策；心智感知

随着智能化时代的即将到来，一些智能机器（Intelligent Machine）或自主机器（Autonomous Machine）已经能够高度自主地在自动驾驶、军事领域、医疗服务等环境中执行任务（Logg et al.，2019）。我们希望智能机器具有更强的功能和自主性以应对复杂的社会环境，这使得智能机器在工作时不可避免地会遭遇各种道德情境，需要快速地做出正确的道德决策。出于慎重起见，人类必须保证智能机器以合乎道德的方式来行动。为此，政府管理部门、学术界一直致力于制定适合人工智能的伦理规则，给予智能机器合适的伦理规范限制。在这一背景下，研究者展开了对人们如何期望智能机器进行道德决策的探讨，以便为人工智能的道德立法提供学理依据。

目前，机器道德研究受到 AI 研究者的广泛关注。为此，本文拟对智能机器道德决策期望的研究进展进行概述性介绍，在此基础上分析已有研究的不足及未来发展方向，旨在为 AI 研究者提供参考和借鉴。

1 智能机器的道德决策期望研究的主题

1.1 智能机器是否具有道德主体地位

与一般的风险决策不同，道德决策往往涉及是非对错。人们都认可智能机器的高速运算能力，接受

基金项目：教育部人文社会科学规划基金项目（19YJA190007）；浙江省科技厅软科学研究计划项目（2020C35080）。
通讯作者：孙晓玲，博士，杭州师范大学心理系副教授，研究方向：道德心理学，E-mail：psysun@163.com。
文章引用：吴明证，严梦瑶，林铭，等．人们对智能机器的道德决策期望研究概述［J］．中国心理学前沿，2022，4（10）：1097-1105.
https://doi.org/10.35534/pc.0410130

智能机器应用于一般风险决策领域，但对于智能机器能否进行道德决策则莫衷一是。探讨人们对智能机器的道德决策期望，前提在于人们是否认可智能机器具有道德主体（Moral Agency）地位。一般认为，道德主体是指能够评估和调控自身行为，并且考虑到可能的潜在危害和可能忽视的职责的个体（Wallach et al.，2008）。一些研究者否认智能机器具有道德主体性，理由在于智能机器对于程序设计师的依赖性（Johnson and Miller，2008），仅具有工具性价值（Martin and Andreas，2011），存在认知、动机和情感的缺失（Brozek and Janik，2019），因此智能机器不能作为道德主体，也不能进行道德自主决策。

但也有许多研究者（如 Floridi and Sanders，2004；苏令银，2019）认为，随着自动化程度和智能算法的不断迭代，智能机器将具有独立的道德地位，在某些情况下可以被视为真正的道德主体。研究者将这些具有道德决策能力的智能机器称为人工道德智能体（artificial moral agents，AMAs），能够在没有人类监督的情况下自主进行道德推理与决策。一些研究者则提出，如果要赋予智能机器以道德主体地位，智能机器需要具备一定的能力。例如，格雷等人（Gray et al.，2007）指出，当人类知觉并意识到智能机器完成各项事务的能力远胜人类时，才可能愿意赋予人工智能道德主体地位，将决策权和控制权交给智能机器。

1.2 智能机器在道德情境中选择作为还是不作为?

研究发现，人们对智能机器道德决策的态度存有分歧。一方面，人们反感智能机器进行道德决策，或者期望智能机器在道德困境中选择不作为（Inaction）。迪特沃斯特等人（Dietvorst et al.，2015）揭示了算法厌恶（Algorithm Aversion）现象，发现即使由计算机驱动的算法所做出的评估决策比人类表现得更好，人类也往往排斥由计算机做出的判断或决策，而选择信任自己的判断，那些对算法有过一些经验并且发现它们并不完美的人尤其如此。研究还发现，即使智能机器在驾驶、医疗、司法和军事等困境中做出的道德决策能够带来积极的结果，被试仍倾向于让人类而非智能机器来做决定（Bigman and Gray，2018）。因此，人们更愿意将涉及社会交互的任务委派给人类，而非智能机器（Gogoll and Uhl，2018），对那些能够自主进行决策的智能机器感觉诡诞（Stein and Ohler，2017），不愿意让智能机器自主进行道德决策。此外，莫德尔等人（Meder et al.，2018）关于人们对自动驾驶汽车在不确定后果下的决策研究则发现，相比于人类司机，被试总是希望自动驾驶汽车在面对两难抉择时选择不作为，保持原定方向，即使这样的选择并不能使伤害最小化。另一方面，也有研究揭示了人们的算法偏好（Algorithm Appreciation）现象。例如，无论是在主观还是客观领域，相较于人类，人们都更信赖智能机器的决策（Logg et al.，2019）。

无论人们是否愿意让智能机器做出道德决策（“作为”），智能机器与人工智能技术的发展趋势都是不可阻挡的，智能机器已经面临着各种现实的道德决策问题，回避这一问题反而可能产生更加严重的影响（Damm，2011）。例如，如果要求自动驾驶汽车在遇到紧急情况时总是将控制权交还给人类司机，而面对突发事件人类司机可能根本来不及做出判断和反应，这将造成难以预料的后果。上述研究发现的矛盾之处表明，也许人们并不是期望智能机器不作为，而是在某些条件下作为。

1.3 人们期望智能机器如何进行道德决策?

大多数机器道德研究采用电车困境及其改进版本作为研究材料，在功利主义和道义论两种道德决策

框架下，探讨人们期望自动驾驶汽车在面对道德两难困境时该如何抉择，以及人们对智能机器道德决策的判断与接受程度（Malle et al.，2015；Bonnefon et al.，2016；Awad et al.，2018；Meder et al.，2018）。例如，博纳丰等人（Bonnefon et al.，2016）改编了电车困境，要求被试分别在牺牲 10 个行人和牺牲 1 个行人、牺牲 1 个行人和牺牲车上乘客，以及牺牲 10 个行人和牺牲车上乘客这三种驾驶决策的两难情境中做出符合自己期望的选择，结果发现，总体上被试都强烈认为，如果牺牲车上的乘客或者 1 个行人可以挽救更多人的生命，那么自动驾驶汽车所做的决策是更符合道德的。博纳丰等人（Bonnefon et al.，2016）还发现，尽管被试普遍认为功利主义决策是更加道德的，但对于具有这种决策模式的自动驾驶汽车，被试却更喜欢采用自我保护模式的自动驾驶汽车，即发生紧急状况时优先保证车上乘客的安全，这一结果暗示人们对于智能机器的道德决策期望可能存在“偏私”现象。

阿瓦德等人（Awad et al.，2018）为了探究人们对智能机器的道德决策期望，开发了名为 Moral Machine 的在线研究平台，以探究人们对自动驾驶车辆面临道德困境时的决策期望。通过对 200 多个国家和地区的人们对自动驾驶汽车道德决策偏好进行分析发现，在世界范围内人们对自动驾驶车辆存在着一些非常普遍的决策偏好，主要表现在三个方面：对保护人类生命的偏好、对保护年轻生命的偏好和对保护更多生命的偏好，即人们希望自动驾驶汽车能够在必要时选择牺牲动物来拯救人类生命，在紧急情况下选择牺牲老人而拯救儿童，在面临与电车困境类似的情境时选择牺牲少数人而挽救多数人。

2　影响人们对智能机器决策期望的因素

综观已有研究，影响人们对智能机器道德决策期望的因素大体分为四类。

2.1　道德决策情境特征

道德决策情境特征影响着人们对智能机器的道德决策期望。研究发现，在电车困境中乘客和行人的数量比、行人是否违法、行人的身份等都显著影响人们对自动驾驶汽车的决策期望（Bergmann et al.，2018；Awad et al.，2018，2020；Bonnefon et al.，2016）。个人道德情境和非个人道德情境也会影响人们对智能机器的道德决策期望。褚华东等人（2019）探讨了人们在这两种不同的道德情境下对智能机器的道德评价是否存在差异，结果发现，在个人道德情境（天桥困境）中，被试做出道义论取向行为的智能机器道德评价更好，而在非个人道德困境中被试对做出功利主义选择的智能机器道德评价更好。

事件结果的可能性也会影响人们对智能机器的道德决策期望。人们在做复杂的道德决策时会同时考虑结果的重要性及发生的可能性，这被称为“期望道德价值”，由于这种心理机制的存在，人们在不确定条件下做出的道德判断不同于在确定条件下（Shenhav and Greene，2010；Shou and Song，2017）。莫德尔等人（Meder et al.，2018）使用改编的电车困境作为研究材料，在该两难情境中自动驾驶汽车要么选择直行牺牲车道上的行人，要么选择转向牺牲路边的旁观者，汽车与车道上行人相撞的可能性有 20%、50% 和 80% 三种情况，而撞向路边旁观者的可能性分为不确定和有风险两种情况：在不确定的情况下，自动驾驶汽车是否会撞到路边旁观者是未知的；而在有风险的情况下，自动驾驶汽车有 50% 的概率会撞到路边旁观者。结果发现，在不确定的情况下，即汽车与路边旁观者相撞的可能性未知时，无论汽车与车道上行人相撞的可能性是多少，被试都倾向于让汽车保持直行；而在有风

https://doi.org/10.35534/pc.0410130

险的情况下，被试的决策期望则与汽车撞到车道上行人的概率有关，当汽车与车道上的行人碰撞概率只有 20% 时，被试仍然期望汽车保持直行，但是当碰撞概率为 80% 时，被试出于结果主义的考虑会选择让汽车转向。

2.2 决策者或评价者特征

决策者或评价者的情绪、认知、心理状态及个体差异等影响着人们对智能机器的道德决策期望(王鹏、方平、江媛，2011 ）。例如，博纳丰等人（ Bonnefon et al.，2016 ）发现，视角影响着被试对机器道德决策态度，相对于行人视角，被试在乘客视角和旁观者视角下选择牺牲行人的可能性明显更高。弗兰克等人（ Frank et al.，2019 ）根据道德决策的双加工模型（ Dual Process Theory ）设计实验操纵，通过限制被试决策时间来启动不同的决策模式：小于 7 秒的条件为直觉模式（ Intuitive Mode ），7 秒到 30 秒的为精加工模式（ Deliberate Mode ）。结果发现，决策模式影响着人们对智能机器道德决策的偏好，直觉决策使得人们转向偏好道义论决策，显著降低了牺牲行人的可能性。

此外，人们对智能机器违背道德规范的安全担忧，以及对智能机器的控制愿望也影响人们对智能机器道德决策期望。远征南（ 2019 ）的研究发现，人们对智能机器违背道德规范的安全担忧及对智能机器的控制愿望在人们对智能机器的道德决策期望中发挥链式中介作用，这表明授权智能机器可以违背道德规范会引起人们的安全担忧，由此引发人们想去控制智能机器的愿望，从而影响了被试对智能机器的决策期望。

2.3 智能机器本身的属性

恐怖谷（ Uncanny Valley ）现象揭示了智能机器的外形如何影响人们的态度，智能机器的外观仿真度越高人们越有好感，但达到一定的临界点时这种好感度会突然降低，机器外观越像人类，人们反而越会感到恐惧（ Mori et al.，2012 ）。迪萨尔沃等人（ DiSalvo et al.，2002 ）也发现，智能机器的面部特征在人类与机器人的交互中起到了重要作用，头部尺寸、面部特征总数等因素会显著影响人们对智能机器的感知。一般根据智能机器的外观可以将机器人分为两种：一种是人形机器人，即模仿人的形态和行为而设计制造的智能机器，外观与人类十分相似；另一种是机械机器人，即仿照各种各样的生物、日常生活物品、交通工具等做出的智能机器，其外观像一个机械装置。马勒等人（ Malle et al.，2016 ）的研究发现，机器外观影响人们对智能机器的道德判断，人们在道德认知上会将机械机器人和人形机器人区别对待，更能接受机械机器人在电车困境中做出功利主义的决策，而不愿接受人形机器人做出功利主义的决策。

3 人们对智能机器的道德决策期望的心理机制

心智感知（ Mind Perception ）是目前解释人们对智能机器道德决策期望的主要理论。格雷等人（ Gray et al.，2007 ）提出，心智感知指的是思考问题、感受情绪以及有意图地实施某个行为的能力，包括主体性（ Agency ）和感受性（ Experience ）两个维度。其中，主体性的高低决定了主体能否承担道德责任，人们通常不会要求儿童、动物承担道德责任，是因为人们知觉到其主体性很低。感受性的高低则意味着一个主体能否共情，能否感受快乐或痛苦等情绪，决定了主体是否需要被保护，是否拥有道德权利。格雷

和瓦格纳（Gray and Wegner，2012）的研究表明，人们对特定主体的心智能力的感知影响着对该主体的道德决策期望。对于智能化程度较低的机器，人们认为它们在复杂的道德情境中无法正确分辨道德和不道德行为，对智能机器的道德决策期望也较低，因而更希望智能机器始终按照道德规范要求行事，不愿意让机器拥有道德决策的自由；而对于智能化程度较高的机器，人们会认为它们具备比人类更优秀的决策判断能力，对机器的道德决策期望较高，可能希望由智能机器代替人类做出决策。

在心智感知中，相较于感受性，对主体性的心智感知可能影响着人们对智能机器的道德决策期望。比格曼和格雷（Bigman and Gray，2018）的研究发现，人们对智能机器的主体性心智感知在人们对智能机器道德决策态度中发挥中介作用，人们之所以不愿意赋予智能机器决策权可能是认为智能机器缺乏像人类一样的思维能力。布罗德本特等人（Broadbent et al.，2013）发现，知觉到智能机器具有更高主体性的人选择使用机器的意愿更低。斯塔福德等人（Stafford et al.，2014）也发现，老人更喜欢那些低主体性（即并不能做很多事）的机器人。这些研究说明，人们可能并不希望机器具有太多的能动性和自主性，而只是希望机器作为一个完成任务的工具而存在。这表明，人们存在着对智能机器的工具化（Instrumentalization）倾向，由此影响着人们对智能机器道德决策的态度。

4　已有研究不足与未来研究方向

4.1　已有研究不足

已有研究以自动驾驶汽车为主要研究对象，难以构建智能机器道德决策的指导性理论。就智能机器的发展而言，社交机器人、医疗护理机器人、自动驾驶汽车、无人战斗机等各种类型的智能机器都将在人类生活中承担不同的角色，发挥不同的作用。这些不同类型的智能机器所面临的道德困境，以及需要遵循的道德法则有所不同。只有通过对承担不同角色的智能机器道德决策期望进行广泛探讨，才能够有望构建智能机器道德决策的指导性理论。

已有研究发现的生态效度具有局限性。已有研究主要采用自我报告法，通过文字呈现智能机器的道德决策情境。这些道德决策情境并非真实的决策情境，一般采用电车困境改编而来，而现实中的道德两难情境更为复杂多样。被试也缺乏与智能机器有直接交互的经验，对决策情境的代入感不足。因此，已有研究发现虽有助于理论探索，却缺乏生态效度。随着大数据、VR 技术的发展，未来研究有必要在接近真实或真实的人机交互情境中展开。

4.2　未来研究方向

4.2.1　探究人对智能机器在不同道德基础情境下的偏好是否不同

道德基础理论（Moral Foundation Theory）认为，人类的道德至少包括关爱 / 伤害、公平 / 欺骗、忠诚 / 背叛、权威 / 服从、神圣 / 贬低、民主 / 压制六个领域，这六种道德有其进化根源、诱发刺激和情绪反应，并作为不同文化下道德规范形成与发展的基础，因此被称为道德基础（Haidt and Joseph，2007）。目前，大多利用电车困境的自动驾驶研究主要涉及关爱 / 伤害维度，未来研究可以在涉及不同道德基础的情境中，系统分析人们对智能机器的道德决策期望。

https://doi.org/10.35534/pc.0410130

4.2.2 人们期望智能机器依照何种道德准则构建自身伦理体系

目前，使智能机器的行为符合道德规范要求的编码方式主要有自上而下（Top-down）和自下而上（Bottom-up）两种进路。其中，自上而下的进路是指将伦理学家或者政府达成共识的道德准则直接编码进智能机器内，使智能机器基于嵌入的道德哲学程序做出道德选择。自下而上的进路是指不给智能机器直接编码道德准则，而是让智能机器通过机器学习、进化算法或强化学习等方式建构出自身的道德原则，例如模仿人类道德直觉的认知架构（Bello and Bringsjord，2012）。陈齐平等人（2019）指出，智能机器依据人类道德规范做出决策会增加人类对机器伦理决策的接受度和信任度，自上而下的编码方式可能更容易被大众所接受。此外，刘纪璐等人（2018）指出，将中国传统的儒家伦理准则纳入智能机器的道德体系，也是一种解决当前人工智能伦理困境的可行方案。人们究竟期望智能机器依照何种道德准则构建自身伦理体系，才能使智能机器在各种复杂的道德情境中灵活做出符合人们期望的反应，是未来研究需要深入探究的重要议题之一。

4.2.3 探究人们对智能机器道德决策期望的影响因素

探究人们对智能机器道德决策期望的影响因素，有利于程序设计师根据人类偏好设计智能机器程序，提高人类对智能机器的信任和接受。如前所述，心智感知影响着人们对特定主体的道德决策期望，因此，人们对技术的态度可能影响人们对智能机器的道德决策期望。一般来说，人们对技术持有工具化倾向和拟人化倾向两种态度。如果人们更多地将智能机器作为人类服务的工具则为工具化倾向，如果认可智能机器的主体性和独立性，期望智能机器像人类一样与人们互动则为拟人化倾向。卡恩等人（Kahn et al，2012）在研究中询问人们如何看待智能机器人，是将其看作一种技术、一种活着的存在，还是介于两者之间的某种存在，结果发现，近一半的被试将机器人看作一种技术，另一半的被试将其看作介于两者之间的某种存在。这意味着，人们对智能机器的看法和态度存在着个体差异，这种差异可能影响人们对智能机器的道德决策期望，未来研究有必要对此展开探讨。

4.2.4 探索智能机器道德决策期望的跨文化差异

道德伦理内嵌于特定的社会文化背景。西方文化个体更看重关爱 / 伤害、公平 / 互惠这两个道德领域，更看重个体价值的平等；而东方文化个体重视权威 / 尊敬、忠诚 / 背叛，对人际关系的依附性更强（Haidt and Joseph，2007）。阿瓦德等人（Awad et al.，2018）将 200 多个国家和地区的数据分为三个道德集群（Moral Clusters），分别是西部集群（北美地区以及信仰天主教和新教的一些欧洲国家等）、东部集群（中国、日本等远东国家和印度尼西亚、巴基斯坦等国家）以及南部集群（中美地区和南美地区的一些拉丁美洲国家等）。这些道德集群内部的国家和地区由于地理、文化的接近，从而产生对机器道德决策的相似偏好，但世界上不同道德集群的人们对自动驾驶汽车的决策期望却存在着一些差异。例如，在电车困境中牺牲老人以拯救年轻人的倾向在东方文化中非常弱，这可能与东方文化普遍尊重老人的传统有关。未来研究可以从跨文化比较，或从文化的动态建构视角出发探讨人们对智能机器道德期望的文化特异性。

随着智能机器在人类日常生活和工作中的不断融入，人工智能产品不再像过去一样仅仅被人类当成工具来看待，更多的新角色和功能被赋予到高智能机器人身上（闫坤如，2018）。现实世界中的许多案例也表明，人们对智能机器也可能产生复杂的情感联结，将其拟人化为伴侣、朋友甚至是救命恩人。智能机器在发挥作用的过程中将会涉及越来越多的道德问题，并承担一定的道德后果。探讨人们对智能机

器的道德决策期望，不仅对于完善机器伦理学具有重要理论意义，也有助于对人工智能伦理法律法规的制定。目前，对机器道德的探讨尚处于起步阶段，亟需更多机器伦理学研究给出答案。

参考文献

[1] 陈齐平，魏佳成，钟陈志鹏，等. 智能机器伦理决策设计研究综述[J]. 计算机科学与探索，2019，13（11）：1801–1812.
[2] 褚华东，李园园，叶君惠，等. 个人—非个人道德困境下人对智能机器道德判断研究[J]. 应用心理学，2019，25（3）：262–271.
[3] 刘纪璐，谢晨云，闵超琴，等. 儒家机器人伦理[J]. 思想与文化，2018（1）：18–40.
[4] 苏令银. 创造智能道德机器的伦理困境及其破解策略[J]. 理论探索，2019（4）：30–37.
[5] 王鹏，方平，江媛. 道德直觉背景下的道德决策：影响因素探究[J]. 心理科学进展，2011，11（4）：119–125.
[6] 闫坤如. 人工智能的道德风险及其规避路径[J]. 上海师范大学学报（哲学社会科学版），2018，47（2）：42–49.
[7] 远征南. 人们对自主机器道德决策期望的探索性研究[D]. 杭州：浙江大学，2018.
[8] Awad E，Dsouza S，Kim R，et al. The moral machine experiment[J]. Nature，2018，563（7729）：59–64.
[9] Awad E，Dsouza S，Shariff A，et al. Universals and variations in moral decisions made in 42 countries by 70，000 participants[J]. PNAS，2020，117（5）：2332–2337.
[10] Bello P，Bringsjord S. On how to build a moral machine[J]. Topoi，2012，32（2）：251–266.
[11] Bergmann L T，Schlicht L，Meixner C，et al. Autonomous vehicles require socio–political acceptance–an empirical and philosophical perspective on the problem of moral decision making[J]. Frontier in Behavioral Neuroscience，2018（12）：article 31.
[12] Bigman Y E，Gray K. People are averse to machines making moral decisions[J]. Cognition，2018，181（1）：21–34.
[13] Bonnefon J F，Shariff A，Rahwan I. The social dilemma of autonomous vehicles[J]. Science，2016，352（6293）：1573–1576.
[14] Broadbent E，Kumar V，Li X. et al. Robots with display screens：a robot with a more humanlike face display is perceived to have more mind and a better personality[J]. PLoS One，2013，8（8）：e72589.
[15] Brozek B，Janik B. Can artificial intelligences be moral agents[J]. New Ideas in Psychology，2019（54）：101–106.
[16] Damm L. Moral machines：teaching robots right from wrong[J]. Philosophical Psychology，2011，25（1）：1–5.
[17] Dietvorst B J，Simmons J P，Massey C. Algorithm aversion：People erroneously avoid algorithms after seeing them err[J]. Journal of Experimental Psychology（General），2015，144（1）：114–126.
[18] DiSalvo C F，Gemperle F，Forlizzi J，et al. All robots are not created equal：The design and perception of humanoid robot heads[C]. Paper presented at the Proceedings of the 4th conference on Designing interactive systems：processes，practices，methods，and techniques. London，England，2002：321–326.
[19] Floridi L，Sanders J W. On the morality of artificial agents[J]. Minds & Machines，2004，14（3）：

https://doi.org/10.35534/pc.0410130

349–379.

[20] Frank D, Chrysochou P, Mitkidis P. Human decision–making biases in the moral dilemmas of autonomous vehicles [J]. Scientific Reports, 2019 (9): 13080.

[21] Gogoll J, Uhl M. Rage against the machine: Automation in the moral domain [J]. Journal of Behavioral & Experimental Economics, 2018, 74 (1): 97–103.

[22] Gray H M, Gray K, Wegner D M. Dimensions of mind perception [J]. Science, 2007, 315 (5812): 619.

[23] Gray K, Wegner D M. Feeling robots and human zombies: Mind perception and the uncanny valley [J]. Cognition, 2012, 125 (1): 125–130.

[24] Haidt J, Joseph C. The moral mind: How 5 sets of innate intuitions guide the development of many culture–specific virtues, and perhaps even modules [M] // Carruthers P, Laurence S, Stich S. The innate mind. New York, NY: Oxford University Press, 2007.

[25] Johnson D G, Miller K W. Un–making artificial moral agents [J]. Ethics & Information Technology, 2008, 10 (2/3): 123–133.

[26] Kahn P H, Kanda T, Ishiguro H, et al. Do people hold a humanoid robot morally accountable for the harm it causes [C]. Paper presented at Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction. Boston, Massachusetts, 2012: 33–40.

[27] Logg J M, Minson J A, Moore D A. Algorithm appreciation: People prefer algorithmic to human judgment [J]. Organizational Behavior and Human Decision Processes, 2019, 151 (10): 90–103.

[28] Malle B F, Scheutz M, Arnold T, et al. Sacrifice one for the good of many: people apply different moral norms to human and robot agents [C]. Paper presented at the Tenth ACM/IEEE International Conference on Human–Robot Interaction. Portland, Oregon, 2015: 117–124.

[29] Malle B F, Scheutz M, Forlizzi J, et al. Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot [C]. Paper presented at the 11th ACM/IEEE International Conference on Human–Robot Interaction. Christchurch, New Zealand, 2016: 125–132.

[30] Martin P, Andreas S. Can technological artefacts be moral agents [J]. Science & Engineering Ethics, 2011, 17 (3): 411–424.

[31] Meder B, Fleischhut N, Krumnau N, et al. How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty [J]. Risk Analysis, 2019, 39 (2): 295–314.

[32] Mori M, Mac Dorman K F, Kageki N. The uncanny valley [J]. IEEE Robotics & Automation Magazine, 2012, 19 (2): 98–100.

[33] Shenhav A, Greene J D. Moral judgments recruit domain–general valuation mechanisms to integrate representations of probability and magnitude [J]. Neuron, 2010, 67 (4): 667–677.

[34] Shou Y, Song F. Decisions in moral dilemmas: The influence of subjective beliefs in outcome probabilities [J]. Judgment and Decision Making, 2017, 12 (5): 481–490.

[35] Stafford R, Mac Donald B, Jayawardena C, et al. Does the robot have a mind? mind perception and attitudes towards robots predict use of an eldercare robot [J]. International Journal of Social Robotics, 2014, 6 (1): 17–32.

[36] Stein J P, Ohler P. Venturing into the uncanny valley of mind–the influence of mind attribution on the acceptance of human–like characters in a virtual reality setting [J]. Cognition, 2017, 160 (1): 43–50.

［37］Wallach W，Allen C，Smit I. Machine morality：Bottom-up and top-down approaches for modelling human moral faculties［J］. AI & Society，2008，22（4）：565-582.

A Review of the Research on People's Expectation for Moral Decision-making of Intelligent Machines

Wu Mingzheng[1] Yan Mengyao[1] Lin Ming[1] Liu Yiyao[1] Sun Xiaoling[2]

1. Department of psychology and behavioral sciences, Zhejiang University, Hangzhou;
2. Department of psychology, Hangzhou Normal University, Hangzhou

Abstract: Intelligent machines refers to the AI system that can perform tasks autonomously or interactively in various environments. In recent years, more and more researchers have paid attention to the moral decision-making of intelligent machine. This review systematically introduces the research status of intelligent machines on three topics: people's expectation of moral decision, influencing factors and psychological mechanism. Limitation and future research directions were also discussed.

Key words: Intelligent machine; Moral decision making; Mind perception

https://doi.org/10.35534/pc.0410130