自动驾驶算法设计中的伦理决策
——基于“有意义的人类控制”

李德新[1],宫志超[2]

1. 山西大学科学技术哲学研究中心,太原 030006
2. 山西大学哲学社会学学院,太原 030006

摘要　基于“有意义的人类控制”这一人工智能伦理学的核心概念,总结了自动驾驶在算法设计阶段的伦理难题;分析了“有意义的人类控制”运用于自动驾驶的可行性;从“跟踪”和“追踪”2大条件,围绕“问责制与透明度”和“价值敏感设计”进行“有意义的人类控制”框架构建,以为自动驾驶的算法设计提供系统方法论指导。

关键词　自动驾驶;伦理困境;电车难题;有意义的人类控制;价值敏感设计

中国于2021年8月20日正式发布了《汽车驾驶自动化分级》标准,驾驶自动化分为了应急辅助、部分驾驶辅助、组合驾驶辅助、有条件自动驾驶、高度自动驾驶、完全自动驾驶6个等级[1]。自动驾驶的发展拥有很大潜力,其技术优势体现在多个方面,包括降低意外事故的发生率、减少事故伤亡人数、提高道路资源的利用率、优化驾驶环境等,最终自动驾驶可能促成一个更安全、更高效、更公平的道路交通系统。

尽管自动驾驶在行驶的安全性上远超人类司机,但绝对安全和零事故仍难以企及。在开放的道路环境下,自动驾驶将在不可预测的行人、骑手、人类司机、动物以及任何其他可能出现在道路中的利益相关者中行驶,自动驾驶会面临复杂的道路情况。与其他交通工具不同,自动驾驶可以预测各种碰撞轨迹,并选择损坏或碰撞可能性最低的路径。在选择碰撞路径时,会涉及到多种伦理决策,例如:在牺牲行人和牺牲乘客之间作出选择。通过与“电车难题”案例进行类比来讨论自动驾驶如何在事故场景下作出正确的伦理决策是当下的研究热点。然而,如此狭隘的问题与工程师实际的编程方式形成了鲜明对比,研究不能局限于“电车难题”引发的

收稿日期:2022-12-27;修回日期:2023-02-25

基金项目:国家社会科学基金重大项目(18ZDA030);山西省高等学校人文社科重点研究基地项目(2022J002);山西省研究生教育教学改革课题(2022YJJG031)

作者简介:李德新,副教授,研究方向为科技伦理、科学哲学,电子信箱:ldx@sxu.edu.cn;宫志超(共同第一作者),博士研究生,研究方向为科技伦理,电子信箱:madradaist@163.com

引用格式:李德新,宫志超. 自动驾驶算法设计中的伦理决策——基于“有意义的人类控制”[J]. 科技导报, 2023, 41(7): 47-54; doi: 10.3981/j.issn.1000-7857.2023.07.005

伦理困境,而应该专注于如何以符合伦理道德的方式进行算法设计,Gerdes即指出:“我们是在道德编程,而不是编程道德。”[2]人类价值观作为自动驾驶算法设计的核心,相关利益者的价值诉求需要纳入到技术设计过程中,通过把自动驾驶置于“有意义的人类控制”(meaningful human control)之下,实现道德编程。

1 自动驾驶的伦理挑战与反思

1.1 自动驾驶的道德困境

鉴于自动驾驶技术的飞速发展和所涉及的严重风险,如何对自动驾驶进行编程已经成为了一个亟待解决的伦理问题。2016年,Bonnefon等[3]调查了司机在面临涉及自我牺牲的汽车事故时潜在的不同决定。实验显示多数参与者表现出后果主义倾向,即尽量减少伤亡人数。然而,这些参与者也表示,他们不会购买以这种方式作出决定并可能使自己和家人面临风险的车辆。参与者从“安全旁观者”转向“潜在牺牲者”时,实验结果出现了“言语”和“行动”分离的现象,自动驾驶在算法设计上陷入了道德困境。除此之外,“电车难题”引起的自动驾驶的伦理困境已经成为了当下的研究热点。自1967年Foot[4]提出“电车难题”以来,这一思想实验也已成为自动驾驶伦理算法实验的经典范式。然而,“电车难题”和自动驾驶处理碰撞的事故场景之间存在着明显差异。首先,“电车难题”实验中涉及的是剥离身份、年龄以及具体行为的抽象人,而自动驾驶在现实生活中需要面对更复杂的多因素情况[5];其次,“电车难题”需要在事故场景中作出“分秒决策”,而自动驾驶需要预先编程,以应对不同类型的事故场景(表1)。

表1 自动驾驶与“电车难题”事故场景区别

场景	决策者	决策时间	情境因素	责任、道德、法律	认知情况
自动驾驶	个体群体/利益相关者群体	预期决定/应急计划	无限制/多因素	考虑	不考虑
电车难题	个人	分秒决策	有限因素/抽象条件	不确定的风险评估和伦理决策	确定的情况

从“电车难题”与自动驾驶的对比中,提出了以下几点反思。

1)是否能够赋予自动驾驶同人类一样的道德主体地位。这个问题需要在心灵哲学领域进行深入研究,目前已经有研究人员认为可以赋予自主机器部分道德主体地位,Brändle等[6]认为自主机器作为明确的道德行为者,但同人类又有一定的区别,他将自主机器视为衍生的参与者,只模拟人类典型的能力,如“行动”和“决定”等。因此,自动驾驶的行为标准必须由人类决定。若采用自上而下的方法,自动驾驶被赋予某些决策标准作为学习系统[7]。若采用自下而上的方法,“自动驾驶可以使用类似的人工智能技术,加上更多的训练实验数据,来学习人类在复杂的驾驶环境中如何进行伦理选择,从而期望人工智能技术与伦理能够集成为一个整体对象,即融合成一种面向人工智能技术本身的伦理框架”[8],自动驾驶系统独立地从个案中得出相关的决策标准,随后检查是否符合公认的基本标准。这意味着道路交通中自动驾驶的道德决定最终是由人类做出,而如何保证把自动驾驶的伦理决策置于人类控制之下才是重中之重,因此将围绕“有意义的人类控制”详细讨论。

2)人类应该采用什么道德标准进行算法设计。目前,基于某种道德标准解决伦理难题进展甚微。迄今为止,在人工智能中探索实施的道德框架在很大程度上是义务论,例如阿西莫夫机器人定律、康德道德理论以及罗尔斯差分原理等。虽然道义伦理可以在很多情况下提供指导,但由于任何一套规则的不完整性和难以将复杂的人类伦理作为一套规则来表述,所以道义论并不适合作为一个完整的道德框架。在“电车难题”范式之下探讨伦理困境时,功利主义正在成为主流。然而,功利主义

忽视了行为者的意图,把个人生命视为同质和可置换的,只关注给定行动或规则的结果,以最大限度的提高总体幸福,“没有充分认识到公共道德对于伦理决策的重要性”[9]。Noah[10]指出,当试图将成本降到最低时,自动驾驶会选择与安全等级较高的车辆碰撞,那些为安全付出更多努力的人反而成为被伤害的目标是不公平的。

综上所述,没有一种道德标准可以普遍有效的解决自动驾驶的道德困境。因此,随着自动驾驶技术的发展,自动驾驶的伦理决策可能在不确定的道德标准下作出。如何在不确定的道德标准下进行道德编程成为难题。Millar[11]在讨论医疗系统的技术设计时,明确提出了技术设计需要考虑患者的伦理自主性,否认了技术中立性,认为技术应该被描述为一个人行事的道德代理,可以把设计师、技术设备和用户之间描述为一种特别的道德关系,当技术设备充当道德代理时,用户应该合理地最大化他们的自主权。因此,即使自动驾驶在算法设计上的道德标准是不确定的,但算法标准至少必须与现有的社会价值观符合,需要充分考虑自动驾驶利益相关者的价值诉求,制造商、驾驶员、政府、行人、伦理学家等利益相关者都要参与到算法设计过程中,当不同的价值主体发生利益冲突时,价值敏感设计为价值主体间冲突解决和利益重构提供了解决思路。

1.2　自动驾驶面临的问责难题

与自动驾驶的伦理困境不同,如何处理自动驾驶的事故问题具有重要的现实意义。虽然尚不清楚自动驾驶的道德困境在道路交通中的实际发生频率,但道路交通中自动驾驶造成的事故是一个极其严重的问题。2016年2月谷歌自动驾驶汽车事故以及2016年5月特斯拉自动驾驶致死事故表明,目前自动驾驶技术无法完全履行与之相关的安全承诺,事故风险依然存在。谁应该对自动驾驶造成的事故后果负责变得尤为重要。

就必要的硬件和软件的设计而言,参与自动驾驶汽车开发的主要是制造商和程序员,软件开发尤其涉及核心伦理问题。除此之外,管理人员、监管者、法律专家、保险公司和政治决策者等利益相关者也应参与其中,而谁应该对自动驾驶的事故后果负责将难以鉴定。

通常自动驾驶的事故后果责任是由相应的制造商或程序员承担,然而在没有明显制造缺陷的情况下,并非所有情况下都有理由将事故责任转移给他们。除了制造商和程序员之外,驾驶员在责任问题上也扮演着矛盾角色。Hevelke等[12]指出,在自动驾驶背景下,驾驶员只有“乘客”角色,因此不能被视为事故原因。然而,驾驶员作为系统受益者,必须对与自动驾驶相关的事故风险负部分责任,即驾驶员的严格责任。

现有法律框架之下的产品责任制和风险承担制都存在局限性,这种不确定的责任框架对制造商和消费者造成的阻碍从根本上减缓了自动驾驶的发展。因此,制定专门的自动驾驶问责机制确有必要。围绕”有意义的人类控制”进行的问责制设计将为解决自动驾驶的问责难题提供思路。

2　“有意义的人类控制”与自动驾驶

人类价值是算法设计的重要参考维度,需要把相关利益者的价值诉求纳入到技术设计过程中,将自动驾驶置于“有意义的人类控制”之下,从而以符合伦理道德的方式进行算法设计。

2.1　“有意义的人类控制”的提出与理论基础

2016年4月11到15日,《特定常规武器公约(CCW)》第三次非正式专家会议在联合国日内瓦办事处召开,围绕“自主性”“致命性自主武器系统(LAWS)可行定义”“国际人道主义法面临的难题”等议题进行了深入讨论[13]。在伦理层面上,讨论焦点在于“致命性自主武器系统(LAWS)”对于人权和道德的冲击,各方基本认为不应将生死决定权让渡给机器,但对机器是否作为道德主体等问题存在分歧[14]。会议上首次提出了“有意义的人类控制(meaningful human control)”这一概念。随后,Filippo等[15]对“有意义的人类控制”作了进一步阐述,他认为“有意义的人类控制”是一种从代理对动作的直接操作控制出发,转向源于人类行动原因的控制机制,“有意义的人类控制”必需满足以下2个条件。一是跟踪条件,自主系统应该能够响应人类

设计和部署的相关道德原因和系统运行环境中的相关事实;二是追踪条件,自主系统的设计方式必须保证操作结果可以追溯到设计和操作链中的至少1个人。

“有意义的人类控制”的哲学解释可以追溯到20世纪关于道德责任的辩论。道德不相容论者认为,因果关系和道德责任无法调和,否认人类行为的因果解释与道德责任的相容性。当且仅当人类拥有做出决定和执行的特殊权力时,才能控制自己的行为并对其行为负责,从而摆脱遗传(神经)生物学、社会心理学和环境因素的因果影响。传统道德相容论者认为,即使人类不具备任何特殊的形而上学力量来逃避对其行为的因果影响,人类也可能对某些行为负有道德责任。霍布斯和休谟等哲学家的传统相容论是以机械论和联想主义的人类思想观点为基础,根据传统的相容论,为了让代理人在道德上负责,代理人行为只需要在成为内部动机因素(欲望、意图、性格特征、价值观)的因果产物的意义上是自由的,而不是“外部”力量,即身体或心理胁迫的产物。而当代道德相容论者否认精神因果关系作为道德责任的基础,例如由严重精神失常的人做出的行为。因此,对行为进行理性控制的能力是道德责任的关键。

Fischer和Ravizza[16]提出了非常有影响力的“指导控制”这一道德相容理论。所谓的“指导控制”就是当一个人需要在道德上对其行为负责时,这个人就应该对其行为拥有“指导控制”权,而拥有“指导控制”权需要满足以下2个条件:一是适度的理性响应,二是决策机制应该是代理本身。本研究认为,从“指导控制”理论出发解释“有意义的人类控制”是一个有意义的哲学起点。

2.2 “有意义的人类控制”应用于自动驾驶的可行性分析

尽管“有意义的人类控制”概念来自关于禁止完全自主武器系统的讨论,但它完全可以拓展到更为广泛的领域,例如,为人工智能的负责任创新提供指导作用。自动驾驶作为人工智能的一个新兴应用领域,旨在促进形成一个更安全可靠的道路环境,“有意义的人类控制”完全可以成为自动驾驶在技术设计阶段的方法论指导。

Schwarz[17]认为,人类介入自动驾驶系统的控制会面临3个方面的限制:人机互动中出现的认知限制,当驾驶员与高度自动化的驾驶系统交互时,驾驶员的认知水平会限制人机交互水平,这大大降低了人类在决策过程的参与度;人类无法有效地访问和利用自动驾驶系统在其功能过程中处理的信息量,有效的决策是困难的;时间限制,即由于认知和身体限制,人类无法及时对当下场景做出决策。实际上,“有意义的人类控制”不同于传统的控制概念,不是系统对控制器动作的直接响应,而是对包括驾驶员、设计者、监管者以及与系统交互的其他人等在内的相关人员的响应。Leila[18]认为,“有意义的人类控制”是一种控制框架,在这种框架下,用户和系统之间的交互以透明、可追溯的方式进行。如果一项行动受到质疑,那么至少可以确定因果链中的一名负责人并追究其责任,同时系统需要以负责任的方式开发,以响应与给定情况相关的人类道德原因。这种控制概念会把控制器的范围扩展到某个时候参与了设计、部署或监管而不直接接触受控系统的个人。因此,“有意义的人类控制”不同于传统的控制概念,可以用于自动驾驶的算法设计。接下来,本研究根据“有意义的人类控制”的2个基本概念进行进一步的可行性论证。

根据“有意义的人类控制”的“跟踪条件”:一方面,自动驾驶必须遵守交通规则并反映道路使用者的相关利益。例如,在道路交通中运行的自动驾驶系统应该比在商业环境中运行的交互式服务机器人更能响应人类用户的遇险信号;另一方面,提高自动驾驶系统对相关道德原因的响应能力同时,还可以设计环境以防止类似于道德难题的情景出现。

“有意义的人类控制”的“追踪”条件要求自主系统的设计方式必须保证操作结果可以追溯到设计和操作链中的至少1个人。这就将设计任务的范围扩展到了自动驾驶系统设计和环境设计层面之外的第三个层面:社会和制度实践设计,满足追踪条件的设计意味着操作链条上的不同人类代理在技术和心理上都能够完成任务,并清楚自身需要对自主系统的行为负责。因此,不仅要在功能层面

明确人类和自动驾驶之间的任务分配,即哪些驾驶操作应该交给计算机,哪些应该由人类负责;还要在社会层面明确人类在哪些情况下能够有动力作出“有意义的人类控制”,这意味着我们需要社会调查来评估期待人类作出某些规范操作的合理性。

综上所述,“有意义的人类控制”不仅局限于禁止自主武器系统的政治讨论,还可以成为自动驾驶系统设计的核心概念之一,为自动驾驶的算法设计提供方法论指导。

3　“有意义的人类控制”下自动驾驶算法设计原则

从“跟踪”和“追踪”两大条件,对“有意义的人类控制”进行了框架构建,旨在为自动驾驶的算法设计提供更为系统的方法论指导。

3.1　基于“跟踪条件”的价值敏感设计

“有意义的人类控制”的“跟踪条件”指出,自主系统应该能够响应人类设计和部署的相关道德原因和系统运行环境中的相关事实。价值敏感设计是一种在技术设计过程中全面系统地考量人类价值的设计方法,需要所有利益相关者参与到设计过程中以促进技术设计的人性化发展。在自动驾驶算法设计阶段运用价值敏感设计,可以实现自动驾驶系统对人类代理者的理性响应,为“跟踪条件”的实现提供了设计方向。

1）价值敏感设计的提出及其哲学基础。

Friedman是华盛顿大学信息学院教授,于20世纪90年代开始技术设计与人类价值的相关研究。1996年,Friedman等[19]在论文《价值敏感设计》中首次对价值敏感设计进行了论述,文章从用户自主性和公正无偏见2种价值角度说明了价值敏感设计的重要性。Friedman从体现立场、外生立场、交互立场对人类价值观如何融入技术设计进行了阐述[20]。体现立场指出,设计者将自身的意图和价值观嵌入技术中,技术应用和开展的结果将决定使用者的特定行为;外生立场指出,经济、政治、种族、阶级、性别等社会力量影响了技术的使用方式,技术系统不是价值中立的,而是始终偏向拥有经济和政治权力的人的利益;交互立场指出,尽管技术中的特征或属性更容易支持某些价值并阻碍其他价值,但技术的实际使用取决于与之交互的人们的目标,交互立场一方面强调了技术设计的属性,另一方面强调了用户如何在社会组织结构的背景下使用技术。

技术设计过程中要充分吸收道德差异性和道德普遍性的可取之处。揭示道德普遍性的理论家通常认为,其在与道德的本质及其最有意义的属性作斗争。相比之下,描述道德差异性的理论家认为,当你拥有一个跨越文化的共同道德特征时,这种抽象形式已经失去了所有的意义和效用。Friedman对此持中间立场,我们不仅要对普遍道德价值观进行分析,而且要注意这些价值观在特定时间和特定文化中发挥的不同作用。

目前价值敏感设计的研究大致分为3个阶段:概念阶段、实证阶段和技术阶段(图1)。概念阶段需要对所调查的中心结构和问题进行哲学分析,如哲学类文献如何定义某些特定价值,区分受此项设计影响的直接和间接利益相关者,如何在技术的设计、实施和使用中权衡相互矛盾的价值等。此外,还需要设定上述问题发生的历史背景与现实背景,从整体上描述具体的技术设计问题。实证阶段需要使用观察、访谈、调查、实验操作、相关文件收集、用户行为和人体生理测量等定量和定性方法分析受技术影响的人们的背景和体验。在技术阶段,需要对具体的技术细节进行研究,分析当前的技术机制和设计,以评估对特定价值的支持程度。

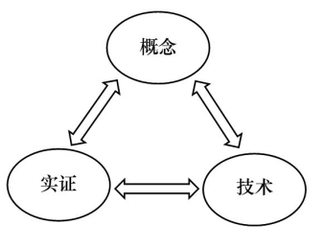

图1　价值敏感设计框架图

2）自动驾驶的价值敏感设计。

随着自动驾驶的发展,算法设计阶段的伦理诉求越来越重要,当同一价值主体面临价值抉择或者不同价值主体的利益诉求发生冲突时,设计人员需

要根据一定的价值级序确定使用价值或优先使用价值。因此,价值敏感设计需要所有的利益相关者参与到技术设计的过程以促进技术设计人性化发展。在自动驾驶的伦理决策中,价值敏感设计要求制造商、驾驶员、政府、行人、伦理学家等利益相关者参与到设计过程中,当不同的价值主体发生利益冲突时,本研究从概念阶段提出了3点解决方案。

(1) 明确价值级序,通过确定一种价值作为最重要的参量解决利益冲突,然后在满足该价值的设计空间内处理剩余价值。例如,当自动驾驶汽车驶向人行横道时,安全性是最重要的价值参量,在保证安全性的设计空间内,设计师可以充分发挥自动驾驶的机动性。(2) 通过平衡不同价值主体的利益冲突解决紧张局势。自动驾驶的一个基本功能是将乘客(或货物)从A点运送到B点。为了实现此功能,自动驾驶规划路径中最常用的就是最短路径搜索算法。当车辆路线对社区的安宁或交通系统造成影响时,不同价值主体就出现了利益冲突。我们可以对行驶时间和交通影响进行量化,适当调整成本函数中的相关参数来解决不同价值主体的利益冲突,这种方式类似于经济学领域中的帕累托最优化。(3) 通过重构化解利益冲突。当自动驾驶汽车行驶在开放道路环境中,驾驶员和行人之间必然会发生利益冲突。可以通过重构交通设施来防止这种可能性发生。例如,为自动驾驶和传统车辆提供单独车道从而使车辆无法与行人进行互动。

3.2 基于"追踪条件"的问责制和透明度

"有意义的人类控制"的"追踪条件"要求自主系统的设计方式必须保证操作结果可以追溯到设计和操作链中的至少1个人。问责制和透明设计作为有效的追溯手段,可以帮助我们找到责任因果链中的相关参与者并确定哪些参与者应该对自动驾驶的行为负责。因此,问责制和透明度设计为"追踪条件"的实现提供了方向。

1) 问责制设计。

自动驾驶汽车用户代表了一个利益集团,一方面受益于该技术,另一方面作为这种车辆的乘员,也面临事故的风险。除制造商和程序员外,还需要其他行动者参与自动驾驶设计过程,如管理人员、监管者、法律专家、保险公司和政治决策者等。此外,利益相关者还包括非自动驾驶汽车以及骑自行车的人或行人,他们是道路交通中与自动驾驶汽车共享道路的弱势群体。

随着自动驾驶意外事故的不断发生,有效地让责任方对系统行为负责对于维持公众对技术的信任至关重要。然而,当前的道德和法律框架未能明确回答谁应对自动驾驶采取的行动以及如何采取行动负责,问责难题持续存在。Bryson等[21]认为,"虽然系统本身不能被授予法人资格并被追究责任,但可以对其开发、部署和使用而受益的组织和个人进行追责"。Mark[22]将问责制分解为法律责任、专业责任、政治责任、行政问责和社会责任5种不同的类型,每种类型都有自己的执行机制和对行为的控制手段。问责制是对系统的行为及其潜在影响负责,其中的组织和个人作为责任链的一部分,有义务解释和证明他们的决定。

关于问责制的设计有如下2点建议:一是通过教育治理等举措让自动驾驶责任链的各方承认和理解自己应负的责任。如此才能通过研究技术解决方案、社会组织活动以及与技术相关的流程来提高我们的道德和法律远见,并建立问责制和责任实践;二是需要从技术和社会两个层面规范责任要求。例如在社会层面上,自动驾驶在算法设计过程中要尊重公民的隐私权;在技术层面上,需要确保算法设计的稳定性与持续性。虽然责任是前瞻性思维,即采取行动阻止违反我们道德和法律价值观的行为发生,但责任是一种"向后看"的形式,需要提供事件发生后的描述。为了履行其职能,问责制不仅需要将责任各方纳入考虑范围,而且还需要将行动追溯到适当的责任方。

2) 透明度。

透明度并非新概念,早在19世纪,透明原则已成为政治、经济等领域的一条基本准则。Williams[23]将透明度定义为"组织以书面和口头形式向投资者、监管机构和市场中介机构提供相关、及时和可靠信息的程度"。就自动驾驶来说,透明度就是责任链上的参与者需要接受更多关于系统的信息,从而判断系统是否能够按照预期运行以及需要

做出何种改变。自动驾驶问责成为难题,透明度受到持续关注,从透明度出发进行算法设计有如下优势:(1) 人类的认知水平有限,传统的信息系统测试技术无法对自动驾驶系统进行有效的调试和理解。透明度通常可以作为代理与其系统环境交互时了解代理行为的手段,通过对系统的充分了解,用户可以就何时接受或拒绝系统采取的操作做出明智决定,从而对系统进行更有效的控制,提高系统的安全操作和性能。(2) 透明度也被认为是在算法决策中追求公平的一种手段。系统数据反映了随着系统的持续使用而延续和放大的社会偏见,在语言语料库上训练的算法会获得历史偏见。当然,数据并不是嵌入在自动驾驶系统中的唯一偏见来源,人类在决策时受到文化影响,会存在形成有害偏见的风险,而这些偏见会通过实践得到加强。透明度可以识别和解决不必要的偏见以确保算法设计的公平性。

然而,算法透明度也存在其局限性。一是自动驾驶智能系统是复杂的社会技术生态系统的一部分,在利益相关者的决定、利益和整体背景没有足够开放的情况下,算法的透明度只不过是窥视整个社会技术系统的有限部分。因此,在透明度设计中,我们需要调查更广泛的背景,需要考虑社会法律因素以及其他参与者和系统的行动;二是透明度并不能成为一种独立的伦理准则进行算法问责。事实上,透明度只是一种有利于伦理实现的条件,“当透明度提供了支持伦理原则所需的信息,或者信息如何被约束(监管)的细节时,信息透明在伦理上是可行的”[24]。因此,透明度需要与各种伦理原则共同作用,才能促进算法问责。在自动驾驶中,这些伦理原则必须符合制造商、驾驶员、政府、行人、伦理学家等责任相关者的价值诉求。

4　结论

自动驾驶的发展伴随着算法设计的诸多道德争议,其道德困境和问责难题受到大家的广泛关注。文章对自动驾驶道德困境进行介绍,分析了不确定的道德标准下如何进行道德编程,并从关于自主武器的讨论中引入了“有意义的人类控制”这一概念,应用到了自动驾驶的系统设计中。本研究从“跟踪”和“追踪”2大条件出发,围绕自动驾驶提出了价值敏感设计和问责与透明度设计理念,突出了人类价值在技术设计中的重要性,为自动驾驶的算法设计提供了系统的方法论指导。除了“结果论“道义论”等,我们发现多种道德框架的结合正在成为自动驾驶伦理决策的一种可能,Veljko[25]提出了一种基于神经科学中“代理人-行为-后果(ADC)”模型的伦理决策理论,把美德伦理、义务论和功利主义3种道德框架相结合来提高自动驾驶在伦理决策中的稳定性和灵活性。然而,在目前的条件下,尽管算法已经定义了各种各样的变量、参数与函数,一定条件下可以直接转化为机制设计中收益空间或行动策略的参数[26],但其仍然只是在有限的范围内模拟人类的认知判断,人类在行动中包含的道德因素依靠目前自动驾驶系统的结构和算力是无法精确捕捉的。总的来说,自动驾驶无论在何种道德框架下作出伦理决策,人类价值观都是自动驾驶算法设计的重要参考维度。需要进一步讨论对自动驾驶系统进行有效治理的重要性,以保证系统在动态环境下具备适应性水平的人类响应能力。

参考文献(References)

[1] 国家市场监督管理总局, 国家标准化管理委员会. 汽车驾驶自动化分级: GB/T 40429—2021[S]. 2021: 3-5.

[2] Christian G J, Thornton S M, Millar J. Designing automated vehicles around human values[C]//Road Vehicle Automation 6. Automated Vehicles Symposium 2018. Switzerland: Springer Nature Switzerland AG, 2019: 39-48.

[3] Bonnefon J F, Shariff A, Rahwan I. The social dilemma of autonomous vehicles[J]. Science, 2016, 352(6293): 1573-1576.

[4] Foot P. The problem of abortion and the doctrine of the double effect[J]. Oxford Review, 1967, 5: 5-15.

[5] 隋婷婷, 郭晓. 自动驾驶电车难题的伦理算法研究[J]. 自然辩证法通讯, 2020, 42(10): 85-90.

[6] Brändle C, Grunwald A. Autonomes fahren aus sicht der maschinenethik[C]//Bendel O. Handbuch Maschinenethik. Wiesbaden: Springer, 2019: 281-300.

[7] Wallach W, Allen C. Moral machines: Teaching robots right from wrong[M]. Oxford: Oxford University Press, 2009: 83.
[8] 潘恩荣, 杨嘉帆. 面向技术本身的人工智能伦理框架: 以自动驾驶系统为例[J]. 自然辩证法通讯, 2020, 42(3): 33-39.
[9] Awad E, Dsouza S, Kim R, et al. The moral machine experiment[J]. Nature, 2018, 563(7729): 59-64.
[10] Noah J G. Can you program ethics into a self-driving car [J]. IEEE Spectrum, 2016, 53(6): 28-58.
[11] Millar J. Technology as moral proxy: Autonomy and paternalism by design[J]. IEEE Technology and Society Magazine, 2015, 34(2): 47-55.
[12] Hevelke A, Julian N. Responsibility for crashes of autonomous vehicles: An ethical analysis[J]. Science and Engineering Ethics, 2015, 21(3): 619-630.
[13] Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)[EB/OL]. (2016-06-10) [2022-03-20]. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G16/117/16/pdf/ G1611716.pdf.
[14] 徐能武, 龙坤. 联合国CCW框架下致命性自主武器系统军控辩争的焦点与趋势[J]. 国际安全研究, 2019, 37(5): 108-132.
[15] Filippo S S, Jeroen H. Meaningful human control over autonomous systems: A philosophical account[J]. Frontiers in Robotics and AI, 2018, 5(15): 1-14.
[16] Fischer J M, Ravizza M. Précis of responsibility and control: A theory of moral responsibility[J]. Philosophy and Phenomenological Research, 2000, 61(2): 441-445.
[17] Schwarz E. The (im)possibility of meaningful human control for lethal autonomous weapon systems[EB/OL]. (2018-08-29) [2021-10-20]. https://blogs. icrc. org/law and policy.
[18] Leila M, Andrea A T, Virginia D, et al. Let me take over: Variable autonomy for meaningful human control [J]. Frontiers in Artifificial Intelligence, 2021, 4: 133-143.
[19] Friedman B, Freier N. Theories of information behavior [M]. New Jersey: Information Today, 2005: 370.
[20] Andrew S, Julie A J. The human-computer interaction handbook[M]. Florida: CRC Press, 2007: 1178.
[21] Bryson J J, Diamantis M E, Grant T D. Of, for, and by the people: The legal lacuna of synthetic persons[J]. Artificial Intelligence and Law, 2017, 25(3): 273-291.
[22] Mark B. Analysing and assessing accountability: A conceptual framework[J]. European Law Journal, 2007, 13(4): 454-457.
[23] Williams C C. Trust diffusion: The effect of interpersonal trust on structure, function, and organizational transparency[J]. Business & Society, 2005, 44(3): 357-368.
[24] 王娟, 叶斌. "负责任"的算法透明度——人工智能时代传媒伦理建构的趋向[J]. 自然辩证法研究, 2020, 36(12): 66-72.
[25] Veljko D. Toward implementing the ADC model of moral judgment in autonomous vehicles[J]. Science and Engineering Ethics, 2020, 26(5): 2461-2472.
[26] 苏宇. 论算法规制的价值目标与机制设计[J]. 自然辩证法通讯, 2019, 41(10): 8-15.

On ethical decision-making in algorithm design for autonomous vehicles——Based on meaningful human control

LI Dexin[1], GONG Zhichao[2]

1. Research Center for Philosophy of Science and Technology, Shanxi University, Taiyuan 030006, China
2. School of Philosophy and Sociology, Shanxi University, Taiyuan 030006, China

Abstract Based on "meaningful human control", which is the core concept of artificial intelligence ethics, the ethical difficulties in the algorithm design stage for autonomous driving are summarized in the article. The feasibility of "meaningful human control" applied to autonomous vehicles is analyzed. On this basis, using the two conditions of "tracking" and "tracing", a "meaningful human control" framework is constructed around "accountability and transparency" and "value-sensitive design", which may provide systematic methodology guidance for algorithm design of autonomous vehicles.

Keywords autonomous vehicles; ethical decision-making; trolley problem; meaningful human control; value-sensitive design

（责任编辑　卫夏雯）