智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

符合伦理的人工智能应用的价值敏感设计:现状与展望

古天龙， 马露， 李龙， 闫茹

引用本文:
古天龙, 马露, 李龙, 等. 符合伦理的人工智能应用的价值敏感设计:现状与展望[J]. 智能系统学报, 2022, 17(1): 2–15.
GU Tianlong, MA Lu, LI Long, et al. Value sensitive design of ethical-aligned AI applications: current situation and prospect[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(1): 2–15.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202105002

DOI: 10.11992/tis.202105002
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20211216.1820.002.html

符合伦理的人工智能应用的价值敏感设计: 现状与展望

古天龙[1,2], 马露[1], 李龙[1,2], 闫茹[1]

(1. 桂林电子科技大学 广西可信软件重点实验室, 广西 桂林 541004; 2. 暨南大学 信息科学技术学院/网络空间安全学院, 广东 广州 510632)

摘 要: 价值敏感设计是一种将伦理嵌入人工智能设计的高效方法, 尤其是其独特的三方方法为人工智能应用提供了规避伦理问题和道德风险的手段, 因此有着广阔的应用前景。本文系统地介绍了价值敏感设计的内涵、原理和方法; 详细描述了价值敏感设计在国内外的研究现状, 并对典型研究内容进行了对比分析; 总结了近年来价值敏感设计在智能机器人、智能运载工具等人工智能领域取得的研究成果, 并探讨了已有研究的优缺点; 最后对后续研究中亟待解决的问题和挑战进行了分析和讨论, 并对未来研究方向进行了展望。

关键词: 人工智能; 人工智能应用; 符合伦理设计; 价值敏感设计; 人类价值; 设计方法; 利益相关者; 三方方法
中图分类号: TP18 **文献标志码:** A **文章编号:** 1673−4785(2022)01−0002−14

中文引用格式: 古天龙, 马露, 李龙, 等. 符合伦理的人工智能应用的价值敏感设计: 现状与展望 [J]. 智能系统学报, 2022, 17(1): 2–15.
英文引用格式: GU Tianlong, MA Lu, LI Long, et al. Value sensitive design of ethical-aligned AI applications: current situation and prospect[J]. CAAI transactions on intelligent systems, 2022, 17(1): 2–15.

Value sensitive design of ethical-aligned AI applications: current situation and prospect

GU Tianlong[1,2], MA Lu[1], LI Long[1,2], YAN Ru[1]

(1. Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China; 2. College of Information Science and Technology/ College of Cyber Security, Ji'nan University, Guangzhou 510632, China)

Abstract: Value sensitive design, as a high-efficiency way to embed ethics into artificial intelligence design, provides a means to avoid ethical problems and moral hazards for the application of artificial intelligence by using its unique tripartite methodology, having broad application prospects. In this paper, the connotation, principle, and method of value sensitive design are introduced systematically. Then the research of value sensitive design at home and abroad is described in detail, and comparative analysis of its typical research components is carried out. And further, this paper summarizes the achievements of value sensitive design in artificial intelligence fields in recent years, such as intelligent robots and intelligent vehicles, and discusses the advantages and disadvantages of the existing studies, and finally, gives discussion and analysis on the current problems as well as challenges and the future research directions.

Keywords: artificial intelligence; artificial intelligence application; ethical-aligned design; value sensitive design; human value; design method; stakeholders; tripartite methodology

近年来, 随着核心算法的突破、计算能力的大幅提高和海量数据的支撑, 人工智能得以迅速发展, 并推动着社会各领域朝着更智能化的方向加速跃升。然而, 人工智能在应用过程中存在着“不确定性”, 这种“不确定性”不仅来自于人工智能自身的技术黑箱, 也来自于人们对法律和伦理道德的忽视, 这引发了如隐私保护、信息安全、公平正义、人类健康、责任区分等方面的伦理问题。为解决上述问题, 在研发人工智能应用设计及使用过程中, 人们逐渐开始探索并使用技术方法来嵌入人类价值, 如参与式设计[1]、以用户为中心的设计[2]、通用设计[3]、包容性设计[4]、价值敏感设计[5](value sensitive design, VSD)等。但是, 前4种方法倾向于重视工具性和功能性价值, 如用

收稿日期: 2021−05−02. 网络出版日期: 2021−12−17.
基金项目: 国家自然科学基金项目 (62172350, U1811264, 61966009, 61961007); 广西省自然科学基金项目 (2019GXNSFBA245049).
通信作者: 李龙. E-mail: lilong@guet.edu.cn.

户友好性和可用性, 在设计符合伦理的人工智能应用方面的使用范围有限; 相比之下, 价值敏感设计不仅关注工具性和功能性价值, 而且着重强调设计中的伦理价值, 如知情同意、信任、公平性等。同时, 价值敏感设计还将人类价值放在技术人工物所处的具体环境中进行考量, 并关注现有技术是如何支持或阻碍人类价值的, 以便能在设计过程中系统地、彻底地考虑人类价值和社会影响。鉴于拥有以上优势特征, 价值敏感设计在人工智能应用的设计中具有独特的优势, 受到更多研究者和设计师的青睐。

目前, 价值敏感设计在如智能机器人、智能运载设备、可穿戴设备、虚拟现实等人工智能领域的伦理研究中广受关注。此外可以预见, 符合伦理的人工智能应用的价值敏感设计是未来很长一段时间的重要研究课题。因此本文拟对本领域内已有研究进行梳理, 并对下一步的研究重点及存在的挑战进行展望, 以期为推动该方面的理论研究和实践应用提供参考。

1　价值敏感设计概述

在价值敏感设计提出之前, 技术研究与开发中涉及的伦理道德已经引发了学者们广泛的讨论, 但并未出现一个总体的理论与方法来处理技术中的价值维度。在这样一种情境下, Batya Friedman 教授等于 20 世纪 90 年代提出了价值敏感设计的概念, 通过在信息系统中考虑人类价值, 为研究者提供了一种有效协调技术与伦理间关系的方法。随着智能信息服务和应用技术在人们生活中的不断渗透, 价值敏感设计受到了更加广泛的关注, 成为研究者考量技术设计中价值维度的重要方法之一。

1.1　基本内涵

价值敏感设计被定义为“一种以价值理论为基础的设计方法, 强调在整个设计过程中以一种有原则的、全面的方式考虑人类价值”[5]。从此种意义上讲, 符合伦理的技术设计不仅仅是一种前瞻性行为, 而且应该贯穿于设计的整个过程, 以有效嵌入人类的道德价值。

价值敏感设计认为技术不是价值中立的, 技术体现着开发人员的价值观, 而且技术人工物会对人类生活的环境产生影响, 包括道德、政治等多个层面。也就是说, 研究人员在设计过程做出的决策具有价值含义[6]。因此, 在设计前充分考虑人类价值极其重要。这里的“价值”是一个广泛的术语, 指的是一个人或一群人认为在生活中重要的东西[5]。就价值敏感设计关注的具体伦理价值而言, 它涵盖了人类福利、所有权和财产、隐私、无偏见、普遍可用性、信任、自主性、知情同意和问责制等; 也涉及在使用过程中侧重用户使用体验的实用性价值(如系统操作的简易化)、公约(如标准化协议)和个人品味(如图形化用户界面中的颜色偏好)等。

价值敏感设计的主要特点可以概述为 3 点: 互动理论、考虑直接和间接利益相关者、三方方法论[7]。1) 互动理论认为价值观产生于技术与用户的交互中, 其中技术作为塑造社会的要素, 需要遵守一定的伦理规范; 人的行为和社会力量会推动技术的更新迭代, 与此同时, 新技术也改变着人的行为与整个社会体系。2) 识别直接和间接利益相关者以及他们的价值观是价值敏感设计的重要组成部分, 其中, 直接利益相关者指的是直接与计算机系统交互的个人或组织, 间接利益相关者是指间接受系统影响的其他方。3) 价值敏感设计为研究者们提供了一套独特的三方方法论, 即概念调查、经验调查和技术调查。这 3 个阶段之间的相互作用是动态的, 某一阶段中进行的设计更改会影响其他两个阶段。例如, 技术解决方案的变更可能会导致新的社会风险出现, 或者利益相关方的新的经验和价值输入可能需要技术设计做出相应的调整。因此, 为实现产品的优化设计, 概念调查、经验调查和技术调查的执行过程需要不断重复和迭代, 如图 1 所示。

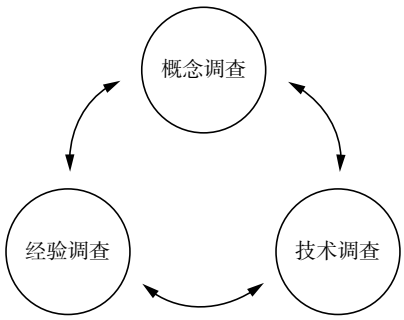

图 1　三方方法论及其迭代

Fig. 1　Tripartite methodology and its iterations

1.2　三方方法论

1.2.1　概念调查

概念调查是价值敏感设计中价值的概念化过程, 这一阶段包含对所调查的问题和结构进行分析性和哲学性的探索, 主要回答以下关键问题: 1)谁是受其影响的直接和间接利益相关者? 2)利益相关者与系统的关系如何? 3)所研究的问题和结构中包含哪些价值? 4)哲学类文献是如何定义某些特定的价值(如信任、知情同意、隐私)的? 5)价值之间有无优劣之分? 6)如何处理价值之

间的冲突(如控制与自主,隐私与获取)？其中,识别间接利益相关者的难度较大,需要设计者做好充分的调研。概念调查将可能的价值定义为适合目标环境的新概念,并纳入了一系列标准的社会科学研究方法,如半结构化访谈、调查、观察、实验设计、探索性调查和纵向案例研究[8]。

1.2.2　经验调查

经验调查通常在概念调查之后进行,以概念调查的研究结果为基础,主要考察技术产品所处的人类环境,并将其量化,为概念调查提供相应的数据支持。经验调查需要借助社会科学研究中使用的定量和定性方法,如:观察、访谈、调查、实验操作、文件收集、用户行为和人体生理特征的测量等。这一阶段主要关注两方面内容:1) 利益相关者在互动环境中对价值观的理解和权衡;2) 设计过程中考虑的伦理价值与人们实际使用中需求的伦理价值之间是否有差距。经验调查的结果不仅可以被用来改进新技术的设计,使之更符合利益相关者的价值观。此外,经验调查还可以放在概念调查之前进行,以帮助研究者在概念调查阶段明确价值的具体定义。

1.2.3　技术调查

与概念调查和经验调查将重点放在利益相关者以及他们所关注的价值上不同,技术调查更具体地指向所讨论技术的设计,通过对现有技术的回顾性分析以及对新技术的前瞻性分析来权衡多个价值,以提供不同的价值适应性。技术调查主要有两种形式,第一种形式侧重于对现有的技术属性和基本机制进行调查,关注技术是如何支持或阻碍人类价值的。例如,一些基于视频的协同工作系统提供模糊的办公室视图,另一些系统则提供清晰的图像,显示关于谁在场以及他们在做什么的详细信息。这两种系统的不同源于它们对两种价值的权衡差异:1)个人隐私;2)群体中成员的存在与活动。第二种形式的技术调查涉及系统的主动设计,以支持概念和经验调查中所发现的价值。例如,Xu[9] 通过概念调查发现采用隐私保护机制有助于保护用户的在线隐私,因此在技术调查阶段为浏览器设计了 3 种隐私增强工具,使用户能对其个人信息进行控制。

尽管三方方法之间的迭代十分重要,但 Winkler 等[10] 调查了 1996—2016 年超过 219 项价值敏感设计的相关研究,发现只有 17 篇论文使用了所有 3 种调查方法。其中,进行了三方方法迭代分析的只有 4 篇文献,仅占研究总量的 1.8%,这提醒学者们在未来的研究中需要充分加强三方方法的迭代使用。为使研究者们更清晰地了解这 3 种方法的任务、特征和不足,本文给出了三方方法的对比分析,如表 1 所示。总体来看,价值敏感设计的三方方法论为研究人员和设计者们进行人工智能伦理研究提供了针对性的指导,也降低了新手使用价值敏感设计的难度。

表 1　三方方法对比分析

Table 1　Contrastive analysis of tripartite methods

三方方法	主要任务	主要特征	不足之处
概念调查	1)设定研究问题的背景 2)识别价值及价值冲突 3)识别直接和间接利益相关者	1)对价值进行哲学性探索 2)重视间接利益相关者的价值 3)关注潜在的价值冲突	1)缺乏识别利益相关者的明确方法 2)由谁来制定价值列表备受争议 3)无法判断价值列表的合理性
经验调查	1)识别利益相关者对互动语境下个人价值的认知 2)识别预期和实际操作的差异	1)关注实际场景 2)关注利益相关者的认知	1)对于价值权重缺乏统一的衡量标准 2)尚未明确利益相关者意见分歧时的解决办法
技术调查	1)对现有技术进行回顾性分析 2)对新技术进行前瞻性分析 3)识别技术如何阻碍或支持某些价值	1)关注技术本身 2)将价值映射到技术中	1)尚未明确需评估的具体技术 2)尚未关联到具体的技术执行环节

2　价值敏感设计研究

国外对于价值敏感设计的研究主要集中在 4 个方面,包括理论反思、框架补充、方法扩展以及工具增添;国内则主要以价值敏感设计理论的引进与推介为主。

2.1　理论反思

价值敏感设计作为技术伦理研究的一种新形式,促进着“技术实践”与“伦理实践”的统一,这个过程中需要明确的伦理理论为解决道德争论、应对道德考虑提供支撑,学者们对此进行了针对性的研究。

Yetim[11] 从话语伦理的角度出发,在价值敏感设计中纳入边界批评方法,并应用不同类型的话语、原则和话语支持方法来补充价值敏感设计。Jacobs 等[12] 提议将中层伦理理论与价值敏感设计

相结合, 以解决价值敏感设计面临的挑战: 1)研究者的主张和价值观不够明确; 2)过分依赖对普世价值的实证研究, 存在犯自然主义谬误的风险; 3)在价值冲突产生时难以进行价值权衡。新设计的伦理理论需要满足 3 个条件: 1)拥有较强的解释力; 2)拥有较强的证明力; 3)简单实用。基于上述研究, Jacobs[13] 将 Nussbaum 所提出的实质性能力理论用于支持价值敏感设计, 实质性能力理论认为, 每个人都应获得 10 种核心能力: 1)能够正常生活; 2)身体健康; 3)保持身体完整; 4)能够运用感官、想象和思考; 5)有情感和情感依恋; 6)拥有实践理性来形成一种善的概念; 7)拥有有意义和受尊重的社会关系; 8)对其他物种表示关心; 9)会玩; 10)控制自己的物质和政治环境。这 10 种核心能力可以为道德主张和考虑提供正当理由和论证来源, 帮助价值敏感设计明确研究者的主张以及应对自然主义谬误的挑战。Umbrello[14] 首次从道德想象理论的角度探讨价值敏感设计, 充分考虑诸如转喻、隐喻、想象和叙事等对道德思考和发展至关重要的想象结构, 并在概念调查中增加了想象工具, 重新整合价值观的概念化过程。

2.2　框架补充

在确定价值和利益相关者、将价值转化为设计等关键阶段中, 需要更具体的框架对其进行指导, 因此 Van [15] 提出了“价值层次”的框架, 如图 2 所示。其中, 规范是指技术为了支持价值而应展示的属性或能力, 且规范通常不止一个。该框架能将价值转化为具体的设计要求, 明确价值判断。基于上述研究, Aizenberg 等[16] 提出了人权设计框架, 如图 3 所示。该框架将设计过程建立在 4 种价值观的基础上, 同时还强调利益相关者的参与和三方方法的迭代。其中, 4 种基本价值观可以进一步细化为更详尽的价值观, 该框架为制定设计要求提供了结构化和包容性的途径。

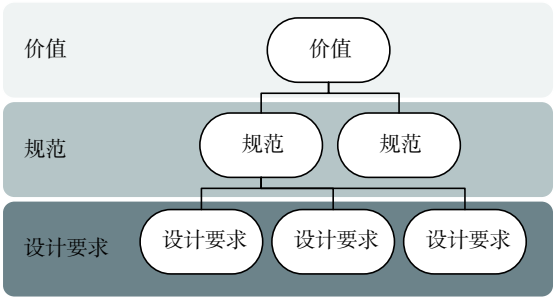

图 2　价值层次
Fig. 2　Values hierarchies

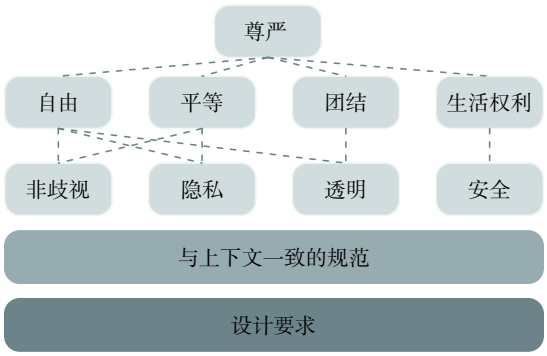

图 3　人权设计框架
Fig. 3　Human rights design framework

为解决新兴技术的本体不确定性问题, de Reuver[17] 提出了 TCRC 框架, 即在价值敏感设计原有的三方方法中加入反射性活动, 用于反思技术开发的目标和价值。这里的反射性是通过“道德沙盒”来实现的, 以识别技术中的道德因素。为检验人们对技术的解释关系, 并从中识别相应的价值, Grünloh[18] 引入了技术框架, 该框架包含 3 个核心部分: 1)技术的本质; 2)技术战略; 3)技术的使用。为支持设计师有意识地、明确地将人类价值观嵌入设计过程, 克服价值敏感设计所提供价值清单的不完整和不通用性, Kheirandish 等[19] 通过经验调查收集了价值观数据, 并对价值观进行分组和聚类, 最终形成了一个全面的价值框架, 该框架包含 4 个主题、9 个价值组、42 个关键价值和 135 个额外价值, 有助于明确与各种设计项目相关的人类价值。

2.3　方法扩展

三方方法论为研究者们提供了一定的指导, 但研究者们面临的设计挑战还需要更明确的实践方法和前进道路。Friedman 等[20] 系统总结了 2017 年以前的 14 种主要价值敏感设计方法的原理和相关基础研究, 并阐述了各种方法的使用技巧, 所提到的 14 种方法: 1)直接和间接利益相关者分析; 2)价值来源分析; 3)技术与社会结构的共同演进; 4)价值情景; 5)价值观的信息维度; 8)面向价值的编码手册; 9)面向价值的模型、原型或现场部署; 10)关于价值观和技术的民族志调查; 11)网上知情同意模式; 12)重视价值坝和价值流; 13)价值敏感行为反射模型; 14)设想卡片。为解决价值敏感设计方法的松散性问题, Greef 等[21] 将情境认知工程(the situated cognitive engineering, sCE)与价值敏感设计方法相结合, 提出了 sCEthics 方法。为解决价值本身的动态变化问题, Poel[22] 提出了价值变化分类法。为衡量不同价值的权重, Van 等[23] 引入了最佳最差方法。

为说明伦理和社会价值如何在连续迭代中体现在具体的技术上，Cawthorne 等[24]引入了基于能力的程序–审议伦理方法。上述 4 种方法的主要特点和应用效果如表 2 所示。

表 2　新兴方法间的对比分析

Table 2　Comparative analysis of emerging methods

方法	主要特点	应用效果
sCEthics[21]	1)添加政策信息(如立法文件与政策指南) 2)提供关于受访者需求的全局概览 3)要求利益相关方就价值的定义达成共识 4)用设计模式为受访者的需求提出解决方案	1)提高了提取价值观的准确性 2)伦理价值的使用更具一致性 3)缩小了想象与实际设计间的差距 4)使利益相关者快速了解各模块间的关系
价值变化分类法[22]	在设计中纳入价值的适应性、灵活性和稳健性这3种技术特征	更好地处理了价值变化
最佳最差方法[23]	根据专家意见对价值进行排序	更有效地确定价值权重
基于能力的程序–审议伦理方法[24]	1)在设计过程中采用参与式协商方法 2)新建立偏功能性的价值观念策略 3)寻求更多VSD专家及设计人员的参与	1)增强了价值与技术在执行上的一致性 2)避免了预先定义的伦理概念框架在使用上的僵化 3)纳入了更多利益相关者，提高了对语言特殊性、多元价值观以及设计规范的关注度

2.4　工具增添

有形的工具能在实践中帮助研究者更好地调查利益相关者及其价值，充分调动参与者的积极性。为促进利益相关者和设计师的沟通，Pommeranz 等[25]提出了价值获取工具，该工具由两个部分组成：1)由利益相关者用于数据收集和现场自我反思的移动应用程序；2)由设计者和利益相关者合作使用的用于分析和交流价值的网站。为更全面地识别利益相关者，Yoo[26]提出了利益相关令牌法，该方法包含 5 个步骤：1)选择参与者；2)选择令牌的材料；3)创建利益相关者列表和标签；4)将标签附加到令牌上；5)简述利益相关者的关系。为更好地识别潜在价值，Logler 等[27]提出了隐喻卡的方法，使用该工具需要 4 个步骤：1)深入熟悉项目领域；2)形成一套生成隐喻；3)制作隐喻卡片；4)将隐喻卡引入设计研究中。为将现有的文化模式融入价值获取过程，以克服文化限制对价值获取造成的障碍，Alshehri 等[28]提出了情景共同创作卡的方法，使用该工具需要 3 个步骤：1)概念设计，即对卡片进行粗略分类；2)物理设计，即为每张卡片设计基于特定文化的图像；3)实用设计，即在半结构化访谈中使用创作卡，通过对不同卡片的组合来创造不同的价值情景，促进参与者表达个人价值观。为促进设计师在设计过程中考虑人类价值，Kheirandish 等[29]提出了 HuValue 工具，该工具包括：1 个价值轮、45 个价值、207 张图片卡。上述 5 种工具的主要特点和应用效果如表 3 所示。

表 3　新工具间的对比分析

Table 3　Comparative analysis of new tools

工具	主要特点	应用效果
价值获取工具[25]	1)采用照片启发法 2)使利益相关者在移动应用程序上记录特定情境下的价值信息 3)价值信息被上传到特定网站，设计师可在该网站与利益相关者进行联系	1)使利益相关者在日常生活中反思自己价值观的正确性 2)促进了利益相关者和设计师的沟通
利益相关令牌[26]	1)以有形工具代表利益相关者，并使利益相关者以角色扮演的方式进行价值敏感分析 2)对利益相关者采取多元化选择方式 3)以标签的方式创建一个整体利益相关网络	1)提高了利益相关者参与设计过程的积极性，并更深层次理解研究背景 2)纳入了先前研究中被忽视的利益相关者 3)更好地理解利益相关者和他们之间的动态关系

续表 3

工具	主要特点	应用效果
隐喻卡[27]	1)鼓励设计师和利益相关者使用隐喻思维 2)让参与者接触隐喻这种抽象概念	1)可更准确识别某些潜在价值 2)实现了将参与者的隐喻体验从描述性转变为生成性
情景共同创作卡[28]	1)在设计中考虑文化因素并使用图像来帮助参与者表达自己的价值观 2)通过组合不同的卡片形成不同的价值场景, 使创作卡的运用具有“游戏性”	1)使参与者能够通过卡片来进行不同价值概念之间的比较和联系 2)减少了参与者的压力, 促进参与者的独特表达及情感体验 3)增加了参与者对讨论的参与度
HuValue工具[29]	1)在价值轮中, 任何人都可以对价值进行重要性排序 2)任何人都可以在每张图片卡底部填写自己的价值认知	1)缩小了人类价值的抽象层面和实际设计层面之间的差距 2)促进设计师思考人类价值的多样性 3)增加了嵌入设计概念中的价值多样性

2.5　国内价值敏感设计研究

国内对于价值敏感设计的研究起步较晚, 目前主要以理论的引进与推介为主。刘瑞琳等[30]系统梳理了价值敏感设计, 介绍了价值敏感设计的研究历程、哲学意蕴以及研究方法, 并讨论了价值敏感设计的实践性、先进性以及局限性。刘宝杰[31]介绍了价值敏感设计的内涵、发展历程与特点, 并以浏览器中的 Cookies 为例, 展示了价值敏感设计三方方法的迭代运用, 此外, 还叙述了荷兰学者对价值敏感设计应用领域的扩展以及当前价值敏感价值设计的不足之处。郭延龙等[32]梳理了智能人工物技术治理的 3 种路径, 包括内在关系进路、混合式系统进路和价值敏感设计进路, 并对价值敏感设计的定义、方法、应用与不足进行了介绍。杜严勇[33]指出, 目前机器人伦理设计的实践探索主要涉及计算机科学进路、认知科学进路和价值敏感设计 3 个方面, 并对以护理为中心的价值敏感设计进路进行了介绍。闫坤如[34]指出, 价值论转向之后的技术哲学发展更加关注技术设计中的伦理与价值, 并介绍了价值敏感设计的理念与应用。孙福海等[35]认为, 要开发出具有道德的人工智能, 系统设计必须考虑道德因素, 并介绍了三种重要方法: 价值敏感设计、透明度设计及道德学习设计, 梳理了价值敏感设计的内涵、原则、研究内容及研究方法。

张贵红[36]认为, 价值敏感设计体现出了当前的大数据技术所具有的伦理与价值维度, 能为大数据伦理提供基础的分析策略, 而且价值敏感设计本身的发展也能推动大数据伦理学与大数据技术的研究与发展。尹文娟[37]指出, 价值敏感设计提供了一种理论支持, 能将利益相关者多元化的价值观嵌入工程设计的过程中。刘培等[38]建议在解决算法伦理问题时考虑使用价值敏感设计方法, 将算法的伦理设计与具体的情景分析相结合。李飞翔[39]指出, 算法工程师和科研工作者群体在对待大数据技术时, 需注重价值敏感设计的使用, 以避免算法过于抽象化。朴毅等[40]认为, 在应对人工智能的价值非中立性伦理问题与决策风险时, 可以采用价值敏感设计方法, 将可预见的偏见问题扼杀在算法开发阶段。陈炜等[41]建议, 在解决劝导性技术的伦理问题时, 可以借鉴价值敏感设计方法。他们认为, 这种方法能提供一个交互性视角, 并能更全面地解决价值冲突。

3　人工智能应用的价值敏感设计

价值敏感设计为解决人工智能应用中可能存在的保密性、隐私性、安全性和可信度等伦理问题提供了一种有效的方式, 能够在人工智能应用的各个设计阶段融入人类价值, 使人工智能应用更符合人类伦理。目前, 价值敏感设计在人工智能的应用主要涉及智能机器人、智能运载工具等方面。

3.1　智能机器人

随着“机器人口爆炸”时代的到来, 智能机器人在人们的生活中扮演着越来越重要的角色, 如情感陪护、聊天、送药等。但智能机器人在决策、行动方面被赋予一定的自主性, 这可能会导致大量的伦理问题出现, 例如: 聊天机器人可能会在言语中透露种族歧视[42]、看护机器人可能会减少人与人之间的亲密接触[43]、军事机器人可能会误伤平民[44]等, 如何设计出符合伦理的智能机器人成为了学者们的研究热点。Cheon 等[45]利用价值

敏感设计方法，研究了机器人和自动化协会网站对人形机器人专家的采访，选择了其中 27 个访谈来定性分析机器人专家的价值观。研究表明，机器人专家所关心的伦理价值包括安全、可靠、透明和人类尊严。基于此，该研究提出了一个将伦理价值融入人形机器人设计的议程：1）调查开发者的价值观；2）调查在机器人的定义中应该接受的伦理价值；3）邀请非传统利益相关者参加机器人设计研讨会。但该研究仅仅是对二次数据的加工，未来还需对机器人专家进行针对性的伦理访谈。

在医疗方面，护理机器人可用于临床诊断、病患护理、病房清理等，成为了有效缓解人工护理压力的新手段。由于护理机器人的服务对象为患者群体，尤其是老龄人等弱势人群，护理机器人所面临的伦理问题具有特殊性。因此，如何设计出符合伦理的护理机器人显得尤为重要。Wynsberghe[46] 基于价值敏感设计方法，提出了一个以关怀为中心的护理机器人伦理评估框架，该框架由 5 个部分组成：1）语境；2）实践；3）相关行动者；4）机器人类型；5）道德因素的体现。以识别在机器人设计过程中需关注的伦理价值。使用上述框架的方法被称为以护理为导向的价值敏感设计（care centered value sensitive design, CCVSD）方法，该方法能帮助设计师和伦理学家对护理机器人进行回顾性伦理评估，并对护理机器人进行符合伦理的前瞻性设计。在后续研究中，Wynsberghe[47] 指出，当服务机器人被集成到护理实践中时，也可以使用 CCVSD 方法对其进行评估与设计，并提出了识别护理实践的两个必要条件：1）该实践必须是对另一个实践的需求的响应；2）通过护理者和护理接受者之间的相互作用来满足他人的需求。此后，Wynsberghe 等[48] 以举重和尿检机器人为例，将活动本质方法用于护理机器人的研究，结果显示，活动本质方法能推动 CCVSD 进一步发展：1）为 CCVSD 提供更坚实的哲学基础；2）为机器人的护理活动提供更精细的描述，以扩展护理伦理；3）提供概念工具来缓解不同伦理价值之间的冲突。

Poulsen 等[49] 基于机器伦理和 CCVSD 方法，提出了运动设计中的价值 (values in motion design, VMD) 方法，用于设计符合伦理的护理机器人。在使用 VMD 方法时，首先需要进行价值敏感设计的概念调查，以建立用户价值模型和机器人设计的初始框架；随后进行经验和技术调查，以区分患者的外在价值和内在价值，其中，当护理机器人尊重患者的外在价值时，即可做出符合护理伦理的决策；而当护理机器人确保为患者提供内在价值时，即可做出符合职业伦理的决策。结果显示，使用 VMD 方法后，研究者们可以通过提供良好的护理伦理决策和职业伦理决策，设计出符合伦理的护理机器人。此后，Poulsen 等[50] 提出了一个 Attento 模型来补充 VMD 方法，该模型可以为特定的患者制定外在价值的优先级列表，并在实践中动态地调整列表，提供定制的患者服务。基于上述研究，Poulsen 等[51] 提出了一种注意框架，该框架建议在护理机器人的设计过程中联合使用 CCVSD 方法和有“计算意识”的信息系统，并在运行时根据外在护理价值的优先级作出设计决策。其中，信息系统包含 3 个关键的过程和元素：1）计算意识过程；2）价值优先级列表；3）价值确认过程。

从上述研究来看，近几年智能机器人的伦理设计研究取得了不错的成就，其中，CCVSD 方法成为了护理机器人伦理研究的奠基之作，具有重大的探索价值。但目前仍存在着一些挑战和不足。首先，护理机器人在护理过程中会采集许多患者的隐私数据，如何确保众多隐私数据不被泄露，是否需要针对不同的弱势群体采取不同的隐私保护强度，其他家庭成员是否有权获取关于患者的全方位隐私数据。其次，患者在使用护理机器人的过程中，智能护理与患者的自主性之间常常产生冲突，如何对两种价值进行权衡，如何准确评估患者的实时状况，以决定护理机器人应在何种程度上保障患者的自主意愿。同时，我国属于多民族国家，如何避免护理机器人在外观设计和交互过程中暗含种族歧视和性别歧视，如何在设计中考虑各民族的风俗和宗教信仰。最后，目前对于智能机器人的研究仅限于人形和护理机器人，如何对送药机器人、移动病人机器人、工业机器人、物流机器人等进行评估与设计，是否可以简单套用护理机器人的设计方法。总之，解决上述问题，是当下智能机器人的进一步研究方向。

3.2　智能运载工具

自动驾驶汽车和无人机能参与组建更为高效的交通运输系统，帮助减轻交通压力，但在研发、应用和推广中产生的一系列伦理问题使其发展遇到了阻力，如隐私泄露[52-53]、责任分配[54]、非法监控[55]。为解决上述问题，科研工作者们进行了一系列的研究。

Flipse 等[56] 基于价值敏感设计和中游调节方法，建立了一种合作创新开发方法，方法分为两步：1）为共同参与设计过程者建立价值档案；2）

组织研讨会来讨论前述价值, 并将它们转化为规范和设计要求。在自动驾驶汽车的案例研究中, 建立的价值档案包括 6 个伦理价值和 2 个非伦理价值: 1)道路安全; 2)负责制; 3)群体平等; 4)自主; 5)隐私和财产安全; 6)平等受益; 7)高交通流量; 8)以不同方式打发时间。上述价值可以转化为具体的设计规范, 例如, 道路安全可被转化为减少交通事故的设计规范, 高交通流量可被转化为减少拥堵的设计规范。该方法为设计师开发符合伦理的自动驾驶汽车提供了有益的参考, 但在建立价值档案时, 对于样本中具有技术背景和受过高等教育的人, 可能存在偏见, 后续研究可纳入更广泛的参与者。Thornton 等[57] 使用价值敏感设计方法, 开发了自动驾驶汽车的速度控制器, 以解决人行横道阻塞的情况。在概念调查阶段, 确定了利益相关者和相关伦理价值, 其中, 伦理价值包括关心和尊重他人、公平和互惠、尊重权威、信任和透明度、个人自主权。并将上述价值与当前算法中已考虑的价值联系起来, 包括安全性和合法性、移动性和效率、平滑性。在技术调查阶段, 通过构建马尔科夫决策过程来设计速度控制器, 以支持概念调查确定的伦理价值。在经验调查阶段, 将确定性比例速度控制器作为基线, 与价值敏感设计速度控制器进行比较, 结果显示, 价值敏感设计速度控制器能有效控制车辆的纵向加速度, 让车辆安全地通过堵塞的人行横道。但控制速度时, 乘客舒适度将会受到影响, 关心和尊重他人这一伦理价值未得到保障, 这为下一步的研究提供了方向。

Cawthorne 等[58] 首次将价值敏感设计应用于无人机平台的技术开发, 在回顾性分析中, 以技术调查为重点, 分析了现有的货运无人机原型是如何支持伦理价值的, 并探讨了当前无人机的局限性, 比如: 1)可能存在安全风险; 2)人们担心无人机会携带摄像头以侵犯他们的隐私; 3)可能会损害租船运营商的经济利益, 从而违背“有益”这一伦理原则。在前瞻性分析中, 开发了一种新型伦理无人机, 这种新型无人机由内燃机驱动, 且采用模块化组件, 能大幅度提高医疗服务能力和安全性。同时, 通过在机身上贴上未携带摄像机的标志, 能降低人们对无人机的不信任感。此外, 无人机可以为租船运营商提供设计、制造和维护等工作, 保障当地居民的物质福利。研究显示, 新型无人机能更好地提升人类福利(身体、心理和物质福利)。随后, Cawthorne 等[59]等引入价值层次方法, 在原有的 3 个层次的基础上新增了一个伦理原则层次。并构建了无人机伦理框架, 包括: 1)有益; 2)无害; 3)自主; 4)公正; 5)可解释性。以促进公共医疗保健中使用的无人机的设计、开发、实施和评估。使用该伦理框架时, 伦理原则需要被转化为具体的人类价值, 比如, 有益的伦理原则可转化为人类福利(身体、心理和物质福利)、工作/人类技能和环境可持续性的价值, 无害的伦理原则可转化为隐私、安全等价值。上述伦理框架被用于 Cawthorne 等[60] 的另一个研究中, 该研究首先评估了丹麦一架商用无人机所面临的伦理挑战, 随后使用该伦理框架重新设计了一架“节俭无人机”, 以提供价格低廉的血液运输方式。这种新型无人机在设计中支持了 5 项伦理原则及其相应的价值观和相关规范, 比如: 无人机使用小型固定翼, 以降低成本; 无人机行驶较为缓慢, 而且重量轻, 能尽量减少伤害等。下一步可以分析更全面的数据, 以提供更完整的无人机分析。

综上所述, 价值敏感设计为智能运载工具的伦理设计提供了一套系统的指导方法, 但目前仍存在以下挑战: 1) 对于自动驾驶汽车来说, 由于其本身存在着诸多伦理困境, 如著名的“电车难题”, 那么是否可以引入价值敏感设计的方法为解决这一难题提供新的思路。为了实现车辆分流, 自动驾驶汽车可能会选择一条更远的路(而非乘客选择的路线), 在这种情况下, 如何保障乘客的自主性。为了保障车内乘客的安全, 自动驾驶汽车可能会收集间接利益相关方(如其他车内的乘客、行人以及车辆)的信息, 那么哪些信息是自动驾驶汽车能够收集的, 谁有权获取这些信息, 如何保护间接利益相关者的隐私。2) 对于无人机来说, 由于其在飞行过程中可能会惊扰鸟类, 如何保障非人类利益相关者的权益。如何避免军用无人机误伤平民(包括儿童、妇女)。商业航空飞机能让乘客在登机时主动接受风险, 那么无人机要如何才能保障地面人员的知情同意权。无人机在飞行过程中可能会干扰其他航班, 致使其他航班备降、返航或延误, 如何避免这一问题的发生? 民用无人机越来越多, 如何保障人们的隐私权不受侵害。解决上述问题, 是当下智能运载工具的进一步研究方向。

3.3　其他人工智能技术

除上述研究外, 价值敏感设计还被应用于可穿戴技术、学习分析技术、身份识别技术等人工智能技术的设计中。Christodoulidou 等[61] 以谷歌眼镜为研究对象, 采用价值敏感设计和手段–目

的链方法，探讨了高等教育背景下可穿戴眼镜的6个用户价值：保证、自主、交流、效率、学习能力、技术功能。为可穿戴眼镜提供了设计启示。研究显示，充分考虑大学生价值的谷歌眼镜能有效提高课堂上的小组讨论与互动效率。Chen 等[62]展示了将价值敏感设计用于学习分析设计的两项研究，第一项研究是对一种学习分析工具的概念调查，通过利益相关者分析和价值分析，揭示了对该工具未来的改进设计有影响的价值（包括伦理价值和非伦理价值）和价值冲突，如：1）学生自主性和工具评估效用；2）隐私；3）学生成功；4）责任；5）可用性；6）社会福祉；7）无偏见等。第二项研究涉及维基百科项目中推荐算法的设计，包含5个步骤：1）文献分析和实证调查；2）原型开发；3）社区参与；4）迭代测试和改进；5）评估算法。以上两个研究显示，价值敏感设计方法可用来支持学习分析系统中的伦理考虑和人类价值观，未来还需探索算法审计、设计优先权、道德设计批判和价值敏感设计之间的协同作用，以全面考虑学习分析中的人类价值。Briggs 等[63]从包容型设计和价值敏感设计的视角出发，探讨了身份识别技术在设计过程中需要考虑的价值。该研究包含两个阶段：1）收集身份管理场景，如智能手机人脸识别系统、智能纹身、指纹识别系统等；2）与6个边缘化人群共同举办研讨会，包括年轻人、老年人、难民、黑人少数民族妇女、残疾人和心理健康服务使用者，以识别他们的价值。该研究呼应了赫茨伯格的双因素理论，其中，卫生因素中包含的伦理价值：数据隔离、数据完整性、数据访问、信任、可靠性、安全性、知情同意、多样性和排斥性。激励因素中包含的价值：方便、个性化、美学。未来可以纳入更多边缘化群体的价值观，以便更细致地考虑不同群体的价值需求。

Bastian 等[64]将价值敏感设计应用于新闻推荐系统算法（algorithmic news recommenders，ANRs），通过对来自荷兰和瑞士两家优质报纸的17名媒体从业者的半结构化采访，探讨了以组织为中心和以受众为中心的新闻价值观与 ANRs 之间的相互作用。研究表明，受访从业者认为新闻价值观在设计和实施 ANRs 过程中非常重要，这些价值观包括透明度、多样性、编辑自主权、广泛的信息提供、个人相关性、可用性和惊喜。此外，受访者对特定价值观的评估因环境而异，需要结合特定的新闻渠道来考虑。未来可以寻找新的算法设计操作方法，对不同价值之间复杂的相互关系进行审查，并观察价值观在不同类型媒体（如公共服务媒体）和背景中的感知作用。Helbing 等[65]指出，在智慧城市的设计和发展过程中，需要“为价值而设计”，并主张使用价值敏感设计方法，以分布式和参与式的方式来组织数字世界，设计价值敏感型智慧城市。这要求设计师们关注除效率和经济增长之外的多种价值观，如环境条件与健康、安全与保障、人的尊严、福祉与幸福、隐私与自决，并赋予公民数据主权，保证“设计民主”。未来还可以考虑智慧城市中的更多伦理问题，学习生物和生态系统的组织原则，以参与式的方式组织数字世界。Stone[66]将城市描述为开放的、不断发展的系统，并基于夜间照明的案例提出了6项原则，以加强价值敏感设计和价值设计方法在城市技术以及智慧城市的应用。6项原则包括：技术特异性、边界条件、历史背景、象征意义、有价值高于价值、放弃概念上的完整性。该原则为实现价值敏感的城市技术创新提供了出发点，但其主要集中于对道德价值观和城市技术之间关系的概念分析，未来还可以往经验和技术层面继续迭代、修改与扩展。

从上述研究来看，对于可穿戴眼镜和学习分析系统这类应用于教育领域的人工智能来说，目前还存在如下挑战：1）在人工智能教育应用的过程中，隐私泄露所导致的网络诈骗、恶意推送等现象颇为常见，如何保护用户的隐私，哪些数据是用户可以自己保有而不会被迫公开让他人知晓的。2）人工智能技术有利于捕捉用户行为和偏好，能对用户的需求进行精准推荐，这可能会使用户在一定程度上对智能技术过度依赖，使学生的视野受到限制，使教师丧失对教学的独特思考，如何解决这一冲突。3）如果人工智能出现违反道德规范的行为，应该由谁来负责。如何在教育人工智能技术中体现“透明性”这一伦理原则。对于身份识别技术来说，现在人们的智能手机大多采取指纹识别技术，但指纹信息能使用胶带等工具恶意获取，如何提高身份识别技术的可靠性和安全性。如今人脸识别技术泛滥，大多数人未经许可即被收集人脸图像，如何保障用户的隐私权和知情同意。相对于有色人种和女性，目前大多数人脸识别技术对于白人男性面孔的识别准确率更高，如何在人脸识别技术中避免性别和种族歧视？

对于新闻算法来说，除了在具体的推荐算法层面考虑和整合价值观外，还需要在组织层面对其予以考虑，但组织层面需要收集什么数据？如何使用收集的数据？如何向用户传达新闻机构意

识到收集和处理个人数据的具体责任？此外, 由于 ANRs 的新闻价值很难脱离某一特定新闻渠道单独讨论, 那么如何识别出一般化的价值概念, 从而为更一般的推荐设计提供信息？另一个挑战是, 来自不同背景的从业者往往致力于不同的概念化和客观价值, 如何处理不同从业者之间的价值冲突？对于智慧城市来说, 由于城市的建设涵盖面较广, 如何确定智慧城市先进与否的衡量标准？我国属于多民族国家, 如何在智慧城市的建设中体现出对不同民族人民的关怀？在这个经济快速发展的时代, 如何避免经济原则主导城市伦理原则的情况出现？大多数智慧城市通常是少数几个规划设计人员和技术精英的事, 如何真正做到让普通民众参与城市管理中来？未来可考虑研究上述人工智能伦理问题的解决方案。

4　挑战与展望

人工智能在促进社会进步的同时, 也带来了诸多伦理问题, 如制造过程中产生废料造成环境污染、医疗成本昂贵拉大贫富差距、信息收集造成隐私泄露、机器人陪护造成孩童出现情感偏差和认知障碍等。然而, 在符合伦理的人工智能的设计过程中, 不论是采取“自上而下”方法, 还是“自下而上”方法, 亦或是混合方法都存在着或多或少的弊端和缺陷, 无法使人工智能产品完全遵守伦理规范。价值敏感设计为解决人工智能伦理问题提供了一种新的思路, 通过使用三方方法论及其迭代, 能够彻底地调查出人工智能产品设计过程中需要考虑的伦理价值, 并将其嵌入到产品的设计过程中去。但目前的研究还处于起步阶段, 研究成果也较少, 仍难以完全解决人工智能快速发展所带来的一系列问题, 还有许多问题尚待解决。现列举未来可能的研究重点与方向, 以期推动本领域研究的发展。

1) 加强人工智能应用自身的基本价值研究。

进行人工智能伦理研究的一个核心环节是确定人工智能技术应遵循的伦理价值, 并且使价值列表尽可能详尽。价值敏感设计的三方方法可以对人工智能应用所涉及的伦理价值进行充分的调研, 这 3 种调查方法调查出的价值可以分别总结为概念化价值、经验化价值和技术化价值。从目前的研究来看, 学者们对于技术化价值的调查比较缺失, 也就是说, 涉及人工智能技术特殊性的伦理价值尚未受到关注。同时, 如何体现 3 种不同类型的价值之间的异同, 当前的研究也甚少考虑。此外, 为了使价值列表更全面, 需要在调查中纳入更广泛的利益相关者, 尤其是弱势人群和边缘人群, 因为这类人群在人工智能技术的发展过程中最容易受到伤害。

2) 加强人工智能应用的设计规范研究

在确定伦理价值之后, 就需要将其转化为设计规范, 而这一过程的难点在于价值冲突的处理, 由于价值是模糊且无法量化的, 因此如何考虑不同伦理价值之间的复杂权衡至关重要。需要注意的是, 伦理的构建条件是当下社会普遍接受的思想和行为, 可见伦理价值冲突的解决应该跟随时代的变化而动态地调整。在不同的人工智能应用领域, 设计规范的构建可能会出现不一致的现象, 这就需要对设计规范进行普适性研究, 以节约时间成本和减少资源消耗。另一个重要的问题就是描述粒度, 若能用更精准的术语和概念描述设计规范, 则能让设计师更容易地将伦理价值嵌入人工智能的设计中。

3) 加强人工智能应用的设计实践研究

使用价值敏感设计方法的最终目的是设计出符合伦理的人工智能应用, 真正把理论转化为实践, 而当前的大多数研究还停留在理论层面, 这不仅是因为人工智能自身面临着一些技术性难题, 还因为某种伦理理论或规范本身就是计算机无法实现的, 这些问题需要研究者们从技术层面上去着手解决。此外, 三方方法是需要不断迭代的, 即使已经设计出“新版”的人工智能应用, 依旧可以重复使用三方方法, 对人工智能应用进行动态调整, 以设计出更符合伦理的人工智能应用, 这一点在目前的研究中还甚少考虑。同时, 如何评估上述伦理价值和规范的有效性, 如何评估人工智能应用的行为是否符合伦理, 如何实现人工智能应用的伦理稳定性, 仍是未来需要进一步考虑的研究方向。另外, 在实践过程中, 还要避免掉入过度使用价值敏感设计的陷阱中。例如, 在国家或集体价值与个人价值产生冲突时, 如果过度强调价值敏感设计, 个人价值可能会被放在首位, 而这却不符合我国国家和集体利益至上的社会主义义利观。

5　结束语

为了使人工智能遵从人类道德主体的道德规范和价值体系, 并在法律和道德的规范下充分发挥其特定的功能, 学界一直致力于人工智能伦理研究, 避免设计出不符合人类价值的人工智能产品。价值敏感设计作为一种考虑人类价值的前摄性方法, 能够帮助人工智能工程师将伦理道德嵌

入人工智能产品中，具有广阔的应用前景。本文对价值敏感设计进行了简要介绍，并概述了价值敏感设计应用于人工智能伦理的研究案例和发展趋势。虽然价值敏感设计已经对人工智能应用的伦理研究做出了一定的贡献，推动了人工智能应用朝着符合伦理道德的方向发展，但仍然有必要继续探索价值敏感设计为人工智能伦理研究带来的更多发展契机。从研究现状来看，现有的研究主要考虑理论层面人工智能应用伦理价值的构建，并没有考虑到实践过程中可能会遇到的难题。未来这方面的工作应该聚焦于如何将伦理价值转化为设计规范，进而将价值和规范嵌入具体的人工智能产品设计环节，以设计出符合伦理的人工智能应用。

参考文献：

[1] KUKLA C D, BINDER T, PORTER W L, et al. Innovation in design: strategies for designing together[C]// CHI'99 Extended Abstracts on Human Factors in Computing Systems. Pennsylvania, Pittsburgh, 1999: 108−109.

[2] ABRAS C, MALONEY-KRICHMAR D, PREECE J. User-centered design[M].BAINBRIDGE W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, 2004: 445−456.

[3] IWARSSON S, STÅHL A. Accessibility, usability and universal design-positioning and definition of concepts describing person-environment relationships[J]. Disability and rehabilitation, 2003, 25(2): 57−66.

[4] LAGE M J, PLATT G J, TREGLIA M. Inverting the classroom: a gateway to creating an inclusive learning environment[J]. The journal of economic education, 2000, 31(1): 30−43.

[5] FRIEDMAN B, KAHN JR P H, BORNING A. Value sensitive design and information systems[M].HIMMA K E, TAVANI H T. The Hand-Book of Information and Computer Ethics. Hoboken: John Wiley & Sons, Inc., 2008: 69−101.

[6] MANDERS-HUITS N. What values in design? The challenge of incorporating moral values into design[J]. Science and engineering ethics, 2011, 17(2): 271−287.

[7] BORNING A, MULLER M. Next steps for value sensitive design[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Austin, USA, 2012: 1125−1134.

[8] DAVIS J, NATHAN L P. Value sensitive design: applications, adaptations, and critiques[M].VAN DEN HOVEN J, VERMAAS P E, VAN DE POEL I. Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Dordrecht: Springer, 2015: 11−40.

[9] XU Heng, CROSSLER R E, BÉLANGER F. A value sensitive design investigation of privacy enhancing tools in web browsers[J]. Decision support systems, 2012, 54(1):424−433.

[10] WINKLER T, SPIEKERMANN S. Twenty years of value sensitive design: a review of methodological practices in VSD projects[J]. Ethics and information technology, 2021, 23(1): 17−21.

[11] YETIM F. Bringing discourse ethics to value sensitive design: pathways toward a deliberative future[J]. AIS transactions on human-computer interaction, 2011, 3(2): 133−155.

[12] JACOBS N, HULDTGREN A. Why value sensitive design needs ethical commitments[J]. Ethics and information technology, 2021, 23(1): 23−26.

[13] JACOBS N. Capability sensitive design for health and wellbeing technologies[J]. Science and engineering ethics, 2020, 26(6): 3363−3391.

[14] UMBRELLO S. Imaginative value sensitive design: using moral imagination theory to inform responsible technology design[J]. Science and engineering ethics, 2020, 26(2): 575−595.

[15] VAN DE POEL I. Translating values into design requirements[M].MICHELFELDER D P, MCCARTHY N, GOLDBERG D E. Philosophy and Engineering: Reflections on Practice, Principles and Process. Dordrecht: Springer, 2013: 253−266.

[16] AIZENBERG E, VAN DEN HOVEN J. Designing for human rights in AI[J]. Big data & society, 2020, 7(2): 2053951720949566.

[17] DE REUVER M, VAN WYNSBERGHE A, JANSSEN M, et al. Digital platforms and responsible innovation: expanding value sensitive design to overcome ontological uncertainty[J]. Ethics and information technology, 2020, 22(3): 257−267.

[18] GRÜNLOH C. Using technological frames as an analytic tool in value sensitive design[J]. Ethics and information technology, 2021, 23(1): 53−57.

[19] KHEIRANDISH S, FUNK M, WENSVEEN S, et al. A

comprehensive value framework for design[J]. Technology in society, 2020, 62: 101302.

[20] FRIEDMAN B, HENDRY D G, BORNING A. A survey of value sensitive design methods[J]. Foundations and trends® in human-computer interaction, 2017, 11(2): 63−125.

[21] DE GREEF T, MOHABIR A, VAN DER POEL I, et al. sCEthics: embedding ethical values in cognitive engineering[C]//Proceedings of the 31st European Conference on Cognitive Ergonomics. Toulouse, France, 2013: 4.

[22] VAN DE POEL I. Design for value change[J]. Ethics and information technology, 2021, 23(1): 27−31.

[23] VAN DE KAA G, REZAEI J, TAEBI B, et al. How to weigh values in value sensitive design: a best worst method approach for the case of smart metering[J]. Science and engineering ethics, 2020, 26(1): 475−494.

[24] CENCI A, CAWTHORNE D. Refining value sensitive design: a (capability-based) procedural ethics approach to technological design for well-being[J]. Science and engineering ethics, 2020, 26(5): 2629−2662.

[25] POMMERANZ A, DETWEILER C, WIGGERS P, et al. Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate [J]. Ethics and information technology, 2012, 14(4): 285−303.

[26] YOO D. Stakeholder Tokens: a constructive method for value sensitive design stakeholder analysis[J]. Ethics and information technology, 2021, 23(1): 63−67.

[27] LOGLER N, YOO D, FRIEDMAN B. Metaphor cards: a how-to-guide for making and using a generative metaphorical design toolkit[C]//Proceedings of the 2018 Designing Interactive Systems Conference. Hong Kong, China, 2018: 1373−1386.

[28] ALSHEHRI T, KIRKHAM R, OLIVIER P. Scenario co-creation cards: a culturally sensitive tool for eliciting values[C]//Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA, 2020: 1−14.

[29] KHEIRANDISH S, FUNK M, WENSVEEN S, et al. HuValue: a tool to support design students in considering human values in their design[J]. International journal of technology and design education, 2020, 30(5): 1015−1041.

[30] 刘瑞琳, 陈凡. 技术设计的创新方法与伦理考量——弗里德曼的价值敏感设计方法论述评 [J]. 东北大学学报 (社会科学版), 2014(3): 232−237.
LIU Ruilin, CHEN Fan. Innovative approaches and ethical considerations of technological design − a review on B. friedman's value sensitive design[J]. Ournal of Northeastern University (social science edition), 2014(3): 232-237.

[31] 刘宝杰. 价值敏感设计方法探析 [J]. 自然辩证法通讯, 2015(2): 94−98.
LIU Baojie. Method of value sensitive design[J]. Journal of dialectics of nature, 2015(2): 94−98.

[32] 郭延龙, 汤书昆. 价值与责任: 智能人工物设计中技术治理问题探析 [J]. 自然辩证法研究, 2020(5): 61−66.
GUO Yanlong, TANG Shukun. On the value and responsibility: analysis of technology governance in intelligent artifact design[J]. Studies in dialectics of nature, 2020(5): 61−66.

[33] 杜严勇. 机器人伦理设计进路及其评价 [J]. 哲学动态, 2017(9): 84−91.
DU Yanyong. Robot ethical design approach and its evaluation[J]. Philosophical trends, 2017(9): 84-91.

[34] 闫坤如. 技术哲学的价值论转向透视 [J]. 学术研究, 2019(3): 12−17.
YAN Rukun. The axiology of philosophy of technology turns to perspective[J]. Academic research, 2019(3): 12−17.

[35] 孙福海, 陈思宇, 黄甫全, 等. 道德人工智能: 基础、原则与设计 [J]. 湖南师范大学教育科学学报, 2021(1): 38−46.
SUN Fuhai, CHEN Siyu, HUANG Puquan, et al. Moral artificial intelligence: foundations, principles and design[J]. Journal of Educational Science of Hunan Normal University, 2021(1): 38-46.

[36] 张贵红. 价值敏感设计与大数据伦理 [J]. 伦理学研究, 2019(2): 114−119.
ZHANG Guihong. Value sensitive design and big data ethics[J]. Studies in ethics, 2019(2): 114−119.

[37] 尹文娟. 工程活动中的“利益相关者”: 必要性、缺席与复位研究 [J]. 科学技术哲学研究, 2019(5): 68−73.
YIN Wenjuan. An analysis of "stakeholders" in engineering activities: its necessity, absence and back to the context[J]. Studies in philosophy of science and technology, 2019(5): 68−73.

[38] 刘培, 池忠军. 算法的伦理问题及其解决进路 [J]. 东北大学学报 (社会科学版), 2019, 21(2): 118−125.
LIU Pei, CHI Zhongjun. Ethical issues of algorithms and

their solutions[J]. Journal of Northeastern University (social science edition), 2019, 21(2): 118-125.

[39] 李飞翔. “大数据杀熟”背后的伦理审思, 治理与启示 [J]. 东北大学学报 (社会科学版), 2020, 22(1): 7−15.
LI Feixiang. Ethical reflections, governance and implications under the background of big data price discrimination[J]. Journal of Northeastern University (social science edition), 2020, 22(1): 7−15.

[40] 朴毅, 叶斌, 徐飞. 从算法分析看人工智能的价值非中立性及其应对 [J]. 科技管理研究, 2020, 40(24): 245−251.
PU Yi, YE Bin, XU Fei. The value non-neutrality of artificial intelligence and countermeasure from the perspective of algorithm analysis[J]. Science and technology management research, 2020, 40(24): 245−251.

[41] 陈炜, 刘郦. 劝导技术道德化实践探索 [J]. 自然辩证法研究, 2020(1): 44−49.
CHEN Wei, LIU Li. Practical exploration of the moralization of persuasion technology[J]. Studies in dialectics of nature, 2020(1): 44-49.

[42] HOWARD A, BORENSTEIN J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity[J]. Science and engineering ethics, 2018, 24(5): 1521−1536.

[43] RANTANEN T, LEHTO P, VUORINEN P, et al. The adoption of care robots in home care − a survey on the attitudes of Finnish home care personnel[J]. Journal of clinical nursing, 2018, 27(9/10): 1846−1859.

[44] BORGES J V. Robots and the military: a strategic view[M].FERREIRA M I A, SEQUEIRA J S, TOKHI M O, et al. A World with Robots: International Conference on Robot Ethics: ICRE 2015. Cham: Springer, 2017: 199−205.

[45] CHEON E, SU N M. Integrating roboticist values into a Value Sensitive Design framework for humanoid robots[C]//11th ACM/IEEE International Conference on Human-Robot Interaction. Christchurch, New Zealand, 2016: 375−382.

[46] VAN WYNSBERGHE A. Designing robots for care: care centered value-sensitive design[J]. Science and engineering ethics, 2013, 19(2): 407−433.

[47] VAN WYNSBERGHE A. Service robots, care ethics, and design[J]. Ethics and information technology, 2016, 18(4): 311−321.

[48] DE SIO F S, VAN WYNSBERGHE A. When should we use care robots? The nature-of-activities approach[J]. Science and engineering ethics, 2016, 22(6): 1745−1760.

[49] POULSEN A, BURMEISTER O K, TIEN D. A new design approach and framework for elderly care robots[C]//Australasian Conference on Information Systems. Sydney, Australia, 2018: 1−12.

[50] POULSEN A, BURMEISTER O K, KREPS D. The ethics of inherent trust in care robots for the elderly[C]// 13th IFIP TC 9 International Conference on Human Choice and Computers. Poznan, Poland, 2018: 314−328.

[51] POULSEN A, BURMEISTER O K. Overcoming carer shortages with care robots: dynamic value trade-offs in run-time[J]. Australasian journal of information systems, 2019, 23: 1−18.

[52] LIN Chao, HE Debiao, KUMAR N, et al. Security and privacy for the internet of drones: challenges and solutions[J]. IEEE communications magazine, 2018, 56(1): 64−69.

[53] NI Jianbing, LIN Xiaodong, SHEN Xuemin. Toward privacy-preserving valet parking in autonomous driving era[J]. IEEE transactions on vehicular technology, 2019, 68(3): 2893−2905.

[54] FAGNANT D J, KOCKELMAN K. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations[J]. Transportation research part A: policy and practice, 2015, 77: 167−181.

[55] DING Guoru, WU Qihui, ZHANG Linyuan, et al. An amateur drone surveillance system based on the cognitive internet of things[J]. IEEE communications magazine, 2018, 56(1): 29−35.

[56] FLIPSE S M, PUYLAERT S. Organizing a collaborative development of technological design requirements using a constructive dialogue on value profiles: a case in automated vehicle development[J]. Science and engineering ethics, 2018, 24(1): 49−72.

[57] THORNTON S M, LEWIS F E, ZHANG V, et al. Value sensitive design for autonomous vehicle motion planning[C]//2018 IEEE Intelligent Vehicles Symposium (IV). Changshu, China, 2018: 1157−1162.

[58] CAWTHORNE D, CENCI A. Value sensitive design of a humanitarian cargo drone[C]//2019 International Conference on Unmanned Aircraft Systems . Atlanta, USA, 2019: 1117−1125.

[59] CAWTHORNE D, ROBBINS-VAN WYNSBERGHE A. An ethical framework for the design, development,

implementation, and assessment of drones used in public healthcare[J]. Science and engineering ethics, 2020, 26(5): 2867−2891.

[60] CAWTHORNE D, ROBBINS-VAN WYNSBERGHE A. From healthDrone to frugalDrone: value-sensitive design of a blood sample transportation drone[C]//2019 IEEE International Symposium on Technology and Society (ISTAS). Medford, USA, 2019: 1−7.

[61] DENG Xuefei, CHRISTODOULIDOU N. Understanding user values of wearable computing[C]//Proceedings of the 36th International Conference on Information Systems. Fort Worth, USA, 2015: 1−10.

[62] CHEN Bodong, ZHU Haiyi. Towards value-sensitive learning analytics design[C]//Proceedings of the 9th International Conference on Learning Analytics & Knowledge. Tempe, USA, 2019: 343−352.

[63] BRIGGS P, THOMAS L. An inclusive, value sensitive design perspective on future identity technologies[J]. ACM transactions on computer-human interaction, 2015, 22(5): 1−28.

[64] BASTIAN M, HELBERGER N, MAKHORTYKH M. Safeguarding the journalistic DNA: attitudes towards the role of professional values in algorithmic news recommender designs[J]. Digital journalism, 2021, 9(6): 835−863.

[65] HELBING D, FANITABASI F, GIANNOTTI F, et al. Ethics of smart cities: towards value-sensitive design and co-evolving city life[J]. Sustainability, 2021, 13(20): 1−25.

[66] STONE T. Design for values and the city[J]. Journal of responsible innovation, 2021: 364–381.

作者简介：

古天龙，教授，博士生导师，主要研究方向为可信人工智能、人工智能伦理、数据治理、形式化方法。获国家教学成果奖二等奖、广西教学成果奖特等奖、广西教学成果一等奖 (第一完成人)。主持国家 863 计划项目、国家自然科学基金重点项目、国防预研重点项目等 30 余项。发表学术论文 300 余篇。

马露，硕士研究生，主要研究方向为人工智能伦理。

李龙，讲师，博士，中国人工智能学会会员，中国计算机学会会员，主要研究方向为人工智能安全和逻辑程序设计。主持国家自然科学基金面上项目子课题、广西省自然科学家基金项目等 6 项。发表学术论文 20 余篇。