

公众人工智能价值敏感性的测度解析

Analysis on Measurement of Value Sensitivity with Artificial Intelligence Among Public

张学义 /ZHANG Xueyi^{1, 2} 周洲 /ZHOU Zhou¹ 郑宇涵 /ZHENG Yuhan¹

(1. 东南大学哲学与科学系, 江苏南京, 211189; 2. 江苏省道德发展智库, 江苏南京, 211189)

(1. Department of Philosophy and Science, Southeast University, Nanjing, Jiangsu, 211189;

2. Jiangsu Province Moral Development Think Tank, Nanjing, Jiangsu, 211189)

摘要:“价值敏感性”源于巴迪亚·弗里德曼的“价值敏感性设计”概念, 指在技术设计、研发、生产和使用中具有伦理意蕴的价值感知能力。课题组创新性地应用“三方方法论”, 通过多指标多因素结构方程模型(MIMIC)测量公众对人工智能的价值敏感性, 评估现有技术是否支持价值敏感设计目标。研究发现:(1) 公众对人工智能的价值敏感排序为: 透明性、算法公正、劳动就业保障、责任归属等, 说明这些价值需在技术设计阶段予以重视;(2) 信息安全和生态失衡未引起广泛关注;(3) 性别和职业对价值敏感性无显著影响, 但年龄和学历差异显著。

关键词: 人工智能 价值敏感性 多指标多因素模型 伦理治理

Abstract: “Value sensitivity” originates from Batya Friedman’s concept of “Value-Sensitive Design”, referring to the ability to perceive ethically relevant values during the design, development, production, and use of technology. The research team innovatively applied the “tripartite methodology” and used the Multiple Indicators and Multiple Causes (MIMIC) structural equation model to measure public value sensitivity toward artificial intelligence (AI) and assess whether the current technologies support value-sensitive design goals. The study found that: (1) Public sensitivity to AI values is ranked as transparency, algorithmic fairness, job security, and responsibility attribution, indicating that these values should be prioritized in the design stage; (2) Information security and ecological imbalance received less attention; (3) Gender and profession showed no significant impact on value sensitivity, but age and education level did show significant differences.

Key Words: Artificial intelligence; Value sensitivity; MIMIC; Ethical governance

中图分类号: TP18; O174.12 DOI: 10.15994/j.1000-0763.2025.01.003 CSTR: 32281.14.jdn.2025.01.003

近年来, 随着人工智能技术的迅猛发展, 其在诸多领域的广泛应用正深刻地改变着人类的生产生活。与此同时, 人工智能技术伴生

的伦理风险亦愈发凸显, 其治理工作尤为迫切。价值敏感性设计主张将正确的价值观嵌入到技术系统之中, 实现人与技术的动态交互。

基金项目: 国家社会科学基金重大项目“负责任的人工智能及其实践的哲学研究”(项目编号: 21&ZD063); 国家社会科学基金一般项目“科学理解的实验哲学研究”(项目编号: 22BZX038); 中央高校基本科研业务费专项资助“人工智能价值敏感性设计的伦理风险及其防范研究”(项目编号: 2242023S20031)。

收稿日期: 2024年4月1日

作者简介: 张学义(1983-)男, 安徽阜阳人, 东南大学哲学与科学系副教授, 江苏省道德发展智库研究员, 研究方向为科学哲学、实验哲学、科技伦理。Email: zxyj0928@126.com

周洲(2003-)女, 江苏南京人, 东南大学哲学与科学系本科生, 研究方向为科学哲学、科技伦理。Email: zoezhou790@gmail.com

郑宇涵(2003-)女, 江苏南京人, 东南大学哲学与科学系本科生, 研究方向为政治哲学、伦理学。Email: zyhdj2021@163.com

越来越多的学者也注意到,公众作为人工智能技术的利益相关者,应该纳入到人工智能治理框架之中,关注其价值需求。课题组聚焦公众对于人工智能技术的价值敏感性程度,对其进行深度解析。该项工作可能的贡献包括:第一,结合人工智能技术的具体情境,对领域内的关键价值进行分解和操作化定义,深化对于人工智能相关价值的理解;第二,通过多指标多因素模型(Multiple Indicators Multiple Causes Model, MIMIC),量化分析公众对于信息安全、人类主体性、价值选择、责任归属、生态环境保护、算法公正、透明性、主体情感平衡、劳动就业保障等九类价值敏感性,考察公众在与人工智能技术互动过程中对各项价值的感知状况,检验现实数据结果与原初设想的差距,从而为具有操作性的技术设计指南提供参考建议。

一、价值敏感性及其测度

“价值敏感性”(Value Sensitivity)一词源自1992年技术哲学家巴迪亚·弗里德曼(Batya Friedman)和彼得·卡恩(Peter H. Kahn)共同发表的“人类主体与责任的计算:对计算机设计的影响”一文^[1]中提到的关于促进负责任计算的设想,而后弗里德曼在其1996年的《价值敏感性设计》对此概念系统阐述。该概念主张在技术的初创阶段就将具有伦理性的价值理念嵌入到设计之中,从而走出一条技术创新与伦理价值相融合的技术实践之路。^{[2], [3]}而“价值敏感性”在这里主要指认知主体在技术设计、研发、生产、使用等过程中具有伦理意蕴的价值感知能力。20世纪70年代以来,随着技术实践的迅速发展,技术哲学领域出现了“设计转向”(Design Turn)思潮,即从专注对基础理论、思想实验的分析转变为对经验层面技术系统设计的关注,旨在通过现实的制度或物质设计来实现价值目标。从历史的角度来看,技术发展与公众价值需求之间的联结日益紧密,从早期阶段的完全不采纳或仅在最低限度上采纳公众价值需求,发展为现阶段的把公众的价值偏好、

社会的公共辩论等都纳入技术系统设计阶段的考虑之中。从技术自身来看,弗里德曼指出,“作为这种人类活动的结果,所有的技术都在某种程度上反映并相互影响着人类的价值观”。^[4]换言之,从设计研发阶段开始,技术就承载了设计者的利益与价值。在这种时代背景之下,价值敏感性设计作为基础性理论诞生并兴起。

价值敏感性设计的主要研究方法被称之为“三方方法论”(Tripartite Methodology),即概念研究(Conceptual Investigation)、经验研究(Empirical Investigation)和技术研究(Technical Investigation)。概念研究指的是从理论层面对于技术应用背景下的价值进行识别与阐释。这里“价值”的内涵是“个人或团体认为重要的东西”。^[5]经验研究则是使用实证分析方法,对技术人工物所处的社会环境进行具体分析。技术研究的重点是技术本身。一方面关注现有的技术特性和潜在机制如何支持或阻碍人的价值,是否满足了实际需求;另一方面对技术系统进行主动设计,旨在使技术符合并支持人们的价值需求。此外,弗里德曼等人在三方方法论的大框架下,还具体提出了17种调查方法,包括直接和间接利益相关者分析、价值情景法、价值草图等等。^[4]

然而,目前国内研究主要集中在从宏观层面探讨价值敏感性设计理论,对于微观层面的实证研究较少。例如,有学者探讨了人工道德智能体价值敏感性设计的必要性和建构方法。^[6]有学者论述了弗里德曼“三方方法论”在人工智能技术设计过程中的应用。^[7]还有学者指出价值敏感性设计可以以一种积极的方式将价值前置,使技术设计从根源上体现价值的诉求。^[8]整体而言,国内关于具体技术层面的经验研究尚较为缺乏。当前,国际研究的主要阵地为华盛顿大学价值敏感设计实验室与荷兰学派3TU技术伦理研究中心,但在人工智能的技术治理和政策实施方面,公众的相关价值需求仍未得到充分重视。^[9]佩德罗·罗伯斯(Pedro Robles)等人的一项研究考察了公众对于人工智能的价值态度与接受程度,但仍存在不足之处,包括数据切口过大,时间滞后等。^[10]因此,

课题组尝试结合三方方法论。在概念研究阶段,综合具体应用情景、使用群体等对人工智能伦理价值进行识别与分析;在经验研究阶段,在问卷调研数据的基础上结合MIMIC模型进行价值敏感性分析,旨在切实建立技术设计者与社会因素的有效链接,减少设计偏差和设计失误。在技术分析阶段,评估现有技术何种程度上支持或阻碍价值敏感设计目标实现,并给出合理建议。

二、研究设计

1. 数据来源

本研究使用的数据源自于课题组2023年进行的“公众的人工智能价值敏感性”的问卷调查。调查采用互联网问卷发放形式与多种抽样方法,包括判断抽样、偶遇抽样、滚雪球抽样和多阶段抽样等,以确保样本的代表性和多样性。最终,课题组收集到1052份有效样本。其中男性占比为49.3%,女性占比为50.7%。同时,受访者的年龄分布、教育背景、职业背景均呈现出多样性。关于使用人工智能技术的频率,

31%的被试表示经常使用,占比最大。25.9%的被试表示很少或从未使用。23.9%的被试表示偶尔使用。19.2%的被试每日使用。总体而言,本次调查的被试群体分布合理,能够有效代表“公众”群体,可用于测度公众对于人工智能的价值敏感性。

2. 变量说明与描述性统计

(1) 解释变量:公众对于人工智能中主要价值的感知能力

人工智能技术关联的价值在测量上是多维的,课题组根据技术特征,参考“技术设计中的价值级序”模型,^[11]制定了人工智能的主要价值级序(见表1)。这一主要价值排序借鉴了亚伯拉罕·马斯洛(Abraham H. Maslow)的需求层次理论和马克思·舍勒(Max Scheler)的价值级序划分,从生命安全价值、心理情感价值到伦理道德价值为逻辑递进的关系,同时侧重关注了伦理道德价值。本研究采用了五级李克特量表^①来测量公众对人工智能中主要价值的感知能力。

(2) 被解释变量:公众的人工智能价值敏感性

表1 人工智能的主要价值级序

程度	价值类型	具体价值	价值内涵
高 ↑ 低	伦理道德价值	信息安全	数据隐私保护、算法和模型保护、合规性等
		人类主体性	确保人工智能系统的发展和应用始终注重人类的需求、价值观和权益
		价值选择	使人工智能系统的目标、决策和行为与人类的道德标准和社会价值观一致
		责任归属	当人工智能系统产生不良结果、错误决策或引发问题时,能够清楚地追溯到责任的具体方
		生态环境保护	在发展和应用人工智能技术时应关注和最大程度地减少对生态环境的不良影响
		算法公正	公平性定义、消除偏见、监管和法规遵循等
		透明性	决策可解释性、算法透明性、数据透明性、模型透明性、用户界面透明性等
	心理情感价值	主体情感平衡	使用人工智能技术来理解、响应和支持人们的情感需求
	生命安全价值	劳动就业保障	确保人们在面对人工智能技术发展时仍能够保持就业机会和劳动权益

①0代表完全其完全不重要,不需要被考虑;5代表其非常重要,需要被重点加以考虑。

价值敏感性在测量上也是多维的,反映型指标(Reflective Indicators)能够从多种维度对作为被解释变量的潜变量进行测量。^[12]课题组选用了“广泛性”“重要性”与“持续性”作为三个反映型指标。^[13]广泛性指的是公众对人工智能价值的认知程度与涵盖范围;重要性指的是公众对于人工智能在伦理道德、心理健康、生命安全等方面的价值是否重要的判断;持续性所测绘的是公众对于上述人工智能价值在未来是否会持续存在的认知。当公众对于人工智能价值持续性的认可程度高时,则表明他们的价值敏感性也相应较高。本文用这三个反映型指标变量来反映与测量“公众的人工智能价值敏感性”。

3. 多指标多因素(MIMIC)模型

多指标多因素是一种特殊的结构方程模型,包括结构模型与测量模型两部分:前者是解释变量(即外生显变量),后者是被解释变量,由一组测量指标检测的内生潜变量组成;该指标模型目前已被广泛应用。^[14]信息安全、人类主体性、价值选择、责任归属、生态环境保护、算法公正、透明性、主体情感平衡、劳动就业保障是客观的“价值”,会影响受访者的“价值敏感性”(潜变量)。而“价值敏感性”则反映在广泛性、重要性、持续性等三个方面。借助MIMIC这一模型,能够有效地将“价值敏感性”的影响因素——客观的“价值”与测量“价值敏感性”的反映型指标区分开来。^[12]这样处理的优点在于,它能够考虑测量误差的影响,从而提高分析的准确性。因此,MIMIC模型不仅考虑了多种影响公众的人工智能价值敏感性的因子,还具备对价值敏感性的内在属性进行测量的能力。

三、实证结果分析

1. 描述性统计结果^①

从统计指标来看,被访者给9类人工智能的主要价值变量的打分介于3.48至3.79分之间。公众对于人工智能的主要价值已经形成了较为明确的感知和理解。反映“价值敏感性”的三个指标中,评分最高的是持续性3.56,大致为最高取值5的70%左右。其次是广泛性3.52和重要性3.48。在判断评估人工智能相关价值时,公众对这项技术随着时间发展所可能产生的持续性影响表示非常关注。这体现了公众的长远目光,也反映了人工智能的广阔前景,表明公众并不局限于短期现状。同时,公众对广泛性的较高打分显示,人们普遍认为人工智能的价值影响并不局限于特定的领域或群体,而是涉及社会结构和个人生活的诸多方面。此外,公众对人工智能价值的重要性也有一定的重视,反映了人们对于人工智能已经及可能产生的负面社会影响的担忧。调查数据显示,被访者对广泛性、重要性等的评价结果与他们给9类人工智能主要价值的打分基本一致。

2. MIMIC模型分析^②

图1展示了MIMIC模型的分析结果,路径线上标注了9类人工智能相关价值因素影响公众价值敏感性的标准化回归系数,其中公众对于算法公正、劳动者就业保障、透明性、责任归属、价值选择、人类主体性、主体情感平衡等评价因素均显著为正。而对信息安全、生态平衡等评价因素则并未通过显著性检验。以上结果说明,被试对于上述7类人工智能相关价值的判断越严重,其对于人工智能的价值敏感性水平就会越高。这种感知体现了公众对于技术影响的担忧。从人工智能相关价值因素影响价值敏感性的原始回归系数估计结果来看,公众对以上7类价值敏感程度的判断每提高一个分值,其价值敏感性得分随之作相应提高。

通过比较标准化系数估计值,能够得出结

①本研究对问卷的信效度与相关性进行了分析,均已通过了检验。在相关性分析中使用的是Pearson相关性分析,结果显示本次问卷所涉及的9个价值要素与价值敏感性都具有正相关关系。并且,每个变量的相关系数均低于0.7,表明不存在严重的多重共线性问题。

②本次问卷调查量表的拟合指数较理想,模型拟合指标数据均符合要求。其中 $CMIN/DF=3.409<5$, $RMSEA=0.048<0.08$; $GFI=0.990$, $AGFI=0.958$, $TLI=0.964$, $CFI=0.990$, $IFI=0.990$ 均达标。

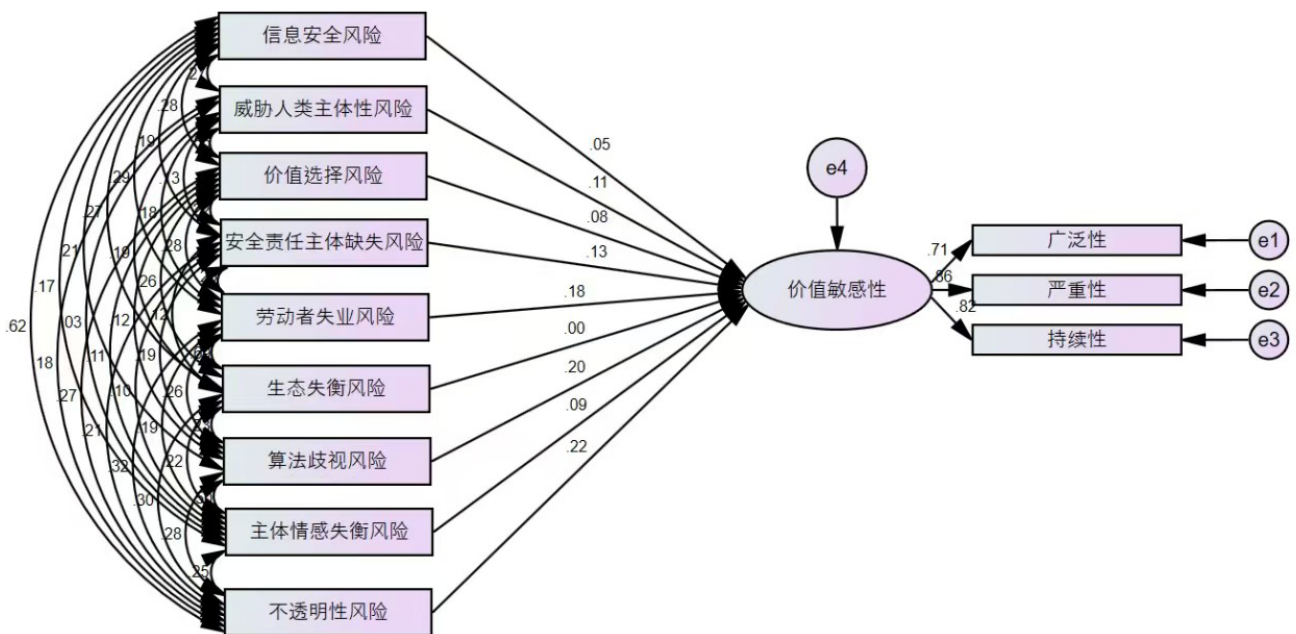


图1 MIMIC模型示意及估计结果

论：对公众价值敏感性影响最显著的是“透明性”这一价值。“算法黑箱”是人工智能的主流技术——深度学习带来的本质特征，它不遵循传统的普遍性规则（如布尔决策规则算法），从数据输入到结果输出，中间过程面临着不透明性和不可解释性，纵使人工智能专家可用标记数据训练以识别数据中的模式或相关性，但该系统内部的具体运行过程依然无法了解和解释。因此，透明性便是公众针对人工智能技术中存在的“算法黑箱”困境而提出的价值诉求。具体而言，“算法黑箱”又分为“客观算法黑箱”与“主观算法黑箱”。其中，“客观算法黑箱”是指算法设计者亦不能完全理解所创造的算法，而“主观算法黑箱”则是指基于某种保密性而受到相关机构部门的保护而对公众具有黑箱性质。^[15]不管是何种黑箱，对于公众而言，其都具有不透明性和不可解释性，会带来潜在的伦理风险与社会恐惧。一方面，公众对于算法透明性价值的高度关注可以看作是对自身知情权的追求，是开展有效的权力保护和责任追求的重要基础。另一方面，公众也担忧在新的“算法社会”秩序当中被架空与被边缘化。技术公司和平台企业正在通过制定网络社会规则等方式，成为新的权力主体，并存在算法越位

的可能。在国家新一代人工智能治理专业委员会发布的《新一代人工智能治理原则——发展负责任的人工智能》中，安全可控原则也明确提出人工智能系统应不断提升透明性、可解释性、可靠性与可控性。^[16]

产生次要重要影响的因素是“算法公正”价值。算法本身并非价值中立的，它负载了特定主体的价值观。选择不当、不完整或具有偏见的数据，设计开发人员有意或无意的价值观偏差，不同利益群体通过设计、编写算法来恶意破坏他人的权益等等诸多因素都可能会带来算法不公正的后果。并且，这种不公正后果借助算法这一特殊的中介具有了隐蔽性、多元性等特性。如果不进行恰当的治理会对社会产生难以估量的危害。日常生活中较为常见的表现形式包括“大数据杀熟”以及招聘广告推荐系统中的性别歧视等。^[15]公正是现代社会中为人们普遍尊重的一项伦理原则。人工智能算法应用具有广泛性和针对性，因而公众也期待算法公正能够渗透到这一领域，从而更好地保障自身的合法权利。

第三位影响因素是“劳动者就业保障”价值。这一因素属于生命安全价值的范畴，其在当前社会背景下的重要性不容忽视。国内学者

的研究表明：“在未来20年中，高达76.76%的总就业人口可能受到人工智能的冲击，而在非农业人口中，这一比例也高达65.58%。”^[17]人工智能技术通过重构产业结构和社会分工模式，正逐步改变劳动力市场的格局。一方面，人工智能的高效性和自动化特性使得许多传统岗位被机器取代，导致大量劳动者面临失业的风险；另一方面，新兴产业的发展又需要劳动者具备新的知识、技能，以适应技术变革带来的挑战。这种结构性失业的现象，对劳动者的就业稳定性构成了巨大的威胁。公众对劳动就业保障价值的关切，实际上是对自身生存和发展的深度忧虑。他们希望在社会和技术变革的大潮中，能够得到合理的保障与支持，确保自身的就业权利和生活质量不受损害。

排在第四位的是公众对“责任归属”的价值敏感性。责任归属指的是当人工智能系统产生不良结果、错误决策或引发问题时，能够准确地追溯到责任的具体方。在传统的技术责任伦理背景下，通常以决策者的行为结果判定其权利与责任的关系。^[18]然而，在人工智能的设计、研发、应用等各环节中存在包括研发人员、生产者、经销商、使用者等多元主体，人工智能系统的复杂性也使得责任归属变得异常困难，例如“在特斯拉、Uber等低级别自动驾驶的伤亡事故发生后，碍于立法和认知的双重缺位，相关责任认定在中外法院审理案件过程中均出现了较大难题和争议”。^[19]明确责任归属是确保技术健康发展的重要保障，能促使相关责任方更加谨慎地设计和使用人工智能技术，推动其向着更加安全、可靠的方向发展。

第五位影响因素是“人类主体性”价值。人类主体性价值强调的是人的主体作用与社会参与程度。有学者指出，人工智能正在借助其拟人特质，尝试取代人类的体力劳动与脑力劳动，使得人类的主体性面临被异化的风险。^[20]一方面，人工智能通过数据化对人进行解构，制造“信息茧房”。^[21]在“信息茧房”下，个体的甄别能力逐渐削弱，更易受到行为偏见的影响，面临着客体化风险，逐渐失去对自身行为和生活的主动控制权，成为算法和数据的

被动接受者。最终个体的主体意识被逐步销蚀，从而陷入对自我认知和自主性的深刻危机之中。另一方面，“人工智能在社会系统中的应用可能造成人对技术的严重依赖，甚至社会治理的角色替代”。^[22]在社会治理方面，政府可能越来越倾向于使用人工智能算法来做出决策，例如用于预测犯罪发生地点和时间、分配社会资源等等。人们逐渐将社会治理的责任和决策权大量交由人工智能系统，而不是依赖人类理性和判断力，从而削弱人类的主体性地位。人不再是治理主体，反而可能成为算法权力中被计算的客体。

排在第六位的是“主体情感平衡”价值。这一价值要求新一代的人工智能系统也具有“感性”的情感连接能力，从而满足人们普遍的心理与情感需求，打破长期以来对人工智能的“高智能、低情商”的标签，然而当下人工智能技术可能无法满足这种需求，进而主体可能面临人机交互导致的“情感失衡”风险。^[23]目前，人工智能技术通常只是基于语言和行为的表面特征（如语音识别、文本分析等）来识别用户的情绪，忽视了背后的情境和个体差异。这导致算法对情感的理解不准确，进而影响其对个体需求和情感状态的判断，使得个体在与人工智能交互时产生误解或不适，形成情感失衡的现象。例如，“在虚拟教学中由于教学主体之间以及教学主体与客体之间的交互作用，仍需要借助数字化中介系统来展开和实现”，所以易造成认知参与但情感遮蔽的“伪参与”现象、^[23]学习体验“单一化”和社会情感学习淡化^[24]等主体情感不平衡的现象。

最后一位影响因素是公众对“价值选择”因素的评价。这一价值的目标是使人工智能系统的目标、决策和行为与人类的道德标准和社会价值观一致，从而能够推动人工智能朝着更为负责任和可持续的方向发展。然而，算法黑箱为隐形权力的运行提供了条件，集中表现为算法的设计、目的、标准体现设计者和开发者的主观意志和价值取向。同时，即使在人工智能系统被赋予了特定的价值理念，它仍可能难以区分正义与非正义行为，因为价值观在不同

文化、社会和个体之间存在显著差异。例如,不同文化和社会对个人隐私权和公共安全的权衡考虑可能存在差异:一些社会可能更加注重个人隐私权的保护,而其他社会则更加重视公共安全。因此,公众较为重视人工智能技术是否能够灵活适应不同文化和社会背景的需求,是否能够最大化平衡个人权益和公共权益。

通过对数据模型的分析,研究发现“信息安全”“生态平衡”等评价因素并未通过显著性检验。在信息安全方面,这一结果可能表明政府和相关部门在信息安全领域,尤其是网络安全、数据泄漏以及隐私侵权等方面,已经采取了有效的措施,如近年来政府加大了对网络安全法律法规的制定和执行力度,企业和组织也提高了对信息安全的重视程度,强化了数据保护。同时,公众教育和意识提升活动的增加,提高了人们对个人信息保护的认识。然而,当前的研究结果反映的仅仅是一个阶段性的成果,并不意味着可以放松对信息安全的关注。随着云计算、大数据、物联网等前沿技术的应用扩展,新的安全漏洞和风险点不断被发现,需要不断更新和加强信息安全管理措施。

其次,在生态平衡方面,当前人工智能技术对生态环境的影响还没有达到公众所能显著感知的程度。这一方面可能是因为人工智能技术在生态资源领域的积极作用与负面影响之间存在复杂的平衡。例如,人工智能技术在提高产业效率、减少资源浪费方面具有巨大潜力,但是与此同时其训练过程中会产生高能耗和需求一些特定的稀有资源。另一方面,生态系统的复杂性和变化的缓慢性意味着人工智能技术带来的生态影响可能需要较长时间才能被充分感知和评估。因此,当前公众对生态平衡问题的敏感性水平并不高。

综合以上数据分析可知,公众对于人工智能领域中的透明性、算法公正、劳动就业保障、责任归属、人类主体性、主体情感平衡、价值选择等因素的价值敏感性较高,是当前公众关注的焦点。这表明了公众对于人工智能的公平正义、责任与算法的透明度等方面高度关切。此外,数据模型还表明,信息安全、生态平衡

等价值因素由于其现实影响相对较小,且具有滞后性和隐蔽性等问题,尚未引起公众的广泛关注,因此公众对其的价值敏感性也相对较低。

3. 差异性分析

(1) 性别

研究发现,不同性别的受访者在伦理敏感性并不具有显著性差异。在信息时代,随着教育的普及和科技发展,男性和女性都有机会接触和学习关于人工智能的应用与知识。这种大背景的相似性可能导致他们在判断人工智能价值时具有趋同的视角和观点。与此同时,男性与女性在处理信息和做出判断时具有许多认知共性。

(2) 年龄

通过单因素方差分析,研究表明不同年龄的受访者在伦理敏感性上具有显著差异。通过进一步多重比较,研究发现17岁及以下、56岁及以上显著高于18岁至35岁、36至55岁。17岁及以下的受访者几乎是在数字时代中成长起来的。他们对于人工智能技术有着天然的亲近感,并能够基于日常的实际使用经验,对人工智能技术形成较为全面深刻的理解。并且,随着社会发展,这一新生群体对于自主权利、社会正义的追求更加强烈,这种思维模式也拓展到了人工智能技术方面,因而表现出更高的价值敏感性水平。56岁及以上的受访者一生中经历了信息技术的巨大变迁。这一经历赋予了他们关于社会变革、伦理责任的深厚理解,使其能够从一个更成熟的角度来评估人工智能技术的影响。他们对技术可能引发的信息安全风险表示出高度的关注,对个人数据的保护尤为重视,因而表现出更高的价值敏感性水平。

(3) 学历

研究还发现,不同学历的受访者在伦理敏感性上具有显著差异。学历越高的受访者的伦理敏感性越高,研究生及以上学历的被试的敏感性水平最高。一方面,高等教育通常鼓励学生发展批判性思维,培养学生的自我价值感和自我意识。另一方面,通过教育,人们往往能够接触到更多关于人工智能技术及其应用的知识信息,使得高学历人群具有更高水平的价值

敏感性。

(4) 职业

最后,本研究还对从事人工智能相关职业与不相关职业的这两大类人群的价值敏感性进行了差异分析。研究结果显示,从事人工智能相关职业与不相关职业的两类人群对于人工智能的价值敏感性并没有形成显著差异。这说明当前人工智能相关价值在不同职业背景的人群中的感知能力与敏感性水平相似,有利于形成构建人工智能治理框架的共识。

四、技术研究分析

在价值敏感性设计理论的“三方方法论”中,“技术研究”作为其核心组成部分,承担着评估现有技术在何种程度上支持或阻碍价值敏感设计目标实现的任务。基于经验研究,我们已经得出结论:公众对于透明性、算法公正、劳动就业保障、责任归属、人类主体性、主体情感平衡、价值选择等7项价值的敏感性水平较高,通过了显著性检验。因此,本文将在技术研究的方法论基础上,对部分人工智能技术进行分析,旨在厘清相关价值是否在人工智能具体技术中得到支持,以及如何得到支持,并给出一些合理的建议。

例如,在通用语言大模型方面,原本的语言模型技术在“透明性”方面存在诸多问题,主要表现在用户难以理解模型内部的决策过程和数据处理机制。这种缺乏透明性的现象导致用户对其结果产生不信任感。为了提高透明性,开发者们采取了一定的改进措施。例如,OpenAI在发布GPT-4时,提供了详细的技术文档和操作手册,帮助用户更好地理解和使用这些技术。与此同时,它还提供了一定程度的数据访问和控制功能。尽管如此,这些改进仍然不足以完全解决透明性的问题,仍需进一步的努力和创新,以确保用户对技术的充分了解和信任。又如在信用评分算法当中,“算法公正”是公众尤为关注的一项价值。过去,由于某些少数族裔和低收入群体在历史上获得贷款和其他金融服务的机会较少,导致这些群体在信用

评分系统中的打分较低,并在申请贷款时遭受了不公正待遇,进一步加剧了经济不平等的现象。这引起了社会的广泛关注与讨论。在此之后,金融公司开始尝试将算法公正这一伦理价值嵌入到信用评估算法当中,包括使用更为多样化的数据集(如租金支付记录和公共事业账单),以更全面地评估申请人的信用风险;引用偏见检测和校正机制,从而实时监测和调整系统的输出等等。

由于篇幅所限,这里不能对公众敏感的7项人工智能相关价值在具体技术中的体现进行详细分析。通过上文,我们能够发现人工智能技术在初级形态阶段往往并不完美。然而,通过技术开发者对公众伦理价值需求的关切,能够在技术迭代过程中将相关的伦理原则嵌入到新技术的设计当中,转化为具有可操作性的实践指南,给人工智能植入一颗“善芯”。但目前的人工智能技术距离“科技向善”的目标仍有一段距离,需要技术开发者、政策制定者和社会各界的通力合作,从而共同推进人工智能技术向善发展。

结 语

通过上文的分析,可以得出如下结论:

首先,公众对人工智能价值敏感性整体水平较高。通过问卷调查显示,公众的价值敏感性判断的平均评分为3.48至3.79之间(最高值为5)。同时,从广泛性、重要性、持续性三个方面测量了“价值敏感性”水平,整体均值为3.52(最高值为5),为最高值的70%。由此可见,当前公众对于人工智能的价值敏感性水平整体较高,并在年龄和学历方面呈现显著性差异:年龄在17岁及以下、56岁及以上的公众价值敏感性显著高于18岁至35岁、36至55岁;学历越高,价值敏感性亦越高。

其次,公众较为敏感的价值原则可测。通过采用MIMIC结构方程模型检测,公众对于透明性、算法公正、劳动就业保障、责任归属、人类主体性、主体情感平衡、价值选择等7项价值的敏感性水平较高,通过了显著性检验,

对“价值敏感性”的影响显著为正。上述7项价值每提高一个分值,其价值敏感性得分将分别作相应提升。而信息安全、生态平衡两项价值则并未对公众的价值敏感性判断产生显著影响。需要指出的是,在进一步进行差异性分析时发现,不同性别或职业人群对于人工智能技术的价值敏感性并未呈现显著差异,而不同年龄或学历的受访者群体之间呈现出显著差异。

透过上述结论,又给我们带来如下启示:

第一,价值敏感性研究为人工智能技术发展提供价值指南。自弗里德曼等人上世纪90年代中期提出价值敏感性设计理论以来,该理论提出的理念即在技术设计阶段,将带来伦理道德意蕴的价值原则融入到技术创新当中。但一直以来,该理论在国内外学界或者以宏观的理论探讨居多,或者只是对其进行引介评述,缺乏与具体技术应用的微观层面的量化研究。课题组立足于学界相关研究成果,创造性地将这一抽象的哲学概念转换为可操作化的量化指标,运用问卷调查和数据建模的方式,调研了人工智能技术应用过程中9项“价值敏感性”指标,并参照价值敏感性设计理论中的“三方方法论”,探索性地开展了“概念研究”“经验研究”与“技术研究”的“三方研究”,进而动态把握了普通大众对人工智能技术应用已经或者可能带来的伦理风险感知状况,并对现有技术何种程度上支持或阻碍价值敏感设计目标实现进行了评估。

第二,拓展以公众关切为导向的人工智能伦理治理的新进路。随着人工智能技术的广泛应用,特别是生成式人工智能的快速发展,人们对其带来的伦理风险愈发关注,随之而来的对人工智能技术应用所产生的伦理治理工作亦变得愈发紧迫。目前,在全球范围内,各个国家都纷纷行动起来,共同应对包括人工智能技术在内的科技伦理治理难题。我国的科技伦理治理工作正逐步迈入正轨,2022年3月中共中央办公厅、国务院办公厅颁布了《关于加强科技伦理治理的意见》,2023年10月,科技部联合10部委印发《科技伦理审查办法(试行)》。为更好地推进科技伦理的治理工作,一方面采

取“自上而下”的研究进路,深刻理解和把握前沿科技发展的内在机制,制定相对完善的伦理审查制度和法律法规体系,做好顶层设计,但这些新兴技术到底带来哪些伦理风险,又如何将抽象的伦理原则贯彻到具体的技术研发应用之中,这是该进路的困难所在。面对如此困境,另一方面采取“自下而上”的研究进路,即动态掌握普通大众对新兴科技的伦理风险感知状况,厘清新兴科技应用的底层逻辑,可为科技伦理治理的顶层设计提供经验参考。本课题即在后一层面进行了实践探索,通过实证研究,具体掌握公众对人工智能技术应用过程中较为关切的伦理风险,并对此进行了价值排序,为人工智能伦理审查提供经验参考,旨在探索一条人工智能伦理治理的新路径。

[参考文献]

- [1] Friedman, B., Kahn, P. H. 'Human Agency and Responsible Computing: Implications for Computer System Design'[J]. *Systems and Software*, 1992, 17(1): 7-14.
- [2] Friedman, B. 'Value-Sensitive Design'[J]. *Interactions*, 1996, 3(6): 16-23.
- [3] 张浩鹏、夏保华. 价值敏感性设计透视: 背景、现状、问题与未来[J]. *自然辩证法研究*, 2023, 39(4): 77-83.
- [4] Friedman, B., Hendry, D. G. *Value Sensitive Design: Shaping Technology with Moral Imagination*[M]. Cambridge: MIT Press, 2019.
- [5] Friedman, B., Kahn, P. H. *The Handbook of Information and Computer Ethics*[M]. New Jersey: John Wiley & Sons, Inc., 2008.
- [6] 张浩鹏、夏保华. 人工智能道德智能何以可行——基于对价值敏感性设计的审视[J]. *自然辩证法研究*, 2021, 37(4): 37-42.
- [7] 闫坤如. 人工智能设计的道德意蕴探析[J]. *云南社会科学*, 2021, (5): 28-35; 185-186.
- [8] 于雪、李伦. 人工智能的设计伦理探析[J]. *科学与社会*, 2020, 10(2): 75-88.
- [9] Wilson, C. 'Public Engagement and AI: A Values Analysis of National Strategies'[J]. *Government Information Quarterly*, 2022, 39(1): 101652.
- [10] Robles, P., Mallinson, D. J. 'Artificial Intelligence Technology, Public Trust, and Effective Governance'[J]. *Review of Policy Research*, 2023, 1-18.

- [11] 刘瑞琳. 价值敏感性的技术设计探究 [D]. 沈阳: 东北大学, 2017.
- [12] 阳义南. 民生公共服务的国民“获得感”: 测量与解析——基于MIMIC模型的经验证据 [J]. 公共行政评论, 2018, 11 (5): 117-137.
- [13] 李森林、张乐、李瑾. 当代青年人工智能风险感知的测度与解析 [J]. 科学学研究, 2023, 41 (10): 1737-1746.
- [14] 阳义南. 结构方程模型及Stata应用 [M]. 北京: 北京大学出版社, 2020.
- [15] 陈雄燊. 人工智能伦理风险及其治理——基于算法审计制度的路径 [J]. 自然辩证法研究, 2023, 39 (10): 138-141.
- [16] 发展负责任的人工智能: 新一代人工智能治理原则发布 [EB/OL], https://www.safta.gov.cn/kjbgz/201906/t20190617_147107.html. 2023-01-20.
- [17] 陈永伟. 人工智能与经济学: 近期文献的一个综述 [J]. 东北财经大学学报, 2018, (3): 6-21.
- [18] 冯永刚、席宇晴. 人工智能的伦理风险及其规制 [J]. 河北学刊, 2023, 43 (3): 60-68.
- [19] 赵志耘、徐峰、高芳. 关于人工智能伦理风险的若干认识 [J]. 中国软科学, 2021, (6): 1-12.
- [20] 苗存龙、王瑞林. 人工智能应用的伦理风险研究综述 [J]. 重庆理工大学学报 (社会科学), 2022, 36 (4): 198-206.
- [21] 郑智航. 人工智能算法的伦理危机与法律规制 [J]. 法律科学 (西北政法大学学报), 2021, 39 (1): 14-26.
- [22] 张铤. 人工智能的伦理风险治理探析 [J]. 中州学刊, 2022, 40 (1): 114-118.
- [23] 冯锐、孙佳晶、孙发勤. 人工智能在教育应用中的伦理风险与理性抉择 [J]. 远程教育杂志, 2020, 38 (3): 47-54.
- [24] 赵磊磊、张黎、代蕊华. 教育人工智能伦理: 基本向度与风险消解 [J]. 现代远距离教育, 2021, (5): 73-80.

[责任编辑 李斌]

