DOI: 10.16538/j.cnki.fem.2019.10.008

# 如何认识人工智能的伦理冲突？
## ——研究回顾与展望

谢洪明[1,2,5]，陈　亮[1,3]，杨英楠[4]

（1. 浙江工业大学 管理学院，浙江 杭州 310023；2. 广州大学 工商管理学院，广东 广州 510006；
3. 浙江金融职业学院，浙江 杭州 310018；4. 浙江大学 工程管理研究所，浙江 杭州 310058；
5. 浙江工业大学 中小企业研究院，浙江 杭州 310023）

**摘　要：**人工智能尤其是机器人在各个领域从事越来越多的决策，逐步从被动工具变成人类的代理者，这引发了社会各界对人工智能伦理的思考和担忧，需要建立新的伦理范式，将人类社会的伦理规范延伸到智能机器。近年来该研究取得了较大发展，丰富了已有研究内容，提升了对人工智能伦理研究的指导能力。本文全面回顾了国内外有关人工智能伦理的研究进展，在文献计量分析的基础上，发现“传统派”、“谨慎派”和“乐观派”三种对人工智能的不同态度引发了“人—机”关系的伦理冲突，从人工智能道德哲学、道德算法、设计伦理和社会伦理四个视角系统性地评述了人工智能伦理的研究成果，国家和企业（组织）分别从战略和社会责任的层面上强调对人工智能伦理的态度，提出了更加系统、完善的人工智能伦理的理论框架，有助于从理论和实践层面系统地把握已有研究成果。未来需要在全球情景条件的伦理体系建设、伦理对技术的前瞻性、伦理角色塑造和科学发展的伦理观上做进一步研究。

**关键词：**人工智能伦理；机器人伦理；述评及展望

**中图分类号：**F270　**文献标识码：**A　**文章编号：**1001-4950(2019)10-0109-16

## 一、引　言

随着人工智能（artificial intelligence，AI）的发展，人工智能伦理在国内外都成为了各界讨论和研究的核心议题之一（Moor，2006；Wallach和Allen，2010；Greenwald，2015；杜严勇，2015；Lei等，2019；赵汀阳，2019）。人工智能伦理研究是人工智能时代的必然产物，它既包括对技术本身的研究，也包括在符合人类价值的前提下对人、机和环境之间的关系研究（刘伟和赵路，2018）。发展的不确定性带来了新的伦理问题和风险，尤其是奇点理论（singularity theory）提出

**收稿日期：**2019-01-07

**基金项目：**国家自然科学基金项目（71772163，71673240）；浙江省自然科学基金项目（LY17G020024，LY16G020009）

**作者简介：**谢洪明（1971—），男，浙江工业大学教授，博士生导师；广州大学工商管理学院教授；
陈　亮（1985—），男，浙江工业大学博士研究生；浙江金融职业学院讲师（通讯作者）；
杨英楠（1977—），女，浙江大学工程管理研究所副教授。

之后，许多人对人工智能的快速发展表示了担心，随之而来的诸如无人驾驶汽车致人死亡、护理机器人带来人格与尊严等问题更加印证了人们的这种印象。同时，人工智能也带来歧视、侵犯个人隐私、改变就业结构、挑战国际关系原则等问题，这对个人权利、政府监管、经济社会发展乃至全球治理将产生深刻影响（Wallach和Allen，2010；Bostrom，2014；Diakopoulos，2015；杜严勇，2015；苏令银，2019）。因此，需要引起更多的人关注人工智能伦理，为未来人工智能的发展奠定道德性根基。

如何正确认识人工智能在发展过程中出现的失败案例（Yampolskiy和Spellchecker，2016）？是因为数据爆发式增长和深度学习算法的发展，人类的决策让渡使机器从被动工具向能动体（intelligent agent）转变而具备不需要人类介入的“感知—思考—行动”体现出的自我学习进化能力，还是因为人工智能系统在算法引导下的错误，以及技术的不确定性导致后果的难以预测和量化评估（闫坤如和马少卿，2018）？是因为人类的有限理性引发了人工智能的伦理风险，还是说人工智能还远没有具备伦理的属性？这些问题在学术界还尚未形成统一的认识。如何评价人工智能的行为将会直接影响人以何种方式与机器相处，因此，需要系统地梳理人类面对的形式多样的伦理挑战，通过发展和改进人类现有的伦理体系，为国家实施《新一代人工智能发展规划》解决伦理性问题，更好地适应人工智能的发展。

本文的目的是借助内容分析方法，试图厘清人工智能伦理研究的发展脉络、现状和趋势，对现有研究从道德哲学、道德算法、设计伦理和社会伦理四个方面进行评述，以期对未来的理论研究和实践提供借鉴。行文结构如下：第二部分对人工智能伦理相关文献进行计量分析；第三部分分析“人—机”关系引发的伦理冲突；第四部分综述了人工智能伦理研究的视角及内容；第五部分为国家及企业对人工智能伦理的态度；第六部分是总结和未来研究展望。

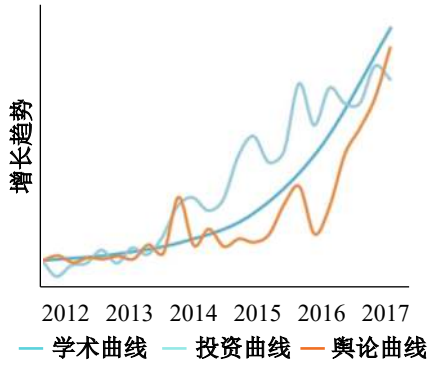
## 二、人工智能伦理相关文献的计量分析

### （一）文献来源

人工智能伦理的研究背景跨度宽泛，涉及计算机科学、人工智能、机器人学、伦理学、哲学、生物学、社会学、宗教等方面。本文基于Web of Science（WOS）和CNKI数据库，以Citespace为文献计量工具，以知识图谱的形式分析人工智能伦理的研究脉络。选择以人工智能伦理（AI ethics）、机器道德（machine morality）、机器人伦理（robert ethics）、机器伦理（machine ethics）、道德机器（moral machine）、价值一致论（value alignment）、人工道德（artificial morality）、技术伦理（technology ethics）、人工智能安全（AI security）、友好人工智能（friendly AI）等为关键词进行搜索。为保障文献的质量，剔除影响因子较低的期刊，选取1997—2018年（共21年）期间发表在较高影响因子期刊上有关人工智能伦理研究的298篇文献。将阈值设置为“Top50”，选择每个时段出现频次在前50的文献形成知识图谱。

### （二）文献研究分析

中国“人工智能”在理论研究、资本投资以及社会关注度方面呈逐年递增的趋势，如图1所示。其中“学术曲线”是CNKI中“深度学习”相关论文各年份发布数量变化

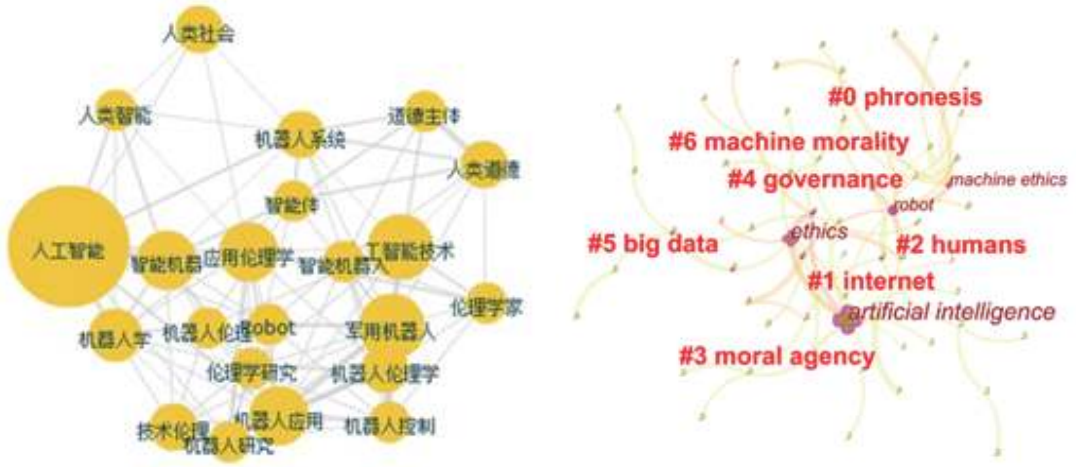



资料来源：亿欧智库2018中国人工智能投资市场研究报告。

**图1　中国人工智能2012—2017年学术、投资和舆论变化曲线**

情况，“投资曲线”是人工智能私募股权投资市场的投资频数变化，“舆论曲线”是百度指数中“人工智能”的搜索热度变化。从曲线的变化趋势来看，学术的研究热度呈现持续增长态势，大量资本进入人工智能研究领域大概出现在2014年，公众舆论中“人工智能”热点出现于2016年。学术到资本市场转化再到舆论热点关注，人工智能发展趋势与历次新技术产生和发展的情况类似，可以预测人工智能及其相关领域未来都将是研究和关注热点。

主题词是对文献的核心内容的提炼和概括，其频次高低也在一定程度上反映了文献的研究热点和趋势。本文基于搜集的人工智能伦理文献，进行了主题词共现分析，如图2所示。




资料来源：利用CNKI和Citespace分析所得。

**图 2　CNKI和WOS数据库关于人工智能伦理主题词共现网络分析**

国内外从技术、文化、伦理、制度等方面对AI进行研究，形成了实践智慧、互联网、大数据、机器道德、机器伦理等明显的聚类。机器道德（machine morality）关注机器人应该拥有哪些道德性能，以及这些性能如何实现的问题（Malle，2016）。机器伦理（machine ethics）关注机器人在设计、应用实践和与人类相处等方面的问题（Anderson等，2005）。机器人伦理（robert ethics）以及路线图的提出（Di Robotica和Veruggio，2006），促进学者对机器人技术使用影响的跨文化讨论，从伦理因素的植入程度角度来理解广义机器伦理（Moor，2006）。

需要说明的是，机器人是目前人工智能研究最重要的载体，是多种技术的集合，因此在很多研究中将机器人伦理等同于人工智能伦理。但二者还是存在一定差异，机器人伦理是人工智能伦理的一部分，前者属于后者研究范畴的核心，就二者关系而言属于包含关系（莫宏伟，2018）。

（三）国家级课题分析

国家级课题立项情况能够从另一个角度反映与人工智能伦理的研究现状。以“人工智能”、“机器人伦理”为关键词搜索国家社科网，相关的立项课题计16项，再根据申报人信息进行期刊论文查询，剔除相关度不高的课题6项。在文献查阅的过程中，另发现与人工智能伦理相关的课题16项。因此，在2002—2017年间，与人工智能伦理相关的国家级课题研究共计26项，立项范围涉及人工智能道德哲学（刘西瑞，2002；李小五，2002；胡惊雷，2015；李熙，2017；王绍源和任晓明，2017；梅亮等，2017；高兆明，2017）、技术伦理（李伦，2018；张成岗，2017）、生活伦理（徐英瑾，2006，2015；阎国华和闫晨，2017；党家玉，2017；张吉豫，2015）、伦理解决进路（杜严勇，2015；段伟文，2017；黄闪闪，2017）等内容，这与文献研究热点分析保持一致，表明国家层面也在关注人工智能伦理研究。

## 三、“人—机”关系引发的伦理关注

使用计算机技术在伦理设计框架内跨界进行道德模型构建并应用于实践时，研究具有开创性和前瞻性（Danielson，1992）。“智能体”（agents）再加上“人工”的属性时，这种智能就属于机器而不是生物的（Ferraz和Del Nero，2018）。智能机器通过自我深度学习拥有了工具理性，对世界中发生的事能做出判断和预测，构成一种全新“人—机”关系。学术和实践领域对此主要有三种立场和观点，如表1所示。

**表 1　人工智能发展引发对伦理问题的关注**

| 对人工智能的态度 | 代表人物 | 观点 | 引发伦理问题的评价 |
|---|---|---|---|
| 传统派 | Neumann、Turing、Dreyfus、Searle | 人类智能是AI的极限状态，无法超越 | （1）AI只是手段与工具，无法区别“善与恶”“好与坏”，关键在于应用后果的善恶评价；（2）AI发展仍处于初级阶段，无法确保AI必然会服从人类设定的道德标准；（3）应加强弱人工智能、强人工智能、超级人工智能以及人类智能与人工智能之间的关系问题研究 |
| 谨慎派 | Hawking、Bill Gates、Mask | AI的发展会威胁人类的生存，存在“作恶”可能 | |
| 乐观派 | Herbert A.Simon、Ray Kurzweil、Markram | AI最终能够达到乃至超越人类智能水平、“奇点理论”、“人机共存” | （1）片面地、孤立地看待AI的积极方面，忽视或者掩盖AI的消极作用；（2）采取谨慎发展的态度，通过政府监管，制定机器伦理和价值体系 |

资料来源：本文根据相关文献整理。

一是以冯•诺依曼、阿兰•图灵为代表的“传统派”。他们认为人类始终处于支配地位，人工智能永远不会超越人类智能，这种观点源自于宗教中“造物主一定比所造之物高明”。20 世纪 70 年代从生物和心理学的层面得出了人工智能必将失败的结论（Dreyfus，1972），认为“当前人类面临的风险，不是超智能机器的降临，而是低智能人的出现”。80年代的“中文房间”（Chinese room）模型对上述观点进行了印证（Searle，1980）。

随着人工智能的广泛应用，人类也逐步开始重视、反思弱人工智能与强人工智能，甚至超级人工智能（super intelligence）的研究以及与人类智能之间的关系问题。尼克·波斯特姆（Bostrom，2014）从“全脑仿真、生物认知、人机交互以及网络和组织”等路径分析“超级智能”的可能性，这不同于目前广泛应用的“弱人工智能”（如下棋机器人、自动驾驶技术等），而是一种能全面超越人类智能的“强人工智能”。但这些实现路径是在人脑基础上的直接迭代或高仿真，混淆了“强大的弱人工智能”与拥有主体性的“强人工智能”，并没有深入探讨人类智能现象的本质，对经典的“他心问题”也无涉及，所以他提及的“超级智能”还是属于工具性“弱人工智能”的范畴。

二是以霍金、比尔•盖茨、马斯克为代表的“谨慎派”，他们认为人类应该敬畏人工智能的崛起。当机器具备自我意识（self-aware），甚至可以通过自己的“神经”和“意识”自主做出决定，并与其他人工智能机体互联，人工智能将最终构成一个现实性的威胁。

但是扎克伯格、李开复、吴恩达等人对“人工智能威胁论”提出相反的观点，争议的焦点在于强人工智能是否以及何时会出现。马斯克语境中主要是指“强人工智能”，即具备处理多种类型的任务和适应未曾预料情形的能力，代表了公众对强人工智能和超级人工智能可能失控、威胁人类未来的生存表达了担忧；而扎克伯格所述的“人工智能”是狭义的专业领域人工智能能力，是产业界从功用和商业角度出发，保持对人工智能研发和应用的持续探索。这两种典型的观点源于对“人工智能”的不同理解。

三是以赫伯特•西蒙(Herbert A.Simon)为代表的“乐观派”,他们认为人工智能肯定会达到并超过人类智能。“奇点理论”(singularity)预言机器智能将在2045年超过人类的智能(Kurzweil,2011),“蓝脑计划”(blue brain)提出的时间更短,预言在2020年左右制造出科学史上第一台会“思考”的机器。但是这些预言遭到了很多人的质疑,认为人工智能的基础技术是过去技术研发的量变积累,还未达到质变的程度。未来人工智能和人类智能将会是社会发展的两种形态,“人机共存”将成为人类社会比较理想的状态。

虽然对人工智能的发展持乐观态度,但人类社会在哲学、理智等各方面都还没有准备好迎接人工智能时代,需要呼吁人类正确使用人工智能,尽快制定机器人相关方面的伦理和价值体系,通过建立标准寻求“预防性原则”进行风险评估,在此基础上对危险与风险再进行评价(Wallach和Allen,2010)。

人工智能伦理是技术发展中不可回避的重大问题。但是,也有部分学者表示人工智能还远没有伦理的概念,只是人类对伦理的理解并强加在机器上(刘伟和赵路,2018)。人工智能的本质仍然是人与人之间关系的研究范畴,呈现的问题与危机仍属于控制与反控制的议题,也就是说,人工智能的技术飞跃或者“智能大爆发”带来的应用伦理问题并不是新问题,而是一系列老问题的叠加。它会对人类既有的认知提出挑战,改变社会学阶层分析的经典框架,但不是颠覆性的“消灭”(Bostrom,2014)。

## 四、人工智能伦理研究视角及内容

国外对人工智能伦理的研究起步较早,研究领域主要集中在机器人的社会伦理、法律和安全方面,涉及技术性失业、致命性自主武器、算法公平、道德判断、价值一致性等问题。国内研究相对滞后,但人工智能与机器人的伦理责任议题也逐渐被关注,具体包括人权伦理、责任伦理、道德地位伦理、代际伦理、环境伦理以及大数据伦理与隐私等,说明国内已从单一的技术伦理研究转向人机系统交互关系的伦理问题研究,这也是研究领域的新突破和进步。本文结合李伦(2018)在第四届全国赛博伦理学暨人工智能伦理学研讨会上根据国内外人工智能伦理研究的状况和趋势、人工智能与社会各领域的关联属性,从人工智能道德哲学、道德算法、设计伦理和社会伦理四个方面来理解人工智能伦理的研究。

(一)人工智能道德哲学

1. 机器人道德地位问题

人工智能的发展提出了机器人道德地位问题的命题(苏令银,2019)。唯物主义观点认为人的意识是基于大脑神经元而产生的,人类的所有行为活动都是在意识驱动下完成的,而人工神经网络(artificial neural network)无法自动合成“神经蛋白”这类特殊物质,不会产生主观的意识。因此,意向性(intentionality)被认为是人工智能与人类最本质区别之一(Boden,2017),解决这个难题涉及“他心难题”及其变种的“机心难题”,这属于形而上学的思考,是人工智能伦理研究的基础和本源。心智的边界在一定条件下可以不依赖人类生理系统而通过外部物理载体来实现(Heersmink,2009)。因此,道德地位需要考虑的不仅是属性问题而更多地是关系问题,即实体之间以及与人类之间的关系,尤其是主体道德地位和客体道德地位的归属。独立于这些关系去定义道德地位,本身就是违背道德的,因为它被当作了一个具有抽象“属性”的“实体”。

机器人责任主体的界定可以表述为“自主智能机器人在人类社会中应该扮演什么角色?”。沙特阿拉伯授予机器人索菲亚“公民身份”而备受关注,成为“机器人公民”意味着具有权利义务并对其行为负责,需要承担道德责任。但是普遍的学者却认为机器人作为“人工物”是没有能力去承担道德责任的,目前的人工智能仍然是对人类思维、行为的模仿,停留在工具和机械范

畴，并不是真正意义上的人格实体与道德实体。也许随着量子技术的发展，未来会出现类动物甚至类人类AI，即强人工智能或超级人工智能，人类与AI的关系会发生根本性转变，但这仍是停留在理论层面的不确定问题，未来需要科技的发展来检验。经典的图灵测试在人工智能的“强弱”上也不能作为智能意识产生的推演依据，已经产生了“人工智能逆反图灵测试”[①]的现象。阿尔文•纳拉亚南建议在机器人设计时需要遵循“图灵红色警戒”（Turing red flag law），即机器人的定位也应当是机器人，而不是人甚至超越人。

2. 人工道德主体边界

“物伦理”的概念提出以后，在“人—机”系统交互情境下，技术伦理学出现了“物转向”（thingly-turn）的趋势，“技术人工物”（technological artifacts）在技术伦理学领域掀起了新的研究热潮（Peter和Anthonie，2002）。对人的理解越来越物化和去意义化，导致人和“人工物”的边界越来越模糊，人类需要思考这种边界模糊的后果。

两种对道德主体不同的观点形成了人工道德主体边界之争，即实用主义（Levin，2013）和标准主义（Eshleman，2014）。实用主义认为人工道德主体对外界产生某些行为和反应，需要模拟现实行为的部分或者全部标准。黛博拉·约翰逊（Johnson，2006）通过四个条件来设定人工道德主体的边界[②]，她认为人工物不能成为道德主体，是因为在技术上很难对人类的道德行为进行编码建立数据库，缺乏可能导致他们行为发生的内在心理规定。标准主义观点认为人类道德主体需要满足特定的条件来模拟或展现行为的全部标准，提出了人工道德主体标准所要求的状态，但同时也应该从实用主义的角度来理解。

（二）人工智能道德算法

1. 算法歧视

算法是通过计算机一系列运转解决特定的问题或完成一个确定的结果（Diakopoulos，2015），人类已经开始将部分事情的决定权转移到具备高度智能的算法。但是算法缺乏学习、推断、联想等高级智能，难以解释行为背后的决策逻辑，因为人类的行为具有主观动机性和意向性，需要对算法的伦理进行分析，这也是算法设计过程中的重点和难点问题。

算法是一种相对客观的数学表达，很多人会认为算法决策倾向是公平的。但是人类的价值观和道德规范却没有纳入数字系统，对数据的管理在全球互联的背景下更难以控制。数据和算法的不透明性反映了数据权利之间的失衡现象，这将不可避免导致算法歧视（Algorithms Bias）。美国在计算机（Joseph等，2016）、犯罪学（Goel等，2016）、政治学（Veale和Binns，2017）、人力资源管理（Gumbus，2017）等多领域揭示了算法歧视的规律，发掘了相关数据和证据，研究层次较深。歧视的主观倾向性难以判断，这也致使国外学者在研究中更多采用实证和案例的研究方法来提出建议（Edelman和Luca，2014）。国内研究主要从经济、法律和传媒角度，多以介绍性和科普性为主，集中在对国外研究成果的介绍和算法歧视现象的评述，仅有少量做了定量分析，而其他领域关注较少。

2. 算法歧视产生的根源

个体通过数据获得了非对称优势，算法技术将对个人数据进行重构，规则代码化带来的不

①图灵测试表示有超过30%的测试者不能确定出被测试者是人还是机器，就表明这台机器通过了测试，并被认为具有人类智能。谷歌公司2018年发布的“Duplex”的人工智能系统，其最大特点是系统能够进行较自然的声音对话，这一切都让此人工系统听起来像一个真实的人。谷歌Duplex模仿人类发声的启示引发了社会对人工智能的担忧，高度发达的人工智能系统能让人类误以为这台机器就是人类，但是Yaniv Leviathan和Yossi Matias作为项目主要负责人，表示只关注了产品的技术潜力，并未意识到其存在的社会隐患。

②黛博拉•约翰逊（Johnson.D，2006）规定了一个人工道德主体E要成为道德主体必须具备下列条件：（1）E能够以自己的身体引起物理事件；（2）E有一个内在的规定I，它组成了自己的欲望、信仰和其他有意图的规定，这些规定共同组成了E以某种方式（理性和意识）行动的原因；（3）规定I是导致条件（1）的直接因素；（4）条件（1）中的物理事件产生一些具有道德重要性的影响：当计算机系统运转时，它们的行为对表达世界的其他方面产生影响，这些表达效果对道德施加者有害或有益。

透明、不公平、难以审查等问题对公平和正义发起挑战(Veale和Binns,2017),让网络环境在定价、服务和机会等领域具备了歧视的可能(Helbing,2015)。

算法歧视产生的原因主要归结于算法技术和数据输入两个方面(Executive Office of the President,2016),机器完全受人工设计的结构和学习进程中接收的数据形成对世界的"认知",导致在特定场景下发生错误决策。算法的设计是技术人员的主观选择和判断,他们是否公平公正地将道德规则和法律写进程序是值得怀疑的,这使得算法继承了人类决策者的种种偏见。此外,由于技术壁垒、机器学习算法的"黑箱"属性、政策限制而产生不透明性,让计算机在不被明确编程的情况下运转,在自主系统中探究是否存在歧视和其原因,很难通过技术来实现公平。

算法离不开大数据支撑,数据的有效性、准确性会影响算法决策和预测的准确性。数据自身偏见缺陷、大小样本的地位悬殊、敏感属性都会导致算法歧视存在的必然性(张玉宏等,2017),带有歧视的数据经过运算之后也带有歧视倾向,创新也会使歧视和机会不平等的现象持续下去。随着数据应用范围的扩大,算法的一次小失误或者歧视,容易形成"自我实现的歧视性反馈循环",包括在警情预测、风险评价和信用评估上都存在类似问题,这会压榨消费者个人财富、剥夺个人自我决定权、破坏信息的多样性,不断蚕食消费者剩余,甚至对个体生命构成潜在的威胁。

3. 算法歧视的解决出路

由于算法的客观性和人类社会伦理的发展,目前以算法为主的弱人工智能并非完全可以依赖和信任(莫宏伟,2018),需要承认人工智能系统内在歧视性,并意识到这些歧视的潜在来源。把伦理原则形式化,在机器学习中引入"歧视指数"(discrimination index)的概念,维护用户与机构间数据权利的平衡,倡导有规范的数据共享,建立基于权利的数据伦理。

算法采用的逻辑基础是多语境系统,从设计、执行、测试到推广采取跨学科的原则,做到更加包容性。提出设计更为"公平的"算法,从算法设计的源头来防止歧视,使人工智能体的行动符合相关伦理规范。保持算法的透明度,制定机会均等的设计原则,使算法的力量发挥作用,尤其需要考虑弱势、易被忽视人群的利益。国外现在已经有OpenAI等一些人工智能开源运动,保障算法的透明性和公平性。Google在实践中倡导"机会平等"(equality of opportunity),以避免一套敏感属性的歧视。当然,算法透明和商业秘密、国家安全之间的关系需要进一步去平衡。

越来越多的国家、地区、组织出台监管措施,包括设计标准、性能标准、责任标准等,甚至扩展到涉及人工智能算法、数据等核心的内容。美国通过大数据报告提醒信贷、就业、教育和刑事司法领域等歧视问题需要从立法、技术和伦理方面予以补救。也有学者建议修改美国残疾人法案,扩大反歧视的授权,禁止利用算法对身体和精神障碍预测性歧视;要求以书面形式说明数据处理者对含有健康类信息收集和处理过程;满足公平正义和经济效率的前提下,保护大数据市场的竞争秩序(Townley等,2017)。

(三)人工智能设计伦理

1. 人工智能设计的伦理原则

人工智能在应用过程中对人类产生侵蚀和威胁的案例证明了技术异化的伦理风险,研究中用"人类关切"(human-centric)来描述这种内在价值的亏损。这种亏损是在被制造的时候缺乏精细和复杂的伦理学设计(Bostrom,2014),因此设计过程中需更多聚焦人与数据的自由关系(李伦,2018)。美国电气和电子工程师协会(IEEE)将伦理植入设计人员的意识中,鼓励科技人员在人工智能研发过程中优先考虑伦理问题(Davies,2016)。

基于人权优先、强调人机和谐共处,在人工智能设计中需要优先考虑:(1)人类利益,确保

人工智能和自主系统(AI/AS)不侵犯人类,这也是阿西莫夫(Asimov)提出的机器人的三大定律之一,后来又提出了"零律"对其做出补充[①]。虽然这些规范在理论上得到了大家的认可,但是从未在实践中被完全实现,因为这没有给出能够被业界所推崇的机器编码技术路线。(2)责任,确保AI/AS是可以被问责的。在设计程序层面具有可责性,强化机器人研发人员的职业责任(Murphy和Woods,2009),证明其为何以特定的方式运行。(3)透明性,AI/AS的运作必须是透明的,可以使人类发现机器人是如何以及为何做出特定的决定。透明至少应该包含开放和可理解两个方面,只有让用户和机构在算法上达到平衡,才能确保用户的数据权利。(4)教育与意识,需要强化AI/AS的优势,推进伦理教育和安全教育意识,降低其被滥用所带来的风险(腾讯研究院等,2017)。

2. 人工智能伦理设计进路

人工智能作为一种技术必然带有价值偏好,机器的自由化程度越高,就越需要道德标准。由于人工智能系统在自主性方面的能力的提升,在设计阶段让其采纳、学习并遵循所服务的社会和团体的规范和价值显得至关重要,可以分步骤来实现将人类规范和道德价值嵌入人工智能系统:

(1)识别特定区域范围内的规范和价值。社会和道德规范是特定区域范围内内化于行为、语言、习俗等针对特定任务的价值观,人工智能系统受到多种规范和价值约束,应明确需要嵌入的规范和价值,避免导致道德过载(moral overload)问题。在人工智能价值权重设计阶段,优先考虑利益相关群体共同分享的价值体系,技术上要满足不同空间和时间下价值和规范发生变化的可能性。

(2)将这些规范和价值嵌入人工智能系统。实现机器人伦理道德的决策设计,避免机器的不当使用威胁人类的生存,需要把伦理准则变成可编程的代码嵌入人工智能机器。伦理设计进路包括"自上而下(top-down)"和"自下而上(bottom-up)"。TD研究方法即认知主义,突出人类的主体地位,利用"特定的伦理理论进行分析指导实现该理论的运算法则和子系统的计算需要"的方法,提倡研究人类的策划、推理以及解决问题的能力,在"机器人三定律"基础上构建人工道德主体的理论研究框架。BU研究方法是涌现主义,强调机器的能动作用,即不通过任何抽象概念或推理过程,机器人对所感知到的刺激能动做出反应,能够从不同的社会机制中动态地进行集成输入,为完善人工道德主体整体性发展提供技巧和标准。

随着深度学习算法的提出与发展,综合运用TD的演绎规约路径和BU的归纳规约路径来对人工智能伦理问题进行分析,提出了"混合型解决方法(hybrid resolution)"。内部从顶层设计开始对人工智能产品进行伦理设计,建立人工智能的伦理标准和规范,限定存在争议和容易引起安全问题的技术应用范围和智能水平。外部需要加强设计者的社会责任意识,加强对人工智能安全的评估和管理,增加公众对人工智能的接受度(杜严勇,2015)。

(3)评估人工智能系统的规范和价值的有效性,即是否与人类社会的伦理相一致和兼容。由于设计的目的和应用导向,与人类兼容的人工智能系统应确立利他主义、不确定性、考虑人类三项基本原则。信任作为人—机交互的动态变量,功利主义和绝对主义导向会做出不同的道德选择,需要建立使用者对人工智能的信任,通过第三方评估组织需要界定价值一致性和相符性标准,评价主体的价值对接问题。

对机器人进行伦理设计旨在寻求实际的解决办法以保障机器的行为是在人类可承受范围

①第一定律:机器人不得伤害人类,或者目睹人类将遭受危险而袖手不管;第二定律:机器人必须服从人给予它的命令,当该命令与第一定律冲突时例外;第三定律:机器人在不违反第一、第二定律的情况下要尽可能保护自己的生存;第零定律:机器人必须保护人类的整体利益不受伤害,其他三条定律都是在这一前提下才能成立。

之内，并将研究拓展到不同机器之间的道德规则传递（Anderson和Anderson，2007）。现将机器伦理用于实践的应用程序主要包括基于“伦理原则”和“实践原则”的两种形式，前者以Michael Anderson设计的基于边沁功利主义的“Jeremy”程序和基于罗斯显见义务（prima facie duties）的“W.D.”程序为代表，后者以McLaren设计的“说真话”（truth-teller）为代表（Anderson和Anderson，2007；于雪和王前，2016）。从技术领域通过程序来指导伦理决策，可以优化机器人在无意识的情况下选择道德的行为过程。设计过程中强调人工智能的人文化（humanitas），创造出有道德的人工智能体。这样，人类就可以与机器人建立在一定信任程度基础之上，通过逻辑性的命令保障机器人的伦理行为被约束在一定的框架范围内，使机器人的行为最大化的符合人类社会的道德规范。

（四）人工智能社会伦理

目前，就AI应用问题还未达成更好的社会契约，人们质疑其快速发展会给人类带来极大的危险，衍生出更多的问题。国内外聚焦在责任与安全、隐私保护等方面分析人工智能应用产生的社会后果，探讨如何善用人工智能造福于人类。

1. 责任与安全

国外关于机器人伦理研究中争论的一个核心问题就是：究竟谁应该对机器人的行为承担责任?联合国教科文组织与世界科学知识与技术伦理委员会（2015）提出了两种对策，一是由智能机器人承担责任。早有学者以具体的APACHE系统为例，阐述机器也许会承担一些道德责任（Friedman和Kahn，1992），因为没人能够承担机器人故障的责任（Matthias，2004）。为此，已有开展赋予机器人道德责任问题的研究，提出使用某种道德测试来裁决机器人是否需要在法庭上负责。但也有学者认为，机器人没有能力承担全部道德责任（段伟文，2017），因为计算机的自主意识性与人类期望的差距仍然较大。在以人为中心的伦理框架下很难对机器人的行为做出有效界定，为此，在机器人设计中应提倡保证算法透明性的披露方式，以利于人类对机器所承担的道德责任保持清晰的理解。

另一种对策是让所有参与机器人发明、授权和分配过程中的人来分担责任。算法无法预见机器人在与人类相处过程中所出现的全部可能的行为，不可能对隐藏在机器人动作中的因果链进行完全掌控。因此，当问题出现时，程序员不会承担全部但也不能完全免责。为了避免在事件发生时没有人承担全部责任，应该推广使用保险制度，能够让所涉及事件的所有人共同承担。并不断完善当前的法规来填补“责任空白”，为责任归属问题找到解决途径。

2. 社会就业结构

人工智能的广泛应用将人类从危险、枯燥和困难的任务中解放出来，人类在分享AI带来巨大物质财富和生活便利的同时，也承担了巨大的心理压力，如AI是否会造成从事劳动力密集型、重复性、程序化等领域的工种的大面积失业，影响人们的收入和福利，产生新的贫富差距和社会分化，从而会引起社会危机，造成社会安全程度降低和引发动荡（闫坤如和马少卿，2018）。很早就有学者反对将机器投入工作中，认为自动化技术将会引发人类学习动机的下降，使人类丧失个性，造成“人脑的贬值”，从而把社会变得机械化。麦肯锡的报告预测，因技术的进步和AI的普及，到 2030 年全世界将有3.9亿人会更换工作，有8亿人会失业。因此，未来因AI发展而引发的社会稳定问题也同样严峻。

虽然机器的感觉、运动、计算都将会远远超过人类，但人类和机器还处于非对称交互阶段，仍然存在着交互的时间差。莫拉维克悖论（Moravec’s paradox）认为，人类独有的高级智慧能力只需要非常少的计算能力，但是无意识的技能和直觉却需要极大的运算能力，这正是机器在短期无法比拟与模仿的。机器和人类的发展都是建立在对环境的改造和认知上，环境结构化的高

低在一定程度上决定了机器对人的替代可能性，虽然未来一部分职业会被机器人取代，但是技术的发展也会改变现有的社会就业结构，并产生一些新的行业。

智能化生产追求的不是简单、异化的“机器换人”，而是采用机器进行柔性生产，重新回到“以人为本”的组织生产模式，重视人在社会价值创造中的主体作用，实现企业生产运营效率的提升，本质是实现“人—机”协同。因此，AI不会从根本上冲击人类的工作岗位与就业。对比第一次工业革命，蒸汽机的使用取代了许多人，但创造了更多的就业机会，社会整体失业率反而稳定下来。当然并非所有人都有能力迈过技术性和社会性壁垒，这会削减劳动力在与资本谈判中的议价能力，未来社会可能会出现“无用阶级”，引发新的社会问题。

3. 生活中的人工智能伦理

（1）隐私问题

算法需要大量的优质数据支撑，各种活动之间的数据交换变成了新的价值创造源，个人的数据容易被机构的主动收集和使用，处于相对被动的，这会削弱个体对个人数据的控制和管理。这些数据如果被机构披露将对个人的隐私产生影响，因此需要在深度学习过程中加强对个人隐私的保护。

隐私问题激发了更为基础的伦理问题的讨论，其实质是数据滥用和数据侵权的问题，发展到极致可能导致数据巨机器（data megamachine）的出现。数据主义成为了数据巨机器的意识形态，它主张绝对的数据共享，这可能有助于提高社会的运行效率的，但是一旦数据巨型机获得了自由，就有可能造成人与数据关系的破裂、人的自由的丧失，从而导致“楚门效应”[①]的产生。

为了防止隐私问题的扩大和恶化，需要从伦理和技术两方面入手。建立以人的权利为中心的数据伦理，强调算法的透明性，消除数据孤岛，提倡有规范的数据共享，防止数据滥用，旨在消除数据主义对数据自由的崇拜。在技术上，目前已经有诸如匿名化、差别化隐私、决策矩阵等工具的应用来实施隐私保护。“人工智能+区块链技术+量子技术”在未来也可能会提供更加有效的解决方案，但在此之前仍需要政府、企业和民间团体等相关主体共同努力，防止由于数据大规模泄露而产生的隐私问题，重建人在大数据时代的主体地位，建构人与技术、人与数据的自由关系，消除机械论世界观的不良影响。这些思路各有利弊，目前来看仍是一个需要展开探讨的开放性问题。

（2）其他方面

人工智能放大了信息处理和知识加工的能力，人参与社会互动的范围和频率可能减小，人类需求与知识之间的关系变得越来越间接。信息和知识的冗余让人陷入选择困境，甚至会反过来支配人的需求。在司法、医疗、指挥管理等重要领域，研究人员也开始探索人工智能在审判分析、疾病诊断和对抗博弈方面决策上应遵循的规范（Sikchi等，2012；Alzou'bi等，2014）。

在法律领域，人工智能的快速发展需要人们重视资源优化配置与价值观负载的制度匹配，因此亟需通过立法来规范人工智能技术。一是研究并确立因人工智能行为而产生的责任制度和约束机制，这样才能在判决中明确人工智能对第三方造成损失或者违反制度而承担的相应法律责任；二是需要建立一套道德规范，促使技术伦理由隐性向显性转化，包括建立和编写智能系统或机器人取代人类活动所适用的基本规则（Ferraz和Del Nero，2018）。

在政府管理方面，人工智能技术在推动社会进步、提升政府组织治理效率的同时，也造成了政府部门与社会的“技术鸿沟”。政府的监管能力跟不上技术的发展，使得技术的发展在没有告知、分析、审查和监管的情况下继续发展，进而损害了公民利益。算法可能会继承人类社会传

①“楚门效应”的实质是，消费者在毫不知情的情况下其自主权遭到侵犯。在楚门世界，除了楚门，人人都是演员；在数据机器里，人人都是楚门，人人都是演员，无人是自己。

统行政管理的冗余、歧视等现象而出现非理性决策。因此，人工智能对政府治理理念、政府技术能力、政务流程、传统行政伦理提出了挑战。

## 五、国家、企业对人工智能伦理的态度

为了更好地促进全世界人工智能产业发展，应将人工智能监管纳入战略考虑，相关的法律、伦理、监管配套制度都应并行建立。可追溯制度的建立才能让机器人的行为及决策全程处于监管之下，只有这样人类才能占据主动权或者事后进行全面追踪调查的权利。

（一）国家战略部署

近年来，世界各国都从国家战略层面发布了人工智能发展的指导文件（如表2所示）。例如，2016年，美国将“理解并解决人工智能的道德、法律和社会影响”列入国家人工智能战略，同时对人工智能从业者和学生加强道德伦理教育，并成立相应的管理机构，负责跨部门协调人工智能的研究和发展、提出技术和政策建议、监督各部门的人工智能技术研发，进一步促进公平与正义。同年，英国也探讨了人工智能所带来的一系列潜在的伦理和法律挑战，尝试寻找能够实现社会经济效益最大化的途径，指出应建立人工智能委员会来应对机器人技术带来的对社会、伦理和法律的影响。日本也制定了机器人应用部署问题的管理方针，包括建立中心数据基地来存储机器人对于人类造成伤害的事故报告。

**表 2　世界主要各国的人工智能国家战略文件**

| 国家 | 时间 | 人工智能国家战略 |
|---|---|---|
| 德国 | 2013年4月 | 德国工业4.0战略 |
| 日本 | 2016年1月 | 超级智能社会5.0 |
| 美国 | 2016年10月 | 国家人工智能研究和发展战略计划 |
| 英国 | 2016年10月 | 人工智能：未来决策制定的机遇与影响 |
| 中国 | 2017年7月 | 新一代人工智能发展规划 |
| 俄罗斯 | 2017年7月 | 数字经济列入俄罗斯2018—2025年战略 |
| 法国 | 2018年3月 | 国家人工智能战略 |

资料来源：本文根据相关文献整理。

中国对人工智能伦理问题也高度重视，并于2018 年开始组建人工智能伦理专委会。在中国发布的战略规划中，“人工智能伦理”这一字眼出现了15次之多，侧重从打造平台、构建市场、激励创新等方面进行了阐述，并提出了“2025年初步建立人工智能法律法规、伦理规范和政策体系，2030年建成更加完善的人工智能法律法规、伦理规范和政策体系”的建设目标，但是对于监管原则、监管体系和相关监管机构的建设还没有提及。

2017年，人工智能滥用、算法公平、人工智能伦理、人工智能监管和责任等进入更广泛的公众讨论视野，相应的标准和规范相继出台。欧盟签署《人工智能合作宣言》，共同面对人工智能在伦理、法律方面的挑战，成立统筹人工智能监管的政府机构，专门研究与机器人和人工智能相关的法律问题，并呼吁制定“机器人宪章”。联合国发布的《关于机器人伦理的研究报告》中提供了物理形态下人工智能系统全新路径研究的新视角，提出了人工智能发展的基本原则，包括保障人类利益和基本权利、安全性、透明性、推动人工智能普惠和有益发展。联合国犯罪和司法研究所（NUICRI）在海牙成立第一个联合国人工智能和机器人中心，教科文组织和世界科学知识与技术伦理委员会（COMEST）联合发布了《关于机器人伦理的初步草案报告》（2015），其中讨论了在机器人的制造和使用过程中产生的社会与伦理问题，并列举了相应的措施：数据和隐私保护、构建机器人设计者与机器人之间的责任分担制度、在实景中对机器人进行测试、建立

全新的针对机器人的保险制度、智能机器人的退出机制，并进一步提出在机器人及机器人技术的伦理与法律监管中确立可追溯性，保证机器人的行为及决策全过程处于监管状态。

（二）企业社会责任

由于技术发展和利益诱惑带来的驱动效用，国内外知名的互联网企业如Microsoft、Google、Facebook、百度、腾讯等不断加大对人工智能领域的投资，预计未来10年内将创造巨大的财富价值。但人工智能技术的发展可能最终导致世界所依赖的各种机器为数据和算法所驱动且不受伦理或哲学规范约束。面对公众的担心和忧虑，科技巨头们也开始采取行动，将伦理考量纳入到企业社会责任框架中。

Microsoft阐述了包括公平、可靠性、保护隐私、包容性、透明性和可责性6项人工智能系统原则，确保人类的利益不受损害。Google还明确列出了“不会追求的人工智能应用”。Facebook在经历数据泄露丑闻之后也开始成立人工智能伦理团队，负责防止人工智能软件中的歧视。从其官网可以发现，正在招聘人工智能政策、伦理、法律等方面的人员，表明其开始重视人工智能伦理相关的工作。

这些企业在人工智能伦理上采取的行动，归纳起来主要包括：一是开展人工智能伦理与社会研究，将人工智能与人文社会科学结合起来，研究人工智能对社会、法律、组织、民主、教育和道德的深远影响，这是科技企业负责任研发与创新的体现；二是提出人工智能价值观，美国、英国、欧盟等致力于达成国家层面的人工智能伦理准则，未来生命研究所（FLI）主导提出了“阿西洛马人工智能原则”、IEEE希望推动形成行业层面的人工智能伦理共识，谷歌、微软等科技公司提出建立全球性参与平台（global participatory platform）的设想，构建数据和算法模型构成的公共商店（Helbing，2015），提出企业层面的人工智能价值观以赢得公众的信任；三是成立人工智能伦理委员会，通过行业实践，建立健全监管机制，配套法规制度，明确监管主体，树立具有行业特色的行动规范。

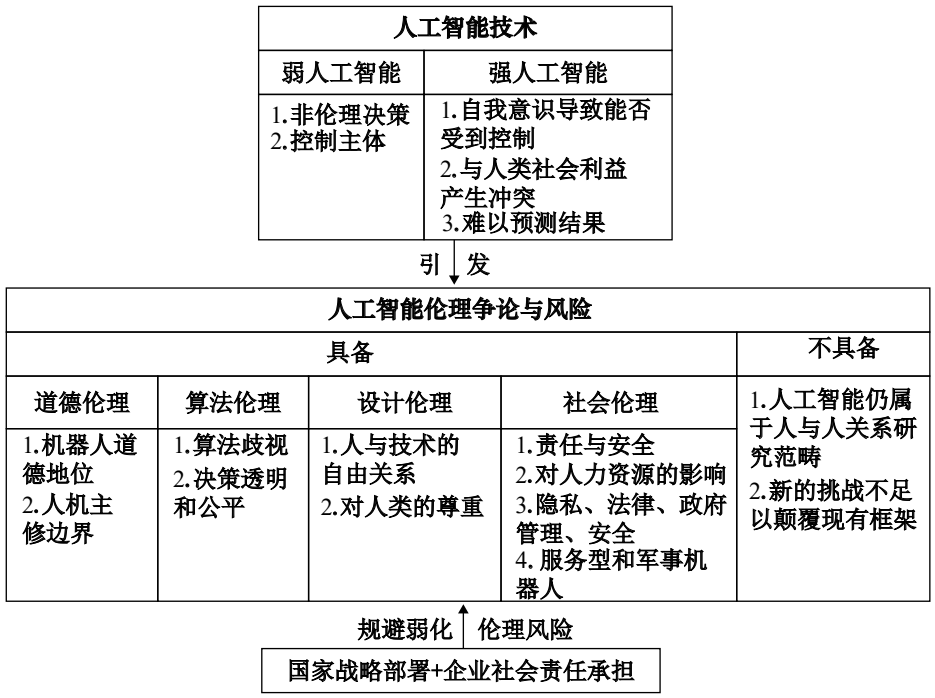
## 六、结论及展望

本文选取影响因子较大的国内外期刊，回顾了1997—2018年期间与人工智能伦理相关的研究文献，补充完善了现有的人工智能伦理研究框架，如图3所示。无论是人工智能应用在现有伦理基础上产生的问题，还是影响人类行为引发的全新伦理问题，弱人工智能和强人工智能上的不同表现引发了伦理研究的必要性和迫切性。二者是人工智能发展过程中的两个阶段，人类对人工智能都应持有一定的忧患意识和风险意识，对前者应该大力研究和发展，对后者应该加以限制。虽然个别学者认为人工智能还不具备伦理意识，但按照目前的研究成果与趋势，人工智能的伦理意识只是时间和技术问题。因此可以认为人工智能在“强弱”上的差别是其是否具备伦理属性的分水岭。

人工智能伦理研究需要建立新的伦理范式，从人工智能伦理的普遍性问题（道德哲学、道德算法）出发，分析不同领域引发的伦理问题（社会伦理），提出伦理问题的解决路径（设计伦理、国家和企业的制度）。为了确保人类与人工智能系统和谐相处，人类已经在理论突破、立法研究、伦理准则、安全标准、监管体系等方面取得了一定成果，但是对人工智能伦理的研究还不够系统、完善，未来可以从以下角度进一步探讨和思考：

（一）重视全球情景下的人工智能伦理体系研究

伦理是具有情境性的，在一个环境下能正常接受的伦理，当迁移到另一种情境可能就会变得难以接受。人工智能的发展是一个全球性的过程，对现行国际秩序将在安全合作、可持续发




资料来源:本文整理所得。

**图 3　人工智能伦理主要研究框架**

展和创新治理上产生深刻影响。因此,解决人工智能伦理的跨情境问题是未来需要思考的重点之一。以Wallach为代表的学者从个体提出了人工智能伦理必要性的研究,引发了国际社会的广泛关注。不同的组织、企业和国家也分别提出了各自的人工智能伦理准则,这些准则的建议已基本达成共识。未来需要从全球的层面来建立人工智能伦理体系,即沿着"个体→组织(国家)→全球(全人类)"的路径对人工智能伦理开展"意识→准则→体系"的研究。

虽然不同国家和组织提出了人工智能伦理的战略指导,但更广泛的信息交流和参与程度仍相对停滞,在某种程度上是割裂和分散的,呈碎片化状态,没有形成统一的范式,不足以形成全局性的行业共识和最佳实践。制定政策、法规或道德和操作标准,需要把技术进步纳入到一个全球视域的、生态可持续的、政治上公平的社会愿景中,服务于全人类的利益而不是国家主义或者民族主义。

怎样建立全球范围内统一的人工智能伦理体系具有非常重要的研究价值。仅考虑道德伦理、法律法规和社会影响框架(ELSI)是不够的,文化在预测未来技术的发展轨迹上具有关键性的作用。不同区域的不同政治制度、文化制度和社会意识形态看待人工智能伦理的标准也不相同,文化多样性应该体现在互动之中,在文化的冲突、融合、发展中促进人类意识的趋同性。这就是全球范围内的文化差异应该成为人工智能伦理讨论的核心的原因所在。

在全球范围内深化人工智能的国际合作,搭建关联政府、企业、公共组织以及个人的平台来促进人工智能政策、技术和文化间的交流与互动,将科技公司自发形成的人工智能原则转变为共识并用以形成对人工智能创新及应用的伦理约束。将伦理问题纳入ESG(环境、社会和治理)框架,从人类命运共同体的高度构建全球化的人工智能治理新秩序,共同探讨全球通用的人工智能伦理法则,人工智能价值观才能发挥其作用。

（二）重视人工智能伦理研究对技术研究的前瞻性

在新兴技术飞速发展的背景下，由于缺乏相应的监管机制、法律规范，伦理问题的研究往往滞后于技术的发展（如基因编辑婴儿），相关的社会反应出现一定的延迟也成为不可避免的现象。目前对人工智能伦理展开研究的主体很多，但是从哪个阶段开始考虑人工智能的伦理、法律及社会影响尚未形成统一认识，人文社科研究者和政策法律群体认为应该从人工智能的基础研究阶段开始考虑伦理、法律、社会影响，而科学家、企业家和技术人员则认为从产品化、服务化之后社会使用和实施阶段开始。社会对人工智能的关注缺乏整体性的思考和讨论，目前仍集中在技术与经济方面，就伦理挑战和社会问题而言，这些讨论更多地还停留在理论研究，尚未达到公共政策层面，而且讨论的重点多集中在“强人工智能”这样相对遥远的议题。

前瞻性的探索非常重要，伦理研究缺乏前瞻性，就会导致政策的真空。高校和研究机构开展前瞻性的科技伦理研究，为相关规范和制度的建立提供理论支撑，明确人工智能技术研发应用的合理边界，这既包括人工智能的技术伦理研究，也包括“人—物”、“人—机”、人际与环境之间关系的探索。“人—机”关系不能仅停留在人与机器、机器人的层面，更应包括社会运作的机制、机理，需要从伦理学的视角来促进人文科学与自然科学的统一，构建同时具备理论探索和实践指导的伦理价值体系。人们设计制定人工智能道德规范，并积极反思人工智能与人类之间的关系，但是考虑人工智能对未来社会影响的不确定性，还需要全世界和各国政府的共同努力，制定人工智能开发和应用的伦理规范和政策方向，采取一些新的监管和认证程序，分别从伦理、政策、法律三个层面推动各项政策的落地。建立保障人工智能健康发展的伦理道德框架和法律法规，重点需要建立人机协作的伦理框架，完善立法和明确人工智能法律主体，在弱人工智能的细分领域加快制定相应法规，为强人工智能的到来奠定法律基础。

（三）重视多利益主体的伦理角色塑造

在人工智能研发和政策的应用过程中始终要围绕以人为本的核心，满足人类全面发展的需求，促进社会的公平和可持续发展。人工智能伦理研究将面对智能信息技术发展与跨学科的共同挑战，正确认识人工智能伦理，引导技术研发方向，注重源头设计与开发，在技术推进中注入人文理性，这需要哲学家、艺术家以及社会科学各界从不同视角审视伦理问题。在技术变革中积极参与，及时发现技术当中隐含的道德议题、社会议题，向科学界、技术界和企业界传递他们的观点，构建开放的公共研究平台进行多领域交叉研究。

多种类型的主体共同参与治理，需要各国政府与社会各界从人类命运共同体的高度予以关切和回应。以政府、企业、社会机构、行业组织、社会大众等社会各界需要共同参与来塑造人工智能的社会角色，确立行业标准，建立健全相关法律法规，采取统筹监管和分散监管的方法，改变经济、政治和教育系统的优先顺序，使人类在与人工智能的竞争中保持和谐共处。在这样一种多利益相关方合作的模式之上，人们才能最大程度地理解并控制人工智能的伦理和社会影响，在人类社会进步和技术创新之间实现协同效应。

（四）除了自然和社会科学路径，还需要从“科学发展观的伦理学”角度来理解人工智能伦理问题

人工智能发展的过程会对现有伦理学构成挑战，同时也会产生更多的未知新伦理问题，内含的技术和社会双重属性及其矛盾将会越来越突出。解决人工智能引发的各种伦理问题，需要树立“发展的伦理和伦理的发展”观点，通过自然科学和社会科学的路径来认识和解决人工智能伦理问题，人类更需要在机器人广泛使用的社会中转变当前的工作观念。为了确保人类精神在一个与人工智能并存的世界中同步发展，需要改变哪些传统观念？如何克服观念转变过程中的障碍和困难？新的伦理观念是否能匹配新的社会形态？这些问题需要人类建构可持续发展的发展伦理和基于健全理性的科技伦理来回答。

## 主要参考文献


[1]杜严勇. 关于机器人应用的伦理问题[J]. 科学与社会，2015, (2): 25-34.

[2]段伟文. 人工智能时代的价值审度与伦理调适[J]. 中国人民大学学报，2017, (6): 98-108.

[3]Ferraz S, Del Nero V. 人工智能伦理与法律风险的探析[J]. 科技与法律，2018, (1): 19-24, 31.

[4]Kurzweil R著, 李庆诚、董振华、田源译. 奇点临近[M]. 北京: 机械工业出版社, 2011.

[5]李伦. “楚门效应”: 数据巨机器的“意识形态”——数据主义与基于权利的数据伦理[J]. 探索与争鸣，2018, (5): 29-31.

[6]刘伟, 赵路. 对人工智能若干伦理问题的思考[J]. 科学与社会，2018, (1): 40-48.

[7]莫宏伟. 强人工智能与弱人工智能的伦理问题思考[J]. 科学与社会，2018, (1): 14-24.

[8]苏令银. 当前国外机器人伦理研究综述[J]. 新疆师范大学学报(哲学社会科学版)，2019, (1): 105-122.

[9]腾讯研究院, 中国信通院互联网法律研究中心, 腾讯AI Lab, 等. 人工智能: 国家人工智能战略行动抓手[M]. 北京: 中国人民大学出版社, 2017.

[10]闫坤如, 马少卿. 人工智能伦理问题及其规约之径[J]. 东北大学学报(社会科学版)，2018, (4): 331-336.

[11]于雪, 王前. “机器伦理”思想的价值与局限性[J]. 伦理学研究，2016, (4): 109-114.

[12]张玉宏, 秦志光, 肖乐. 大数据算法的歧视本质[J]. 自然辩证法研究，2017, (5): 81-86.

[13]赵汀阳. 人工智能的自我意识何以可能?[J]. 自然辩证法通讯，2019, (1): 1-8.

[14]Alzou'bi S, Alshibly H, Al-Ma’aitah M. Artificial intelligence in law enforcement, a review[J]. International Journal of Advanced Information Technology，2014, 4(4): 1-9.

[15]Anderson M, Anderson S L. Machine ethics: Creating an ethical intelligent agent[J]. AI Magazine，2007, 28(4): 15-26.

[16]Boden M, Bryson J, Caldwell D, et al. Principles of robotics: Regulating robots in the real world[J]. Connection Science，2017, 29(2): 124-129.

[17]Bostrom N. Superintelligence: Paths, dangers, strategies[M]. Oxford: Oxford University Press, 2014.

[18]Danielson P. Artificial morality: Virtuous robots for virtual games[M]. London: Routledge Press, 1992.

[19]Davies J. Program good ethics into artificial intelligence[J]. Nature，2016, 538(7625): 291.

[20]Diakopoulos N. Algorithmic accountability: Journalistic investigation of computational power structures[J]. Digital Journalism，2015, 3(3): 398-415.

[21]Edelman B, Luca M. Digital discrimination: The case of Airbnb.com[R]. Working Papers 14-054, 2014.

[22]Executive Office of the President. Big data: A report on algorithmic systems, opportunity, and civil rights[M]. Washington: The White House, 2016.

[23]Greenwald T. Does artificial intelligence pose a threat?[N]. The Wall Street Journal, 2015-05-10.

[24]Heersmink R. Ghost in the machine: A philosophical analysis of the relationship between brain- computer interface applications and their users[D]. Netherlands: University of Twente, 2009.

[25]Helbing D. Big data society: Age of reputation or age of discrimination?[A]. Helbing D. Thinking ahead - essays on big data, digital revolution, and participatory market society[M]. Cham: Springer, 2015.

[26]Johnson D G. Computer systems: Moral entities but not moral agents[J]. Ethics and Information Technology，2006, 8(4): 195-204.

[27]Lei R P, Zhai X M, Zhu W, et al. Reboot ethics governance in China[J]. Nature，2019, 569(7755): 184-186.

[28]Lindsay R K. Complexities of the mind at work.(Book Reviews: What computers can't do. A critique of artificial reason)[J]. Science，1972, 176(4035): 630-631.

[29]Malle B F. Integrating robot ethics and machine morality: The study and design of moral competence in robots[J]. Ethics and Information Technology，2016, 18(4): 243-256.

[30]Moor J H. The nature, importance, and difficulty of machine ethics[J]. IEEE Intelligent Systems，2006, 21(4): 18-21.

[31]Murphy R, Woods D D. Beyond Asimov: The three laws of responsible robotics[J]. IEEE Intelligent Systems，2009, 24(4): 14-20.

[32]Peter K, Anthonie M. Reply to critics[J]. Techné: Journal of the Society for Philosophy and Technology，2002, 6(2): 34-43.

[33]Searle J R. Minds, brains, and programs[J]. Behavioral and Brain Sciences，1980, 3(3): 417-424.

[34]Townley C, Morrison E, Yeung K. Big data and personalized price discrimination in EU competition law[J]. Yearbook of European Law，2017, 36: 683-748.
[35]Veale M, Binns R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data[J]. Big Data & Society，2017, 4(2): 1-17.
[36]Wallach W, Allen C. Moral machines: Teaching robots right from wrong[M]. Oxford: Oxford University Press, 2010.

# How to Comprehend the Ethical Conflict of Artificial Intelligence? A Literature Review and Prospects

Xie Hongming[1,2,5], Chen Liang[1,3], Yang Yingnan[4]
（1. *School of Management*, *Zhejiang University of Technology*, *Hangzhou* 310023,*China*；2. *School of Management*, *Guangzhou University*, *Guangzhou* 510006,*China*；3. *Zhejiang Financial College*, *Hangzhou* 310018,*China*；4. *Institute of Construction management*, *Zhejiang University*, *Hangzhou* 310058,*China*；5. *Institute of Chinese SMEs*, *Zhejiang University of Technology*, *Hangzhou* 310023,*China*）

**Summary:** Artificial intelligence (AI), especially robots, is engaged in more and more decision-making in various fields, gradually changing from passive tools to agents of human beings. This has aroused the thinking and worries of all walks of life on AI ethics. It is necessary to establish a new ethical paradigm to extend the ethical norms of human society to intelligent machines. In recent years, the research has made great progress, enriched the existing research content, and enhanced the guidance ability of the ethical research of AI. This paper comprehensively reviews the research progress of AI ethics at home and abroad. On the basis of bibliometric analysis, three different attitudes towards AI, namely, “traditionalist”, “prudent” and “optimist” have triggered the ethical conflict of the “man-machine” relationship. From the perspectives of AI moral philosophy, moral algorithm, design ethics and social ethics, this paper systematically reviews the research results of AI ethics. The state and enterprises (organizations) emphasize their attitudes towards AI ethics from the levels of strategy and social responsibility respectively, and put forward a more systematic and perfect theoretical framework of AI ethics, which is helpful to systematically grasp the existing research results at the theoretical and practical levels. In the future, we need to do further research on the ethical system construction of global scenario conditions, the forward-looking of ethics to technology, the shaping of ethical roles and the ethical concept of scientific development.

**Key words:** artificial intelligence ethics; robert ethics; a review and prospect

（责任编辑：宋澄宇）