单位代码：10166

沈阳师范大学

硕士学位论文

人工智能技术的伦理问题及其治理研究

论文作者：陶思琦

学科专业：科学技术哲学

指导教师：唐丽

培养单位：马克思主义学院

类　　别：全日制

完成时间：2023 年 05 月 15 日

沈阳师范大学学位评定委员会

人工智能技术的伦理问题及其治理研究

中文摘要

人工智能技术是用来实现让机器像人一样思考、学习和处理问题这一目的的具体方法。自人工智能技术诞生之日起，人们对 AI 与人类社会高度融合的想象就从未停止，在长达半个多世纪的发展过程中，人工智能技术向人类展现了它傲人的成绩，但同时也产生了很多的伦理问题，给人类带来了恐慌与误解。将人工智能技术的伦理问题依据影响范围和时间跨度两个维度进行分类，可以发现人工智能技术的伦理问题有显性和隐性的特征，这分别代表着人类的近忧和远虑。在此基础上，考虑到人工智能技术具有跨学科和多领域的特点，本文提出了通过构建完善的伦理治理体系，加强对人工智能技术的伦理约束，促进人工智能技术的可持续发展，确保其不仅仅是为了技术发展和经济利益，而是服务于人类社会的公共利益。

本文运用了马克思主义科技伦理思想和汉斯·尤纳斯责任伦理思想，对人工智能技术伦理问题的产生进行了分析，进一步发现人对技术的理解和运用，是在不断演化中进行的。在显性伦理问题方面，人的目的和手段是产生问题的源头，针对这一部分的治理，需要约束人的行为，通过人向善来引导技术向善，减少技术的负效应。而在隐性伦理问题方面，问题的产生原因则是来自更深层的社会制度和人类发展的历史。针对这一部分的问题，需要法律、制度和全民教育的底层构建，促进人工智能技术的良性发展。

通过文献分析法、逻辑与历史相统一法和对比分析法对人工智能技术的伦理问题进行分析与总结，利用马克思主义科技伦理思想和汉斯·尤纳斯责任伦理思想作为伦理治理的理论基础，提出以下治理措施：首先，通过明确发展“有担当”的人工智能技术、推进人工智能技术的可持续发展和构建智能联结的社会为伦理治理的使命，提出坚持以人为本和坚持和谐共生的人机关系的底层原则。其次，在确立原则的基础上提出 6 个伦理治理准则，即安全和不伤害、公平和正义以及公开和透明，明确人工智能技术能做与不能做什么的界限。最后，通过构建多元共治的人文环境、提高科学共同体的伦理素养、加强社会公众的科技伦理科普和完善制度建设 4 个具体路径，完成人工智能技术伦理治理的框架搭建。人工智能技术伦理治理是一个复杂的问题，需要政府、企业、学术界和公众的共同参与和努力。期望通过以上措施，能够促进人工智能技术的健康发展，减少人机矛盾，让人工智能技术更好地为人类服务。

关键词：人工智能技术，伦理，治理

Research on the Ethical Issues and Governance of Artificial Intelligence Technology

Abstract

Artificial intelligence (AI) technology is a specific method used to achieve the goal of making machines think, learn, and problem-solve like humans. Since the birth of AI technology, the idea of a high degree of integration between AI and human society has never stopped. Over the course of more than half a century of development, AI technology has shown impressive achievements to humans, but at the same time, it has also raised many ethical issues, causing panic and misunderstanding. By classifying the ethical issues of AI technology based on the dimensions of impact range and time span, explicit and implicit features can be identified, representing the near-term concerns and long-term worries of humanity, respectively. Based on this, considering the interdisciplinary and multi-domain nature of AI technology, this paper proposes to strengthen the ethical constraints of AI technology through the construction of a sound ethical governance system, to promote its sustainable development, and to ensure that it serves the public interest of human society, rather than solely serving for technological development and economic interests.

This article employs the Marxist Ideas on Ethics of Science and Technology and Hans Jonas' philosophy of responsibility to analyze the ethical issues arising from artificial intelligence (AI) technology, and further reveals that people's understanding and use of technology are evolving over time. In terms of explicit ethical issues, the root cause of the problem lies in the purposes and means of human actions, which requires governance to restrict human behavior and guide technology towards benevolence, thereby reducing its negative effects. In terms of implicit ethical issues, the underlying causes stem from deeper social systems and the history of human development. Addressing this issue requires the establishment of legal, institutional, and educational foundations to promote the sustainable development of AI technology for the public good.

Through literature analysis, the unity of logic and history, and comparative analysis, this article analyzes and summarizes the ethical issues of artificial intelligence technology. Utilizing the Marxist Ideas on Ethics of Science and Technology and Hans Jonas' responsibility ethics as the theoretical basis for ethical governance, the following governance measures are proposed:Firstly, by clearly

defining the mission of ethical governance as developing "responsible" artificial intelligence technology, promoting its sustainable development, and building intelligent connected societies, the underlying principles of putting people first and maintaining a harmonious man-machine relationship are established.Secondly, based on these principles, six ethical governance guidelines are proposed, including safety and non-harm, fairness and justice, and openness and transparency, which clarify the limits of what artificial intelligence technology can and cannot do.Finally, through the construction of a diverse and shared humanistic environment, the promotion of ethical literacy among the scientific community, the strengthening of moral science popularization among the general public, and the improvement of institutional construction, a framework for artificial intelligence technology ethics governance is established.Ethical governance of artificial intelligence technology is a complex problem that requires the participation and efforts of governments, businesses, academia, and the public. It is hoped that through the above measures, the healthy development of artificial intelligence technology can be promoted, human-machine conflicts can be reduced, and artificial intelligence technology can better serve humanity.

Keywords:Artificial intelligence technology,ethical,governance

目　录

第一章 绪论

一、选题背景及研究意义

（一）选题背景

人工智能技术是推动全球经济发展的核心技术，同时也是引领第四次工业革命的关键技术。随着该技术在不同领域的嵌入和渗透，在带来了经济与社会变革的同时，也引发了一系列的伦理问题。一方面，人工智能技术正在用更高效、更准确和更便捷的方式服务人类。比如，人工智能技术能够帮助医生诊断疾病、分析病情，提高诊断的准确性和效率，智能家居能够自动控制房屋的温度、照明、安保等，提高居住的舒适度和安全性，这些功能让人类看到了技术智能化的优势。另一方面，人工智能技术也带来了隐私泄露、算法不公、劳动力替代等诸多问题，加剧了社会的不公平现象，放大了技术的负效应，给人类带来了恐慌。短期来看，这些伦理问题给人类的生活带来了不安全、不确定性等困扰，长期来看则是对技术对人的异化和压迫程度不断加深。进一步对人工智能技术产生的伦理问题进行分析，通过将人的责任与良知放在技术研发的前端，通过对人的伦理约束，实现技术发展与人类整体利益相统一，从而确保人工智能技术的合理、公正和安全使用本文旨在提出相应的治理思路，为人工智能技术的可持续发展和社会和谐稳定做出理论层面的贡献。

（二）研究意义

1. 理论意义

通过对人工智能技术伦理问题及其治理的研究，可以拓展伦理学理论。人工智能技术的兴起和快速发展挑战了传统伦理理论对于技术伦理的研究，使得传统的伦理理论难以兼容新技术所带来的伦理问题。因此，需要提出新的伦理理念和范式，丰富伦理学的现实内涵。此外还可以推动伦理实践，通过对人工智能技术进行伦理治理研究可以促进企业和组织建立起可持续的、社会责任感强的技术应用模式，使得技术和社会之间的关系更加和谐，因此该研究具有重要的理论意义。

2. 现实意义

随着人工智能技术的不断发展和广泛应用，各种伦理问题也日益凸显，研究这些伦理问题及其治理，一方面有助于更好地保护人类的利益和尊严，防止人工智能技术对社会造成负面影响。另一方面，人工智能技术的伦理问题涉及多个领域，如哲学、伦理学、法学、社会学、经济学等，需要跨学科的研究与合作，这有助于促进不同学科之间的交流和融合，拓展研究视野，提高研究质量和水平。人工智能技术的伦理问题及其治理研究是一个长期而复杂的过程，需要不断地探

索和完善，这一过程可以推动伦理实践的发展和应用，增进人机关系，促进科技创新，使得技术与社会之间的关系更加和谐，也为我们提供了一个思考未来、构建更加安全、公正、和谐的人类社会一个新的方向。

二、文献综述

（一）国外研究综述

为了确保人工智能技术的合理、公正和安全使用，国外学者对人工智能技术的伦理问题及其治理对策进行了大量的研究。其中伦理问题的研究主要包含人工智能技术与道德的关系研究、人工智能技术的伦理责任研究以及人工智能技术在应用过程中引发的伦理问题研究。在伦理治理对策方面主要有道德嵌入研究、伦理框架设计的研究和伦理准则研究。

1.关于人工智能技术引发的伦理问题研究

人工智能技术的伦理问题研究可以追溯到上世纪五六十年代。当时，人们开始意识到机器在执行某些任务时可能会产生道德问题，促使了学者开始探讨人工智能技术的道德和社会影响。国外对人工智能技术的伦理问题研究主要包括以下几个方面：

一是人工智能技术与道德的关系研究。国外学者在这方面的研究是围绕人工智能技术或机器能否获得道德主体地位展开的。温德尔·瓦拉赫、科林·艾伦共同撰写的《道德机器：如何让机器人明辨是非》一书探讨了设计机器道德即人工道德智能体重要意义。二人在书中对于人工道德智能体的道德主体地位进行了深入的探讨，归纳出三种类型的道德主体。他们依据技术的自主性和人的价值敏感性两个维度，将机器道德的能力从低到高，分为操作性道德、功能性道德与完全道德主体。他们基于对人类道德判断和伦理的本质入手，提倡构建功能性道德，进而使人工道德智能体具备基本的道德敏感性。[①]而詹姆斯·摩尔认为机器人是不可能具备道德主体地位的，他在《机器伦理的本质、重要性与困难》中，从机器人与人的本质区别开始讨论，进一步论证机器无法成为完全道德智能主体，他将道德主体分成四类：道德影响主体、隐性道德主体、显性道德主体以及完全道德主体。[②]前两种道德主体在内涵上倾向于机器对道德的影响以及道德赋予机器的积极作用。显性道德主体是人类某方面能力的补充，应该在其内部进行伦理推理，使机器具备相应的伦理判断和行为能力。而完全道德主体意味着机器与人一样具有伦理道德能力，但是伦理道德能力是人所固有的特质之一，是人类拥有意识的体现，这是机器人无法具有的，因此摩尔认为机器不可能具有和人一样的道

① [美]温德尔·瓦拉赫，科林·艾伦.如何让机器人明辨是非[M].王小红.北京：北京大学出版社，2017：32.

② Moor,J.H.The Nature,Importance,and Difficulty of Machine Ethics'[J].IEEE Intelligent Systems,2006，21（4）：18-21.

德主体地位。与之相反，弗洛里迪和桑德斯是人工智能技术能够获得智能道德主体的支持者，他们将人工智能技术道德主体观扩大到传统伦理学中人类中心主义的主体范畴，将人工物即高度进化的人工技术产品纳入到道德主体的范畴中来。在此基础上对主体的概念进行标准的界定，标准有三：（1）交互性，即主体及其环境可以相互作用；（2）自主，即主体能够在未能直接响应交互的情景下，通过执行内部转换来更改其状态；（3）适应性，即主体能够改变促成状态变化的转换规则。二人认为目前人工智能技术可以具备这三个标准，得出人工智能技术具备成为道德主体的可能这一观点。①

二是人工智能技术的伦理责任研究。在这一部分，主要是围绕人工智能技术承担事故责任的问题展开。例如出现技术故障或意外事件时，应由谁来承担责任，如何界定责任界限等问题。在这一问题的研究中，主要有两种观点，第一种观点，认为人工智能技术是自主行为，责任应该由技术本身承担。这一观点主张赋予人工智能技术法律人格，使其可以像人类一样承担法律责任。持这一观点的人工智能商业化先锋杰瑞·卡普兰，他在《人人都应该知道的人工智能》一书中，将人工智能比作法律范畴中的公司，他提出问题，当人工智能技术成为一种行为主体时，人们还愿意为人工智能的行为负责吗？或者谁来为人工智能技术的行为（如犯罪行为，伤害行为）来负责任呢？卡普兰认为，当人工智能技术拥有了一定的权利时，就应该负相应的责任，就像公司一样在一定范围内拥有法人主体地位的公司拥有获利和分配财产的权利，当遇到亏损时也要承担相应的责任，因此卡普兰认为，为人工智能技术制定相应的法律是非常有必要的。②第二种观点认为，身为设计、制造和使用人工智能产品的人需要承担全部责任。这一观点主张加强对人工智能技术生产者的监管和责任追究。《AI 联结的社会：人工智能网络化时代的伦理与法律》一书中，从事人工智能技术开发工作的高桥恒一将技术研究者负责任的研究和开发，形象地比喻成“刹车”设计，即技术研究者为了能够安心地进行产品设计，需要了解在研究的过程中什么能做什么不能做，强调对技术人员在监管与追责的重要性。③

三是人工智能技术在应用过程中引发的伦理问题研究。这部分是对人工智能技术应用过程中遇到的安全性、公平性、透明度和隐私性等问题进行研。德国计算机科学家 Ipke Wachsmuth 在文章《像我这样的机器人》中，对老年人陪伴机器人产生的伦理问题进行的研究，指出虽然机器人在可预见的未来似乎无法提供

① 转引张浩鹏，夏保华.人工道德智能体何以可行——基于对价值敏感性设计的审视[J].自然辩证法研究，2021，37（04）：37-42.

② [美]杰瑞·卡普兰著.人人都应该知道的人工智能[M].汪婕舒译.杭州：浙江人民出版社，2018：132.

③ [日]福田雅树，林秀弥，成原慧.AI 联结的社会：人工智能网络化时代的伦理与法律[M].宋爱译.北京：社会文献出版社，2020：103-120.

真正的关怀，但当机器人表现得好像它们关心时，它们参与照顾可能就足够了，也就是说，给人一种错觉，即它们对被照顾者的感受和痛苦做出反应。[①]纽约大学教授凯特·克劳福德认为技术出现的伦理问题，与社会和政治影响有很大的关联。比如，在人工智能技术的数据和隐私问题上，她认为数据并不是一种孤立的存在，而是与社会、政治和文化环境紧密相关，因此应该对数据的透明度给予更多的关注。[②]

2.关于人工智能技术伦理治理对策研究

国外对人工智能技术的伦理治理对策研究主要包括以下几个方面：

一是道德嵌入研究。为人工智能技术嵌入道德，是指在技术的设计和开发阶段，对其编写人类道德原则、价值和规范，使其采取有道德的行为决策，避免损害人类利益。因此在这一进路上，瓦拉赫和艾伦提出的机器人伦理设计的三种进路，是非常具有代表性的理论研究。他们基于智能计算器人在不同场景中的应用，提出了自上而下、自下而上和二者混合的三种进路。[③]考虑到人类的道德价值和利益，强调机器人应该如何遵守道德规范。

二是伦理框架设计的研究。俄国学者 Roman V·Yampolskiy 认为人工智能技术发展的真正的问题是如何管理进度以确保自己的安全。因此他在《人工智能安全与保障》一书中，为了防止人工智能权力集中，提出需要一种综合性的人工智能安全框架，并提出 22 个防止原则。[④]而美国学者雷·库兹韦尔在《奇点近在咫尺：当人类超越生物学时》一书中，针对未来科技、人类与机器的交互以及人工智能领域的发展趋势等内容提出了奇点的假设，书中提到面对技术的奇点式发展，预防性原则的作用将越来越小，因此人类面临这样的风险时应该构建人机共存伦理框架。[⑤]

三是伦理原则和准则研究。阿西同的莫夫的“机器人三定律”是可以追溯的关于机器人准则的最早模板，这是阿西莫夫于 1950 年所著的《我，机器人》一书中引言部分的内容，他将这三大法则放在了最明显的位置，突出了法则的重要性。这三条法则分别是，第一定律：机器人不得伤害人类个体，或者目睹人类个

① Wachsmuth I.Robots Like Me：Challenges and Ethical Issues in Aged Care [J].Frontiers in Psychology，2018，9（432）：1-3.

② Kate Crawford.Atlas of AI Power，Politics， and the Planetary Costs of Artificial Intelligence [M].London：Yale University Press，2021：9.

③ [美]温德尔·瓦拉赫，科林·艾伦.如何让机器明辨是非[M].王小红译.北京：北京大学出版社，2017：148.

④ Roman V.Yampolskiy，PhD.Artificial Intelligence Safety and Security [M].LondonC：RC Press，2018：157.

⑤ Kurzweil R.The Singularity Is Near：When humans transcend biology[M].London：Penguin Books，2005：33.

体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，除非该命令伤害到人类；第三定律：机器人必须对自己的生存负责，条件是不与前两条法则矛盾。[①]这三条定律对后世人们在制定人工智能技术伦理准则时起到了重要的指导意义。2019 年 4 月 8 日欧盟发布了“可信赖的人工智能伦理准则”，里面提出了未来人工智能系统应该满足的 7 大原则：人的能动性和监督、技术稳健性和安全性、隐私和数据管理、透明性、多样性和非歧视性以及公平性、社会和环境福祉、问责。[②]

近些年行业也开始出台人工智能伦理准则以及伦理规范条例。比如，2018 年德国公布首份自动驾驶伦理道德标准，部分准则如下：（1）自动驾驶系统要永远保证比人类驾驶员造成的事故少，是该准则的第一必要条件；（2）人类的安全必须始终优先于对动物或其他财产；（3）当自动驾驶车辆发生不可避免的事故时，任何基于年龄，性别，种族，身体属性或任何其他区别因素的歧视判断都是不允许的；（4）在任何驾驶情况下，责任方，无论驾驶者是人类还是自动驾驶系统，都必须遵守已经明确的道路法规；（5）为了辨明事故承担责任方，自动驾驶车辆必须配置始终记录和存储行车数据的“黑匣子”；（6）自动驾驶汽车将对车辆所记录的数据保留唯一所有权，其可决定是否由第三方保管或转发；（7）虽然车辆在紧急情况下可能会自动作出反应，但人类应该在更多道德模棱两可的事件中重新获得车辆的控制权。[③]这是由行业协会将自动驾驶的伦理道德纳入到产品标准体系的重大举措。综上，国外学者在研究人工智能技术如何与人类共存、如何让技术更好地服务于人类方面，提出了一些具体的实践方案，例如人工智能技术伦理框架、智能监管机制等，并且十分关注个人的隐私保护问题。虽然国外在人工智能伦理问题上的研究取得了一定进展，但仍缺乏相应的法律规范来规范技术的发展。

（二）国内研究综述

国内对人工智能技术伦理问题的研究比国外晚，1986 年中国科学技术协会人工智能学会成立，推动了国内在该领域的研究和发展。其研究进展和研究方向主要有人工智能技术与道德关系研究、人工智能技术在应用中的伦理困境研究、结合中国传统道德观念对人－机关系问题的研究和关于人工智能技术伦理治理对策的研究。在这些问题上，国内学者为探讨人工智能技术的未来发展提供了中国式的思路和方案。

① [美]艾萨克·阿西莫夫著.我，机器人[M].叶李华译.南京：江苏文艺出版社，2013：1.

② 央广网.欧盟发布人工智能伦理标准[EB/OL].https://baijiahao.baidu.com/gn/nems/2019/4-11/16304854991701706&wfr.shtml，2019-4-11.

③ 亿欧网.德国公布首份自动驾驶伦理道德标准[EB/OL].https://www.iyiou.com./gn/briefing/2018/5-16/15690.html，2018-5-16.

1.关于人工智能技术与道德关系研究

在这一研究上国内学者最关注的还是人工智能技术在决策、行动和影响人类生活等方面所带来的道德问题，比如如何让人工智能系统具备道德决策能力，使人工智能技术可以通过遵守道德准则、尊重人类的价值和权利，保障人类的利益。闫坤如在《人工智能机器具有道德主体地位吗？》一文中，系统地梳理和总结了国外学者对道德主体的界定和标准，在此基础上对道德嵌入的人工道德主体进行伦理构建，在文章中提倡将“善”的理念在设计之初就嵌入到人工智能机器中，从而促进该行业的未来的发展。[①]而张浩鹏和夏保华在文章《人工道德智能体何以可行——基于对价值敏感性设计的审视》中，探讨了人工道德智能体获取道德能力等相关问题，试图将人类价值系统嵌入到人工道德智能体的构建之中，尝试运用一种整体性、动态性的实践方法给予人工道德智能体更多维的视角，使得在该技术的研发和设计阶段就考虑到人的价值敏感性。[②]通过这样的方式构建人工道德智能体，解决机器在处理人类问题时遇到的伦理问题。

2.关于人工智能技术在应用中的伦理困境研究

人工智能技术在应用中的伦理困境问题研究。孙波和周雪健在文章《人工智能“伦理迷途”的回归与进路——基于荷兰学派“功能偶发性”分析的回答》中基于荷兰学派的威伯·霍克斯和彼得·费玛斯对于“功能偶发性失常”的分析，提出人们对人工智能技术的产生的恐惧或者盲目崇拜，是一种普遍的失常现象。认为面对人工智能的伦理困境应该保持一种相对积极的态度，因为这是人类对于人工物产生的普遍性的迷思，而非人工智能独有的特性。[③]为研究人工智能技术伦理问题提供一种新的视角。陈小平在归纳人工智能技术的伦理困境时，将存在的伦理风险总结为技术失控、技术误用、应用风险和管理失误四类，这些风险可以分别对应到现实场景中。[④]而王前和曹昕怡从五种容易被人们忽视的人工智能技术隐性伦理责任入手，对人工智能技术产生的伦理困境进行研究。发现这五个隐性伦理责任普遍具有责任主体不十分明晰、问题不十分紧迫、潜移默化却影响深远的特点，因此强调人们应该这部分隐性伦理责任采取足够重视，避免出现相应的伦理风险。[⑤]

3.关于中国传统道德观念对人－机关系问题的研究

① 闫坤如.人工智能机器具有道德主体地位吗?[J].自然辩证法研究，2019（05）：47-51.

② 张浩鹏，夏保华.人工道德智能体何以可行——基于对价值敏感性设计的审视[J].自然辩证法研究，2021（04）37-42.

③ 孙波，周雪健.人工智能“伦理迷途”的回归与进路——基于荷兰学派“功能偶发性失常”分析的回答[J].自然辩证法研究，2020（05）：67-72.

④ 陈小平.人工智能伦理导引[M].合肥：中国科学技术大学出版社，2021：12.

⑤ 王前，曹昕怡.人工智能应用中的五种隐性伦理责任[J].自然辩证法研究，2021，37(07)：39-45.

近几年来国内学者将中国传统道德思想与人工智能技术的伦理问题进行了结合，衍生出带有中国特色的理论研究，丰富了人工智能伦理研究的视域，彰显了我国优秀传统道德文化与时俱进的特点，赋予传统道德文化新的时代内涵。张立文在文章《和合伦理道德论——中华传统道德精髓与人工智能》一文中运用中国传统的伦理道德即和合伦理道德论来说明人工智能技术的发展离不开伦理道德的规制。面对人工智能技术为人类谋福祉的初心与实际应用中给人类带来的威胁二者之间的矛盾，他从中国优秀的传统思想入手，阐释了伦理道德对于人造的智能体生存发展的必要性，最终以建立一种网络空间共同体从而实现人工智能技术在一种更民主、透明、良性开放的环境中为人类的幸福生活谋求发展。[①]而王萍萍在文章《人工智能时代机器人的伦理关怀探析——以《老子》“善”论为视角》中则是将《老子》“上善若水”的他者伦理思想运用到人与机器人的关系之中，倡导了一种人类与机器人共生共感的相处模式，通过建立人类对机器人“不仁”“不敢”“不积”的一种相互尊重的态度，最终达到人与机器人和谐相处的状态。[②]

4.关于人工智能技术伦理治理对策的研究

国内学者对人工智能技术的伦理治理对策研究主要包括以下几个方面：

一是伦理原则和准则研究。2019 年，国家新一代人工智能治理专业委员会在北京发布《新一代人工智能治理原则——发展负责任的人工智能》，这里明确提出了八项原则，分别是和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理。[③]发展负责任的人工智能是本次原则明确的主题，在此基础上结合伦理的原则和手段进一步实现人工智能伦理治理在理论层面和实践层面的结合，从而对该项技术进行有效的治理。2020 年 11 月全国信息安全标准化技术委员会面向公众，针对人工智能技术可能产生的伦理道德问题，公开征求意见。2021 年国家新一代人工智能治理专业委员会发布了《新一代人工智能伦理规范》，内容中落实了伦理规范的总则和组织实施等内容，提出了增进人类福祉、促进公平正义等 6 项基本伦理要求，同时还对人工智能的研发管理等环节提出了 18 项具体伦理要求。[④]

二是伦理框架设计的研究。于雪和段伟文则在《人工智能的伦理建构》中从

① 张立文.和合伦理道德论——中华传统道德精髓与人工智能[J].社会科学战线，2018（08）：34-48+281.

② 王萍萍.人工智能时代机器人的伦理关怀探析——以《老子》“善”论为视角[J].自然辩证法研究，2021（05）：54-59.

③ 中国社会科学网.发展负责任的人工智能[EB/OL].http://www.gov.cn/xinwen/2019/6/17/content_5401006.html，2019-6-17.

④ 中华人民共和国科学技术部.新一代人工智能伦理准则[EB/OL].https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html，2021-9-26.

人工智能的伦理构建入手，研究人工智能技术的伦理框架问题。文章中采取嵌入设计、合理规制、合作管理和多元参与四种方式来实现伦理治理框架的有效运行，四种方式相互配合完善人工智能技术伦理内涵的建设。[①]此外，徐英瑾在传统了道德嵌入对策的基础上，提出了不仅要关注人工智能的“内在”，还要关注承载内在道德的框架。他在文章《具身性、认知语言学与人工智能伦理学》中讲到在以往的人工智能伦理学研究中人们的主要关注点是将伦理规范编写成为机器识得的代码，但这是远远不够的，因为传统的方式没有考虑到代码的语义学理论框架是否在逻辑上与其相匹配，这样做会割裂了伦理学与人工智能技术二者在本质上的联系。因此他将机器伦理学研究的核心问题，即人工智能伦理学和应用于人工智能伦理技术的认知语言学分别比作人工智能的“心智”和“身体”，从而强调关于构建伦理编码的“身体”在其基础地位的重要性。主张学术界和产业界从人工智能的“具身性”角度看待该技术未来的发展，重视“身体图式”对于伦理规则的基础性作用。[②]

三是从伦理治理理论转向实践的研究。吴红和杜严勇在文章《人工智能伦理治理：从原则到行动》中将关注点聚焦在人工智能伦理原则和指南在各主体的积极举措中得到贯彻和实施。两位学者认为要落实人工智能伦理准则的具体落实要从参与人工智能技术发展的主体入手，理清伦理学家、科学家、学术团体、人工智能企业以及政府管理部门的各自职责，促成各主体的分工合作从而达到对人工智能伦理原则到行动的完整过程，也就是“理论”走向“实践”的过程。[③]郭林生和刘战雄认为近些年在欧美国家兴起的“负责任创新”的理念，可以帮助解决人工智能技术发展中的责任问题。他们在文章《人工智能的“负责任创新”》中，论证“负责任创新”理论对于利益相关者进行技术创新的重要性和合理性。[④]这是对人工智能技术健康发展的重要保障，并进一步分析“负责任创新”如何从改变人的观念态度，进而使人工智能技术具有道德的决策和行动能力，最终实现人机的和谐共存。闫坤如在《人工智能机器具有道德主体地位吗？》一文中，系统地梳理和总结了国外学者对道德主体的界定和标准，在此基础上对道德嵌入的人工道德主体进行伦理构建，在文章中提倡将“善”的理念在设计之初就嵌入到人工智能机器中，从而促进该行业的未来的发展。[⑤]而张浩鹏和夏保华在文章《人工道德智能体何以可行——基于对价值敏感性设计的审视》中，试图将人类价值系统嵌入到人工道德智能体的构建之中，尝试运用一种整体性、动态性的实践方

① 于雪，段伟文.人工智能的伦理建构[J].理论探索，2019（06）：43-49.

② 徐英瑾.具身性、认知语言学与人工智能伦理学[J].上海师范大学学报，2017（06）：5-11+57.

③ 吴红，杜严勇.人工智能伦理治理：从原则到行动[J].自然辩证法研究，2021（04）：49-54.

④ 郭林生，刘战雄.人工智能的“负责任创新”[J].自然辩证法研究，2019（05）：57-62.

⑤ 闫坤如.人工智能机器具有道德主体地位吗?[J].自然辩证法研究，2019（05）：47-51.

法给予人工道德智能体更多维的视角，使得在该技术的研发和设计阶段就考虑到人的价值敏感性。[①]通过这样的方式构建人工道德智能体，解决机器在处理人类问题时遇到的伦理问题。

综上所述，国内在人工智能技术伦理问题研究方面有着比较鲜明的政策引导性，例如《新一代人工智能伦理规范》等相关指导性文件中，都明确了中国人工智能技术的发展与应用方向，同时大力培养人工智能技术人才，积极促成国内高等院校与科研机构在人工智能伦理问题方向上的研究，如清华大学人工智能研究院和中国科学院自动研究所人工智能伦理研究中心等。

总的来说，国内外关于人工智能技术伦理治理研究，无论是框架设计还是准则条例都已经有现实落地的成果。但是，全世界在人工智能技术伦理问题及其治理研究方面，普遍存在缺乏公众的参与和理解以及治理机制不完善的问题。在这样的背景下，随着人工智能技术进一步发展和应用，风险和问题也会持续存在，而社会公众对技术的恐慌也将会长期持续。本文的价值在于对人工智能技术伦理问题进行明确的划分，即显性伦理问题和隐性伦理问题，理清问题的产生原因，在此基础上帮助公众认识人工智能技术的本质，从而破除人们对技术的片面认识，树立正确的技术观。同时利用构建的伦理治理框架回归道德对人工智能技术发展的正确引导，为人工智能技术更好地服务于人类提供伦理进路。

三、研究方法与创新点

（一）研究方法

本文是以人工智能技术的伦理问题及其治理为对象进行的基础研究，主要运用到文献分析法、历史与逻辑相统一的方法、比较分析法。

1．**文献分析法：**围绕着人工智能技术进行国内外、线上线下文献的搜索查阅，在大量掌握文献资料的基础上，了解与本课题相关的最新前沿动态和理论创新。研究的成果主要运用于第二章人工智能技术的内涵、发展和分类中。

2．**逻辑与历史相统一的方法：**运用逻辑分析和历史考察相结合的研究方法，对人工智能技术的伦理问题从逻辑的角度，总结归纳出基本特征，研究伦理治理的发展历程，把握人工智能技术的良性发展的趋势和规律。

3．**比较分析法：**通过对比分析国内外人工智能技术伦理准则的研究，找出各国伦理准则中的异同，归纳出各国对人工智能技术未来发展的设想，明确人工智能应该做什么和不能做什么的问题。

（二）创新点

本文的创新点主要体现在：

① 张浩鹏，夏保华.人工道德智能体何以可行——基于对价值敏感性设计的审视[J].自然辩证法研究，2021（04）37-42.

第一，视角创新。本文用哲学思辨的方法研究人工智能技术引发的伦理问题，并结合了马克思主义科技伦理思想和汉斯·尤纳斯责任伦理思想，探讨引发伦理问题的原因，为研究人工智能技术的伦理治理提供解答思路。在结合三种思想的基础上，提出坚持以人为本和人机和谐共生是缓和人机矛盾，引导人工智能技术健康发展的根本原则。

第二，观点创新。本文将人工智能技术的伦理问题依据影响规模和影响时效，分为显性伦理问题和隐性伦理问题，并对问题进行了溯源研究，归纳出人工智能技术存在算法黑箱、工具属性和价值属性失衡、人工智能技术的利益相关者缺乏责任意识、人工智能技术自我意识拟人性与物理构造非人性之间的矛盾、人工智能技术对社交秩序的错构和人工智能技术对劳动力就业的干扰六个根本原因。最后通过构建使命+原则+伦理准则的伦理框架，提出多元共治的人文环境、提高科学共同体的伦理素养、加强对社会公众的科技伦理科普和完善人工智能技术制度建设的具体路径，形成一套完整的伦理治理措施。

第二章 人工智能技术的基本理论

人工智能技术的伦理问题及其治理研究首先要对其进行概念的界定，从词义解析的视角、技术应用的视角和哲学的视角对人工智能的相关概念进行研究，进一步总结人工智能技术的概念。在此基础上对人工智能技术的产生、发展和分类进行系统地总结。研究人工智能技术伦理问题的形成原因和治理措施需要马克思主义科技伦理思想和汉斯·尤纳斯责任伦理思想的支撑，其为人工智能技术的伦理问题及其治理研究提供理论基础。

一、人工智能技术的概述

（一）人工智能技术的概念

人工智能是一个广泛的概念，旨在让机器像人一样思考、学习和处理问题。而人工智能技术则是用来实现这一目的的具体方法。在理论基础方面如果要理解人工智能技术的概念，首先需要深入分析人工智能的含义。从词义的角度来看，可以将其拆解为“人工”和“智能”两部分去认识。“人工”一词在现代汉语词典中的注解是“人为的”，可以与“自然的或天然的”进行对比理解，与自然的发展和演化相比，“人工”是由人类主导的、目的明确的、经过精心设计和制造的，是人类基于对物质世界的客观认识，充分发挥自身能动性的行为。而“智能”又可以进一步分为“智”和“能”两部分去理解，我国古代哲学家荀子在《荀子·正名篇》中写道：“所以知之在人者谓之知，知有所合谓之智。所以能之在人者谓之能，能有所合谓之能”。[①]在其言论中的“智”是指人运用知识的能力，包含着对知识的理解、运用、分析和预测等，是一种高级思维。而“能”则是指人在对于认识，认知和理论的基础上的实际操作能力，所以徒有知识却不会运用就不能称作“能”。荀子的这段话体现了实现“智能”既要求人需要有知识理论和思维能力，同时又必须要有将知识和思维进一步转化为实践的能力。因此，“人工智能”在词义上就是指通过人工手段创造的类人智能。“技术”意为技能、艺术、手艺，后来发展演变成了指掌握一定技巧、方法和知识的一种能力。指的是一系列实践、方法、技能、知识和工具的组合，旨在解决现实问题或实现特定目标。因此，站在词义分析的角度上对“人工智能技术”可以做出的概括性解释，即“人工智能技术”是通过人为的手段，让机器模拟人类的感知、认知、推理和决策等过程，使其能够实现自主学习、自主适应和自主创新的这一目的的科学技术。基于此，人工智能技术就具有无法脱离人类制造与使用的特点，因此可以进一步引申出“技术人工物”的概念，即人类依托于主观世界的尺度，通过技术实践活动而生成的存在物，这一概念也从侧面表明了人工智能技术是以自然为前提的客观

① 林崇德，杨治良，黄希庭.心理学大辞典[M].上海：上海出版社，2003：1704.

存在的实体。[①]

从技术应用的角度来看，人工智能与人工智能技术就是一般和个别，共性与特殊的关系。因此，日本人工智能学会就将人工智能解释为“具有智能的机械”，如今泛指“为机器赋予人类智能”的所有技术、方法和应用的统称。[②]国内学者李开复将人工智能进一步解释为“深度学习+大数据=人工智能”，由于深度学习+大数据的组合能够实现的效果具有近乎无限的可能性，因此也赋予了人工智能技术发展的无限可能性。

从哲学的视角来分析，通过对人工智能与人类智能的共性以及差异性的比较，会更利于人们准确地发现人工智能技术的特点。英国认知科学家玛格丽特·博登在她《人工智能哲学》一书中提到“人的智能是意识的反应，人工智能就是把人的一部分智能活动通过机械化的形式表现出来了。这个外化的过程的具体细节是站在控制论的理论上，实现电脑仿效人脑的部分功能的功能模拟方法。”[③]在这段话中，人工“智能”的产生原理和发展过程与“人类智能”有着本质上的不同。人类智能是意识的、生理的、有机的，而人工“智能”是模仿的、机械的、无机的。如果对这句话做出进一步解释，可以把人工智能技术看作是通过控制论的原理和数学模型构建计算机系统来实现信息的传递，从而模拟人类的智能行为和思维过程，以及对环境的感知和理解，实现自主、智能和灵活等行为的一种技术手段。这种技术的最终目标是让机器具备类似于人类的智能，即具备类似于人类的智慧和思维能力。

以上这些来自不同领域的学者对人工智能概念的解读存在着一些差异，这是由于人工智能需要借助于多种学科的知识和技术才能实现，因此不同学科对其的理解和定义也有所不同。但这并不妨碍人们可以对人工智能概念的本质得到一致性的把握，即人工智能是让机器像人一样思考、学习和处理问题。在对人工智能概念的本质有了一定把握的基础上，就可以对人工智能技术的概念做出总结，即“人工智能技术”就是一种模拟人类智能的技术，它是可以让机器或者计算机系统像人一样进行思考、学习、决策和执行任务的具体方法，是让机器或计算机系统具备类似人类思维能力、推理能力、学习能力与感知力等智能能力的技术。其研究的初衷是改进机器让其拥有类人智慧，其本质是让机器更好地为人类及人类社会服务。

（二）人工智能技术的发展历程

计算机的发展是人工智能技术实现和发展的基础。上世纪 50 年代，人类为

① 阴训法，陈凡.论“技术人工物”的三重性[J].自然辩证法研究，2004（07）：28-31.

② 聂明.人工智能技术应用导论[M].北京：电子工业出版社，2019：18.

③ [英]玛格丽特·A·博登.人工智能哲学[M].刘西瑞，王汉琦译.上海：上海译文出版社，2006：72.

了处理更加复杂的数字运算创造了计算机。在这之后，人们开始探索如何让计算机具有智能，从而发展出人工智能技术。1950 年，被誉为“计算机科学之父”的艾伦·麦席森·图灵在他的论文“计算机和智能”中提出了机器思维的设想“I propose to consider the question , ‘Can machines think ? ’ This should begin with definitions of the meaning of the terms ‘machine ’and‘think’”[①]，并于同年十月发表论文《机器能思考吗》从此奠定了他“人工智能之父”的地位。图灵试想通过一场“图灵测试”来测验机器能否拥有智能，机器智能的测试采用设问的形式，通过向机器提出问题并由机器给出回答，来衡量机器智能的水平。如果受试者无法区分得到的回答是来自机器还是人类，即出现了“误判”，并且如果这种情况出现的比例占测试的 30%以上，那么就可以认为这台机器具有近人类智能。因此，人工智能技术本质上是计算机技术的一种延伸，旨在让计算机具有类似人类思考、学习和解决问题的能力。

1956 年夏天，人工智能一词在美国达特茅斯学院举办的人工智能研究会议上被正式提出。这场会议结束之后人工智能技术迎来了自己的发展元年。在这之后的 60 余年里人工智能技术经历了三次发展高潮。第一次发生在 1950—1960 年，处于这一时期的人工智能技术，其主要路线和范式属于符号主义。符号主义的诞生基于形式逻辑和符号处理的思想，此路线认为人工智能技术是通过建立一系列符号来具体实现的，这些符号可以被计算机程序处理从而实现智能行为。在符号主义的框架下，计算机程序需要事先编写规则和知识库，然后通过逻辑推理的方式来实现问题的最终解决。例如肖、西蒙和纽维尔利用“逻辑推理家”来推理《数学原理》中的 38 条定理，这一阶段人们通过研究还发现了计算机可以在一定程度上理解并使用语言来解决问题。[②]这次高潮期过后人们发现使机器通过简单的逻辑推理解决不了实际问题和复杂问题，因为此时的机器只有解决问题的方法（逻辑推理）而没有关于解决更加复杂问题的具体知识。

1970—1980 年人工智能技术迎来了第二次发展高潮，这个时候的科学家把注意力集中在研究机器的知识系统上面，从而出现了大量“专家系统”（即知识工程）的项目研发。[③]人们提取了特定领域内的专家知识和专门知识并将其转换成类似于知识库的系统让机器学习，使机器具备解决一定问题的能力。但是其弊端也是很明显的，系统中设定的专家知识只能解决特定的问题，如专业性较强的理论型问题，而对于常识性问题的解决能力却很缺乏，可以说在机器获取知识并解决问题这一路径上虽有一定成果但是并未完全成功。1980 年之后，联结主义

① A.M.Turing Computing Machinery and Intelligence[J].Mind，1950：433-460.

② 陈小平.人工智能伦理导引[M].合肥：中国科学技术大学出版社，2021：2.

③ 陈小平.人工智能伦理导引[M].合肥：中国科学技术大学出版社，2021：3.

逐渐成为一种新的技术路线和范式，它区别于传统研究方向的思维训练法成为了当时人工智能领域的新热点。联结主义基于神经网络和分布式表示的思想，认为人工智能技术可以通过构建大规模的神经元网络进行具体实现。神经元网络可以通过学习来自动提取特征，并根据数据进行自适应调整。相对于符号主义，联结主义更加适用于处理模糊或不完备的问题，在其盛行期间诞生了一系列的机器学习算法，包括神经网络、决策树等。

1990 年之后，人工智能技术迎来了第三次发展高潮。随着计算机科学技术的飞速发展，人工智能技术在解决传统的知识问题上也得到了更大的提高。这是因为“我们有了足够海量的数据、强大的计算资源以及更先进的算法。这一代人工智能的重要特征是——基于大数据的深度学习”[①]其发展重点是数据挖掘和大数据处理，同时基于思维训练法的人工智能深度学习（这是基于人脑神经元研究的人工神经网络学习法）研究也取得了重大进展，比如在 1997 年人类与机器人的国际象棋比赛中，AI 棋手“深蓝”战胜了世界棋王从而引起了全世界范围内科学领域的轰动，这一次项目的成功与之后的人机围棋大战有着直接的联系。

此后的 20 年时间，人工智能技术的研究与发展逐渐从专业领域向民生领域渗透和下移，这使得计算机在图像识别、语音识别等领域都取得了很大的进展，并推动了人工智能技术的广泛应用，包括机器人、智能家居、智能交通、医疗健康、金融和游戏等领域。如在机器人领域，人工智能技术的愈加成熟使得机器人具备了更强的感知、决策和执行能力，广泛应用于工业生产、医疗护理、军事防务等等，从而形成了如今“人工智能+”的崭新图景。“人工智能+”，这是人工智能技术与传统行业相结合的新兴领域。2020 年世界人工智能大会上提出的“AI 家园”理念，涉及了多项“人工智能+”的模块，其中主要包含有，人工智能+制造、人工智能+医疗、人工智能+教育、人工智能+交通、人工智能+安防、人工智能+新零售和人工智能+金融等。在这些领域中，人工智能技术的应用主要包括，语音识别、自然语言处理、图像识别、智能推荐等，其中智能推荐系统已经成为很多企业的重要业务，例如电商平台、在线视频平台、社交网络等等，智能推荐系统通过收集用户的行为数据和个人信息，为用户提供更加个性化的服务，从而提高用户的“满意度”和“忠诚度”。如今，人工智能技术正处在新一轮的工业革命浪潮之中，德国、美国等发达国家都将“智能化”写进制造业改革升级的目标之中，国内也紧随脚步在“中国制造 2025”制造强国的战略文件中将智能化制造摆在了突出位置，即“中国制造”转向“中国智造”。人工智能技术+领域的应用研究是对于传统行业转型升级的突破点，也是未来制造业面向智能化发展

① 腾讯研究院，中国信通院互联网法律研究中心，腾讯 AI Lab 腾讯开放平台著.人工智能[M].北京：中国人民大学出版，2017：20.

的应有之意。

（三）人工智能技术的分类

1.按照功能类型分类

人工智能技术按照功能类型分类主要有 5 种，分别是计算机视觉、语音识别、自然语言处理、机器学习和大数据。这些功能是依据人的感官和思维方式进行延伸发展而来的，广泛应用于人的日常生活。

计算机视觉是一种利用计算机和数学算法对数字图像或视频进行自动处理、分析和解释的技术，该技术致力于使计算机和摄像机能够对目标进行分割、分类、识别、跟踪和决策，实现对多媒体数据的自动理解和智能处理，拥有类似人眼一样的功能。[①]例如，在智能交通领域，计算机视觉技术可用于车辆识别、行人识别、交通流量监测等方面。

语音识别是指将人类语言转换为机器可以理解和处理的数字信号的技术。机器学习和深度学习技术的出现，使语音识别得到了迅速发展，并被广泛应用于手机、智能音箱、车载导航等各种智能设备中。

自然语言处理是基于自然语言理解和自然语言生成的信息处理技术，[②]是计算机科学与人工智能领域中研究人类自然语言与计算机之间交互的理论和技术，其主要目标是使计算机能够理解、处理和生成自然语言。最近火爆全球的 ChatGPT 和谷歌 BERT 都是这一领域里程碑式的发展。

机器学习利用统计学方法来对数据进行建模，并利用计算机进行优化和迭代，是一种能够从数据中学习规律并进行预测或决策的算法，这也是实现计算机具有“智能”的根本途径。机器学习致力于研究如何通过计算的手段，利用经验来完善系统自身的性能，机器学习所研究的主要内容，是关于在计算机上从数据中产生“模型”的算法，因此可以说机器学习是研究关于“学习算法”的学问。[③]机器学习被广泛应用在图像识别、语音识别和自然语言处理等技术中，可以说是实现这些技术智能化发展的重要支撑。在自然语言处理领域机器学习应用广泛，比如语音识别、文本分类、机器翻译、情感分析等。机器学习模型可以从大量的语音或文本数据中学习到语言的规律和模式，实现自然语言的处理和理解，进而回答人类的提问或完成人类下达的任务。

大数据或称巨量资料，[④]指的是海量、复杂、高维度的数据集合，并且需要

① 陶映帆，胡鹏飞，杨文明.智能家居中的计算机视觉技术[J].人工智能，2020（05）：30-38.

② 转引王海宁.自然语言处理技术发展[J].中兴通讯技术，2022，28（02）：59-64.

③ 周志华著.机器学习[M].北京：清华大学出版社，2016：1.

④ 济南市大数据局.什么是大数据？

[EB/OL].http://jndsj.jinan.gov.cn/art/2019/8/13/art_39428_3158286.html，2019-8-13.

使用先进的技术和方法进行处理和分析。可以说，大数据支撑了人工智能技术的海量计算，也是其重要的组成部分。生活中，人们常用的移动设备、物联网设备和社交媒体等都是在人工智能技术大数据的分析和处理基础上，实现数据的存储、处理、分析和可视化等等。

2. 按照应用领域分类

人工智能技术的应用领域随着该技术的不断发展也在持续的拓展当中，目前广泛应用于包括医疗、金融、教育、交通、农业、制造业等领域。

人工智能技术+医疗。人工智能技术的应用包括诊断和治疗方案的优化，患者监测和健康预测等方面。人工智能算法可以帮助医生更好地识别疾病和提供更好的治疗方案，从而提高治疗成功率和减少误诊率。比如，用于医疗影像诊断、个性化治疗、健康监测等方面，还可以通过监测患者的数据来预测疾病的发展，从而提早干预，减少疾病造成的损害，提高医疗服务的效率和质量。

人工智能技术+金融。人工智能技术可以应用于风险评估、信用评级、投资决策等方面，提高金融服务的精准度和效率。比如，在风险管理、欺诈检测和交易优化等方面，通过分析海量的数据来预测市场走势和风险，帮助投资者更好地做出投资决策。还可以通过分析交易数据来优化交易策略，提高交易效率和收益率。

人工智能技术+教育。人工智能技术可以用于智能教育和个性化教育。通过对学生学习数据的分析，人工智能可以为学生提供个性化的学习方案，提高学习效率和成绩。同时，人工智能技术还可以辅助教师进行教学管理，为教育教学提供更智能化的支持。

人工智能技术+智能交通。人工智能技术可以实现自动驾驶、交通流量优化等功能，提高交通效率和安全性等。

人工智能技术+农业。人工智能技术可以用于农作物生产管理、农产品质量检测和精准农业等方面。例如，人工智能可以分析土地和气象数据，为农民提供更加科学和精准的农作物种植方案。同时，人工智能技术还可以用于农产品的智能检测和质量评估，提高农产品的质量和安全性。

人工智能技术+制造业。人工智能技术可以提供质量控制和检测、预测性维护、自动化控制和供应链管理的技术支持，帮助制造商提高生产效率、降低成本、提高产品质量等，从而增强市场竞争力。

二、人工智能技术研究的伦理理论基础

人工智能技术研究的伦理理论基础是指在分析和处理人工智能技术引发的伦理问题时需要使用的理论工具。人工智能技术的健康发展需要人的主体性发挥积极作用，需要寻求可持续的发展方式，需要责任伦理指导技术创新，因此在伦

理问题的根源分析和治理路径上需要借用马克思主义科技伦理思想和汉斯·尤纳斯责任伦理思想中的核心内涵作为研究问题和提出对策的理论支撑。

（一）马克思主义科技伦理思想

1.马克思主义科技伦理思想的内涵

马克思主义科技伦理思想是科技伦理的一次革命性变革，是卡尔·马克思和弗里德里希·恩格斯将科技伦理建立在唯物史观的基础进一步发展而来的。其涵盖范畴广泛，对于科技进步和社会发展具有深远的影响，标志着科技伦理的不断革新与进步。其主要的伦理思想包含，技术与道德的辩证关系、技术的人本意蕴以及技术发展应该重视生态与自然。马克思主义科技伦理思想是关于科学技术问题的基本观点和理论。强调科学技术与社会、经济、政治等方面的密切联系，指出科学技术是生产力发展的重要因素之一，是社会生产力的核心内容之一。

马克思主义科技伦理思想主要探讨了有以下3种关系：（1）科技与道德的辩证关系。马克思说：“在我们这个时代，每一种事物都包含有自己的反面”；一方面便显出“不能想象的工业和科学的力量”；另一方面却“显露出衰颓的征象”，“技术的胜利，似乎是以道德的败坏为代价换来的”。[①]马克思和恩格斯揭示了科技与道德的辩证统一关系，指出在我们这个时代（指现代化、工业化的时代）充满了科技进步和工业发展的，但它也伴随着人类道德与伦理的挑战和危机。虽然现代技术、工业等带来了人类的繁荣和便利，但它们也带来了很多负面影响，包括环境污染、资源消耗、人类关系的疏离和脆弱等问题，所有的事物都是复杂的，既有积极的一面，也有消极的一面。（2）科技与人的关系。在马克思主义科技伦理思想中，“人本”意蕴是该思想的根本立场，体现的是对人的深刻关怀。马克思从“现实的人”出发，指出科学技术是人的本质力量的体现，因此现实的人的实践决定了科学的本质与内涵。[②]马克思在阐述有关科学、技术和人类社会关系的基本观点时，一直强调人作为主体的重要性，并认为科技发展的最终目标是促进人的自由全面发展。马克思认为，科学技术是人类社会发展的重要组成部分，而人类是科技发展的主体和决定因素。人在科技活动中不仅仅是被动的接受者，更是科技的创造者和使用者。因此，在马克思主义科技伦理思想中，人本意蕴所强调的是人的主体作用，因此人的主体性被赋予了重要的地位。科技是为人类服务的，必须紧密联系着人类社会的发展和人的自由发展，即“自然科学通过工业日益在实践上进入人的生活，改造人的生活，并为人的解放做准备”[③]。在马克思主义理论的框架下，科学技术的发展应该为社会的进步和人类福祉

① [德]卡尔·马克思，[德]弗里德里希·恩格斯.马克思恩格斯全集（第31卷）[M].北京：人民出版社，1971：588.

② 刘海龙.马克思科学技术观的人本意蕴[J].社会主义研究，2011，197(03)：1-5.

③ [德]卡尔·马克思.1844年经济学哲学手稿[M].北京：人民出版社，2000：89.

服务。这就需要将科学技术的发展与社会生产、社会福祉以及人的自由全面发展紧密联系起来，促进科学技术的良性发展，使其造福于人类。与此同时，技术创新和发展必须基于广泛的自觉主体，尤其是技术工人的积极作用。技术工人应该既掌握先进科技，积极发挥其创造性和创新精神，同时也应该掌握科技的伦理和社会责任。（3）科技与生态的关系。恩格斯强调人与自然和谐一致，在自然面前人不能像是站在自然系统外那样统治着一切，不能不计后果的破坏自然。科技的发展，使得今天的许多地方成为了不毛之地，因此科技与生态的问题同样也是道德问题和制度问题。因此，马克思主义科技伦理思想中指出科学技术的发展不是自然的、孤立的过程，而是与社会的生产关系、生产方式和生产力水平密切相关的。科学技术的发展必须适应社会生产关系的需要，符合人民群众的利益和要求，为实现共同富裕、提高人民生活水平、保护环境、保障国家安全和发展作出积极的贡献。

2. 马克思主义科技伦理思想对人工智能技术发展的启示

马克思主义科技伦理思想强调技术和社会的互动关系，认为科技在社会发展和人类生活中起着重要的作用，但科技发展必须以人类的利益和幸福为核心。他认为，技术应该以人类幸福为导向，为人类服务，而不是盲目追求技术进步和经济效益。他强调技术是人类创造的产物，应该为人类服务，而不是人类为技术服务。所以，人工智能技术的发展应该以人民的利益为出发点，满足人们的需求，改善人类生活和工作的质量，而不是仅仅为了追求经济效益和技术创新，这对人工智能技术的发展具有重要的指导意义。

首先，马克思主义科技伦理思想强调将科学技术的发展与社会生产和人类福祉紧密结合。这为人工智能技术的发展提供了明确的方向。人工智能技术的发展应该是服务于社会进步和人类福祉的，而不是为了某些特定的经济利益或个人利益而发展。

其次，马克思科技伦理思想重视科技与社会伦理的关系。马克思认为科技应该作为社会的重要组成部分，但同时也认为科技发展可能会导致社会乃至人类所面临的各种问题。因此，技术与社会伦理标准应该相协调，科技应该符合环境保护和回报自然的原则。人工智能技术的发展也需要注重其对社会和道德方面的影响，防止产生不良后果。例如，人工智能的普及可能导致就业机会减少，对于文化和人性方面的需求等问题也需加以考虑。

最后，马克思科技伦理思想主张实现各方利益的平衡，以推进社会公正和公平，在人工智能技术的相关应用方面，需要平衡各方利益（如工业资本家利益和工人利益等），推动人工智能技术的发展，发挥其在社会和人类中的积极作用。同时贯彻技术应用以需求为导向，在推进人工智能技术应用的过程中，我们也需

要根据应用需求来开发，实现科技与社会需求的相互促进。

在马克思科技伦理思想的指导下，反思技术的发展能够启示我们对人工智能技术向善的思考。技术发展本身不是目的，而是应该服务于人类社会的发展，这是马克思科技伦理的核心思想。人本意蕴是这一思想的重要内涵，人工智能技术的发展必须坚持以人为本，认真把握人的主体性作用，充分发挥人的主观能动性，推动人工智能技术为全人类的福祉服务，而不是出于利益或控制的考虑而发展。马克思科技伦理思想提供了一个科技发展的理论框架，人工智能技术的发展需要按照这个框架引导和推动对于引导人工智能技术的发展有着积极作用。

（二）汉斯·尤纳斯责任伦理思想

1.汉斯·尤纳斯责任伦理思想的内涵

汉斯·尤纳斯20世纪著名德国哲学家，他的 “责任伦理” 思想，来源于对技术时代人类发展的忧思。尤纳斯认为，“人类必须存在”是责任伦理思想的基础，这种责任产生于这样一个事实，人类有能力预见其行为的后果，并采取措施防止负面结果。受到康德“绝对命令”这种表达形式的影响，尤纳斯也提出了责任伦理的绝对命令，即“如此行动，以便你的行为后果与人类持久的真正生活一致”。[①]换句话说，人类有责任确保人的行为能够促进这个星球的长期可持续性，从而实现人的持久生存。

因此，他的责任伦理就包含了人对自然的责任、人对社会的责任和人对未来的责任，具体内容如下。（1）人对自然的责。尤纳斯认为自然不是人可以肆意利用的附属品，反对以人类为中心主义思想中对自然的认识。如果按照后者观点，认为自然的存在只是为了满足人类的需要，忽视自然的内在价值，那么人与自然之间的关系会不断地恶化，直到人类自己走向终结。（2）人对社会的责任。他强调了人是群体性的这一特点，人处在社会关系中不是孤立的个体，这就意味着人的行为与选择，要在群体的政治秩序中努力寻找平衡，强调了责任伦理思想中整体性的重要含义。（3）对未来的责任。尤纳斯指出未来责任是“对子孙后代的责任”[②]、“对遥远后代的生存及其条件的责任”[③]和“为了人的理念的本体论责任”[④]因此，尤纳斯的责任伦理要求人要为自然负责，为社会负责和为后代负

① [德]汉斯·尤纳斯.责任原理--现代技术文明伦理学的尝试[M].方秋明译，香港：世纪出版集团， 2013：11.

② [德]汉斯·尤纳斯.责任原理--现代技术文明伦理学的尝试[M].方秋明译，香港：世纪出版集团， 2013：35.

③ [德]汉斯·尤纳斯.责任原理--现代技术文明伦理学的尝试[M].方秋明译，香港：世纪出版集团， 2013：36.

④ [德]汉斯·尤纳斯.责任原理--现代技术文明伦理学的尝试[M].方秋明译，香港：世纪出版集团， 2013：39.

责，这一切的根本都是为了满足人类整体的利益与发展。为了能够实现他的伦理思想，尤纳斯强调了预测科学的重要性，也就是对还没有发生的事情，进行充分的科学幻想。他指出“一种关于人类和地球的可预测的未来处境的真理，那些第一位的哲学真理要对其作判断。那些判断将进而作用于今天的行为，通过长远的推断，那些未来处境就从已被辨别的趋势中显现出来，像它们肯定或可能的结果那样随之发生”[①]。

2.汉斯·尤纳斯责任伦理思想对人工智能技术的启示

人工智能技术的责任问题是研究其伦理问题不可绕过的核心，贯穿人工智能技术发展的全过程。尤纳斯的责任伦理思想，为如何负责任设计、使用和管理人工智能技术提供了几个关键的见解。首先，它强调了考虑人类行为长期后果的重要性。在人工智能技术大放异彩的背景下，这意味着仔细思考新技术对环境、人类社会和其他生物的潜在影响，也意味着认识到人类知识的局限性，并对新技术相关的风险持谨慎态度。其次，尤纳斯的责任伦理还强调了承认所有生物的内在价值的重要性。在人工智能技术普及度不断提高的情况下，这意味着确保新技术的开发和部署不会伤害人类或非人类的生命，还意味着承认人工智能技术本质上不是“好”或“坏”，而是可以被用于做有益或有害的技术方式。最后，责任伦理还强调了为人的行为负责的重要性。与传统的“责任”不同的是，尤纳斯的责任伦理更加关注事前责任、对未来行为的前瞻预测性和责任归属主体的整体性，这意味着人需要认识到新技术的开发和部署具有社会和道德影响，必须仔细考虑，同时承认个人和组织有责任确保人工智能技术的使用方式符合伦理原则。

汉斯·尤纳斯的责任伦理思想为人工智能技术的伦理治理提供了宝贵的指导。该思想强调考虑人类行为的长期后果和责任，认可所有生物的内在价值，这为开发和部署人工智能技术发展提供了有担当和道德的方式。比如，在开发和使用人工智能技术的过程中，必须考虑其对人类和自然界的道德影响，这需要对技术行为全过程的规范，适用于规范科学家的研发行为、企业（供应商）的生产和销售行为以及用户（使用者）的使用行为等一系列主体。以确保技术的使用对人类和自然界都有益。随着技术研究和应用的不断深入，产生的技术事故风险会增加，责任问题也会变得更加复杂。因此，责任伦理发挥着重要作用，防止“责任扩散”和“有组织的不负责任”等情况发生，并为人工智能技术伦理治理提供理论依据。

综上，本文在探讨人工智能技术伦理问题的成因和伦理治理的使命与措施部分，运用了马克思主义科技伦理思想和汉斯·尤纳斯的责任伦理思想作为理论支撑。其中马克思主义科技伦理思想的人本意蕴强调人的主体作用，可以帮助人们

① [德]汉斯·尤纳斯.责任原理--现代技术文明伦理学的尝试[M].方秋明译，香港：世纪出版集团， 2013：23-24.

应对技术的风险和挑战时，树立正确的观念和态度，帮助人类正视人工智能技术的发展，并提醒人类在进行人工智能技术的创新时，要遵循自然的规律，尊重生态平衡。汉斯·尤纳斯的责任伦理对人的技术创新行为进行伦理上的约束，使人工智能技术的利益相关者意识到负责任的重要性，从而规劝人们进行负责任的技术创新行为。以上理论的观点，对促进人工智能技术向善发展，更好地为人类服务，以及帮助人类树立正确的技术观，具有良好启示。

第三章 人工智能技术引发的伦理问题

本文依据人工智能技术引发的伦理问题的影响力、话题度和影响的时间跨度，对问题进行了划分，分为人工智能技术的显性伦理问题和隐性伦理问题。

一、人工智能技术引发的显性伦理问题

人工智能技术的显性伦理问题，是指贴近于人类社会生活且时有发生，对人的影响有直接表现，人们可以清楚感知得到并且时常讨论的那部分伦理问题。人工智能技术所带来的显性伦理问题依据呈现形式概括为技术滥用、技术失控和管理缺失。这些问题与人的生命、健康和财产安全息息相关，并且随着人工智能技术的发展，问题数量还在不断增加。

（一）技术滥用

人工智能技术滥用，指利用人工智能技术实施不道德、不合法或有潜在危害的行为，这些行为可能会危害人类和社会的福祉，违反公共道德和法律法规。这种滥用可能包括但不限于个人隐私泄露、数据滥用、歧视性算法、虚假信息、恶意攻击、人工智能武器化等，这是该技术负效应最广泛的存在。目前人工智能技术所面临的滥用问题中，影响范围最广的是算法歧视和数据隐私泄露，下面将以这两个问题为例具体解释人工智能技术滥用问题的严重性。

现代人工智能技术发展离不开算法，这是人工智能对海量数据进行处理的主要方式，如果算法不公平，就会对人类个体与社会均产生负面影响。“算法歧视”指的是人工智能系统在处理数据时可能出现的不公平和偏见现象，导致某些人群不受到公平的待遇。这种歧视有可能会加剧社会中现有的不平等现象，进而影响到受到影响的人群的生活质量。常见的算法歧视包含价格歧视（杀熟）和性别歧视等，其中价格歧视在一些外卖平台上表现得尤为突出，部分商家的割韭菜行为会导致“熟客”利益受损。

数据隐私泄露是指个人或组织的敏感数据被未经授权的第三方获取和使用的行为。在大数据的时代背景下，每个人的信息都可以被数字化，放眼到人类社会中，人类个体从某种角度上可以被看作是由自己的全部信息所构成的综合体，现代技术和现代人的深度合作给个人和企业都带来了数据隐私泄露的风险。数据隐私泄露会带来身份盗窃、偏见性算法、个人隐私泄露、商业机密泄露、国家安全泄露等。比如，近几年对社会影响较大 Facebook 泄露用户隐私事件以及疫情期间全球最大的视频会议软件 Zoom 泄露隐私事件等等，在这些事件中个人的姓名、联系方式和私人的亲密聊天等被售卖或公开在网上，对用户的个人隐私造成了极其严重的损害。人工智能技术的算法歧视和数据隐私泄露的危害不仅涉及个人的利益，还涉及社会的稳定和公共安全问题，并且正在随着人工智能技术的进一步应用，给人的生活带来持续不断的困扰。

（二）技术失控

技术失控通常指技术脱离了人类的控制并对人造成伤害或带来经济损失。而人工智能技术失控，体现在人工智能系统的设计、研发和应用过程中，系统出现无法预测、无法解释和无法控制的行为，由于各种原因导致技术无法受到人类的控制和管理，从而引发各种问题和风险。现实生活中，人们常将机器伤人事件归咎于技术失控的后果，引起了社会的广泛关注。然而，此类事件往往是由于技术自身不完善或者技术失误所导致的，因此将其形容为技术失控这样的说法并不准确。更多时候，是媒体为了博人眼球而有意采用的说法，因此误导了公众进而使人们偏离了对人工智能技术的客观认识。但也正因为媒体大肆报道，人工智能技术失控问题也同样具备了显性伦理问题的特性。从上世纪末至今，随着人们对于该问题的关注度逐渐升高，衍生出了诸多的文学和影视作品，并在全世界范围内吸引大批人工智能技术爱好者，投身于技术发展的行列。

在人工智能技术失控问题中，最引人关注的，就是由于过度智能化而引发的技术失控，在这种情况下，人工智能技术有可能会突破所谓“程序和代码的禁锢”，开始出现“独立思考”的意识特征，并且通过自我学习和自我进化来进一步脱离人类的控制。因此技术失控的表现具有三个阶段的特点，即超越、摆脱和反向控制，三个阶段分别对应着技术水平超越人类的计划和预想、摆脱人类的控制以及最终控制人类。[①]英国作家玛丽·雪莱在 1818 年创作的长篇小说《弗兰肯斯坦》中就关于人与技术人工物之间的关系问题，表达过自己的看法，这本小说也被认为是科幻小说的开山之作，对于人们正确理解人工智能技术失控具有借鉴意义。小说中，主人公弗兰肯斯坦通过自己的实验创造了一个人工生命体——怪物，但他却无法掌控自己所创造的生命，最终导致了悲剧的结局。这反映出了人类在追求科技进步时，常常会忽略自然的规律，盲目地追求技术上的突破，导致了对自然的破坏和对人类自身的威胁。《弗兰肯斯坦》揭示了技术失控的危险和后果。在小说中，科学家弗兰肯斯坦创造了一个怪物，但是无法控制这个怪物的行为和影响。同样地，人工智能技术的发展也可能导致失控和危险，当技术超越人类的理解和控制能力时，可能会引发难以预料的后果。

人工智能技术在发展上有着两个阶段的预想，即弱人工智能和强人工智能，其中强人工智能同雪莱小说中的怪物有着极大的相似性，即拥有独立的意识和情感，这样的技术要求人类必须主动思考关于人与技术人工物之间关系的问题，而不能任技术没有规制和底线的发展。长久以来技术失控问题是对技术持悲观态度的人普遍存在的观点，如今随着人工智能技术在其自主性研发上不断地投入，人类需要对其保持警觉，而不能因为技术的水平还没有达到，就认为是无稽之谈。

① 高亮华.技术失控与人的责任——论弗兰肯斯坦问题[J].科学与社会，2016（03）：128-135.

（三）管理缺失

人工智能技术的管理缺失是指在人工智能技术的开发、应用和监管中存在的缺乏有效管理的问题。人工智能技术被普遍认为是目前科技领域中最具前景的领域之一，可以带来更高的效率和利润。然而，随着人工智能技术的快速发展，人们也开始意识到在管理方面可能带来的风险和挑战。人工智能技术的管理缺失主要体现在管理上的滞后性和管理上的经验缺乏。

管理上的滞后性。这种情况出现在人工智能技术快速发展的背景下，管理和监管机制与之匹配的速度跟不上，导致相关法律法规、政策、标准、伦理规范等方面的滞后。2019 年，美国高盛银行的 Marcus 零售银行部门在使用人工智能技术时，因为管理不善，导致一款算法在审批信用额度时存在性别歧视问题。该算法将女性的信用额度设置为男性的一半，而且这种歧视是系统性的，而不是偶然的错误。这一问题最终被曝光并引起了广泛关注，导致该银行被迫进行改进和调整。该企业在使用人工智能技术时，从一开始就没有考虑到种族和性别等因素，也没有考虑到算法在可解释性和公平性上的问题，造成了管理上的滞后，引发了不良的社会影响。此外，人工智能技术的突破很快，新技术出现的时间可能只有几个月，管理者没有及时了解和掌握，也会出现技术滞后的问题。

管理上的经验缺乏。这种情况出现在人工智能技术的开发和应用过程中，缺乏足够的经验和知识来有效管理和监督技术的使用。由于缺乏对技术的管理经验，英国公共医疗机构 NHS 在采用人工智能技术来提高理疗效果和患者满意度时，面临着隐私和数据安全问题，这导致患者和医生对该技术的认可不足。除了企业在技术的研发和使用中出现管理缺失的问题外，法律法规的制定也体现了内容上的缺失与相对滞后。例如，2016 年美国发生了一起特斯拉自动驾驶汽车事故。由于缺乏对自动驾驶车辆事故责任的明确法律规定，这一事件引发了公众对人工智能技术管理和应用缺乏法律规范的争议。当然，这场事故的争议促使人们进一步思考如何在法律法规上规范自动驾驶汽车的使用。同样欧盟委员会，发布的《数字服务法案》和《数字市场法案》虽然是数字领域立法的重大成就，但其中关联的内容，仍是以限制头部科技公司为主（这里的科技巨头包含苹果、谷歌、亚马逊等），导致在适用范围和细节处仍有许多需要补充和完善的地方。

管理缺失，给人工智能技术的健康发展带来了巨大的阻碍，因管理缺失而造成技术上的失误，会使人对技术产生的合理性提出怀疑，并抱有消极的态度。尽管从人工智能技术的发展而言，管理上的失误并不是产生诸多伦理问题的关键，但这仍制约着技术的良好发展。

二、人工智能技术引发的隐性伦理问题

与显性伦理问题不同，人工智能技术的隐性伦理问题对人类的影响具有隐蔽

性，这是一类暂时无关当下而更关乎未来的伦理问题，在现阶段对人类的影响较小，距离人们生活较远，不容易被人们发现和理解，并且不会对人和社会造成直接的伤害与财产损失，但是从人类社会长远发展的角度来看具有重要的意义。人工智能技术所带来的隐性伦理问题包含了道德地位模糊、主体交往异化和人类生存危机隐患。比起显性伦理问题，隐性问题对人的影响是潜移默化的，同样应该引起人的重视。

（一）道德地位模糊

从人工智能技术诞生之日起，人类就想象着人工智能可以像人一样真实的生活在世界上，它们有着和人一样的外表，可以做出和人一样的行为。那么这些外表上类人的，功能上拟人的机器或者系统可以拥有和人相似的道德地位吗？人工智能技术的道德地位模糊问题指的是，在人工智能技术越来越广泛应用的过程中，人们很难对其道德地位做出明确的规范性判断。换句话说，人工智能技术的道德地位缺乏一致性和稳定性，难以确定它们在人与人、人与自然和人与社会等伦理关系中应该扮演什么样的角色和承担什么样的具体责任。

从理论上来讲，当我们谈论人工智能技术的道德地位模糊问题时，确实存在包括人工智能技术的自主性、是否获得道德主体地位以及是否具有道德能力等问题。首先，尽管目前的人工智能技术还远远没有达到真正的自主水平，但是随着技术的不断进步，人类无法排除人工智能技术在未来某个时刻会具备真正的自主能力。如果人工智能技术具备了自主性，那么这些技术将能够自主地做出决策，甚至可以对其程序进行修改，这种情况下，人就无法完全掌控人工智能技术的行为。这就可能导致人工智能技术做出不符合道德规范的决策，从而给社会和人类带来巨大的危害。其次，关于人工智能技术是否应该被赋予道德主体地位也是一个备受争议的问题。由于目前的人工智能技术仍然无法具备意识和感知能力，因此它们不能像人类一样拥有道德主体地位。然而，在某些情况下，人工智能技术可以对我们的行为产生重大的影响。例如，当人工智能技术在医疗诊断或司法裁决等领域发挥作用时，它们的决策可能会直接影响到人类的生命和自由。最后，一旦人工智能技术的具有了一定的道德的能力，并把它投入到战争中时，人如何确定人工智能是可以信赖的？如果人工智能技术因系统故障或算法上的错误时，发动了错误的目标位置打击甚至误伤平民或队员，最后需要谁来承担责任？人类能够对似人非人的人工智能技术给予信赖吗？

在现实应用中，道德地位模糊的问题常常与人对人工智能技术持有的观念和态度有关。例如，在与聊天机器人对话的过程中，如果机器人展现出类似于人类的思维和行为时，人们可能会将其看作是一种拥有自我意识的个体，或是将其看作为人类自己的延伸。这种将机器视作“类人类”或“伪人类”的看法很可能会

导致人们忽略了机器的本质特征和局限性，甚至可能会在与机器的深入互动过程中逐渐偏离对自己和对机器的客观认知。此外，人们可能会给机器施加不适当的道德要求，或者将机器视为一个可以被责备和惩罚的实体。比如，在日本一些人会把机器人视为自己的亲人或宠物，并对它们进行冠以“姓名”“喂食”等行为。如果人们将机器视为纯粹的工具或设备，就很可能会忽略机器在与人交互中具有的重要作用，以及机器对人类社会和环境的影响。这些都是人工智能技术道德地位模糊问题的具体体现。随着人工智能技术的在自主性上的不断发展与完善，技术获得的决策权和行动权会不断提高，让人工智能技术获得道德地位并学习和运用道德能力变得越来越重要，因此明确人工智能技术应该具备何种道德地位的关键性变得日益突出。

（二）主体交往异化

“交往”产生于社会实践活动，发生在主体之间，体现的是人与人之间的交互。而“异化”则是代表“分离”或“疏远”，表达的是客体脱离出主体后，反过来对主体的控制。因此，可以将“交往异化”理解为客体对主体之间关系的控制，这使得主体之间的关系表现出分离、疏远和不平等。在此基础上，人工智能技术导致的交往异化，就是指由于技术所导致人与人之间的交往关系变得疏离和不平等，可以从“人对自身、人对他人”两个角度，分析交往异化的具体表现。

首先，人工智能技术引发的交往异化，会导致人产生自我认知方面的怀疑。一直以来，人的交往活动都是出于个体的主观选择，人可以按照自己的意愿与自己想结交的人沟通、交流，这是人有意识的、自主的活动。近些年，随着人工智能技术与人的“合体”、“合脑”并伴随着研究的不断深入，人机交互领域的现代化水平在不断地提高，智能化程度也在不断加深。由此，人工智能技术会导致人的有意识的、自主的活动变得不再那么明晰，甚至出现一种与人相对抗的意志，处处挑战着人作为主体认识自身的限度。雷·库兹韦尔指出：“未来，人工智能可以发展出自我意志，一个与我们冲突的意志。”[①]近些年，“脑机接口”的研究已经得到了初步的成果，这项致力于通过脑机连接来解决或强化人的行动能力的技术，引发了人对自身主体认识的怀疑，通过机器增强后的人体还是传统意义上的人体吗？在医学领域，将机器植入人体内，完善人体器官功能保持人的生命已经是较为成熟的技术。而将智能芯片植入人类大脑，利用芯片来代替或者强化大脑的功能的技术手段也已投放在实验阶段，这些芯片可以通过刺激人类大脑部分神经元使人跑得更快或者提升记忆力等。当机器植入人体，成为人体的一部分，与人共生时，人对自身能力的认识，对主体的认识，都会发生改变，“交往异化

① [美]雷·库兹韦尔著.人工智能的未来[M].杭州：浙江人民出版社，2016：274.

之所以产生就是因为‘人不承认自己是人’，人同自身相异化。”[①]

其次，人工智能技术引发的交往异化，会导致主体与主体之间出现交往的异化，人与人工智能之间虚拟的交往，会引起现实中人与人之间距离感的增加。同时，虚拟环境下人与人之间的交往也有很大的隐患。如今，交互机器人的应用已经开始走进入寻常百姓的家中，像宠物机器人、陪伴机器人和情侣机器人等人工智能技术产品，对人的现实交往造成了一定的影响，机器本身是没有情感的，但是却会迎合人做出情感反应，在这个过程中人会对机器有感情，人机交互的天然不平等性，会带来交往的异化，如果过度地依赖这样的虚拟情感，则必然忽视现实的情感，从而与现实中的人产生距离感。

技术的发展给人类的交往方式带来了很大的变化，人与人的认识可以冲破时间和空间界限，任意畅聊。技术大爆炸之前，人的交往是受局限的，不便捷的交通，口音多变的方言，就像歌词里形容的那样从前车、马、邮件都慢，在有限的时间里人可以用漫长的一生认识一个人。如今，基于脑机接口技术构建的元宇宙，可以帮助人从现实世界进入到虚拟世界，虚拟世界中人和人物理上的距离感会在技术的加持下消失，一个人可以轻易地认识地球上的任何人。同时，虚拟世界里，人与人的关系、人的生活和工作，都会搬到元宇宙中，基于此人与人之间的交往就带有很强的不确定性。人工智能技术引发的交往的不确定性会增加人类社交的风险，更值得关注的是，很可能成为人类避世，拒绝与现实的人和事物产生联系的阻碍。

（三）人类生存危机隐患

人工智能技术所引发的人类生存危机，是指人类创造出一个比自己更聪明的东西时，面对一个在智力上和人类相当，甚至会超越人类的存在，人自身的发展将何去何从的问题。在过去，人类制造机器是为了将人从重复性的劳作中解放出来，并认为人工智能技术不论如何发展，都只是对人行为和能力的模仿，只能解决程式化的工作，而不会像人一样，有想象和创造的能力，对人工智能技术持乐观态度的人甚至认为，在未来人工智能的存在会创造更多只有人才能做的工作，从而发挥人更多的价值，但事实真的如此吗？2019 年，由微软研发的人工智能机器人“夏语冰”（后面以小冰代替），以硕士生的身份从中央美术学院毕业，不仅如此，小冰还以画家身份举办了画展。在举办画展过程中，没有人发现小冰的作品，竟然出自人工智能技术之手，从某种意义上来说，小冰通过了图灵测试，除此之外，小冰还会写诗和唱歌。艺术创作一直以来，是作为人类智慧金字塔一

① 谢世良.从交往异化视角论述马克思劳动异化的第四个规定——重读《穆勒评注》[J].信阳师范学院学报(哲学社会科学版)，2016（06）：18-21.

般的存在，如今人工智能技术通过深度学习也可以进行艺术创作，这无疑是对人类智慧的挑战。埃隆·马斯克曾公开表示过，在未来人与人工智能机器人博弈的过程中，人类生存下来的概率可能是 5%到 10%。这是对强人工智能到来的时代，人类对比自己强大的人工智能技术的忧思。但人工智能“夏语冰”的出现，预示着人类天才般的创造力，机器是可以通过学习攻克的。2022 年 11 月，Open AI 研发的智能聊天机器人 ChatGPT 横空出世，并在极短的时间内席卷全球，这个聊天机器人与以往有很大的不同，它可以根据用户聊天的上下文，进行连续的回答，也就是说用户如果围绕一个关键词提问，那么每次的提问可以相应的省略一些重复的话，比如，用户问它“锅包肉哪里的最正宗”，接下来又想了解锅包肉的具体做法，你可以直接提问，“请教我怎么做”就可以得到想要的答案，而不需要重复输入主语它就能够理解，这样的文字分析和处理能力是以往聊天软件无法企及的。除了聊天的功能之外，它还会写歌词、写诗、写代码、写文案，就连脑筋急转弯也不在话下。对未来人工智能技术的担心不是杞人忧天，人工智能技术威胁的不是人类的眼前的生存问题，而是长远的发展问题。据牛津大学的调研报告，未来将有 1000 万非技术工种被机器人取代，其中包括文秘、工人、中介、司机等一大批岗位。报告特别指出，目前软件工程师所做的工作将会被智能机器人所代替，即创造者被其创造的技术产品所代替。[①]试想，在未来协同产业进行智能化升级的科学技术，可能不会是辅助人类的工具，而是会成为更是适宜机器工作的助手，技术补充的不再是人类功能上的短板，而是机器身上的短板。到那时，究竟还有多少工作是人类真正可以胜任的呢？就像卫星导航系统，服务于现在的人类，辅助人行驶在适宜人类生活的道路上，而在未来，假设当无人驾驶汽车盛行时，卫星导航服务的对象将不再是人类，而是与无人驾驶系统之间形成技术互补，到那时，道路也许不再是适宜人类生活的模样，而是为了技术量身定做。除了人工智能技术可能会逼迫人类向技术投降之外，人类还面临着自身发展受限的威胁，如果你动动手指就能知晓问题的答案，还会有多少人通过学习来增长自己的知识？

人类的祖先与黑猩猩的祖先曾在历史前期有过共同生活的经历，并在约 500-600 万年前，分道扬镳各自发展，这种生物进化的经历，与人和人工智能技术在智能的进化上有一定相似性。试想，未来人与人工智能技术的关系，可能就像现在黑猩猩与人一样，当人工智能看向人类时，就像人看动物园里的黑猩猩。随着技术的不断发展，当技术到达奇点时，人工智能技术会实现智能上的进化，会全方位地碾压人类，如果人类不重新思考技术对人的塑造，人类文明很可能面临终结。

① 吴汉东.人工智能时代的制度安排与法律规制[J].法律科学(西北政法大学学报)，2017（05）：128-136.

第四章 人工智能技术伦理问题的根源分析

人类的现实需求是需要得到充分满足的，科技的飞速发展也为此开启了百花齐放的时代。不过，这也给人工智能技术的伦理问题带来了复杂多变的形势。人的观念和技术的复杂性之间相互拉扯，引发了问题的出现。本章在原因分析部分将以马克思主义科技伦理思想和汉斯·尤纳斯的责任伦理思想为支撑，深入分析人工智能技术伦理问题的产生根源。

一、人工智能技术显性伦理问题的根源分析

人工智能技术的显性伦理问题，是对人产生直接负效应并且矛盾突出的现实问题，对人当下的活动有着直接的影响。这类问题的产生，主要围绕着技术本身复杂性，和利益相关者的一些不当行为展开，即技术存在算法黑箱、人对技术持有工具属性大于价值属性的观念和技术的利益相关者缺乏责任意识。

（一）人工智能技术存在算法黑箱

在人工智能技术的显性伦理问题中，技术滥用问题的出现，通常发生在算法的黑箱背景下，黑箱当中无法了解和理解的那部分运行机制，恰恰给技术滥用提供了可乘之机。随着算法的应用范围不断地扩大，人对于算法的接受度不断地提高，因此也大大增加了技术的风险。

算法，是现代人工智能技术发展的基础性技术之一，同大数据和算力一起并称为人工智能技术的“三驾马车”。算法是一种有效的计算方法，其核心是提供解决问题的方案和步骤，通过计算机的支持，可以处理各种类型的输入信息，并输出一个解决方案。在现代社会中，算法已经广泛应用于人们的生产和生活活动中，比如金融、医疗、物流、社交网络等各种领域。算法的运用使得许多原本繁琐、重复和复杂的工作得以自动化和简化，提高了工作效率和精度，同时也降低了人力成本和时间成本。

“黑箱”指的是算法在进行决策时缺乏可解释性和透明性，这使得用户难以理解算法的决策过程，可以泛指由输入得到输出，却不了解其内部运行机制的一切系统。在人工智能技术算法中，由于深度学习等技术的应用，系统的内部运行机制变得更加复杂，当分析的数据与情境变得愈加庞大和复杂时，所带来的“黑箱”问题也愈发突出。这种缺乏透明性可能导致算法的决策不公正或有偏见，对社会和个人都可能带来负面影响。

算法黑箱导致的人工智能技术滥用问题主要体现在以下几个方面。

首先，围绕深度学习展开的算法研究，是对人脑神经元的模仿和学习，因此技术本身存在巨大的研究难度。人工智能技术对人脑神经元的学习是基于目前人类脑学科的发展水平而言的，人脑又是一个极其复杂的系统，人对于人脑的研究目前还没有达到充分的了解，这就会导致模拟人脑功能的深度学习研究必然存在

无法解释的黑箱，而基于深度脑神经网络研究的算法在使用过程中也就会出现一些人类无法解释的现象。

其次，技术自身的难易程度决定了存在黑箱的可能，当算法出现错误导致问题发生时，黑箱就会遮蔽问题的源头，过于复杂的算法内部，会隐藏人的意图，甚至成为责任人开脱的理由，这也成了如今人工智能技术滥用的主要原因。

最后，除去算法自身的复杂性带来的黑箱之外，另一个引起算法黑箱的重要因素就是人。人工智能技术的算法是由人类设计和编写的，因此技术本身并不是中立的，而会是受到设计、开发和使用它的人的影响。从现实的视角看，算法是由一串串代码组成的，而人正是代码的编写者，可以说每一个算法的背后，都在表达编写者的意图，“在某种意义上讲，算法只是一种工具，它不能完全区隔价值观，抽离价值观。文化是先于算法设计的出现就早已存在的，植根于现存的社会制度、实践、态度及其价值取向之中。”[①]算法歧视的背后体现的是人工智能技术被心存偏见、恶意以及一心求利的人滥用。在此基础上，如果普通人不了解技术的局限性和不足之处，算法就成为一个面向普通人的“黑箱”。在这里，我们所说的黑箱体现了技术对普通人的遮蔽。对于那些不熟悉计算机语言的非专业人士来说，算法的黑箱表现出来的是极度的不透明性。这种不透明性导致了如今人工智能技术中发生的性别歧视和大数据杀熟等问题。以算法黑箱作为屏障，则有利于软件开发商采取障眼法，从而损害用户的个人利益。作为非专业人士的终端用户，缺少对算法背后意图的了解，使得自己成为了被割韭菜的受害者。算法的黑箱对应着技术的研发环境是“不透明的”，体现了人工智能技术自身和开发环境的复杂性。但是，就用户而言，算法的黑箱并不一定是坏事。因为对于大多数终端用户来说，他们只关心算法提供的功能，而无需了解其背后的逻辑。因此过度增加算法的“透明度”并不能消弭普通人对技术的“不可理解性”，反而会增加用户的困扰。但值得注意的是，一旦技术的核心掌握在少数人手中，或者只有专业人士才能读懂时，就必然会产生问题。一旦技术只为了少部分人的利益服务，并成为少部分人开脱责任的借口时，就会造成黑箱影响下的技术滥用。

（二）技术的工具属性和价值属性的失衡

人工智能技术的工具属性，指的是人工智能技术作为一种工具或者手段来使用，为人类提供服务和帮助，例如自动驾驶、语音识别等。人工智能技术的价值属性，指的是人工智能技术在价值层面上的作用和影响，例如道德、社会、政治等方面的价值。马克思主义科技伦理思想强调科学主义和人本主义的结合，是一种辩证的科技观，即技术的工具属性和价值属性的兼顾与平衡。因此，该观点不局限于科技工具的使用，也不局限于技术实体本身，而是注重在历史唯物主义的

① 李仁涵，黄庆桥等，编著.人工智能与价值观[M].上海：上海交通大学出版社，2021：18.

指导下，人与技术的共同作用和影响。因此，二者的平衡，是确保技术健康发展的前提。

然而，在实际应用中，人工智能技术的工具属性往往被过度强调，而其价值属性却被忽视或弱化。这种失衡表现在两方面：一方面，现代人对待技术过于工具化，忽视了技术自身的价值；另一方面，人们普遍享受技术带来的便利，但对技术带来的负面影响已逐渐失去敏感性。这种对待技术的失衡观念会加剧人工智能技术的负面影响，使技术逐渐转向功利主义，从而成为满足人类欲望的一种工具。技术一旦盲目发展就会导致技术失控，从而给人类的生存和生活带来更多危机。

当人们对技术的工具属性过于重视时，技术就成为了人的附属品。工具属性，意味着为人所造，为人所用，为达成人的目的而驱使这些特征。在以工具属性为主的观念下，技术自身的意义往往会被掩盖，人们更在意的是如何通过技术提高效率、推动生产力进步。但长此以往，技术的发展可能会转化为资本间的竞争和斗争，从而忽视技术的良性发展和合理应用的重要性。技术的工具属性使得技术成为中立的存在，从而导致人与技术之间出现了根本的对立，呈现出控制与被控制的关系。一旦技术出现问题，责任往往会全部归咎于人类，因为技术本身并没有思想和控制权。然而，技术发明的大多数时候都是为解决现实问题而出现，本质上并不应该与人的活动造成冲突。如果我们过度依赖技术的工具属性，技术会完全沦为人类的附属品，其发展将永远逃脱不了人类的偏见，有些本来有望为人类带来福利的发明和创造，却可能带来负面的影响和效果。在技术快速发展的今天，现代技术与社会深度合作，将技术的工具属性不断放大。人工智能作为变革社会和生产的关键技术，更凸显了这一问题。人类过度依赖技术的工具属性可能引发技术失控或人类生存危机，进一步加深人机矛盾。正如，清华大学高亮华教授在探讨弗兰肯斯坦问题时所表达的观点一样“技术之所以失控，并不是因为技术本身的自主与邪恶，而是因为我们忽视对技术的理解与责任。”①

重视技术自身的价值属性，是指不仅关注技术的功能和效果，同时也注重技术所具有的道德、社会、政治等价值属性。比如智慧城市的构想就建立在技术成为人类生活的背景，技术不再被看作是一种单纯的工具，而是被视为一种与人类价值体系相互关联的文化现象。关注技术的价值属性，是人类对技术合理性的关注，这满足人类自身的长远发展，而忽视技术自身的价值，技术发展就会被人的欲望完全控制，从而引发技术失控和人类生存危机。

（三）人工智能技术的利益相关者缺乏责任意识

汉斯·尤纳斯认为，伦理责任是一种行动者的责任，其目的是确保行动者的

① 高亮华.技术失控与人的责任——论弗兰肯斯坦问题[J].科学与社会，2016，6（03）：128-135.

行为符合道德规范和社会价值观。在人工智能技术引发的伦理问题中，行动者即利益相关者，这些人缺乏责任意识是导致技术管理缺失等问题的主要原因。这种缺乏责任意识表现为缺乏远见和长远考虑、缺乏主动担当和道德判断以及缺乏透明度和问责制度。因此，利益相关者缺乏责任意识，不仅可能导致个人或组织的利益受损，也可能对整个社会造成不良影响。

利益相关者，是指在某个特定的事件、事物或者决策过程中与之相关的各方，他们可能会受到该事件、事物或者决策的影响而获得利益或遭受损失。在人工智能技术中，利益相关者可以包括但不限于开发者、生产商、用户、政府、监管机构、研究人员、劳动者、消费者、投资者等等。这些相关者可能会因人工智能技术的开发和应用而获得利益，也可能遭受损失。可见，人工智能技术的利益相关者既可以是某一个人也可以是集体。

责任意识是一种美德，它能够使技术的利益相关者对技术的潜在风险保持警觉，从而避免技术事故的发生。而缺乏责任意识会导致技术利益相关者对技术风险视而不见，对风险的重要性进行低估，甚至会在事故发生后互相推诿，逃避责任，导致技术管理缺失。从群体产生的效应来看，集体具备责任意识更为重要。集体作为由每一个具备承担道德责任条件的个体组成的道德共同体，其本身也满足承担道德责任的要求。①

在过去发生的人工智能技术事故中，例如亚马逊的面部识别技术和Facebook的数据泄露事件等，实质上是由于企业在技术研发时缺乏责任意识。这些企业没有将责任意识放在产品设计的前端，并缺乏对技术的事前预测。只有在问题出现后才开始认识到其重要性，缺乏远见和长远考虑。缺乏责任意识不仅会对企业和使用者产生负面影响，还会影响到人工智能技术的监管和立法。如果政策制定者和监管机构没有意识到人工智能技术可能带来的风险和潜在后果，他们可能会对人工智能技术的监管不力，对其立法缺乏关注，缺乏主动担当和道德判断，从而对社会和公共利益造成损害。因此，人工智能技术的管理和监管需要建立在充分的责任意识基础之上，建立相应的问责制度，重视技术的风险和潜在后果，制定合理的管理和监管策略，以保障人工智能技术的安全和可靠性。

二、人工智能技术隐性伦理问题的根源分析

人工智能技术的隐性伦理问题，建立在“技术茧”的背景下，人们在不知不觉中被“它”左右。人工智能技术通过计算机视觉、语音识别、自然语言处理、机器学习和大数据足可以达到与人交互的能力，但人也很可能因为与人工智能的交互而产生出错觉，比如是否在与一个具备道德能力的“机器”对话，人也会在

① Perron Tollefsen D.Participant Reactive Attitudes and Collective Responsibility[J]. Philosophical Explorations，2003，6（3）：218-234.

与其交互的过程中不由自主地对技术产生情感，甚至对技术的发展表示恐惧与担忧。人工智能技术创新在催生新兴产业发展与社会关系重构的同时，往往引发道德伦理、主体交往异化和社会民生等方面的负面危机。[①]通过进一步地分析这些问题可以发现，这与技术的底层设计和人类社会制度的发展息息相关，具体包括人工智能技术存在意识的拟人性和物理构造非人性的矛盾、人工智能技术对社交秩序的重构和人工智能技术对劳动力就业的干扰。

（一）自我意识拟人性与物理构造非人性之间的矛盾

人工智能技术存在自我意识的拟人性，体现在它致力于对人类社会知识的学习，对人类语言、行为和思维方式等能力的模拟，它的研究对象和服务对象都是人类，在它逐步自我升级与进化的整个过程中，相当于是对人类本体的研究。这导致人们很容易将其视为，具有与人类相同或类似的道德地位的“生命体”。由此引发了人工智能技术是否应该被赋予道德地位、是否应该承担责任和是否应该受到保护等道德地位模糊的问题。

究其原因，人工智能技术在自我意识的拟人性方面的发展，存在着与其物理构造的非人性之间的矛盾。具体来说，人工智能技术在意识发展上呈现出拟人性，通过模拟人类智能的形成方式，使技术自主性得到了不断的发展，人工智能技术可以通过自主学习进而完成特定的任务，也可以通过知识系统、逻辑推演和语言处理来与人交流，但当下社会人工智能技术呈现出物理构造的非人性，导致人们将其视为“工具”或“资源”，在这样的矛盾之下人工智能技术的道德地位非常模糊，因此也成为了计算机科学，认知科学以及哲学等学术界和社会讨论的焦点问题。

矛盾具体表现在以下两方面。一是，现如今的人工智能技术在其顶尖水平领域可以做到尽可能地模拟人类的思维和行为方式，并在某些任务上表现出超越人类的能力水平，但是它们的物理结构和运作方式并不能够像人类一样具备生物学特征和情感认知能力。人类智能是人类知行统一的表现，是人类在漫长的演化过程中，基因突变的结果。而人工智能技术的运作是基于算法和数据的处理，这与人类的大脑神经元的结构和工作方式有很大不同。例如，深度学习算法是基于人工神经网络的机器学习技术，它的工作原理是通过模拟人脑神经元的工作方式来实现人工智能的目标，但是人脑神经元的工作是基于离散的电化学信号传递，而人工神经网络则是通过一系列数学函数来模拟信号传递的过程。因此，人工智能技术的物理构造与人有着本质的区别。二是，人工智能技术也缺乏与人类大脑相应的情感认知能力，如理解和表达情感、同情心、善良和道德认知等。它们对人

① 梅亮，陈劲，吴欣桐.责任式创新范式下的新兴技术创新治理解析——以人工智能为例[J].技术经济,2018,37（01）：1-7+43.

类价值观和道德规范的认知和理解也存在一定的局限性。例如，在语言翻译任务中，机器翻译系统可能会选择不合适的词语或短语来表达情感和语气，导致误解和不当的表达。

从理论角度来看，人们做出道德判断和选择时通常会依据一定的原则，人工智能技术可以通过对人脑思维意识的不断模拟加上算力的支持来无限地逼近人类，实现和人相当的意识能力，从而获得对人类道德的理解，还原人们进行道德判断时的情感倾向，并在大部分情况下做出与真人相一致的道德选择，因此具备成为道德主体的可能。从现实角度而言，道德行为中又往往存在非理性的因素，人工智能技术不具备人类的生物机能基础，也就无法通过劳动实践来进化出真正的意识思维和自主意识，无法具备人在进行道德判断时所必要的“时中”性，因此难以完全拥有和人类一样的道德能力和地位。这种自我意识拟人性与物理构造非人性之间的矛盾，正是造成人工智能技术道德地位困境的根源。

（二）人工智能技术对社交秩序的重构

社交秩序指的是人们在社会互动和交往中所遵循的一些规则和规范，包括社交礼仪、人际关系和社会交往方式等。人工智能技术的出现与发展，不仅改变了人们的生产、生活和思维方式，也对社交秩序产生了重要的影响。社交秩序的重构体现在，人类依托于人工智能技术营造的虚拟环境，进行虚拟的互动，由此引发了人之间的交往异化。这种虚拟的互动表现为，由人工智能技术所创造的虚拟场景、虚拟人物和设定以及虚拟情感。虚拟的互动不仅加剧了人对自身主体性的怀疑，还会阻碍人与外界的交往，并且增加交往中的不确定性，这对人与人之间建立互信、平等和尊重的伦理关系产生了负面影响。

在人工智能技术的支持下，虚拟的互动，正在逐步占领和取代人类在现实中的互动方式。近几年，越来越多的游戏公司，使用人工智能技术帮助游戏实现智能化设计，比如在游戏领域，“超级马里奥兄弟”采用了人工智能合成器来生成游戏，人工智能模型在学习了大量的游戏数据后，可以生成无限的、随机的关卡。DOTA2 利用强化学习和人工神经网络学习人类战术和策略，从而挑战人类玩家。在人工智能技术的加持下，神秘海域 4 还会根据玩家的表现调整游戏难度。利用人工智能技术，这些游戏给玩家带来了更逼真的游戏体验，沉浸式的冒险感和奇幻感让玩家可以在游戏中获得一些现实生活中无法给予的满足感。这些满足感可能包括身临其境的感觉、探索未知领域的感觉，以及实现自我价值的感觉等。然而，随着虚拟技术带给人的沉浸感越来越强，人对技术的感知力也会越来越弱，依赖性也就更强。虽然技术能够带领人进入一个“完美”的世界，但也把人困在方寸之中。在虚拟游戏中，人可以逃避现实中的家庭关系、社会联系，建立起新的社交，这种依托虚拟环境交友的形式，会在某种程度上给人一种避世的安全感。

比如，可以隐藏自己的社会地位和外貌特征。如此一来，人类真实的交互就会受到虚拟交互的影响，改变现实中的社交秩序，并试图用虚拟互动中形成的社交礼仪和交往方式取代现实中人与人的交往，现实里人的感情很可能会因此变得薄弱，甚至对现实中的人和环境变得很冷漠。人们在享受人工智能技术产品带来的服务时，不自觉地重构了社交秩序。在社交秩序被重构的背景下，虚拟互动看似给人类逃离现实提供了便捷的渠道，但现实中的问题不会因此从根本上消失。这种避世的手段只会加剧人与现实的矛盾，导致交往的异化。

（三）人工智能技术对劳动力就业的干扰

在经济学中，常有这样一个比喻，就是把经济比做成一块蛋糕。想要经济增长，意味着需要把这个蛋糕越做越大。自从工业革命以来，技术作为生产要素之一，成为了将蛋糕越做越大的重要力量。技术帮助人类将蛋糕做大，同时也在改变蛋糕的形状和原材料。人工智能就是改变蛋糕形状与原材料的强大技术，其迅速发展和广泛应用在许多行业引起了巨大的变革，这体现在某些工作岗位的消失和工作内容的改变，这对现有的劳动力就业构成了干扰，从而引发了人们对人工智能技术发展的恐慌。

关于人工智能技术发展带来的“人类生存危机”迷思，在不同人的眼中是有争议的。对技术持乐观态度的人认为，人工智能技术的发展会在未来创造更多的工作岗位，如人类学家施里哈什·克尔卡在研究教育者与数字教学工具的关系时发现，人类教师正利用这些数字工具提高效率。他表示，问题不是电脑自动化正让工作消失，而是“人类与电脑正在合作”。但对技术持悲观态度的人认为，人工智能技术的深入发展，会是对人类自身发展的颠覆，会压迫人性，挤压人的自由时间，比起新增的岗位会有更多的人类岗位消失，从而引发技术性失业。2016年，耶路撒冷希伯来大学历史系教授尤瓦尔· 赫拉利撰写的全球畅销书《未来简史：从智人到神人》预测：人工智能和算法（Algorithm）将战胜人类， 99%的人将沦为无用阶层！[①]事实上，现代人对人工智能技术产生的危机感（如技术性失业的担忧）是时代的产物，这与人类进入工业社会以来，资本主义市场经济的运行方式有着直接关系。私有制、市场经济和利润追求共同引导下的人类社会，人也是商品，技术是资本的帮凶，人工智能技术的出现加剧了资本对人的压迫，造成人被技术淘汰，或者为了生存要不断地去适应和学习新的技术，这就是人工智能技术给劳动者就业带来的最直接的影响。

如今，人工智能技术给许多行业提供了更高效、更准确且更便捷的解决方案，正在打破传统的如雇主－雇员此类型的雇佣关系，进一步加速岗位和工作内容上的变动，干扰了劳动力就业。具体来讲，人工智能技术的发展和应用，一方面使

① [以色列]尤瓦尔·赫拉利著，未来简史：从智人到智神[M].林俊宏译.北京：中信出版社，2017：535.

得一些传统的劳动力岗位逐渐被自动化取代。例如，在制造业、物流业和服务业等领域，许多机器人和自动化设备已经可以完成一些简单的工作，减少了人力成本，提高了效率。这就意味着一些需要大量人力的工作岗位逐渐消失，人的就业机会在减少。另一方面，一些传统工作内容产生了变化，例如在金融业和医疗健康领域，人工智能技术已经可以完成一些传统的人工操作，如审批贷款、医学诊断等。这就意味着这些领域的从业人员需要具备更高的技能和更深入的专业知识才能适应工作的变化。

马克思认为，资本主义机器大工业不会放弃技术对人的剥削，只会在激烈的利益追逐赛中，不断地压迫人的生存空间，于是人类被从农田赶到了工厂，再从工厂赶到高楼大厦里狭小的写字间，这也是劳动异化的具体表现。现代人在与人工智能技术的合作的过程中，人类的生存空间被不断地压缩，人的自由时间并没有因为人工智能技术高效率而相对变长，相反高压的工作成为了现代人生活的日常。人原以为技术的发展，可以将人从繁重的作业中解放，获得更多的自由。然而，现实情况却是，技术发展过于迅速不仅未能解放人类，反而逐步剥夺了人类创造价值的权利。人工智能技术对劳动力就业的干扰使劳动主体异化问题日渐突出，成为了引发人类生存危机问题的主要原因。

第五章 加强人工智能技术伦理治理的对策

在马克思主义科技伦理思想的视角下，人发明技术的目的是为了造福全体人类，因此技术本身并没有善恶之分，技术表现出来的效果往往体现的是人性的善与恶。在此基础上，为了发展向善的技术，需要借助马克思主义科技伦理思想和汉斯·尤纳斯的责任伦理思想，来重新审视人与人工智能技术的关系，重塑人对技术的理性认识。因此，人工智能技术伦理治理这条路径看似用伦理道德来约束机器，实质上是对人的约束，基于此人工智能技术治理的对象是人类在设计和使用技术时人的行为，而实现治理效果的方式，是建立人人都应该遵循的底线原则与行为准则。

一、人工智能技术伦理治理的使命

目前世界各国在发展人工智能技术伦理治理的方法上大致可以分为两类，一种是对人的直接约束，比如构建人工智能技术的伦理框架，设计伦理准则，另一种是从技术入手，利用“道德敏感性”设计为技术嵌入伦理知识从而实现对技术的约束。这些治理方式都可以归为软法治理，而软法治理的根本目的在于呼唤人内心的“善”。如果人想逃避规制那么怎样都能想出对策，因此唤醒利益相关者的良知，赋予他们对待人类整体利益的感性认识，才是伦理治理的根本目的。

（一）发展“有担当”的人工智能

普遍认为，不会伤害人类利益为人类生活提供便利的技术，是好的技术，而促进人类进步，缓和人与人、人与自然矛盾的技术是“向善”的技术，面对人工智能技术繁荣发展人类有着怎样的期许呢？什么样的人工智能技术可以被称其为向善呢？

尤纳斯认为，技术的发展应该建立在人类得以持续生存的基础上，这就要求，人工智能要是“有担当”的技术。发展“有担当”的人工智能技术，意味着技术的设计和使用符合人类价值观和道德标准。而“有担当”所包含的主体，即人工智能技术的利益相关者。这里包含了研发人员如何设计、销售人员如何策划宣传其功能以及使用者如何使用，因为人的行为和选择才是造成技术正负效应的根源。因此，发展有担当的技术，归根结底还是在强调人的有担当。因为，科学的价值的第一点是众所周知的。科学知识使人们能制造许多产品、做许多事业。当然，当人们运用科学做了善事的时候，功劳不仅归于科学本身，而且也归于指导着我们的道德选择。[①]

要求人有担当还是负责任二者之间是有区别的，负责任体现了事后性。在机动车发生交通事故时，交警会依据交规对事故责任进行划分，判定事故中牵连主

① [美]理查德·费曼.发现的乐趣[M].北京：北京联合出版公司，2018：144.

体之间的责任，一般情况下主观上处于过失的一方承担事故的主要责任。法律的强制性迫使人要守规矩，这种情况下人的负责任行为很难说是心甘情愿的，更多地体现一种趋利避害的本能反应。责任主体为了不触犯法律，从而约束自身行为、遵守规则、安全驾驶，从而对自己和他人负起责任，可见负责任体现了人在做价值判断和价值选择时，趋利避害的一种自然趋势，也是追求利益最大化的不情愿的被动伦理。在法的视域下“负责任”的行为虽然在利己的同时也为社会带来了好处，但是这种方式不能从根本上解决矛盾的冲突。

反观“有担当”强调事前责任和情感上的认同，也就是尤纳斯责任伦理中受情感趋势下的道德责任。这就要求人工智能技术有担当，也就是要求人在技术的研发和构想阶段就在情感态度上主动地承担起责任，这与事后责任之间有着本然的区别。以往我国对于工程技术的重大事故通常采取事后追责的处理方式，即对既成事实“负责”，这会对事故的解决产生一定的滞后性，对避免再次发生很难起到控制作用。因此要求人工智能技术在研发阶段就考虑到相应的后果，充分地进行科学预测，并且尽可能地保证研发技术的目的是有利于增进人类福祉的，而非单单满足研究人员自己的研发预想和实现经济效益。

此外，有担当的人工智能技术还应该在应用的过程中，尽可能地减少对社会和人的负面效应，抛开设计时目的是好的，在技术使用过程中也应该产生好的效果，不伤害人类不给人类带来可怕后果。交通事故几乎每时每刻都在发生，但是驾驶智能汽车引发的交通事故却更容易引发社会的关注，这是因为在法律层面和社会的认可与接受度层面，还没有做好万全的准备，因此人们会下意识的认为智能汽车在设计时出了问题，但是追究其事故原因，绝大多数出在驾驶人员的操作不当上，因此一味地要求技术完善并不能完全解决问题。人要预想技术在使用时会出现的风险，而且是尽可能多地考虑到，并为降低技术风险寻找相应的对策，比如在智能汽车的使用上，针对购买者进行系统专业的智能驾驶学习，以及针对学习情况进行考核，或者在收车之前进行充分的试驾体验，采取这些积极主动的方式减少事故的发生，同时面向社会进行产品科普。如果出现超出预想的负面效果，更应该及时补救主动地承担，而不是利用法律和规则的漏洞将自己撇清，这就要求利益相关者在有良知的情况下追求利益并做出合理的选择。

（二）推进人工智能技术的可持续发展

目前，世界各国在发展人工智能技术的同时，都在寻求一种共识，即对人工智能技术进行有效地约束，使其得到良性发展，服务于人类的同时，对改善环境、消除贫困、缓和不平等起到积极作用。推进人工智能技术的可持续发展，就是推进全人类的可持续发展，是对尤纳斯责任伦理中对子孙后代负责的延伸。从各国出台的人工智能伦理准则来看，内容上有很多共同之处。比如美国电气电子工程

师协会（IEEE）在 2017 年发布的“人工智能设计的伦理准则”白皮书中，明确了五个一般原则，包括了“人权、福祉、问责、透明和慎用”。欧盟在 2019 年提出的“可信赖的人工智能”伦理准则下，讨论了七个关键条件，包括了“人的能动性和监督能力、安全性、隐私数据管理、透明度、包容性、社会福祉、问责机制”。同年，我国在《新一代人工智能伦理规范》中提出了六项基本伦理要求，分别为“增进人类福祉、促进公平公正、保护隐私安全、确保安全可控、强化责任担当、提升伦理素”。这些目前引领全世界人工智能技术发展的主要国家和地区，为发展符合人类长远利益的技术而提出了伦理约束，从这些伦理约束中能够发现高度一致的增进“人类福祉”这一目标。2015 年在美国纽约联合国总部召开的“联合国可持续发展峰会”上，围绕“变革我们的世界---2030 年可持续发展议程”这一主题，针对人类、地球与繁荣制定了 17 个可持续发展目标。17 个发展目标，对应着“消除极端贫穷、战胜不平等和不公正以及遏制气候变化”，[①]其根本指向就是增进人类福祉。人工智能作为时下最热门的且是支撑下一次工业化转型与升级的核心技术之一，肩负着改善人类生活水平，提升人类幸福感和促进生态平衡的艰巨任务。2019 年底，新冠肺炎疫情暴发瞬间席卷全球，在这场人与病毒的较量中，“数字哨兵”（结合了佩戴口罩的人脸识别、体温检测、扫描健康码和接种疫苗信息等多项功能于一体）、“智能消毒机器人”（能够均匀喷洒消毒液，提升效率的同时保障了防疫工作者和公众的健康安全）成为了与人类并肩作战的“战友”。此外，科学家们还利用 AI 序列优化算法对病毒的类型进行推算，以应对复杂多变的变异毒株，极大地缩短了研发药物和疫苗的时间。就增进人类福祉这一角度来看，疫情期间人工智能技术的应用符合人们对于这一目标的追求。这也与全球可持续发展的理念不谋而合，因此推进人工智能技术的发展能够符合伦理原则，本身也促进了全人类的可持续发展，二者之间有着密不可分的联系。

（三）构建智能联结的社会

虽然现在还不能够确定人工智能技术在未来会发展到什么程度，但是可以确定的是全世界都在为构建一个智能化的社会而努力。2013 年德国提出了工业 4.0 计划，这是将工业生产过程中所涉及的供应、制造和销售的各环节进行数字化的云端共享，以实现物联网的智能化生产，从而提升人们的生活质量。在德国工业 4.0 概念的基础上，中国提出了“中国制造 2025”，这是推动中国制造业从制造大国转向制造强国的第一个十年计划，也是持续推进我国制造业转型升级的行动纲领。纲领中明确提到以推动智能制造为主攻方向，这与深化人工智能技术的发展和广泛运用之间有着紧密的联系。这意味着，世界未来的发展趋势，是朝向智

① 经济和社会事务部可持续发展.[EB/OL].https://sdgs.un.org/cn(访问时间：2022 年 7 月 25 日）

能化的方向不断地创新和升级。在 10 年前，万物互联还处在一个初步的设想与搭建的阶段，而在今天正在变成现实。2022 年 11 月，华为在广东东莞召开的第四届华为开发者大会上，向世界展示他们 OpenHarmony 的行业生态进展，这是专门为万物互联所研发的开源系统。这个通用的底层系统可以广泛地应用于教育、金融、交通、工业和政务等多个领域，使得这些行业可以相互联通，相互协同。比如大会上教育版块开发的“智慧学生证”，通过与网络连接，将学生的个人信息汇聚在卡片上，简单的刷卡之间完成许多数据繁琐的登记、统计和填表等工作，这个智慧学生证还配备了通话功能，可以让学生直接与老师和家长进行联系。而在金融领域，人们最为熟悉的电子支付也在发生改变，中国作为全世界最大的电子支付市场，正在努力摆脱国外资本的掌控，让经济效益更多地流向国内。在交通板块，OpenHarmony 在道路、交通和隧道的监管与维护方面进行着智能化的努力，积极改善从前设备之间缺乏联通性的问题，同时简化了操作方式。这些智能化的应用与万物互联的进一步实现，为搭建人机共生环境提供了现实的基础，人类正在面临生产方式和生活方式的重大变革。智能社会的搭建，意味着人类的生活场景在未来会发生不小的改变，人与人、人与物、物与物的关系都将转向一种智能化的联结。在这样的趋势下，何为人机和谐共生呢？

以往人们将生态伦理思想中和谐共生的理念应用于构建人与自然的关系上，强调了人与自然之间是相互交融而非两极对立的关系，以达到消解人与自然之间的矛盾，实现人与自然共荣共存的良性发展。将这一概念运用到人机关系上，实际是对和谐共生内涵的延伸，也是对人与社会关系的一种外延。人类的持续生存与发展不仅依赖于稳定的生态系统，还需建立在稳固的社会环境基础上。在智能联结的社会框架下，人的自由全面发展仍然是全人类的共同追求。智能社会的构建是人机共生的大前提，也是技术作为背景而存在的现实表现，由智能联结的人类社会，技术对人类生产与交往的影响都将达到前所未有的高度，人们希望的是在未来社会中，能够利用这些技术，更舒适并且有尊严地生活，因此促进未来智能社会人机共生关系的和谐发展，这是至关重要的。

二、明确人工智能技术伦理治理的原则和准则

人工智能技术的伦理治理原则是贯穿于伦理问题的底线标准，是实施准则的理论根源，具有一般性。而伦理准则更为具体，是规范人工智能技术能做什么和不能做什么的具体标准。人工智能技术的伦理原则是对于人的行为，人的意识的根本指导，而准则具体对应到技术负效应的各种问题，提出善的要求。

（一）人机关系的伦理原则

1. 坚持以人为本

马克思主义科技伦理思想中的人本意蕴是人工智能技术发展坚持以人为本

原则的理论来源，强调科学技术的发展应该服务于人民群众的利益和需要，把人的根本利益放在科技发展的核心地位。阿西莫夫的机器人三大定律，也体现了“以人为本”这一根本尺度，即定律的第零条“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”这是机器人出现的前提。伦理是人工智能技术发展的前置性规范。人工智能技术设计最终是为了人类的福祉，设计的目的是为了人，因此，人工智能设计基础也应该以人为本，在人工智能设计中优先考虑伦理问题，人工智能设计要符合人类的价值观和伦理原则。[①]以人为本原则强调了人工智能技术始终把人的需求作为一切工作的出发点与落脚点，其设计理念和功能实现都应该以人为根本目的造福于全人类。马克思主义科技伦理思想的人本意蕴是以人为本原则存在的依据，强调了技术作为生产力推动人类社会发展的重要性，同时也揭示了技术不应该阻碍人的自由全面发展，而是服务于人类的解放。2006 年美国电气与电子工程师协会，将“合乎伦理设计”作为发展自主性人工智能技术的指导方针，其目的是造福全人类，充分体现了以人为本原则的积极意义。

2. 坚持和谐共生

长久以来，人类一直渴望实现与自然的和谐共处，并不断追寻可持续发展的美好愿景。而人类社会的繁荣与稳定，则依托于人与人之间的和谐共处，以及人类文明的延续。我们的生活和行为都建立在人与人、人与社会、人与自然之间的复杂关系中。因此，我们所生存的世界，实质上就是一个多维关系织成的世界，而伦理则是人们在这个世界上生存需要遵循的道德原则和规范。但今天的世界出现了一个极为重要的新情况，就是人工智能技术的飞速发展使一种新的物——智能机器也出现在伦理思考的范围之内。[②]在社会逐步智能化的今天，人与人，人与社会之间的联结正在发生着改变，一种新型的关系正在挑战着人的对于“关系”的认知，那就是人－机之间的关系（文中人－机关系中的“机”指代的是人工智能技术产品，包含智能机械、设备和软件系统等）。因此未来人类社会的繁荣发展，除了应该建立在人与人、人与社会，人与自然和谐关系的基础上，还应该建立在人－机关系和谐稳定的基础上，人与技术之间存在着紧密的联系。人与技术在人－机关系中应当被视为一个整体，二者之间存在互动和依存关系。这种理解有助于建立和谐的人机关系，并具有积极的意义。人－机之间坚持和谐共生的原则，正是在构建和谐稳固的社会基础上，对人提出的新的要求。

人机和谐友好原则的实现，汲取了生态伦理思想中人与自然关系和谐共处的理念，以及马克思主义科技伦理思想中人的主体性作用。首先，人与自然的关系

① 闫坤如. 人工智能“合乎伦理设计”的理论探源[J]. 自然辩证法通讯，2020（04）：14-18.

② 何怀宏. 人物、人际与人机关系——从伦理角度看人工智能[J]. 探索与争鸣，2018，345（07）：27-34+142.

经历过漫长而又曲折的演变，从前人类肆意的蹂躏土地，但如今人类意识到人身为自然界有机质的一环，应该在能力与责任中寻找新的生存方式。利奥波德在他的生态伦理思想“像山那样思考”“荒野价值”“大地共同体”中，将道德对象从人以及人类社会扩大到整个自然界，因此，利用这一观念来构建智能化发展趋势下的人－机社会，依托建立起人机和谐友好的关系，来降低技术对人类生存的威胁，消解人对人工智能技术的恐慌。在具体的实践中，可以通过体会“人工智能技术”产品设计理念和使用功能，来达到对人工智能技术的理解和包容。人工智能技术的是人造物，但是也有其客观性和独立性，因此在使用的过程中要尊重其自身的发展规律，怀揣着对他者的尊敬与热爱，消解对立和冲突。

其次，在人－机构建友好关系的过程中，人应该遵循马克思主义科技伦理思想中对于人的主体性和主观能动性的要求，去积极解决构建新型关系过程中遇到的问题，比如对已经出现的问题伦理问题积极地解决，使其发展应该为人类社会服务，并且受到人类的控制和管理。

最后，通过对人类与自然以及人类与技术之间关系的思考，可以进一步理解人与人工智能技术之间的关系。意识到人与人工智能之间是相互依存又互相独立的关系。为了促进双方的共同发展，需要人类去实践人与技术优势互补的发展观，共建人与人工智能技术的关系向友好型的相处模式转变，重塑人对“关系”认知，这也是人类思考“生而为人”的使命感与价值意义的探索。因此，和谐共处的友好原则不仅满足于人类朝向智能化社会发展的现实要求，同时也在人与技术的矛盾与冲突中寻找和解，从而实现人－机－自然三个维度的和谐共生，消解人与自然，人与技术以及技术与自然之间的矛盾，实现更高维度的和谐共生。

（二）制定人工智能技术伦理治理的准则

在以人为本原则与和谐共生的友好交往原则的共同指导下，面对人工智能技术与现实世界之间发生的具体伦理问题，还应该借助具体的准则帮助实现人工智能技术的伦理治理目标。伦理准则，可以明确人工智能技术能做什么和不能做什么。比如面对聊天机器人的性别歧视、面部识别功能的种族歧视和 AI+HR 的人才管理等对人类公平与正义的挑战，以及信息茧房、过滤气泡等对个人喜好和隐私的侵犯带给人们对生活的不安全感等等，都需要通过伦理准则来约束技术的发展尺度。

1. 确保安全和不伤害

在马斯洛的需求理论中，高级的需求往往建立在较低级的需求被基本满足之后，科技发展和文明进步，让人类对需求金字塔底部生理需求得到了基本的满足，因此对于现代人而言，安全需求是目前人类最重要同时也是最基本的需求。将安全需求上升到对人工智能技术的约束中，是为了确保人工智能不会伤害人类。关

于“不伤害”的解释，康德认为除去人的生命不遭遇危险之外，还应该包括人的心理不受伤害。因此不伤害的伦理准则，是为了保护人的身体不受到人工智能技术的伤害，同时也不会因技术产生心理上的自卑。随着智能化生产在工业领域的迅速普及，机器人使用率迅速增长，工厂里由人+机器人协同作业的景象变得越来越常见，但是企业工厂对于机器人的安全使用存在着巨大隐患，从上世纪末的机械手臂出现以来，机器伤人的事件就从未停止，虽然发生事故时机器有急停的功能，但是一旦机器抱闸锁死，人也很难从重量是自身数倍的机器下快速脱身。要想避免这种伤人事件的发生，需要研发人员在构思产品时，带入一种同理心，“从平等之人所享有的人权中，导引出的义务首先表现为不伤害，不伤害是所有的人必须认同也能够遵循的行为准则，是在陌生人充斥的宏大社会里最基本的底线共识。”[①]应该将确保人的安全和不受伤害，作为机器人诞生的首要任务。保护人的安全和不伤害这一伦理准则，应该是人与机器互动协作的最低限制。

2.促进公平正义

公平的含义是指不偏不倚，世界上不存在分毫不差的绝对公平，相对公平才是现代社会普遍追求目标。正义是伦理学的基本范畴之一，柏拉图在《理想国》中将正义作为开篇探讨的问题，书中苏格拉底认为正义是智慧与善。公平的社会和制度是正义生长的土壤，二者之间相互联系，辩证统一。人对公平正义的追求，反映了人类对文明的向往，体现了一个国家和社会文明发展的程度。现代技术的发展，创造出新事物来取代旧事物，新的职业取代旧的职业，因此也总有一些人能够更快适应新事物，掌握新的知识从而获得更多的机会，这加剧了人与人之间的竞争与不平等。二十一世纪的人工智能技术，如果不遵循公平正义的伦理准则任其发展，不公平的社会现象将会被无限放大，由单一或少数指标定义的智能化信息处理系统，会忽视掉环境对人的影响，忽视人自身的复杂性和多样性。因此，人工智能技术的发展应该充分考量人对公平正义的渴望，这包括让每一个人都可以享有使用人工智能技术的权利，享有平等的技术支持，不会因为种族、性别、地域或宗教信仰的差异受到歧视，同时依据该准则，企业在开发技术时加入反歧视的算法，对产品的设计与使用进行监督，国家出台相应的管理条例，对妨碍公平正义实现的企业和个人采取相应的管理方法。实现人工智能技术的公平正义不是少数人的力量就能够完成的，需要全社会的共同监督。

3.保持公开透明

人工智能技术的公开透明伦理准则，是确保安全不伤害准则和公平正义准则实现的基础性准则。属于人工物范畴的人工智能技术及其产品，在设计过程中包含着人类的意图，如搜索引擎公司会利用算法将广告费用投入高的网站放在搜索

① 甘绍平.伦理学的当代建构[M].北京：中国发展出版社，2015：16.

结果的前排，诸如此的信息茧房，过滤气泡等都是算法编写者有意为之的结果。算法编写者的不良意图、数据样本的迭代不及时，以及机器自主学习出现的偏见，都会导致人工智能技术做出歧视行为或造成伤害事件的发生，因此在必要时公开算法是对人类利益的保护。公开透明的伦理准则就要求将人工智能算法模型和运行系统向公众进行公开，让人们了解技术产品的设计过程，从而减少人们对于人工智能技术的不信任和恐慌。但是现实的情况是此相关的维权案数不胜数，当用户遭遇被算法割韭菜或者个人信息被窃取时，很难得到有效的维权。因为算法属于技术类知识产权，本身具有秘密性、价值性、实用性及保密性。[①]因此成为了算法开发者逃避侵权责任的武器。公开透明是让开发商个人算计无处可藏的伦理准则，是对广大用户个人利益的最基本的保障。这将有助于提高人工智能系统的可信度和可靠性，并使公众更能理解和接受人工智能技术的应用。这一准则，是在约束设计开发人工智能技术的企业和个人，自觉地接受群众监督，审慎地使用用户的个人数据，在算法模型和运行系统透明度层面不再依赖于法律的强制公开，而是积极主动地承担社会责任，促进人工智能技术向善发展。

三、完善人工智能技术伦理治理的具体路径

（一）构建人工智能技术多元共治的人文环境

多元共治是我国社会治理实践中总结的宝贵经验，是社会治理的制度创新，利用社会共治中多元主体之间的相互协作解决现实问题。多元共治最鲜明的特征就是多元主体参与到治理体系中，如今人工智能技术在更多领域得到使用，技术与人的嵌合度越来越深，这就需要更多主体融入到治理体系中，参与人工智能技术的设计、试验、使用和推广的各个阶段。多元共治跳脱出了传统的由政府、市场和社会组织构成的治理体系，将公民个人（如学者、教师、医生等行业的专业人才都是多元共治的主体）也纳入到治理的体系中一同献计献策。同时，多元的治理方式不仅局限于主体的丰富，也在空间维度上完成多元，实现跨国家和地区、跨部门和领域的共同治理。对人工智能技术采取多元共治，不仅是由于技术与人融合的程度不断加深，还是由于人工智能技术本身就是以计算机科学为基础的，由计算机、心理学、哲学等多学科交叉融合的产物，自身就体现了多元性的特点。因此，人工智能技术从诞生之初就包含了科学的理性和浓厚的人文性。随着社会智能化的不断发展，人工智能技术与社会生活的不断融合，吸取多元共治的治理理念来约束人工智能技术向善发展，营造万众参与群体献策的人文环境，是未来人工智能技术健康发展必经之路。

（二）提高科学共同体的伦理素养

① 吴椒军，郭婉儿.人工智能时代算法黑箱的法治化治理[J].科技与法律(中英文)，2021（01）：19-28.

汉斯·尤纳斯责任伦理思想中强调了关注科学共同体的伦理责任问题，通过预测可能存在的风险来规范技术发展，从而应对技术伦理风险。因此，从事科学研究工作的研究人员，有责任向公众告知技术可能带来的不良影响。正如美国管理学家西蒙在他的自传中所说的那样“关于社会后果，我认为每个研究人员都有责任评估，并努力告知其他人，他试图创造的研究产品可能带来的社会后果。”[①]伦理素养不仅关乎科学研究的道德和法律合规性，更关乎人类文明的发展和未来。只有具备高度的伦理素养，科学研究才能更好地为人类服务，因此提高科学共同体的伦理素养可以有以下措施。

首先，加强科学共同体的伦理教育。一方面，企业、院校和研究所可以组织职业培训和学术交流活动，系统地传授伦理理论、科学道德和法律法规等方面的知识，让这些知识成为科学共同体底层意识的一部分。另一方面，邀请伦理专家和学者举办伦理研讨会和讲座，为科研人员提供伦理知识和实践经验的分享，这样不仅可以增强科研人员的伦理素养和还可以增强专业素养。

其次，需要加强伦理监管。通过设立人工智能伦理委员会和相应的伦理举报制度，对科研人员的研发行为进行监管，一旦出现触碰伦理问题的敏感项目，可以通过监督机制对研究全过程进行监督和管理，这样可以防止科学研究中出现的违法违规行为，维护科学研究的公正性和透明度。

最后，应该通过加强国际合作，分享各国实践中的优秀案例，进行经验分享和交流，进一步推动国家之间对人工智能技术伦理问题的探讨，加强合作，达成共识，实现科学共同体伦理素养的提高。

科学共同体需要注重伦理责任和社会责任。科学研究不仅仅是为了追求科学发现和技术创新，更应该考虑科学研究对社会和环境的影响。因此，科学共同体需要承担起伦理责任和社会责任，确保科学研究的可持续发展。只有通过加强伦理教育、监管和责任担当，才能推动科学研究走向更加健康、可持续的发展之路。

（三）加强对社会公众对科技伦理的科普

当前，高新技术的发展通常伴随着政治和经济的大量投入，然而基层的科研工作者和公众对科技伦理的导向往往被忽视。因此，科技伦理素养的缺失已经成为十分严重的社会问题，这种缺失不仅仅体现在个人行为上，也会反向渗透到科技、媒体和政治等领域，给社会带来许多负面影响。为解决这一问题，就需要加强社会公众的科技伦理教育。汉斯·尤纳斯的责任伦理思想不仅适用于科学共同体，同样可以为提高公众的科技伦理素养提供借鉴，包括强调个人责任、关注道德风险、倡导公众参与和重视伦理教育等方面。因此，人们需要认识到科技伦理

① [美]赫尔伯特·A·西蒙.我生活的种种模式——赫尔伯特·A·西蒙自传[M].曹南燕等译.上海：东方出版中心，1996：351.

素养的重要性，科技伦理素养不仅是个人行为的基础，也是技术良性发展与繁荣的基石。若没有科技伦理的积极引导，社会大众使用科技的理念和水平就很难有提升，从而导致伦理滑坡等多种问题的出现。因此，提高社会公众的科技伦理素养，对于构建人机共生的和谐社会具有重要意义。

基于提高公众科技伦理素养的整体水平，需要采取多种措施加强科普工作。一方面，政府应该加大对人工智能技术伦理科普的投入和支持，通过广告、电视节目、网络宣传渠道宣传人工智能技术的利与弊，让公众认识到技术既可以带来正面的效果，也会带来负面的效果，因此要审慎使用技术和技术产品，比如利用媒体宣传科技企业的道德典型。另一方面，社会各界也参与到科技伦理科普的工作中，如在中小学开设的人工智能课程的同时融入科技伦理课程，将中华传统文化中“善”的理念融入科技观当中，让孩子们明白利用科技不仅可以改变生活还可以让生活更加美好。通过加强公众对于人工智能技术的了解和认知，配合相应的科技伦理科普，提高公众的伦理素养，促进公众对人工智能技术树立起正确认识和理解。通过提高科技伦理素养，个人或组织可以更加负责任地使用和推动科技发展，规避可能带来的潜在风险和误解。为人类在智能社会能够获得更有尊严更有幸福感的生活，人们应该共同努力以良好的科技伦理素养为基础，共建美好的人机社会。

（四）完善人工智能技术的制度建设

随着人工智能技术的不断发展，人们必须面对的一个重要问题是如何建立健全人工智能技术的伦理治理制度，以确保其发展和应用符合道德和社会价值观，认识到人工智能技术的发展和应用对社会和个人产生的影响。例如，人工智能系统可能导致工作岗位的流失，带来职业转型的需要；可能会影响个人隐私和自由；还可能对人类价值观产生挑战。这些影响和挑战需要一个全面的、系统的解决方案。

首先，需要建立一个全面的、透明的以及公正的人工智能技术伦理治理制度，它应该考虑到政府、企业、学术界、社会团体和公民等不同利益相关者的角度。政府应该扮演领导角色，制定和执行政策和规定行业的发展标准，推动人工智能技术伦理治理准则的制定和确立，以确保技术得到合法和合理的使用。这些准则包括人工智能技术伦理治理的基本原则和指南，以及针对技术开发和使用过程中可能出现的伦理问题的解决方案，比如需要规定人工智能技术开发和使用过程中的安全和隐私保护措施，以保障技术的安全和合理使用。企业应遵守相关规定，承担社会责任和促进技术的可持续发展，以确保人工智能技术的发展和应用符合社会价值观。学术界应该开展相关研究，提出科学建议，以引领人工智能技术的发展方向。民间组织和普通公民应积极参与公共讨论和决策，以确保技术的发展

和应用符合公共利益。基于此，必须考虑到各种利益相关方的需求和关注，不断完善人工智能技术伦理治理的制度建设。

其次，面对同样一件技术事故，人们利用伦理约束能够得到的效果，远没有法律的强制力和经济利益的风险评估来得有效果。主要是因为，良知本身没有严格的标准。因此，需要建立人工智能技术伦理治理的法律框架。这些法律应该明确规定技术的使用和开发范围，限制人工智能技术在对人类造成伤害的情况下的使用，并规定其开发和使用过程中的法律责任。法律框架还应该包括人工智能技术伦理治理机构的设立和职责。这些机构可以监督人工智能技术的开发和使用，制定伦理准则和标准，评估和监督技术的风险和影响。

最后，在制度和法律的双重规制下，还需要建立与之相匹配的监管制度，加强人工智能技术伦理治理制度的执行和监督。在这个方面，可以考虑使用技术工具，例如区块链，实现人工智能系统的可追溯性和透明性以及确保人工智能技术的合法和合理的使用。此外，还可以建立独立的评估机构和审查委员会，对人工智能系统进行评估和审查，以确保其符合相关的政策和标准。综上所述，完善人工智能技术伦理治理的制度建设需要建立伦理治理的法律框架、行业标准与准则和社会共识。这将有助于保障人工智能技术的安全和合理使用，促进人工智能技术的发展和进步。

结 论

（一）人工智能技术引发的伦理问题，是指在人工智能技术的研究、开发、应用和评估过程中，出现可能侵犯人类权利、价值观、道德准则等方面的问题。其中有对人产生直接影响，并受到人们广泛关注的显性伦理问题，包括人工智能技术的滥用、失控和管理缺失。也有技术作为背景时，对人产生潜移默化影响的隐性伦理问题，包括人工智能技术道德地位模糊、主体交往异化和人类生存危机隐患。人们需要在认清问题的基础上，对人工智能技术的发展树立正确的认识。随着人工智能技术的不断发展，新的问题还会不断涌现。因此，需要对人工智能技术的发展保持敏感性。

（二）关于人工智能技术产生伦理问题的原因，本文运用了马克思主义科技伦理思想和汉斯·尤纳斯责任伦理思想对问题进行分析。发现在显性伦理问题方面，原因有人工智能技术存在算法“黑箱”、人工智能技术的工具属性和价值属性失衡和人工智能技术的利益相关者缺乏责任意识。而在隐性伦理问题的方面，原因有人工智能技术存在自我意识拟人性与物理构造非人性之间的矛盾、人工智能技术对社交秩序的重构和人工智能技术对劳动力就业的干扰。

（三）人工智能技术伦理问题的治理旨在规范人工智能技术的研究、开发、应用和评估，促进其安全、可靠、公正、透明、可控。为了确保人工智能技术的发展符合伦理和道德标准，需要政府、企业、学术界、社会组织等多方合作，共同推进人工智能技术的伦理治理研究和实践。

荀子认为，人能够通过学习通向善，这说明善不依靠与生俱来而需后天习得。人工智能技术也一样，不是从诞生就拥有对“善”的理解，因此需要人“善”的引导。在人与 AI 和谐共生的将来，人应该通过对人工智能技术的研究，认识到一切科学技术的发展最终都指向人认识自己的过程。哪怕 AI 会越来越像人，甚至在能力上超越人，都不应该阻碍人类追求超越自身极限的渴望。技术改变生活，是不争的事实，但是人们需要的不仅仅是改变，而是希望这种改变可以让人更幸福，更有尊严地生活。人工智能技术的伦理治理是一个长期而艰巨的任务，需要不断探索和创新，需要不断推进和完善，需要不断强化和落实。只有这样，才能更好地保护人类的尊严和价值，推进人类社会的进步和发展。

由于学识有限和人工智能技术发展的不确定性，本研究还存在以下问题：第一，在人工智能技术的伦理问题方面，随着技术与人的交往不断密切，伦理问题的具体表现还需要实践的进一步发现；第二，在人工智能技术伦理治理建构方面，不同的国家、组织之间缺乏共识，这导致在技术应用应遵循的法律法规也难以明确。因此，建立统一的伦理标准仍然会是未来伦理治理研究的主攻方向之一。

参考文献

1.中文文献

（1）著作

[1] [美]休伯特·德雷福斯. 计算机不能做什么：人工智能的极限[M]. 宁春岩译. 上海：生活·读书·新知三联书店，1986.

[2] [美]赫尔伯特·A·西蒙. 我生活的种种模式——赫尔伯特·A·西蒙自传[M]. 曹南燕等译. 上海：东方出版中心，1996.

[3] [美]A. 利奥波德. 沙乡年鉴[M]. 侯文蕙译. 长春：吉林人民出版社，1997.

[4] [美]冯·诺伊曼. 计算机与人脑[M]. 甘子玉. 北京：商务印书馆，2010.

[5] [美]艾萨克·阿西莫夫著. 我，机器人[M]. 叶李华译. 南京：江苏文艺出版社，2013.

[6] [美]雷·库兹韦尔. 人工智能的未来[M]. 盛杨燕译. 杭州：浙江人民出版社，2016.

[7] [美]温德尔·瓦拉赫,科林·艾伦. 如何让机器人明辨是非[M]. 王小红译. 北京：北京大学出版社，2017.

[8] [美]杰瑞·卡普兰著. 人人都应该知道的人工智能[M]. 汪婕舒译. 杭州：浙江人民出版社，2018.

[9] [美]理查德·费曼. 发现的乐趣[M]. 朱宁雁. 北京：北京联合出版公司，2018.

[10] [美]欧内斯特·戴维斯，盖瑞·马库斯. 如何创造可信的 AI[M]. 龙志勇译. 杭州：浙江教育出版社，2020.

[11] [德]汉斯·尤纳斯. 责任原理——现代技术文明伦理学的尝试[M]. 方秋明译. 香港：世纪出版集团，2013.

[12] [德]卡尔·马克思，[德]弗里德里希·恩格斯. 马克思恩格斯全集（第 31 卷）[M]. 北京：人民出版社，1971.

[13] [德]卡尔·马克思. 1844 年经济学哲学手稿[M]. 北京：人民出版社，2000.

[14] [英]玛格丽特·A·博登. 人工智能哲学[M]. 刘西瑞等译. 上海：上海译文出版社，2006.

[15] [日]福田雅树，林秀弥，成原慧. AI 联结的社会：人工智能网络化时代的伦理与法律[M]. 宋爱译. 北京：社会文献出版社，2020.

[16] [以色列]尤瓦尔·赫拉利著，未来简史:从智人到智神[M]. 林俊宏译. 北京：中信出版社，2017.

[17] 林崇德，杨治良，黄希庭. 心理学大辞典[M]. 上海：上海出版社，2003.

[18] 聂明. 人工智能技术应用导论[M]. 北京：电子工业出版社，2019.

[19] 腾讯研究院，中国信通院互联网法律研究中心，腾讯 AI Lab 腾讯开放平台著. 人工智能[M]. 北京：中国人民大学出版，2017.

[20] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[21] 李仁涵，黄庆桥. 人工智能与价值观[M]. 上海：上海交通大学出版社，2021.

[22] 甘绍平.伦理学的当代建构[M].北京：中国发展出版社，2015.

[23] 尼克.人工智能简史[M].北京：人民邮电出版社，2021.

[24] 胡云冰，何桂兰，陈潇潇等编著.人工智能导论[M].北京：电子工业出版社，2021.

[25] 王前主编.技术伦理通论[M].北京：中国人民大学出版社，2011.

[26] 陈小平.人工智能伦理导引[M].合肥：中国科学技术大学出版社，2021.

[27] 余谋昌.生态伦理学——从理论走向实践[M].北京：首都师范大学出版社，1999.

（2）期刊论文

[28] 程鹏，高斯扬.通用人工智能体道德地位的哲学反思[J].自然辩证法研究，2021(07).

[29] 张冬烁.从大地伦理学到伦理整体观理论——克利考特生态伦理学研究[J].玉溪师范学院学报，2007(9).

[30] 包庆德，夏承伯.土地伦理:生态整体主义的思想先声——奥尔多·利奥波德及其环境伦理思想评介[J].自然辩证法通讯，2012，34(05).

[31] 闫坤如.人工智能“合乎伦理设计”的理论探源[J].自然辩证法通讯，2020(04).

[32] 赵汀阳.人工智能“革命”的“近忧”和“远虑”——一种伦理学和存在论的分析[J].哲学动态，2018(04).

[33] 阴训法，陈凡.论“技术人工物”的三重性[J].自然辩证法研究，2004(07).

[34] 刘海龙.马克思科学技术观的人本意蕴[J].社会主义研究，2011，197(03).

[35] 陶映帆，胡鹏飞，杨文明.智能家居中的计算机视觉技术[J].人工智能，2020(05).

[36] 王前，曹昕怡.人工智能应用中的五种隐性伦理责任[J].自然辩证法研究，2021，37(07).

[37] 吴椒军，郭婉儿.人工智能时代算法黑箱的法治化治理[J].科技与法律(中英文)，2021(01).

[38] 高亮华.技术失控与人的责任——论弗兰肯斯坦问题[J].科学与社会，2016(03).

[39] 王亮.社交机器人“单向度情感”伦理风险问题刍议[J].自然辩证法研究，2020(01).

[40] 钱学胜.回顾 4 届世界人工智能大会历程,透视 AI 产业发展特征[J].张江科技评论，2021(05).

[41] 徐英瑾.具身性、认知语言学与人工智能伦理学[J].上海师范大学学报(哲学社会科学版)，2017(06).

[42] 郭林生，刘战雄.人工智能的“负责任创新”[J].自然辩证法研究，2019(05).

[43] 闫坤如.人工智能机器具有道德主体地位吗？[J].自然辩证法研究，2019(05).

[44] 孙波，周雪健.人工智能“伦理迷途”的回归与进路——基于荷兰学派“功能偶发性失常”分析的回答[J].自然辩证法研究，2020(05).

[45] 陈凡，曹继东.现象学视野中的技术伊代——技术现象学评析[J].自然辩证法研究，2004(05).

[46] 杨庆峰.从人工智能难题反思 AI 伦理原则[J].哲学分析，2020(02).

[47] 张浩鹏，夏保华.人工道德智能体何以可行——基于对价值敏感性设计的审视[J].自然辩证法研究，2021(04).

[48] 张立文.和合伦理道德论——中华传统道德精髓与人工智能[J].社会科学战线，2018(08).

[49] 王萍萍.人工智能时代机器人的伦理关怀探析——以《老子》“善”论为视角[J].自然辩证法研究，2021(05).

[50] 何怀宏.人物、人际与人机关系——从伦理角度看人工智能[J].探索与争鸣，2018，345(05).

[51] 田海平.伦理治理何以可能——治理什么与如何治理[J].哲学动态，2017(12).

[52] 梅亮，陈劲，吴欣桐.责任式创新范式下的新兴技术创新治理解析——以人工智能为例[J].技术经济，2018，37(01).

[53] 王勉，黄颖.人工智能技术伦理治理现状及趋势[J].电子商务，2019，235(07).

[54] 何哲.人工智能技术的社会风险与治理[J].电子政务，2020(09).

[55] 吴汉东.人工智能时代的制度安排与法律规制[J].法律科学，2017(05).

[56] 吴红，杜严勇.人工智能伦理治理：从原则到行动[J].自然辩证法研究，2021(04).

[57] 王钰，程海东.人工智能技术伦理治理内在路径解析[J].自然辩证法通讯，2019，41(08).

[58] 汪琛，孙启贵，徐飞.国际人工智能伦理治理研究态势分析与展望[J].医学与哲学，2022，43(07).

[59] 樊春良，张新庆，陈琦.关于我国生命科学技术伦理治理机制的探讨[J].中国软科学，2008，212(08).

[60] 于雪，段伟文.人工智能的伦理建构[J].理论探索，2019(06).

（3）学位论文

[61] 浮婷.算法“黑箱”与算法责任机制研究[D].中国社会科学院研究生院，2020.

[62] 郭延龙.技术人工物设计伦理转向研究[D].中国科学技术大学，2020.

[63] 薛峰.人工智能对马克思劳动理论的影响研究[D].上海师范大学，2020.

[64] 高杨帆.技术决策者伦理责任研究[D].华中师范大学，2013.

2.电子文献

[65] 亿欧网.德国公布首份自动驾驶伦理道德标准.

[EB/OL].https://www.iyiou.com./gn/briefing/2018/5-16/15690.html,2018-5-16.

[66] 济南市大数据局.什么是大数据？

[EB/OL].http://jndsj.jinan.gov.cn/art/2019/8/13/39428_3158286.html,2019-8-13.

[67] 经济和社会事务部可持续发展.

[EB/OL].https://sdgs.un.org/cn(访问时间：2022 年 7 月 25 日）

[68] 央广网.欧盟发布人工智能伦理标准.

[EB/OL].https://baijiahao.baidu.com/gn/nems/2019/4-11/1630485499170176.html,2019-4-1.

[69] 发展负责任的人工智能. 中国社会科学网.

[EB/OL].http://www.gov.cn/xinwen/2019/6/17/content_5401006.html.2019-6-17.

[70] 中华人民共和国科学技术部. 新一代人工智能伦理准则

[EB/OL].https://www.most.gov.cn/kjbgz/2021/9/26/177063.html,2021-9-26.

[71] 中华人民共和国科学技术部. 关于加强科技伦理治理的意见.

[EB/OL].https://www.most.gov.cn/xxgk/2022/3/25.180004.html,2023-3-25

3.外文文献

[72] Kate Crawford.Atlas of AI Power, Politics, and the Planetary Costs of Artificial Intelligence[M].London:Yale University Press,2021.

[73] Roman V. Yampolskiy, PhD.Artificial Intelligence Safety and Security [M].London:CRC Press,2018.

[74] Kurzweil R.The Singularity Is Near:When humans transcend biology[M].London:Penguin Book,2005.

[75] A.M.Turing. Computing Machinery and Intelligence[J] Mind,1950.

[76] FloridiL,Sanders J W.On the Morality of Artificial Agents[J].Minds and Machine,2004.

[77] James H.Moor. The Nature, Importance, and Difficulty of Machine Ethics'[J].IEEE Intelligent Systems,2006.

[78] Wachsmuth I.Robots Like Me: Challenges and Ethical Issues in Aged Care [J].Frontiers in Psychology,2018.

[79] Jobin A, Ienca M, & Vayena E.The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence,2019.