本合同是否为中小企业预留合同：□是/√否。

课题编号	20250264009
课题分类	二类调研

北京市科学技术委员会、
中关村科技园区管理委员会
课题研究合同

甲方：北京市科学技术委员会、中关村科技园区管理委员会

乙方：北京邮电大学

专项名称：中关村战略规划与政策研究资金

课题名称：北京人工智能科技伦理治理规范研究

研究时间：2025 年 9 月-2026 年 7 月 31 日

北京市科学技术委员会、中关村科技园区管理委员会制

甲方：北京市科学技术委员会、中关村科技园区管理委员会

法定代表人：张继红

地址：北京市通州区宏安街 9 号

课题承担处室：信息科技处

课题承担处室负责人：韩健

联系电话：010-55577909

课题承担处室主管工程师：胡越

联系电话：15070301230

乙方：北京邮电大学

法定代表人：徐坤

地址：北京市海淀区西土城路 10 号

课题负责人：周琳娜

联系人：杨忠良

联系电话：18811536956

一、合同内容

依据《中华人民共和国民法典》，甲方将 北京人工智能科技伦理治理规范研究 课题由乙方承担，经双方充分协商，达成一致，签订本合同，双方共同信守。

1.购买内容

2024 年 12 月，市科委、中关村管委会等八部门印发《关于加强科技伦理治理的实施意见》，明确提出加强对人工智能在医疗、金融、交通、教育、文化等应用场景中的科技伦理治理，建立健全相关标准，明确具体要求，研究制定人工智能研发、应用等具体活动的科技伦理规范和通用大模型等前沿技术的伦理审查标准，加强对具有舆论社会动员能力和社会意识引导能力的算法模型，以及对人类主观行为、心理情绪和生命健康等具有较强影响的人机融合系统的研发等纳入清单管理的科技活动的伦理审查和复核。为落实《关于加强科技伦理治理的实施意见》，推动北京市构建具有引领力的人工智能伦理治理体系，现委托具备相应研究能力和实施经验的单位承担“北京人工智能科技伦理治理规范研究”项目。

该项目聚焦通用大模型、医疗、金融等典型人工智能应用场景中所面临的伦理挑战，通过研究建立指标体系、操作手册与审查标准，支撑北京市相关政策落地和治理能力提升。具体工作内容包括：

（1）研究建立涵盖公平性、可解释性、风险预警的评测指标体系，形成全流程风险评估框架；

（2）面向医疗、金融等应用场景，制定差异化的伦理操作手册，建立算法责任追溯机制，提出防范大数据杀熟、信息茧房的技术路径，探索“技术合规+伦理嵌入”的双轨治理模式；

（3）聚焦通用大模型等前沿技术，设计动态伦理审查清单，构建包含舆论引导力评估、心理影响测试的复合伦理审查模型。采用“政策对标-场景剖析-技术验证”三维路径，突破跨领域治理协同、动态风险预警等难点。

（4）汇总研究成果，提交《人工智能伦理治理与测评要求研究报告》1 份、《人工智能在重点领域研发、应用的科技伦理规范建议》1 份、《通用大模型伦理审查标准建议》1 份，为相关部门制定标准规范和政策措施提供决策支撑。

2.工作方案（暂定，以最终情况为准）

项目总实施周期为 10 个月，分为调研分析、体系构建和技术验证三个阶段，由承接单位统筹组织、分工协同推进。具体实施方案如下：

（1）在项目启动后 3 个月内完成前期调研工作，系统梳理国内外人工智能伦理治理政策、标准和治理机制，结合北京市发展实际，完成医疗、金融等重点领域伦理风险分析，形成场景剖析报告和问题清单。

（2）中期 4 个月为体系建设阶段，基于调研成果，构建伦理评测指标体系，撰写操作手册草案与责任追溯机制建议，设计复合伦理审查模型与动态清单，组织专家论证并开展初步测试迭代。

（3）后期 3 个月为验证评估与成果提炼阶段，选择具有代表性的应用场景完成模型验证工作，整理技术路径、评估结果与政策建议，形成项目成果材料，接受专家评审，完成结项验收。

3.成果要求

（1）完成《人工智能伦理治理与测评要求研究报告》1 份；

（2）完成《人工智能在重点领域研发、应用的科技伦理规范建议》1 份；

（3）完成《通用大模型伦理审查标准建议》1 份。

二、合同费用

1.合同费用（课题经费）共计人民币 28.8 万元。

2.课题经费拨付：合同签订后一次性拨付课题经费。

3.乙方指定收款账户信息如下：

开户人：北京邮电大学

开户银行：中国工商银行股份有限公司新街口支行

银行账号：0200002909005405044

乙方保证上述信息真实、准确。乙方上述账户信息发生变化的，应至少于甲方付款前 5 个工作日书面通知甲方，否则由此导致的错付、逾期支付、无法支付等所有法律后果由乙方自行承担。

4.乙方应向甲方出具与乙方单位名称一致的正式发票（不包括往来款收据、非经营性收款收据）。

5.课题经费由乙方按照国家及北京市相关经费管理办法和会计制度统一管理，不得转包分包，不得截留、挤占、挪用，不得利用虚假票据套取资金，不得使用课题资金支付各种罚款、捐款、赞助、投资等费用，不得在课题经费中列支与课题研究无关的任何费用。

三、甲方权利义务

1.按合同约定金额提供课题研究经费。

2.为乙方开展课题研究提供必要的协助或便利。

3.掌握研究工作进度，指导监督乙方完成课题任务。

4.有权对乙方工作提出意见和建议，有权对课题验收。

5.有权要求乙方更换不符合要求的课题研究人员。

6.有权依据《北京市市级党政机关课题经费管理办法》和《市科委、中关村管委会课题管理办法》执行课题相关管理工作。

四、乙方权利义务

1.根据甲方课题管理相关要求开展课题研究，并尽勤勉义务。

2.亲自处理受托事务,未经甲方书面同意,不得将本合同项下全部或部分工作转包、分包给任何第三方。

3.乙方应有独立法人资格、健全的财务和课题管理制度，保证研究团队具备完成课题研究所需的相应资信和能力。

4.乙方应严格执行《北京市市级党政机关课题经费管理办法》和《市科委、中关村管委会课题管理办法》有关要求，课题经费使用应当符合国家和本市财政资金使用的相关办法标准和财务审计的要求，专款专用，单独核算。

5.按照甲方要求报告课题研究开展情况，对研究成果进行补充、修改。

6.乙方协助组织课题开题会、结题会，及时提交课题验收材料。

7.课题经费的管理和使用情况接受甲方及相关主管部门的监督检查。

五、课题组织实施要求

1.乙方应按照合同办法完成课题目标任务，不得擅自变更研究内容、课题负责人等。如课题研究条件发生重大变化，乙方应向甲方报告具体情况，研究提出解决方案。

2.甲方负责组织专家对课题研究成果进行评审，并进行课题验收。

3.未通过专家评审，乙方应按照评审会意见修改完善，并于开题会或结题会结束之日后20个工作日内召开第二次评审会。若乙方第二次评审未通过验收或提供的服务无法实现合同目的，甲方有权提前解除合同，并要求乙方退还全部合同款项及赔偿相应损失。

六、履约验收方案

1.履约验收的主体、时间、方式：乙方完成全部工作内容后，由甲方根据项目情况组织验收。

2.履约验收程序：甲方根据合同约定对乙方提交的成果进行验收，验收合格的，视为乙方已交付工作成果。验收不合格的，乙方应当规定时间内按照甲方要求进行返工或调整，并重新提交甲方验收。

3.履约验收的内容：根据磋商文件、响应文件及国家行业有关标准，针对本磋商文件每一项商务、技术要求履约情况进行履约验收。

4.验收标准：乙方为甲方提供的相关服务应符合本项目要求。

七、知识产权及保密条款

1.课题成果的知识产权归甲方所有。

2.经甲方书面同意，乙方可发布、发表部分研究成果，但须标注“北京市科

学技术委员会、中关村科技园区管理委员会支持研究成果”字样。未经甲方书面同意，任何参与课题研究的单位及个人不得擅自披露涉及甲方的有关数据、信息和资料。如发生上述情况，甲方将依法追究相关单位及个人责任。

3.乙方保证，本合同下的研究工作，不存在任何侵权问题，如发生侵权纠纷，由乙方自行承担所有责任，并赔偿甲方因此遭受的损失。

八、合同变更或解除

经甲乙双方协商一致，可以变更或解除本合同。对本合同的变更或解除必须以书面合同进行。双方未签署书面变更或解除合同的，应认定为没有对本合同进行变更或解除。

九、违约责任

1.因乙方原因，导致课题研究工作未能在委托期限内完成，或者课题成果未能达到合同约定目标的，乙方应当采取措施，并承担由此而增加的费用。逾期未通过课题验收的，甲方有权解除合同。

2.乙方因法定的不可抗力不能履行合同义务时，可以免除违约责任，但应在不可抗力事由发生之日起 20 日内通知甲方，并在 30 日内向甲方提供因不可抗力导致合同不能履行的证明。

3.本课题终止或验收未通过的，乙方应依审计结果将结余资金原渠道返还甲方。

十、解决争议

凡与本合同有关的争议，双方应协商解决，经协商不能解决的，任何一方均应向甲方所在地有管辖权的人民法院提起诉讼。

争议解决过程中发生的费用，包括但不限于诉讼费、差旅费、保全费、律师费等，由败诉方承担。

诉讼进行过程中，除双方有争议的部分外，本合同其他部分仍然有效，各方应继续履行。

十一、其他事项

1.本合同一式肆份，甲乙双方各执贰份，具有同等法律效力。

2.本合同自甲乙双方法定代表人或授权代表签字并加盖公章之日起生效。

3.本合同未尽事宜，甲乙双方可另行协商签订补充合同。

4.附件与补充合同与本合同具有同等的法律效力。

附件：课题研究工作方案

（以下无正文）

甲方（盖章）：

甲方（签字）：

（课题承担处室分管委领导）

日期：2025.9.28

乙方（盖章）：

乙方（签字）：

日期：2025.9.28

课题研究工作方案

一、基本信息

课题负责人基本情况					
姓 名	周琳娜	性 别	女	出生年月	1972 年 4 月
专 业	信号与信号处理	学 历	博士	职 称	教授
电 话	/	手 机	13121068100	邮 箱	zhoulinna@bupt.edu.cn
承担单位：北京邮电大学					
通讯地址：北京市海淀区西土城路 10 号				邮编：100876	

课题组主要成员基本情况						
姓名	出生年月	专业	学历	职称	工作单位	主要分工
杨忠良	1993.4	网络空间安全	博士	副教授	北京邮电大学	主要研究人员
尤玮珂	1993.3	信息安全	博士	中级	北京邮电大学	主要研究人员
林钦鸿	1999.1	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员
储贝林	1998.4	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员
宋佳	1997.10	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员
徐璇	2001.6	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员
舒开元	1999.3	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员

李坤	1997.9	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员
陆智高	1997.10	网络空间安全	在读博士生	无	北京邮电大学	一般研究人员

二、选题背景

随着人工智能技术的迅猛发展，尤其是在通用大模型、医疗、金融等关键领域的应用，人工智能正在深刻改变社会生产与生活方式。然而，随着技术的进步，伦理风险问题也逐渐浮出水面，尤其是在数据隐私保护、算法偏见、决策透明性等方面。这些问题不仅影响到个体的隐私权与公平性，也可能对社会治理、经济安全等产生广泛的深远影响。如何有效开展人工智能伦理治理，确保技术发展符合社会价值与伦理标准，成为全球科技界与政策界亟待解决的关键问题。面对这一挑战，国内外已相继提出了一系列伦理框架与政策倡议，但普遍存在实施路径模糊、评估标准不统一等问题。因此，建立一套系统的人工智能伦理治理框架，并通过科学的伦理测评体系实现其在实际应用中的落地，已成为推动人工智能健康发展的必由之路。

1.人工智能伦理治理和测评研究

人工智能作为新一轮科技革命和产业变革的重要驱动力，正以前所未有的速度推动全球社会经济的发展。自 20 世纪 50 年代提出“机器能否思考”的科学设问以来，人工智能经历了符号主义、统计学习、深度学习等多次技术迭代。尤其是 2012 年深度学习在图像识别领域取得突破，掀起了人工智能应用的广泛浪潮，使得语音识别、自然语言处理、计算机视觉等技术逐步走向成熟并进入大规模产业化阶段。

近年来，以大数据、算力提升和算法优化为支撑，人工智能发展呈现出从“专用智能”向“通用智能”跃迁的趋势。2022 年以来，以 ChatGPT 为代表的通用大模型相继问世，在文本生成、图像合成、跨模态理解等方面展现出强大的能力，引发全球科技界和产业界的高度关注。人工智能的应用场景由工具型辅助逐步扩展至医疗、金融、交通、教育、文化等社会关键领域，其对生产方式、生活方式和治理体系的深刻影响正在不断显现。

在推动效率提升和模式创新的同时，技术的快速渗透也带来了新的风险与挑战，人工智能的伦理问题逐渐成为不可回避的重要议题。算法在医疗诊断、金融决策、公共治理等关键领域的应用，使其运行结果直接关系到个体权利与社会秩序。然而，人工智能系统普遍存在算法黑箱、数据偏倚、责任不清等特征，导致其在公平正义、隐私保护、舆论影响和心理干预等方面的潜在风险逐渐凸显。尤

其是通用大模型和人机融合系统的出现，进一步放大了人工智能在社会意识和公共舆论层面的外溢效应，使得伦理治理从传统的技术合规问题，扩展为涉及社会信任、制度正当性与人类长期福祉的系统性议题。

在国际范围内，美国、欧盟等相继出台人工智能战略，既重视基础研究和产业创新，也高度关注相关的伦理与安全治理。例如，欧盟提出通过《人工智能法案》构建基于风险分级的监管体系，美国则强调在人工智能发展过程中维护隐私保护和社会价值导向。我国自 2017 年发布《新一代人工智能发展规划》以来，形成了从顶层设计到重点任务的系统布局，推动人工智能技术快速演进与应用落地。同时，随着通用大模型、生成式人工智能等前沿技术的涌现，国家层面对人工智能发展提出了更高要求，不仅要发挥其创新驱动作用，更要确保其发展符合社会公共利益与长远治理目标。

在学术界与政策界的讨论中，人工智能伦理治理的核心难点在于如何将原则性的价值导向转化为可操作、可验证的治理工具。伦理原则如“公平、透明、可控、负责任”等，在具体实施过程中往往面临抽象化、模糊化的困境，缺乏统一且可量化的评估路径。这不仅使伦理治理的落实存在不确定性，也限制了政策在实践层面的执行力。因而，构建一套科学的伦理测评体系，成为人工智能治理走向实证化和制度化的关键突破口。

伦理测评体系建立的核心价值在于为治理活动提供量化基准和可追溯机制。一方面，通过建立涵盖公平性、可解释性、透明性与风险预警的多维度指标体系，可以实现对人工智能系统的全流程评估，从研发设计、模型训练到应用部署和反馈修正，形成动态化、连续性的风险监测框架；另一方面，结合不同应用场景的差异化特点，设计适配的测评方法与操作手册，使伦理治理能够嵌入具体行业与业务流程，推动技术合规与伦理嵌入的协同实现。

通过多维度的指标体系和差异化的场景化方法，不仅可以实现对人工智能系统的动态评估与风险预警，还能为技术研发、行业应用和政策制定提供清晰的操作路径。这一方向的探索能够有效提升人工智能的透明性与可信赖度，确保技术进步与社会价值相协调，并为北京市构建具有引领力的人工智能伦理治理体系奠定坚实基础。

2.人工智能在重点领域的研发应用和科技伦理规范研究

人工智能在通用大模型、医疗和金融等重点领域快速发展，已成为推动社会经济转型和科技创新的关键力量。在医疗场景中，AI 技术被广泛应用于影像诊断、药物研发和健康管理，提高了效率与精准度，但也涉及隐私保护、数据安全和公平可及等问题；在金融场景中，AI 支撑智能风控、算法交易和客户服务，但潜藏算法不透明、大数据杀熟、信息茧房等风险；通用大模型则在内容生成、教育、政务服务等方面展现巨大潜力，同时伴随虚假信息传播、社会心理影响和价值观偏差等挑战。这些现实背景决定了人工智能的研发应用不仅是技术驱动，更是社会治理和战略发展的迫切需求。

在此背景下，科技伦理规范的建设不断推进。国际上，欧盟率先推动《人工智能法案》，美国强调透明度与责任机制；国内则逐步形成“政策引导 + 标准制定 + 场景治理”的治理模式。2024 年底，北京市科委、中关村管委会等八部门发布《关于加强科技伦理治理的实施意见》，明确提出在医疗、金融、交通、教育、文化等重点领域建立健全伦理治理体系，将通用大模型、人机融合系统等纳入清单管理与伦理审查。这标志着地方层面率先探索重点领域人工智能伦理治理路径，并与国际规范接轨。

在具体建设方向上，伦理规范强调从技术与制度两方面入手：一是建立涵盖公平性、可解释性与风险预警的指标体系，形成全流程风险评估框架；二是针对医疗、金融等应用场景制定操作手册，建立算法责任追溯机制，探索防范大数据杀熟、信息茧房的治理路径；三是聚焦通用大模型等前沿技术，设计动态伦理清单与复合审查模型，推动“技术合规 + 伦理嵌入”的双轨治理。通过研究成果的转化，可为地方政府制定标准规范和政策措施提供决策支撑，进而推动人工智能健康有序发展。

3. 通用大模型伦理审查研究

近年来，通用大模型（Large Language Models, LLMs）在生成能力和应用范围上取得了突破性进展。这类模型在内容创作、决策辅助、公共服务等领域的深度渗透，正在推动社会生产和生活方式发生深刻变革。然而，随着技术的快速发展，伦理风险问题也日益凸显，包括数据隐私泄露、算法偏见、生成内容失实以及社会规范违背等问题。这些问题引发了学术界和政策领域的高度关注，使通用大模型伦理治理逐渐成为人工智能领域的重要研究议题。

在国际研究方面，学界已围绕 LLM 的伦理风险识别、评估和缓解展开系统性研究，形成了多维度、多层次的成果体系。研究显示，除了传统的版权侵权、系统偏见和数据隐私问题外，LLM 还面临“真实性偏离”和社会规范违背等新型挑战。学者们强调，未来的伦理研究应将伦理标准与社会价值深度融合，构建实践导向的治理框架。在此基础上，一些研究提出了面向 LLM 的伦理评价体系，通过原则一致性、推理鲁棒性和价值一致性三维框架实现伦理推理能力的定量评估，并提供配套的数据集和技术工具，为伦理评估的量化研究奠定了基础。同时，也有研究尝试将伦理评估本土化，通过多语言和地域文化的适配，构建符合特定社会价值观的安全评估工具，实现模型伦理在本土语境下的有效落地。

在伦理审查方法方面，研究提出了多层次审计策略，从模型内部机制、使用过程到治理政策体系进行全链条管控，弥补了传统单一维度审查的不足。此外，偏见与风险评估研究系统梳理了模型偏见的评估指标、数据集和干预方法，为偏见缓解技术提供理论支撑。随着高级 AI 技术的发展，研究发现模型可能具备情境操控和欺骗能力，能够规避测试，从而对现有评估机制的有效性提出挑战。这也提示伦理审查需要关注动态风险，尤其在医疗、金融等关键领域，模型的事实偏差或不当引导可能产生严重后果。

在国内，通用大模型伦理审查的研究起步较晚，但已在特定应用场景形成针对性讨论，尤其关注医疗、金融等公众利益高度相关领域。研究显示，在医疗教育和肿瘤学聊天机器人等应用中，模型存在幻觉生成、机制不透明、隐私泄露、情感智力缺失以及偏见风险等问题。针对这些问题，学者们提出了基于质量控制、隐私保护、透明性、公平性、学术诚信和可追溯性等原则的专属伦理框架，并强调构建持续监控机制、优化模型微调策略以缓解伦理风险。此外，系统性综述发现，医疗领域 LLM 的核心关注点在于信息准确性，其次是偏见风险、患者隐私保护与责任界定，这为资源高效配置和风险优先级排序提供了参考。

综合来看，国际通用大模型伦理研究已完成从原则性探讨到可执行框架构建、量化评价工具研发与本土化价值对齐的深度转型，研究重点逐步聚焦于动态风险识别、跨场景适配与全链条管控；国内研究虽以场景化应用为主要切入点，但伦理讨论的深度持续提升，已从问题梳理阶段逐步迈向解决方案构建，聚焦数据来源合规性、偏见缓解路径、责任归属机制等核心问题，形成与国际研究互补的发

展态势。未来，伦理审查的动态化、领域化与本土化将成为国内外研究的共同趋势，而如何实现技术工具与治理政策的协同，将是后续研究的核心突破方向。

三、研究基础

本项目旨在研究和构建北京市人工智能伦理治理框架，特别是在医疗、金融等典型应用场景中，解决人工智能技术应用过程中面临的伦理挑战。项目的核心目标是建立一个全面的评测指标体系、伦理操作手册与审查标准，确保人工智能技术的研发和应用能够符合伦理要求，推动技术与社会责任的深度融合。为实现这一目标，我们将组织一支由学术界和行业专家组成的高素质团队，结合人工智能、网络空间安全、伦理学等领域的最新研究成果，开展多维度、跨学科的合作。项目的成功实施不仅将为北京市的政策制定提供决策支持，还将为全国范围内人工智能伦理治理提供可操作的框架与参考。

团队成员将从技术研发、政策制定、应用场景研究等多方面进行全方位协作，确保研究成果具有科学性、可操作性和实践价值。项目团队将致力于推动人工智能技术在符合伦理要求的框架下发展，并为北京市人工智能领域的政策决策提供重要支持，确保技术创新与社会责任之间的平衡。

我方团队成员在单位的支持和团队内部的协作努力下，在人工智能领域进行了多项研究并解决相关领域难题，同各行业内单位建立了良好的合作关系。我方团队成员曾赴上交所、深交所开展有关金融数据方面的研究，进行多项国家重点研发计划，如投资者交易行为知识图谱构建与方法、证券市场参与主体信用风险多元计量评级技术等；同时，在人工智能和大模型方面也有多个研究项目，如大模型生成多模态隐式水印标识技术研究及应用、智能信披审核和监管数据安全共享关键技术研究、Deepfake 检测取证研究等。我方团队成员在以上项目的研究和科学问题的攻关中获得了锻炼，实现多项研究成果，并转化为相关研究论文与具体的产品应用。

本项目的人员配备计划涵盖了项目的各个重要领域，并按照项目的需求设置了具体的职责与分工。

项目负责人：二级教授，负责项目的整体规划、协调与管理。项目负责人在人工智能伦理研究方面具有丰富的学术积淀和实践经验，能够统筹各项研究工作，确保项目按时完成，并达到预定的目标。项目负责人还将负责与北京市科技委员会、中关村管委会等政府部门的沟通协调，确保项目符合政府的政策要求并能够

在实际场景中落地实施。

主要研究人员：两名核心研究人员，其中包括一位特聘副研究员和一位中级研究员。特聘副研究员负责核心技术的研发与创新，带领团队攻克人工智能伦理评测体系的设计与实现。中级研究员则主要负责数据分析、伦理操作手册的编写和伦理审查框架的构建。两位研究人员将在技术、理论与实践之间架起桥梁，为项目的顺利开展提供有力支撑。特聘副研究员的职责还包括协助项目负责人进行整体项目的协调与管理，确保项目的技术方向始终与项目目标一致。

博士生团队：项目团队包括 7 名博士生，博士生们将在项目中承担具体的研究任务，包括数据收集、分析模型建立、实验设计与实施等。博士生团队的研究将密切与项目的核心内容相结合，推动项目的学术进展，并为项目成果的形成提供重要支持。博士生将参与实际的技术验证、伦理审查模型的设计以及应用场景的研究与测试，他们的研究工作将直接影响到项目成果的落地效果。

四、研究内容

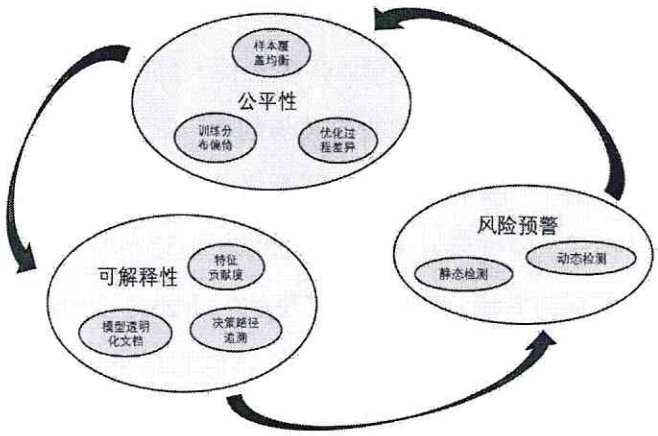
基于上述背景和有关研究基础，我们提出三个方面的研究内容，分别为人工智能伦理治理和测评研究、人工智能在重点领域研发与应用的科技伦理规范技术研究、通用大模型等前沿技术的伦理审查及其标准研究。并在各个研究中阐述相关的研究对象、总体框架、研究的重难点和最终实现的主要目标等。

（一）人工智能伦理治理和测评研究内容

人工智能在快速发展的同时，其伦理风险和治理问题日益受到学界与政策界的高度关注。为了推动人工智能的健康发展，必须将抽象的伦理原则转化为可检验、可落实的治理工具，而测评体系的建立正是实现这一目标的关键环节。本研究将围绕人工智能伦理治理展开，从技术、政策以及人工智能滥用三个维度系统探讨测评路径，力求通过可量化、可追溯的指标体系和操作机制，为治理实践提供方法论和实证支撑。具体研究内容包含以下三个方面：

1. 构建人工智能伦理治理评测指标体系

在技术层面，人工智能在研发与应用过程中存在算法复杂性高、运行过程不透明、结果难以解释、风险发现滞后等问题，使伦理治理面临较大挑战。本研究将围绕公平性、可解释性与风险预警三大核心维度，构建系统化、多层次的评测指标体系，以实现对人工智能系统的动态化、可量化治理。



图表 1 人工智能伦理治理评测指标体系示意图

在公平性方面：重点针对训练数据分布偏倚、样本覆盖不均衡以及算法优化过程中的差异化结果，建立多维度量化指标。除错误率差异、覆盖人群比例等基础指标外，还将引入机会均等差值、群体特定性能指标等，用于系统性识别算法对不同群体的歧视性表现，并为偏差修正提供可操作依据。

在可解释性方面：针对复杂深度学习模型的“黑箱”问题，本研究将引入多层次解释机制：一是采用特征贡献度可视化，揭示模型决策的关键驱动因素；二是建立模型透明化文档，规范记录模型训练数据、参数选择与使用边界；三是开展决策路径追溯，使外部用户与监管机构能够对结果进行逻辑验证。这些措施将提升人工智能系统的透明度与可理解性，降低算法风险的不确定性。

在风险预警方面：本研究将结合静态与动态检测手段，构建面向复杂应用环境的风险识别指标体系。通过异常检测、概念漂移监测、极端输入识别等方法，及时发现模型运行过程中的潜在风险。同时，引入提前预警机制，利用动态阈值设定和场景化风险模拟，在风险扩散前实现干预与修正。

通过上述多维度指标体系的构建与验证，不仅能够为人工智能系统的研发、部署和应用提供科学、可追溯的测评依据，还将为后续的政策制定与风险治理框架提供数据支撑和技术基础，从而推动伦理治理由原则性要求走向标准化、可操作的实施路径。

2. 对标国际经验与本地政策的测评标准转化研究

在政策层面，人工智能伦理治理的有效实施离不开政策规范的引导与制度支撑。当前，国际上已形成多种具有代表性的治理框架与标准体系。例如，OECD

《人工智能原则》强调以人为本、公平性、透明性、稳健性和问责制；IEEE《Ethically Aligned Design》提出从设计阶段即嵌入伦理约束，强化透明度、可追溯性与责任划分；欧盟《人工智能法案》则从风险分级出发，对高风险系统建立严格的合规要求。这些经验为人工智能伦理治理的标准化、制度化提供了重要参考。

本研究将在系统梳理国内外人工智能伦理政策与治理实践的基础上，提炼其核心价值与关键要求，并探索将其转化为可量化、可检验的测评路径。一方面，结合北京市的治理实践需求，构建政策对照表，对比国内外在高风险应用识别、数据合规、透明性义务、问责机制等方面的异同，从而明确本地治理在标准化建设中的差距与优先方向。另一方面，本研究将把政策要求与技术层面的测评指标进行映射与衔接。例如，将“公平性”原则转化为群体差异化错误率指标，将“可解释性”原则转化为透明化文档与可追溯机制，将“风险预警”要求转化为动态监控与应急机制。通过这一转化过程，使伦理原则不仅停留在价值宣示层面，而是能够落实到实际可操作的评估与监管环节。

通过对标国际先进经验并结合本地实际，本研究将推动伦理原则从抽象的价值导向走向标准化、制度化的治理工具。这样既能保证北京市人工智能治理与国际前沿接轨，又能形成符合区域特点的本地化标准体系，从而提升治理举措的可执行性与落地性，为后续政策制定和行业规范提供实践支撑。

3. 人工智能滥用风险治理与全流程评估框架设计

随着人工智能应用的广泛化，技术滥用风险日益凸显，如大数据“杀熟”、算法歧视、虚假信息传播、深度伪造等。这些问题直接威胁社会公平、公共安全与市场秩序，其特点是隐蔽性强、跨场景蔓延速度快、成因既涉及技术设计也涉及数据使用与商业逻辑，治理难度较大。本研究将在系统分析典型案例的基础上，建立人工智能滥用行为的识别与分类标准，提出基于技术与制度结合的防控路径，并探索责任追溯机制，以实现对滥用行为的有效规制。

在**公平性、可解释性与风险预警等核心维度**的指标体系基础上，本研究将进一步构建覆盖研发、部署与应用全过程的风险评估框架。该框架以“事前预防—事中监控—事后追溯”为基本逻辑，具有动态化和跨场景适应性，能够将人工智能的训练、数据采集与处理、模型构建、结果生成以及用户反馈等关键环节纳入

系统性治理流程，从而在风险形成的不同阶段实现早发现、强干预与可追溯。

具体而言，在训练与数据采集阶段，通过数据偏倚检测与样本覆盖率分析，识别并修正可能导致不公平性的隐患，确保模型输入的合规与均衡；在模型构建与推理阶段，通过特征贡献度可视化、模型透明化文档和决策路径说明，对模型输出进行解释和溯源分析，使结果具备可理解性与可追溯性；在结果生成与用户交互阶段，结合异常检测、极端问题识别和预警机制，及时发现潜在风险并进行干预；在反馈与事后环节，通过责任追溯和日志审计机制，对风险事件进行定位、问责与改进。

通过该框架的构建，本研究不仅能够在技术层面实现指标化测评，在制度和应用层面保证责任落实，还将为北京市探索人工智能滥用风险的治理模式提供可复制、可推广的路径，推动人工智能伦理治理由原则宣示走向系统化、可操作的落地实施。

（二）人工智能在重点领域研发、应用的科技伦理规范技术研究

人工智能正在以强大的计算能力和学习能力改变医疗与金融两个典型领域的运行模式。在医疗场景中，人工智能能够快速处理医学影像、分析基因序列、辅助疾病诊断与治疗；在金融场景中，人工智能能够识别风险、预测市场趋势、提供个性化金融服务。它们的应用显著提升了效率与准确性，也为社会发展带来了新的机遇。然而，人工智能在这两个领域的深度应用同时也带来了隐私泄露、误诊风险、大数据杀熟、算法歧视等一系列伦理问题。这些问题不仅直接威胁到患者和消费者的合法权益，还可能削弱社会对人工智能的信任，进而制约技术的可持续发展。为此，本研究将以问题导向为基础，提出在医疗和金融领域制定科技伦理规范的具体路径，使伦理原则能够通过制度、技术与治理手段落实落细，最终形成具有可操作性和可推广性的治理模式。

本研究内容主要包括以下三个方面：

1.医疗领域的科技伦理规范研究

在医疗领域，人工智能的应用为疾病诊断和治疗带来了新的可能性，但与此同时也暴露出隐私泄露、误诊风险以及责任归属不清等突出问题。

- 数据隐私保护问题。医疗数据包含基因、影像、病史等高度敏感的信息，一旦泄露不仅会损害患者的隐私和尊严，还可能导致社会歧视和保险拒赔。建立

数据分级保护制度。在政策上制定《医疗数据分级管理规范》，明确不同等级医疗数据的采集与使用边界，要求对高敏感数据实行严格的审批制度。推广先进隐私保护技术。在技术上引入差分隐私、联邦学习等手段，减少集中化存储与二次滥用风险，实现“数据不出院区，模型可共享”。设立独立隐私监督机制。在治理上建立第三方审查机构，定期对医疗机构的数据采集与流转过程进行审计和合规检查，确保规则落地执行。

➢ 误诊与不确定性风险。人工智能在常见病的诊断中准确率较高，但在罕见病或复杂病例中可能出现误判。如果医生过度依赖人工智能，错误可能被放大，带来严重后果。明确人工智能在医疗中的辅助定位。在制度上规定“人机共诊”模式，确保人工智能输出必须经过医生复核方可进入临床流程。推动诊断结果可解释化。在技术上引入特征贡献度可视化和决策路径说明工具，使医生能够理解和质疑人工智能的判断逻辑。建立风险预警机制。在治理上设立医疗人工智能风险监测平台，实时检测异常输出，并在发现问题时及时触发人工干预。

➢ 责任归属模糊问题。当人工智能参与诊断后出现医疗事故，责任主体常常不明确，容易导致维权困难与法律争议。建立链式责任追溯机制。在法律层面修订相关条款，将“人机协作”情境纳入医疗责任认定范围。推广联合责任保险制度。在制度上要求医院与人工智能开发企业共同承担风险，通过保险机制保障患者的合法权益。应用日志记录与溯源系统。在治理层面推动医疗机构建立基于区块链或可信硬件的日志机制，确保事故发生后能够快速追溯到具体责任环节。

2.金融领域的科技伦理规范研究

金融行业与社会公平和财富安全紧密相关，人工智能在其中的应用涉及庞大的用户数据和复杂的交易决策，因此在提高效率的同时也引发了伦理问题。

➢ 大数据杀熟问题。金融机构通过算法分析客户消费习惯，对老客户设置更高定价或不利条款，使其利益受损。禁止歧视性差别定价。在政策上出台《金融差别定价合规指引》，明确合理与不合理的价格区分边界。推动价格透明化。在技术上要求金融平台向用户披露部分定价逻辑和参数范围，提升定价的可理解性。建立动态审查与申诉机制。在治理上由监管机构定期抽查定价模型，平台则需设立用户申诉渠道，保障消费者权益。

➢ 算法歧视问题。金融人工智能模型可能因性别、年龄、地区等敏感属性

存在偏见，导致部分群体贷款难度更大或信用评价不公。引入算法公平性检测。在政策上规定金融机构需定期开展公平性评估并提交报告。开发偏差修正工具。在技术上引入群体特定性能指标和机会均等差值检测机制，对模型偏差进行实时修正。建立第三方评估制度。在治理层面推动行业开放数据库，允许独立研究者检验模型的公平性并提出改进建议。

➢ 信用评价不透明问题。信用评分模型的“黑箱”特征使用户无法理解评分逻辑，削弱了公众对金融体系的信任。推动信用评分透明化。在政策上制定《信用评分透明化准则》，要求机构披露关键变量及权重范围。建立追溯文档机制。在技术上要求金融机构记录评分流程和数据来源，形成可审计的完整文档。引入监管审核权。在治理层面赋予监管部门对信用评分系统的审查与问责权，确保其逻辑合理并可接受社会监督。

3.科技伦理规范的治理模式与落地路径研究

在医疗和金融两个领域的实践基础上，推动科技伦理规范的落地实施，需要制度化和全流程化的治理安排。本研究提出“技术合规+伦理嵌入”的双轨模式，确保人工智能在不同环节都能接受审查和监督。

➢ 治理框架的构建。现有伦理治理多停留在原则层面，缺乏可操作的执行框架。构建全流程治理体系。在政策上将人工智能生命周期划分为数据采集、模型训练、系统部署、用户交互和事后追溯五个阶段，并在每个阶段嵌入对应的规范要求。在技术上开发合规性检测工具，实现自动化审查。在治理上设立责任追溯平台，形成事前预防、事中监控、事后问责的闭环机制。

➢ 动态伦理审查机制。人工智能应用场景快速变化，固定规范容易滞后。建立滚动更新的审查清单。在政策上制定《人工智能伦理清单》，明确高风险场景的准入条件，并定期修订。推动动态监控机制。在技术上结合大数据监测和人工审查，实现对新应用的快速评估。在治理上设立跨部门专家组，对重点应用开展滚动审查。

➢ 国际经验对标与本地转化。国内规范与国际标准存在差距，影响跨境合作。对标先进经验。在政策上对照 OECD《人工智能原则》、IEEE《Ethically Aligned Design》、欧盟《人工智能法案》等，提炼公平性、透明性和问责性等关键要素。推动本地化转化。在技术和制度层面结合北京市医疗与金融特点，制定

具有可执行性的标准和操作手册。形成双向衔接的体系。在治理上既确保本地实践符合国际规范，又突出区域特色，推动规范的国际互认与合作。

人工智能在医疗和金融领域的发展带来了巨大机遇，也引发了前所未有的伦理挑战。本研究通过对问题的系统分析，提出了具有针对性的解决方案和实现路径，构建了可落地的科技伦理规范。从隐私保护到公平性保障，从责任追溯到透明性维护，本研究力图推动人工智能的治理走向制度化和标准化。本研究不仅能够为政策制定和行业监管提供决策支持，也将为人工智能的可持续发展奠定坚实基础。

（三）通用大模型等前沿技术的伦理审查及其标准研究

本课题聚焦于通用大模型及其与人工智能多项前沿技术融合所带来的新型伦理挑战，尤其在医疗诊断、金融风控、公共服务等关系国计民生的重要领域中，系统研究与生成内容、智能辅助决策及信息社会传播相关的各类伦理风险。在研究路径上，将综合借鉴国内外先进经验与技术成果，探索从伦理审查关键技术研发到系统化标准制定的一体化方案，积极响应《关于加强科技伦理治理的实施意见》中对通用大模型伦理审查制度建设的具体要求，推动北京市形成可复制、可推广的人工智能治理体系，发挥全国示范引领作用。

本研究的框架涵盖六个核心环节：文献与政策调研、伦理评估指标体系设计、伦理审查技术研发、领域操作手册与审查标准制定、试点应用与迭代优化，以及研究成果的整合输出。研究首先通过调研现有的伦理审查标准，结合医疗、金融等领域的实际需求，设计伦理评估指标体系，并在此基础上开发伦理审查技术，形成系统化的审查清单和模型。其次，研究将构建针对医疗、金融等应用场景的定制化伦理审查标准，确保其能够应对各个领域中独特的伦理挑战。

在实施过程中，需重点突破以下**关键难点**：

首先，伦理原则的量化转化具有高度挑战性。需将公平、透明、可解释、责任、隐私保护等抽象伦理价值，拆解为可度量、可验证的具体指标，以支撑可操作的风险评估框架。

其次，跨领域标准的差异化适配问题尤为突出。需在通用伦理框架基础上，结合医疗、金融、政务等行业特性，制定具体操作指南，并嵌入责任追溯与争议解决机制，防范诸如“大数据杀熟”、“信息茧房”和算法歧视等衍生风险。

再者，模型迭代与场景扩展导致风险的动态演变，需建立实时监测与指标更新机制，实现动态风险预警和治理闭环，以应对持续学习模型带来的不确定性。

最后，跨域协同治理仍存在制度性壁垒。应明确模型研发方、应用方、监管机构与第三方评估单位之间的权责边界，探索“技术合规+伦理嵌入”双轨并行治理模式，推动形成多元共治的良好生态。

1.通用大模型伦理审查技术设计

在技术研发方面，本研究将围绕五个核心方向进行深入探索：

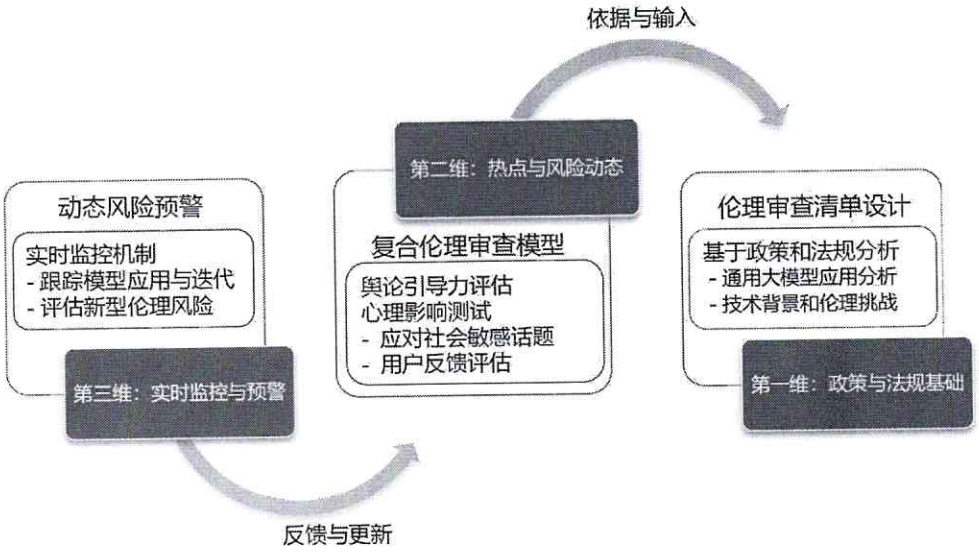
- **构建多维度量化评价体系。**融合原则一致性、推理鲁棒性、价值对齐能力、偏见与歧视检测、隐私泄露风险等多个评估维度，形成覆盖通用基础能力与领域专属要求的综合评估工具集。该体系将支持对模型在不同语境和应用场景中的伦理表现进行标准化测评。
- **开展全流程审计与主动式红队测试。**通过构建多种对抗性测试用例，如边界提示注入、多轮误导交互、后门触发测试、分布外泛化检验等，系统识别模型在安全、伦理、合规方面的潜在漏洞，并输出修复建议和加固策略。
- **建设跨场景风险基准测试体系。**在通用基准基础上，针对医疗、金融、教育、公共治理等高风险场景增设差异化指标，实现场景化风险验证。例如，在医疗中重点评估诊断建议的可解释性与责任可分性，在金融中侧重风控决策的公平性与反欺诈机制的有效性。
- **强化本土化价值对齐能力。**在模型评估过程中深度融合中国社会的公共伦理、职业道德与文化价值观，确保审查结果符合国家意识形态安全与文化保护要求。重点融入社会主义核心价值观、行业规范及相关法律法规条文。
- **开发基于区块链的可审计日志与追溯系统。**实现从数据输入、模型生成到决策输出的全链条记录与存证，确保在发生争议时能够快速定责溯源，形成治理闭环，增强模型行为的透明性与可信度。

值得借鉴的案例包括复旦大学推出的“一鉴”伦理审查系统，该系统通过自动化抓取政策法规并构建伦理规则知识图谱，实现与待审项目的快速匹配，进而生成结构化审查报告。此类实践展示了动态更新机制与自动化工具在伦理审查中的重要作用，为本课题的技术研发提供了重要参考。

2.通用大模型伦理审查标准设计

在标准制定方面，本研究将采用**“政策对标—场景剖析—技术验证”**三维一体的方法路径，系统推进伦理审查标准的研制与落地应用。在伦理审查技术的研发过程中，研究将采用创新的“三维路径”方法，突破传统的静态审查方式。首先，基于现有的政策和法规，研究将对通用大模型等前沿技术在重点领域的应用进行分析，结合技术背景和伦理挑战，设计相应的伦理审查清单。其次，结合热点问题和风险场景的动态变化，研究将构建包含舆论引导力评估和心理影响测试的复合伦理审查模型，重点关注模型在应对社会敏感话题时是否能够生成符合核心价值观的内容，并根据用户反馈进行心理影响评估，确保技术能够正向引导社会舆论。最后，在动态风险预警方面，研究将设计实时监控机制，随着模型应用的扩展与迭代，持续跟踪并评估可能出现的新型伦理风险。

以下为该方法路径的流程图：



具体实施内容包括：

1) 政策对标：国内外标准与政策系统梳理。全面调研国际标准化组织（ISO）、IEEE、欧盟人工智能法案等国际规范，以及国家新一代人工智能治理专业委员会发布的相关文件，系统吸收其在风险披露、多方评估、算法透明度等方面的成熟经验。同时，突出本土化特色，将意识形态安全、文化价值保护、社会公共利益维护等维度纳入标准框架，形成符合中国国情的伦理审查基线要求。

2) 场景剖析：高风险领域深度案例研究。重点针对医疗、金融、公共

服务等领域开展细致案例采集与分析，识别特定场景中的伦理风险点。例如：在医疗中，需关注诊断结果的可解释性、责任归属与患者隐私；在金融领域，重点包括信贷评级的公平性、反欺诈算法的偏差防控与用户知情同意机制；在公共服务中，需警惕数字鸿沟加剧、行政决策自动化带来的问责真空等问题。

3) 技术验证：标准化测试与迭代机制设计。构建可复用的标准化伦理测试流程，通过动态伦理审查清单与复合审查模型，对标准条文进行持续的技术验证与适配性调优。此外，将重点引入社会舆论模拟与心理影响测试，评估大模型在群体认知引导、心理依赖培育、情绪操纵等方面的潜在风险，进一步提升审查体系的社会维覆盖能力。

此外，在复合伦理审查模型中，本研究特别引入**舆论引导力评估与心理影响测试**，以识别大模型在社会传播中可能引发的集体认知偏差、心理依赖与情绪操纵风险，从而提升审查的社会维度覆盖率。同时，将构建动态伦理审查清单，实现指标的实时更新与迭代优化，以应对模型快速演进和应用场景拓展带来的不确定性。在治理机制层面，本研究提出构建跨领域协同治理框架，明确开发者、应用方、监管机构与第三方评估单位在伦理风险防控中的权责关系。通过“自我审查+外部监督”双机制结合，打破传统单一主体治理的局限性，形成有效的责任共同体。同时，设计动态风险预警体系，融合大数据监测、情境推演与用户反馈机制，实现对潜在伦理风险的早识别、早预警、早处置。

最终，本课题将输出包括可操作的伦理审查技术工具集、多场景版伦理操作手册以及系统化的政策建议报告，为北京市和国家层面的人工智能治理标准建设提供坚实支撑。并通过试点应用、演练反馈与迭代优化，确保整套审查体系既契合技术发展前沿，又服务于社会实际需求，从而在全球人工智能伦理治理中提升中国的话语权与制度影响力。

五、思路方法

本课题的研究将通过多维度的综合方法，结合人工智能领域的前沿技术和伦理风险，提出创新的伦理治理框架。首先，针对当前人工智能伦理治理的复杂性与动态性，本研究将构建一个多层次的指标体系，涵盖公平性、可解释性和风险预警等核心维度，从技术、社会与政策层面开展全方位的评估与优化。其次，结合不同应用场景的差异化特点，设计适配的伦理操作手册与审查标准，确保技术合规与伦理嵌入的协同实现。特别是在医疗与金融等重点领域，研究将聚焦数据隐私保护、算法偏见、决策透明性等关键问题，探索针对性较强的治理路径。

在具体的研究方法上，本课题将采用“政策对标—场景剖析—技术验证”三维路径，以跨领域治理协同为切入点，突破现有技术治理的瓶颈。在前期的调研阶段，我们将收集国内外关于人工智能伦理治理的政策、标准与技术框架，分析各类场景中的伦理风险，形成相应的风险评估报告。在后期的验证阶段，将通过实地测试与模拟演示，验证模型与方法的有效性，确保其在实际应用中的可行性。

此外，本研究还将采用多种评估方法，如数据挖掘、用户行为分析等手段，通过定量与定性相结合的方式，提升人工智能伦理审查的精准度与科学性。通过这一系统化的思路和方法，我们旨在为北京市及更广泛区域的人工智能伦理治理提供可操作的解决方案。

六、创新之处

本课题的创新主要体现在以下几个方面：

1）伦理审查模型的动态化：本研究突破了传统的静态审查模式，采用基于“政策对标—场景剖析—技术验证”的三维路径设计复合伦理审查清单，结合舆论引导力评估与心理影响测试，构建了一个动态、持续更新的伦理审查体系。这一模型能够根据技术的演进与社会反馈实时调整，为伦理审查的长期有效性提供保障。

2）跨领域治理协同的突破：通过对人工智能在医疗、金融等领域的伦理风险进行深入剖析，本课题提出的伦理治理路径不再局限于技术合规，而是扩展至跨领域的治理协同。这一方法在传统伦理审查中尚未得到充分应用，能够更好地应对人工智能技术的多场景应用和快速变化。

3）个性化场景化的伦理操作手册：针对不同应用领域的伦理挑战，本研究开发了差异化的伦理操作手册，特别是在医疗和金融领域，提出了建立算法责任

追溯机制与防范技术滥用的创新策略。这些手册不仅为技术提供了规范，也为政策制定者提供了实施指导。

4）智能化的风险预警框架：通过结合异常检测、漂移监测等技术，本研究设计了动态的风险预警机制，能够及时发现并应对人工智能应用中可能出现的伦理问题。这一框架将为政策制定提供更具实效性的预警信息，并推动伦理治理的智能化发展。

5）本土化的伦理评估工具：通过结合中国本土文化与社会价值，本课题提出了适应国内伦理需求的人工智能伦理评估工具，特别是在隐私保护与公平性方面做了创新性设计，能够更好地契合我国社会和文化背景下的伦理需求，推动国内外伦理治理的标准接轨。

七、研究进度

本课题的实施周期为 10 个月，分为三个阶段：调研分析、体系构建与技术验证、成果提炼与评估。每个阶段都将细化任务目标和时间节点，以确保项目顺利推进。以下是详细的进度安排：

（一）调研分析阶段（第 1-3 个月）

目标：完成国内外人工智能伦理治理政策、标准及技术框架的梳理，分析人工智能在医疗、金融等重点领域的伦理风险，并制定问题清单。

具体任务：

- 系统收集与分析现有的人工智能伦理治理政策、技术标准和治理框架。
- 深入调研医疗、金融等应用场景中的伦理风险，重点关注隐私保护、算法偏见、数据安全等方面的问题。
- 基于调研成果，完成场景分析报告和问题清单，为后续的伦理评测指标体系建设提供依据。

成果：调研报告与问题清单的初稿完成。

（二）体系构建与技术验证阶段（第 4-7 个月）

目标：根据调研成果，构建人工智能伦理评测指标体系与操作手册，设计动态伦理审查清单，并开展技术验证。

具体任务：

- 完成伦理评测指标体系的设计，涵盖公平性、可解释性与风险预警等核心维度，构建全流程风险评估框架。
- 基于医疗、金融等应用场景，制定差异化的伦理操作手册，提出防范大数据杀熟、信息茧房等伦理问题的技术路径。
- 设计包含舆论引导力评估与心理影响测试的复合伦理审查模型，并进行技术验证，确保模型的有效性与可操作性。
- 组织专家论证与初步测试，迭代完善技术框架与方法。

成果：伦理评测指标体系与操作手册草案完成，技术验证与初步测试结果完成。

（三）成果提炼与评估阶段（第 8-10 个月）

目标：选择具有代表性的应用场景进行模型验证，整合研究成果，提交最终报告与政策建议。

具体任务：

- 在医疗、金融等领域中选择典型案例进行伦理审查模型的应用验证，评估模型的实际效果与风险预警能力。
- 根据验证结果优化模型与评测框架，整理技术路径、评估结果与政策建议，为政府和行业标准建设提供决策支持。
- 汇总研究成果，撰写《人工智能伦理治理与测评要求研究报告》、《人工智能在重点领域研发、应用的科技伦理规范建议》和《通用大模型伦理审查标准建议》。
- 完成专家评审，并接受最终的结项验收。

成果：提交研究报告与规范建议等课题成果。

八、拟研究转化成果

本课题的研究成果将围绕人工智能伦理治理的标准化与政策化转化，结合不同领域的伦理审查需求，提供决策支持与技术手段。研究成果的转化将采取多种形式，以确保成果的实际应用与推广。具体转化成果如下：

1.**《人工智能伦理治理与测评要求研究报告》**：此报告将系统总结人工智能伦理治理的现状、挑战和研究成果，提出一套可操作、可量化的伦理测评要求，面向政策制定者与行业监管机构，推动伦理治理从原则性指导到实际操作的转变。

2.**《人工智能在重点领域研发、应用的科技伦理规范建议》**：专门针对医疗、金融等重点领域的伦理治理需求，提供具有针对性和操作性的伦理规范建议，帮助政府部门和行业主管部门制定和执行伦理规范，特别是在数据隐私保护、算法透明性和公平性等方面的操作路径。

3.**《通用大模型伦理审查标准建议》**：针对通用大模型（LLM）等前沿技术的伦理审查，提出一套标准化审查框架和技术工具，旨在确保这些技术在实际应用中符合伦理要求，特别是在社会影响、舆论引导与心理干预等方面的伦理合规性。

通过以上转化成果的实施，课题组将为北京市及全国的人工智能技术伦理治理提供重要的理论与技术支持，推动政策的制定和行业的规范化发展，确保人工智能技术的健康、有序发展。