CAICT 中国信通院

蓝皮报告

人工智能伦理治理研究报告

（2023 年）

中国信息通信研究院知识产权与创新发展中心

中国信息通信研究院科技伦理研究中心

2023年12月

版权声明

前　言

伴随人工智能的迅速发展和广泛应用，人类正在进入一个“人机物”相融合的万物智能互联时代，人工智能技术的应用在给人们带来生活和工作便利的同时，也带来了系列伦理挑战。2022 年 3 月，中共中央办公厅国务院办公厅印发《关于加强科技伦理治理的意见》，对科技伦理治理工作进行了系统部署，将人工智能列入科技伦理治理的重点领域。

人工智能伦理是开展人工智能研究、设计、开发、服务和使用等活动需要遵循的价值理念和行为规范。人工智能技术的突破发展，引发了技术应用的伦理争议，特别是生成式人工智能技术的发展应用引发了偏见歧视、隐私侵犯、责任不明、虚假内容传播等伦理挑战。

为应对人工智能技术应用带来的风险，世界各国积极推动人工智能伦理国际治理合作。各国政府通过出台人工智能伦理原则、发布人工智能伦理治理指引、提供技术治理工具等加强本国本地区的人工智能伦理治理监管。我国通过积极完善人工智能伦理制度规范，探索人工智能伦理治理技术化、工程化、标准化落地措施，加强人工智能治理国际合作等举措推动人工智能向善发展。

人工智能伦理治理是多主体协作的全流程治理，是以敏捷机制协调人工智能发展与安全的重要治理模式。未来一段时期，人工智能伦理治理将与产业创新活动增强协调；多学科多主体参与、分类分级治理、技术工具开发等措施将有效推动人工智能伦理治理机制完善；全民科技伦理素养的提升将有效防范人工智能伦理风险；全球人工智能

伦理治理合作也将推动人工智能技术造福人类。

中国信息通信研究院首次发布《人工智能伦理治理研究报告》蓝皮书。本报告在总结分析人工智能伦理治理相关特点的基础上，对人工智能生成内容、自动驾驶、智慧医疗三个典型应用场景的伦理风险进行分析，并结合国内外人工智能伦理治理实践，提出人工智能伦理治理的四点展望，以期为更加广泛深入的讨论提供参考。

目 录

一、人工智能伦理治理概述

（一）人工智能伦理的概念与特点

“伦理”是人的行为准则，是人与人之间和人与社会的义务，也是每个人源于道德的社会责任[1]。伦理作为价值规范，为不同场景的行为提供引导。在科技活动中，伦理从价值引导和实践规范层面指导技术研发应用。

人工智能伦理是开展人工智能研究、设计、开发、服务和使用等科技活动需要遵循的价值理念和行为规范。人工智能伦理关注技术的“真”与“善”，并为人工智能发展提供更广阔的讨论空间。人工智能伦理包含价值目标与行为要求两个方面。在价值目标上，人工智能伦理要求人工智能各阶段活动以增进人类福祉、尊重生命权利、坚持公平公正、尊重隐私等为目标。在行为要求上，人工智能伦理要求人工智能技术做到安全可控、透明可解释，在人工智能研发应用各环节强化人类责任担当，提倡鼓励多方参与和合作。

人工智能伦理呈现出哲学性、技术性、全球性三大特点。**一是人工智能伦理拓展了人类道德哲学反思的边界。**人工智能伦理蕴含了人与机器相互关系的伦理思考，拓展人类关于善、理性、情感等问题的探索。人工智能伦理的讨论既包含了对人工智能主体、人格、情感方面的本体伦理问题研究，也关注人工智能应用是否符合社会道德要求。关于人工智能伦理的讨论体现着当代人对社会生活的价值理想，将人与人交往的伦理规范扩展至人与技术交互的反思。**二是人工智能伦理**

[1] 辞海编辑委员会.辞海[M].上海：上海辞书出版社，1979:221.

与人工智能技术的发展应用密切相关。从 1940 年人工智能第一次浪潮中阿西莫夫提出“机器人三原则”，到 2004 年人工智能第三次浪潮中机器人伦理学研讨会正式提出“机器人伦理学”[2]，再到目前，人工智能伦理已成为政府、产业界、学界等共同关注的议题。伴随深度学习算法的演进以及人工智能技术应用领域的拓展，人工智能伦理关注算法技术风险和技术应用危机的防范。在“强人工智能”时代到来前，人工智能伦理主要关注歧视偏见、算法“黑箱”、技术滥用、数据不当收集等问题。但随着大模型技术的发展，人工智能伦理讨论的议题也不断深化。**三是增进人类福祉是人工智能伦理的全球共识。**不同于传统伦理观念受地区历史传统文化的影响产生的差异，人工智能技术的发展和应用带来了全球性伦理挑战。目前，社会偏见、技术鸿沟、多样性危机等成为国际社会面临的共同挑战。以人为本、智能向善、促进可持续发展等已成为全球人工智能伦理共识。

（二）人工智能伦理治理的必要性

面对人工智能带来的风险，通过人工智能治理的多种机制促进人工智能健康发展已成为普遍共识。**人工智能伦理治理是人工智能治理的重要组成部分，主要包括以人为本、公平非歧视、透明可解释、人类可控制、责任可追溯、可持续发展等内容。**人工智能伦理治理能根据人工智能技术发展和应用情况，及时提出调整人与人工智能关系和应对人工智能风险的方法。人工智能伦理治理重点不在于关注对创新主体的最低义务要求，而在于推动“智能向善”的价值目标的实现。

[2] 杜严勇.人工智能伦理风险防范研究中的若干基础性问题探析[J].云南社会科学,2022(03):12-19.

1.通过伦理治理，加深关于“人工智能体”的哲学探讨

人工智能技术的发展和应用带来了新的主体“人工智能体”，即能在一定条件下模拟人类进行自主决策，与人和环境开展交互的技术体。大语言模型的发展，使得关注和应对“人工智能体”带来的伦理问题变得更加迫切。**一是人类自主性受到挑战。**工业时代，机械化的发展降低了人类体力型、重复型劳动的比例；智能时代，机器开始替代人类进行决策分析，从自动化向自主决策发展。人工智能从人类操控为主的“人在决策圈内”转向“人在决策圈外”[3]的智能体自主决策。**二是人类自我认知受到冲击。**人类因所具有的学习能力、创造力、个体的多样性、饱含情感等与机器存在不同。当前，大模型技术已具有自然语言交互和开展专业任务的能力，并能根据反馈进一步学习提升，与人的差异进一步缩小，对人类自身价值的认知面临冲击。**三是人机关系进一步复杂化。**与工业时代可视、可理解、可控制的工具不同，人工智能体具有自主学习能力，但可解释性不足，也可能产生不可控的人机伦理风险。同时，具身智能机器人、自动驾驶等的发展，推动人机交互实体化。随着人工智能技术向通用人工智能发展，有关人工智能体意识、人工智能体是否成为社会主体、人工智能体的法律地位等问题受到更多关注。有关人工智能与人类关系的问题引起包括技术专家、哲学家、科技企业家、科幻小说作者等的关注。如科幻小说作家艾萨克·阿西莫夫提出“机器人三定律”，提出人与机器交互的基本底线；人机协同工作模式的探索等。

[3] 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,2017,31(06):98-108.

2.通过伦理治理，应对人工智能应用风险

人工智能被视作促进经济发展、产业升级的重要推动力。然而，人工智能技术本身是否无害、人工智能技术应用的潜在危害仍然需要广泛探讨，伦理治理的包容性、跨学科性为风险的评估和应对提供空间。**一是关于技术本身是否承载观念意志仍存在争议。**同意技术中性（neutrality）的观点认为，技术是纯粹的科学应用，是自然规律与科学原理的反映。反对技术中性的观点认为，技术由人创造，具有社会属性和价值观念[4]。**二是技术应用的“不中立”“非中性”不存在争议，技术的应用根据场景不同，其争议度有明显区别。**一方面，人工智能技术应用被视作推动经济发展的重要力量，为工业、农业、服务业等升级转型提供了技术支撑。另一方面，人工智能伦理风险与其他技术风险叠加，且不同场景下的人工智能技术应用的伦理风险有明显差异。如在农业智慧化的场景下，人工智能技术与物联网等技术相结合，可对土壤、光照、病虫害等进行监测分析，实现对农业作业的指导、管理、优化等，对提高农作物产量和质量起到重要作用，总体风险较小。在互联网信息推送场景下，企业主体将数据与算法相结合，对用户偏好进行分析，进行个性化推荐，实现用户体验的提升与广告分发效率的提高；然而，个性化推荐可能造成的隐私、监控、信息茧房等伦理问题受到关注。

3.通过伦理治理，实现敏捷有效的风险规制

人工智能伦理治理具有灵活敏捷、包容开放的特点，与人工智能

[4] 参见吴致远.有关技术中性论的三个问题[J].自然辩证法通讯,2013,35(06):116-121+128.

全流程治理理念契合，成为人工智能治理的关键模式。**从机制的灵活性上看**，人工智能伦理治理是一种高适应性规范，与科技发展和伦理事件动态互动，能够进行快速调整。**从参与主体上看**，人工智能伦理治理是政府、产业、学界等多主体合作应对风险的机制，是以多主体参与为基本实践的治理模式。**从治理工具上看**，人工智能伦理治理包括了原则指导、规范指南、技术工具等丰富机制，能与人工智能研发和应用不同阶段的需要相配合。**从治理介入阶段看**，人工智能伦理治理融入研发到应用的全生命周期中，既能有效发挥向善价值的前置引导作用，也可借助风险评估和风险反馈机制及时调整技术发展与应用，为新技术匹配适宜的治理方式。

二、人工智能伦理治理关切

（一）人工智能伦理挑战

目前，人工智能引发的伦理挑战已从理论研讨变为现实风险。根据 OECD AI Incidents Monitor 的统计，仅 2023 年 11 月 1 个月，人工智能事件超过 280 件[5]。对人工智能伦理问题的关切覆盖人工智能全生命周期的各个阶段，需要从技术研发和应用部署[6]层面分析评估人工智能伦理风险。

在技术研发阶段，由于人工智能技术开发主体在数据获取和使用、算法设计、模型调优等方面还存在技术能力和管理方式的不足，可能产生偏见歧视、隐私泄露、错误信息、不可解释等伦理风险。偏见歧

[5] 数据来源： OECD AI Incidents Monitor，最后访问日期：2023 年 12 月 11 日.

[6] 国际标准化组织《人工智能系统生命周期过程》（ISO/IEC WD5338）将人工智能系统全生命周期概括为初始、设计研发、检验验证、部署、运行监控、持续验证、重新评估、废弃八个阶段，本部分将上述环节概括为工程技术研发和应用开发部署两大环节.

视风险是由于训练人工智能的数据集存在偏见内容、缺乏多样性等数据集质量问题，以及算法对不同群体进行了非公平性设计等，产生歧视性算法决策或内容输出。隐私泄露风险是指使用包含未经同意的个人数据进行模型训练，继而引发模型输出内容可能产生侵犯隐私的风险。错误信息风险主要发生在人工智能基础模型中，由于大模型是根据前序文本对下一个词进行自回归预测生成，预测内容受前序文本影响较大，大模型可能产生“幻觉”（Hallucination），继而生成错误不可靠的内容。不可解释风险是指由于人工智能算法的复杂性和“黑箱性”，导致人工智能决策原因和过程的无法解释。**在产品研发与应用阶段**，人工智能产品所面向的具体领域、人工智能系统的部署应用范围等将影响人工智能伦理风险程度，并可能产生误用滥用、过度依赖、冲击教育与就业等伦理风险。误用滥用风险是指由于人工智能技术使用便利度提高、任务完成能力增强，人工智能容易被用于不当任务的风险，具体包括快速生成大量虚假内容、生成恶意代码、被诱导输出不良信息等。过度依赖风险是指由于人工智能技术能力的提升，使用者对人工智能产生过度的依赖和信任，包括在未进行事实核查情况下对大模型生成内容的采信，甚至因长时间交互产生情感依赖等。冲击教育与就业风险是指人工智能便捷性的提升，使得学生可以借助机器完成作业论文，影响教育学习的基本方法，而青少年与人工智能广泛的直接互动也可能带来心理健康风险。同时，人工智能的就业冲击已不止于数字化替代，还可能冲击从事艺术、咨询、教育等领域的专业

人员，加速教育投入的折旧[7]，引发进一步的就业替代冲击。

需要注意，包括隐私泄露、偏见歧视、产生虚假错误信息、归责不明等伦理风险的发生原因既可能产生在研发阶段，也可能产生于应用阶段，甚至是两者叠加产生的负面后果。识别人工智能伦理风险需要进行全生命周期的评估。

（二）典型应用场景的人工智能伦理风险

根据人工智能具体应用场景的不同，主要伦理风险、风险影响对象与范围、伦理治理的客体存在较大差异。目前，包括图文生成、自动驾驶、智慧医疗等应用领域面临的典型伦理风险有一定差异，需要分场景分析与讨论。

1.人工智能生成内容

伴随大模型的发展，文本生成、图片生成、代码生成等生成式人工智能技术应用快速发展，包括 ChatGPT、Claude、Stable Diffusion、Midjourney 等生成式人工智能应用成为 2023 年人工智能应用热点。但使用大模型生成内容具有三大突出伦理风险。**一是误用滥用风险。**生成式人工智能技术应用普及快、使用门槛低，可能成为制作深度伪造内容、恶意代码等的技术工具，引发虚假信息大量传播以及网络安全问题。**二是数据泄露与隐私侵犯风险。**生成式人工智能使用的训练数据集可能包含个人信息，继而被诱导输出有关信息。同时，在用户在使用过程中上传的个人信息、企业商业秘密、重要代码等都有可能成为生成式人工智能训练的素材，进而产生被泄露的风险。**三是对知**

[7] 段伟文. 准确研判生成式人工智能的社会伦理风险 [J]. 中国党政干部论坛, 2023, (04): 76-77.

识产权制度带来挑战。生成式人工智能技术对知识产权体系造成了冲击。在训练数据的使用上，哪些数据能用于模型训练还存在争议，关于“合理使用”是否能适用于大模型数据训练还在讨论，且已有艺术家开始使用技术工具[8]阻止未经允许的模型训练。在生成内容的权利归属上，人工智能技术是否仅能发挥工具作用还有探讨空间。

误用滥用案例：生成假视频、图片、声音以实施诈骗

2023 年 2 月 28 日至 3 月 2 日的三天内，加拿大至少有 8 名老人因为利用深度伪造内容实施的诈骗损失大量资金。犯罪分子利用人工智能技术，快速克隆声音、图片等，并结合详细的个人信息，使得受害者产生错误判断，按犯罪分子要求支付费用。

隐私侵犯案例：泄露用户信息

2023 年 11 月 28 日，来自谷歌、华盛顿大学等研究团队发现 ChatGPT 数据泄露漏洞。文章指出，让 ChatGPT 多次重复一个词后，模型可能会输出个人信息。2023 年 3 月 25 日，OpenAI 发文证实部分 ChatGPT Plus 服务订阅用户的姓名、电子邮件地址、支付地址、信用卡的后四位和信用卡到期时间等被泄露。

知识产权争议案例：是否属于“合理使用”的争议

2023 年 9 月，美国作家协会对 OpenAI 发起集体诉讼，起诉 OpenAI 在未经许可的情况将受版权保护的作品用于大语言模型训练。2023 年 7 月，数千名作家签署公开信，要求 OpenAI 等人工智能公司停止在未经许可的情况下使用其作品训练大模型。

[8] 如 Nightshade、Glaze 等通过改变人眼无法识别的像素影响机器学习模型的工具.

2.自动驾驶

自动驾驶是人工智能、物联网、高性能计算等新一代信息技术深度融合的产物，是人机交互领域的重要实践。自动驾驶在载人、载货等方面的技术持续发展，并从封闭场景、封闭道路运行向复杂社会化场景进行拓展。自动驾驶伦理与风险控制、生命价值衡量、责任分配等问题关系密切。**一是自动驾驶技术复杂性使得风险控制更具挑战。**自动驾驶汽车的风险来自以自动驾驶汽车为载体的各类软硬件技术，包括自动驾驶算法偏差、软件安全漏洞、硬件架构不可靠等，并具有风险叠加的危机。**二是“电车难题”[9]“隧道难题”[10]成为现实的伦理风险。**目前，有条件自动驾驶已上路测试，与软件信息服务、封闭场景下的运行的机器人等风险不同，自动驾驶技术作为面向不确定开放环境，且能够造成巨大物理损害的技术应用，其风险影响范围不断增大、在紧急情况下，自动驾驶算法将可能直接对人的生命利益作出选择。**三是责任分配在自动驾驶场景下更具复杂性。**在传统情况下，损害责任可根据因果关系、过错程度并结合具体场景进行认定。但在自动驾驶场景下，自动驾驶算法具有一定的自主性，同时，涵盖自动驾驶汽车的关键责任方包括制造商、汽车设计主体，软件服务主体、使用者等，承担责任的主体繁多且难以确认，因果链条更加复杂，社会风险责任的分配具有显著的复杂性。

[9] 电车难题（Trolley Problem）由 Philippa Ruth Foot 提出，讨论是否能因为追求多数人的利益而牺牲少数人的利益。参见 Philippa Foot,The Problem of Abortion and the Doctrine of the Double Effect, 1967. Oxford Review, No. 5.

[10] 隧道难题（Tunnel Problem）由 Jason Millar 提出，讨论牺牲汽车乘客还是牺牲行人的两种判断。参见 Jason Millar (2016) An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars, Applied Artificial Intelligence,30:8, 787-809.

责任归属案例：驾驶员和汽车厂商的责任划分

2018 年 3 月美国亚利桑那州，一辆正在测试中的 Uber 自动驾驶汽车以 69 公里时速撞死了一位横穿马路的行人。2020 年 9 月 15 日，亚利桑那州陪审团以过失杀人罪起诉当时 Uber 自动驾驶汽车前安全驾驶员 Rafaela Vasquez，建议判处该安全员 2.5 年有期徒刑。检察官认为：“当驾驶员操作汽车方向盘时，他们有责任以遵守法律的方式安全地控制和操作汽车”。

2019 年，一特斯拉 Model 3 撞树起火，并造成驾驶司机死亡和乘客受伤。事后，两名受伤乘客认为特斯拉在出售汽车时知道其自动驾驶系统 Autopilot 有缺陷并提起诉讼。2023 年 9 月，美国加州地方法院陪审团认为车辆没有制造缺

陷，特斯拉方面称该案件是由驾驶员人为因素导致。

自动驾驶汽车算法设计问题：紧急情况下的处置不当

2023 年 10 月，一行人被人类驾驶的轿车撞倒并滚落至车道上。此时，由 Cruise 运营的自动驾驶出租车 Robotaxi 迎面驶来，该行人再次遭到撞击并被卷入车下。该 Robotaxi 在紧急制动后又启动了靠边停车操作，以 20 英里每小时的速度将这名行人拖行了 7 英尺，导致其全身多处严重损伤。

3.智慧医疗

人工智能在医疗卫生领域的应用带来显著的机遇，为病灶诊断、药物研发、疾病风险预测等提供了重要的工具。根据斯坦福大学最新

发布的《人工智能指数报告》，2022 年与医疗健康领域相关的人工智能投资超过 60 亿美元，是吸引最多投资的人工智能应用领域。然而，由于在医疗卫生领域应用人工智能技术直接影响生命健康，需要谨慎评估人工智能在该领域应用的伦理风险，特别要关注人类监督、隐私保护和医患知情权。**一是缺乏人类监督引发安全控制风险。**由于人工智能可用于医学图像、诊断特征匹配等医学场景，能够影响医疗决策。在直接面向患者的智慧医疗领域，缺乏专业人员介入和监督的医疗诊断风险极大。**二是隐私泄露与数据保护风险。**医疗活动产生和收集大量个人信息，医生也需要依靠患者既往病史等做出判断。随着电子病历、基因检测等应用，以及生成式人工智能在医疗领域应用开发，敏感个人信息[11]被大量收集和处理，一旦发生泄露、非法使用、篡改等，将可能对患者隐私、患者诊疗等产生严重危害。**三是缺乏透明风险。**目前谷歌、亚马逊、微软等公司已与医疗保健组织开展合作，开发应用于医疗保健领域的大模型，但若急于将缺乏完整理解和透明性的大模型用于医疗，将可能会影响专业人员的诊断，对患者造成损害。

确保人类监督和决策案例：禁止人工智能替代医师和生成处方

2023 年 8 月，北京市卫生健康委发布《北京市互联网诊疗监管实施办法(试行)》（征求意见稿），明确人工智能软件不得代替医师提供诊疗，禁止使用人工智能等自动生成处方。

[11] 根据《中华人民共和国个人信息保护法》，医疗健康数据属于敏感个人信息.

三、人工智能伦理治理实践

人工智能成为各国科技发展的重要战略，各国通过伦理规范明确人工智能技术研发和应用的基本伦理要求，加强全球人工智能伦理治理合作，构建人工智能有序发展的治理机制。全球科技伦理治理主要包括全球合作下的共识性人工智能伦理原则，各国因地制宜的人工智能伦理规范、指南、工具等，并关注增进人类福祉、可持续发展、保护隐私、防止和减少偏见等。

（一）国际组织人工智能伦理治理方案

人工智能技术应用可能对人类社会产生广泛负面影响，成为全球面临的共同风险。国际社会正加紧推进人工智能伦理治理领域的合作。

在人工智能伦理共识性原则层面，联合国教科文组织 193 个成员国于 2021 年 11 月达成《人工智能伦理问题建议书》，提出“将伦理视为对人工智能技术进行规范性评估和指导的动态基础，以人的尊严、福祉和防止损害为导向，并立足科技伦理”的要求；明确尊重、保护和促进人的权利、基本自由、人的尊严，环境和生态系统发展，确保多样性和包容性，生活在和平、公正的互联网社会中，4 项人工智能价值观；确立 10 项人工智能原则；并提出人工智能伦理治理的 11 项政策建议。同时，ISO、IEC、IEEE、ITU 等国际标准化组织积极推动以人工智能技术为代表的科技伦理标准研制，如 2022 年 8 月 ISO/IEC 发布人工智能伦理和社会问题概述标准（ISO/IEC TR 24368）。**在人工智能具体应用领域的伦理规范层面，**世界卫生组织于 2021 年 6 月发布《卫生健康领域人工智能伦理与治理》指南，分析在卫生健康领

域使用人工智能的机遇和挑战，并提出在医疗领域使用人工智能的伦理政策和确保人工智能为所有国家的公共利益服务的 6 项原则。

目前，人工智能伦理已成为全球人工智能治理讨论的重要议题，以人为本、公平公正等人工智能伦理原则在国际合作机制中不断深化。**在联合国层面**，2023 年 5 月 25 日，联合国发布《我们的共同议程》政策简报 5“全球数字契约—为所有人创造开放、自由、安全的数字未来”。2023 年 10 月 26 日，联合国高级别人工智能咨询机构成立，就人工智能可能产生的偏见歧视等关键问题开展讨论。12 月，该咨询机构发布临时报告《以人为本的人工智能治理》，将包容性、公共利益等伦理原则作为设立人工智能国际治理机构的指导原则。**在区域合作层面**，2023 年 8 月，金砖国家领导人第十五次会晤上同意尽快启动人工智能研究组工作，推动有广泛共识的治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性、公平性。2023 年 9 月，二十国集团（G20）发布《G20 新德里领导人宣言》，提出“负责任地使用人工智能以造福全人类”。

（二）域外国家和地区人工智能伦理治理机制

为协调人工智能技术发展与风险防范，各国政府制定符合其人工智能治理理念、与技术发展情况相适应的宏观规划，并通过颁布人工智能伦理原则、指南、工具包等指导行业实践。其中，既有以美国为代表的以面向市场和创新为导向的监管模式，也有以欧盟为代表的积极介入模式。本部分主要介绍美国、欧盟、德国、新加坡在人工智能伦理治理领域的实践。

1.美国发展以鼓励创新为基础的可信赖人工智能

美国长期将人工智能视作其保持全球竞争优势的重要技术，并在近年来关注可信赖人工智能建设，在联邦政府层面规划和具体事务层面制定相应政策，但目前仍然缺乏以人工智能为规制对象的具有约束力的法律规范。**在行政规划层面，**2023 年 10 月，美国总统拜登发布《关于安全、可靠、可信赖地开发和使用人工智能》行政令，明确进行负责任人工智能技术开发，提出安全可靠、保障权利、隐私保护等伦理要求，并对各行政部门如何促进负责任人工智能技术开发和应用作出安排。此前，2022 年 5 月，拜登政府成立国家人工智能咨询委员会，围绕人工智能安全、防止和减少偏见等开展工作，推动负责任且具有包容性的人工智能技术发展。白宫科技政策办公室于 2020 年提出《人工智能应用的监管原则》，对政府机构是否监管以及如何监管人工智能提出 10 项原则，并强调公众参与及机构间的协调；2022 年 10 月发布《人工智能权利法案蓝图》，确定人工智能的 5 项原则以指导自动化系统的设计部署。同时，美国政府积极鼓励行业自律，2023 年 7 月和 9 月两次推动包括 OpenAI、谷歌、微软、英伟达等 15 家企业就开发避免偏见歧视、保护隐私的人工智能作出承诺。**在具体领域，**美国联邦政府不同部门也依据主责主业发布相应的伦理原则和治理框架。2020 年，美国情报体系发布《美国情报体系人工智能伦理原则》，要求运用人工智能要遵守法律、确保安全、客观公正、透明负责等。2023 年 1 月，美国商务部下属美国国家标准与技术研究院（NIST）发布《人工智能风险管理框架》，细化开发部署负责任人工智能技术

的指南。2023 年 12 月，美国审计署发布报告，公布对联邦 23 个机构负责任使用人工智能的情况评估。

2.欧盟通过健全监管规制落地人工智能伦理要求

欧盟在数字时代积极探索对新技术、新业态的规制方式，着力推动以人工智能为对象的立法，将人工智能伦理原则转化为规范要求，努力在人工智能治理方面保持全球影响力。**在伦理框架方面，**2019 年 4 月，欧盟高级专家组发布《可信人工智能伦理指南》，提出可信人工智能的概念和 7 项关键要求，包括保证人类监督、鲁棒性和安全性、隐私和数据治理、透明度、多样性、社会和环境福利、问责。**在治理落地方面，**2023 年 12 月，欧盟委员会、欧洲理事会、欧洲议会达成共识，就《人工智能法案》达成临时协议，提出基于风险的监管方式，将人工智能系统带来的挑战分为不可接受的风险、高风险、有限风险、基本无风险四个级别，并形成对应的责任和监管机制。同时，针对通用人工智能提出具体义务要求，并强调以人为本、保证透明度等伦理价值要求。

3.德国关注人工智能具体应用领域伦理风险规制

德国联邦政府制定国家人工智能战略，发展德国在智能制造领域的优势，并积极推进人工智能算法、数据以及自动驾驶等领域伦理规制，强调以人为本，发展负责任、以公共利益为导向的人工智能。**在伦理治理机制方面**，2018 年 9 月，德国联邦政府成立数据道德委员会，从政府层面制定数字社会的伦理道德标准和建议。2023 年 11 月 7 日，德国联邦教育和研究部更新人工智能行动计划，提出采用适当、

灵活的人工智能治理，推动值得信赖的人工智能发挥作用。**在治理落地方面**，从数据、算法、自动驾驶应用领域进行分类规范。在数据领域，德国数据伦理委员会强调以人为本的价值导向，提出数据权利和数据义务的认定需要考量不同数据主体的权利、对数据的贡献、公共利益等。在算法领域，德国数据伦理委员会将与人类决策相关的人工智能系统划分为基于算法的决策、算法驱动的决策和算法决定的决策三种类型，并提出人工智能算法风险导向型“监管金字塔”，将人工智能算法风险进行 1-5 级评级，对于评级为无潜在风险的人工智能算法不采取特殊监管措施，并逐级增加规范要求，包括监管审查、附加批准条件、动态监管以及完全禁止等措施。在自动驾驶领域，德国联邦交通与数字基础设施部推出全球首套《自动驾驶伦理准则》，提出了自动驾驶汽车的 20 项道德伦理准则，规定当自动驾驶车辆对于事故无可避免时，不得存在任何基于年龄、性别、种族、身体属性或任何其他区别因素的歧视判断，认为两难决策不能被标准化和编程化。

4.新加坡积极探索人工智能伦理治理技术工具

新加坡政府通过发布“智慧国家”“数字政府蓝图”等国家政策从多维度提升人工智能发展与应用，着力推动社会数字转型有利于人的发展。**在政府规划方面，**自发布“智慧国家”政策以来，新加坡已有超过 20 个行政机构提交人工智能应用规划。2023 年 12 月，新加坡发布国家人工智能战略 2.0，提出人工智能服务公共利益的愿景，再次强调建立一个可信赖和负责任的人工智能生态。**在工具方面，**2022 年 5 月，新加坡通信媒体发展局和个人数据保护委员会发布《人

工智能治理评估框架和工具包》，成为全球首个官方的人工智能检测工具。该工具结合技术测试和流程检查对人工智能技术进行验证，并为开发者、管理层和业务伙伴生成验证报告，涵盖人工智能系统的透明度、安全性及可归责性等人工智能伦理要求，并积极吸收不同机构的测试建议，完善评估工具。2023 年 6 月，新加坡成立由政府、企业等组成的 AI Verify 基金会，以打造人工智能治理开源社区，为人工智能测试框架、代码库、标准和最佳实践的使用和发展提供支持。

（三）我国人工智能伦理治理实践

1.确立科技伦理治理体制机制

科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范，是促进科技事业健康发展的重要保障。我国将科技伦理规范作为促进技术创新、推动社会经济高质量发展的重要保障措施，并逐步完善科技伦理治理顶层设计。2022 年 1 月 1 日起施行的《中华人民共和国科学技术进步法》第一百零三条从法律层面确认“国家建立科技伦理委员会，完善科技伦理制度规范”并明确禁止违背科技伦理的科学技术研究开发和应用活动，在第一百一十二条明确对违背伦理的活动需要承担的法律责任。2022 年 3 月，中共中央办公厅国务院办公厅印发《关于加强科技伦理治理的意见》，提出强化底线思维和风险意识，明确科技伦理原则和科技伦理治理要求，提出加强科技伦理治理五项措施。2023 年 10 月 8 日，科技部、教育部、工业和信息化部等十部门联合发布《科技伦理审查办法（试行）》，对科技伦理审查的基本程序、标准、条件

等提出要求，规范科学研究、技术开发等科技活动的科技伦理审查工作，要求从事人工智能等科技活动的单位设立科技伦理（审查）委员会。《科技伦理审查办法（试行）》将对人类主观行为、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，具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发，面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发等 7 项科技活动列入需要开展伦理审查复核的清单。

2.细化人工智能伦理要求

我国人工智能伦理治理历经发展规划的认可、伦理原则的确立和伦理规范的细化。2017 年国务院印发《新一代人工智能发展规划》，指出人工智能发展的不确定性带来挑战，影响涵盖就业、法律与伦理、个人隐私、国际关系等，必须高度重视人工智能可能带来的挑战，提出到 2025 年初步建立人工智能伦理规范，并结合法律法规和政策体系，共同促进人工智能安全评估和管控能力，提出建立伦理道德多层次判断结构、人机协作的伦理框架、人工智能产品研发人员道德规范和行为守则、人工智能潜在危害与收益的评估、复杂场景下突发事件解决方案等，并重视国际合作的重要性。2019 年 2 月，国家新一代人工智能治理专业委员会成立，成员涵盖高校、研究院所和企业专家，着力推动产学研在人工智能治理方面的合作。2019 年 6 月，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则—发展负责任的人工智能》，提出了人工智能治理的框架和行动指南，提出

人工智能发展相关各方需要遵循和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷协作的原则。2021 年 9 月，国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》，提出人工智能活动的伦理规范包含管理规范、研发规范、供应规范、使用规范，细化 18 项具体伦理要求。

同时，行业主管部门推动人工智能应用领域的伦理规范指引建设，发布实施意见、牵头制定行业标准等，明确具体领域伦理标准和细化措施，促进行业人工智能治理体系的完善。2022 年 10 月，中国人民银行发布《金融领域科技伦理指引》行业标准，提出在金融领域开展科技活动需要遵循守正创新、数据安全、包容普惠、公开透明、公平竞争、风险防控、绿色低碳等 7 个方面的价值理念和行为规范。2023 年 4 月，工业和信息化部成立工业和信息化部科技伦理委员会、工业和信息化领域科技伦理专家委员会，关注人工智能等重点领域的科技伦理治理，提出从加强科技伦理审查和监管，组织制定重点领域科技伦理审查规范和标准，开展重点领域科技伦理敏捷治理，强化科技伦理管理培训和宣传教育，加强人才队伍建设等方面提升伦理治理能力。

3.行业积极探索人工智能伦理治理落地措施

人工智能产业主体作为人工智能伦理管理责任主体，承担着人工智能向善发展的重要责任。**在行业层面，**行业组织积极推动人工智能伦理原则落地，包括中国人工智能产业发展联盟等开展人工智能伦理技术和管理标准制定、监测认证、典型案例分析汇编等，关注实现人工智能系统可解释性、隐私保护、公平等技术实施路径，促进人工智

能应用相关行业伦理规范的提升。**在企业层面，**多家企业积极落实人工智能伦理管理要求，包括建立科技伦理委员会、人工智能伦理委员会等内部机构；加强企业人工智能科技伦理管理机制的完善；加强外部多学科合作，通过引入法学、哲学、管理学、人工智能技术等多个方面的外部专家，提升人工智能伦理治理的外部监督；定期发布企业人工智能伦理治理情况与研究，增强公众沟通；积极探索人工智能伦理问题技术解决方案，通过技术创新和技术能力升级提升个人隐私的保护力度、算法可解释性、模型可靠性等。

4.积极参与人工智能伦理治理国际合作

近年来，我国不仅在人工智能技术研发应用领域加强技术创新合作，也积极参与全球人工智能伦理治理。2023 年 8 月，在金砖国家领导人第十五次会晤上，金砖国家已同意尽快启动人工智能研究组工作，拓展人工智能合作，形成具有广泛共识的人工智能治理框架和标准规范，提升人工智能技术的安全性、可靠性、可控性、公平性。2023 年 10 月，我国发布《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理中国方案，提出坚持伦理先行的原则，提出建立并完善人工智能伦理准则、规范及问责机制。同时，我国专家也积极参与联合国、世界卫生组织等国际机构的人工智能伦理规则构建。2023 年 10 月，我国两位专家入选联合国高级别人工智能治理机构，积极参与联合国层面的人工智能治理讨论，提出全球人工智能治理建议。

四、人工智能伦理治理展望

（一）协调人工智能产业创新发展与伦理治理

人工智能伦理治理需要关注产业科技创新活动与伦理风险防范协调发展，通过伦理治理对人工智能可能存在的风险进行全生命周期的指导，促进人工智能朝着有利于人类发展、尊重和保障各群体权益的方向发展。**重视人工智能技术创新。**坚持促进创新与防范风险相统一，鼓励人工智能基础技术创新和突破，发展人工智能芯片、数据、算法等基础核心技术，加强人工智能治理产品开发，以人工智能技术防范人工智能风险。**探索建立敏捷的人工智能伦理治理机制。**建设快速响应、创新主体与相关政府部门有效沟通的治理合作，增强科技伦理治理的有效性和科学性，推动产业创新与伦理风险防范相协调。

（二）完善人工智能伦理治理举措

1.健全多学科多主体合作的治理体系

形成多学科共建的人工智能治理合力。人工智能伦理治理是多学科合作的治理实践，要推动技术发展、政策引导、管理规范、伦理研究、法律规制等一系列专业知识的交流配合，实现人工智能伦理治理落地。**建立多主体参与的人工智能治理生态。**人工智能伦理治理是监管主体、创新主体、公众和其他利益相关协同合作的多元治理。人工智能伦理治理需要监管主体完善审查监管规范，制定可预期的人工智能伦理研发、应用伦理规范。人工智能创新主体需要履行科技伦理管理主体责任，建立科技伦理日常管理机制，及时化解新技术带来的风险，将人工智能伦理风险在源头予以防范。公众可以提升伦理素养，

对可能影响个人思维、生活的人工智能技术应用进行了解，并反馈发现的人工智能伦理问题。

2.建立分类分级伦理治理机制

根据人工智能伦理治理原则和目标识别人工智能应用的伦理风险。以增进人类福祉、尊重生命权利、公平公正、公开透明、控制风险等人工智能伦理原则要求为基础，对人工智能从技术研发到部署应用进行全生命的风险识别，并根据技术和产品发展及时进行更新调整。**根据人工智能伦理风险大小和影响范围确定责任义务和监管规则。**对于基本不具有伦理影响的人工智能技术应用，简化监管程序，鼓励技术创新和发展；对于具有一定伦理影响的人工智能技术，设定伦理要求，建立风险评估机制，明确责任；对于具有长期且广泛影响的人工智能高风险应用领域，加大监管力度，通过多种监管手段防范伦理风险；对于具有不可接受伦理影响的人工智能应用，考虑禁止部署。

3.推动人工智能伦理治理技术化、工程化、标准化

支持人工智能伦理治理工程化和技术化。将人工智能伦理原则转化为工程问题，用技术工具推动伦理原则落地，以人工智能技术应对人工智能风险。积极发展人工智能产品伦理风险评估监测工具，推动应对人工智能偏见歧视、隐私泄露、不可解释等伦理风险的技术工具研发。**研制人工智能伦理治理有关标准。**研制关于人工智能公平性、透明性、可解释性等技术指标的国际标准、团体标准、行业标准，发挥标准引领性作用，形成人工智能伦理治理经验示范。

（三）提升各主体人工智能伦理风险应对能力

符合人工智能伦理的开发和应用实践有赖于研发、设计、应用人员承担起个人责任，也需要公众加强对人工智能伦理风险的认知及负责任使用。**支持高校开设科技伦理课程引导学生和科技工作者提升人工智能伦理素养。**通过必修课、选修课等课程设置提升学生科技伦理素养，特别是加强人工智能相关专业人员伦理知识，并为伦理学、法学等专业学生提供人工智能通识课程，形成人工智能伦理学习研讨氛围。**要求企业持续推进员工人工智能伦理培训。**将开展面向员工的科技伦理培训作为企业落实科技伦理管理主体责任的必要环节，特别关注技术研发、产品开发等员工培训，加强入职培训和日常的伦理宣传。**引导行业加强科技伦理自律与合作。**推动联盟等行业平台促进人工智能伦理治理合作，引导制定行业自律举措，形成人工智能行业、典型应用领域等的伦理实践指南。鼓励企业将科技伦理经验和工具向产业链上下游企业、中小企业进行分享，提升行业伦理管理能力。支持科技类社团和行业平台开展政策解读、座谈交流等，加强人工智能等重点领域科技伦理教育宣传培训。**加强面向社会公众的科技伦理教育。**引导社会公众认识科技风险，倡导合理正确使用生成式人工智能等技术工具。通过小说、影视剧、短视频等多种形式普及科技伦理知识，推动科技伦理问题讨论，形成全社会尊重科技伦理的氛围。

（四）加强人工智能伦理治理国际交流合作

人工智能技术具有全球性、时代性，人工智能伦理治理是全球科技治理的重要组成部分，也已成为复杂国际竞争环境中各国合作的重

要切入点。**积极参与全球科技伦理治理双多边合作。**积极参与联合国等多边机构的人工智能伦理治理交流合作，推动科技伦理治理国际共识、国际标准的研讨制定。加强区域性科技伦理治理合作，加强与有关国家的科技伦理治理交流，深化科技伦理治理经验分享与合作。**鼓励国内企业和专家学者参与国际人工智能伦理治理交流合作。**鼓励产业界积极分享国内人工智能伦理治理实践，推动学术界与研究机构加强与域外机构对人工智能伦理治理议题的深入研讨，形成全球人工智能伦理治理合作生态。

中国信息通信研究院 知识产权与创新发展中心

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62304212

传真：010-62304101

网址：www.caict.ac.cn