人工智能应用中的伦理风险与信息监理机制

王骆明
浙江天航咨询监理有限公司　浙江杭州　310000

【摘　要】伴着科技飞速前行，人工智能深深融入社会生活各处，像医疗诊断、金融交易、交通出行以及智能安防等，它所给予的高效和便捷众人皆见。不过，在人工智能运用持续扩展之际，不少伦理隐患也慢慢浮现。诸如算法偏差、隐私遭侵、责任划分不明这类问题，愈发受到重视。本文着眼于人工智能应用里的伦理风险，深度探究其形成缘由，同时尝试搭建有效的信息监管机制，目的是在促使人工智能技术进步的过程中，维护社会伦理规范和信息安全。

【关键词】人工智能；伦理风险；信息监理；算法；隐私

Ethical risk and information supervision mechanism in the application of artificial intelligence

Wang Luoming
Zhejiang Tianhang Consulting Supervision Co., LTD. Zhejiang Hangzhou　310000

【Abstract】With the rapid advancement of technology, artificial intelligence has deeply integrated into various aspects of social life, such as medical diagnosis, financial transactions, transportation, and intelligent security. The efficiency and convenience it brings are widely recognized. However, as the application of AI continues to expand, numerous ethical concerns have gradually emerged. Issues like algorithm bias, privacy invasion, and unclear liability are receiving increasing attention. This article focuses on the ethical risks associated with AI applications, delving into their causes and attempting to establish effective information regulatory mechanisms. The aim is to promote the progress of AI technology while maintaining social ethical norms and information security.

【Key words】artificial intelligence; ethical risk; information supervision; algorithm; privacy

引言

人工智能堪称现今极具变革性的技术之一，正以空前之速重塑整个世界。智能家居赋予生活便捷，智能物流提高供应链效能，智能教育达成个性化学习，各类应用场景持续拓展，为人类带来庞大发展契机，大幅增进生产效率，改善生活品质。然而，技术向来利弊兼具，人工智能在应用时所暴露的伦理难题不容小觑。此类伦理隐患，既可能对人工智能技术的良性发展形成阻碍，又会对社会公平性、个人权益等方面产生不良影响。像部分司法辅助的人工智能系统，因算法存在偏差，或许会致使对特定群体出现不公正审判趋向；智能摄像头的普遍应用，也蕴含侵犯居民隐私的潜在危机。深入钻研人工智能应用中的伦理风险，构建与之契合的信息监管机制，成为当下急需解决的关键课题。

一、人工智能应用中的伦理风险表现

（一）算法偏见导致不公平决策

人工智能算法依靠海量数据展开训练，一旦数据存有偏差或者不够完整，就极有可能引发算法偏见。于招聘场景而言，部分招聘平台所运用的人工智能筛选算法，由于历史数据存在性别、种族歧视等因素，或许会针对某些求职者给出不公平的筛选成果。像某些地区长久以来存在的职场性别歧视状况体现于过往招聘数据之中，算法学习这些数据后，说不定会无意识地降低女性求职者简历筛选的优先等级，即便这些女性拥有出色的专业技能以及丰富经验。此类基于算法偏见所做出的决策，有悖公平公正准则，制约了部分群体的发展契机，还可能进一步加深社会不平等程度，致使弱势群体在职场竞争里愈发艰难。

（二）隐私侵犯与数据滥用

人工智能系统的运转离不开海量数据的收集，这些数据包含用户个人信息、行为习惯这类敏感内容。不少企业或机构在数据收集、存储以及使用环节，缺少严谨的隐私保护手段，很容易造成用户隐私的泄露。像是智能音箱、智能摄像头等智能设备，可能在用户毫无察觉时收集数据，并且将数据用于商业用途。曾有报道表明，某些智能音箱在用户日常交谈期间，会悄悄记录下关键词，随后把这些数据传送给关联企业，以便进行精准广告投放。数据滥用的状况同样显著，

诸如把用户数据卖给第三方用于非法活动。有些不法分子通过购买用户的医疗数据，实施保险欺诈等违法操作，极大地损害了用户的权益。

（三）责任界定模糊引发信任危机

一旦人工智能系统做出决策或者引发不良后果，责任究竟该由谁承担常常难以清晰界定。就拿自动驾驶汽车发生事故的情况来说，究竟责任在于汽车制造商、编程人员，还是车辆使用者呢？打个比方，曾有一起自动驾驶汽车在某种特定路况下未能及时刹车从而导致追尾事故[1]。汽车制造商宣称这是因为编程算法对复杂路况的识别能力存在局限；编程人员却认为是传感器数据采集不够精确；而车辆使用者觉得自己完全是依照正常操作流程在驾驶。这种责任界定的模糊状态，致使受害者很难获得合理的赔偿，同时也使得公众对人工智能技术的可靠性和安全性心生疑虑，进而对人工智能技术的推广应用造成影响，阻碍行业的发展进程。消费者在考虑购置自动驾驶汽车时，会因为对责任不明确的担忧而举棋不定。

二、伦理风险产生的根源

（一）数据质量与偏差问题

数据采集阶段，采样手段不当、样本欠缺典型性等现象颇为常见。于社会调研数据采集实践中，若仅着眼城市繁华地段采样，却忽视偏远地区与农村等群体，依此数据训练出的人工智能模型，在解析社会整体态势时难免产生偏差。部分数据受采集技术制约，难以完整、精确展现真实状况。就图像识别算法训练来讲，若训练数据集中某类图像占比甚微，算法识别此类图像时便容易失误，从而引发伦理隐患。在医疗影像识别范畴，若训练数据里罕见病的影像样本稀缺，算法对罕见病的诊断精准度就会锐减，极可能延误患者诊疗时机。

（二）技术黑箱与不可解释性

诸多人工智能算法，像深度学习里的神经网络，架构繁杂，内部运作机理恰似 “暗箱” 一般。就连开发者都难以透彻明晰算法究竟怎样做出决策。拿谷歌的 AlphaGo 来说，它在与人类棋手对弈之际，落子决策流程难以借传统逻辑予以清晰阐释。这般不可解释特性致使算法一旦出现伦理问题，追溯根源并加以修正困难重重[2]。在医疗诊断领域，倘若人工智能给出的诊断结论无法阐明依据，医生与患者便很难安心采纳，同样制约人工智能在关键领域的应用程度。医生参考人工智能诊断建议时，因无从得知其推理路径，或许不敢贸然更改原有的治疗方案。

（三）利益驱动与监管缺失

不少企业在人工智能技术的研发与运用里，过于看重商业盈利，轻忽伦理方面的思考。为节省成本、增进效率，对于数据安全维护、算法优化改进等环节，投入可能并不充分。像有些小型互联网企业，为快速推出产品抢占市场，在数据加密技术上不舍得多花钱，让用户数据安全面临的隐患大幅增加。当下，关于人工智能伦理的法律法规与监管体系并不完备，存在诸多空白之处。缺乏有效的外部管束，使一些企业在碰到伦理风险时，难以及时做出纠正行动。在数据跨境传送方面，因为没有清晰的法律规定，部分企业随意把用户数据传至境外，有着数据泄露的潜在危机。

三、信息监理机制的构建原则

（一）公正性原则

信息监理机制在应对人工智能伦理问题期间，理应坚守公正原则，对所有利益相关方一视同仁[3]。当审视算法有无偏见时，需运用客观且公正的尺度予以评判，针对那些因算法偏见而遭受不公待遇的群体，给予恰当补偿并加以纠正，以此维护社会公平正义。以公共资源分配的人工智能系统为例，一旦察觉算法对低收入群体存在资源分配欠缺的偏见，监理机制便要责令相关部门重新优化算法，并对先前遭受不公的群体实施资源倾斜补偿。

（二）隐私保护优先原则

拟定严谨的数据采集、存储以及运用规则，清晰界定企业与机构于数据处理流程里的责任和义务。强制规定它们在收集数据之前务必取得用户的明确许可，运用诸如加密等技术手段来捍卫数据安全，避免隐私遭到泄露与滥用。拿社交平台来说，在采集用户位置信息之际，一定要向用户详尽说明采集目的、使用方式，且要让用户通过主动勾选予以同意。于数据存储阶段，运用先进的加密算法，例如 AES 加密技术，对用户数据实施加密存储，防范数据被非法盗取。

（三）透明度与可解释性原则

积极推进人工智能算法的透明度构建工作，规定开发者必须将算法原理、数据出处、训练流程等关键信息予以公开披露[4]。针对复杂的算法模型，要给出便于理解的阐释说明，使得用户与监管者能够知悉算法做出决策的依据。在金融范畴的智能风控体系里，需要向用户阐释信用评估算法的运作逻辑，以此提升公众对人工智能系统的信赖程度。以银行运用人工智能评估用户信用额度为例，银行应当向用户说明算法主要参照的因素，诸如还款记录、消费习惯等，并且用通俗易懂的方式讲解这些因素怎样对信用额度的评定产生作用，让用户清晰知晓自身信用得分的得出过程。

四、信息监理机制的具体措施

（一）建立伦理审查委员会

伦理审查委员会集合了计算机科学、伦理学、法学等诸多领域的专家。计算机领域专家对算法潜在风险了如指掌，伦理学界学者负责深入剖析伦理层面的争议点，法学专家则致力于确保项目全程合法合规推进。拿某互联网企业启动的智能客服项目来说，特意邀请高校教授、伦理学者以及法律专业人士共同参与其中。高校教授全力对算法进行优化改良，伦理学者仔细排查项目里可能存在的不当引导情况，法律专家严谨审查数据相关环节。在项目立项初始阶段，委员会针对项目需求文档展开详细分析，并对伦理风险展开全面评估，尤其是针对青少年学习辅助产品，会着重严格审查其价值观导向以及数据收集方式。在项目开发过程中，委员会定期对算法代码进行审查，同时仔细核查数据来源。项目成功上线之后，专门设置了反馈通道，根据用户反馈意见对算法加以优化，就像处理智能客服回复问题的优化工作。在医疗人工智能产品正式上市之前，委员会同样会进行全方位审查。以糖尿病诊断软件为例，需要评估数据样本的完整性，审查数据加密手段是否妥当，检查结果呈现方式是否合理，一旦发现算法存在缺陷，便会要求开发者进行优化，直至达到相应标准。

（二）完善法律法规与行业标准

立法机构宜尽早颁布人工智能伦理专门法规，界定数据保护、算法责任、隐私侵权等法律责任领域，诸如限定企业存储用户数据的时长、依据算法情形确定相关方责任、加重侵权处罚力度[5]。行业协会应当拟定统一规范，对开发应用流程予以规范，比如明确算法透明度、规定数据安全等级等，并借助培训之类方式加以推广。鉴于人工智能具有全球性特征，各个国家有必要强化国际协作，借鉴其他国家的经验，共同应对难题，可举办研讨会等活动，一同商讨全球人工智能伦理准则。

（三）加强技术监管手段

要达成对人工智能系统实时监测与评估，我们借助先进技术打造出专用的算法审计工具。其能自动察觉算法偏见以及数据滥用情形，通过深度剖析算法的输入数据、输出成果及运行中的中间参数来精确判断潜在风险，比如在电商平台商品推荐算法审计时，工具会研究推荐结果与用户浏览、购买数据的联系以及不同用户群体推荐的差别，一旦发现小众品牌商品推荐频次低，可能存在品牌偏见，便马上发出预警以促使平台优化算法；与此同时，我们构建起数据安全监测平台，运用网络流量监测技术实时把控数据流向，严格审核向外部第三方的数据传输行为。监测到大量用户敏感数据流向陌生 IP 地址，平台即刻报警并阻断传输，并且平台还监测数据存储环境，包括服务器物理安全和数据加密状态，定期生成数据安全报告助力企业评估自身数据安全状况；我们运用区块链技术实现数据可追溯，在数据收集环节，采集时间、采集者、数据内容等信息都记录在区块链上形成不可篡改的记录，科研数据收集便是如此操作。数据使用期间，每次访问、使用目的、产生结果等信息也被记录下来，让数据使用过程清晰明了，一旦出现数据滥用或者隐私侵犯问题，可凭借区块链追溯数据流转轨迹，明确责任主体。

结语

人工智能技术的进步极大地推动了社会发展，然而伦理风险也相伴而生。深入探究伦理风险的具体呈现及根源所在，以公正性、隐私保护优先、透明度与可解释性为原则构建信息监理机制，通过设立伦理审查委员会、健全法律法规与行业标准、强化技术监管手段等切实举措，可有效防控人工智能应用过程中的伦理风险。有利于促进人工智能技术朝着健康、可持续的方向发展，还能切实保障社会的公平正义，维护个人的合法权益，从而让人工智能更好地造福人类社会。

参考文献

[1]张铤.人工智能的伦理风险治理探析[J].中州学刊，2022，（01）：114-118.

[2]薛桂波.生成式人工智能的伦理风险及其防范对策[J/OL].南京林业大学学报（人文社会科学版），2025，（01）：1-11[2025-04-11].

[3]林雨佳.生成式人工智能对信息治理的挑战与应对[J].苏州大学学报（哲学社会科学版），2025，46（02）：104-115.

[4]覃露.基于人工智能的信息安全防护系统研究[C]//中国智慧工程研究会.文化传承与现代化治理学术交流会论文集.广西民族大学；，2024：316-317.

[5]郭佳楠.生成式人工智能大模型的伦理风险与规制策略——以 DeepSeek 为例[J/OL].华北水利水电大学学报（社会科学版），1-10[2025-04-11].