Dispute Settlement 争议解决, 2024, 10(5), 175-180
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ds
https://doi.org/10.12677/ds.2024.105268

Hans汉斯

总体国家安全观视域下人工智能风险防范与规制

李依阳

华东政法大学国际法学院，上海

收稿日期：2024年4月9日；录用日期：2024年4月17日；发布日期：2024年5月20日

摘　要

随着大数据、云计算等先进技术的迭代优化，各类人工智能深度嵌入公共领域已是大势所趋。在对社会发展不断提供技术支撑之余，新兴人工智能技术在意识形态、网络安全等多方面诱发的国家安全风险日益引发担忧。国家政策战略引导之下，中国人工智能产业正呈蓬勃态势发展，但也面临着前所未有的发展阻力。鉴于日趋严峻的国内外形势，我们亟须加深对新兴技术可能诱发多领域安全风险的思考，并在总体国家安全观思想引领下，通过对现有治理框架进行秩序重构以提升中国对新兴技术发展的动态适应能力，强化人工智能时代的国家安全治理。

关键词

人工智能，总体国家安全观，风险防范

Prevention and Regulation of Artificial Intelligence Risks in the Context of Overall National Security Outlook

Yiyang Li

School of International Law, East China University of Political Science and Law, Shanghai

Received: Apr. 9th, 2024; accepted: Apr. 17th, 2024; published: May 20th, 2024

Abstract

With the iterative optimization of advanced technologies such as big data and cloud computing, the deep embedding of various types of artificial intelligence in the public sphere has become a

文章引用: 李依阳. 总体国家安全观视域下人工智能风险防范与规制[J]. 争议解决, 2024, 10(5): 175-180.
DOI: 10.12677/ds.2024.105268

general trend. While continuously providing technical support for social development, emerging AI technologies are increasingly raising concerns about national security risks induced by ideology, network security, and other aspects. Under the guidance of national policies and strategies, China's AI industry is thriving but also facing unprecedented development resistance. Because of the increasingly difficult situation at home and abroad, we must deepen our thinking about the security risks emerging technologies may induce in various fields. Under the guidance of the overall concept of national security, we should enhance China's ability to dynamically adapt to the development of emerging technologies by restructuring existing governance frameworks to strengthen national security governance in the era of artificial intelligence.

Keywords

Artificial Intelligence, Overall National Security Concept, Risk Prevention

1. 引言

1956 年，人工智能的概念首次在达特茅斯会上被麦卡锡、明斯基等科学家提出。新一代工业智能技术在经历了三次“飞跃期”后，在海量数据积累、算力提升和算法优化的加持下，成为推动新一轮科技革命的关键力量。在数字经济和科技全球化的当下，人工智能已经深度嵌入社会生活视野，并衍生出大量的传统安全和非传统安全风险，逐渐从科技领域向国家安全领域蔓延。依据 2023 年 4 月美国斯坦福大学人工智能研究所(Stanford HAI)发布的《2023 年人工智能指数报告》(Artificial Intelligence Index Report 2023)，人工智能滥用事件数量正处于上升态势。通过追踪人工智能道德滥用相关事件的人工智能算法和自动化事件和争议(AIAAIC)数据库发现，从 2012 年起，10 年内人工智能相关争议事件增加了近 26 倍[1]。

为应对我国日益复杂严峻的国家安全态势，2014 年中央国家安全委员会第一次会议时首次提出总体国家安全观这一重大战略思想。总体国家安全观丰富了党的国家安全思想，是对传统国家安全理念的重大突破。作为一种具有丰富的内涵和外延，且能够统筹传统与非传统两个方面的安全的非传统安全观，总体国家安全观能够不断吸收新的要素，适合作为分析人工智能安全风险的理论工具从而作为人工智能治理的思想指引。

2. 人工智能应用背后的国家安全风险

2.1. 人工智能应用产生的传统国家安全风险

基于算法和数据的生成式人工智能易引发意识形态风险。2023 年 2 月，OpenAI 公司宣布为 ChatGPT 加入价值观设定，即宣扬并维护其认为的“普世价值”，包括保障人权、公正、诚信、自由和平等。换言之，ChatGPT 的数据来源于西方国家的语料库，以西方民众价值观为核心取向。此类情形下可能导致通用大型模型被利用为“认知战”工具，从而加速西方国家在全球范围内推广其意识形态。北京理工大学计算机学院研究人员曾于功能测试中发现，ChatGPT 生成的内容中存在大量对中国的偏见言论，并且对涉及中国政治的言论并未避免或拒绝回答，表明训练数据的构建过程中未对这部分言论进行筛查[2]。

目前阶段下 ChatGPT 难以实现对中文内容进行准确的识别判断和内容过滤，若部分境外势力借此大量提供如对华偏见和涉黄涉暴等不良信息，那么“高度拟人”的生成式人工智能，最终或沦为外国对中国进行意识形态渗透和舆论攻击的工具。

军事方面人工智能也为国家安全乃至国际法治理增添了复杂性。首先，人工智能技术的嵌入为致命性自主武器系统(Lethal Autonomous Weapon Systems, LAWS)的扩大适用提供了新的契机。在现有国际军控机制不能有效防治的情况下，致命性自主武器系统存在法律责任的归属模糊不清、技术不成熟以及违背军事必要性等问题，动摇人类社会战争伦理和国际人道主义规范。2019 年《特定常规武器公约》(Convention on Certain Conventional Weapons, CCW)第八次政府专家组会议提出了规范致命性自主武器系统开发及应用的涉及技术、法律规范、道德伦理等方面的 11 条指导原则。尽管国际社会尚未就全面禁止人工智能武器达成全球共识，但明显已开始重视防范和规避人工智能武器衍生的风险问题。此外，随着人工智能在现代战争中的应用，可能引发新一轮军备竞赛，对传统军事力量格局造成影响，导致新的“国际安全困境”的出现。

2.2. 人工智能应用孕育的非传统安全风险

人工智能受到全球热捧的同时，不免对数据的收集和处理合规、数据流动与信息保护等具有更高的要求。在数据输入环节，由于逐一征求用户同意难以实现，故而在前期训练和模型优化阶段存在数据源合规风险；在数据储存环节，存在着数据泄露风险，可能引起公民隐私权益侵害、不正当竞争纠纷以及对国家安全的威胁。此前，微软以及亚马逊等公司便已对内部员工进行警告，禁止员工向 ChatGPT 分享公司机密或敏感数据，以防被 ChatGPT 当作迭代训练数据并用于反复的模型训练。保持警惕实为必要，OpenAI 公司就曾于 2023 年 3 月 24 日发表声明称，有 1.2%的会员用户数据存在泄露风险，其中包含姓名、聊天记录、邮箱和地址等多项隐私信息[3]。

算法安全隶属于广义国家安全概念中的科技安全。确保算法安全要求算法本身的体系保持完整性和有效性，特别在国家重点领域要维护核心算法技术的安全性和可控性。整体而言需要算法具备保障持续安全状态的能力，以达到维护国家核心利益和确保国家安全的目标[4]。近年来，美国政府、国防部及智库等核心机构陆续发布了系列算法与人工智能安全领域的专业报告，这深刻反映了美国在新兴技术安全保障方面所展现出的前瞻布局与重视态度。为此为鉴，在新安全格局下系统完善新技术领域国家安全制度，是维护国家安全的必要之举。

在交通领域，陆运汽车与航运船舶均不可避免地面临着由人工智能等新兴科技所推动的产业结构调整，然而，这种变革也带来了由关键技术所引发的交通运输安全风险。2019 年 2 月，浙江宁波发生的“东方盛”轮与“浙象渔 46102”轮碰撞事故，其根源在于自动舵技术的缺陷。根据航行数据记录，当两船相距约 4.6 海里时，“东方盛”轮的二副便已通过视觉、雷达及电子海图观测到“浙象渔 46102”轮。然而，直至两船距离缩减至 1.9 海里时，自动舵才启动避让行为，但所采取的措施并不足以避免碰撞，最终导致了“浙象渔 46102”轮的沉没，造成 5 人失踪的严重后果[5]。

在医疗产业中，人工智能已广泛渗透到疾病预测、辅助诊断、医学影像分析以及健康管理等多个环节，成为医疗领域的新兴助力。特别在医学影像领域，智能诊疗仪器通过扫描胸部 CT 照片，能够精准识别肺结核等疾病，显著将误诊率从 3.5%降低至 0.5% [6]。美国食品药品管理局已批准多款人工智能工具，用于诊断中风和脑出血、检测心房颤动以及解读脑部核磁共振成像[7]。然而，在传统医疗服务模式受到新兴技术冲击的背景下，从法律视角审视，医疗人工智能领域面临着侵权责任界定困难、人工智能主体及算法设计者责任承担不明等规则疏漏。从医疗活动角度看，医疗人工智能还存在诊疗活动自主性受限、诊疗过程透明性不足、医疗算法可能存在的偏见性诱导以及实践环节安全性缺失等多重风险。

3. 中国人工智能国家安全风险治理现有路径分析

3.1. 中国人工智能治理的制度框架

中国对人工智能风险治理的规制已初步构建成体系化框架，自上而下囊括法律法规、部门规章、国家标准与行业自律标准等多维度治理规范，形成硬法与软法互为补充的规范化治理格局[8]。中国对人工智能风险治理的战略政策主要包括强制类、市场类、引导类、自愿类几类[9]，并引导法律法规、伦理道德规范以及技术标准的制定方向。2015 年 5 月，《中国制造 2025》作为我国制造强国战略的首个十年行动纲领，明确提出了加强智能制造与工业互联网建设的目标，为人工智能技术的实践应用提供了有力支撑。随后，在 2018 至 2019 年间，国家进一步加大了对人工智能发展的引导力度。2018 年发布的《全国人工智能标准化战略规划》提出了标准制定与推广应用等具体举措，有助于促进人工智能行业的规范有序发展。自 2020 年起，政策发展进入稳定阶段。《国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》明确了未来五年乃至更长时期的经济社会发展目标，其中特别强调了人工智能、数字经济等战略性新兴产业的发展目标和政策支持。从宏观战略层面来看，中国对人工智能的重视程度日益提升，政策颁发单位逐渐扩展至国家级、省部级政府单位，政策类型多样化且数量显著增加，涉及领域也日趋广泛。

中国积极致力于构建人工智能建设与治理的法律体系，并陆续出台了一系列专门性标准规范与法律法规。2020 年，多个部门联合发布了《国家新一代人工智能标准体系建设指南》，旨在推动人工智能产业技术研发与标准制定的协同发展。在法律法规方面，中国已构建起以《网络安全法》《个人信息保护法》及《数据安全法》等为基础的法律框架，同时辅以《互联网信息服务管理办法》和《网络信息内容生态治理规定》等法规与规章，为人工智能技术的健康发展提供了坚实的法律保障。在技术标准领域，中国不仅发布了包括《信息安全技术步态识别数据安全要求》在内的多项国家标准，还涵盖了电力、电信、金融等 18 个行业的网络安全标准，为人工智能技术在各领域的广泛应用提供了技术支撑与规范指导。

在伦理治理日益获得国际共识的背景下，中国高度重视将伦理准则深度融入人工智能的全周期发展之中。2019 年发布的《人工智能伦理指南》从伦理维度切入，为人工智能的研发与应用确立了伦理原则与规范，着重强调了所应承担的社会责任与风险管理的重要性。随后，2021 年 9 月 25 日推出的《新一代人工智能伦理规范》进一步细化了人工智能在管理、研发、供应、使用等特定环节中的具体伦理要求，充分响应了当今社会在促进公平公正、保护隐私安全、提升伦理素养等方面的广泛关切与期待。这些伦理规范不仅为中国人工智能产业的健康发展提供了伦理指引，也为全球人工智能伦理治理贡献了中国智慧与方案。

3.2. 中国人工智能治理的路径反思

首先，中国在人工智能战略政策方面虽取得进展，但统筹协调机制仍有不足。纵向维度看，人工智能相关政策的制定主要集中在国务院层面，其他中央机构、地方政府组织及部门则在此基础上进行补充完善，并负责具体执行。然而，在实际推进过程中，各地区在政策制定与实施时可能出现目标不一致、资源不共享等问题，这在一定程度上影响了政策的有效执行和整体效果。横向维度看，各地区在推进人工智能发展时所制定的政策及制度有时会出现重叠、冲突等现象。这种政策间的差异性和不一致性不仅可能导致资源浪费，还可能给企业和研究机构带来困扰，使得其难以在统一的市场环境中进行公平竞争和创新发展。

其次，中国的人工智能法律制度尚待进一步完善。由于法律的滞后性，面对日新月异的人工智能技术和应用场景，现行法律在监管和规制方面难免力不从心。以自动驾驶技术为例，当涉及侵权责任的认

定时，传统的法律框架难以有效应对算法决策带来的复杂问题。算法造成的损害往往难以用传统的因果关系和主观过错标准来分析，而算法设计者也可能利用技术局限性进行抗辩。中国在人工智能监管方面仍存在较大的空白，亟须建立更为完善和健全的监管体系，以加强对人工智能技术和应用的监督和管理。

最后，相较于其他主要发达国家或地区，中国的科技伦理监管立法总体上滞后。虽然已有《人工智能伦理指南》《关于加强科技伦理治理的意见》等文件作为理论支撑，但对人工智能伦理道德问题的规范仍较为宽泛，缺乏具体、有效的指引和规范。在针对数字科技不同领域的专门性立法中，科技伦理的相关内容也鲜有涉及，这导致在实际操作中难以对人工智能的伦理问题进行有效监管。

4. 总体国家安全观视域下我国人工智能风险防范进路

4.1. 均衡发展安全的整体治理思路

就如何从总体国家安全观视域下看待新兴技术，习总书记从多个角度进行了重要阐述。一是要准确把握当前国家安全所面临的内外形势。二是要深刻理解发展与安全的辩证统一关系[10]。在总体国家安全观的指导下推进人工智能治理，要将全面治理与系统治理理念融入全过程。同时，人工智能研发与应用安全风险防范与化解需以国家整体利益与安全为导向，实现“安全”与“发展”双线并行。在国际竞争激烈的背景下，中国需双管齐下应对挑战。一方面，要增强科技硬实力，重点研发关键领域核心技术，如基础理论、算法、芯片等，以提升国际科技话语权，减轻外部制裁影响。另一方面，完善人工智能风险评估与保障体系至关重要，需建立事前预防、事中控制、事后处置的完整机制，全面监控治理各环节风险。建立诸如技术风险评估机制、灾难性风险管理机制和技术错误纠正机制等相关可管理安全(managed security)机制[11]。同时，统筹法律法规、伦理规范与技术标准，实施差异化治理，构建中央与地方协调配合的监管体系，并支持区域试点探索。通过这些举措，确保人工智能安全可控，促进国家安全和可持续发展。

4.2. 引导算法向善与多方协作治理

可靠、安全、稳定的人工智能算法设计是确保其平稳运行的关键。在算法的设计过程中，必须坚守“以人为本”的基本原则，将道德价值观念融入其中，使之成为具有社会认同的智能体。具体而言，算法设计需充分考虑保护人身财产利益、个人隐私以及维护公序良俗等伦理规范，并将这些规范转化为程序语言，嵌入到算法的运行过程中。这不仅是对算法行为边界的明确，更是作为研判评估算法潜在威胁与风险的警戒线。同时，面对意识形态领域的潜在风险，必须对人工智能保持高度警惕，严格厘清与审查人工智能的应用场景，防止其被用作政治策反或传播西方价值观的工具。在此基础上积极引导和改造人工智能算法，使其更好地服务于社会主义事业和人民的需求。

在算法治理层面，实现多方参与和共治是确保协同合作治理的关键。政府及相关机构应扮演掌舵者与监督者的角色，加强与平台之间的合作，建立算法风险预警和跟踪研判机制。通过及时敦促平台遵守强制力规范与伦理道德标准，预防和减轻潜在的权益损害。在此过程中，第三方力量可成为治理框架中的重要辅助。学术性组织、非营利机构及自媒体等可适当参与，为算法治理提供多元化的视角和建议。德国政府的先进经验值得借鉴，如“监控算法”(Algorithm Watch)这样的非营利性有限责任公司(gGmbH)，通过评估、监控算法决策过程，揭示潜在风险，并推动透明度和问责机制的建设[12]。

4.3. 维护数字人权与提升用户素养

维护数字人权，应当对人工智能在弱人工智能时代的辅助工具属性有明确的认知，人工智能算法技术虽然有高效特性与自我学习能力，但拥有决策权与自主地位的仍应是作为使用主体方的人类，人工智

能应用对象的自主权不容忽视。目前阶段人工智能大部分只起辅助参考作用，更多作为一种工具手段而不具有观念上与法律上的主体地位。同时要进一步细化用户在数字空间的权利，比如数据隐私权、数据知情权、数据公开使用权、数据财产权等事关数字身份的子权利，细化具体数字人权，以制度保障权利行使，使人们更易跨越“数字鸿沟”而寻得公平正义。

从使用端考虑，需要逐渐培养用户认知与风险防范意识，提升用户算法素养与自主性。从理论层面，努力使用户充分了解人工智能运行的原理以及固有缺陷；从实践层面，提升用户在使用人工智能服务时的隐私保护意识与安全防范意识，使得用户具有一定的独立辨别能力与批判意识，自主妥善保护个人信息与敏感数据，考量数据来源的可靠性与真实性，避免被虚假信息轻易误导。

5. 结语

作为一种框架性的底层技术，人工智能将从根本上变革国家以及国际社会的发展进程。人工智能在提升一国生产力、竞争力和国防军事实力的同时，也将为国家安全治理与社会的公共管理带来广泛深刻的结构性挑战，有效克服人工智能的负面效应是确保未来中国国家安全的关键。

对于中国而言，统筹规划人工智能技术发展与国家安全治理，把握科学技术领先性与自主性，做好相应的风险研判和预防措施是保障未来国家安全的重要命题。中国亟须从总体国家安全观视角出发，贯彻全面治理与系统治理思维，透彻理解人工智能技术不足与局限，搭建兼顾安全与发展的协同治理机制，在满足不同主体安全需求的同时提升风险治理效能，从秩序整合视角完善人工智能治理体系，助推国家治理体系和治理能力现代化水平不断提高。

参考文献

[1] Stanford Human-Centered Artificial Intelligence (2024) Artificial Intelligence Index Report 2023. https://aiindex.stanford.edu/report/

[2] 张华平, 李林翰, 李春锦. ChatGPT 中文性能测评与风险应对[J]. 数据分析与知识发现, 2023, 7(3): 16-25.

[3] March 20 ChatGPT Outage: Here’s What Happened (2023) https://openai.com/blog/march-20-chatgpt-outage

[4] 贾珍珍, 刘杨钺. 总体国家安全观视域下的算法安全与治理[J]. 理论与改革, 2021(2): 135-148+156.

[5] 浙江海事局. 浙江宁波“2.23” “东方盛”轮与“浙象渔 46102”轮碰撞事故调查报告[EB/OL]. https://www.zj.msa.gov.cn/ZJ/zwgk/gkml/xzqz/201912/t20191224_590231.html, 2024-04-06.

[6] 刘建利. 医疗人工智能临床应用的法律挑战及应对[J]. 东方法学, 2019(5): 133-139.

[7] Topel, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, **25**, 44-56. https://doi.org/10.1038/s41591-018-0300-7

[8] 毕文轩. 生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J]. 比较法研究, 2023(3): 155-172.

[9] 曾坚朋, 张双志, 张龙鹏. 中美人工智能政策体系的比较研究——基于政策主体、工具与目标的分析框架[J]. 电子政务, 2019(6): 13-22.

[10] 中共中央党史和文献研究院. 习近平关于总体国家安全观论述摘编[M]. 北京: 中央文献出版社, 2018: 90.

[11] 阙天舒, 张纪腾. 人工智能时代背景下的国家安全治理: 应用范式、风险识别与路径选择[J]. 国际安全研究, 2020, 38(1): 34.

[12] Algorithm Watch (2024) Vision Mission & Values. https://algorithmwatch.org/en/vision-mission-values/