人工智能时代安全保密信息技术管控

李大勇[1] 李欣桐[2] 张秋琳[1] 殷彬权[1] 付 刚[3]
1. 北京北方车辆集团有限公司 北京 100072
2. 北京政法职业学院 北京 102628
3. 北方置业集团有限公司 北京 100080

摘 要：随着人工智能技术的飞速发展，其在各领域的广泛应用带来了安全保密信息管控的诸多挑战。本文深入剖析人工智能时代安全保密信息在技术、管理、社会层面面临的核心挑战，如数据安全、算法安全、跨域数据流动合规困境等。详细探讨数据安全、算法安全、网络防御等关键管控技术，并从技术、管理、法律与伦理、国际合作维度提出针对性对策建议，为保障人工智能时代安全保密信息提供参考。

关键词：人工智能时代；安全保密；信息技术管控

引言：在当今数字化浪潮中，人工智能以其强大的创新力和变革力，重塑着社会生产生活的诸多方面。然而，人工智能在推动社会进步的同时，也给安全保密信息管控带来了前所未有的复杂局面。海量数据的汇聚、智能算法的演进以及网络攻击手段的升级，使得安全保密信息面临着诸多潜在风险。深入探究人工智能时代安全保密信息技术管控策略，已成为保障信息安全、维护社会稳定的关键课题。

1 人工智能时代安全保密信息的核心挑战

1.1 技术层面挑战

（1）数据安全：大数据聚合引发的隐私泄露风险。人工智能运转依赖海量数据投喂，多源数据聚合分析易突破个体信息碎片化限制，通过关联挖掘还原个人身份、行为轨迹等敏感信息，即便单条数据脱敏，聚合后仍可能造成隐私泄露，给信息保密带来基础层面的冲击。（2）算法安全：面临对抗样本攻击、模型逆向工程、后门植入等多重威胁。对抗样本可通过微小扰动误导AI决策，模型逆向工程能窃取算法核心逻辑，后门植入则可能让AI在特定指令下泄露机密，严重动摇AI应用的安全根基。（3）网络攻击升级：AI驱动的自动化攻击工具日益泛滥，如AI钓鱼能精准模拟可信身份生成定制化诱饵，智能勒索软件可自主探测系统漏洞、动态调整攻击策略，大幅提升攻击效率与成功率，传统防御体系难以应对。

1.2 管理层面挑战

（1）跨域数据流动的合规性困境。人工智能的跨场景应用需实现数据跨地域、跨行业流动，但不同区域、领域的隐私保护与数据保密规则存在差异，易出现合规冲突，增加信息管控难度。（2）供应链安全中的第三方风险。AI产业依赖多元供应链，开源模型、第三方数据服务等环节存在漏洞，部分开源模型可能隐藏安全隐患，第三方服务商的保密措施参差不齐，易成为信息泄露的薄弱环节。（3）人工智能伦理与法律监管的滞后性。AI技术迭代速度远超伦理规范与法律体系更新速度，现有规则难以覆盖深度合成、自主决策等新兴应用场景，导致安全保密责任界定模糊，监管存在真空地带[1]。

1.3 社会层面挑战

（1）深度伪造技术对舆论安全的冲击。AI驱动的深度伪造可生成逼真的虚假音视频、文本，易被用于传播虚假信息、操纵舆论，扰乱社会认知，甚至危害国家安全，且识别难度较大。（2）算法歧视与公平性争议。AI算法可能固化或放大社会偏见，在资源分配、风险评估等场景中产生歧视性结果，引发社会不公争议，同时也可能因算法不透明导致信息使用的合理性备受质疑。（3）公众对AI技术的信任危机。频发的AI安全事件与隐私泄露案例，降低了公众对AI技术的信任度，公众既担忧个人信息被AI滥用，也质疑AI决策的可靠性，不利于安全保密体系的协同构建。

2 人工智能安全保密信息管控的关键技术

2.1 数据安全技术

（1）联邦学习与隐私计算：作为数据“可用不可见”的核心实现路径，有效破解AI训练中数据聚合的隐私泄露难题。通过分布式训练架构，各参与方在本地保留原始数据，仅共享模型参数更新信息，无需暴露敏感数据本体，既满足AI对海量数据的需求，又实现数据隐私的全周期保护，为跨域数据协同训练提供安全支撑。（2）同态加密与差分隐私：精准平衡数据效用与隐私保护的核心技术。同态加密允许在加密状态下直接进行数

据运算，无需解密即可完成模型训练，从根本上杜绝数据运算过程中的泄露风险；差分隐私通过向数据中添加微小扰动，模糊个体信息边界，既能保证数据统计分析的有效性，又能防止个体信息被精准定位，适配多场景数据共享需求[2]。（3）区块链技术：构建去中心化的信任机制保障数据安全。依托去中心化存储、不可篡改、可追溯等特性，区块链可记录数据全生命周期操作轨迹，实现数据来源可查、去向可追、责任可究。同时，通过智能合约自动执行数据访问权限管控，减少中心化管理的漏洞，为敏感数据的存储与流转提供可信环境。

2.2 算法安全技术

（1）对抗训练与鲁棒性增强：针对性防御对抗样本攻击的核心技术。通过在模型训练过程中引入对抗样本，让模型在与“攻击样本”的对抗学习中提升泛化能力，增强对微小扰动的识别与抵御能力。同时，优化模型结构设计，采用正则化、数据增强等辅助手段，进一步提升模型的鲁棒性，降低对抗样本导致模型决策失误的风险，保障算法输出的可靠性。（2）模型水印与溯源技术：实现算法知识产权保护与责任追溯。在AI模型训练过程中嵌入特定水印信息，该水印可在模型输出结果中稳定提取，且不影响模型正常性能。通过水印信息能够精准界定模型的权属归属，有效防范模型盗版、篡改等侵权行为；同时，当模型出现安全问题时，可通过水印溯源定位问题源头，明确责任主体，为算法安全管控提供追溯支撑。（3）AI安全评估框架：建立算法透明性与可解释性标准。构建涵盖数据质量、算法逻辑、决策过程、输出结果的全维度评估体系，通过可视化技术、逻辑拆解等手段，将复杂的算法决策过程转化为可理解的信息。明确算法透明性的评估指标，要求模型对决策依据、数据依赖等关键信息进行清晰呈现，既便于检测算法潜在的安全漏洞与偏见，也为算法安全审查提供标准化依据[3]。

2.3 网络防御技术

（1）AI赋能的威胁情报分析：实现实时检测异常行为的智能防御技术。依托AI的深度学习能力，对网络流量、系统日志等多源数据进行实时分析，快速识别AI驱动的钓鱼攻击、自动化渗透等新型威胁，相比传统防御手段，大幅提升威胁检测的时效性与准确率，实现风险早发现、早处置。（2）自动化攻防演练系统：通过模拟AI攻击场景提升防御能力的重要载体。系统可自动生成多样化AI攻击脚本与场景，常态化开展攻防演练，帮助防御团队熟悉AI攻击逻辑与手段，针对性优化防御策略；同时，可沉淀攻防经验，推动防御技术的快速迭代升级，强化整体网络防御韧性。（3）零信任架构：基于身份的动态访问控制技术，重构网络安全防御边界。秉持“永不信任、始终验证”理念，打破传统网络的内外网边界划分，对每个访问请求进行身份认证、权限校验与环境评估，动态分配访问权限。有效防范内部人员越权访问与外部非法入侵，为AI系统及敏感数据构建全方位、精细化的安全防护体系。

3 人工智能时代安全保密信息技术管控的对策建议

3.1 技术维度

（1）推动AI安全技术标准化建设。以国际通用标准为参照，主动对接ISO/IEC等国际权威组织的AI安全标准体系，结合我国人工智能产业发展实际与安全保密需求，牵头或参与制定涵盖数据安全、算法安全、模型评估等关键领域的国家标准与行业规范。明确AI安全技术的核心指标、实现路径、测试方法与认证流程，统一技术应用的安全基准，解决不同企业、不同场景下AI安全技术应用不规范、不兼容的问题。同时，建立标准动态更新机制，紧跟AI技术迭代步伐，及时修订完善标准内容，确保标准的时效性与前瞻性，为AI安全保密信息管控提供坚实的技术标准支撑。（2）加强AI安全攻防技术研发与开源社区协作。加大科研投入力度，聚焦对抗样本防御、模型漏洞挖掘、隐私计算等关键核心技术领域，支持高校、科研院所与龙头企业开展产学研协同创新，突破一批具有自主知识产权的AI安全核心技术与装备，提升核心技术自主可控能力。积极搭建开源社区协作平台，鼓励企业、科研机构开放AI安全相关的技术成果、数据集与工具组件，促进技术交流与共享。通过开源社区汇聚全球创新力量，快速响应新型安全威胁，形成“研发-应用-反馈-迭代”的良性循环，推动AI安全攻防技术的快速升级，构建多元化、协同化的技术研发体系。

3.2 管理维度

（1）完善数据分类分级保护制度。立足AI时代数据多源化、复杂化的特点，进一步细化数据分类分级标准，明确不同级别数据的安全保护要求与管控措施。对涉及国家秘密、商业秘密、个人隐私的高敏感数据实施重点保护，严格规范数据的采集、存储、传输、使用、销毁全流程管理。建立数据分类分级动态调整机制，根据数据价值变化、应用场景拓展等情况及时优化分类分级结果。强化数据管控责任落实，明确数据处理各环节的责任主体，建立健全数据安全审计与监督检查机制，对违反数据分类分级保护要求的行为实施严厉追责，从管理层面筑牢数据安全防线。（2）建立AI产品全生命周期安全审查机制。将安全审查贯穿AI产品的设计、

研发、生产、部署、运维及淘汰的全生命周期。在设计阶段，要求企业开展安全风险评估，明确安全设计目标与防护方案；研发阶段，强化对数据来源合法性、算法安全性的审查；部署应用前，实施严格的安全测试与认证，未通过审查的产品不得投入使用；运维阶段，建立常态化安全监测与漏洞修复机制，及时处置安全隐患；淘汰阶段，规范数据销毁与设备处置流程，防止数据泄露。同时，明确审查主体、审查流程与审查标准，引入第三方专业审查机构，提升审查的客观性与权威性，确保AI产品全流程安全可控[4]。

3.3 法律与伦理维度

（1）制定《人工智能安全法》等专项立法。结合AI技术发展特点与安全保密需求，加快推进人工智能领域的专项立法进程，制定出台《人工智能安全法》等法律法规。明确AI安全保密的法律边界，界定各主体的权利与义务，规范AI技术的研发、应用与管理行为。重点规制深度伪造、算法歧视、数据滥用等违法行为，明确相应的法律责任与处罚标准，形成强有力的法律震慑。同时，做好专项立法与《网络安全法》《数据安全法》《个人信息保护法》等现有法律法规的衔接，构建系统完备、协调统一的AI安全法律体系，为安全保密信息管控提供坚实的法律保障。（2）构建AI伦理审查委员会与问责制度。组建由政府部门、高校、科研院所、行业协会、伦理专家等多方代表组成的AI伦理审查委员会，明确委员会的职责与权限，对AI技术研发与应用中的伦理问题开展审查评估。建立伦理审查标准与流程，要求重大AI项目在立项前必须经过伦理审查，重点评估项目是否符合公平、公正、透明、无害等伦理原则。同时，健全AI伦理问责制度，对因违反伦理要求导致安全保密事故的单位与个人，实施严格的问责追究。加强伦理宣传教育，提升全社会的AI伦理意识，引导企业与科研机构自觉遵守伦理规范，推动AI技术健康有序发展。

3.4 国际合作维度

（1）参与全球AI安全治理规则制定。主动融入全球AI安全治理体系，积极参与联合国《人工智能伦理建议书》等国际规则与标准的制定过程，充分发挥我国在AI产业领域的优势，提出符合我国国情与国际公平正义的治理方案与主张。加强与世界各国、国际组织的沟通协商，推动形成包容、开放、公正的全球AI安全治理共识。积极分享我国在AI安全保密领域的实践经验与技术成果，提升我国在全球AI安全治理中的话语权与影响力，为构建全球AI安全治理新格局贡献中国智慧与中国力量。（2）建立跨国数据安全流动与应急响应机制。针对AI时代数据跨国流动日益频繁的特点，加强与主要国家和地区的数据安全合作，协商建立跨国数据安全流动规则与互认机制，明确数据跨境传输的安全要求与责任划分，保障跨国数据流动的安全可控。建立跨国AI安全应急响应机制，与相关国家和地区建立信息共享、协同处置的合作渠道，及时通报新型AI安全威胁与漏洞信息，联合开展应急处置演练，提升应对跨国AI安全事件的协同能力。加强跨国执法合作，严厉打击跨国AI相关的网络攻击、数据窃取等违法犯罪行为，维护全球AI安全生态。

结束语

人工智能时代安全保密信息技术管控是一项长期且艰巨的任务。面对技术、管理、社会等多层面的挑战，我们需综合运用数据安全、算法安全、网络防御等关键技术，从技术标准、管理机制、法律伦理、国际合作等多维度发力。只有各方协同共进，不断完善管控体系，提升应对能力，才能在充分发挥人工智能优势的同时，筑牢安全保密防线，为人工智能健康有序发展与社会稳定提供坚实保障。

参考文献

[1]王艺洁.“互联网+”背景下的计算机网络信息安全防护研究[J].家电维修,2025,(05):77-79.

[2]郭金龙.网络信息安全防护体系中信息管理技术的应用研究[J].中国管理信息化,2025,28(08):156-158.

[3]丘业.信息化背景下计算机网络信息安全防护策略分析[J].信息与电脑,2025,37(06):90-92.

[4]王红岩.大数据技术的计算机网络信息安全防护对策[J].中国宽带,2025,21(03):52-54.