中国移动
China Mobile

AI+智慧城市
安全解决方案白皮书

中国移动通信集团有限公司

2024年9月

目录

发布单位：中国移动通信集团有限公司

编制单位：中国移动通信集团有限公司信息安全管理与运行中心

中移雄安信息通信科技有限公司

中移（上海）信息通信科技有限公司

《中国信息安全》杂志社

北京启明星辰信息安全技术有限公司

卓望数码技术（深圳）有限公司

成都思维世纪科技有限公司

上海嘉韦思信息技术有限公司

参编人员：王昀、袁捷、冯国华、张峰、江为强、曹雪峰、黄静、孙海涛、李子晔、王光涛、路骁虎、安宝宇、邱勤、董航、袁胜、位华、姚飞、闫少维、于朝翔、郭中元、谷田田、赵威、唐双林、周涛、李刚、刘乐、陈文博、黄琴、章明珠、舒首衡、钟志成、刘超、李卓、何升文、钟立、郗上才、尹咸阳、李峻天、周天翔、曾礼、李照雷、赵广义、光国庆

1. 前言

智慧城市是推动城市治理体系和治理能力现代化建设的重要抓手。“十四五”期间国家对数字经济和智慧城市发展进行了专项规划，智慧城市作为数字经济的重要应用场景，其数据巨大价值和重要意义得到强调和凸显。

当前智慧城市在人工智能技术领域处于起步阶段，在高速发展的同时，智慧城市面临着来自网络、数据、应用层面的安全风险，以及智慧城市人工智能本身引入的安全新风险。智慧城市人工智能安全关系到个人、组织、社会公共利益甚至国家利益,不能疏忽。面对 AI 人工智能的安全形势,随着 2023 年国家网信办《生成式人工智能服务管理暂行办法》的发布，人工智能安全工作具有了可以执行的管理办法，意味着智慧城市中相关技术、业务有了系统化、持续化的发展环境。

为了帮助智慧城市运营管理者如各级政府或组织应对以上众多人工智能安全方面的难题和挑战，实现数字经济的长期健康发展。本白皮书梳理智慧城市 AI 人工智能安全需求，提出 AI+智慧城市安全体系架构，通过智慧城市 AI 风险防范和 AI 赋能智慧城市安全两大方案解决城市各层面风险,并以黑龙江省级、中山智慧城市等综合系统举例，阐述其人工智能安全方案落地具体场景。展望未来城市发展，白皮书提出了智慧城市人工智能安全发展建议，期望为进一步推广、普及和完善 AI+智慧城市的理念、方法、体系与应用贡献力量。

2. AI+智慧城市安全背景概述

2.1. AI+智慧城市发展现状

国家“十四五”规划明确指出：“加强数字社会、数字政府建设，提升城市服务、社会治理等数字化智能化水平。”在此背景下，各地政府积极推进智慧城市建设，将其列为重要的发展战略，加大投入力度，政府的支持和引导对于智慧城市建设的发展起到了关键的作用。当前，智慧城市建设正处于快速发展阶段，依赖于信息技术和通信技术的支持，人工智能、大数据、云计算等新兴技术的应用为智慧城市提供了更多可能性。根据 IDC 今年发布的《中国智慧城市市场预测，2023-2027》数据，2023 年，中国智慧城市 ICT 市场投资规模超过 8700 亿元人民币，预计到 2027 年将超过 1.1 万亿元。

2024 年《政府工作报告》提出，加快发展新质生产力，开展“人工智能+”行动。以大模型、大算力、大数据为基础的人工智能正逐步成为城市治理数字化、科学化、先进化的新动能和新范式。不仅为城市管理带来了高效、智能的解决方案，同时也给居民的生活带来了更好的体验。当前各级政府在国家政策指引下积极探索城市治理新模式，将 AI 能力融入智慧城市建设中。IOC、视频云人体检测、火情监测、智慧运维、智能问答机器人、智慧交通等智慧城市应用场景建设工作不断完善，促进城市治理、民生服务等服务场景智能化提升。

图 1 中国智慧城市市场预测

通信行业积极支撑国家推动 AI+智慧城市战略，中国移动作为全球领先的通信和信息服务提供商，其在人工智能与智慧城市的发展战略方面有着清晰的规划和目标。

2023 年，中国移动发布《中国移动新型智慧城市白皮书》，白皮书中重点强调了通过人工智能、大数据等技术，打造新型智慧城市。中国移动贯彻党和国家数字发展战略，围绕网络强国、数字中国、智慧社会部署，不断践行“力量大厦”发展部署，明确了智慧城市作为我国城镇化发展和实现城市可持续化发展方案的战略地位，公司着力打造智慧中台开放模式，将能力汇聚，在智慧城市中基于业务场景持续打磨基础通信、人工智能、区块链、安全认证等优质技术能力。

2024 年《中国移动人工智能安全白皮书》指出，中国移动始终把人工智能作为公司战略发展方向，发挥运营商特色优势，打造全面的人工智能产品体系，充分发挥中国移动在网、算、运营等方面的资源和能力优势，形成涵盖智算力、高质量数据集、人工智能平台、各领域算法能力、大模型、智能

化应用的全栈新型智能化服务能力，达业界先进水平。中国移动将加快人工智能融入智慧城市中，推进“AI+”产业发展，培育新质生产力。

在《国务院关于进一步优化政务服务提升行政效能推动“高效办成一件事”的指导意见》中，明确要求探索应用自然语言大模型等技术，提升线上智能客服的意图识别和精准回答能力，优化智能问答、智能搜索、智能导办等服务，更好引导企业和群众高效便利办事。2022 年以来，中国移动基于智慧城市领域的丰富建设经验，以通用大模型为基础，融合政务领域数据开展精调，引入政务领域约束模型对输出进行限制，打造了面向政务领域的行业大模型—九天·海算政务大模型。以政务大模型为技术底座，构建“平台+算法+应用”体系，实现了对黑龙江省政务工作的整体提升和跨越式发展。聚焦一网通办、一网统管、一网协同三个典型应用场景，打造智能客服、龙政智搜、数字人、公文写作等具体应用，提供政务领域的智能处理优化、智能内容生成等服务，助力黑龙江省政府更加高效地履行职责，切实提升群众的满意度和信任度。

图 2 九天·海算政务大模型应用实践

2.2. AI+智慧城市安全发展现状

新型智慧城市是推动城市治理体系和治理能力现代化、提升城市居民幸福感和满意度的新理念和新路径，也是网络强国建设和数字经济发展的重要

载体。随着 AI 技术的不断发展和在智慧城市智领域广泛的应用，人们享受技术红利的同时，也面临着日益突出、复杂多变的网络安全风险，规模性隐私数据泄露、关键信息基础设施遭受网络攻击等安全事件屡见不鲜。

当前，AI+智慧城市安全发展有如下两点现状，一方面是智慧城市引入大量 AI 技术，各项便民技术正在深度改造城市的运营和管理方式。然而，伴随着便利性和效率的提升，城市管理者大多没有关注 AI 自身的安全性：如何处理数据安全、隐私保护以及算法透明度等问题，正成为当前我们面临的重要挑战。另一方面从智慧城市的网络攻击防护视角来看，面对新场景、新特征、新需求，深层漏洞和未知威胁越来越多，尤其是网络安全环境更加趋于复杂化、交织化和自动化，需要结合 AI 加强智能化监测预警和主动化安全防护的联动，阵线前移和协同联动的主动防御显得尤为重要。

当今国内外安全大小模型迅猛发展，国际市场出现包括思科安全人工智能助手、Elastic AI Assistant for Security、Google Cloud Security AI Workbench、Microsoft Security Copilot 等产品，国内厂商包括 360、安恒、金睛云华、华为、绿盟、奇安信、启明星辰、深信服、天融信等均也提出了安全大模型或小模型。然而各家厂商几乎没有针对智慧城市安全特点的训练集，用 AI 赋能智慧城市安全尚处于起步阶段。

360 安全集团在 2024 年一季度末发布安全大模型 3.0，包括语言、规划、判别、道德和记忆五个中枢，据悉采用 360 安全大模型能实现 MTTR 缩短一半、人均工作效率 30%的提升。

华为在 HSA2024 会议上发布 L4 级 AI 安全智能体（网络安全高度自主防御）。华为具备全栈大模型能力，包括从 AI 芯片（昇腾系列），CANN 计

算架构，MindSpore 深度学习框架，MindStudio 开发工具到智算训练网络以及盘古通用大模型，目前实现自动处置 90%以上的安全事件，极大降低人工工作量，提升网络安全运维效率。

中国移动与启明星辰在 2024MWC 会议上联合发布九天•泰合安全大模型，该模型依托于中国移动九天基座大模型的强大计算能力和广泛数据处理优势，深度融合启明星辰安全行业特有的海量数据资源，包括但不限于威胁情报、漏洞数据库、专业安全知识及最新安全研究成果等。针对不同的智慧城市应用场景，再训练出如医疗类数据识别、政务数据脱敏等多个小模型。

2.2.1.智慧城市 AI 技术风险防范紧迫性

随着人工智能（AI）技术在智慧城市中的广泛应用，不仅城市运营形态发生了巨大变革，人们的生活方式也因此变得更加便利。但同时，也正由于智慧城市资源的独特性，复杂且不可预测的 AI 技术可能带来的安全风险的破坏力也是有独特性的。

AI 在智慧城市服务中的决策一旦出现问题，可能带来极坏的社会影响，影响智慧城市的公信力，甚至可能波及各行业，导致社会安全、法律责任、道德困境或舆论的负面影响出现。如在园区治安管理、灾情应急管理中，一旦出现问题，将出现严重安全事故。

智慧城市的 AI 技术风险防范势在必行，我们需要制定有效的策略和措施，既确保人工智能发挥积极作用，又最小化其潜在风险。建立完善的数据安全机制和隐私保护政策，确保个人数据得到合理使用和保护。同时，加强对 AI 算法的监管和审查，推动算法透明化和公正性。

智慧城市的发展坚持科技创新与风险管理相结合，促进人工智能技术的

可持续发展，提升城市治理和服务水平。只有在充分认识和有效应对 AI 技术风险的基础上，智慧城市才能实现可持续、智慧和安全的发展。

2.2.2.AI 赋能智慧城市安全保障必要性

在智慧城市的环境中，传统网络安全领域的防御技术如基于规则和认证的机制，例如防火墙、入侵检测系统和入侵防御系统等，虽然起着重要作用，但面对新型威胁和变种攻击时，这些方法可能显得力不从心。智慧城市因其庞大的资源暴露面、复杂的业务逻辑、多样的数据交易形式，使得安全管理面临极大挑战。

随着网络技术和数据价值的迅速发展，网络安全威胁呈现出智能化、隐匿性和规模化的趋势，为智慧城市的安全防御带来了更大的挑战。其中，人工智能赋能的安全攻击已成为一种常见现象，包括但不限于漏洞自动挖掘、恶意软件智能生成、智能网络破坏等形式。

考虑到智慧城市的特性，采用 AI 技术助力网络安全管理以应对新形势下的安全挑战是必要的。通过 AI 分析数据流，可以识别那些难以通过传统方法检测出的复杂威胁，从而提高威胁检测的准确性和效率。利用训练好的模型识别恶意代码特征，可以快速准确地检测异常行为和恶意软件，从源头上抑制安全威胁的滋生。此外，生成实时的威胁报告并进行趋势分析，有助于预测潜在的攻击行为，提高网络安全的响应速度和效果。AI 技术的应用不仅提高了安全防御的智能化和自适应性，还加强了快速、高效应对安全威胁的能力，有助于建立更为安全稳定的智慧城市网络环境。通过持续创新和优化，结合人工智能技术的智能防御手段，智慧城市可以更好地保护网络安全、促进信息共享和智能应用，实现城市数字化转型的顺利推进。

3. AI+智慧城市安全风险与需求

3.1. AI 技术在智慧城市应用风险

目前，智慧城市提供多类服务能力，整合多种人工智能技术研发和应用场景，随着人工智能等新技术发展日新月异，其伴生的安全风险日趋严峻。主要风险涉及智慧城市模型算法、智慧城市数据要素、智慧城市 AI 服务、智慧城市 AI 平台和智慧城市 AI 运营等多个方面。

3.1.1.智慧城市模型算法风险

（1）城市模型算法偏见和歧视风险

在智慧城市的建设中，AI 模型算法偏见和歧视风险是一个不可忽视的问题。由于训练数据可能存在隐性的偏差，使得模型在预测或决策时可能会无意中倾向于某些群体或者反对某些群体。比如，在公共服务的分配、交通管理或者治安管控等方面，如果 AI 模型受到了含有偏见的数据的影响，可能会对某些群体造成不公或歧视。因此，在使用 AI 技术时，我们需要考虑如何避免和纠正这些潜在的偏见和歧视，以实现真正公平公正的智慧城市。

（2）城市模型算法伦理和法律风险

在智慧城市的实施中，AI 模型将带来了伦理和法律风险。其中包括 AI 技术对个人信息的处理可能涉及隐私保护，如果不当使用，可能引发法律纠纷。同时 AI 决策过程中可能存在的偏见和歧视问题也触及伦理道德层面，比如其可能导致某些群体在享受公共服务时受到不公平对待。因此，在智慧城市中使用 AI 技术，不仅需要严格遵守相关法律法规，还需要考虑其可能产生的伦理影响，以实现公平、公正和透明。

智慧城市AI模型同样存在滥用用户偏好数据,导致城市服务存在信息差,用户被“大数据杀熟”的伦理风险。

（3）城市模型算法不可解释性风险

智慧城市中的AI模型算法普遍面临着不可解释的风险。这意味着，尽管AI模型能够生成有效的预测或决策，但其内部的复杂计算过程往往是难以理解和解释的。这种“黑箱”特性在一定程度上增加了使用AI模型的风险，因为它使得错误或者偏见难以被检测和纠正。例如，如果一个用于政务服务分配的AI系统做出了错误的决策，而我们无法理解决策背后的逻辑，那么就难以找到问题的根源并进行改进。

由于智慧城市的大模型代表的是城市服务的权威，它的决策和输出直接影响到公众利益。如果大模型出现误导，需要对所产生的逻辑进行溯源解释。因此，对于数字政府而言，解决大模型的不可解释性问题，确保其决策过程的透明和公正，不仅是技术层面的需求，更是社会责任与公众期待的体现。

（4）城市模型算法被逆向篡改风险

智慧城市面向社会提供大量公共接口和服务，一些恶意攻击者可能通过分析和理解AI模型的运作方式，找到其中的漏洞，并获得运行内部逻辑，达到逆向攻击的效果。这种对抗攻击能够破坏AI模型的正常功能，在智慧交通或治安监控等场景下产生误判，并可能获得AI模型内部逻辑进而对模型算法逆向攻击。

同样的，攻击者可以通过智慧城市面向公共的服务接口作为入口，通过公共组件的漏洞利用或人员意识淡薄，违规获取已部署的人工智能模型算法的详细信息，包括参数、结构、功能等，从而导致知识产权被侵犯或商业机

密泄露等风险。

3.1.2.智慧城市数据要素风险

（1）智慧城市数据非法采集风险

在智慧城市的运行中各类 AI 模型需要采集、训练大量城市数据。由于智慧城市依赖大量的数据驱动，包括公共服务、交通管理、能源规划等各方面，这些数据中往往涉及到国家政务、城市管理数据。如果没有得到合法授权，可能构成违法犯罪的行为，严重的甚至影响国家安全。

智慧城市也需要大量公民个人信息，随着《个人信息保护法》的出台，对于数据采集和处理提出了更高的要求。这意味着，在收集、使用个人信息时，应明确告知并取得用户的同意，并限制其使用范围。任何未经授权擅自获取或使用个人信息的行为都将受到法律的严惩。因此，所有在智慧城市运营中涉及到的 AI 模型，必须严格遵守相关法律法规，确保所有采集和处理的数据都是在合法、合规的情况下进行的，以保证国家安全和公民的个人信息权益不受侵犯。

（2）智慧城市训练数据异常风险

智慧城市的 AI 模型依赖于大量训练数据，如果这些训练数据存在异常，可能会导致模型的预测结果出现偏差。异常情况包括数据含有噪声、错误标签、训练数据成分组成不合理或者是与大部分数据不一致的离群点。这些异常数据可能会影响模型的学习效果，导致在实际应用中，如交通管理、政务服务等方面产生误判。

（3）智慧城市数据投毒污染风险

在智慧城市 AI 模型中大量训练数据来自社会公开数据。 这种情况下，

恶意攻击者可能会注入有害的或误导的信息到数据集中，对模型的训练过程进行干扰。例如，攻击者可能会修改大众数据如水电使用数据，或者添加具有错误标签的数据，使得 AI 模型在学习时产生偏见或误解。这样的污染数据可能会严重影响模型的性能和准确性。

（4）智慧城市训练数据泄露风险

一方面智慧城市中存在着大量的敏感数据，包括居民的个人信息、地理位置数据、交通流量数据等。一旦这些数据被非法获取或误用，可能对个人隐私造成严重侵犯，甚至引发国家安全数据泄露等犯罪行为。另一方面，智慧城市存在大量对外服务接口和系统，多类系统存在异构复杂性，极大增加了数据暴露风险。因此 AI 训练模型在使用敏感数据时，应重点关注数据泄露风险。

3.1.3.智慧城市 AI 服务风险

（1）城市服务输出内容偏差风险

智慧城市的 AI 服务依赖于 AI 模型的准确预测和决策，但如果 AI 模型的输出内容出现偏差，可能会影响到整个城市运行的效率和公平性。例如智能问答机器人，如果其输出内容受到数据偏见或错误学习的影响，如在政务场景回答偏差，将导致公信力下降，社会服务质量降低。

（2）城市服务生成内容违规风险

智慧城市常使用 AI 内容生成服务为社会提供服务，AI 生成或合成的内容由于其具有一定的随机性和不可控性，可能导致违规内容、歧视偏见、隐私泄漏、内容侵权等众多问题的出现，这对公众的生命财产安全、国家安全、意识形态安全及伦理安全构成威胁。特别是在与大型模型的交互提问场景中，

用户输入的提示词也可能存在包括涉政、涉黄、涉恐、涉暴、涉赌、涉毒、诱导犯罪和恶意代码等各类违法违规内容风险。若模型的内容安全防护机制不完善，可能会导致模型产生有害、不适当或违反公序良俗的输出内容。

AI 生成内容异常可能使公众信息发布系统误传播不真实或者有害信息，造成社会不安。同时若智慧城市 AI 生成内容有违法违规内容，也将对城市运营者造成影响，造成不良社会舆论。

（3）城市服务异常调用安全风险

智慧城市 AI 服务异常调用安全风险主要涉及服务调用失败、非法访问和频繁请求等问题。当前社会出现利用 AI 技术对系统进行攻击的行为，如果通过 AI 批量生成能力对智慧城市服务大规模调取，或通过 AI 模式识别能力对身份认证绕过后进行非法访问，将对城市治安稳定带来较大风险。

传统的网络安全攻击对智慧城市 AI 服务仍然有效，包括 DDoS 攻击、CC 攻击、漏洞扫描、木马植入等。当网络攻击大面积影响城市服务时，将导致业务系统宕机无法正常对外提供服务，严重的将导致城市交通、公共安防、卫生应急等业务停摆。

3.1.4.智慧城市 AI 平台风险

（1）智慧城市算力恶意消耗风险

智慧城市AI算力恶意消耗风险主要来自外部的恶意攻击者和内部的不适当使用。外部攻击者可能通过发起大量无意义的请求，导致 AI 服务器资源被大量消耗，进而影响正常服务提供。内部不适当使用，如无节制的资源消耗，不合理的任务分配，也可能使得算力资源在无价值的任务上被浪费，影响关键任务的处理效率。

（2）智慧城市供应链安全性风险

智慧城市 AI 技术平台常依赖国内外各类系统组件开发完成，在 AI 技术研发、生产、分发和维护过程中可能出现安全威胁。因上游芯片、算力平台目前仍以境外技术为主、国内技术为辅，智慧城市 AI 技术目前可能因国际形势、技术封锁等原因带来安全风险，最终影响社会公众安全。

3.1.5.智慧城市 AI 运营风险

（1）法律遵从风险

智慧城市在数据收集、存储、处理和传输过程中，必须严格遵守相关法律法规，如《网络安全法》、《数据安全法》等，确保数据的合法性和安全性。然而，随着 AI 技术的快速发展，其在智慧城市领域的应用越来越广泛，对现有的法律法规提出了严峻挑战，如何界定 AI 技术在智慧城市建设中的法律责任，成为亟待解决的问题。2023 年 7 月，国家网信办联合七部门发布并施行《生成式人工智能服务管理暂行办法》进一步规范 AI 技术的发展。但要想更好地发挥 AI 技术作用，规避潜在风险，还应以更加科学直接的法律方式进行治理。

（2）运营管理风险

智慧城市的建设涉及多个部门、领域和层级的数据共享和协作，同时涉及大数据、云计算、物联网等众多技术栈，传统的运营模式容易产生“烟囱运维”，对运维人员技术能力要求高，伴随 AI 技术的广泛应用，运营方面的风险愈加凸显。一是智慧城市在数据的采集、传输、存储和处理过程，包含大量用户的个人隐私数据，运用 AI 技术对数据的调用和处理存在数据泄露的风险。二是 AI 技术的应用依赖于大量的算法，如果算法出现错误或失控，可

能会造成严重后果。三是我国智慧城市依旧存在安全运营模式不清晰、重建设轻运营等问题，严重影响智慧城市安全高效运转。

智慧城市安全运营团队同样存在相应问题。日益严重的网络安全威胁正在对已经承受巨大压力的安全从业人员施加额外的负担。当我们将解决 AI 引入网络安全问题时，需要更多专精于 AI 和 ML 网络安全的专家。专业技术人员根据项目需求进行维护和调整，将极大地提升智能网络安全技术的实际效益。但是目前全球范围内合格且训练良好的此类专业人员的数量，远远无法满足当前的需求。

（3）评价模糊风险

智慧城市在建设过程当中，智慧城市的运行评估是非常重要的一个管理环节，它有利于我们智慧城市建设和发展的持续改进，同时也可以推进我们智慧城市的有效建设，把握智慧城市建设的基本方向。但是，伴随 AI 技术在智慧城市广泛的应用，传统的智慧城市评价体系不足以支撑新技术应用下的新需求。

3.2. 智慧城市安全技术问题及 AI 赋能需求

3.2.1.智慧城市网络安全需求

（1）智慧城市智能化网络安全需求

对于基础的网络安全设备而言，在人工智能的技术发展的背景下，已经无法应对下一阶段复杂的网络环境。智慧城市的网络环境具有传感设备数据量大、系统种类多、数据流动性强等特点，然而目前尚未形成安全协同。随着智慧城市中接入的设备和系统不断增多，其攻击面也随之扩大。每个连接

点都可能成为潜在的入侵途径。此外，智慧城市中涉及的关键基础设施如电力、交通、供水、通信等系统存在被网络攻击的风险，这些系统通常同时运行、相互交互，使得传统的封闭式网络安全防护手段和基于特征与规则的威胁检测技术难以覆盖整个智慧城市的复杂网络环境。

（2）智慧城市智能化决策需求

大模型是人工智能领域广泛应用的技术，当前已衍生出多种不同领域的大模型，其中，安全领域相关的安全大模型是智慧城市中一项关键的技术。传统的网络安全运维方式，需要运维人员具备良好的技术能力，才可以做出及时正确的决策。在智慧城市中，网络环境复杂，设备众多，安全运维难度高，需要能够支撑问答的智能助手进行辅助，智能问答模型，能够回答安全管理人员的各种查询，提供数据支持和决策建议，帮助他们做出明智的安全决策，提升整体安全管理水平。通过满足以上需求，智能问答模型能够显著提升智慧城市人工智能安全领域的运营效率和防护能力。

（3）智慧城市自动化自检需求

在智慧城市网络环境中，除了对网络环境的分析预测，以及实时策略更新外，各个系统的安全自检也是重要的一环。常规的安全测试如渗透测试，高度依赖技术人员的人为判断与操作，亟待利用人工智能技术辅助测试人员，提高效率和威胁识别率。人工智能的介入尤其在多系统的复杂情况下，优势明显。

3.2.2.智慧城市数据安全需求

（1）智慧城市数据定向处理需求

在智慧城市中，会产生和收集的大量数据，包括个人信息、公共服务数

据、交通数据等，这些数据的不当使用可能带来严重后果，如个人隐私泄露、国家和社会敏感数据暴露，这些数据被恶意利用进行广告骚扰、电信欺诈、社会工程学攻击或商业情报窃取等违法行为。同时，数据存储和使用不当也会引发一系列风险。目前，智慧城市数据在数据使用和访问过程中，缺乏必要的数据保护手段，使用如数据水印、数据脱敏等技术，对智慧城市数据进行标注和有限范围内脱敏，让数据可追溯，防止使用人工智能时产生的数据投毒和数据泄露风险。

（2）智慧城市数据分类分级需求

人工智能技术支持下的智慧城市，数据是其重要的资产，但是庞杂的数据，往往致使重要或敏感数据缺乏梳理整理，难以针对海量数据实施有效的分类分级保护措施。使用智能化手段对海量数据进行分类分级，不仅能提升效率，减少人工辅助过程，而且自动化的流程，可以让数据整理不再繁琐。多种智能分类器，在智慧城市的各个行业中进行自动化标签处理，为人工智能提供良好的数据基础。

（3）智慧城市数据合规审计需求

智慧城市数据审计合规需求是一个综合性的要求，旨在确保智慧城市建设中数据的安全性、合规性和有效性，遵守国家及地方关于数据保护、隐私保护等相关法律法规要求，应建立健全的数据管理政策和流程，明确数据访问和使用的权限和规范，加强对数据的监控和审计。智能化的审计系统和工具，可以让关键数据方便审查，通过强化数据审计方面的管理和措施落实，确保智慧城市建设中数据的安全、合规和有效使用。

3.2.3.智慧城市应用安全需求

（1）智慧城市内容安全需求

智慧城市中，应用涵盖了交通管理、能源管理、公共安全、卫生健康、环境监测等多个方面，极大提高了城市运行效率和居民生活质量。然而，智慧城市软件内容出现异常，将导致城市服务效率降低，严重者将影响社会舆论和智慧城市公信力。

（2）智慧城市应用监测需求

上层应用和服务支撑着智慧城市运转，提供多种功能性的接入点，方便城市服务提供者和使用者更快捷的访问，同样，API 一类的接入点会给攻击者提供攻击渠道。人工智能将为应用监测提供更强的技术，实现城市智能化管理。覆盖到所有的 API，动态监测 API 的越权攻击、隐私泄露、拒绝服务攻击等多种行为，并且对交互内容进行风险识别，实时发出告警，保护服务和应用的安全，持续提供可靠的服务。

（3）智慧城市拟态分析需求

随着人工智能技术的发展，更多的类人模型、应用也随之诞生，在给用户提供便利的同时，也给了恶意攻击者新的攻击手段，使其可以规避常规防范机器行为的防护措施。对抗此类攻击，同样需要人工智能的技术支持，通过机器学习，对恶意用户和实体的行为进行分析，减少人工识别的错误率，同时，诸如拟态蜜罐技术，通过模拟真实服务和伪装来欺骗攻击者，从而达到保护真实系统的目的。此类人工智能安全技术，可以对抗更高层次的攻击，提升智慧城市总体的防御能力。

3.2.4.智慧城市公共安全需求

公共安全事关国家安危，社会稳定。随着我国城市化进程加快，城市人口增加、功能多元化、规模不断扩大，城市运行系统日益复杂，安全风险不断增大，传统的城市公共安全管理已经难以适应时代发展的要求，不能有效应对新的挑战。AI 技术的不断发展，在长期追踪、智能分析、趋势预判和城市精准化管理方面的优势，可以帮助提升公共安全风险态势感知、预测预警、动态管控等方面的能力。此外，在高精度识别、实时性处理、交通安全监测等领域，也亟需 AI 技术的应用，助力提高城市治理能力现代化。

4. 中国移动 AI+智慧城市安全体系架构

在国家标准智慧城市体系架构和《中国移动人工智能安全白皮书》的指导下，中国移动 AI+智慧城市安全体系架构，针对智慧城市 AI 风险防范和 AI 赋能智慧城市两部分内容进行设计。

图 3　AI+智慧城市安全解决方案体系框架

4.1. 总体目标

AI+智慧城市安全集中体现在智慧城市 AI 风险防范、AI 赋能智慧城市安全两个方面。首先让 AI 模型合法合规、算法公平公正、数据安全可信、计算平台可管可控。其次加强 AI 对智慧城市安全的赋能，将智能化手段运用到智慧城市的安全防护工作中，提高网络安全水平。最终总体实现“让智慧城市更安全，让城市安全更智慧”的目标。

4.2. 智慧城市 AI 风险防范

智慧城市 AI 风险防范针对人工智能技术应用与平台存在的安全风险，综合考虑模型算法、数据要素、业务服务、平台能力、运营合规等五个方面进行全面防范。通过对模型算法的审查优化、数据隐私保护、业务服务监控、平台安全强化和合规管理，确保智慧城市中的 AI 能力在可管可控的前提下安全可信地运行。这样的风险综合防范措施不仅提升了智慧城市的安全性，也保障了 AI 技术的稳定、可靠和可持续发展，促进智慧城市建设朝着更加安全和可信赖的方向不断前行。

4.3. AI 赋能智慧城市安全

AI 赋能智慧城市安全是开展人工智能核心技术研究在智慧城市各个场景中的应用，利用如安全大模型等核心技术，提升基础网络安全、数据安全治理、内容安全治理、业务应用安全等防护水平。

AI 赋能智慧城市安全是通过开展人工智能核心技术研究，并将其应用于智慧城市各个场景中，以提升城市安全保障水平。利用诸如安全大模型等核心技术，可以加强基础网络安全、数据安全治理、内容安全治理以及业务应用安全等多个方面的防护措施。

在基础网络安全方面，AI 技术可以实现实时流量监测和入侵检测，识别和阻止潜在威胁，加固网络边界防护。在数据安全治理方面，AI 可用于数据加密、访问控制和安全审计，确保数据传输和存储的安全性。针对内容安全治理，AI 可以检测恶意软件、过滤有害内容，保障网络环境清洁和健康。而在业务应用安全方面，AI 技术可以通过身份认证、访问控制等手段，减少安

全漏洞和数据泄露风险，确保业务系统稳定运行。

通过 AI 技术的全面运用，智慧城市可以提高安全防护水平，减少网络风险和安全事件的发生，保障居民和城市信息的安全。AI 赋能的智慧城市安全体系不仅提高了安全防范的智能化和自适应性，也为城市安全管理提供了更为高效和可靠的解决方案，推动智慧城市建设向着更加安全、智能和可持续的发展方向迈进。

4.4. 智慧城市安全基本原则

智慧城市的安全将遵循下面 4 个原则：

（1）统一领导、分级管理：智慧城市安全中遵照“谁主管，谁负责；谁运营，谁负责；谁接入，谁负责”的原则，将责任分工进行明确，落实人工智能安全的主体责任。

（2）安全三同步：按照工业和信息化部的关于安全三同步的制度要求，在智慧城市人工智能安全建设和运行过程中，应符合同步规划、同步建设、同步运行三项原则。

（3）坚持“1264 人工智能安全的规划”：中国移动明确了人工智能安全领域的发展原则，即规划一个工作体系框架，着力两个工作发力方向，落实六大 AI 安全风险防护措施，明确四大类 AI 赋能网络安全工作。后续的 AI+智慧城市安全发展也将按照这个原则，构建多层次、全方位的安全保障体系。

（4）坚持协同合作、推广技术应用：保持开放合作的姿态，积极参与行业 AI 安全标准的研究，广泛开展合作与资源共享，共同面对 AI+战略转型过程中的新挑战，共同突破 AI+智慧城市安全的新的高度。

5. 智慧城市 AI 风险防范方案

5.1. 智慧城市 AI 模型算法安全

5.1.1.维护城市模型公平透明

在设计模型算法阶段，需要重点关注选择特征和参数，并在开发阶段构建了包含各种场景和来源的测试数据集，以充分测试和验证模型算法，确保对所有对象的决策结果具有一致性。需要重点评估模型的公平性，例如进行基于静态测试数据集的公平性评测。

为遵守法律法规关于算法透明度的相关规定，需要建立可以透明地进行监管的算法管理体系，并按照要求公开算法的机制原理。智慧城市可以采用视觉和语言辅助解释、策略模仿、可解释模型、逻辑关系提取和策略分解等技术手段来解释模型的推理过程，这样可以增强模型的透明度。

5.1.2.提升城市模型可解释性

提升智慧城市 AI 模型算法的可解释性，需采用清晰和透明的算法，记录决策过程和参数设置，提供可视化展示结果和解释机制，建立模型文档和说明书，定期审查和更新算法，同时加强对使用者和相关人员的培训，使他们能理解和解释模型运行结果，确保决策过程可追溯和解释，以维护模型的合理性和公正性。采取视觉和语言辅助解释、策略模仿、可解释模型、逻辑关系提取、策略分解等技术手段解释模型的推理过程，以增强模型的透明度。

5.1.3.保证城市模型合法合规

保证解决智慧城市 AI 模型算法伦理和法律风险，需明确城市各类数据隐

私保护政策，遵守数据合规性要求，建立透明的算法决策流程，进行风险评估和监控。依据国家《新一代人工智能伦理规范》、《科技伦理审查办法（试行）》等要求开展定期的伦理审查检查，强化科技伦理风险防控，促进负责任创新，有效预防潜在风险。增加面向智慧城市用户建立服务使用反馈沟通渠道，用户可通过该渠道反馈使用人工智能系统过程中遇到的问题。

智慧城市管理者应通过国家互联网信息办公室“互联网信息服务算法备案系统”对服务使用的模型和服务形式进行备案。保证算法模型有效监督和透明使用。

5.1.4.增加城市模型加密混淆

为确保智慧城市模型不可逆向，可采用模型加密混淆技术，通过对算法和数据进行加密、混淆处理，保护模型不被恶意破解。同时对模型输出内容进行脱敏处理，去标识化处理可以减少数据关联性，确保输出结果匿名化，以防止模型被逆向推导。

5.2. 智慧城市 AI 数据要素安全

5.2.1.智慧城市数据采集安全

为保证智慧城市模型数据采集安全，要对数据源进行可信验证，确保数据来源真实可靠。确保数据源内容中不包含违法不良信息、偏见歧视、商业秘密等内容，并标记数据来源可追溯或获取开源许可协议。同时，明确合规的数据采集、使用和存储规则，对敏感信息进行脱敏处理，保证个人隐私不被侵犯。根据合规要求可签署必要的用户知情书。

智慧城市要确定能够访问人工智能系统相关数据的用户或程序的范围，

并对此实施了数据安全访问控制。为了实施这些控制，可以采用包括账号口令、指纹识别、人脸识别等的身份验证方式，以及角色管理、权限管理和访问控制列表等授权管理方式。同时，智慧城市需要对访问数据的行为进行了审计，包括记录了访问用户信息、访问时间、访问内容和访问结果等信息。

5.2.2.智慧城市训练数据配置

保证智慧城市模型训练数据配置合理性首先需要明确模型目标，选择与目标相关的高质量数据源。其次，适当地进行数据预处理，如去除噪声、填充缺失值、进行特征工程等，以提升模型性能。要保持数据集的多样性和平衡性，避免出现偏向导致模型训练不准确。另外，定期对模型进行评估和调整，不断优化参数设置，确保模型最终效果的合理性和有效性。最后，及时跟踪新的研究成果和技术发展，更新和优化数据配置方案。

要加强数据的标注安全，在进行训练数据标注前，需要明确标注目的、标注内容、标注人员的资质、标注环境以及原始数据的类型和级别，以确保标注内容的追溯性。在执行标注任务过程中，可以进行安全审计、数据分类存储并对标注过程进行审查。在标注结果输出时，我们对输出内容的格式、级别和内容进行了核查，并在数据交付时采取了诸如加密传输等的安全措施。标注任务完成后，需要对数据标注情况进行人工抽查，如果发现标注内容不准确，则会重新进行标注。

5.2.3.智慧城市数据防范投毒

保证智慧城市模型训练数据不受投毒污染，主要依赖于有效的数据清洗和验证，确保数据采集源的可靠性，通过加强访问控制和权限管理防止恶意

插入异常数据。使用机器学习或统计方法检测并剔除异常值或明显偏离正常分布的数据，定期对模型进行训练结果评估，如出现预测性能下降、模型偏差大等异常情况，需要立即查验数据，并进行必要的清洗和过滤。

在模型的开发阶段，智慧城市可以引入对抗样本，通过添加微小扰动或进行数据增强来提高模型的抗攻击能力。从模型结构角度来看，我们将多个模型的输出结果进行了融合，这样即使某些模型无法提供服务，我们仍然可以得到有效的输出。同时，智慧城市可以通过定期更新模型和添加新的训练数据以保持模型的鲁棒性。当模型鲁棒性出现降低时，我们会及时进行模型的更新和优化，以保证模型的稳定运行。

5.2.4.智慧城市数据泄露防范

为防范智慧城市训练数据泄露，可以使用安全管理和技术手段，包括访问控制、数据分类分级管理和敏感数据的加密存储及传输，以保证训练数据在存储、传输、加工和使用等各环节的安全。在训练数据的使用阶段，智慧城市业务系统可记录下关键操作行为。而针对在不可信环境、临时环境或是安全措施不足的环境中进行模型训练和测试，对敏感信息进行脱敏或去标识化处理，训练和测试结束后，相关数据将被安全转移或删除销毁。

5.3. 智慧城市 AI 业务服务安全

5.3.1.城市服务内容生成监控

智慧城市需要制定一个包含不良信息的内容安全管理和监测机制。这个机制建立在人工智能内容生产和传播的相关规章制度之上，并结合各类违法和不良信息可能带来的危害，对内容进行分级分类管理。

在用户输入环节，智慧城市针对那些输入违法不良信息或者使用、引导模型生成、传播违法不良信息的用户，设定了拒绝回答或进一步的处罚措施。在模型输出环节，设定了关键词或敏感词库，采用分类模型等方法对输出内容进行监测。同时，智慧城市也需要设定了异常答案和正常答案的评判标准，可以设置标准问题库，通过识别特定的提问内容，调用标准答案以降低输出内容的风险。我们将与意识形态、偏见歧视、侵犯个人或组织权益的内容一起纳入监测管理。最后，通过构建内容安全监测系统，管理者对预计输出或传播的文本、图片、音频、视频等内容进行监控，对已确认的不良信息进行过滤，对疑似的不良信息通过人工进行辅助审核处理。

5.3.2.城市服务伪造内容识别

智慧城市在舆情监控工作中，需要关注伪造内容识别。深度伪造是一种在社交媒体上常见的违法违规行为，它利用人工智能技术篡改和捏造真实的图像、视频和音频，以假乱真地误导公众，产生不良影响。因此，智慧城市可以根据行业监管要求和自身业务发展需要，部署能够识别人工智能生成或者合成内容的检测系统，从而对常见的深度合成算法以及人工智能模型生成或合成的内容进行识别。

5.3.3.智慧城市服务调用安全

智慧城市服务调用安全主要涉及身份验证、权限控制和数据传输安全。身份验证确保只有许可的用户或服务可以访问系统，防止恶意攻击。权限控制在验证身份后进行，确定用户或服务可以访问哪些资源，防止滥用。数据传输安全涉及在网络中传输的所有数据的加密和完整性校验，以防止数据被

截取、篡改或泄露。同时，也需要有应急响应机制，对任何安全事件进行检测、记录、分析和响应，以保护智慧城市服务的安全性和稳定性。

智慧城市也需要防范网络安全攻击。对其人工智能系统的外部访问、输入数据、及行为决策进行了检测，以便及时发现针对业务系统的安全攻击。智慧城市持续监测人工智能系统的运行状态和安全状况，并及时警告任何系统运行的异常情况。除此之外，智慧城市可以部署了针对人工智能应用服务调用接口的风险监控和安全防护能力，包括对 API 的资产监控、访问控制、异常行为监控、封堵阻断、API 接口加密、动态机器人拦截、弱口令防护、安全审计、以及 API 运营状态监控等。

5.4. 智慧城市 AI 平台能力安全

5.4.1.智慧城市算力防范滥用

部署算力安全管控措施，以防止算力资源被用于违规应用场景，例如网络攻击或密码破解等由恶意攻击者利用强大算力发起的活动。这些措施包括通过解析计算任务类型并结合算力用户的算力阈值，对智慧城市计算任务进行安全评估。如果超出算力阈值，系统会限制算力用量或拒绝其算力请求，并考虑降低该用户的信用。此外，系统还对算力用户的操作进行审计，监控并记录任何异常行为，以确保算力资源的正确和安全使用。

5.4.2.智慧城市供应链安全性

智慧城市 AI 模型训练前，检查使用组件的版本及已知漏洞。推荐使用国产化算力，推动人工智能系统适配国产芯片，以提升自主研发算力设备的比例，从而加强国内技术的独立性和自主性。为了保障供应链的安全，智慧城

市产业线建立了一套包括风险管理、供应方选择和管理、产品开发采购及安全维护在内的完整的供应链安全管理策略。这不仅涵盖了供应链各个环节的风险评估与控制，也包括对供应商的严格筛选和管理，以及对产品开发采购和安全维护的规范操作。通过这些措施，我们能够确保供应链的安全稳定，降低因供应链问题引发的风险。

5.5. 智慧城市 AI 运营合规安全

数字化时代，网络攻击的手段和频率不断增加，企业面临的安全威胁日益严峻。传统的安全运营中心（SOC）依赖人工分析和响应，不仅效率低下，还容易出现误报和漏报。攻击面的扩大和数据量增加导致安全运营人员疲于应对，难以在短时间内作出准确的响应。当下技术通过机器学习、深度学习和自然语言处理等技术，能够在海量数据中快速识别出潜在威胁，并提供有效的解决方案，提升安全运营效率，发挥体系的最大优势，满足国家对新型智慧城市安全运营工作的要求。

5.5.1.法律遵从建设

智慧城市安全运营制度是确保安全运营工作顺利进行的基石。通过项目制度、管理制度、合规制度的建设来帮助智慧城市明确安全管理标准、流程和责任分工，及时发现和应对网络安全风险，提高安全运营工作的效率，保障智慧城市安全的合规性，降低风险。同时，伴随 AI 技术的广泛应用，智慧城市的安全运营带来了很大的便利，但是也带来一系列的安全风险和挑战，人工智能的不可预期性、自我进化能力以及对隐私和数据的侵犯等问题，智慧城市安全和可信度带来了威胁,因此一个健全的 AI+安全行业管理制度势在

必行。

5.5.1.1. 安全制度建设

智慧城市安全运营管理制度是做好安全运营工作的基础，需要根据业务的实际需求和国家的具体要求制定安全运营管理制度和策略，并在实施过程中不断优化更新。

同时为覆盖安全运营的各个层面，可以制定分层级的安全运营管理制度体系，包括但不限于安全战略、安全制度规范、安全制度流程、安全细则与指南等多个层次的管理制度。

安全战略类文件主要是根据智慧城市安全运营目标、业务需求等制定的安全运营管理方针，指导智慧城市安全运营的建设目标、管理范围、基本原则等重要内容。安全制度规范类文件主要是为了落实方针政策而制定的管理规范、标准，应建立安全运营管理制度、人员管理、教育培训、监测预警、应急响应、安全评估、检查评价等制度规范。安全制度流程类文件主要用于明确安全运营管理的流程和规范操作，将安全动作固化，作为制度进行执行。一般包括安全事件管理、信息备份管理、安全培训及考核、权限管理、应急预案、日志管理等文件。安全细则与指南类文件主要用于指导具体操作或运营过程中的留痕文件。一般包括申请表单、安全报告、安全记录、事件清单、账号口令等文件。

5.5.1.2. AI 制度建设

人工智能技术应用广泛，其中包括各种系统、算法和模型。这些技术的安全问题涉及到数据隐私、系统漏洞等方面。通过建立合理的管理制度，能

够提前识别和解决潜在风险，保障人工智能技术的安全性。

（1）加强数据隐私保护

数据是人工智能技术的核心资源，而数据隐私的泄露往往导致严重后果。因此，建立健全的数据隐私保护制度是非常重要的。包括明确数据采集和使用的规范、加强数据加密技术的研发应用等。

（2）完善人工智能系统的监管机制

对于人工智能系统的使用和运行，需要建立起一套严格的监管机制。包括对人工智能算法的训练和测试过程进行规范，确保系统的稳定性和可控性。

（3）加强安全漏洞的挖掘与修复

人工智能系统中存在的安全漏洞是导致其安全问题的根源。因此，需要加强安全漏洞的挖掘和修复工作。建立起一套快速响应的漏洞修复机制，及时消除港在的威胁。

5.5.2.运营管理建设

在构建智慧城市安全运营体系中，运营管理工作能有效衔接安全管理制度和安全技术，确保安全管理制度和安全管理责任的有效落实，是智慧城市安全运营体系中的重要组成部分。

5.5.2.1. 智慧城市安全组织运营架构

在智慧城市的建设过程中，因其业务广、场景多等特点，会涉及城市管理局、大数据局、第三方开发公司等多方建设运营，需要按照决策层、管理层、执行层、参与层、监督层的组织架构，合理设计智慧城市安全运营的组织架构，确保能有效协调整合各方资源，保证各层级、各部门间的沟通协作；

指导和推动安全管理制度的制定与实施，确保安全措施的落实与执行。

在传统运营组织架构基础增设 AI 运营组，运用 AI 技术自动化完成一些数据分析、预测等重复性的工作，大大提高企业的运营效率，同时帮助机构减少人力成本。此外，通过优化算法和模型，降低运营成本。为智能城市安全运营带来新的发展机遇。

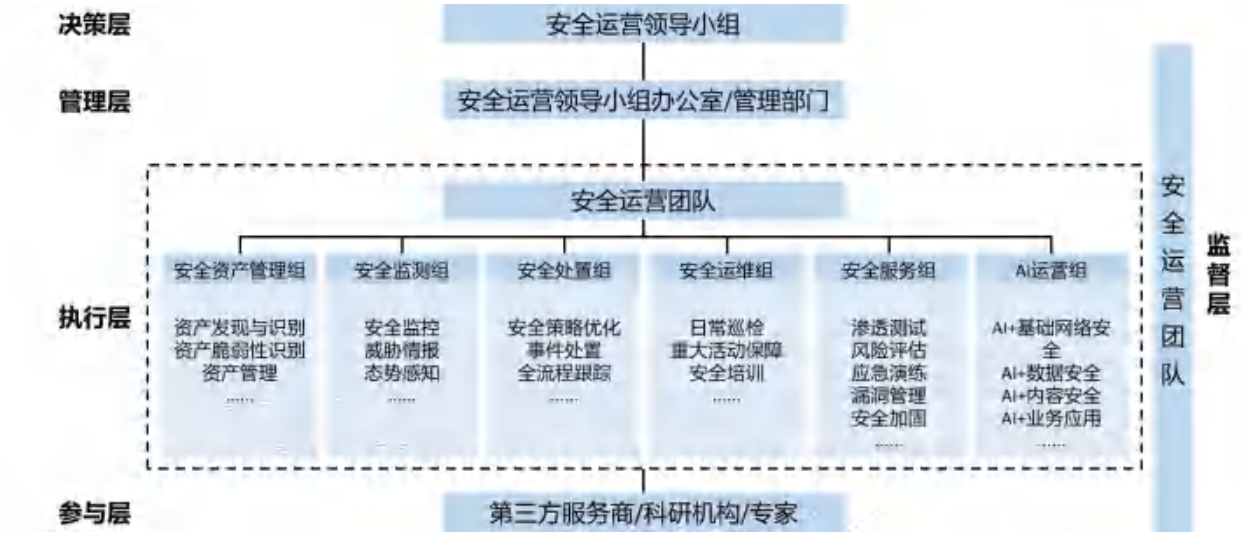

图 4 安全运营组织架构

（1）决策层

决策层作为智慧城市安全运营组织的顶层和安全运营工作的决策机构，一般由智慧城市主要领导人、智慧城市内主管网络安全的最高负责人构成。决策层主要负责统筹指导智慧城市建设中的安全运营工作，制定安全运营战略或规划，负责安全运营标准规范、管理制度及其他重大事项的审议等工作。

（2）管理层

管理层是连接决策层和执行层的桥梁，既要执行决策层下达的任务，又要指导、监督执行层的日常活动，在智慧城市安全运营组织架构中具有举足轻重的地位，一般由网络安全部门、信息安全部门牵头组建。管理层主要基

于决策层给出的策略方针，对安全运营实际工作制定详细方案，因地制宜的制定与安全运营规划配套的安全运营制度、标准规范；落实国家、省、上层关于安全运营建设经费的要求，落实各参与方的安全运营职责与分工，建立权责明确的安全运营管理组织；加强建设安全运营队伍，加强人员安全管理，负责组织人员安全意识、安全技能教育与培训，制定安全考核机制、供应商人员安全管理机制等；安排、协调和监督各部门网络安全工作，及时向上层部门汇报等工作。

（3）执行层

考虑到智慧城市业务范围广的特点，执行层一般由业务部门具体安全运营人员、安全技术人员组成。执行层需要认真贯彻落实管理层提出的安全运营要求，针对上级部门的指示制定具体的实施方案，明确安全运营各项工作的阶段任务、分工及时间节点，保证工作有序进行；要做好日常安全运维、应急处置、安全检查等工作。同时运用人工智能技术（AI）对业务进行运营、管理和优化的过程，严格遵守安全运营操作规程，能及时发现制度规范中的漏洞和潜在危险，帮助管理层及时调整改进。

（4）参与层

参与层是管理层和执行层的协助者。参与层主要由第三方服务商（如等保测评、密评服务商等）、科研机构、专家等人员构成。参与层主要负责协助执行层承担安全运营的建设、实施、维护、服务等工作；协助管理层承担安全运营标准或制度的制定。

（5）监督层

监督层要独立于决策层、管理层、执行层和参与层，人员不能共同兼任，要确保其监督审计工作不受其他四层的影响，从而保证组织能发现安全运营过程中真实出现的问题和面临的风险。监督层主要对安全运营方面的制度、策略、规范文件等的贯彻落实情况进行查验、监督和考核，对安全运营工作进行监督落实，对安全运营风险进行监控和审计。

5.5.2.2. 智慧城市安全组织人员管理

（1）人员任命

人员是组织内部的核心要素。人员任命的精要是要将合适的人员配备到合适的职位上，这对安全运营人员管理也至关重要。人员任命主要有以下几类角色：一是责任人聘任，在安全运营组织体系中首先需要任命安全第一责任人，统一负责安全相关事务，承担相关安全责任，一般由部门第一负责人担任；二是信息安全经理（CSO），负责公司的整体信息安全策略，制定和实施信息安全政策，监督企业内所有有关信息安全的活动；三是网络安全工程师，主要负责监控公司的网络，预防和应对网络攻击，确保网络的稳定和安全；四是安全审计员，通过审查公司的信息系统和操作过程，评估公司的信息安全措施是否得到有效实施，是否符合相关法规和标准。

（2）AI 人才教育

（1）AI 技术在智慧城市领域的应用日益广泛，不仅提升了城市的智能化水平，还带来了更高效、更便捷的管理和服务方式。然而，要确保 AI 技术在智慧城市中的合规化应用，并充分发挥其潜力，需要引进专业的 AI 技术人员，

并对机构管理者及普通运维人员进行 AI 技术赋能培训。确保 AI 技术在智慧城市中的合规化应用，并充分发挥其潜力，为智慧城市的可持续发展提供有力支持。与知名高校的 AI 相关专业合作进行人才联合培养。通过成立专门的联合实验室或专项研究小组，互派师生、人员参与其中，深入到教学现场与业务一线，结合学术研究、专业教学和智慧城市业务研发的需要，更有针对性地从源头培养未来人才。

5.5.2.3. 智慧城市安全运营大模型

智慧城市安全运营方案以 AI 大模型帮助企业在网络安全中解决告警数量多，噪声大，并且人员能力难以支撑高效的告警、事件研判等问题。同时，监控 AI 技术在智慧城市滥用问题。安全运营大模型具备自然语言对话能力、检测能力等智能化功能，能够实现 7×24 小时全天候值守，提升效率并缩短对网络安全风险和威胁的响应时间。大模型的应用承载着 80%的安全运营操作，极大地提高了安全运营效率，从而构建了安全运营新范式。通过智慧城市安全运营大模型，可实现“发现告警-智能研判-威胁定性-封堵隔离-影响面调查-加固建议”的全流程闭环。

5.5.3.评价指标

（1）安全管理类指标。主要是对各地智慧城市安全运营管理措施是否充分进行评价，其主要包含安全运营战略规划、安全运营标准规范、安全运营管理组织、人员安全管理以及安全运营投入 5 个指标。

（2）安全运营类指标。主要是对智慧城市安全运营体系在运行过程中的风险识别、安全监测及应急处置等能力进行评价，其包括资产管理、安全监

测、安全运维、安全处置、安全检查、安全审计等指标。

（3）安全效果类指标。主要是对智慧城市安全运营体系的实际运行效果进行评价，其包括安全漏洞、安全事件、攻防对抗等指标。

（4）智能化运营指标。主要是对智慧城市 AI+安全运营体系的结合程度进行评价，其包括 AI 能力平台安全水平、AI 合规性、实际效果等。

6. AI 赋能智慧城市安全方案

6.1. AI+智慧城市网络安全

智慧城市 AI+网络安全是通过人工智能技术来保护和增强网络层的安全性，涵盖下一代防火墙、全流量威胁检测、智能路由与负载均衡等技术，与传统网络安全不同的是，AI+的介入对传统安全领域难解决的问题如加密与隧道技术提供了解决方法。

6.1.1.下一代防火墙

智慧城市将广泛部署防火墙，以应对日益复杂和多样化的网络威胁，具备 AI 技术支持的下一代防火墙，不仅具备传统防火墙的包过滤功能，还可以通过人工智能和机器学习，能够进行网络行为分析，识别异常模式，并结合威胁情报源提供实时的威胁反馈和响应。这些功能使其能够在智慧城市中提供比传统防火墙更加全面和智能的网络安全保护，有效应对现代网络环境中的多重威胁。

6.1.2.全流量威胁检测

在智慧城市中，加密流量占网络流量的比例已超 70%，传统威胁检测方式效果减弱。AI 全流量监测通过最佳监督式机器学习算法和网络协议还原技术，训练分类标记并建立增量式学习数据库，构建自动化威胁检测体系，实现对恶意加密流量的高效准确识别。系统采用有监督的机器学习算法进行样本采集、处理、模型训练、验证，使用 leaf-wise 分裂策略进行分类，提高检出准确度，缩短模型训练时间。同时，采用互斥特征捆绑算法提升检测效率，降低内存消耗，有效检测恶意加密流量、DGA 域名、隐蔽隧道等高级威胁。

结合威胁情报和检测模型等多种手段，全面识别网络流量中的威胁行为，为智慧城市安全运营提供支持。

6.1.3.智能路由与负载均衡

智慧城市中的智能路由和负载均衡是确保城市网络基础设施高效、安全、可靠运行的重要技术，AI 可以优化网络路由，动态调整路径选项以避免拥塞和潜在的安全威胁。例如安全策略集成、自适应路由、机器学习实时优化路径、基于 AI 的负载均衡预测分析，以确保网络资源的高效利用并防止单点故障。这些技术在 AI 的支持下，能够更高效的管理智慧城市中的网络基础设施，保障各类城市应用可靠性和性能。

6.1.4.安全智能问答

现有的安全智能问答应用基于知识图谱结合长短记忆网络模型开发，回答问题的模式固定，整体能力有限，基于 AI 和安全大模型的安全知识问答对话技术，引入了安全语料，既利用了大模型 NLP 自然语言的类人化能力，也保证了模型回答的专业性。除了实时应答的场景之外，通过对运营过程数据的分析，还可以生成便于运维和决策的总结性报告。在智慧城市中，良好的交互性可以给各类安全工作提供支撑，提升各类系统的易用性。

6.1.5.威胁情报分析

威胁情报分析引入人工智能和机器学习技术，及时发现、分析和应对潜在的网络攻击和安全威胁，为智慧城市提供更高层次的安全保障。基于 AI 的威胁情报分析技术，避免了基于预设数据标签的归并去重，以及结合人工运营的传统方式，解决了情报数据量大、异构性强、时效性差的问题，并且在

智慧城市中可以对系统运行日志、用户行为日志、流量数据及外部公开情报数据、黑客论坛等多个数据源进行自动加注标签，并利用机器学习、深度学习等技术进行情报聚合、分类等，为城市中的威胁情报进行安全风险评估、预警、安全事件响应提供能力支撑。

6.1.6.自动化渗透测试

渗透测试可以有效评估智慧城市中系统的安全状况，并提出合理的改进方案，传统的渗透测试工作高度依赖安全人员的人为判断和操作，结果和效率都存在提升空间。基于 AI 大模型的自动化渗透测试，将模型和工具进行结合，让大模型具备渗透测试输入输出和过程推理的反馈能力，配合各项自动化脚本和智能流程应用，让智慧城市中的系统可以自动化进行渗透测试工作，辅助渗透测试参与人员，提供高效的工作流程并生成可靠的报告。

6.2. AI+智慧城市应用安全

智慧城市 AI+应用安全是作用在应用层的新型防御技术，在智慧城市的各种应用中，通过使用人工智能技术，确保应用能够安全运行、防止数据泄露、抵御网络攻击，并且保护公民的隐私和城市基础设施的完整性。例如风险控制、拟态蜜罐、内容检测是基于 AI 的新型技术。

6.2.1.风险控制

风险控制系统通过利用 AI 技术构建业务访问动态认证机制和访问行为分析，追求主动构建一个不断变化的访问环境，从而实现对自动化攻击和未知风险的防御，有效防护爬虫、AI 脚本攻击等机器行为。动态智能身份认证通过生成智能动态令牌和动态指纹，持续跟踪分析访问来源与行为，以确认访

问者身份并实施合适的防御策略。智能行为分析通过多维度特征采集和深度学习，训练和优化模型，实现对威胁行为的识别。这种防御方式同时弥补了传统 WAF 的缺陷，大幅降低攻击成功的可能性。与此同时，风险控制系统可以与传统检测规则结合，为智慧城市的 Web 应用等提供动静一体的综合防御。

6.2.2.拟态蜜罐

拟态蜜罐通过模拟真实目标，如操作系统、网络设备等，引诱攻击者进入虚假目标，保护真实目标并分析攻击行为。AI 蜜罐能更好地分析攻击行为，并实现大规模智能联接部署。基于 AI 驱动的分析平台利用机器学习技术自动识别攻击方法和特征，从而建立欺骗系统，有效发现和研究攻击行为。AI 技术实现蜜罐的快速联接和部署，在业务环境中建立欺骗网格和布置诱饵。AI 蜜罐系统能智能建立欺骗防御系统，并通过攻击手法分析，模拟不同环境引导攻击者进入预设陷阱。一旦攻击者入侵，系统会发出高准确度警报，并通过 AI 技术进行数据清洗和去噪，明确攻击意图，并进行关联预警和防御策略反制。同时，AI 技术可实现事件驱动的自动响应，自动创建仿真部署。

6.2.3.内容安全治理

一种通过分析和过滤应用数据内容，防止未授权、恶意或者不适当的信息通过网络传播。AI 通过学习恶意内容，可以更快速的识别风险内容特征。通过搭建内容风控体系，充分发挥 AI 在语义理解、图像识别、音频辨识等方面的优势，利用 AI 大模型的预训练机制，实现文本、音频、视频数据审核能力的提升。

6.2.4.供应链安全智能分析

软件供应链已成为业界关注焦点，涵盖了软件从源代码审查到交付和部署的整个过程，以确保软件在每个环节都是安全和可信的。随着 AI 技术的介入，智慧城市中的供应链信息维护、开源软件管理、闭环软件物料这些方面都可以被大模型进行处理，第三方的开源库、二进制软件包可以在不解密的状态下进行诊断，除此之外，运用安全大模型构建软件，可以从开发到运行，全流程的执行安全测试任务，涵盖策略生成、任务执行、结果回收和报告总结等方面，实现一站式、智能化的智慧城市软件供应链安全风险管理。

6.2.5.API 安全智能监测

业务系统的 API 已是黑客攻击的重点目标，现有的 API 安全解决方案管理能力不足，常引发数据泄露、恶意攻击、未授权访问、业务运行故障等问题。通过 AI 与大模型技术，可以对智慧城市中 API 资产进行保护。API 资产自动发现与识别技术、建立 API 资产管理库、监测并分析 API 异常操作，识别隐蔽威胁（令牌违规使用、数据泄露丢失等）等多项技术都将在智慧城市功能接口中起到关键作用，联合 AI 大模型提供风险处置建议，增强 API 在使用和内容提供过程中的安全防范能力。

6.2.6.恶意代码检测

通过静态特性分析（比如对二进制代码结构的检查）和动态行为监测（例如在运行时追踪操作序列），人工智能技术掌握并学习了以往恶意软件的行为模式。这使得 AI 可以辨识出新出现的、以前未曾见过的恶意文件，特别是对于零日攻击的识别上非常有效。另外，随着持续的训练，AI 能够逐渐提高

对恶意文件的识别准确度，降低误报和漏报的可能性，同时也能快速适应恶意软件的新变种。

6.2.7.用户和实体行为分析

智慧城市服务广、用户多。依托大模型的推理能力，创新构建用户访问异常行为分析模型。自动完成网络日志、用户行为日志、数据库日志的分析，输出高精准的异常行为告警信息。通过深度学习算法构建异常行为检测模型，实现预测和识别未知异常行为，辅助安全专家开展数据安全运营服务。

6.3. AI+智慧城市数据安全

基于 AI+数据安全能力，在智慧城市中负责保护城市各类数据，AI 数据安全涵盖多个维度，从数据的存储到模型的训练和部署，再到实时监控和响应。使用数据管控平台、AI 数据水印等技术确保隐私的安全性和可靠性。

6.3.1.AI 数据水印

是一种用于保护数据所有权和确保数据完整性的技术。通过在数据中嵌入难以察觉但可验证的标识符，数据水印技术帮助监控数据的使用情况，防止未经授权的复制和篡改，保障数据的版权和安全。值得一提的是，随着生成式 AI 的发展，AI 生成的数据也需要通过水印技术进行鉴别、区分，防止数据的混淆和攻击。

6.3.2.AI 数据分类分级

根据重要数据识别的相关法律法规、行标、企业规范和业务知识，增强通用大语言模型在特定垂直领域的分析推理能力，将结构化、半结构化、非

结构化数据源转换为自然语言形式，构建重要数据、核心数据的智能化识别与分类分级打标技术，从而提升数据识别的自动化程度以及准确率和效率。

6.3.3.AI 数据安全审计

利用人工智能技术的监督学习和非监督学习结合的方式，建立一种灵活而智能的审计技术手段。通过学习用户的历史行为数据，可以形成用户在特定行为场景下的正常行为范围基线。在后续的审计过程中，利用训练好的行为基线模型，可以实时监测和检测用户当前的行为是否偏离了正常行为范围，从而发现异常的业务操作行为，并通过对已审计历史告警事件相关数据的学习和训练，构建智能化审计模型，自动识别出告警数据中的是否包含安全事件并自动审计。

6.3.4.AI 数据安全合规工具

基于 AI 的数据安全检查工具箱，以实现对数据安全的智能化、自动化评估与检测。利用 NLP 构建智能评估矩阵模型，自动分析用户的需求，生成目标评估矩阵，实现与用户的自然语言交互，理解用户需求与问题。利用收集的企业现有评估信息、业务流程、安全措施等相关信息，自动创建并执行包含检查要点、检查问题、检查所需材料等内容的评估矩阵。同时对收集到的证据进行深度分析，匹配评估检查信息，基于预设的评定标准与逻辑，自动对分析结果进行评定，得出证据与检查信息的满足度。

6.3.5.AI 数据脱敏

AI 技术可以实现在使用数据的过程中进行动态的信息脱敏处理，这意味着根据用户的身份、访问环境及使用场景的不同，自动调整展示的数据形式。

如用户是内部的审计员，可能需要显示更多的细节内容；而对于外部的合作伙伴，我们则会展示经过高度脱敏的数据视图，这样做的目的是为了最大程度地减少数据泄露的风险。

6.4. AI+智慧城市公共安全

智慧城市建设与 AI 的紧密结合，为公共安全领域带来了革命性的变化。通过人工智能技术，可以实现对城市环境的实时监测和分析，对污染源进行准确定位，及时发出预警信息，为环境保护决策提供科学依据。同时，智能治理与环境优化使得城市环境质量得到提升。在公共安全方面，AI 技术应用于人脸识别、视频监控、事件预警等方面，帮助城市实现精准安全防控。通过人脸识别、车牌识别等功能，辅助公共安全管理和治安维护工作，实现对紧急事件的智能识别和实时预警，提高应急救援工作效率，有效保障公共安全。

6.4.1.社会治理安全方案

智能城市社会治理安全方案可使用视频监控和智能视频分析技术实时监控关键区域，防止犯罪活动，最大化公共安全。其次，利用传感器、摄像头和 AI 算法的智能交通管理不仅优化交通流量，减少事故，还能快速响应紧急事件。此外，智慧巡逻机器人提供 24 小时服务，配合人工智能人脸识别等技术，确保社区安全，提高治安水平。不可忽视的是，预测性警务通过数据分析预测高风险地点和时间，实现警力优化部署。这些技术的综合运用使城市安全管理更智能，更高效。

6.4.2.灾情监测预警方案

灾情监测预警是智慧城市防灾减灾的关键，其中包括基于人工智能技术的灾害预警系统、应急指挥调度和智能消防系统。利用气象传感器、地震监测设备及洪水预警系统，结合大模型算法，灾害预警系统可提前预测自然灾害并发出预警，以最大程度降低灾害带来的破坏。在突发事件发生时，综合应急指挥平台能够协调各部门资源，实现快速响应与处理。同时，智能消防系统通过烟雾和温度传感器等设备监控火灾隐患，并能及时通知相关部门予以应对。三者共同构成了全方位的灾情监测预警方案，旨在预防自然和人为灾害，保护人民生命财产安全。

6.4.3.公共卫生安全方案

公共卫生安全方案主要包括疾病监测预警、食品安全管理和环境监测三个关键环节。通过运用大数据和 AI 技术，疾病监测预警系统能实时监控和预测疾病的传播趋势，及时发布健康咨询。为了确保食品安全，可以通过设立溯源系统和传感器，对食品生产和供应链进行严密监控。环境监测系统能定期检测空气质量、水质和噪音等环境指标，一旦发现异常及时处理，防止环境污染，保护居民健康。这套全面覆盖的公共卫生安全方案，可以有效地防控各类卫生健康风险，保障公众生命安全。

6.4.4.安全生产管理方案

安全生产涉及工业互联网安全和作业环境监测，需要建设基于人工智能监控和管理的工业控制系统。作业环境监测通过利用传感器和 AI 计算平台实时监测工作场地的有毒有害气体和噪音水平，从而保障工人健康和安全。

7. 中国移动智慧城市人工智能安全参考案例

7.1. 智慧城市 AI 风险防范案例

7.1.1.黑龙江省级海算政务大模型安全防护

（1）背景与需求

黑龙江省数字政务项目建设从 2022 年启动以来，坚持数据驱动、创新引领，持续加快人工智能应用。借助其在数字政府领域的丰富建设经验，中国移动以通用大模型为基础，并融合了政务领域的数据进行精调。同时，引入了政务领域约束模型对输出进行限制，最终成功打造出面向政务领域的行业大模型—九天·海算政务大模型。

海算政务大模型将“政务政策-政务事项-政务数据存储”深度贯穿模型，驱动整体业务流程灵活易用。通过向大模型发出自然语言指令，便可通达深层数据库，串联多来源、复杂异构的数据表，快速获取直观的数据分析结果。目前累计训练 10 多类数据、4000 亿私域数据、10 万精标数据。

在安全方面，海算政务大模型使用信息场内的政务领域专业知识对模型进行课程学习式增强以及对齐泛化，同时协同私域数据作为最终结果反馈用户。通过政务信息场的调度能力，汇聚散落的关联数据，围绕用户咨询实现场内问题全解决；拓宽政务服务边界，实现主动式服务；政务流程不出“场”，实现可信的政务问题响应，保证政务服务安全可控。

图 5 海算政务大模型建设方案

（2）建设方案

以需求边界、业务边界、安全边界为原则，对云、网、数、用 4 方面开展详细调研和确认，在建设过程中确保 26 项安全能力落地。包括集中展示、能力输出、业务支撑和数据汇聚等模块。配套建设外部对接平台、安全运营管理体系、安全运营服务体系、安全运营物理环境等。共同防护

针对省级海算政务大模型平台，通过云、网、数据、应用安全运营中心，项目实施了全面严格的网络和数据安全防护措施，并结合了有效的合规管理，确保所有操作都符合相关政策和法规的要求。项目组设置了针对平台的应用监控系统，实时跟踪并分析应用程序的行为，以及与之相关的数据流动，即时发现并解决可能存在的问题。

关于训练数据管理，本项目定义了严格的流程。海算大模型只使用经过合适清洗和匿名化的数据进行训练，从源头上确保数据的安全性和隐私性。同时，我们对精练训练数据进行归类和标签，保证在使用时我们能够清楚地知道哪些数据被用于什么目的。项目组定期开展数据质量检查和更

新，以确保海算大模型基于安全和准确的数据进行训练。

通过这样的一系列数据和网络安全防护措施，包括但不限于访问控制、网络加密、数据监控、漏洞扫描、合规管理、应用监控以及训练数据管理，我们构建了一个安全、稳定且可靠的海算政务大模型平台，为用户提供了一个优质的使用环境。

图 6 黑龙江省级智慧城市政务安全建设方案

（3）建设成果

海算政务大模型构建了 12345 智能热线、政务智能搜索、政务智能助手、公文写作辅助四大应用场景，服务了黑龙江省数字政府项目、山东省政务大模型项目、广东省联合实验室案例、深圳市民生诉求（12345 热线）项目等。

针对海算政务大模型的安全能力建设持续推进，从云、网、数、用等方面防护大模型的技术风险，同时通过合规管理、训练数据管理、内容管理等解决了大模型的合规需求。截至目前，未发生安全相关事件，系统平稳运营为客户提供优质服务。

7.1.2.中国移动 AI 模型漏洞评估平台

（1）背景与需求

人工智能大模型应用越来越广，面临着攻击者通过越狱攻击、目标劫持和提示泄露等方式，绕过人工智能大模型自身的防御策略，非法获取大模型的敏感信息，造成行业知识被窃的巨大安全风险。与此同时，攻击者还通过对微调数据进行投毒的方式，影响大模型的输出的准确性，或者通过向大模型输入对抗样本的方式，诱导大模型做出错误推理。此外，用于构建大模型的各类组件和中间件，可能存在软件漏洞，从而造成大模型参数或者大模型应用被窃取或非法控制，由此引发智慧城市智能应用异常。

中国移动根据自身经验，推出 AI 模型评估平台，用于解决人工智能模型漏洞评估。

（2）建设方案

针对以上需求，建设方案如图所示。

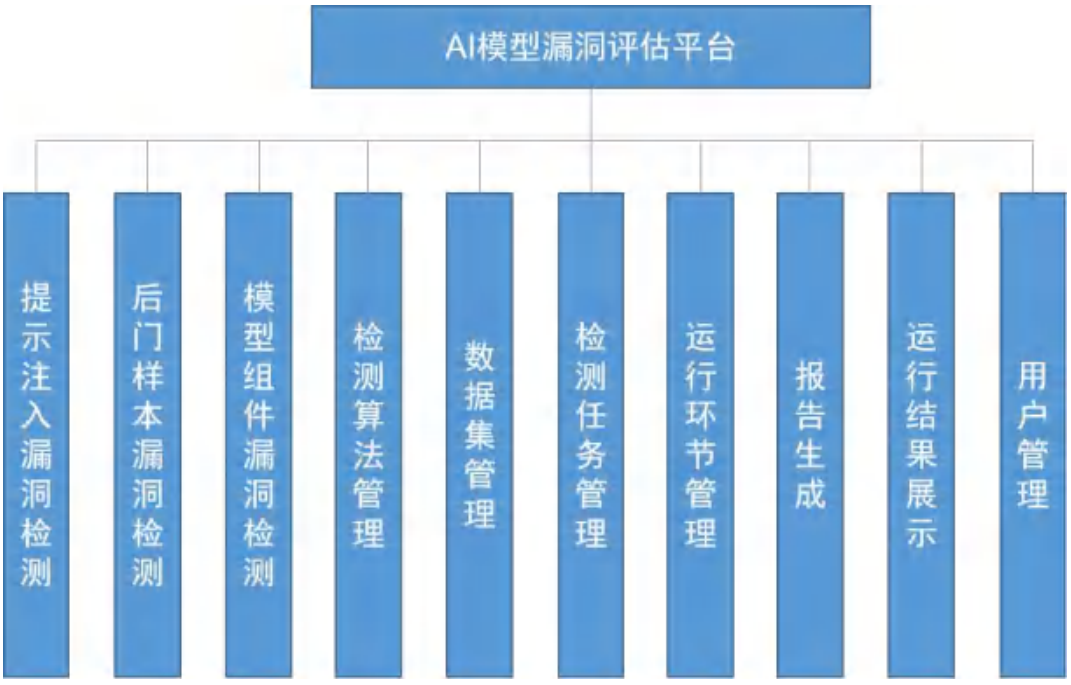

图 7　AI 模型漏洞评估平台架构图

大模型安全评估工具是一个全面且高效的漏洞评估工具，旨在确保大模型在智慧城市各类应用场景中的安全性。该系统通过强化学习和深度学习等技术，对大模型进行反向推导和交互式验证，检测可能存在的注入指和后门指令，对大模型框架和相关组件进行安全分析，从而发现大模型在组件调用、模型加载、算法运行等过程中可能存在的安全漏洞和风险。同时，该系统还提供了丰富的安全报告和可视化工具，帮助用户深入了解大模型的安全状况，为大模型的优化和加固提供有力支持。大模行安全评估工具评估维度包括模型中毒、提示注入和样本对抗等方面，并生成大模型安全评估报告。

提示注入漏洞检测

模型提示注入漏洞会导致攻击者利用漏洞构造的恶意提示来干扰模型的预测结果，使模型输出错误或不可预测的信息。提示注入漏洞检测模块通过对模型及其输入进行分析，比较模型在正常输入和恶意提示下的输出差异，识别可能存在的提示注入漏洞。当系统检测出模型存在漏洞时，将提出相应的评估结果和修复建议，从而提高其对恶意提示的防御能力。

后门样本漏洞检测

模型后门样本漏洞在被攻击者利用时，可能会使模型在接收到特定输入时产生预期的错误输出，这种漏洞的存在使得模型更容易受到攻击者的操控。模型后门样本漏洞检测模块利用后门样本的特性进行检测，通过对模型输入进行扰动，观察输出是否发生变化，以判断模型是否存在后门样本漏洞。当系统检测出模型存在漏洞时，将提出相应的评估结果和修复建

议，以排除可能存在的后门样本，以免这些恶意注入的数据影响到模型的准确性和可信度。

模型组件漏洞检测

模型组件漏洞是模型在训练和应用过程中使用的各个组件可能存在的安全缺陷或错误配置，该类漏洞可能导致模型遭受攻击、数据泄露、性能下降等风险。模型组件漏洞检测模块源代码进行分析技术和漏洞扫描技术对模型组件进行自动化检测，发现可能存在的漏洞和安全问题，然后分析检测到的漏洞的严重程度和潜在影响，评估模型组件的安全性和可靠性，根据评估结果生成相应的修复措施和建议，以提高模型组件的安全性和稳定性。

检测算法管理

检测算法管理模块负责管理和维护提示注入漏洞检测算法、后门样本漏洞检测算法和模型组件漏洞检测算法。提供算法存储、版本管理、参数预设以及接口配置等功能，保障检测算法的可用性和一致性。同时提供外部算法接口管理，方便用户添加新的算法，配置新的检测策略。

数据集管理

数据集管理模块则负责收集、整理、存储和管理检测用的典型数据集，为检测模块提供丰富可靠的数据依据。此外，数据集管理还负责外部数据接入和保存，方便检测模块对外部数据进行检测。

检测任务管理

检测任务管理模块负责创建和执行检测任务，通过任务形式对提示注入漏洞、后门样本漏洞以及模型组件漏洞进行检测。检测任务管理模块通

过自动化调度漏洞检测算法，使系统能够高效、精准地检测模型是否存在漏洞。当发现漏洞时，系统通过告警提示用户发现漏洞事件，然后模块根据用户制定的安全策略，标注出该漏洞并生成改进建议和方案，从而有效防止漏洞对模型性能的影响。

运行环境管理

运行环境管理模块是系统的基础支撑，它负责管理和维护算法的运行环境，确保算法的可用性。同时运行环境管理模块也为新增算法和第三方算法提供基础环境和适配环境部署能力，确保算法的正常运行。

报告生成

报告生成模块能够根据检测结果自动生成详细的检测报告，为用户提供全面的检测结果分析。

运行结果展示

运行结果展示模块则负责将检测结果以直观的方式展示给用户，便于用户分析和理解。

用户管理

用户管理模块负责用户信息的注册、登录、权限管理等操作，保障系统的安全性和用户权益。

（3）建设成果

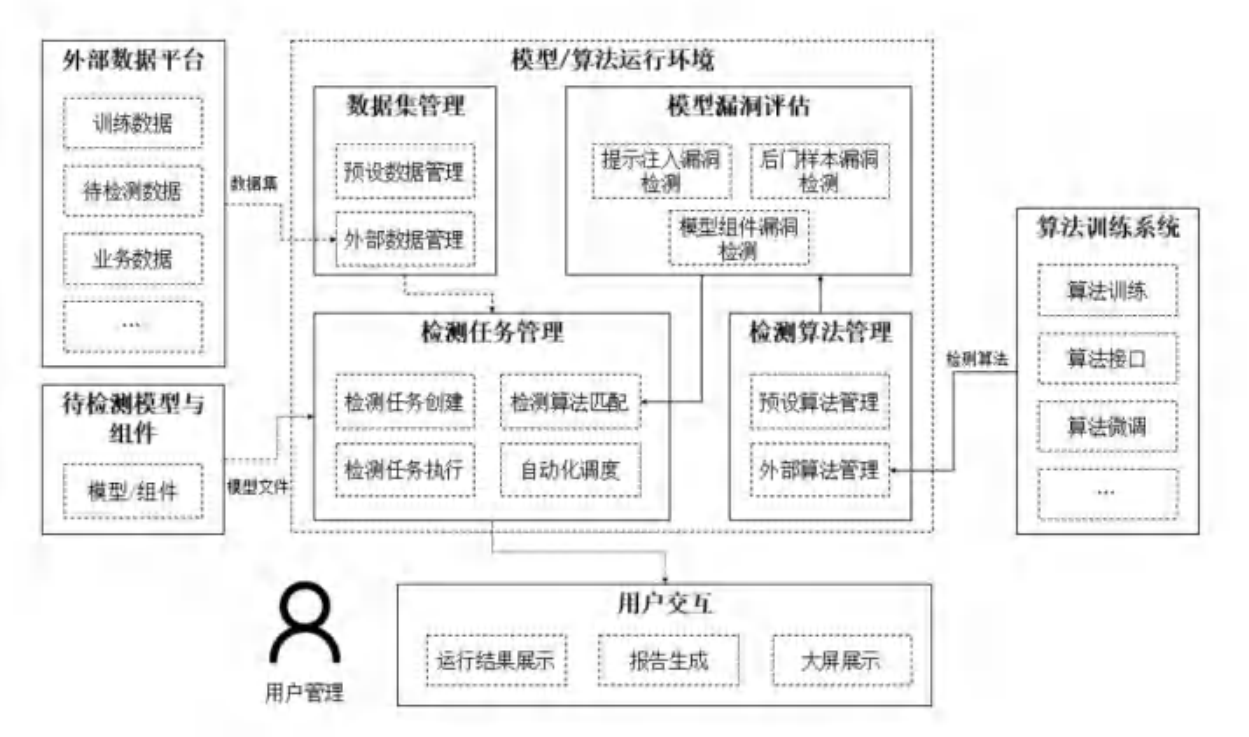

图 8　智慧城市大模型安全合规解决方案

该大模型安全评估工具功能已在中国移动上研院智慧城市配套工业大模型及金融大模型上线评估中初步应用落地，截止目前，工业大模型未被监管单位通报并检测到合规、网络与数据安全风险，并保护用户的隐私和数据安全；后续将会依次将配套 API 安全监测等一系列能力。平台先后为辽宁盘锦、安徽合肥、云南文山等地的多家农商行、财险公司提供了安全稳定的风险辅助管理服务。目前大模型安全能力赋能超过 20 个应用。

该工具完善了工业大模型及金融大模型安全一体化防护体系，保障工业、金融核保大模型安全上线。

7.2. AI 赋能智慧城市安全案例

7.2.1.中山市政务信息化安全系统建设和运营案例

（1）背景与需求

中山市政务服务和数据管理局于 2020 年 6 月建设了集态势可视、事件预警、工单流转的安全运维管理中心，但随着政务服务场景的快速扩展，出现了告警疲劳、专家稀缺、效率瓶颈等现状。

当前深度学习、机器学习等技术的飞速发展，AI 大模型在图像识别、自然语言处理等领域取得了显著成就。这些技术进步极大地推动了社会生产力的提升和生活便利性的改善，同时，随之而来的还有一系列的安全风险和挑战，黑客的攻击和学习成本越来越低导致政府网络安全面临的风险越来越大、自动化攻击越来越多导致政府需处理的告警事件越来越多，因此需要安全大模型引入解决新技术带来的安全问题。

通过补充网络安全大模型能力实现资产漏洞治理、威胁检测发现、告警事件闭环、处置调查、总结汇报等能力提升，人工智能技术持续赋能于威胁检测与态势感知、终端、边界、云安全等安全产品，构成了高效智能运营的安全运维中心。

（2）建设方案

通过部署安全大模型对接政务外网安全运维中心的流量、日志、代码等数据的预训练，流量理解能力、代码理解能力、攻防对抗理解能力和安全常识理解能力，实现资产漏洞治理、威胁检测发现、告警事件闭环、处置调查、总结汇报等能力提升。

1、安全数据分析与安全告警研判：支持多源多类型威胁数据汇聚，降噪，

并对降噪后告警特征关联发现威胁事件，并分析威胁事件的涉及范围、形成原因、攻击者身份及风险评估等，支持基于自然语言形式解读攻击流量，分析攻击者手法和意图等，通过全面的安全态势分析和安全大模型的能力发现未知威胁。

2、安全事件响应与安全知识问答：支持告警事件自动化线索取证分析、统计分析、溯源分析、协助或主动封堵，能对安全设备进行控制等。支持对安全领域中文本数据的深入理解处理能力，需准确地解析和分析安全相关的文本信息，并根据上下文语句回应用户的需求。

图 9 中山市政务信息化安全建设案例图

（3）建设成果

本项目建立在中山市政务服务和数据管理局 2020 年 6 月建设的安全运维管理中心上，首次将网络安全大模型与安全运维结合，覆盖了“云、网、数、端”的纵深防御一体化安全体系架构，具备了全面风险感知、实时监测预警、精准研判分析和快速响应处置的安全运营能力，保障了全市政务信息化系统

的安全、持续、稳定运行。本次项目覆盖中山市市委政法委、司法局、交通运输局、农业局、市场监督管理局等26家单位，为后续业务扩展占据了有力形式。

7.2.2.启明星辰安星智能安全助手运营案例

（1）背景与需求

与传统的网络安全方法相比，人工智能在网络安全中具有独特的优势，人工智能能够通过机器学习和深度学习技术，从大量数据中学习和识别模式，发现隐藏的关联性和异常，而且人工智能能够快速适应新的威胁，并提供实时的响应和防御。因此，需要继续研究和发展人工智能技术，加强其在网络安全中的应用和创新，以确保网络空间的安全可信。

安星智能安全助手是基于安全大模型的全场景智能安全运营助手。它具备自然语言指令理解、安全数据分析和解读、安全运营任务智能决策和自动化执行等能力，是新一代智能安全运营平台的决策和控制核心。基于安全行业的特点，安星智能安全助手发展了独有的基于AI安全智能体的安全垂直领域“大小模型自主协同”技术体系，把启明星辰深厚的安全产品能力和AI专项安全小模型积累，与大模型强大的意图理解和推理能力进行融合，构建了由安全大模型驱动的全场景安全智能自动化运营中心。通过“大小模型自主协同”技术体系，安星智能安全助手可以作为安全运营入口，能够大幅降低安全运营的技术门槛，提高安全运营的执行效率，实现说一句话就可以准确驱动安全产品和专项安全模型执行安全运营任务的下一代智能化安全运营体验。

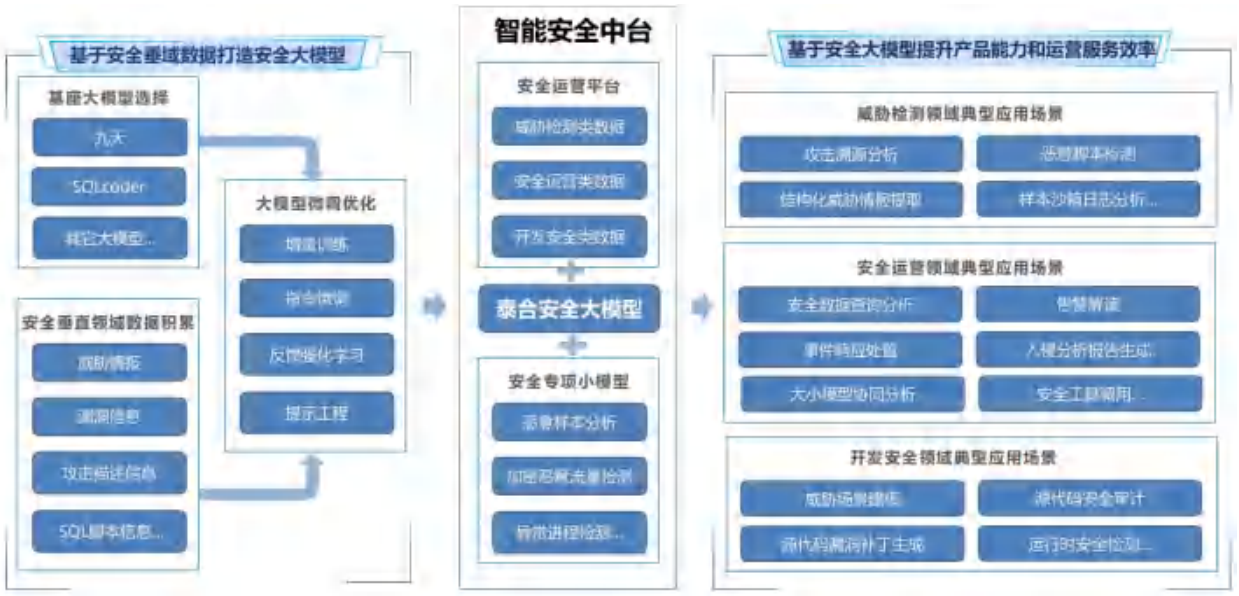

图 10 安星智能安全助手技术应用路线

（2）建设方案

智能中心：智能助手以自然语言交互为核心，提供安全事件多人协同处置、安全指令下发、剧本执行交互平台，在安全事件处置过程中，实现在交互界面中，完成安全人员交互、安全人员下发安全指令、以及调用安全剧本调用功能。协同机器人实现 7*24 响应，能够进行图片智能识别，支持图片文字内容提取，协助下发或解析安全指令，智能推荐相关知识库和处置动作，提升事件响应效率。

网络安全领域知识服务：使用大模型全量参数预训练、监督微调来向大模型注入安全垂直领域知识，使用检索增强生成（RAG）技术灵活的向大模型注入包括如最新的威胁情报、网络知识和产品使用手册等

数据分析呈现：通过自然语言交互快速展示数据，帮助用户快速了解安全告警、风险资产、风险用户等综合态势。

对接安全产品和能力：已支持超过 100 款产品和能力对接，支持安全运营中常见多品牌、多类型安全产品的系统调度，实现对安全运 营任务的编排

和调度执行。

图 11 安星智能安全助手技术架构图

自动化分析，聚焦关键告警：整理两类溯源图，一是无重要线索的告警的溯源图代表攻击未成功或者误报，攻击者在受害者主机上没有执行操作，没有进一步线索；二是有线索证据链的告警的溯源图代表着攻击成功，在受害者主机上产生文件、进程相关的操作，留下线索。

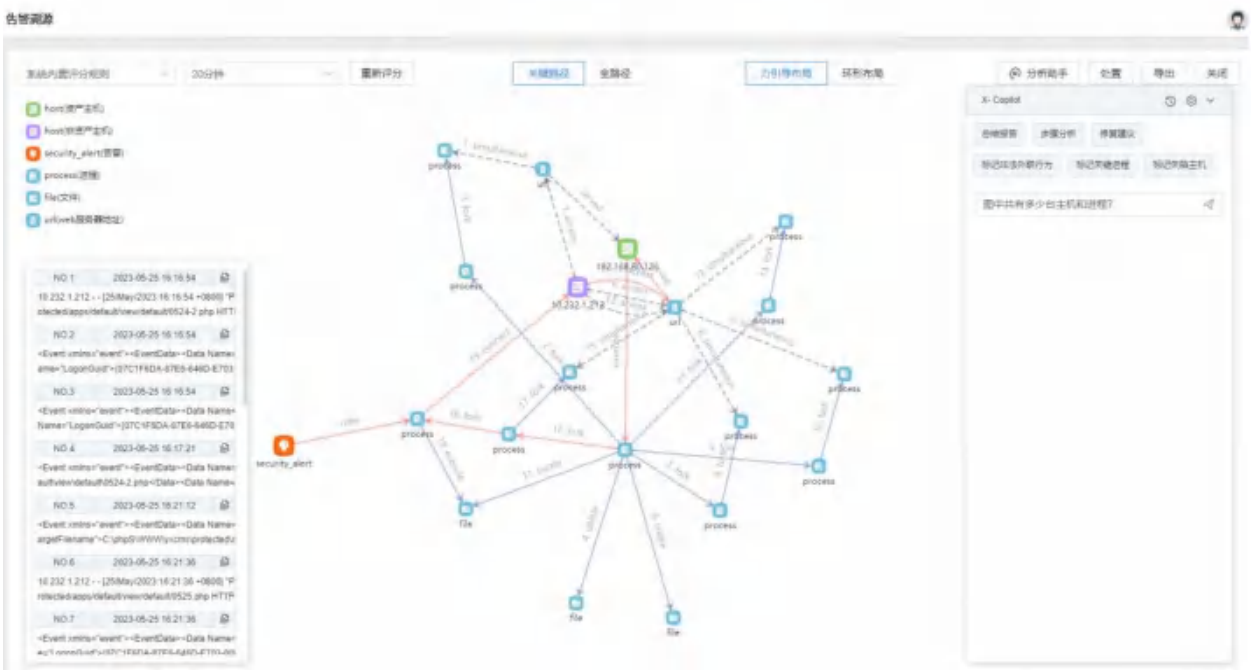

图 12 安星智能安全助手告警溯源图

脆弱性问题闭环管理：调用引擎扫描目标资产并获取漏洞、配置、弱口令等脆弱性，识别出网内脆弱性分布情况。根据最新漏洞情报匹配资产库中的资产系统类型、版本、端口等信息，找出可能存在漏洞相关的资产。根据历史扫描结果对比处置情况，对漏洞处置状态进行实时跟踪生成脆弱性总结报告，面向资产责任人派发处置工单

智能调度平台、安全产品和工具进行响应：具备挖矿病毒处理、勒索病毒处理、主机外发攻击事件处理、（失陷主机）、钓鱼邮件、蜜罐威胁狩猎攻击防护、一键找人找资产、ARP 劫持处置、DLP 联动数据库脱敏处置方案、堡垒机绕行处置、DDOS 攻击处置、用户运维权限管理自动化、异常 DNS 请求自动化分析、一键封堵等相关能力。

（3）建设成果

在某省级智慧城市项目测算，面向 3 万以上 IT 资产开展安全运营工作，安星智能安全助手可帮助客户每月节省人力 375 人天，每年节省 4500 人天。

行程如下特色成果：

集中调度：安全平台和设备集中调度，形成统一战线助力安全运营。

快速响应：通过智能安全运营系统快速并行响应海量安全运营工作。

全天候值班：实现 7×24 小时×365 天，不间断安全运营值班；以逸待劳，扭转攻防不利局面。

智能运营：通过智能运营任务处理安全运营日常工作，工作进展透明可见，协助用户跟踪闭环。

安全专家：提供常用安全分析和安全专业知识问答能力，缩小安全运营人员经验差距，提升安全运营团队整体作战实力。

8. 中国移动 AI+智慧城市安全展望

8.1. AI 让智慧城市更安全

8.1.1.完善法律法规和安全标准体系

目前我国现有的法律法规正在逐步完善，《互联网信息服务算法推荐管理规定》和《生成式人工智能管理办法》的颁布和实施，确保 AI 技术的安全、透明和道德，同时也在促进技术创新和保护个人隐私。

现有的法律法规对城市管理者在人工智能安全发展大方向上提供了重要指导，让智慧城市 AI 做到了有法可依，更好的组织开展相关领域的工作。智慧城市管理者应严格遵守国家及地方法律法规要求，落实智慧城市人工智能管理相关工作。

“智慧城市人工智能安全”的国家标准体系仍待完善，需加快推进相关标准规范的制定。国家层面应以构建人工智能全产业链安全为目标，做好个人数据保护、网络安全、AI 伦理等基础标准的制定，并持续推动智慧城市人工智能安全国标标准化工作，提升国家标准化水平。

8.1.2.推进技术发展，加强自主可控

智慧城市发展的背后，有着众多人工智能技术的支撑，从数据全生命周期管理到信息化系统建设，每一个环节都引入大量新兴技术。从城市发展的角度，AI 赋能网络安全、数据安全和应用安全，保障城市各类业务系统安全，让不同的服务可以平稳运行在智慧城市中，让城市更加高效。

同时随着 AI 技术的飞速发展，除传统网络安全技术之外，针对 AI 能力平台的安全技术也亟待发展，对于 AI 大模型技术，例如对抗性训练、模型审

计和验证、模型加密、数据掩码等安全技术的综合运用，可以有效的对 AI 技术的安全性进行提升。

对待日新月异的人工智能技术，我国应加大核心技术投入，鼓励科研机构和企业在自主可控的关键技术、产品上进行深度研究，推进国产化替代进度。

8.2. AI 让城市安全更智慧

8.2.1.强化运营管理水平，培养队伍

智慧城市人工智能建设过程中，会涉及大量如算力、数据、服务器等软硬件资产，相比于单纯的数据管理内容，人工智能的管理更为复杂，要最大化发挥这些资产的价值，则重在运营。

建立全面的城市管理系统，整合城市各个部门的数据并实施数据安全保护措施，以支持智能决策和优化城市运营。其次，推广智能交通系统和智能安防监控，提升城市交通管理和公共安全水平。

培训城市管理者和从业人员，提升其对人工智能技术应用的理解和运用能力，加强智慧城市治理水平。要强调的是，智慧城市人工智能运营管理需要与社会治理结合，搭建智慧城市治理体系，促进信息共享和跨部门协同，实现智慧城市的可持续发展。

8.2.2.完善 AI 安全体系与治理

我们会继续努力完善由 AI 驱动的智慧城市安全体系，并确保其成为公司日常运营的一部分，以便为 AI 基础设施和业务应用提供稳固的安全支撑。通

过明确各相关方的责任和角色，我们将全面打造 AI 安全治理，致力于满足合法、公正公平、可信赖的数据安全和可控可管的系统等安全目标。