ICS 35.030

中华人民共和国国家标准化指导性技术文件

GB/Z XXXXX—XXXX

网络安全技术 人工智能应用安全 应急管理

Cybersecurity technology—Artificial intelligence application security Emergency management

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布

国家市场监督管理总局
国家标准化管理委员会 发布

目　　次

前言 …… III
引言 …… IV
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　概述 …… 2
5　基本安全原则 …… 2
6　系统安全要求 …… 3
6.1　概述 …… 3
6.2　部署安全 …… 3
6.3　运行安全 …… 4
6.4　下线安全 …… 4
6.5　供应链安全 …… 5
7　应用场景安全要求 …… 5
7.1　概述 …… 5
7.2　执法监管类 …… 5
7.3　监测预警类 …… 6
7.4　指挥救援类 …… 6
7.5　政务管理类 …… 7
7.6　社会动员类 …… 7
8　安全管理要求 …… 8
8.1　概述 …… 8
8.2　应急准备 …… 8
8.3　监测预警 …… 8
8.4　应急响应 …… 9
8.5　复盘改进 …… 9
附录A（资料性）　应急管理行业人工智能应用安全事件典型案例 …… 10
A.1　数据投毒致森林火情漏报情景 …… 10
A.2　对抗样本致危化品泄漏误判情景 …… 10
A.3　模型篡改致救援路径偏移情景 …… 10
A.4　算法偏差致救援资源分配不均情景 …… 10

A.5 智能体失控与工具越权调用情景 …… 10
A.6 生成式人工智能幻觉致次生灾害研判错误情景 …… 11
附录B（资料性） 应急管理行业人工智能安全事件预警指引 …… 12
B.1 红色预警 …… 12
B.2 橙色预警 …… 12
B.3 黄色预警 …… 12
B.4 蓝色预警 …… 12

前　言

本文件为规范类指导性技术文件。

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由全国网络安全标准化技术委员会（SAC/TC260）提出并归口。

本文件起草单位：应急管理部大数据中心、中国科学院信息工程研究所、中国电子技术标准化研究院、北京神州绿盟科技有限公司、天翼安全科技有限公司、深信服科技股份有限公司、中国科学院计算技术研究所、华北计算机系统工程研究所（中国电子信息产业集团有限公司第六研究所）、清华大学、南京工业大学、北京天融信网络安全技术有限公司、工业和信息化部电子第五研究所、北京公安局人工智能安全研究中心、浪潮电子信息产业股份有限公司、中国信息安全测评中心、杭州安恒信息技术股份有限公司、启明星辰信息技术集团股份有限公司。

本文件主要起草人：宋宇宸、查达仁、刘丽敏、张卫伟、贾渊博、张晓伟、薛明、张妍婷、贾世杰、吕飞霄、郭少帅、冷锐、张广兴、邵慧泉、白燕妮、蓝志辉、王子涛、安鹏、吕孝礼、王龙、姜海洋、吕扬、刘北水、田小龙、刘文赛、李鑫泉、费凡芮、张锐卿、安高峰、万夕里、夏嘉远、杨欢、郭晋、郭洪铭、韩浩田、郑红霞、李勇、吴建英、韦国文、蒋发群、郭影起、黄冬秋、高功成。

引　言

本文件是人工智能应用安全系列文件中的行业领域文件，聚焦应急管理领域人工智能应用的特殊安全要求。通用系统安全、网络安全要求由系列文件中的通用文件或相关国家标准规定。本文件落实《人工智能安全治理框架 3.0》关于“促进重要行业规范应用”的部署，是该框架在应急管理领域的行业落地文件。

随着人工智能技术在应急管理相关业务中的持续应用，人工智能系统已逐步用于监测预警、风险分析评估、指挥协调、现场处置与恢复重建等关键环节。训练数据投毒、对抗样本攻击、模型篡改、算法偏差与幻觉、系统失控、数据泄露和供应链风险，可能影响预警准确性、救援公平性、指挥连续性以及人员和数据安全。

本文件围绕上述能通过技术和管理措施防范的风险，结合全生命周期与现阶段典型业务类别，给出基本安全原则、系统安全要求、应用场景安全要求和安全管理指引，并通过资料性附录提供典型事件情景案例。本文件不替代自然灾害、事故灾难、公共卫生等领域的专业应急预案、行政执法制度及其他人工智能安全标准化文件。

网络安全技术 人工智能应用安全 应急管理

1 范围

本文件给出了应急管理领域人工智能应用安全的基本安全原则、系统安全要求、应用场景安全要求和安全管理要求，并提供了应用系统分级、人工接管、部署运维、业务安全保护、智能体工具调用和安全测评的参考。

本文件适用于应急管理部门、承担应急管理职责的相关单位、科研院所、技术支撑机构、人工智能系统开发运营单位以及第三方评估机构开展人工智能应用的安全建设、安全运行、安全评估和应急管理。本文件所涉网络安全事件处置主要针对人工智能应用自身及其支撑环境，不替代通用网络安全事件应急响应制度，以及自然灾害、事故灾难、公共卫生等领域的专业应急预案、行政执法制度和其他人工智能安全标准化文件。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本（包括所有的修改单）适用于本文件。

GB/T 20985.1—2026 网络安全技术 网络安全事件管理 第1部分：原理和过程

GB/T 20986—2023 信息安全技术 网络安全事件分类分级指南

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 25069—2022 信息安全技术 术语

GB/T 28827.3—2012 信息技术服务 运行维护 第3部分：应急响应规范

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 37228—2025 安全与韧性 应急管理 突发事件管理指南

GB/T 39786—2021 信息安全技术 信息系统密码应用基本要求

GB/T 41479—2022 信息安全技术 网络数据处理安全要求

GB/T 42888—2023 信息安全技术 机器学习算法安全评估规范

GB/T 43697—2024 数据安全技术 数据分类分级规则

GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法

GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求

GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能应用安全事件 artificial intelligence application security incident

网络安全事件，以及由模型缺陷、输出错误、算法偏差、系统故障、误用或人工操作不当等引起的非网络安全事件

3.2

核心人工智能应用系统 core AI application system

支撑跨地区、全国性或涉及重大生命救援、重要应急资源调配等核心业务运行，一旦失效可能严重危害国家安全、社会稳定或公众生命安全的人工智能应用系统。

3.3

重要人工智能应用系统 important AI application system

支撑区域性、行业性重要应急管理业务运行，一旦失效可能危害区域或行业安全稳定、显著影响应急救援效率的人工智能应用系统。

3.4

一般人工智能应用系统 general AI application system

支撑本地、局部应急管理业务运行，失效通常仅影响局部业务能通过人工方式及时替代的人工智能应用系统。

3.5

智能体 AI agent

能感知环境、规划任务并调用模型、数据、应用程序或外部工具完成特定目标的人工智能应用实体。

4 概述

人工智能技术在应急管理领域的应用覆盖监督管理、监测预警、指挥救援、政务管理、社会动员等业务活动，涉及规划、设计、开发、测试验证、部署、运行、维护和退役。应用系统能通过模型、数据、工具接口、业务协同和智能装备形成不同自动化程度的业务应用，其安全风险可能沿数据、模型、组件、接口和云环境传导，并在极端应急条件下被放大。

本文件按照“基本原则—系统安全—场景应用—安全管理”的逻辑给出指引。第 5 章规定基本安全原则，第 6 章规定系统部署、供应链、运行和下线安全要求，并将智能体保护、业务协同与数据交换安全纳入运行安全，第 7 章按照监督管理、监测预警、指挥救援、政务管理、社会动员五类业务应用规定场景安全要求，第 8 章规定应急准备、监测预警、应急响应和复盘改进等安全管理要求。各单位可结合业务职责、系统类别和风险等级选择适用措施，并形成可检查、可追溯的记录。风险等级参照《人工智能安全治理框架 3.0》附件 1 确定。

5 基本安全原则

应急管理行业人工智能应用的基本安全原则如下：

a）人工智能应用宜遵循安全可靠、透明可追溯和人在回路原则，符合法律法规、应急管理政策和相关强制性标准要求，在极端应急条件下具备降级运行和人工替代能力。涉及预警发布、生命救援、资源分配、行政监管和智能装备控制等关键活动时，明确人工复核、确认、干预和接管责任，不应将人工智能输出作为高风险应急决策的唯一依据。

b）宜结合 3.2～3.4 的系统分类以及业务后果、影响范围、数据敏感度、决策自主程度和可替代性开展风险分析，确定系统类别和风险等级，并采取与风险程度相匹配的控制措施：高风险或核心系统采取严格的人在回路、复核会商、上线前专项安全测评、持续监测、人工接管和独立备份等措施，其他系统根据风险等级配置人工抽检、定期安全评估、权限复核、异常告警和降级运行等措施，应用场景、模型版本、部署方式或自主决策权限发生显著变化时，重新评估系统类别和风险等级。

c）宜对人工智能应用实施全生命周期安全管理，覆盖规划、设计、开发、测试验证、部署、运行、维护和退役阶段，明确各阶段的安全活动、责任人员和检查记录，数据分类分级保护、密码应用、训练数据与数据标注安全等符合 GB/T 43697—2024、GB/T 39786—2021、GB/T 45652—2025、GB/T 45674—2025 等标准的要求，部署、运行、维护和退役阶段的安全措施按照第 6 章实施。

d）宜明确人工智能应用安全管理的责任部门与责任人，建立业务、技术、数据、模型、云服务和供应商之间的责任清单，落实人工复核、工具调用审批、接管和恢复责任，并定期开展安全培训，提升人员风险意识和人工处置能力。

e）宜开展人工智能伦理审查，根据《人工智能科技伦理审查与服务办法（试行）》建立人机权责清单，明确人工智能输出建议、人员复核、授权决策和异常接管的责任边界。对高风险活动开展科技伦理和算法影响评估，关注算法偏差、公平性、透明度和弱势群体权益，必要时引入专家会商、异议复核和人工纠偏，超出模型适用边界或无法解释关键输出时，及时降低自动化等级或转由人工处置。

f）宜开展人工智能安全评估，评估时机覆盖上线前、重大变更后、定期检查和重大事件后，评估指标覆盖数据、模型、算法、系统、日志审计、供应链、备份恢复和应急响应，评估准则可综合采用本文件及 GB/T 42888—2023、GB/T 45654—2025 等适用标准，评估结果形成问题清单和整改建议，重大风险未整改前不扩大应用范围或提高自动化等级，评估活动不替代依法开展的等级保护测评、密码应用安全性评估及其他法定评估。

6　系统安全要求

6.1　概述

本章规定人工智能应用系统安全要求，涵盖部署、运行、下线和供应链安全，各单位可依据职责、业务风险和技术发展增加、合并或调整具体应用，并纳入相同的生命周期与风险管理过程。

6.2　部署安全

部署安全要求如下：

a）各单位宜依托主管部门认定的行业统一基础模型构建知识库和应用场景智能体，充分发挥其行业统一基础服务作用。核心或重要系统以及处理重要数据、涉密地理信息的服务，非必要情况下不使用其他大模型的模型底座。

b）敏感业务应部署于受控环境，采取专网运行、网络隔离和敏感业务数据不出域等措施，并核验访问控制、日志审计、备份恢复和人工接管能力，尽量避免使用公有云或混合云。

c）模型下沉至灾害现场终端、移动指挥车、无人机、视频监控等边缘设备时，宜支持断网弱网离线推理、模型签名、本地完整性校验和结果回传，恢复通信后开展数据一致性核验，防止旧数据覆盖现场最新态势。

d）采用多级协同部署时，宜与应急指挥体系相匹配，明确模型分级更新、数据分级汇聚、版本兼容、地方适配审核和回滚机制，保持统一安全基线。涉及跨区域、跨部门协同的，宜明确地方适配和安全验证责任。

e）应适配国产芯片、操作系统、数据库和推理框架，宜开展功能、性能、安全和兼容性测试，保持密码、访问控制、日志和备份机制有效。国产化适配不替代供应链评估和业务安全验证，涉及现场终端和移动指挥设备的，宜验证极端条件下可用性。

6.3 运行安全

运行安全要求如下：

a）宜对数据处理、模型运行、系统操作和智能体工具调用等关键活动进行日志记录，定期开展安全审计，日志留存期限符合适用法律法规和单位制度。宜结合监测预警、指挥救援、执法监管等业务类别监测模型有效性和运行风险，发现异常时及时校准、降级或停用，并保留用于事件复盘和责任界定的记录。

b）宜对模型接口实施身份鉴别、调用审计和异常调用检测，监控推理资源消耗，必要时启用限流、降级或弹性扩容，防范模型拒绝服务。预警发布、指挥调度等关键业务接口宜设置优先保障和故障隔离措施，避免影响应急指挥和救援通信。

c）生成合成内容用于预警、方案或公众信息发布时，宜对检索语料与知识库实施异常检测和来源复核，防范知识库投毒和系统提示词泄露，涉及预警发布和公众信息发布的，宜经授权人员确认。

d）宜为智能体分配唯一身份标识，明确功能边界、责任主体和授权范围。涉及指挥救援、执法监管等领域业务时，应对高风险操作分级授权并设置人工确认或复核，防范过度代理和未经授权的关键操作；涉及智能装备控制的，宜设置紧急停机、指令熔断和人工强制接管等硬阻断措施。

e）应防范提示注入、越狱攻击、恶意工具返回和敏感信息外发，对智能体输入输出和工具返回内容进行安全检测，必要时隔离执行环境，插件与工具接口实施输入校验、权限隔离和异常回滚，避免在系统提示词和工具描述中写入敏感规则、接口凭证和业务逻辑。

f）宜建立智能体行为审计、异常检测和熔断机制，记录任务目标、调用链、授权人和执行结果，行为偏离预设目标或出现不可控情况时，及时暂停并触发人工接管。人工接管宜与应急值守和指挥流程衔接，明确接管责任人、接管时限和恢复条件。

g）多智能体协同或跨业务域部署时，明确各智能体职责边界、协同规则和跨域任务审批责任，按业务边界划分数据源、工具和记忆访问范围。涉及行政执法、预警发布、资源调度和公众信息发布等输出，由相应业务授权人员确认，不因跨域自动流转省略人工复核，跨部门协同宜明确指挥关系和数据共享边界。

h）交换数据应携带来源、时间、质量和核实状态等必要信息，对来源冲突、更新延迟和异常数据进行提示，未经核实的信息不宜直接触发预警发布、资源调度或公众发布。涉及灾情速报、现场核查和电话直报信息的，宜标明核实状态并与现场核查结果关联。

i）断网、弱网或跨域接口中断时，保留本地缓存、人工交接和离线处置机制，恢复连接后开展数据一致性核验，防止旧数据覆盖最新业务状态。应急指挥链路和救援队伍通信优先保障宜作为硬约束，恢复通信后宜对回传结果开展来源鉴别、去重和冲突校验。

6.4 下线安全

下线安全要求如下：

a）应用系统退役或智能体停用下线时，应制定下线方案并评估对在用业务的影响，对数据、模型、密钥、日志和介质进行安全迁移或不可恢复销毁。涉及监测预警、指挥救援等业务的，宜提前明确替代系统、人工流程和应急保障措施。

b）使用云服务时，应核验租户数据及备份删除情况，完成模型、配置和日志的迁移或删除验证，保留退出交接记录。关键业务系统下线前宜验证替代流程在断网弱网和极端条件下的可用性。

c）下线涉及跨业务域数据共享和统一模型服务底座的，应同步通知相关业务系统，更新数据目录、接口清单和依赖关系，避免影响应急业务连续性。涉及预警发布、指挥调度等业务的，宜提前开展替代流程演练。

d）应保留下线审批、处置和验证记录，涉及安全事件的，完成事件调查和责任界定后再实施销毁。重大突发事件应对期间原则上不宜实施影响应急业务连续性的下线操作，确需实施的应履行审批并落实替代保障。

6.5 供应链安全

供应链安全要求如下：

a）对算力服务、算法模型、组件工具、智能体扩展组件、训练数据和知识库等供应来源开展安全评估，评估供应商安全能力、服务连续性和事件协同能力。涉及监测预警、指挥救援等关键业务的，宜重点评估极端灾害和突发事件期间的服务保障、应急联络和替代供应能力。

b）选择符合适用法律法规和国家相关安全标准的供应商，建立供应链安全应急机制，明确漏洞通报、服务变更、服务中断、数据删除和退出迁移责任。关键供应商宜提供应急联络渠道和临时保障方案，防范供应链中断影响应急业务连续性。

c）宜建立模型、数据集、组件和工具清单，记录来源、版本、依赖关系和完整性校验信息，对外部预训练模型、开源组件、第三方知识库及智能体扩展组件开展恶意代码、后门、漏洞和许可风险检查。

d）模型、数据、组件和接口更新前宜开展影响分析、安全测试和业务回归验证，保留可信版本和回滚能力。统一模型服务底座版本切换时，宜向相关业务系统提供变更通知、兼容性说明和灰度验证条件；涉及预警发布、指挥调度和执法监管业务的更新，宜选择业务低峰期并具备快速回退能力。

e）供应商或云服务退出时，宜完成模型、数据、配置、日志和密钥的迁移或删除验证，撤销接口凭证与工具权限，保留交接记录，防止服务依赖和残留访问影响应急业务连续性。关键业务宜提前准备替代模型、替代工具或人工处置流程。

7 应用场景安全要求

7.1 概述

本章按照执法监管、监测预警、指挥救援、政务管理、社会动员五类业务应用规定安全要求，各单位可依据职责、业务风险和技术发展增加、合并或调整具体应用，并纳入相同的生命周期与风险管理过程。

7.2 执法监管类

监管执法类应用场景安全要求如下：

a）建立企业基本信息、隐患、执法、事故和监测数据的可信接入、质量审核、更新和追溯机制，保持训练样本代表性、均衡性和时效性，对异常上报数据结合现场核查验证，定期纳入新工艺、新设备和新隐患类型数据开展适配验证。

b）使用经审核并记录效力状态的法律法规、标准和案例知识库，生成内容提供可定位的引用原文，失效或无法核验的内容不进入检查清单、执法文书和整改要求。

c）企业风险画像和分级监管建议说明主要数据来源、指标权重和风险变化原因，监管人员结合现场检查、企业申诉和专家会商校正结果，不应仅依据模型评分实施差别化监管。对裁量建议开展一致性、公平性和偏差检查，保留异议复核和人工校正渠道。

d）风险分级、执法文书、裁量建议和整改复核结论由具备相应职责的人员复核，模型输出不应作为行政执法的唯一依据，不应替代法定程序，不应自动作出行政处罚决定。隐患识别和整改复核关联原始影像、点位、时间和置信度，重大或低置信度结果经现场核查，不应仅凭单张图片自动通过。

e）对企业生产数据、工艺参数、案件材料、个人信息和现场影像实施分类分级、最小必要、脱敏和访问审计，记录模型版本、引用来源、人工修改和最终确认情况。

7.3 监测预警类

监测预警类应用场景安全要求如下：

a）对传感器、视频、遥感等数据源实施身份认证、完整性校验、时间同步和质量监测，建立数据投毒检测与过滤机制，对异常数据自动识别与隔离。

b）多源融合研判保留数据来源、时间戳、空间位置、质量标识和置信度，明确数据源权重、冲突处理和缺失数据降级规则，开展复杂天气、极端灾害和数据源中断等场景鲁棒性测试，低置信度、重大或来源冲突的预警结合现场核查和其他数据源确认。

c）视频识别和点位告警保留原始片段、设备位置、采集时间和模型版本，遥感变化检测标明数据来源和处理过程，重大或低置信度结果经人员回看或现场核查，定期评估误报、漏报并开展适配验证。

d）重大风险预警经人工确认和部门会签后发布，保留机器研判与人工判断不一致时的处置记录以及人工巡查、电话直报和离线告警通道。

e）对重要数据、涉密地理信息、人口分布和敏感目标实施分类分级、加密、脱敏和访问控制，对边缘监测设备实施固件加固、模型签名和本地完整性校验，模型性能衰退时及时再训练、校准或降级运行。

7.4 指挥救援类

指挥救援类应用场景安全要求如下：

a）人工智能输出作为接报、态势掌握、推演评估和调度决策的辅助参考，重大指挥决策、调度指令和装备控制由授权人员确认，推荐结果说明数据来源、约束条件、优先级和预计影响。

b）接报结构化关联原始记录并标明关键要素来源和置信度，摘要不自动删除伤亡、被困、危险物质等关键信息。态势摘要标明数据截止时间和来源，区分已核实事实、模型推断和待核查信息。

c）调度推荐展示队伍能力、道路、装备库存和预计到达时间等依据，开展资源分配公平性检查。救援路径结合实时数据动态复核并设置高风险区域硬约束，智能装备部署紧急停机、指令熔断和人工强制接管等硬阻断措施。

d）灾害趋势推演和影响评估与历史案例、专业仿真和现场核查比对验证，明确关键假设和不确定性，重大结论经专家会商。方案对比展示关键假设、约束条件和风险，不满足硬约束的方案不进入执行流程，危化品处置、恢复重建等专业方案经相应专业人员确认。历史案例和生成内容经事实性核验后使用，转移安置分析核验人口、容量、道路和物资数据并保障重点人群需求。

e）建立人工干预、一键接管、紧急停机和传统指挥机制，断网弱网下保留离线模型完整性校验、人工组网和传统通信手段，指挥链路和救援队伍通信优先保障作为硬约束，恢复通信后对回传结果开展来源鉴别、去重和冲突校验。

f）对指挥敏感数据、位置轨迹、人口和基础设施等信息实施加密、访问控制与审计，跨部门共享明确指挥关系、数据边界和脱敏要求，保留模型版本、人工调整、专家意见和最终决策记录，方案执行后开展复盘。

7.5 政务管理类

政务管理类应用场景安全要求如下：

a）对项目、预算、合同、人事、专家和档案数据实施分类分级、最小权限、脱敏和访问审计，向外部模型或第三方服务提供材料前进行必要性和脱敏评估。

b）使用经审核的制度、模板和业务规则库，记录版本和生效状态，规则更新后开展回归测试，模型输出提供来源和依据。

c）审核意见、专家遴选结果等政务输出由授权人员复核确认，不以模型输出替代法定程序或内部审批程序。回避关系、专业资质和评审数量限制等设置为硬约束，无法确认回避关系时转人工核查。

d）定期检查推荐规则的公平性、一致性和可追溯性，对专家个人信息和履职记录实施最小必要使用和访问审计，记录推荐依据、人工调整和最终确认情况。

7.6 社会动员类

社会动员类应用场景安全要求如下：

a）对公众信息、互动记录、灾情速报和社会救援力量信息实施最小必要采集、分类分级、脱敏和留存期限管理，使用他人形象、声音、作品或数据时核验授权。

b）政策解读、科普和灾情信息来源于经审核的权威材料，建立内容审核、专业复核和更新机制，多语种内容经具备相应能力的人员校核。

c）建立多源信息交叉验证和恶意注入识别机制，重大、敏感或高热度信息经人工复核，不应仅依据传播热度自动定性。灾情线索跨业务域共享时标明来源与核实状态，未经核实的信息不直接进入预警或指挥系统。

d）发布账号、素材和接口分级授权，生成合成内容的显式标识和隐式标识符合 GB 45438—2025，数字人内容和互动界面显著提示合成属性，不冒用真实人员身份。

e）设置互动问答边界和敏感问题转人工机制，无法核验或超出权威材料范围时明确提示不确定性，建立纠错、撤回、人工接管和应急发布机制，重大或敏感内容经人工确认后发布。

8　安全管理要求

8.1　概述

本章按照应急准备、监测预警、应急响应和复盘改进规定安全要求，各单位可依据职责、业务风险和技术发展增加、合并或调整具体应用，并纳入相同的生命周期与风险管理过程，需与应急值守、预案体系、应急联动机制和现场指挥流程相衔接。

8.2　应急准备

应急准备安全要求如下：

a）预案制定：相关单位宜针对人工智能应用安全事件制定专项应急预案，明确不同级别事件的响应流程、责任分工、处置措施、人工接管和恢复条件。预案宜与本单位、本行业总体应急预案相衔接，并参照GB/T 20985.1—2026、GB/T 28827.3—2012建立事件管理和响应机制。涉及监测预警、指挥救援、执法监管、政务管理和社会动员的，宜明确现场指挥部与远程指挥、业务部门与安全部门之间的协同关系。

b）组织与队伍建设：宜建立人工智能应用安全事件响应小组，明确业务负责人、安全负责人、模型与数据技术人员、云服务或供应商联络人以及人工接管负责人的职责。应急小组宜与应急值守和指挥体系衔接，定期组织极端条件、模型失控、工具越权、断网弱网和指挥链路中断等情形的应急演练，提升协同处置能力。

c）备份与恢复：宜建立数据、模型、配置、知识库和系统的备份机制，定期验证备份的完整性、可用性和恢复时效。备份数据宜采用合规密码技术加密存储并校验完整性，防范备份被攻击、篡改或不可用。核心和重要系统宜具备离线备份或异地备份能力，保障极端灾害条件下关键业务可恢复。

d）资源准备：宜储备应急处置所需的技术工具、算力资源、通信资源、可信模型版本和人工替代资源，保障处置过程中的资源需求。涉及应急指挥和现场救援的，宜准备移动指挥终端、离线模型、传统通信手段和人工处置力量。

8.3　监测预警

监测预警安全要求如下：

a）风险监测：宜建立覆盖数据、模型、系统和业务的全链路安全监测机制，识别数据篡改与投毒、对抗攻击、模型失控、网络攻击、服务中断以及业务输出异常等风险，并关联模型、数据、接口和人工操作记录。涉及监测预警和指挥救援业务的，宜重点监测预警准确性、指挥连续性、救援公平性和关键输出异常。

b）预警分级：针对应急管理领域人工智能安全事件划分事件等级，预警级别采用红色、橙色、黄色、蓝色四级标示，参照国家突发事件预警分级体系确定。情况紧急时可先行发布、后续复核，风险尚未形成事件但可能快速演化的可提前预警，预警级别宜由业务、安全和技术人员结合监测证据会商确认。

c）预警发布：监测到风险后，宜根据预警级别及时向相关负责人、业务人员、应急值守人员和人工接管负责人发布预警信息，内容包括风险类型、级别、影响范围、模型与数据版本、处置建议、复核时限和反馈渠道。涉及跨部门、跨区域应急联动的，宜同步通报相关联动单位。

8.4 应急响应

应急响应安全要求如下：

a）事件上报：发生人工智能应用安全事件后，发现人员宜立即向事件响应小组上报，事件响应小组根据事件级别和本单位制度向上级主管部门报告。属于网络安全事件的，报告时限和报告程序按照《国家网络安全事件报告管理办法》执行。涉及预警发布、指挥救援等核心业务的，宜同步报告业务主管部门和应急指挥机构。

b）应急处置：根据人工智能应用安全事件类型和预警级别，采取止损隔离、人工接管等措施进行处理，并按照分级响应要求调整指挥层级、处置范围和恢复条件。处置过程中宜优先保障应急指挥链路、救援队伍通信和关键业务运行，必要时暂停自动控制并转由人工处置。

c）协同处置：对于跨部门、跨区域的人工智能应用安全事件，宜建立协同处置机制，联动网络安全、业务主管、公安、云服务提供者和其他相关单位，共同开展事件处置，并明确指挥关系和数据共享边界。协同处置宜与现有应急联动机制衔接，避免多头指挥和信息冲突。

d）响应终止与恢复：引发事件的攻击、异常或故障已得到控制，风险不再扩散且残余风险处于可接受范围时，可考虑终止响应。事件响应小组宜提出终止建议，由业务、安全和技术人员共同确认，并按照响应启动权限履行确认程序。恢复前宜验证模型、数据、接口和人工接管能力，逐步恢复自动化等级，并保留恢复过程和验证记录。

8.5 复盘改进

复盘改进安全要求如下：

a）事件复盘：响应终止后，宜组织开展事件复盘，分析事件原因、人工接管与恢复效果、处置过程中的问题，总结经验教训。涉及监测预警和指挥救援业务的，宜重点评估对预警准确性、指挥连续性、救援公平性和业务连续性的影响。

b）优化改进：根据复盘结果，对人工智能应用进行优化，包括修复安全漏洞、优化模型、完善数据安全机制等。同时宜对应急预案、处置流程、人工接管方案和降级运行方案进行更新完善，必要时重新评估系统类别、风险等级和自动化等级，提升应对同类事件的能力。

c）信息共享：将事件情况、处置经验上报相关主管部门，参与行业内的安全信息共享，提升行业整体的安全应急能力。涉及跨区域、跨部门协同处置的，宜形成可复用的案例和处置指引。

附 录 A
（资料性）
应急管理行业人工智能应用安全事件典型案例

本附录案例为依据典型风险归纳的情景示例，不对应特定真实事件，用于说明风险发现、人工接管、响应恢复和改进方法。

A.1 数据投毒致森林火情漏报情景

某单位使用多源遥感和视频数据训练森林火情识别模型，攻击者通过受控数据源注入错误标注的烟雾样本，导致模型将真实早期烟雾判为云霾并漏报。值守人员发现现场直报与模型结果持续不一致后，启动人工复核和事件上报，暂停受影响数据接入，切换至经验证的历史模型和人工巡查。事后通过来源追溯、异常样本隔离、模型回滚和增量安全测试恢复服务，并完善数据投毒检测、多源交叉验证和低置信度告警。

A.2 对抗样本致危化品泄漏误判情景

某危化品园区的气体监测与视频融合模型受到刻意构造的输入扰动影响，将局部泄漏特征误判为正常工况。现场人员依据便携式检测仪和人工巡检确认异常，立即隔离模型输出、启动人工预警并按既有预案处置。恢复前开展对抗样本测试、传感器交叉校验和输入完整性检查，将人工仪表读数设为高风险告警的独立核验来源。

A.3 模型篡改致救援路径偏移情景

某现场救援终端在弱网环境下运行，离线模型文件被未授权替换，导致推荐路径偏向未核实区域。安全校验发现模型签名和版本不一致后，终端停止自动路径推荐，现场指挥员切换至纸质预案和人工导航。技术人员从可信备份恢复模型并验证密码杂凑值与数字签名，小范围试运行后恢复辅助功能，并补充离线完整性校验、版本白名单和定期巡检。

A.4 算法偏差致救援资源分配不均情景

某资源调度模型主要依据历史出动记录推荐优先级，对历史数据较少的偏远地区和特殊人群需求权重不足，造成资源分配建议不均衡。指挥人员会商发现偏差后，人工调整优先级并记录纠偏依据。后续通过补充代表性样本、公平性测试、最低保障约束和异议复核机制改进模型，明确高风险资源分配由授权人员最终决策。

A.5 智能体失控与工具越权调用情景

某应急信息智能体在处理恶意构造的外部文档时受到提示注入影响，尝试调用超出任务范围的数据导出和消息发布工具。权限策略拦截部分调用，审计告警触发后，管理人员暂停智能体、撤销会话凭证并切换人工信息发布。事后通过工具白名单、参数校验、高风险操作人工确认、网络隔离和调用频率限制降低风险，并开展回归测试和应急演练。

A.6 生成式人工智能幻觉致次生灾害研判错误情景

某辅助决策系统生成灾害链推演方案时，虚构不存在的上游设施并给出不符合现场约束的处置建议。专家复核发现事实错误后，阻止方案进入指挥流程，保留人工研判和传统仿真工具。后续引入权威知识库检索、事实性校验、物理约束检查、低置信度拒答和专家会商，并按要求设置生成合成内容显式与隐式标识，防止未确认输出被误用。

附 录 B
（资料性）
应急管理行业人工智能安全事件预警指引

人工智能应用安全事件预警是针对异常征兆、风险演化或事件发展趋势发出的前置警示，不等同于事件定级。预警级别采用红色、橙色、黄色、蓝色四级标示，参照国家突发事件预警分级体系确定，按各判定指标中的最高风险等级确定，由业务、安全和技术人员结合监测证据会商确认。情况紧急时可先行发布、后续复核。风险尚未形成事件但可能快速演化的，可提前预警。

B.1 红色预警

人员生命安全受到严重威胁，核心人工智能应用系统（见 3.2）或跨地区、全国性应急业务严重受损，影响快速扩散且现有措施难以控制。宜立即启动最高级别应急响应，按规定向单位负责人和主管部门报告，暂停或隔离受影响应用及相关工具、数据和接口，实施人工接管和跨部门协同，未经重新授权和安全验证不恢复自动控制。

B.2 橙色预警

重要人工智能应用系统（见 3.3）或区域性、行业性业务受到严重影响，影响已向多个部门或业务环节扩散，但经人工接管、隔离等措施尚可控制。宜启动部门或区域联动响应，限制非必要功能和外部工具调用，隔离异常数据、模型或接口，对关键输出逐项人工复核，根据研判结果决定是否升级。

B.3 黄色预警

局部业务或部分功能受到较大影响，影响范围相对明确，通过跨岗位协同和局部降级可以控制。宜由业务与安全人员联合处置，暂停受影响功能，核验输入数据、模型版本和输出结果，启用备用流程并加强监测，处置无效或影响扩大时升级。

B.4 蓝色预警

出现局部、轻微异常或风险征兆，尚未造成明显业务损害，通过人工复核、参数调整或局部修复即可控制。宜由运行人员核实并记录，开展局部修复和持续观察，确认风险消除后解除，异常重复出现、影响扩大或无法控制时升级。