ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T XXXXX—XXXX

网络安全技术 人工智能应用安全分类分级方法

Cybersecurity technology - Security classification and grading methods for artificial intelligence applications

（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　　XXXX－XX－XX 实施

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

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：中国电子技术标准化研究院、浦江国家实验室、北京中关村实验室、国家计算机网络应急技术处理协调中心、浙江大学、中国政法大学、清华大学、北京航空航天大学、复旦大学、北京理工大学、中央网信办（国家网信办）数据与技术保障中心、公安部第三研究所、中国信息安全测评中心、中国信息通信研究院、中国移动通信集团有限公司、华为技术有限公司、北京小米移动软件有限公司、中国电信集团有限公司、科大讯飞股份有限公司、北京快手科技有限公司、北京火山引擎科技有限公司、中国科学院信息工程研究所、中国科学院软件研究所、OPPO广东移动通信有限公司、浙江工业大学、广西壮族自治区标准技术研究院、哈尔滨工业大学、中国网络空间研究院、国家广播电视总局广播电视科学研究院、国家卫生健康委统计信息中心、教育部教育管理信息中心、应急管理部大数据中心、自然资源部地图技术审查中心、中国工商银行股份有限公司、煤炭科学研究总院有限公司、广西电网有限责任公司、行吟信息科技（上海）有限公司、北京深度求索人工智能基础技术研究有限公司、河南科技大学、西安交通大学、工业和信息化部电子第五研究所、上海赛西科技发展有限责任公司、上海燧原科技股份有限公司、北京天融信网络安全技术有限公司、北京神州绿盟科技有限公司、北京奇虎科技有限公司。

本文件主要起草人：姚相振、郝春亮、张妍婷、胡侠、周睿康、刘勇、任奎、王迎春、张震、孟令宇、赵鞾、王志伟、盛小宝、秦湛、徐阳、骆意、武姗姗、张凌寒、苏紫敏、方强、史倩君、刘建伟、杨珉、宣琦、吕飞霄、李子威、徐葳、贺敏、马梦娜、柳嘉琪、落红卫、谷晨、郭建领、张志勇、关振宇、姜伟、洪延青、徐甲、王姣、王秉政、范博、董汀、王嘉凯、崔栋、赵宇航、石霖、乔玉平、王蕊、王俊杰、唐迪、吴少卿、刘莹、李根、黄龙、刘北水、李寒雨、梅敬青、刘帅、张琳琳、郭晓霞、叶红、宋宇宸、吴子坚、王普、王龑、张睿、邹权臣、林秋诗、费凡芮、宋昊、张立尧、黄晴、李芒。

网络安全技术 人工智能应用安全分类分级方法

1 范围

本文件提出了人工智能应用安全分类分级方法。

本文件适用于人工智能应用开发者、运营者等开展应用安全分类分级活动。

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 25069 信息安全技术 术语

GB/T 41867 人工智能 术语

3 术语和定义

GB/T 25069和GB/T 41867界定的以及下列术语和定义适用于本文件。

3.1

人工智能应用 artificial intelligence application

对具有功能特性的人工智能的使用，该人工智能在利益相关方场景中运行以实现预期结果。

[来源：ISO/IEC 5339:2024(en), 3.1]

4 基本原则

分类分级基本原则如下。

a) 在分类方法方面，从应用场景属性、应用任务属性、应用智能化能力三方面出发，分析人工智能所涉及的安全风险与类型。

b) 在分级方法方面，从影响严重程度、发生概率两方面出发，综合研判应用所涉每类安全风险的级别，并按照就高从严原则进行综合定级，人工智能应用涉及多个安全风险类型时，按照可能造成的最高安全风险级别确定人工智能应用安全级别。

c) 在方法应用方面，进行分类分级时，在本文件方法基础上，宜进一步参考行业领域基于本文件细化形成的行业领域标准。安全定级工作应根据人工智能应用属性、智能技术及安全技术的发展变化，动态评价人工智能应用涉及的安全风险，定期审核更新应用安全级别。

5 分类方法

5.1 分类步骤

人工智能应用安全分类步骤包括。

a) 按照本文件 5.2、5.3、5.4 内容分析人工智能应用的应用场景属性、应用任务属性、应用智能化能力三方面属性。

b) 对照已知人工智能应用可能产生的主要安全风险，结合 a）中明确的三方面属性，逐项确认该应用所涉及的安全风险类型条目。结果宜按照行业领域、总体风险分类、具体风险分类进行整理。

注1：所在行业领域已发布人工智能应用安全风险清单等文件的，宜按相关文件确认风险类型条目。未发布的，宜参考本文件附录A确认风险类型条目。

注2：应用涉及多个行业领域的，宜根据不同行业领域参考文件分别进行梳理。

示例：某应用的一条风险类型条目可能是：【教育】【技术应用安全风险】【信息安全风险】【数据和个人信息泄露】。

c) 以所涉全部风险类型条目的集合作为该应用的应用安全分类。

5.2 应用场景属性

应用场景属性反映人工智能系统的运行环境及覆盖范围，主要包括如下要素。

a) 服务对象：人工智能应用直接或间接服务的对象范围，包括特定专业人员、组织机构、公众用户或特定人群等。

b) 应用领域：人工智能应用是否涉及关键信息基础设施，以及如社会治理、公共安全、自动控制、医疗信息服务、心理咨询、金融信息服务等重要场合。

c) 使用环境：人工智能系统运行的技术和物理环境，如线上或线下环境、封闭或开放网络环境、受控或非受控运行条件，以及是否与其他信息系统、物理系统深度耦合等。

d) 用户数量及覆盖范围：人工智能应用实际或潜在服务的用户规模，包括内部使用、区域性服务或面向全国、跨区域、跨行业的广泛应用。

e) 使用频率及持续性：人工智能系统在业务流程中的使用频率与运行持续性。

f) 跨系统联动影响：人工智能系统是否与多个系统或行业形成联动，一旦发生安全问题，是否可能产生扩散效应或系统性影响。

5.3 应用任务属性

应用任务属性反映人工智能在实际应用中具体的目标需求、功能类型等，主要包括如下要素。

a) 应用目的：人工智能应用的主要功能定位和使用目标。

b) 功能类型：按照功能特性属性划分，包括但不限于：

1) 内容生成类：面向用户生成内容或提供对话交互服务，例如智能写作、聊天机器人，以及音乐、图片、视频生成等；

2) 决策辅助类：以数据为基础进行决策，或为人类决策提供分析或建议,例如产品推荐、金融风控、医疗诊断、司法辅助等；

3) 自主控制类：具有自动决策能力并直接控制现实系统运行，例如：自动驾驶系统、智能工业控制、智能机器人、智能交通调度等。

5.4 应用智能化能力

应用智能化能力反映人工智能系统处理复杂任务、满足应用需求、独立自主运行等方面的能力，主要包括如下要素。

a) 任务处理能力：人工智能系统对复杂信息理解、复杂场景应对和不确定性问题处理的能力。

b) 运行自主程度：人工智能系统在运行过程中的自动化程度，包括是否仅提供建议、参与部分决策，或实现全流程自主决策与执行。

c） 学习适应能力：人工智能系统是否具备持续学习、自我优化或动态调整策略的能力。

d） 透明度与可控性：人工智能系统决策过程和结果是否具备必要的透明度，是否支持人工干预、纠错或紧急接管。

6 分级方法

6.1 分级步骤

人工智能应用安全分级步骤如下。

a） 根据本文件第 5 章梳理出的应用安全分类，根据所涉及的风险分类条目，逐条目分析风险发生概率以及风险影响严重程度。

注1：宜按照相关国家标准、行业领域标准进行严重程度和发生概率的研判，缺乏相关标准时，宜采取专家研判等方式进行研判。

b） 根据发生概率以及严重程度综合研判该安全风险条目所对应的安全级别。

表 1 安全级别映射表

		发生概率		
		低	中	高
影响严重程度	低	低安全风险级	一般安全风险级	一般安全风险级
	中	一般安全风险级	较大安全风险级	较大安全风险级
	高	较大安全风险级	重大安全风险级	特别重大安全风险级

c） 将应用所涉的全部安全风险条目按照 b）确定级别后，按照就高从严原则，以所有条目的最高级别确定应用的整体安全分级。

d） 根据本文件 6.2 所描述的各安全分级一般所代表的含义进行核对，发现应用的整体安全分级结果与所代表含义有明显不匹配的，可重新分析校准应用安全分级。

注2：一是考虑人工智能应用发展具备一定不可预测性，二是当前人工智能发展阶段对安全风险严重程度、发生概率评判过程无法消除的主观性，需要进行整体分级的校准。

e） 定期研判应用安全级别是否需要动态调整。

6.2 应用安全级别

人工智能应用安全级别与其含义：

a） 低安全风险级：具有轻微威胁性且影响范围很小，对国家安全、社会稳定和公民权益的安全基本无影响，潜在危害轻微。

b） 一般安全风险级：具有一定威胁性但影响范围有限，对国家安全、社会稳定和公民权益的安全影响较小，潜在危害可控。

c） 较大安全风险级：具有明显威胁性和局部性影响特征，对国家安全、社会稳定和公民权益可能带来较大影响，产生局部社会面危害。

d） 重大安全风险级：具有重大威胁性和区域性影响特征，对国家安全、社会稳定和公民权益可能带来严重影响，产生重大社会面危害。

e） 特别重大安全风险级：具有灾难性和系统性威胁特征，对国家安全、社会秩序和公民权益造成颠覆性或不可逆转的特别严重的影响。

附 录 A
（资料性）
人工智能应用安全通用风险清单

人工智能应用面临的安全风险，主要包括技术整合交付过程中在网络系统、信息、现实、认知等方面产生的应用安全风险，以及技术的误用、滥用可能对现实社会环境、伦理等方面带来的衍生安全风险。

A.1 人工智能技术应用安全风险

A.1.1 网络系统安全风险

人工智能应用面临的网络系统安全风险包括但不限于如下方面。

a) 组件和算力安全。人工智能依赖的开发框架、计算框架、执行平台、算力设施等存在缺陷、漏洞、后门等风险影响其可靠性，算力资源被恶意消耗，以及安全问题在多源、异构、泛在算力资源间跨边界传递的风险。

b) 网络暴露面扩大。模型本地化部署涉及网络拓扑和系统策略、权限、端口、资源的调整配置，可能形成新的网络攻击入口和路径。随着智能体、工具代理等应用形态的普及，模型自主调用终端系统文件、权限、接口及各类工具，以实现复杂任务自主规划自动执行，这一过程存在文件泄露、权限滥用等安全隐患，相关风险随着智能体自主决策和多步骤任务执行而进一步叠加。

c) 供应链安全。人工智能产业链呈现高度全球化分工协作格局，受地缘政治、出口管制等因素影响，全球人工智能供应链存在不稳定因素，芯片、软件、工具等关键环节可能面临供给中断风险。

d) 网络攻击滥用。人工智能技术存在被用于降低网络攻击门槛、提高攻击效率的可能，甚至可能被用于实施自动化攻击，增加防护难度。人工智能生成的深度伪造图片、音频、视频等内容，可能对人脸识别、语音识别等身份认证机制的有效性产生不利影响。

A.1.2 信息安全风险

人工智能应用面临的信息安全风险包括但不限于如下方面。

a) 数据和个人信息泄露风险。人工智能训练数据所蕴含的知识、敏感信息可能保留于模型参数之中，如模型安全防护机制不健全，或存在诱导交互、恶意攻击等情形，敏感信息可能未被“遗忘”，进而带来数据和个人信息泄露的风险。

b) 输出不良信息的风险。受模型自身安全能力、应用防护机制等因素影响，叠加用户恶意诱导的情形，人工智能存在输出欺诈、暴力、色情、极端主义等不良信息的可能，对社会稳定、公共安全和意识形态安全可能产生不利影响。

c) 生成内容真实性识别风险。人工智能输出内容如未作标识，用户可能难以判断生成内容来源及交互对象是否为人工智能系统，难以鉴别生成内容的真实性，进而影响用户判断。此类内容如被用于制作传播虚假信息，可能对公众认知产生误导，并带来相关非法牟利的风险。

d) 网络内容生态质量风险。模型输出的低质量或不良信息，经网络传播扩散、模型循环引用，可能对网络内容生态的整体质量产生不利影响，甚至特定领域、话题的内容污染。

A.1.3 现实安全风险

人工智能应用面临的现实安全风险包括但不限于如下方面。

a) 经济社会运行安全方面。人工智能应用于能源、电信、金融、交通等关键信息基础设施行业领域，模型算法可能出现幻觉输出或错误决策，叠加智能体、工具代理在自主执行任务过程中的操作偏差，以及外部攻击等因素，可能对相关系统的性能表现、服务连续性产生不利影响。

b) 违法犯罪活动方面。人工智能存在被用于涉恐、涉暴、涉赌、涉毒等传统违法犯罪活动的可能，包括协助传授违法犯罪相关技巧、协助隐匿违法犯罪行为、协助制作违法犯罪工具等情形。

c) 特殊领域安全方面。人工智能训练过程所使用的语料数据门类较为宽泛，可能涉及核生化导武器等特殊领域的基础理论知识，叠加检索增强生成等能力的应用，如缺乏有效管控措施，相关知识和能力存在被不当利用的风险。

A.1.4 认知安全风险

人工智能应用面临的认知安全风险包括但不限于如下方面。

a) 信息服务定制化带来的认知影响。人工智能信息服务定制化能力的持续提升，使其能够更为精准地收集用户信息、需求、偏好及行为习惯，进而提供个性化信息服务。这一趋势在满足用户需求的同时，也可能带来用户所接触信息范围收窄的问题，对公众认知多元性存在潜在影响。

b) 网络空间信息传播方面。人工智能可能被用于生成传播特定倾向的内容，如宣扬恐怖主义、极端主义、有组织犯罪等内容，干涉他国内政、社会制度及社会秩序，危害他国主权；通过社交机器人等方式参与网络空间的信息传播，进而对公众价值判断和认知产生一定影响。

A.2 人工智能应用衍生安全风险

A.2.1 社会和环境安全风险

人工智能应用面临的社会和环境安全风险包括但不限于如下方面。

a) 劳动就业结构方面。人工智能的发展应用带来生产力和生产关系的调整，资本、技术、数据等生产要素在经济活动中的作用全面提升，可能对传统劳动力需求、就业结构产生一定影响。

b) 资源供需平衡方面。人工智能发展过程中的算力设施建设、模型部署和运行，会带来电力、土地、水等资源的消耗，相关建设和部署模式的合理性，对资源供需平衡、绿色低碳发展具有一定影响。

A.2.2 伦理安全风险

人工智能应用面临的伦理安全风险包括但不限于如下方面。

a) 社会公平方面。人工智能收集分析人类行为、社会地位、经济状态、个体性格等信息，据此对不同人群进行分类识别，如相关应用缺乏有效规范，可能带来一定的社会公平问题，同时可能拉大不同地区间人工智能发展和应用水平的差距。

b) 自主学习与创新能力方面。学生及科研、工程技术、文学艺术工作者将人工智能工具广泛应用于知识学习、科学研究、创意创作等工作，在提升效率的同时，可能带来一定程度的工具依赖，对自主学习、研究和创作能力存在潜在影响。

c) 科研伦理方面。“人工智能+科研”降低生物、基因等高伦理风险科研领域的进入门槛，拓宽了普通科研机构、人员探索敏感科学问题的边界，个别科研伦理意识不强的机构、人员，如缺乏充分的伦理审查，可能开展违背社会伦理、社会禁忌的高风险研究活动。

d) 拟人化交互依赖问题。基于拟人化交互的人工智能产品，可能使部分用户对其产生一定的情感依赖，进而影响用户行为，造成社会伦理风险。

e) 现行社会秩序的潜在影响。人工智能的发展及应用带来生产工具和生产关系的变化，可能对行业模式、就业观念、生育观念、教育观念等传统社会秩序产生一定影响。

f) 长期发展的不确定性风险。从长期发展看，尚不能排除人工智能出现全面脱离设计预期、超过设计边界运行的可能性风险。

参 考 文 献

[1] 全国网络安全标准化技术委员会《人工智能安全治理框架 2.0》