人工智能安全标准体系

（V1.0）

（征求意见稿）

全国网络安全标准化技术委员会秘书处

2025 年 01 月

目录

一、体系架构

人工智能安全标准体系旨在支撑落实《人工智能安全治理框架》（以下简称“《框架》”），围绕《框架》中明确的模型算法安全、数据安全、系统安全三类内生安全风险，以及网络域、现实域、认知域、伦理域四类应用安全风险，系统梳理了可帮助防范化解相关人工智能安全风险的重点标准，同时与网络安全国家标准体系进行有效衔接，加强人工智能安全标准工作顶层设计，以科学、合理的标准布局前瞻应对各类风险挑战，促进人工智能技术及应用健康发展。体系内各项标准与《框架》中各类风险的映射关系见附件1。

人工智能安全标准体系主要由基础共性、安全管理、关键技术、测试评估、产品与应用等5个部分组成，体系框架如图1所示。

1、基础共性类标准是以标准工作支撑落实《框架》的重要保障，主要规范了人工智能安全术语定义、分类分级、通用要求、参考架构等方面内容，是人工智能安全的基础性、总体性标准。

2、安全管理类标准围绕《框架》中明确的模型算法安全、数据安全、系统安全三类内生安全风险，以及在人工智能系统开发、应用、运行、维护等生命周期各环节面临的安全风险，提供了覆盖全过程全要素的安全管理标准。

3、关键技术类标准紧扣人工智能相关技术发展情况，主要规范了生成式人工智能安全、智能体安全、具身智能安

全、多模态安全、生成合成安全、安全对齐、安全围栏等方面内容，为人工智能技术健康发展保驾护航。

4、测试评估类标准主要规范人工智能安全能力测试、模型安全性测试、产品服务安全测试、场景应用安全测试、安全测试基准等方面内容，以测试评估工作帮助提升人工智能安全水平。

5、产品与应用类标准主要规范个人应用、重点行业应用等方面内容，保障人工智能技术在各行业、各领域的安全应用。

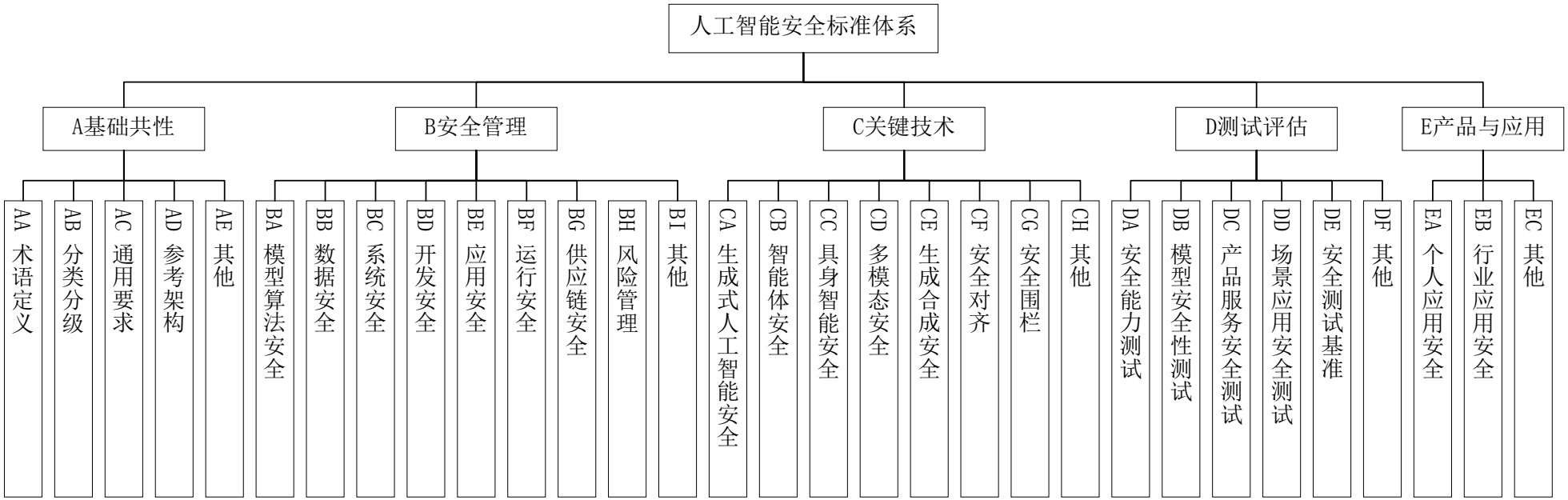

图 1 人工智能安全标准体系框架图

二、重点领域

（一）基础共性

基础共性类标准是推动人工智能安全标准体系建设、落实《框架》各项措施的重要保障，是人工智能安全的基础性、总体性标准，包括术语定义、分类分级、通用要求、参考架构等研制方向。基础共性标准子体系如图 2 所示。

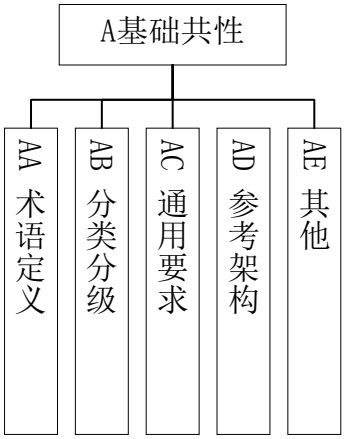

图 2 基础共性标准子体系

1、术语定义。规范人工智能安全相关的概念定义，明确人工智能安全的内涵及范畴，有助于统一行业理解，支撑人工智能安全相关标准的研制。

2、分类分级。规范人工智能应用安全分类分级的基本原则、框架、流程，以及分类方法、分级方法等，并给出人工智能应用安全分类分级参考示例。

3、通用要求。针对人工智能常见安全风险，总结跨领域、跨场景人工智能技术研发应用的共性规律，提出人工智能安全通用要求，解决人工智能安全治理措施碎片化局面以及整体性、系统性、协同性不足的问题。

4、参考架构。规范人工智能研发全生命周期中，基础供应者、技术支持者、服务提供者、服务使用者等相关角色的安全职责，以及在人工智能落地应用时，相关工具、插件、环境、知识库的安全要求。

（二）安全管理

安全管理类标准围绕《框架》中明确的模型算法安全、数据安全、系统安全三类内生安全风险，以及人工智能开发、应用、运行、维护等各环节面临的安全风险，提供了覆盖全过程全要素的安全管理标准，包括模型算法安全、数据安全、系统安全、开发安全、应用安全、运行安全、供应链安全、风险管理等研制方向，基础支撑标准子体系如图 3 所示。

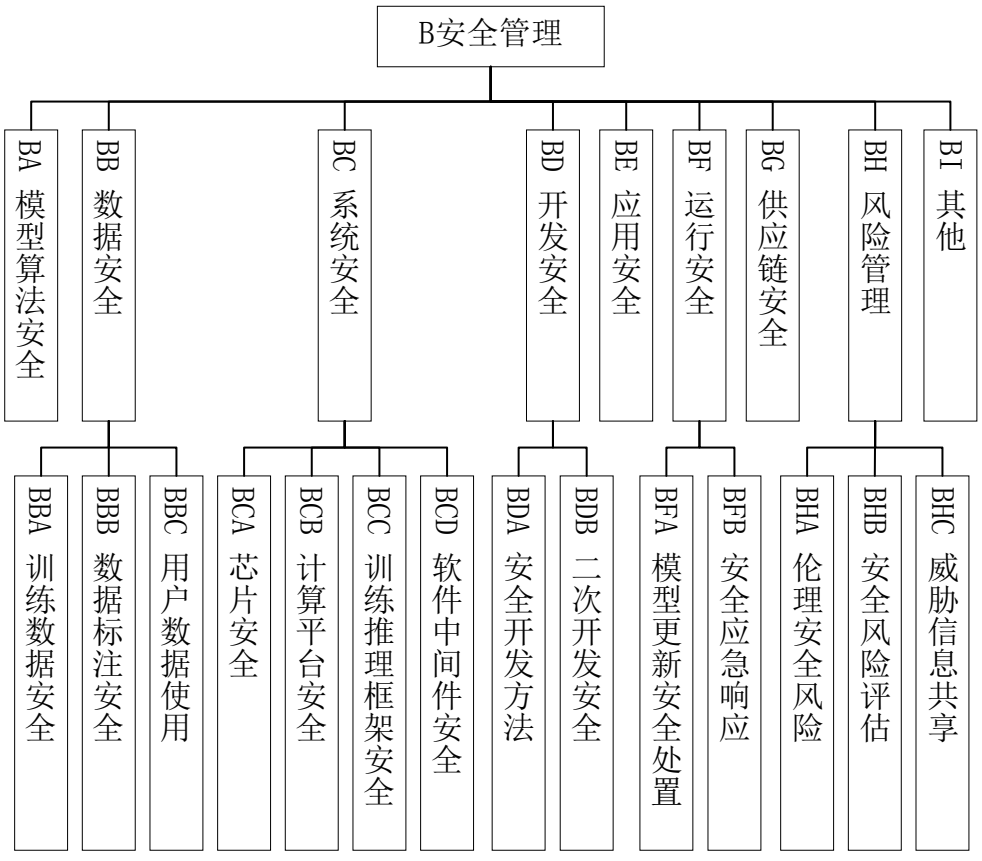

图 3 安全管理标准子体系

1、模型算法安全。规范机器学习算法技术和服务的安全要求和评估方法，解决现有机器学习算法在全生存周期中存在的个人信息泄露、决策偏见、算法难以抵御外部恶意攻击以及算法服务不合规等安全问题。

2、数据安全。规范人工智能研发、应用等过程中的数据安全要求，指导数据处理活动符合相关法律法规及政策文件要求，包括训练数据、数据标注、用户数据使用等方面。

3、系统安全。规范人工智能系统层的软硬件安全要求，包括芯片安全、计算平台安全、训练推理框架安全、软件中间件安全等方面。

4、开发安全。规范人工智能系统的开发管理流程等，确保技术和产品在整个生命周期内的安全性，为人工智能系统的安全开发提供操作指南，并指导开发者做好基于第三方基础模型的安全二次开发。

5、应用安全。规范人工智能应用的安全评估、产品选型、安全建设、安全使用、人机协同管理、安全检测，以及安全审计及问题改进等方面，提升人工智能应用安全水平。

6、运行安全。规范人工智能相关产品、服务的网络运行安全，从全生命周期不同阶段应对人工智能服务的数据泄露、模型篡改、服务中断、算法偏见等问题，包括模型更新安全处置、安全应急响应等方面。

7、供应链安全。规范人工智能软硬件供应链在安全方面的要求，包括供应商评估、生产过程控制、软硬件供应管

理、风险识别和防范等方面。

8、风险管理。规范人工智能产品或服务的研究开发、设计制造、部署应用等活动中安全风险管理，包括伦理安全风险防范、安全风险评估、安全风险威胁信息共享等方面。

（三）关键技术

关键技术类标准紧扣人工智能相关技术发展及应用情况，明确各项关键技术的安全保障要求，为人工智能技术健康发展保驾护航，包括生成式人工智能安全、智能体安全、具身智能安全、多模态安全、生成合成安全、安全对齐、安全围栏等研制方向，关键技术标准子体系如图 4 所示。

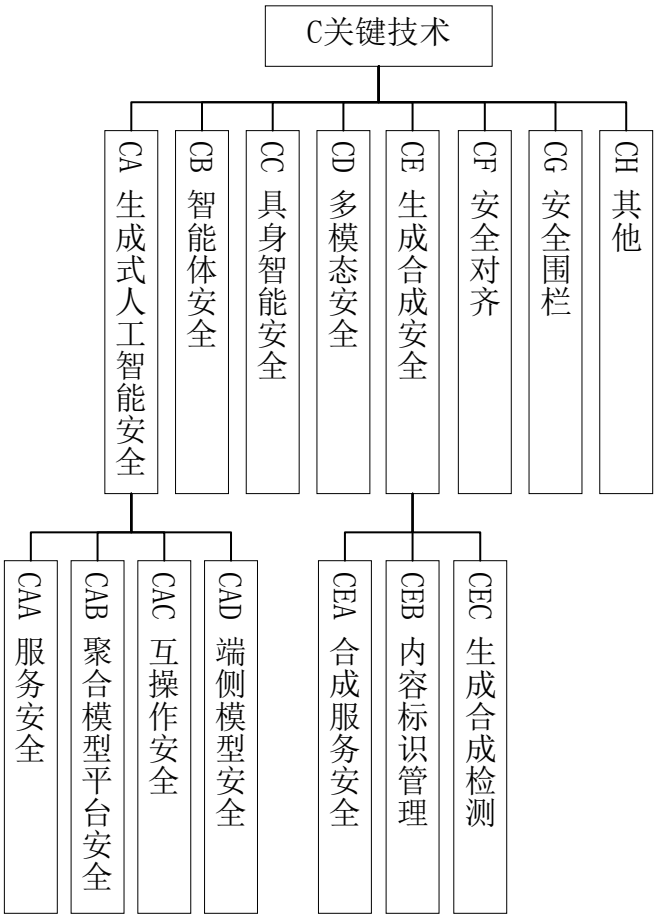

图 4 关键技术标准子体系

1、生成式人工智能安全。规范生成式人工智能研究开发者、服务提供者在模型开发、部署、运行、维护等生命周期以及提供服务时的安全要求，包括生成式人工智能服务安全、聚合模型平台安全、人工智能系统互操作安全、端侧模型安全等方面。

2、智能体安全。围绕智能体形态、场景、数据流程以及特有安全风险，规范智能体安全开发和运行过程，包括感知安全、模型决策安全、接口调用安全、数据安全和个人信息保护、安全责任划分等方面。

3、具身智能安全。规范具身智能的安全开发与应用，给出具身智能系统的安全保障框架及具体安全要求，包括系统架构安全、通信安全、数据存储安全、人机交互安全、自动化更新机制等方面，提升系统抗风险能力。

4、多模态安全。规范多模态大模型在处理跨模态数据及生成内容过程中的安全保障要求，针对多模态生成内容给出模型在模态转换中的全链路安全指引，适用于研发、部署、应用多模态人工智能技术的各类场景。

5、生成合成安全。规范各类人工智能生成合成服务，以及各类生成合成内容的标识和检测验证，防止生成合成内容的滥用、误用、恶意使用。包括合成服务安全、内容标识管理、生成合成检测等方面。

6、安全对齐。规范人工智能系统的人机协同开发，明确系统设计目标和行为边界，制定目标对齐的评估方法，确

保系统输出符合伦理、法律及社会价值观，防止因设计缺陷、数据偏差或外部攻击导致系统偏离安全目标。

7、安全围栏。规范人工智能安全围栏的建设，围绕人工智能模型输入输出安全风险，提出输入检测、提示词安全、模型输出安全等方面的安全要求，指导人工智能企业开展人工智能安全围栏建设。

（四）测试评估

测试评估类标准旨在以测试评估工作帮助提升人工智能安全水平，包括安全能力测试、模型安全性测试、产品服务安全测试、场景应用安全测试、安全测试基准等研制方向，测试评估标准子体系如图 5 所示。

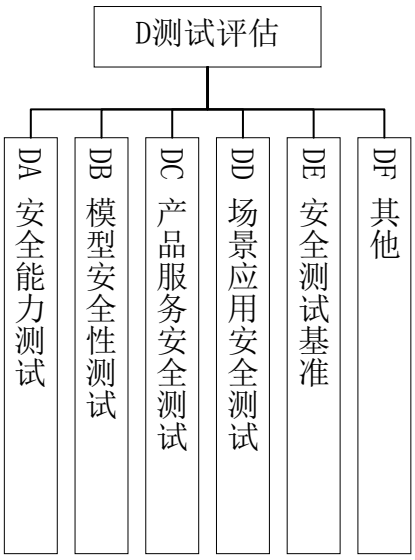

图 5 测试评估标准子体系

1、安全能力测试。规范组织的人工智能安全能力成熟度模型，给出人工智能系统设计、研发、训练、测试、部署、使用、维护等生命周期各环节的安全能力成熟度等级要求以

及评估方法，指导组织开展人工智能安全能力建设。

2、模型安全性测试。规范人工智能模型安全性测试评估框架、流程，以及具体测评方法，包括模型数据安全、模型鲁棒性、输出可靠性，以及训练数据泄露、偏见歧视、注入攻击、后门攻击、对抗攻击等方面。

3、产品服务安全测试。规范人工智能产品和服务在设计、开发、部署及运行过程中各类安全问题的测试方法，包括用户数据保护、产品服务输出安全性、未成年人安全保护措施、服务响应能力、错误处理机制及应急响应方案等。

4、场景应用安全测试。规范不同领域应用场景对人工智能系统的特定安全要求，针对人工智能在特定领域应用时的安全需求，从应用场景适配性、领域合规性以及应用可靠性等维度明确安全测试内容，并给出安全测试方法。

5、安全测试基准。规范人工智能安全评测基准的建设，围绕人工智能主要安全风险，给出安全评测基准数据集建设的安全要求，指导生成式人工智能技术研发者、系统开发者、服务提供者或第三方评估机构开展安全评测基准建设。

（五）产品与应用

产品与应用类标准旨在保障人工智能在各行业、各领域的安全应用，包括个人应用安全、行业应用安全等研制方向，产品与应用标准子体系如图 6 所示。

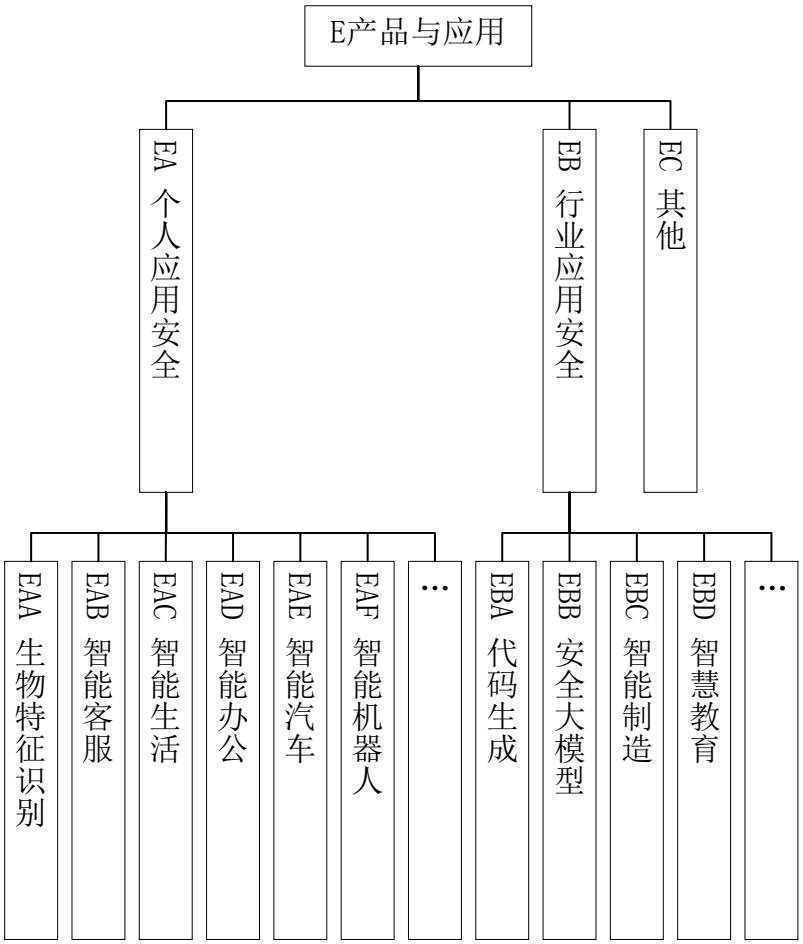

图 6 产品与应用标准子体系

1、个人应用安全。规范人工智能个人应用方面的安全要求，围绕人工智能导致的数据泄露、不当内容输出、服务非法利用等问题给出安全指南，包括生物特征识别、智能客服、智能生活、智能办公、智能汽车、智能机器人等方面。

2、行业应用安全。规范人工智能在各行业领域的安全应用，保障相关环节涉及的人工智能系统安全运行，帮助提升各行业领域智能化安全水平，包括人工智能代码生成、安全大模型、智能制造、智慧教育等方面。

三、组织实施

一是统筹协调、共同推进。统筹推进人工智能安全标准

体系建设，组织开展国家标准制修订工作，各标准化技术组织、行业协会、产业技术联盟、企事业单位等产业各界协调配合，有序地推进开展人工智能安全标准化工作，建设国标为主、行标细化、团标为辅的标准供给体系。

二是急用先行、规划引领。按照本文件明确的研制方向和重点任务，坚持需求导向、注重轻重缓急，尽快制定业界急需和缺失的人工智能安全关键国家标准，完善跨行业、跨领域的标准沟通协调机制，加强规划建设，保持标准先进性。

三是标准宣贯、强化实施。标准宣贯与实施应用是标准化工作的重要组成部分，加强人工智能安全标准宣贯培训力度，面向企业、科研机构、地方主管部门开展标准宣讲，结合重点领域应用推动标准实施，提高人工智能安全标准实施效果。

四是国际合作、创新发展。深化国家标准化战略改革，着力抓好国际标准化。更加深入参与到国际标准化组织和活动中，积极与国外人工智能安全相关组织开展标准化交流与合作，支持企事业单位参与国际标准化组织（ISO）、国际电工委员会（IEC）、国际电信联盟（ITU）国际标准化活动，推动相关国际标准制定。

附件 1：

《人工智能安全治理框架》标准映射表

《人工智能安全治理框架》（以下简称“《框架》”）基于风险管理理念，分析了人工智能技术特性以及在不同行业领域应用场景，梳理了人工智能技术本身及其在应用过程中面临的各种安全风险隐患，并针对不同类型的人工智能安全风险，从技术、管理两方面提出防范应对措施。

为更好以人工智能安全标准化工作支撑落实《框架》，本标准体系围绕《框架》中明确的各类安全风险，结合人工智能技术发展及应用情况，提出了可帮助防范化解各类人工智能安全风险的重点标准，以系统性、前瞻性的标准化工作应对各类风险挑战。本标准体系中各项标准与《人工智能安全治理框架》中各类安全风险的映射关系见表 1。

表 1 《人工智能安全治理框架》映射表

安全风险			对应标准
内生安全风险	模型算法安全风险	可解释性差的风险	机器学习算法安全评估规范、人工智能安全开发指引、人工智能模型更新安全处置指南、人工智能模型安全性测试评估方法、生成式人工智能安全评测基准数据规范
		偏见、歧视风险	
		鲁棒性弱风险	
		被窃取、篡改的风险	

安全风险			对应标准
		输出不可靠风险	
		对抗攻击风险	
	数据安全风险	违规收集使用数据风险	人工智能系统个人信息使用安全指引、车外画面局部轮廓化处理效果验证，以及生物特征识别安全相关 10 项国家标准
		训练数据含不当内容、被“投毒”风险	生成式人工智能预训练和优化训练数据安全规范
		训练数据标注不规范风险	生成式人工智能数据标注安全规范
		数据泄露风险	人工智能安全开发指引、人工智能安全应用指引、智能体安全要求、生成式人工智能系统互操作安全规范
	系统安全风险	缺陷、后门被攻击利用风险	机器学习算法安全评估规范、人工智能计算平台安全框架、聚合模型平台安全指引、人工智能训练及推理框架安全、人工智能软件中间件安全
		算力安全风险	人工智能芯片安全要求、人工智能异构芯片互联安全技术规范
		供应链安全风险	人工智能供应链安全要求

安全风险			对应标准
应用安全风险	网络域风险	信息内容安全风险	生成式人工智能服务安全基本要求、人工智能安全围栏建设指南、互联网信息服务深度合成安全规范
		混淆事实、误导用户、绕过鉴权的风险	人工智能生成合成内容标识方法、人工智能生成合成内容检测技术框架
		不当使用引发信息泄露风险	人工智能安全应用指引
		滥用于网络攻击的风险	生成式人工智能服务安全基本要求、人工智能代码生成服务安全要求、网络安全大模型建设指南
		模型复用的缺陷传导风险	基于第三方基础模型二次开发安全实践指引
	现实域风险	诱发传统经济社会安全风险	人工智能应用安全分类分级方法、人工智能安全风险威胁信息共享指南、生成式人工智能服务安全应急响应指南、人工智能安全能力成熟度评估方法、个人应用安全标准、行业应用安全标准
		用于违法犯罪活动的风险	生成式人工智能服务安全基本要求、人工智能安全风险评估指南

安全风险			对应标准
		两用物项和技术滥用风险	生成式人工智能服务安全基本要求、端侧大模型网络安全指南、网络安全大模型建设指南
	认知域风险	加剧“信息茧房”效应风险	机器学习算法安全评估规范
		用于开展认知战的风险	人工智能生成合成内容标识方法、生成式人工智能服务安全基本要求
	伦理域风险	加剧社会歧视偏见、扩大智能鸿沟的风	人工智能伦理安全风险防范指引、机器学习算法安全评估规范
		挑战传统社会秩序的风险	人工智能伦理安全风险防范指引
		未来脱离控制的风险	人工智能安全围栏建设指南、具身智能安全要求

附件 2：

人工智能安全现行及在研标准文件

在人工智能安全标准文件方面，主要包括三种形式：国家标准（包括强制性国家标准、推荐性国家标准）、技术文件、实践指南。目前已发布 12 项推荐性国家标准、1 项技术文件、3 项实践指南，正在推动 1 项强制性国家标准、6 项推荐性国家标准、1 项实践指南的研制工作。

一、标准文件明细表

总序号	分序号	标准名称	标准/计划号	状态
B 安全管理				
BA 模型算法安全				
1.	1）	信息安全技术 机器学习算法安全评估规范	GB/T 42888-2023	已发布
BB 数据安全				
BBA 训练数据安全				
2.	1）	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	20242095-T-469	送审稿
BBB 数据标注安全				
3.	1）	网络安全技术 生成式人工智能数据标注安全规范	20242097-T-469	送审稿

总序号	分序号	标准名称	标准/计划号	状态
BC 系统安全				
BCB 计算平台安全				
4.	1）	网络安全技术 人工智能计算平台安全框架	20230249-T-469	征求意见稿
BF 运行安全				
BFB 安全应急响应				
5.	1）	网络安全标准实践指南—生成式人工智能服务安全应急响应指南	/	征求意见稿
BH 风险管理				
BHA 伦理安全风险				
6.	1）	网络安全标准实践指南—人工智能伦理安全风险防范指引	TC260-PG-20211A	已发布
C 关键技术				
CA 生成式人工智能安全				
CAA 服务安全				
7.	1）	生成式人工智能服务安全基本要求	TC260-003	已发布
8.	2）	网络安全技术 生成式人工智能服务安全基本要求	20241752-T-469	送审稿
CE 生成合成安全				
CEA 合成服务安全				

总序号	分序号	标准名称	标准/计划号	状态
9.	1）	网络安全技术 互联网信息服务深度合成安全规范	20240395-T-46	制定中
CEB 内容标识管理				
10.	1）	网络安全标准实践指南—生成式人工智能服务内容标识方法	TC260-PG-20233A	已发布
11.	2）	网络安全技术 人工智能生成合成内容标识方法（强制性国家标准）	20241842-Q-252	报批稿
E 产品与应用				
EA 个人应用安全				
EAA 生物特征识别				
12.	1）	信息安全技术 人脸识别数据安全要求	GB/T 41819-2022	已发布
13.	2）	信息安全技术 声纹识别数据安全要求	GB/T 41807-2022	已发布
14.	3）	信息安全技术 基因识别数据安全要求	GB/T 41806-2022	已发布
15.	4）	信息安全技术 步态识别数据安全要求	GB/T 41773-2022	已发布
16.	5）	信息安全技术 生物特征识别信息保护基本要求	GB/T 40660-2021	已发布

总序号	分序号	标准名称	标准/计划号	状态
17.	6）	信息安全技术 远程人脸识别系统技术要求	GB/T 38671-2020	已发布
18.	7）	信息安全技术 基于生物特征识别的移动智能终端身份鉴别技术框架	GB/T 38542-2020	已发布
19.	8）	信息安全技术 虹膜识别系统技术要求	GB/T 20979-2019	已发布
20.	9）	信息安全技术 指纹识别系统技术要求	GB/T 37076-2018	已发布
21.	10）	信息安全技术 基于可信环境的生物特征识别身份鉴别协议框架	GB/T 36651-2018	已发布
EAE 智能汽车				
22.	1）	信息安全技术 汽车数据处理安全要求	GB/T 41871-2022	已发布
23.	2）	网络安全标准实践指南—车外画面局部轮廓化处理效果验证	TC260-PG-20241A	已发布
EB 行业应用安全				
EBA 代码生成				
24.	1）	网络安全技术 人工智能代码生成服务安全要求	/	制定中

二、已发布标准文件

1、GB/T 42888-2023《信息安全技术 机器学习算法安全评估规范》

该标准规定了机器学习算法技术和服务的安全要求和评估方法，以及机器学习算法安全评估流程。适用于指导机器学习算法提供者保障机器学习算法生存周期安全以及开展机器学习算法安全评估，也可为监管评估提供参考。

2、GB/T 41871-2022《信息安全技术 汽车数据处理安全要求》

该标准规定了汽车数据处理者对汽车数据进行收集、传输等处理活动的通用安全要求、车外数据安全要求、座舱数据安全要求和管理安全要求。适用于汽车数据处理者开展汽车数据处理活动，适用于汽车的设计、生产、销售、使用和运维，也适用于主管监管部门和第三方评估机构等对汽车数据处理活动进行监督、管理和评估。

3、TC260-003《生成式人工智能服务安全基本要求》

该技术文件规定了生成式人工智能服务在安全方面的基本要求，包括语料安全、模型安全、安全措施等，并给出了安全评估要求。适用于服务提供者开展安全评估、提高安全水平，也可为相关主管部门评判生成式人工智能服务安全水平提供参考。

4、TC260-PG-20241A《网络安全标准实践指南—车外画面局部轮廓化处理效果验证》

该实践指南给出了车外画面中人脸、车牌局部轮廓化处理效果的验证流程、方法及指标。适用于汽车数据处理者对车外画面进行人脸、车牌局部轮廓化处理效果的自行验证，也适用于第三方机构对局部轮廓化处理效果的验证。

5、TC260-PG-20233A《网络安全标准实践指南—生成式人工智能服务内容标识方法》

该实践指南给出了生成式人工智能服务提供者对生成内容进行标识的方法。适用于生成式人工智能服务提供者利用生成式人工智能技术向公众提供生成文本、图片、音频、视频等内容时对生成内容进行标识。

6、TC260-PG-20211A《网络安全标准实践指南—人工智能伦理安全风险防范指引》

该实践指南针对人工智能可能产生的伦理安全风险问题，给出了安全开展人工智能研究开发、设计制造、部署应用等相关活动的规范指引。适用于相关组织或个人开展人工智能研究开发、设计制造、部署应用等相关活动。

7、生物特征识别相关安全标准

人脸、声纹、步态、基因等 10 项生物特征识别安全标准，相关标准包括《信息安全技术 生物特征识别信息保护基本要求》、《信息安全技术 人脸识别数据安全要求》、《信息安全技术 声纹识别数据安全要求》、《信息安全技

术 基因识别数据安全要求》、《信息安全技术 步态识别数据安全要求》、《信息安全技术 远程人脸识别系统技术要求》、《信息安全技术 虹膜识别系统技术要求》、《信息安全技术 指纹识别系统技术要求》、《信息安全技术 基于生物特征识别的移动智能终端身份鉴别技术框架》、《信息安全技术 基于可信环境的生物特征识别身份鉴别协议框架》，旨在降低生物特征识别技术应用带来的各类安全风险。

（1）GB/T 41819-2022《信息安全技术 人脸识别数据安全要求》

该标准规定了人脸识别数据的安全通用要求以及收集、存储、使用、传输、提供、公开、删除等具体处理活动的安全要求。适用于数据处理者安全开展人脸识别数据处理活动。

（2）GB/T 41807-2022《信息安全技术 声纹识别数据安全要求》

该标准规定了声纹识别数据的收集、存储、使用、传输、提供、公开、删除等活动中，对数据处理者的安全要求。适用于规范数据处理者的声纹识别数据处理行为。

（3）GB/T 41806-2022《信息安全技术 基因识别数据安全要求》

该标准规定了基因识别数据及关联信息的收集、存储、使用、加工、传输、提供、公开、删除等数据处理活动的安全要求。适用于基因识别数据及关联信息的处理者规范数据

处理活动，也可为监管部门、第三方评估机构对基因识别数据处理活动进行监督、管理、评估提供参考。

（4）GB/T 41773-2022《信息安全技术 步态识别数据安全要求》

该标准规定了步态识别数据收集、存储、传输、使用、加工、提供、公开、删除等数据处理活动的安全要求。适用于步态识别数据处理者规范数据处理活动，监管部门、第三方评估机构对步态识别数据处理活动进行监督、管理、评估参照使用。

（5）GB/T 40660-2021《信息安全技术 生物特征识别信息保护基本要求》

该标准规定了各类生物特征识别信息控制者开展收集、存储、使用、委托处理共享、转让、公开披露删除等生物特征识别信息处理活动应遵循的基本原则和安全要求。适用于规范各类生物特征识别信息控制者开展生物特征识别信息处理活动，也适用于第三方机构对生物特征识别信息处理活动进行测评。

（6）GB/T 38671-2020《信息安全技术 远程人脸识别系统技术要求》

该标准规定了采用人脸识别技术在服务器端远程进行身份鉴别的信息系统的功能、性能和安全要求、安全保障要求。适用于采用人脸识别技术在服务器端远程进行身份鉴别的信息系统的研制和测试，系统的管理可参照使用。

（7）GB/T 38542-2020《信息安全技术 基于生物特征识别的移动智能终端身份鉴别技术框架》

该标准规定了基于生物特征识别的移动智能终端身份鉴别的技术框架，包括技术架构、业务流程功能要求和安全要求。适用于基于生物特征识别的移动智能终端身份鉴别系统的设计、开发与集成。

（8）GB/T 20979-2019《信息安全技术 虹膜识别系统技术要求》

该标准规定了采用虹膜识别技术进行身份识别的虹膜识别系统的结构、功能、性能、安全要求及等级划分。适用于虹膜识别系统的设计与实现，对虹膜识别系统的测试、管理也可参照使用。

（9）GB/T 37076-2018《信息安全技术 指纹识别系统技术要求》

该标准规定了采用指纹识别技术进行身份鉴别的指纹识别系统基本级和增强级的功能、性能、安全要求和等级划分。适用于指纹识别系统的设计与实现，对指纹识别系统的测试、管理也可参照使用。

（10）GB/T 36651-2018《信息安全技术 基于可信环境的生物特征识别身份鉴别协议框架》

该标准规定了基于可信环境的生物特征识别身份鉴别协议框架，包括协议框架、协议流程、协议规则以及协议接

口等内容。适用于生物特征识别身份鉴别服务的开发、测试和评估。

二、在研标准文件

1、《网络安全技术 生成式人工智能服务安全基本要求》

拟解决问题：该标准是《生成式人工智能服务管理暂行办法》的配套标准，拟针对当前生成式人工智能服务带来的传播虚假信息、非法留存数据、造成偏见歧视、导致用户沉迷，以及侵害他人肖像权、名誉权，个人隐私、商业秘密等安全风险，提出细化安全技术要求，帮助服务提供者提高安全水平。

主要内容：该标准拟提出面向我国境内公众提供生成式人工智能服务的基本安全要求，具体包括生成式人工智能服务相关的算法模型安全、训练数据安全、数据标注安全、防范虚假信息、防止用户沉迷、防范歧视、保护用户隐私等方面内容。

2、《网络安全技术 生成式人工智能数据标注安全规范》

拟解决问题：该标准是《生成式人工智能服务管理暂行办法》的配套标准，拟针对生成式人工智能产品研制中的人工标注环节，详细描述清晰、具体、可操作的人工标注规则，标注人员培训，标注内容正确性等方面的具体要求。

主要内容：该标准拟针对生成式人工智能产品研制中人工标注环节的标注规则、标注人员培训、标注内容正确性等方面提出安全规范。

3、《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》

拟解决问题：该标准是《生成式人工智能服务管理暂行办法》的配套标准，拟对生成式人工智能产品预训练和优化训练数据来源的合法性进行具体规范，详细描述符合法律法规要求、不含侵犯知识产权内容、保护个人信息等方面的具体要求，阐释数据的真实性、准确性、客观性、多样性要求。

主要内容：该标准拟针对生成式人工智能产品预训练和优化训练数据来源合法性，符合法律法规要求，不含侵犯知识产权内容，保护个人信息，保证真实性、准确性、客观性、多样性等方面提出安全规范。

4、强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》

拟解决问题：该标准拟解决人工智能生成内容的误用风险以及误导性内容的传播风险，通过显式标识提示用户当前内容由人工智能生成，避免用户误用生成内容，并通过隐式标识对生成内容来源进行确认，降低误导性生成内容的传播风险。

主要内容：该标准规定了人工智能生成合成内容显式标识和隐式标识的种类、要素和格式，给出了人工智能生成合成内容标识方法。

5、《网络安全技术 人工智能计算平台安全框架》

拟解决问题：该标准拟解决人工智能应用开发运行过程

面临的底层硬件安全问题，以及上层应用面临的多场景基础共性安全问题，保护人工智能模型及数据。

主要内容：该标准拟提出人工智能计算平台安全框架以及相应的安全模块和机制，以保障用户数据与人工智能模型数据安全。

6、《网络安全技术 互联网信息服务深度合成安全规范》

拟解决问题：该标准细化补充《互联网信息服务深度合成管理规定》各项要求，解决深度合成服务提供者对规定细化理解不到位，以及实践落实规定各项条款时执行不到位的问题。

主要内容：该标准拟从互联网信息服务生命周期的信息生成、处理、发布、传播、存储、销毁等环节，以及技术算法生命周期的设计开发、验证测试、部署运行、维护升级、退役下线等五个阶段，对深度合成服务提供者和技术支持者提出开展互联网深度合成服务在安全方面的通用要求以及证实评估方法。

7、《网络安全技术 人工智能代码生成服务安全要求》

拟解决问题：该标准拟解决人工智能生成代码质量参差不齐，存在输出漏洞代码、泄露用户数据等问题，指导相关组织机构开展自评估，提升人工智能代码生成类互联网信息服务和应用安全水平。

主要内容：该标准拟针对利用生成式人工智能技术所构建的代码生成类互联网信息服务，提出文书签署、代码审查、

过程披露、风险提示等方面安全要求。

8、《网络安全标准实践指南—生成式人工智能服务安全应急响应指南》

拟解决问题：该指南旨在解决生成式人工智能服务在安全应急响应方面的问题。针对生成式人工智能服务安全事件，从事件管理全生命周期不同阶段应对人工智能服务的数据泄露、模型篡改、服务中断、算法偏见等问题，确保人工智能服务提供者在面对安全事件时能够快速、有效地响应。

主要内容：该指南围绕生成式人工智能服务安全事件给出了安全事件的分类和分级建议，并给出了生成式人工智能服务安全事件应急响应过程，包括应急准备、监测预警、应急处置、总结改进阶段的管理措施和技术方法，可用于指导生成式人工智能服务提供者提高安全应急响应能力。

附件 3：

人工智能安全标准重点方向明细表

A 基础共性
AA 术语定义
人工智能安全术语
AB 分类分级
人工智能应用安全分类分级方法
AC 通用要求
人工智能安全治理基本要求
AD 参考架构
人工智能安全参考架构
AE 其他
B 安全管理
BA 模型算法安全
机器学习算法安全评估规范（已发布）
BB 数据安全
BBA 训练数据安全
生成式人工智能预训练和优化训练数据安全规范（送审稿）
BBB 数据标注安全
生成式人工智能数据标注安全规范（送审稿）

BBC 用户数据使用
人工智能系统个人信息使用安全指引
BC 系统安全
BCA 芯片安全
人工智能芯片安全要求、人工智能异构芯片互联安全技术规范
BCB 计算平台安全
人工智能计算平台安全框架（征求意见稿）
BCC 训练推理框架安全
人工智能训练及推理框架安全要求
BCD 软件中间件安全
人工智能软件中间件安全
BD 开发安全
BDA 安全开发方法
人工智能安全开发指引
BDB 二次开发安全
基于第三方基础模型二次开发安全实践指引
BE 应用安全
人工智能安全应用指引、面向未成年人的人工智能安全应用指南
BF 运行安全
BFA 模型更新安全处置
人工智能模型更新安全处置指南
BFB 安全应急响应

生成式人工智能服务安全应急响应指南（征求意见稿）
BG 供应链安全
人工智能供应链安全要求
BH 风险管理
BHA 伦理安全风险
人工智能伦理安全风险防范指引（已发布）
BHB 安全风险评估
人工智能安全风险评估指南
BHC 威胁信息共享
人工智能安全风险威胁信息共享指南
BI 其他
C 关键技术
CA 生成式人工智能安全
CAA 服务安全
生成式人工智能服务安全基本要求（送审稿）
CAB 聚合模型平台安全
聚合模型平台安全指引
CAC 互操作安全
生成式人工智能系统互操作安全规范
CAD 端侧模型安全
端侧大模型网络安全指南
CB 智能体安全

智能体安全要求
CC 具身智能安全
具身智能安全要求
CD 多模态安全
多模态大模型安全要求
CE 合成内容管理
CEA 合成服务安全
互联网信息服务深度合成安全规范（制定中）
CEB 内容标识管理
人工智能生成合成内容标识方法（报批稿）、生成式人工智能服务内容标识方法（已发布）、人工智能生成合成内容标识检测及验证方法、人工智能生成合成内容元数据标识安全防护指南、人工智能生成合成内容隐藏水印标识指南
CEC 生成合成检测
人工智能生成合成内容检测技术框架
CF 安全对齐
人工智能安全对齐指南
CG 安全围栏
人工智能安全围栏建设指南
CH 其他
D 测试评估
DA 安全能力测试
人工智能安全能力成熟度评估方法

DB 模型安全性测试
人工智能模型安全性测试评估方法
DC 产品服务安全测试
人工智能产品服务安全测试评估方法
DD 场景应用安全测试
人工智能场景应用安全测试评估方法
DE 安全测试基准
生成式人工智能安全评测基准数据规范
DF 其他
E 产品与应用
EA 个人应用安全
EAA 生物特征识别安全
生物特征识别信息保护基本要求、人脸识别数据安全要求、声纹识别数据安全要求、基因识别数据安全要求、步态识别数据安全要求、远程人脸识别系统技术要求、虹膜识别系统技术要求、指纹识别系统技术要求、基于生物特征识别的移动智能终端身份鉴别技术框架、基于可信环境的生物特征识别身份鉴别协议框架（10 项均已发布）
EAB 智能客服
智能客服网络安全指南
EAC 智能生活
智能生活助手网络安全指南
EAD 智能办公
智能办公助手网络安全指南

EAE 智能汽车
汽车数据处理安全要求（已发布）、自动驾驶汽车服务安全要求、自动驾驶智能算法安全要求、智能座舱个人信息保护指南、车外画面局部轮廓化处理效果验证（已发布）
EAF 智能机器人
智能机器人网络安全防护指南
……
EB 行业应用安全
EBA 代码生成
人工智能代码生成服务安全要求（制定中）
EBB 安全大模型
网络安全大模型建设指南
EBC 智能制造
智能制造安全
EBD 智慧教育
智慧教育安全
……
EC 其他