人工智能安全标准化白皮书

（2023版）

全国信息安全标准化技术委员会
大数据安全标准特别工作组

2023年5月

人工智能安全标准化白皮书

（2023版）

全国信息安全标准化技术委员会
大数据安全标准特别工作组

2023年5月

引　言

人工智能是人类科学技术发展的重要成果，是信息时代向前演进的关键动力。运用好、发展好、治理好人工智能，让人工智能持续、安全地造福人类社会，已经成为世界各国的基本共识。

党的十八大以来，在习近平新时代中国特色社会主义思想指引下，我国人工智能保持了安全、有序、快速的发展进程，在政治、军事、医疗、工业、经济等领域作用日益关键，已逐渐成为国家和社会的技术支柱。

当前，人工智能发展再一次迈入关键时期，以生成式人工智能为代表的新技术、新应用不断打破人们对于人工智能的固有认知，也带来了大量网络意识形态安全、数据安全、个人信息安全等方面新风险、新挑战，化解安全风险、统筹发展和安全成为重大难题。

贯彻总体国家安全观，全国信息安全标准化技术委员会大数据安全标准特别工作组坚持发挥标准化工作基础性、规范性作用，开展了一系列人工智能安全标准化工作，为推动人工智能发展贡献力量。

面对人工智能安全新形势，为全面介绍人工智能安全标准化工作进展情况，分享相关工作经验，在《人工智能安全标准化白皮书（2019版）》等前期研究成果基础上，特发布本白皮书。

编写单位

中国电子技术标准化研究院

复旦大学

浙江大学

西安交通大学

清华大学

上海商汤智能科技有限公司

中国科学院信息工程研究所

腾讯云计算（北京）有限责任公司

北京计算机技术及应用研究所

华为技术有限公司

阿里巴巴（北京）软件服务有限公司

蚂蚁科技集团股份有限公司

公安部第三研究所

中国科学院自动化研究所

上海燧原科技有限公司

北京百度网讯科技有限公司

国际商业机器（中国）有限公司

北京天融信网络安全技术有限公司

中电长城网际系统应用有限公司

OPPO广东移动通信有限公司

编写人员

姚相振　上官晓丽　王建民　郝春亮　许晓耕　任　奎　杨　珉
沈　超　胡　影　金　涛　王　姣　秦　湛　张世天　王　蕊
张晓寒　蔺琛皓　王秉政　蒋　慧　陈　恺　梅敬青　彭骏涛
李　实　王志波　徐　浩　张雨桐　李　前　程海旭　王　龑
张堃博　孟国柱　王小璞　张妍婷　张宇光　张　骁　刘　炎
刘　楠　郭　敏　周晨炜　薛　科　刘继顺　刘海涛　贾依真

目录 *CONTENTS*

目录 *CONTENTS*

一、人工智能发展现状

过去十余年，依托全球数据、算法、算力持续突破，人工智能全面走向应用，已成为社会生产生活的支柱性技术。2020年后，当自动驾驶、人脸识别等热门应用发展逐渐放缓、社会对人工智能整体发展预期日益冷静时，大模型技术潜力的释放以最振聋发聩的方式宣告了人工智能第三次高速发展期远未结束，当前正是攀登发展高峰的关键时期。

另一方面，当人工智能可以通过人类最严格的考试、同时执行多种工作命令、具备一定的推理规划能力、生成以假乱真的照片、模仿人类与人聊天不被发现时，其安全问题也更为复杂棘手，传统安全考虑以及管理方法需要重新审视。在此背景下，人工智能是否安全、如何保障安全成为全球焦点，统筹安全与发展是其中关键。

（一）人工智能技术特点

技术发展方面，随着谓词推理、专家系统、知识树和向量机学习等传统技术的发展日趋放缓，促使以联结主义和概率统计等理论为基础的深度学习加速发展，迈入了以人工神经网络为基础、以大模型为典型应用的新发展阶段。

在模型方面，大规模人工智能模型逐步成为业界主流。以生成式人工智能为例，具备数百亿参数的模型已非罕见，并随着模型规模增长产生了接近人类的“高级”能力，使人们相信通用人工智能或将到来。Stable Diffusion、Midjourney等视觉生成模型具有类似人类的视觉创作能力，ChatGPT等文本生成模型具有高度近似人类的语言推理和规划等能力。有研究认为，这些能力是随着模型参数达到数百亿级别后逐渐产生的，虽其

技术原理尚未明晰，但进一步推动了模型越来越大的技术趋势。

在训练方面，有人类参与的指令微调技术是近年来人工智能的另一大技术特点。指令微调主要有三种实现形式，以预训练语言模型为例：一是引入人工撰写的大量对话数据对模型进行微调训练；二是人工对微调后模型面向同一提示词生成的多个备选答案进行价值排序，训练价值评分模型；三是在价值评分模型的奖励信号下，微调模型进行强化学习训练，不断改进模型的表现。通过该部分技术，可将在海量语料库上训练的模型与复杂的人类价值观实现对齐，期望人工智能可以生成正确、有用、无害的内容。

（二）人工智能应用趋势

应用发展方面，人工智能进一步与社会各方面融合。跨领域、面向通用的人工智能应用持续发展，各领域处理独立任务的人工智能应用更加深度嵌入产业生态。未来，预期形成以通用人工智能应用为基座，专用人工智能应用环绕的新人工智能“生态圈”。

1、人工智能与实体经济融合发展

近年来，人工智能与实体经济融合愈发深入，融合形式愈发多样，对产业促进作用明显，推动新型业态逐步形成。

当前，人工智能在多个行业领域广泛应用，在制造领域的运营管理优化、制造过程优化等环节，智能家居领域的身份鉴别、功能控制、安全防护等环节，智能交通领域的动态感知、自动驾驶、车路协同等方面，智能医疗领域的辅助诊断、治疗监护、疫情防控等方面，教育领域的虚拟实验室、虚拟教室、课件制作、智能判卷、教学效果分析等方面，金融领域的金融风险控制等方面，都推动了相关产品服务的新一轮变革。

2、人工智能作为助手融入新领域

人工智能的发展不仅颠覆了数字内容生产方式、处理方式和消费模式，而且极大丰富了人们的数字生活，虚拟试装增加购物体验、虚拟主播增强广告效果、智能客服提升反馈效率、虚拟教师增强师生交互、智能办公助手提高各类文档的撰写效率、智能编程助手降低编程时间与人力成本、智能翻译降低沟通壁垒，人工智能应用已成为人类生产生活中必不可少的电子助手。

（三）人工智能安全属性

伴随着人工智能应用的常态化，人工智能安全问题的研讨也持续开展。除了网络安全基本属性，即人工智能系统及其相关数据的机密性、完整性、可用性以及系统对恶意攻击的抵御能力之外，讨论人工智能安全一般还需要考虑以下属性。

1、可靠性：指人工智能及其所在系统在承受不利环境或意外变化时，例如数据变化、噪声、干扰等因素，仍能按照既定的目标运行、保持结果有效的特性。可靠性通常需要综合考虑系统的容错性、恢复性、健壮性等多个方面。

2、透明性：指人工智能在设计、训练、测试、部署过程中保持可见、可控的特性，只有具备了透明性，用户才能够在必要时获取模型有关信息，包括模型结构、参数、输入输出等，方可进一步实现人工智能开发过程的可审计以及可追溯。

3、可解释性：描述了人工智能算法模型可被人理解其运行逻辑的特性。具备可解释性的人工智能，其计算过程中使用的数据、算法、参数和逻辑等对输出结果的影响能够被人类理解，使人工智能更易于被人类管控、更容易被社会接受。

4、公平性：指人工智能模型在进行决策时，不偏向某个特定的个体或群体，也不歧视某个特定的个体或群体，平等对待不同性别、不同种族、不同文化背景的人群，保证处理结果的公正、中立，不引入偏见和歧视因素。

5、隐私性：指人工智能在开发与运行的过程中实现了保护隐私的特性，包括对个人信息和个人隐私的保护、对商业秘密的保护等。隐私性旨在保障个人和组织的合法隐私权益，常见的隐私增强方案包括最小化数据处理范围、个人信息匿名化处理、数据加密和访问控制等。

二、人工智能安全风险分析

近年来，人工智能保持快速发展势头，但人工智能所带来的安全风险也不容忽视。

（一）用户数据用于训练，放大隐私信息泄露风险

当前，人工智能利用服务过程中的用户数据进行优化训练的情况较为普遍，但可能涉及在用户不知情情况下收集个人信息、个人隐私、商业秘密等，安全风险较为突出。一方面，人工智能模型日益庞大，开发过程日益复杂，数据泄露风险点更多、隐蔽性更强，人工智能所使用开源库漏洞引发数据泄露的情况也很难杜绝。另一方面，交互式人工智能的应用降低了数据流入模型的门槛。用户在使用交互式人工智能时往往会放松警惕，更容易透露个人隐私、商业秘密、科研成果等数据，例如企业员工在办公时容易将商业秘密输入人工智能寻找答案，继而导致商业秘密的泄露。为应对该问题，特别是为保护个人信息安全，部分欧洲国家甚至已开始着手禁止ChatGPT等人工智能应用。

（二）算法模型日趋复杂，可解释性目标难实现

长期以来可解释性都是制约人工智能用在司法判决、金融信贷等关键领域的主要因素，时至今日问题尚未解决、且变得更为棘手。由于深度模型算法的复杂结构是黑盒，人工智能模型天然缺乏呈现决策逻辑进而使人相信决策准确性的能力。为提升可解释性，技术上也出现了降低模型复杂度、突破神经网络知识表达瓶颈等方法，但现实中效果有限。主要是因为当前模型参数越来越多、结构越来越复杂，解释模型、让人类理解模型的

难度变得极大，目前部分研究正朝借助人工智能解释大模型的方向探索。同时，由于近年来人工智能算法、模型、应用发展演化速度快，如何判断人工智能是否具备可解释性一直缺乏统一认知，难以形成统一判别标准。

（三）可靠性问题仍然制约人工智能关键领域应用

由于现实场景中环境因素复杂多变，人工智能难以通过有限的训练数据覆盖现实场景中的全部情况，因此模型在受到干扰或攻击等情况下会发生性能水平波动，严重时甚至可引发安全事故。尽管可通过数据增强方法等方式提高人工智能可靠性，然而由于现实场景的异常情况无法枚举，可靠性至今仍然是制约自动驾驶、全自动手术等关键领域应用广泛落地的主要因素。

（四）滥用误用人工智能，扰乱生产生活安全秩序

人工智能在对加速社会发展、提升生产效率等方面产生极大促进作用的同时，也出现了被滥用误用、恶意使用的现象，引起威胁社会安全、人身安全等负面事件。近年来，滥用误用人工智能方面，出现了物业强制在社区出入口使用人脸识别、手机应用扎堆推送雷同信息构筑信息茧房等问题。恶意使用人工智能方面，出现了利用虚假视频、图像、音频进行诈骗勒索、传播色情暴力信息等问题。

（五）模型和数据成为核心资产，安全保护难度提升

人工智能训练数据的获取以及模型开发已经逐渐变成重资产投入、重人力投入的工作，算法模型、参数、加工后的训练数据已成为核心资产，不免遭到觊觎。通过模型窃取、成员推理等攻击手段反向获取模型、数据，或者利用人工标注、数据存储等环节的安全管理漏洞套取数据的情况时有发生。

（六）网络意识形态安全面临新风险

人工智能的目标是模拟、扩展和延伸人类智能，如果人工智能只是单纯追求统计最优解，可能表现得不那么有“人性”；相反，包含一些人类政治、伦理、道德等方面观念的人工智能会表现得更像人、更容易被人所接受。事实上，为了解决人工智能面对敏感复杂问题的表现，开发者通常将包含着开发者所认为正确观念的答案加入训练过程，并通过强化学习等方式输入到模型中，当模型掌握了这些观念时，能够产生更能被人接受的回答。然而，由于政治、伦理、道德等复杂问题往往没有全世界通用的标准答案，符合某一区域、人群观念判断的人工智能，可能会与另一区域、人群在政治、伦理、道德等方面有较大差异。因此，使用内嵌了违背我国社会共识以及公序良俗的人工智能，可能对我国网络意识形态安全造成冲击。

三、人工智能安全政策与标准现状

在人工智能技术快速发展、应用规模迅速扩大的背景下，世界各国纷纷开展人工智能安全政策与标准相关工作。

（一）国内外人工智能安全战略与政策法规

1、联合国持续关注人工智能伦理安全

联合国教科文组织于2021年11月发布《人工智能伦理问题建议书》，旨在为和平使用人工智能系统、防范人工智能危害提供基础。建议书提出了人工智能价值观和原则，以及落实价值观和原则的具体政策建议，推动全球针对人工智能伦理安全问题形成共识。2023年3月31日，该组织号召各国立即执行《人工智能伦理问题建议书》。

2、欧盟严格人工智能监管，持续推进立法进程

欧盟专门立法，试图对人工智能进行整体监管。2021年4月，欧盟委员会发布了立法提案《欧洲议会和理事会关于制定人工智能统一规则（人工智能法）和修订某些欧盟立法的条例》（以下简称《欧盟人工智能法案》），在对人工智能系统进行分类监管的基础上，针对可能对个人基本权利和安全产生重大影响的人工智能系统建立全面的风险预防体系，该预防体系是在政府立法统一主导和监督下，推动企业建设内部风险管理机制，但如何判断评估人工智能系统风险等级是《欧盟人工智能法案》的重点和难点。

欧盟及时填补通用目的人工智能监管空白。2023年5月11日，欧洲议会的内部市场委员会和公民自由委员会通过了关于《欧盟人工智能法案》

的谈判授权草案，新版本补充了针对“通用目的人工智能”和GPT等基础模型的管理制度，扩充了高风险人工智能覆盖范围，并要求生成式人工智能模型的开发商必须在生成的内容中披露“来自于人工智能”，并公布训练数据中受版权保护的数据摘要等。

3、美国强调安全原则，鼓励企业行业自律

相较于欧盟，美国监管要求少，主要强调安全原则。美国参议院、联邦政府、国防部、白宫等先后发布《算法问责法（草案）》《人工智能应用的监管指南》《人工智能道德原则》《人工智能权利法案》《国家网络安全战略》等文件，提出风险评估与风险管理方面的原则，指导政府部门与私营企业合作探索人工智能监管规则，并为人工智能实践者提供自愿适用的风险管理工具。

美国鼓励企业依靠行业自律，自觉落实政府安全原则保障安全。美国企业通过产品安全设计，统一将美国的法律法规要求、安全监管原则、主流价值观等置入产品。以生成式人工智能企业提高内容安全水平为例，工作一般集中在三个方面，一是在产品设计阶段加入符合安全要求的定制化内容作为重点训练数据；二是在产品运行阶段的人机交互环节加入自动化内容过滤机制；三是在每个用户使用产品时置入隐藏的安全前提引导生成内容安全合规。

4、其他国家均构建各自人工智能战略

（1）英国：支持创新，建立监管框架，暂缓立法

2021年9月，英国数字、文化、媒体和体育部（DCMS）发布《国家人工智能战略》，旨在推动英国成为人工智能领域的大国。该文件阐述了英国人工智能战略愿景，并提出了人工智能发展、人工智能作为经济支柱、人工智能治理和监管的三个方面的核心目标与行动建议。

2021年12月，英国政府发布《国家网络战略2022》。该文件阐述了英国将如何巩固网络强国地位、保障网络安全、提升网络空间行动能力。其中明确了英国将积极引领人工智能等七项优先技术领域的安全发展。

2023年3月，英国政府发布了人工智能新监管框架的提案《支持创新的人工智能监管方法》。该文件围绕以下五个方面展开：一是自身安全、应用安全和健壮性，二是适度的透明和可解释性，三是公平性，四是问责制和治理，五是竞争和赔偿，并表示近期不会将上述原则立法。

（2）加拿大：聚焦人工智能应用的民众权益保护

2022年6月，加拿大政府公布《2022年数字宪章实施法案》，旨在加强对加拿大私营部门的管理，为负责任的人工智能开发和使用创建新规则。其中拟议法案《人工智能与数据法》敦促各公司在开发和部署人工智能系统时以减轻伤害和偏见风险为前提，进而维护加拿大民众的权益。

（3）俄罗斯：关注人工智能对国家安全影响

2019年，俄罗斯总统令批准《2030年前国家人工智能发展战略》，旨在加快推进俄人工智能发展与应用，确保国家安全，提升经济实力。

2020年，俄联邦政府批准《至2024年人工智能和机器人技术监管构想》，旨在积极探索俄罗斯法律、人、机器之间的相互适应关系，为人工智能和机器人技术的安全应用和法律监管提供指导。

（4）新加坡：关注人工智能应用安全治理和评估

2019年11月，新加坡金融管理局（MAS）宣布与多家金融机构共同设立Veritas计划框架，旨在帮助金融机构评估人工智能和数据分析解决方案，保证其遵循“公平、道德、负责和透明”的安全准则。

2022年5月，新加坡资讯通信媒体发展局（IMDA）和个人数据保护委员会（PDPC）共同发布人工智能安全治理评估框架和工具包—A.I.VERIFY，旨在结合人工智能系统的技术评估和程序检查，提高评估主体与利益相关者之间的透明度。

（5）日本：同时关注人工智能正面和负面影响

2022年4月，日本政府发布了《人工智能战略2022》，旨在推动人工智能克服自身社会问题、提高产业竞争力。其中提出以人为本、多样性、可持续三项原则，围绕社会安全、流行疾病、重大灾害等安全问题提出了具体方针。

5、我国高度重视，有效平衡发展和安全

2021年12月和2022年11月，国家互联网信息办公室先后发布《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》，针对利用人工智能算法从事传播违法和不良信息、侵害用户权益、操纵社会舆论等问题，加强安全管理，推进算法推荐技术和深度合成技术依法合理有效利用。

2023年4月，国家互联网信息办公室发布了《生成式人工智能服务管理办法（征求意见稿）》，统筹安全与发展，提出生成式人工智能产品或服务应当遵守的规范要求，保障相关技术产品的良性创新和有序发展。

（二）国内外人工智能安全标准

1、人工智能安全国际标准以基础通用为主

国际标准组织（ISO）在人工智能领域已开展大量标准化工作，并专门成立了ISO/IEC JTC1 SC42人工智能分技术委员会。目前，与人工智能安全相关的国际标准及文件主要为基础概念与技术框架类通用标准，在内容上集中在人工智能管理、可信性、安全与隐私保护三个方面。

在人工智能管理方面，国际标准主要研究人工智能数据的治理、人工智能系统全生命周期管理、人工智能安全风险管理等，并对相应的方面提出建议，相关标准包括ISO/IEC 38507:2022《信息技术治理 组织使用人工智能的治理影响》、ISO/IEC 23894:2023《人工智能 风险管理》等。

在可信性方面，国际标准主要关注人工智能的透明度、可解释性、健壮性与可控性等方面，指出人工智能系统的技术脆弱性因素及部分缓解措施，相关标准包括ISO/IEC TR 24028:2020《人工智能 人工智能中可信赖性概述》等。

在安全与隐私保护方面，国际标准主要聚焦于人工智能的系统安全、功能安全、隐私保护等问题，帮助相关组织更好地识别并缓解人工智能系统中的安全威胁，相关标准包括ISO/IEC 27090《人工智能 解决人工智能系统中安全威胁和故障的指南》、ISO/IEC TR 5469《人工智能 功能安全与人工智能系统》、ISO/IEC 27091《人工智能 隐私保护》等。

电气与电子工程师协会（IEEE）在人工智能安全方面主要聚焦伦理安全风险、可解释人工智能、深度学习评估、人工智能责任化等安全问题。最新的标准和报告有IEEE P7000系列标准、IEEE 2841-2022《深度学习评估过程与框架》，在研项目有IEEE P2840《责任化人工智能许可标准》、IEEE P2894《可解释人工智能的体系框架指南》等。

2、欧洲陆续发布多份人工智能安全指南文件及标准需求

欧洲电信标准化协会（ETSI）近期关注的重点议题包括人工智能数据安全、完整性和隐私性、透明性、可解释性、伦理与滥用、偏见缓解等方面，已发布多份人工智能安全研究报告，包括ETSI GR SAI 004《人工智能安全：问题陈述》、ETSI GR SAI 005《人工智能安全：缓解策略报告》等，描述了以人工智能为基础的系统安全问题挑战，并提出了一系列缓解措施与指南。

欧洲标准化委员会（CEN）、欧洲电工标准化委员会（CENELEC）

成立了新的CEN-CENELEC联合技术委员会JTC 21“人工智能”，并在人工智能的风险管理、透明性、健壮性、安全性等多个方面提出了标准需求。

3、美国关注可信任可解释研究，出台企业标准

美国国家标准与技术研究院（NIST）关注人工智能安全的可信任、可解释等问题。最新的标准项目有：NIST SP1270《建立识别和管理人工智能偏差的标准》，提出了用于识别和管理人工智能偏见的技术指南；NIST IR-8312《可解释人工智能的四大原则》草案，提出了可解释人工智能的四项原则；NIST IR-8332《信任和人工智能》草案，研究了人工智能应用安全风险与用户对人工智能的信任之间的关系；NIST AI 100-1《人工智能风险管理框架》，旨在为人工智能系统设计、开发、部署和使用提供指南。

2022年3月，谷歌更新《人工智能原则》，提出人工智能对社会有益、避免制造或加强不公平的偏见、建立并测试安全性、对人负责、结合隐私设计、坚持科学的高标准等原则。该文件同时声明谷歌不会将人工智能技术应用于武器开发，也不会将人工智能用于可能侵犯人权的活动。

2022年6月，微软发布新版《负责任人工智能标准》，提出公平性、可靠性和内部安全性、隐私和外部安全性、包容性、透明度和问责制六项基本原则，用于指导人工智能工作。

4、我国提前布局，即将出台多项人工智能安全标准

2020年7月，国家标准委、中央网信办、发展改革委、科技部、工业和信息化部联合印发了《国家新一代人工智能标准体系建设指南》，形成了标准支撑人工智能高质量发展新格局。

（1）研制人工智能安全基础标准

我国首个人工智能安全国家标准《信息安全技术 机器学习算法安全评估规范》即将发布，规定了机器学习算法技术在生存周期各阶段的安全要求，以及应用机器学习算法技术提供服务时的安全要求，并给出了对应评估方法。

2022年，全国信息安全标准化技术委员会（TC260）启动编制《信息安全技术 人工智能计算平台安全框架》国家标准，规范了人工智能计算平台安全功能、安全机制、安全模块以及服务接口，指导人工智能计算平台设计与实现。

（2）推动关键应用方向安全保护标准

在生物特征识别、智能汽车等人工智能应用领域，针对网络安全重点风险，多项国家标准已经发布。生物特征识别方向，发布了GB/T 40660—2021《信息安全技术 生物特征识别信息保护基本要求》，以及人脸、声纹、基因、步态等4项数据安全国家标准。智能汽车方向，发布了国家标准GB/T 41871—2022《信息安全技术 汽车数据处理安全要求》，有效支撑《汽车数据安全管理若干规定（试行）》，提升了智能汽车相关企业的数据安全水平。

四、人工智能安全标准需求分析

（一）人工智能安全属性定义和度量指标

随着人工智能安全工作开展，安全属性定义逐步形成一些共识，国际标准以及相关技术文件也给出了部分安全属性的定义描述，但不同文件之间仍有差异。可靠性、可解释性、公平性等重要安全属性定义的标准统一，以及其度量指标的规范化，将对人工智能的发展有着重要促进意义。

（二）用户输入数据安全保护相关规范

在人工智能通常收集用户输入数据用于训练的背景下，如何保障用户输入数据的安全亟需技术标准。根据操作场景的不同，用户输入的数据可能包含人脸、身份证号、家庭住址等个人信息，可能包括个人健康情况、情感状况等个人隐私，可能包括企业技术和经营活动有关的商业秘密，甚至可能包括国家秘密等。需要落实《数据安全法》《个人信息保护法》等法律法规，提出可以切实解决用户输入数据安全问题的相关标准规范。

（三）人工智能服务网络安全防护相关指南

围绕人工智能服务过程中，可能会面临的对抗样本攻击、爬山攻击、模型窃取、供应链攻击等新型攻击威胁，需要研究在数据集防护、算法模型保护、抗逆向攻击等方面的安全技术措施指南，帮助人工智能服务提供者保护业务数据以及人工智能模型参数等的机密性和完整性。

（四）人工智能安全评估相关规范

当前，随着人工智能在各行各业的应用逐步深入，需要以确认其基本安全水平作为提供产品或服务的基础。研究人工智能安全评估规范标准，将有助于确定人工智能产品或服务的安全水平，促进安全隐患提前防范，推动人工智能应用与行业的进一步融合发展。

（五）生成式人工智能安全标准

为应对生成式人工智能带来的安全挑战，以促进生成式人工智能发展为基本目标，统筹发展和安全，亟需针对生成式人工智能的网络安全问题出台专门标准，包括但不限于生成式人工智能训练数据安全、人工标注过程安全等方面的标准规范。

五、人工智能安全标准化工作建议

面对未来人工智能发展关键时期，我国人工智能安全标准化工作将继续发挥基础性、规范性、引领性作用。

（一）持续完善人工智能安全标准体系

面向人工智能变化快、安全挑战新等特点，加大研究力度，强化顶层设计，结合当前人工智能面临的安全风险，完善人工智能安全标准体系，统筹规划人工智能安全的基础共性、技术系统、管理服务、测试评估、产品应用等方面标准研制工作。

（二）大力开展基础共性安全标准研究

围绕现阶段人工智能安全发展所需的基础共性标准，加快开展标准化研制。一是加快推动《信息安全技术 机器学习算法安全评估规范》《信息安全技术 人工智能计算平台安全框架》等通用性标准编制发布。二是大力开展技术标准研究，围绕统一人工智能安全属性和度量指标、保护用户输入数据安全、人工智能服务网络安全防护等方面做好标准预研。

（三）加快出台产业发展急需安全标准

面向当前人工智能安全发展的痛点、堵点问题，聚焦以安全促发展，细化支撑《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理办法》等法律法规，加快推动《信息安全技术 生成式人工智能预训练和优化训练数据安全规范》《信

息安全技术 生成式人工智能人工标注安全规范》《信息安全技术 互联网信息服务深度合成安全规范》《基于个人信息的自动化决策安全要求》等标准的编制发布。

附录A

附录A：标准列表

A.1 国内人工智能安全相关标准列表

A.1.1 人工智能安全标准

负责/归口	标准类型	标准编号	标准名称	阶段
全国信安标委（TC 260）	国家标准	20211000-T-469	信息安全技术 机器学习算法安全评估规范	报批稿
全国信安标委（TC 260）	国家标准	20230249-T-469	信息安全技术 人工智能计算平台安全框架	征求意见稿
全国信标委人工智能分委会（TC28/SC42）	国家标准	20221791-T-469	人工智能 管理体系	立项
中国电子工业标准化技术协会（CESA）	团体标准	T/CESA 1193-2022	信息技术 人工智能 风险管理能力评估	发布

A.1.2与人工智能直接相关的安全标准

负责/归口	标准类型	标准编号	标准名称	阶段
全国信安标委（TC 260）	国家标准	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	发布
全国信安标委（TC 260）	国家标准	GB/T 38671-2020	信息安全技术 远程人脸识别系统技术要求	发布
全国信安标委（TC 260）	国家标准	GB/T 40660-2021	信息安全技术 生物特征识别信息保护基本要求	发布
全国信安标委（TC 260）	国家标准	GB/T 41819-2022	信息安全技术 人脸识别数据安全要求	发布
全国信安标委（TC 260）	国家标准	GB/T 41807-2022	信息安全技术 声纹识别数据安全要求	发布
全国信安标委（TC 260）	国家标准	GB/T 41806-2022	信息安全技术 基因识别数据安全要求	发布
全国信安标委（TC 260）	国家标准	GB/T 41773-2022	信息安全技术 步态识别数据安全要求	发布
全国信安标委（TC 260）	国家标准	GB/T 41871-2022	信息安全技术 汽车数据处理安全要求	发布
全国信安标委（TC 260）	国家标准	20230253-T-469	信息安全技术 基于个人信息的自动化决策安全要求	立项
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 41815.1-2022	信息技术 生物特征识别呈现攻击检测 第1部分：框架	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 41815.2-2022	信息技术 生物特征识别呈现攻击检测 第2部分：数据格式	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 41815.3-2023	信息技术 生物特征识别呈现攻击检测 第3部分：测试与报告	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 37036.3-2019	信息技术 移动设备生物特征识别 第3部分：人脸	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 37036.8-2022	信息技术 移动设备生物特征识别 第8部分：呈现攻击检测	发布
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	GB/T 5271.37-2021	信息技术 词汇 第37部分：生物特征识别	发布

负责/归口	标准类型	标准编号	标准名称	阶段
全国信标委生物特征识别分委会（TC 28/SC37）	国家标准	20221220-T-469	信息技术 生物特征识别 人脸识别系统应用要求	立项
中国通信标准化协会（CCSA）	行业标准	YD/T 4087-2022	移动智能终端人脸识别安全技术要求及测试评估方法	发布
中国通信标准化协会（CCSA）	行业标准	2023-0041T-YD	人工智能开发平台通用能力要求 第2部分：安全要求	立项
中国通信标准化协会（CCSA）	行业标准	2023-0039T-YD	面向人脸识别系统的人脸信息保护基础能力要求	立项
中国通信标准化协会（CCSA）	行业标准	——	人脸识别线下支付安全要求	草案
中国通信标准化协会（CCSA）	行业标准	2021-0630T-YD	电信网和互联网人脸识别数据安全检测要求	立项
上海市市监局	地方标准	——	人工智能数据通用安全要求	征求意见稿
上海市市监局	地方标准	——	人脸识别分级分类应用标准	草案
中国电子工业标准化技术协会（CESA）	团体标准	T/CESA 1124-2020	信息安全技术 人脸比对模型安全技术规范	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 110.1-2020	人工智能视觉隐私保护 第1 部分：通用技术要求	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 110.2-2022	人工智能视觉隐私保护 第2部分：技术应用指南	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 113-2021	生物特征识别服务中的隐私保护技术指南	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	T/AI 111-2020	生物特征模板的安全使用要求	发布
新一代人工智能产业技术创新战略联盟（AITISA）	团体标准	2023011205	信息技术 数字视网膜系统 第11部分：安全与隐私保护	草案

A.2 国外人工智能安全相关标准列表

负责/归口	标准类型	标准编号	标准名称	阶段
ISO IEC/JTC1/SC27	国际标准	ISO/IEC AWI 27090	Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems	立项(AWI)
ISO IEC/JTC1/SC27	国际标准	ISO/IEC AWI 27091	Cybersecurity and Privacy — Artificial Intelligence — Privacy protection	立项(AWI)
ISO IEC/JTC1/SC27	国际技术报告	ISO/IEC TR 27563	Security and privacy in artificial intelligence use cases — Best practices	发布
ISO IEC/JTC1/SC42	国际标准	ISO/IEC 22989:2022	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	发布
ISO IEC/JTC1/SC42	国际标准	ISO/IEC 23894:2023	Information technology — Artificial intelligence — Guidance on risk management	发布
ISO IEC/JTC1/SC42	国际技术报告	ISO/IEC TR 24368:2022	Information technology — Artificial intelligence — Overview of ethical and societal concerns	发布
ISO IEC/JTC1/SC42	国际预研项目	ISO/IEC PWI 17866	Artificial intelligence — Best practice guidance for mitigating ethical and societal concerns	预研(PWI)
ISO IEC/JTC1/SC42	国际技术报告	ISO/IEC DTR 5469	Artificial intelligence — Functional safety and AI systems	报批(DTR)
ISO IEC/JTC1/SC42	国际标准	ISO/IEC FDIS 42001	Information technology — Artificial intelligence — Management system	报批(FDIS)
ISO/TC 199	国际技术报告	ISO TR 22100-5:2021	Safety of machinery — Relationship with ISO 12100 — Part 5: Implications of artificial intelligence machine learning	发布
ISO/TC22/SC32	国际标准	ISO/AWI PAS 8800	Road Vehicles — Safety and artificial intelligence	立项(AWI)
IEEE	国际协会标准	IEEE 2790-2020	IEEE Standard for Biometric Liveness Detection	发布
IEEE	国际协会标准	IEEE 2801-2022	IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence	发布
IEEE	国际协会标准	IEEE P2894	Guide for an Architectural Framework for Explainable Artificial Intelligence	草案

负责/归口	标准类型	标准编号	标准名称	阶段
IEEE	国际协会标准	IEEE P2986	Recommended Practice for Privacy and Security for Federated Machine Learning	草案
IEEE	国际协会标准	IEEE P3156	Standard for Requirements of Privacy-preserving Computation Integrated Platform	草案
IEEE	国际协会标准	IEEE P3169	Standard for Security Requirement of Privacy-Preserving Computation	草案
IEEE	国际协会标准	IEEE 7002-2022	IEEE Standard for Data Privacy Process	发布
IEEE	国际协会标准	IEEE P7012	Standard for Machine Readable Personal Privacy Terms	草案
NIST	美国标准	NIST. AI.100-1	Artificial Intelligence Risk Management Framework	发布
NIST	美国标准	NIST. IR.8269	A Taxonomy and Terminology of Adversarial Machine Learning	草案
NIST	美国标准	NIST. IR.8312	Four Principles of Explainable Artificial Intelligence	发布
NIST	美国标准	NIST. IR.8330	Trust and Artificial Intelligence	发布
NIST	美国标准	NIST. IR.8367	Psychological Foundations of Explainability and Interpretability in Artificial Intelligence	发布
NIST	美国标准	NIST. SP.1270	Towards a Standard for Identifying and Managing Bias in Artificial Intelligence	发布
ETSI	欧洲标准	ETSI GR SAI 001	AI Threat Ontology	发布
ETSI	欧洲标准	ETSI GR SAI 002	Data Supply Chain Security	发布
ETSI	欧洲标准	ETSI GR SAI 003	Security testing of AI	草案
ETSI	欧洲标准	ETSI GR SAI 004	Problem Statement	发布
ETSI	欧洲标准	ETSI GR SAI 005	Mitigation Strategy Report	发布
ETSI	欧洲标准	ETSI GR SAI 006	The role of hardware in security of AI	发布