人工智算法安全评估规范

编制说明

标准起草工作组

2025 年 4 月

目　录

1 必要性

本标准在于构建一套系统化、标准化的人工智能算法安全评估框架，以确保人工智能技术的安全性、可靠性和可控性，从而在广泛的应用领域中推动其健康、有序发展。本标准为算法开发者和应用者提供了一套清晰的技术标准和评估方法，确保在算法设计、训练和部署过程中能够充分考虑安全性、透明性和可解释性，从而有效减少算法偏见和安全漏洞的风险。其次，本标准有助于规范人工智能算法的开发流程，推动行业在算法安全性方面的协同创新，建立起健全的算法治理机制，为社会提供更可靠、更安全的人工智能应用。再次，本标准通过建立统一的评估标准，将有效促进人工智能技术的广泛应用，推动数字经济的高质量发展，提升产业的智能化水平，激发经济发展的新动能。

2 工作简述

2.1 任务来源

本标准根据四川省网络空间安全协会数据安全团体标准制修订计划立项，由四川省网络空间安全协会归口，由中国电子科技集团公司第三十研究所牵头组织编制。

2.2 起草单位

本标准牵头起草单位：中国电子科技集团公司第三十研究所；

本标准参加起草单位：中国电子科技网络信息安全有限公司、成都四方数安信息技术有限公司、全域数据信息安全重点联合实验室西南实验室。

2.3 起草过程

2024 年 8 月，中国电子科技集团公司第三十研究所向四川省网络空间安全协会提交《人工智能算法安全评估规范》团体标准项目建议书；

2024 年 12 月，由四川省网络空间安全协会邀请专家对《人工智能算法安全评估规范》立项评审，标准立项，成立标准起草工作组；

2025 年 3 月，完成了团体标准《人工智能算法安全评估规范》草案稿编写；

2025 年 4 月，专家对意见修改稿进行了评审，团体标准《人工智能算法安全评估规范》文本质量达到征求意见稿发布要求。

3 标准编制原则和主要内容

3.1 编制原则

本标准的制定工作遵循合规性的原则。

在标准制订过程中，严格遵循国家已颁布的相关法律法规，如《网络安全法》、《数据安全法》和《个人信息保护法》，并与相关国家标准《GB/T 41867-2022 信息技术 人工智能 术语》、《GB/T 42888-2023 信息安全技术 机器学习算法安全评估规范》和《GB/T 45225-2025 人工智能 深度学习算法评估》等保持一致。

3.2 主要内容

本标准共分为8章，包括范围、规范性引用文件、术语和定义、概述、对抗攻击测试方法、判别式人工智能算法安全评估要求和评估方法、生成式人工智能算法安全评估要求和评估方法、人工智能算法安全评估实施。

1．范围

2．规范性引用文件

3．术语和定义

4．概述

5．对抗攻击测试方法

6．判别式人工智能算法安全评估要求和评估方法

7．生成式人工智能算法安全评估要求和评估方法

8．人工智能算法安全评估实施

4 技术论证与效果

在评估规范的制定过程中，广泛参考了《GB/T 41867-2022 信息技术 人工智能 术语》、《GB/T 42888-2023 信息安全技术 机器学习算法安全评估规范》和《GB/T 45225-2025 人工智能 深度学习算法评估》等相关国家标准，确保本标准与现行的相关国家标准保持一致，避免冲突。

本标准为人工智能算法安全的规范化管理提供了重要依据，有助于增强人工智能算法在各个环节的安全性和合规性，进一步推动人工智能算法安全体系的建设。本标准通过系统性、可操作性的指导，协助组织识别人工智能算法各阶段的潜在安全风险，提供安全性评估手段，从而减少数据泄露、篡改、模型滥用等事件的发生。本标准能够提升社会对人工智能算法安全的整体认知水平，推动人工智能算法治理的高效执行，并支持人工智能算法的安全可信和可持续应用，为企业和社会的智能化转型提供坚实的安全保障。

5 对标情况

《网络安全法》作为我国网络空间安全的基本法律，明确了网络运营者在保障网络运行安全、网络信息安全方面的主体责任。本标准在制定过程中，充分对照《网络安全法》关于关键信息基础设施保护、网络产品和服务安全、网络数据处理等方面的要求，围绕人工智能算法可能引发的网络安全风险，提出相应的评估技术方法，强化算法安全可控性。通过标准实施，助力算法开发和应用单位落实网络安全主体责任，提升智能系统整体防护水平。

《数据安全法》强调对数据全生命周期的管理，明确数据处理活动的安全责任与义务。本标准在设计中参照该法关于数据分类分级保护、安全风险监测、应急响应等规定，结合人工智能算法在数据获取、训练、推理阶段中存在的数据泄露、污染等安全隐患，构建了针对性评估框架。标准的实施将有助于提升人工智能算法对数据安全风险的感知与响应能力，推动智能系统在数据处理环节的合规性、安全性和可控性。

《GB/T 41867-2022 信息技术 人工智能 术语》作为人工智能领域基础术语标准，为统一行业概念、提升沟通效率提供了依据。本标准在术语使用和定义上，广泛采纳并严格对照该

标准所界定的相关术语，确保概念表达准确、一致，便于与其他人工智能领域标准实现兼容对接。

《GB/T 42888-2023 信息安全技术 机器学习算法安全评估规范》为我国机器学习算法安全评估提供了系统方法，是人工智能算法安全评估的重要技术支撑之一。本标准在制定过程中充分参考其结构框架与评估流程，结合深度学习、大模型等技术特点进行拓展与深化，提出更具适应性的细化评估内容，进一步丰富了我国人工智能算法安全评估标准体系。

《GB/T 45225-2025 人工智能 深度学习算法评估》聚焦于深度学习算法的评估框架与指标体系，覆盖模型性能、可靠性、可解释性等多个维度。本标准在制定过程中与之形成互补关系，将其侧重的算法能力评估内容与本规范关注的安全风险评估内容结合，构建涵盖攻击防护、模型鲁棒性、数据安全等方面的完整安全评估流程。标准之间的一致性和互补性，有助于推动人工智能技术在安全与能力评估领域的系统化发展，形成全面的标准支撑体系。

6 标准实施建议

建议从制度建设、人员培训、过程监督与持续改进四方面推进标准实施。首先，结合单位实际制定配套制度与实施细则，确保标准有章可循。其次，开展有针对性的培训，提升相关人员对标准的理解与执行能力。第三，建立全过程监督机制，及时发现问题并加以纠正，确保标准在实践中不偏离、不变形。最后，注重实施效果评估与经验总结，结合反馈不断优化实施路径，实现标准的动态更新与持续改进，确保标准真正落地、发挥实效。

7 需要说明的主要问题

本标准在编制过程中未出现需要说明的主要问题。

8 其他说明事项

无