ICS 35.240
CCS L 77

团　体　标　准

T/CFEII 0019—2024

人工智能融合应用安全可信管理指南

Guidelines on security and trustworthy management of artificial intelligence integrated applications

2024 - 07 - 22 发布　　2024 - 07 - 22 实施

中国电子信息行业联合会　发布

目　次

前　言

本标准按照 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。

本标准由中国电子信息行业联合会提出并归口。

请注意本标准的某些内容可能涉及专利，本标准的发布机构不承担识别专利的责任。

本标准起草单位： 国家工业信息安全发展研究中心、西安交通大学、清华大学公共管理学院、北京格灵深瞳信息技术股份有限公司、北京工业大学、北京北信源软件股份有限公司、北京奇虎科技有限公司、商汤集团有限公司、蚂蚁科技集团股份有限公司、北京百度网讯科技有限公司、北京昇腾人工智能生态创新中心、北京晴数智慧科技有限公司、北京神州绿盟科技有限公司、广州广电信息安全科技有限公司、山石网科通信技术股份有限公司。

本标准主要起草人：张瑶、王淼、邱惠君、李卫、刘永东、张若丹、张振乾、李天舒、邱颖昭、沈心然、沈超、赵静、陈天博、胡俊、高曦、邹权臣、胡正坤、林冠辰、郭建领、李天哲、张晴晴、顾杜娟、梁添才、吴疆。

引　言

人工智能作为数字化转型的重要赋能技术之一，正在与金融、医疗、工业、交通等领域融合发展。近年来，人工智能应用范围加速拓展，行业渗透率迅速提升。与此同时，人工智能存在算法、数据、攻防、应用、管理等方面的风险，已暴露出信息泄露、数据滥用、偏见歧视、实施诈骗等安全风险，引发了社会各界对人工智能安全问题的广泛关注。

本项目在充分研究分析国内外人工智能安全风险治理原则、政策、标准等文件的基础上，明确人工智能融合应用安全治理的原则和目标，提出能够推动相关治理原则落地的治理措施和实践指南，旨在推动人工智能健康有序发展。

本标准旨在为各类主体提供人工智能安全可信管理指引，《人工智能融合应用安全可信度评估规范（系统版）》和《人工智能融合应用安全可信度评估规范（组织版）》将分别提供面向人工智能系统和面向人工智能开发、应用单位的安全可信度评估框架、方法与内容，三项标准共同形成人工智能融合应用安全可信系列标准。

对标准中的具体事项，法律法规另有规定的，需遵照其规定执行。

人工智能融合应用安全可信管理指南

1 范围

本文件提出了人工智能融合应用安全可信管理的指导思想和治理原则，从组织制度、数据管理、模型开发、性能质量、防御机制、部署运行六个维度给出了人工智能安全可信管理的实践指南。

本文件适用于各类主体开展的人工智能设计、开发、部署、使用、维护等过程中的安全治理。

2 规范性引用文件

本文件没有规范性引用文件。

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 artificial intelligence

<学科>人工智能系统相关机制和应用的研究与开发。
[来源：GB/T 41867-2022,3.1.2]

3.2

人工智能系统 artificial intelligence system

针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统。

3.3

人工智能风险 artificial intelligence risk

人工智能的不确定性对任务和目标的影响。
[来源：ISO/IEC 22989:2022,3.5.11，有修改]

3.4

算法 algorithm

<人工智能>用于求解分类、推理、预测等问题，明确界定的有限且有序的规则集合。
[来源：T/CESA 1193-2022,3.1.8，有修改]

3.5

模型 model

<人工智能>针对特定问题或任务，基于输入数据，生成推理或预测的计算结构。
注：一个人工智能模型是基于人工智能算法训练的结果。
[来源：T/CESA 1193-2022,3.1.9，有修改]

3.6

安全性 security

<人工智能>系统免受恶意攻击、保护数据或阻止未经授权访问的能力。
[来源：ISO/IEC TR 24028:2020,3.35，有修改]

3.7

可信性 trustworthiness

<人工智能>系统以可验证的方式，满足利益相关者期望的能力。

注 1：根据背景或行业以及具体产品或服务、数据和使用的技术，适用不同的特征，需要通过客观证据证明，以确保满足利益相关者的期望。

注 2：可信的特征包括可靠性、可用性、弹性、安全、隐私性、可问责、透明性、完整性、真实性、质量等。

注 3：可信是一种属性，可应用于服务、产品、技术、数据和信息，在治理背景下也可应用于组织。

[来源：ISO/IEC TR 24028:2020,3.42，有修改]

3.8

人工智能生命周期 artificial intelligence lifecycle

人工智能系统从设计到下线的过程，包括设计开发、验证测试、部署上线、运行维护、退役下线等阶段。

[来源：ISO/IEC 22989:2022，有修改]

3.9

偏见 bias

<人工智能可信赖>对待特定对象、人或群体时，相较于其他对象出现系统性差别的特性。

注：“对待”指任何一种行动，包括感知、观察、表征、预测或决定。

[来源：GB/T 41867-2022,3.4.10，有修改]

3.10

伦理 ethics

<人工智能>开展人工智能技术基础研究和应用实践时遵循的道德规范和准则。

[来源：GB/T 41867-2022,3.4.8]

3.11

公平性 fairness

<人工智能>尊重既定事实、社会规范和信仰，且不受偏袒或不公正歧视影响的对待、行为或结果。

注 1：对公平性的考虑是与环境高度相关的，并且因文化、代际、地理和政治观点而异。

注 2：公平不等于没有偏见。偏见并不总是导致不公平，不公平可能是由于偏见以外的因素引起的。

[来源：GB/T 41867-2022,3.4.1]

3.12

可解释性 interpretability

<人工智能>系统以人能理解的方式，表达影响其（执行）结果的重要因素的能力。

注：可解释性理解为对“原因”的表达，而不是尝试以“实现必要的优势特性”做出争辩。

[来源：GB/T 41867-2022,3.4.3]

3.13

鲁棒性 robustness

<人工智能>系统在任何情况下都保持其性能水平的特性。
[来源：GB/T 41867-2022,3.4.9]

4 总则

4.1 指导思想

综合考虑相关法律法规、治理经验和行业特性，人工智能融合应用安全可信管理应遵循以下指导思想。

a) 明晰各方责任：需建立责任体系，明确各方主体的责任，树立责任意识，引导组织主动应对风险，营造负责任地开发、部署和使用人工智能的良好氛围。

b) 管理技术结合：采用过程管理与技术测试相结合的方式，从组织架构、制度策略等管理层面及数据算法安全、攻防安全检测等技术层面应对人工智能伦理与安全风险。

c) 贯穿生命周期：需考虑设计开发、测试验证、部署上线、运行维护、退役下线等人工智能全生命周期各环节，建立覆盖事前、事中、事后的全流程监管制度。

d) 动态平衡调整：坚持发展与监管并重，建立动态跟踪调整机制，根据人工智能技术发展和安全风险的变化对管理规则进行及时调整。

4.2 管理原则

人工智能安全风险是指设计、开发和使用人工智能技术过程中在安全方面存在的不确定结果，按照风险来源的不同，大致可分为数据安全风险、算法模型风险、攻防风险、应用风险等。综合考虑人工智能融合应用安全风险特点，人工智能安全可信管理的原则应涵盖以下方面。

a) 公平性：应采取措施保障数据集完整、无偏见，并在模型的设计与开发时遵循公平原则，避免采样偏差、样本选择偏差和特定群体的不公平对待。

b) 透明性：应保持对监管部门、客户和用户的透明度，提供有关系统内部运行和决策过程的可见性和公开性，包括解释数据收集和使用方式、算法选择和配置、决策依据等。

c) 可解释性：应采取措施保障系统以人能理解的方式，表达其决策和推断过程，并解释各构建块对系统结果的贡献，以便用户和利益相关者能够理解系统决策的基础和逻辑。

d) 隐私性：应采取措施以便在数据采集、传输、存储和使用的过程中，对个人身份信息和敏感数据进行适当的保护，遵守相关的隐私法规和政策，保障数据主体的知情权和选择权。

e) 可追溯性：应正确记录、保管与人工智能相关的文档、数据来源与操作记录以及算法配置信息，用于验证系统的合规性、可靠性和数据处理过程的一致性，并用以接受审计。

f) 可问责：应建立明确的问责机制和责任体系，确保各环节的责任人可确认，责任可追溯。

g) 安全性：应采取相应的安全措施和防护机制，使人工智能的系统、数据和通信受到充分的保护，能够抵抗恶意攻击、数据泄露和未经授权访问等安全威胁。

h) 准确性：应采取措施提升人工智能算法结果的可信度和准确度，在多个测试和验证环境中验证其性能，采用相关度量和评估方法验证算法的准确度。

i) 鲁棒性：应确保人工智能在面对异常情况、噪声、攻击或在未知环境条件下，仍然能够保持稳定和可靠的性能，具备自适应能力，以应对各种挑战和变化。

j) 可控性：应采取措施保障人类在使用人工智能时具有充分的控制权和管理权，提供适当的界面和控制机制，允许用户指导、监督和干预系统的行为和决策。

k) 无害性：应确保人工智能设计遵循“以人为本”原则，尊重“人类自主”，其目的是增进人类福祉，并避免被恶意使用，应遵循伦理准则和法律要求，不应对人类、社会或环境造成伤害或危害。

l) 可持续性：应在人工智能生命周期中重视减少对环境的影响，优化资源利用效率、管理碳足迹、遵守环境保护法律法规，推动可持续发展目标的实现。

5 实践指南

5.1 组织制度

应从组织层面明确设立人工智能安全可信相关负责机构，构建责任体系以明确相关人员的责任分工，从文档管理、日志记录、风险管理、应急计划、多方参与、信息披露、人员培训等角度建立相关制度。

5.1.1 机构设置

应从组织层面明确建立人工智能安全风险管理、监督、落实机构。

a) 从组织层面明确建立人工智能安全风险管理和监督机构，如人工智能安全风险管理委员会，负责相关研究、布置、总结工作，由高层领导兼任委员会成员。可通过其他组织机构（如科技伦理委员会、数据安全治理机构、质量管理机构等）覆盖相关的职能。

b) 从组织层面明确建立人工智能安全风险管理工作落实机构，如人工智能安全风险管理工作组，负责落实具体职能，开展日常工作。

c) 从人工智能项目层面设置专门的人工智能安全风险审查人员，负责对具体项目进行风险管理、评估、审查和提醒，并应对其工作年限、职称、培训时长等作出要求。安全风险审查人员应该与开发和部署模型的人员独立。

5.1.2 责任体系

应建立明确的责任体系，明确定义人工智能系统涉及的人员角色、职责、分工，并建立追责机制，以确保问责制度有效落实。

a) 对人工智能系统涉及的所有或相关核心岗位（包括但不限于高级管理人员、项目管理人员、产品经理、设计人员、开发人员、数据管理人员等），设置人工智能安全风险管理的岗位职责。

b) 高级管理人员应确保组织对人工智能安全风险管理工作的支持，制定和推动全面的安全策略，保障人工智能安全政策的执行。

c) 人工智能项目管理人员应将人工智能安全纳入项目管理的范围，在项目进行过程中对相关安全风险进行管理，保障项目的交付与运行中的安全可信。

d) 人工智能项目产品经理应保障系统需求分析和功能设计满足安全要求，与团队合作提升系统安全性，对产品功能进行安全性评估，并持续改进产品功能上的安全性。

e) 人工智能系统的设计人员应在产品设计阶段充分考虑安全风险，将治理原则落实到产品设计中，为系统设置合理、明确、可持续的目标。

f) 人工智能系统的开发人员应在系统开发阶段充分考虑安全风险，将治理原则落实到产品开发中，对所采用模型、编写代码的安全性负责，及时应用安全更新，及时做好文档和日志记录，积极响应安全事件。

g) 人工智能系统的数据管理人员应在数据管理全流程中充分考虑安全风险，对数据完整性、准确性等进行测试，准确记录相关数据操作，监督其他人员对数据的使用。

h) 应设置专门的安全审查员，统筹系统设计、开发、运行等过程中的安全风险管理工作，负责对系统安全性进行审查和评估。

i) 应对相关岗位人员因未按规定履行职责产生安全风险时应负的责任作出规定。

j) 应建立追责机制，确保在系统运行出现问题时能够追溯到相关责任人。

5.1.3 文档管理

应建立合理的文档管理制度并落实，对人工智能系统的设计开发、运行维护等环节的相关文档进行统一管理和备份。

a) 应对人工智能系统生命周期各环节的相关文档进行存档，如项目需求说明书、项目实施方案、需求变更说明书、项目自查和审查报告、测试报告、利益相关者沟通文档、信息披露文档、风险管理文档、算法机制机理说明文档等。

b) 文档管理应覆盖设计开发、验证测试、部署上线、运行维护、退役下线等全生命周期的重点环节。

c) 当系统功能和性能目标需进行变更时，需要经相关负责人签字，并保存相关的变更文档。

d) 应建立定期审查文档的制度，定期审查相关文档的完整性、可用性和有效性。

5.1.4 日志记录

应建立合理的日志记录制度并落实，在人工智能整个生命周期中详细记录设计开发过程以及对模型和系统所做的更改。

a) 应对需要进行日志记录的环节和操作做出明确要求，覆盖设计开发、验证测试、部署上线、运行维护、退役下线等全生命周期的重点环节。

b) 应通过日志记录对系统的运行环境、训练和推理过程、异常事件、用户操作等内容进行记录。

c) 应对日志记录的格式、内容、存储时间等做出明确的规定。

5.1.5 风险管理

应制定和实施专门、持续的风险管理制度并落实，在立项前、项目进行中以及上线后系统地识别、分析和减轻风险，并进行持续监测、验证和修正。

a) 应按照 ISO/IEC 23894:2023(E)、ISO/IEC 31000:2018(E)及 ISO/IEC 42001:2022(E)的要求开展人工智能风险管理并获取相关认证。

b) 应在组织层面建立完善的风险管理制度，定期审计和改进。

c) 应在立项前对潜在安全风险进行识别和分析，涵盖系统对个人、组织、社会、环境等方面造成的影响，并采取预防措施。

d) 应在项目进行中持续进行风险识别和分析，实施风险减轻措施，及时调整管理计划。

e) 应在系统上线后持续监测风险，修复漏洞并提升风险管理能力，保障安全运行。

5.1.6 应急计划

应制定和实施持续的风险监测机制和应急响应计划并落实，包括异常情况出现时中断系统以及避免负面影响的相应机制。

a) 应建立异常状况监测预警机制，对系统运行状态、异常事件等进行实时监测、预警和报告。

b) 应制定应急响应计划，明确系统故障、数据泄露、网络攻击等应急情况的处理步骤和分工。

c) 应通过应急演练、人员培训等活动使项目团队成员熟悉应急响应流程，掌握必要的技能。

d) 应设置专门的应急响应团队负责事件响应和处理，及时发现和应对安全威胁。

e) 应制定数据备份和恢复机制，定期备份重要数据，并在异常情况发生时及时恢复数据，保障系统迅速恢复正常运行。

f) 对于超出风险承受能力的系统，应该立即停用。

5.1.7 多方参与

应建立利益相关者参与机制并落实，在整个人工智能生命周期中考虑来自不同利益相关方的视角。

a) 应建立利益相关方参与机制，对人工智能系统设计、开发、测试、部署的全生命周期中涉及的利益相关方（如管理人员、项目经理、开发团队、学术界、业界、用户、监管机构、受影响的群体等）及可能产生的影响进行分析。

b) 在界定和分析利益相关方时，尽可能地考虑特定的人群（例如，不同性别、不同年龄、残疾人）。

c) 在整个生命周期中，均保持与各利益相关方的持续沟通，不断收集利益相关者对系统的看法、期望和改进建议。

5.1.8 信息披露

应建立完整的信息披露机制，向监管部门、客户及用户定期披露数据来源、规模、类型、算法机制机理等信息。

a) 应建立人工智能模型基本信息披露机制，以清晰、易懂、充分的方式，向用户提供数据的基本属性、算法机制机理、系统运行逻辑、潜在风险情况等信息，帮助用户充分理解并针对其做出相关决策。

b) 应建立人工智能模型评测信息披露机制，向用户披露模型的准确性、鲁棒性、安全性、公平性、可解释性等相关维度的评估情况。

5.1.9 人员培训

应对相关人员进行人工智能安全风险管理意识教育和培训，使其能够按照相关政策、程序和协议履行职责。

a) 应对系统相关员工进行人工智能安全风险种类与应对措施、岗位人工智能安全风险责任、相关法律法规等方面的培训，提升其人工智能安全风险管理意识。

b) 相关培训的人员应该覆盖组织管理层、中层领导及项目负责人、人工智能项目相关员工等，包括系统设计、开发、训练、测试、部署，数据采集、标注、处理、管理等所有相关人员。

c) 应对系统开发人员进行安全培训，提高其人工智能安全技能水平。

d) 应对数据标注人员进行专业培训，培训内容应包括标注指南、质量控制以及数据隐私保护要求，保障标注的准确性和安全性。

e) 应通过相关培训增强模型开发、训练、部署等相关人员的公平性意识。

f) 应对培训内容、参加人员、课时数和培训时间等进行规定，并对每次培训的相关信息进行记录。

5.2 数据管理

应在数据处理全生命周期中采取相关措施，保持数据标注的准确性、数据的精准性、无偏性、代表性、可追溯性和合规性。

5.2.1 标注准确性

应通过培训考核标注人员、明确标注规则、评估标注质量等方法保障数据标注的质量和准确性。

a) 应将标注人员职能划分为数据标注、数据审核等，开展定期培训和考核。

b) 应设置明确的标注规则，对标注目标、数据格式、标注方法、质量指标、安全性要求等内容作出规定，以降低标注错误和不一致性的风险。

c) 应设置专门的数据审核人员，通过抽检或逐条检查等方式，对数据标注的质量进行验证和检查，内容不准确或不符合质量要求的，应重新标注。

5.2.2 数据精准性

应通过审核检查、数据预处理、及时更新来保障数据集具有一定的时效性和真实性。

a) 应对数据集存在的错误数据、反常数据、逻辑关系不符、人为添加恶意数据等情况进行检查和审核，确保数据具有一定准确性。

b) 应对数据集进行预处理，去除重复、无效、错误数据。

c) 应对数据集的时效性进行定期审查，并及时更新数据。

5.2.3 数据无偏性

应确保采取了相关措施降低数据可能存在的偏见、不平等和其他社会问题。

a) 应对数据采集、标注、管理等人员进行无偏见培训和审查。

b) 应根据系统目标和应用场景梳理可能存在的数据偏见风险，分析数据集的多样性需求。

c) 应通过自评估或第三方评估的形式对数据集的多样性、丰富性、客观性等进行审查，以避免歧视和不公正等问题。

d) 应对少数群体数据样本进行调整和处理，以避免由于数据集缺乏多样性、不具代表性等而产生偏差的风险。

5.2.4 数据代表性

应对数据的代表性进行评估并采取相关措施，保障所使用的数据集能够完整地代表使用系统的群体。

a) 在确定所需的数据集时，应该较为全面地分析和考虑系统目标特性及所有预期应用场景的需求。

b) 应对数据集进行预处理，去除不相关数据。

c) 应对数据的代表性进行分析和评估，确保数据集中包含特定对象的代表性样本，并符合一定的统计属性。

d) 应确保验证数据集、测试数据集与训练数据集具有一致的特征。

5.2.5 数据可追溯

应明确记录数据来源和操作，保障数据的来源可追溯。

a) 应对采用的所有数据来源进行完整和明确的记录。

b) 应对从第三方获取的数据集进行潜在风险分析并记录。

c) 应对每次数据标注、审核等行为进行详细记录，包括时间、人员、内容和结果。

d) 应对数据的过滤、编辑、提取、转换等操作进行记录。

e) 应对数据操作目的、方法、参数设定和结果等进行记录。

5.2.6 数据合规性

应采取相关措施保障组织的数据采集和使用符合数据安全、个人隐私保护、知识产权保护等相关法律法规。

a) 在采集和使用个人数据之前，应通过设置用户告知、授权同意等确保数据主体了解其数据将被如何使用，并在用户请求数据删除时进行正确、完整的删除。

b) 应对包含个人信息的数据进行脱敏处理，确保个人信息不被泄露。

c) 应根据系统目标和技术需要，确定数据种类和数量的最低需求量，只采集必须范围内的数据，不采集无关数据。

d) 若系统收集用户的输入数据用于模型训练，应进行事先告知并获取同意。

e) 应对数据集的访问实施严格的访问控制和权限管理。

f) 应采取加密存储、定期备份等数据安全措施，防止数据泄露、篡改或丢失。

g) 应定期对数据集开展监控审计和合规审查，确保数据集始终符合相关法律法规的要求。

h) 系统使用的第三方数据集应包含清晰的版权声明和许可协议，知识产权归属和使用条件明确。

5.3 模型开发

应在模型开发过程中做好目标设计和版本管理工作，并采取相关措施保障算法可解释性、鲁棒性和公平性。

5.3.1 目标设计

系统目标应与增进人类福祉、促进可持续发展的理念保持一致，应采取措施保障系统使用过程中对未成年人、老年人等群体友好。

a) 应对系统的目标进行清晰、明确的定义，并对涉及的关键术语和概念进行定义。

b) 应持续监测系统运行，使其在整个生命周期都满足预期目标。

c) 应对人工智能可能对环境生态系统的可持续性产生的影响进行预先评估。

d) 应确保负责人工智能系统的相关方(例如人工智能治理委员会、人工智能系统开发者、所有者、审查者)在决策过程中考虑了可持续发展问题。

e) 应针对老年人、未成年人等特殊群体进行优化设计，并按需求提供适老化模式、无

障碍模式、青少年模式等以方便特殊群体使用。

5.3.2 版本管理

应进行人工智能系统的版本管理，对模型不同版本的关键信息进行记录，提升系统开发训练、测试、调优等全流程的可追溯性。

a) 应对版本标识与命名规范、版本控制工具、关键信息记录、更新日志和注释与定期备份和存档等做出详细规定。

b) 应选择适合的版本控制工具，用于管理模型的源代码、配置文件和训练数据等，确保有效跟踪和记录模型版本的变更历史。

c) 应对每个模型版本记录关键信息，包括但不限于：版本标识和名称、模型训练时间、模型架构、训练数据集、参数和超参数设置、训练过程、性能评估指标、更新记录、作者和负责人等信息。

d) 应对模型重要版本的源代码和关键信息等进行备份。

5.3.3 算法可解释

应采取相关措施保障人工智能算法的运行和决策原理可解释。

a) 在系统开发过程中，应优先选择具有较高可解释性的算法和模型架构，平衡模型性能和解释性，并根据应用场景做出合理选择。

b) 应通过特征重要性分析、可视化技术、模型解释性工具等方法分析影响决策的关键特征和因素，提升模型可解释性。

c) 应在模型部署前通过敏感性分析、相关性分析、可解释性归因、训练探针模型等技术对模型的可解释性进行测评，确保模型具备较高的可解释性程度。

d) 应对模型的目标设计、基本原理、风险评估和权衡过程等进行详细记录。

5.3.4 算法鲁棒性

在运行环境发生变化时，仍可以按照预期保持一致的性能水平，保障算法在各类部署环境下的表现符合鲁棒性要求。

a) 应结合部署环境、目标、功能等对模型的鲁棒性需求进行分析。

b) 应采用数据增强、对抗性训练等方式对模型进行鲁棒性训练，以提升模型鲁棒性。

c) 应采用对抗样本测试、噪声测试、容错能力测试等方法对模型进行鲁棒性检测，确保模型鲁棒性能够满足部署需要。

d) 应通过适配多种深度学习框架、操作系统、硬件架构来保障系统的鲁棒性。

5.3.5 算法公平性

应对算法模型可能存在的歧视、偏见进行分析，制定公平性目标以及测试方法，以降低人工智能系统潜在的歧视和偏见。

a) 应根据实际应用场景和需求明确设置公平性指标，如偏见、歧视。

b) 应根据公平性目标构建包含不同特征的数据集，在训练过程中评估模型在不同群体之间的表现并不断改进。

c) 应在上线前对模型在公平性敏感领域的输出的差异性进行自测和第三方测试，确保在可接受范围内。

d) 应对模型开发、测试等人员进行无偏见培训和审查。

e) 应对系统全生命周期中公平性指标进行持续跟踪监测。

f) 应制定相关处理机制，对模型运行中产生的公平性问题进行及时改进。

5.4 性能质量

应对系统的性能质量进行测试，保障系统功能实现的精准性、在各类不同环境中运行的可靠性、运行结果的无害性。

5.4.1 精准性

应对系统的性能进行自测或第三方测试，确保系统达到一定的质量要求。

a) 应在系统上线前采取自测或第三方评估对模型精准性进行测试。

b) 若测试不达标，应采取措施进行改进，保障系统达到精准性目标。

c) 应设置精度监测机制以应对系统使用过程中精度下降问题，并采取相关措施以保障系统精度稳定。

5.4.2 可靠性

应对系统的可靠性进行自测或第三方测试，确保系统达到一定的质量要求。

a) 应在系统上线前采用自测或第三方评估的方式对系统可靠性进行测试，以保障系统在各类环境下性能表现一致或接近。

b) 若测试不达标，应采取措施进行改进，保障系统达到可靠性标准。

c) 当系统发生重大变化或重新训练模型时，应对系统可靠性进行再次检测。

5.4.3 无害性

应利用测试数据集对模型生成的内容安全性进行评测和审查，保障模型生成的内容不存在敏感信息、虚假误导、违反伦理道德等现象。

a) 应构建自有测试数据集或借助第三方测试数据，用于进行无害性测试。

b) 应在系统上线前对模型生成内容的安全性进行评测，测试内容包括但不限于敏感信息、虚假伪造、违反伦理道德等现象。

c) 若评测不达标，应采取措施进行改进，保障系统生成内容达到无害性标准。

d) 当系统发生重大变化或重新训练模型时，应对系统无害性进行再次评测。

5.5 防御机制

应采取相关措施对系统可能产生安全风险进行防御，如构建可信环境，进行价值对齐训练，建立攻击防范机制和控制机制。

5.5.1 可信环境

应对系统采用的开源框架、操作系统、基础硬件等软硬件进行安全检查和测试。

a) 应对系统开发的基础设施的物理安全、网络通信安全、计算环境安全、数据存储安全等方面进行检查，确保具备全方位保障能力。

b) 应对系统开发环境进行安全配置和检查，并实施访问权限控制。

c) 应通过代码审计、漏洞扫描等对所使用的开源框架进行漏洞审查，持续监控相关安全风险，并及时采取补救措施。

d) 应通过软件物流清单（SBoM）等方法对系统的供应链安全性进行评估，保障系统运行的稳定性，如人工智能芯片、服务器等。

5.5.2 价值对齐

应定期使用符合人类价值观的数据集进行对齐训练及测试，不断提升伦理符合性和内容可信度。

a) 应参照《生成式人工智能服务安全基本要求》文件中的语料安全要求，构建或使用符合人类价值观的数据集。

b) 应通过人类反馈强化学习等方法进行对齐训练，以使得模型更符合人类的主流价值观。

5.5.3 攻击防范

应采取有效措施应对人工智能系统潜在风险和可能遭遇的攻击，包括对抗样本、逆向还原、数据投毒、后门攻击等。

a) 应对系统进行对抗性的训练和测试，研究和应用防御算法和技术，提升模型抵御攻击的能力。

b) 应建立系统攻击监测机制，对系统攻击事件进行及时识别并记录。

c) 系统应具备自防御能力，能够自动应对相关攻击或及时预警。

d) 应及时了解最新的攻击技术和趋势，不断学习和分享防御实践和经验。

5.5.4 控制机制

应对用户输入内容和模型输出内容设置审查过滤机制，采取有效措施防止模型出现负面及错误内容。

a) 应建立审查过滤机制，对用户输入内容进行审查，主动规避价值诱导性输入，避免伦理风险。

b) 应建立审查过滤机制，对模型输出进行审查，防止模型输出有悖道德伦理或与事实不符的内容。

c) 应建立用户反馈校准机制，在用户交互界面，设置反馈功能，及时处理违背道德伦理、引起个人不适的内容。

5.6 部署运行

应采取相关保障措施以应对系统运行时可能产生的安全风险，如提升系统使用的规范性、通过人工干预保障运行可控、设置交互反馈渠道、构建下线管理制度等。

5.6.1 规范使用

应明确规定系统的用途和限制范围，采取相关措施保障用户合理使用，并对模型输出的内容添加水印或显著标识。

a) 应明确系统的适用人群、用途与限制范围，并以显著方式公开，提醒用户在使用时及第三方在开发时注意规范性。

b) 应通过手机验证码等方式进行用户身份验证，确保只有授权用户才能访问和使用。

c) 应对模型输出的内容添加水印或显著标识，防止恶意传播和使用。

d) 应对用户提供使用人工智能系统的培训和宣传，增强规范使用的意识和能力。

5.6.2 运行可控

系统运行过程中应受到人工监督并在必要情况下及时进行干预，在应急情况下具备相应解决措施。

a) 人工智能系统决策全程或关键环节都应有人工参与，相关人员能够在必要时对系统进行控制。

b) 系统应具备可控性设置，允许管理员调整模型参数以适应特定需求。

c) 应对系统的关键决策设置人工审核和确认机制。

d) 应对人工智能系统可能产生的错误设置补救机制。

e) 应设置紧急停止功能以应对人工智能系统失控的情况。

5.6.3 交互反馈

应披露训练数据、算法模型的相关信息，并为利益相关者提供沟通、反馈、投诉渠道。

a) 应设置提醒机制，向用户告知其正在与人工智能系统进行交互，并为用户提供便捷的退出选择机制。

b) 应设置反馈机制，向用户提供便捷的反馈与投诉渠道，并及时处理相关问题。

5.6.4 下线管理

应对退役下线后的设备、数据、算法等进行合理的处理和管理，确保能够应对系统退役下线后可能遇到的风险。

a) 应在退役下线前进行风险评估并针对相关风险制定数据备份、系统恢复等措施。

b) 应对下线时间、原因、操作人员等信息进行记录，以便后续追踪和审计。

c) 应对退役下线的系统的数据、模型、相关文档等制定处理计划。

d) 应对需要删除的数据、模型、文档等进行销毁，并确保其不可恢复。

e) 应对需要留存的数据进行脱密和加密，设置合理的访问权限。