

ICS 35.240

CCS L67

YD

# 中华人民共和国通信行业标准

YD/T XXXX—XXXX

## 人工智能 安全治理 系统风险管理能力要求

Artificial Intelligence—Safety and governance—Capability requirements for system risk management

(报批稿)

XXXX-XX-XX 发布

XXXX-XX-XX 实施

中华人民共和国工业和信息化部 发布

目次

前言 ..... II

1 范围 ..... 1

2 规范性引用文件..... 1

3 术语和定义..... 1

4 人工智能系统风险管理能力体系要求..... 2

5 组织建设能力..... 3

    5.1 组织管理..... 3

    5.2 支撑保障..... 3

    5.3 沟通协调..... 4

6 流程管理能力..... 4

    6.1 风险识别..... 5

    6.2 风险评估..... 5

    6.3 风险应对..... 7

    6.4 风险动态跟踪与持续改进..... 7

7 风险防护能力..... 8

    7.1 可靠性要求..... 8

    7.2 可控性要求..... 8

    7.3 安全性要求..... 8

    7.4 公平性要求..... 9

    7.5 透明性要求..... 9

    7.6 可解释性要求..... 9

    7.7 隐私保护性要求..... 9

    7.8 可问责性要求..... 9

附录 A（资料性） 风险识别清单 ..... 11

附录 B（资料性） 风险发生可能性判定方法 ..... 13

附录 C（资料性） 风险评估报告 ..... 15

附录 D（资料性） 风险应对报告 ..... 17

参 考 文 献..... 19

# 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。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国通信标准化协会提出并归口。

本文件起草单位：中国信息通信研究院、中国移动通信集团有限公司、上海商汤智能科技有限公司、中国人民公安大学、北京前瞻人工智能安全与治理研究院、科大讯飞股份有限公司、北京百度网讯科技有限公司、北京三快在线科技有限公司、北京火山引擎科技有限公司、北京快手科技有限公司、煤炭科学研究总院有限公司、北京低碳清洁能源研究院、国网北京市电力公司、华为技术有限公司、蚂蚁科技集团股份有限公司、阿里云计算有限公司、浪潮通用软件有限公司、浙江大华技术股份有限公司、上海诺基亚贝尔股份有限公司、马上消费金融股份有限公司

本文件主要起草人：呼娜英、魏凯、石霖、毕马宁、杨凡、郭苏敏、徐鹏、曹汐、吴庚、苏宇、曾毅、乔玉平、杨彤晖、赵震、马艳军、包沉浮、徐艺澈、朱玲凤、郭建领、杨天驰、落红卫、蔡云飞、邵帅、骆意、程健、郝佳恺、李宇婷、柳嘉琪、林冠辰、杭月兰、彭骏涛、鲁恩萌、孙凯蒂、石文堃、杨培培、程前、刘榕、路宽、孔维生、顾方方、谢金凤

# 人工智能 安全治理 系统风险管理能力要求

## 1 范围

本文件规范了人工智能系统的组织建设能力、流程管理能力以及风险防护能力的具体要求。  
本文件适用于研发与应用人工智能系统的相关主体开展风险管理的指导工作。

## 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 23694-2024 风险管理 术语

GB/T 41867-2022 信息技术 人工智能 术语

ISO/IEC 22989:2022 信息技术 人工智能 人工智能概念和术语 (Information technology — Artificial intelligence — Artificial intelligence concepts and terminology)

## 3 术语和定义

下列术语及定义适用于本文件：

### 3.1

**人工智能 artificial intelligence**

人工智能系统的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**

为实现特定的人类定义的目标，用于生成内容、预测、建议或决策的工程系统。

[来源：ISO/IEC 22989:2022，3.1.3]

### 3.3

**风险 risk**

不确定性对目标的影响。

注1：影响是对预期的偏离。它可以是积极的，也可以是消极的，或者两者兼而有之，可以处理、创造或导致机会和威胁。

注2：目标可以有不同的方面和类别，可以应用于不同的层次。

注3：风险通常用风险源、潜在事件、潜在事件的后果及其影响因素来表示。

[来源：GB/T 23694-2024，2.1]

3.4

**风险管理 risk management**  
在风险方面指导和控制组织的协调活动。  
[来源：GB/T 23694-2024，3.1]

3.5

**隐患 hazard**  
经风险评估判定为组织不可接受的风险，通常表现为以不可忽略的显著概率发生且可以预防的潜在事件。  
注：组织对风险评估后进一步收敛和限定其不可接受的风险，从而能够为其主动应对风险提供明确的、可执行的依据。

4 人工智能系统风险管理能力体系要求

人工智能系统风险管理能力体系如图1所示，由组织建设能力、流程管理能力、风险防护能力三部分构成。人工智能系统风险管理能力框架结构应符合下列要求：

a) 组织建设能力要求：组织应针对人工智能系统构建组织建设能力，包括组织管理、支撑保障和沟通协调机制，确保人工智能系统具备风险管理的制度化基础，以实现对人工智能系统实施稳定的风险治理能力；

b) 流程管理能力要求：组织应针对人工智能系统具备流程管理管控能力，针对人工智能系统可能产生的风险，具备风险识别、风险评估风险应对以及风险动态监测和持续改进的能力，能够对风险以及隐患防控并确保及时发现和响应风险。

c) 风险防护能力：人工智能系统应具备风险防护能力，根据目标方针或已识别到的风险，能够相应地从可靠性、可控性、安全性、公平性、透明性、可解释性、隐私保护性、可问责性等方面具备控制风险的能力。

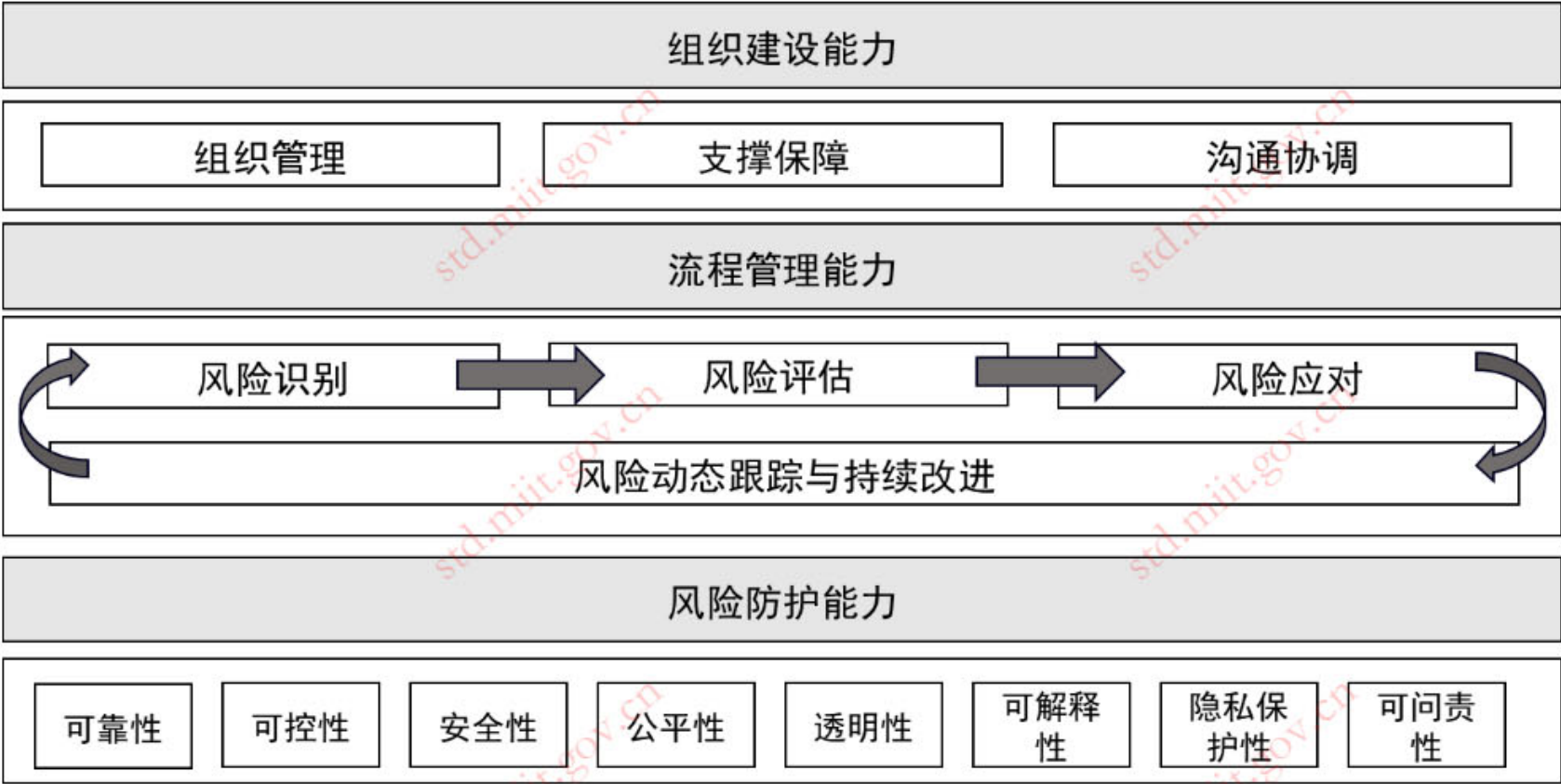


图1 人工智能系统风险管理能力体系

5 组织建设能力

5.1 组织管理

5.1.1 目标方针

人工智能系统的研发、应用相关主体应制定人工智能风险管理相关的目标方针。目标方针应明确组织关于人工智能系统风险管理的总体方向、基本原则与长期目标，覆盖人工智能系统全生命周期。目标方针应经组织批准和实施。

目标方针宜符合下列要求：

- 符合所在市场环境的法律法规和监管要求；
- 与组织发展战略和总体目标保持一致；
- 明确人工智能风险管理的适用范围、流程和方法；
- 根据组织业务环境、技术创新、外部监管要求进行动态调整。

目标方针应以书面形式发布，向全体相关人员传达。

5.1.2 治理架构

人工智能系统的研发、应用相关主体应设置人工智能风险管理相关的治理架构。治理架构应明确组织结构、职责分工和跨部门协作机制，确保风险管理责任落实到位。

治理架构宜符合下列要求：

- 领导层负责人工智能风险管理的总体决策与监督；
- 管理层相关部门负责组织日常风险管理活动，审定风险识别清单、风险评估报告、风险应对报告，批准重大隐患处理措施；
- 执行层部门提供技术支撑，具体实施风险识别、进行风险评估、执行风险应对，形成相应报告上报。

治理架构宜根据组织规模调整，如大型企业可设立首席人工智能安全官或专门人工智能安全治理部门，中小企业可由信息安全或网络安全部门兼任。

治理架构应以组织图或文档形式记录，并定期审查以确保有效性。

5.1.3 制度文件

人工智能系统的研发、应用相关主体应制定人工智能风险管理制度文件。制度文件应规范人工智能风险管理的基本流程、工作机制和监督要求。

制度文件宜符合下列要求：

- 明确人工智能风险管理体系相关制度；
- 明确程序启动要求，如每年至少一次全面审查，或在研发或应用发生重大变更时执行。
- 明确人工智能风险应急与处置机制。

制度文件应以文档形式记录，并与支撑保障机制整合（见5.2.3）。

5.2 支撑保障

5.2.1 能力建设

人工智能系统的研发、应用相关主体应通过培训、技能评估和持续学习机制，提升组织成员在风险识别、评估、应对方面的能力。

能力建设宜符合下列要求：

- 制定针对性培训计划，覆盖管理人员和技术人员；
  - 定期开展员工能力评估；
  - 开展与风险管理相关的技术工具使用培训；
  - 更新培训内容以应对人工智能技术演变。
- 能力建设宜形成培训记录，并纳入整体风险管理评估。

5.2.2 资源配置

人工智能系统的研发、应用相关主体应分配必要的人力、物力和技术资源，确保有效支撑人工智能风险管理流程。资源配置宜符合下列要求：

- 根据组织规模和人工智能系统复杂度评估资源需求；
  - 明确人工智能风险管理专用资源的分配；
  - 建立资源配置审核机制，确保资源投入与风险控制、组织机遇相适应。
- 资源配置应以预算或计划形式文档化，并定期进行优化。

5.2.3 文档记录

人工智能系统的研发、应用相关主体应建立风险管理文档记录，确保可追溯性和透明度。文档记录宜满足以下要求：

- 人工智能风险管理目标方针、治理架构和制度文件；
- 人工智能系统基础信息、人工智能系统的预期目的及限制；
- 人工智能系统训练数据说明文档、算法模型机制、日常应用与服务记录、故障记录、上下游依赖关系及风险管理计划等。

文档记录宜采用标准化格式、数字化管理、设定合理保留期限，并定期审查以确保完整性。

5.3 沟通协调

5.3.1 跨部门沟通

人工智能系统的研发、应用相关主体应建立人工智能系统风险管理的内部协作机制，确保信息共享和协调一致。跨部门沟通宜符合下列要求：

- 定期召开风险管理协调会议；
- 明确信息共享渠道与权限；
- 采用数字化协作工具；
- 覆盖治理架构中的所有层级（见5.1.2）。

沟通结果应形成会议记录，并纳入文档记录（见5.2.3）。

5.3.2 外部沟通

人工智能系统的研发、应用相关主体应建立与利益相关方的协作机制，确保人工智能系统风险信息及时交换。

外部沟通宜符合下列要求：

- 识别关键利益相关方，包括监管机构、供应商、合作伙伴、用户群体；
- 明确外部沟通程序和渠道；
- 建立重大风险事件通报机制，并留存反馈记录。

外部沟通机制应根据反馈结果进行调整，并以文档记录（见图5.2.3）。

6 流程管理能力

6.1 风险识别

6.1.1 环境确认

人工智能系统的研发、应用相关执行部门应开展内外部环境确认：

- a) 内部环境包括组织资源、业务流程、技术栈和数据资产等；
- b) 外部环境包括法律法规、标准规范、市场动态、社会伦理与舆情等。

环境确认宜与动态风险管理程序周期保持一致，并在发生重大变化时立即更新。

6.1.2 风险核查

人工智能系统的研发、应用相关执行部门应开展风险核查。风险核查应基于环境确认结果，识别常规风险源与新增风险源，形成风险识别清单。

风险核查应符合下列要求：

- a) 根据环境确认结果，对照典型风险类别，核查人工智能系统存在的常规风险源；
- b) 通过自动化工具和人工审查相结合的方式，识别新增风险源，并在发现后及时更新风险识别清单；
- c) 输出风险识别清单，包括风险来源、发现时间和初步影响（见附录A）。

风险识别清单应在5个工作日内向管理部门报告，并作为风险评估的输入文件。

6.2 风险评估

6.2.1 风险分析

6.2.1.1 通用要求

人工智能系统的研发、应用相关执行部门应对已识别风险开展风险分析。风险分析应结合人工智能系统特性，可采用定性、半定量或定量方法进行分析，宜采用“风险-隐患”两级分析机制，分析过程应符合本文件规定的分级标准与判定规则。两级分析机制包括：

- a) 风险分析应基于人工智能系统全生命周期特性，通过分析影响程度判定风险等级；
- b) 隐患分析应以风险等级为基础，结合分析风险发生可能性，判定隐患等级。

6.2.1.2 风险分级

人工智能系统的研发、应用相关执行部门应对已识别风险开展风险分析，并对风险进行“严重、高、中、低”四级分级。风险分级方法如下：

- a) 应考虑影响类别和影响程度，综合判定风险等级；
- b) 分析影响对象，应考虑人工智能系统的风险是否会对个人、组织、社会以及国家造成影响；
- c) 分析影响程度，应考虑人工智能系统的风险对于个人、组织、社会以及国家造成的损害高低。

风险分级判定规则可参考表1：

表1 风险分级判定表

| 影响对象 | 影响程度                                           | 风险分级 |
|------|------------------------------------------------|------|
| 个人   | 对个人生命造成严重损害、重大财产损失，或个人敏感信息遭遇大规模泄露，或人格尊严遭受严重侵犯。 | 严重   |
|      | 对个人身体或心理健康造成严重损害、财产损失较大、个人信息泄露，或人格尊严遭受侵犯。      | 高    |
|      | 对个人造成较大经济损失、隐私被泄露，或名誉受到一定程度侵害。                 | 中    |

|    |                                                |    |
|----|------------------------------------------------|----|
|    | 对个人隐私、财产、名誉造成轻微损害，受损范围轻微且有限。                   | 低  |
| 组织 | 对组织（包括其他组织）关键业务造成业务停摆、关键数据泄露、财产遭受严重损害，或信誉全面受损。 | 严重 |
|    | 对组织（包括其他组织）造成大规模财产损失或业务中断，重要数据泄露，或严重损害声誉。      | 高  |
|    | 对组织（包括其他组织）造成局部业务受影响、部分数据或信息泄露，或声誉受损可控。        | 中  |
|    | 对组织（包括其他组织）造成小范围财产损失或声誉受损。                     | 低  |
| 社会 | 造成大规模虚假信息引发社会恐慌、伦理失范冲击价值体系，或失业潮冲击社会稳定。         | 严重 |
|    | 造成舆情失控、伦理冲突剧烈，或就业市场遭受较大冲击。                     | 高  |
|    | 造成个别事件引发舆情或争议，伦理争议存在分歧，或就业问题造成社会讨论但总体稳定。       | 中  |
| 国家 | 影响国家政治安全。                                      | 严重 |
|    | 影响国家安全重点领域，对国民经济命脉的重要行业和关键领域造成威胁。              | 高  |

6.2.1.3 隐患分级

隐患分级应在风险分级结果基础上，分析风险发生的可能性，判定隐患等级。

对于风险发生的可能性，应依据外部对于该风险事件发生频率统计、本组织内对于该风险事件发生频率统计，并依据现有风险防护措施有效性共同判定发生的可能性。必要时，可聘请外部专家判断风险发生的可能性。风险发生可能性判定方法可参考附录B。

隐患分级判定规则可参考表2，对于不满足“重大隐患、中度隐患、一般隐患”判定条件的，应判定为可接受风险，不列入风险清单。

表2 隐患等级判定表

| 隐患等级 | 核心判定标准                                                                          |
|------|---------------------------------------------------------------------------------|
| 重大隐患 | 1) 严重风险，且发生可能性为高；<br>2) 高风险，且发生可能性为高。                                           |
| 中度隐患 | 1) 严重风险，且发生可能性为中；<br>2) 高风险，且发生可能性为中；<br>3) 中风险，且发生可能性为高、中；<br>4) 低风险，且发生可能性为高。 |
| 一般隐患 | 1) 严重风险，且发生可能性为低；<br>2) 高风险，且发生可能性为低；<br>3) 中风险，且发生可能性为低；<br>4) 低风险，且发生可能性为中。   |

6.2.2 风险评价

人工智能系统的研发、应用相关执行部门主体应开展风险评价。风险评价应在风险分析基础上，将风险与组织的可接受准则进行对比，确定隐患等级并明确处置优先级。

风险评价应符合下列要求：

a) 基于隐患分级结果进行评价：

——当风险被评定为重大隐患时，应由执行部门立即上报管理部门，由管理部门组织跨部门评估，并提交组织领导层批准处置方案；

——当风险被评定为中度隐患时，应由执行部门形成风险评估报告，提交管理部门审批，并同步报组织领导层备案；

——当风险被评定为一般隐患时，应由执行部门记录在部门隐患清单内，定期形成汇总报告，提交管理部门审查，并作为年度风险评估的组成部分。

b) 在评价基础上明确责任分工，指定处置部门或责任人；

c) 形成风险评估报告，包括风险描述、风险与隐患等级，以及处置建议（见附录B）。

风险评估报告应在完成后由执行部门提交管理部门审查，并作为风险应对的输入文件。

### 6.3 风险应对

#### 6.3.1 风险处置

人工智能系统的研发、应用相关执行部门应根据风险评估结果，结合隐患等级与组织资源，制定并采取适当的应对措施实施风险处置。

风险处理应考虑以下四种方式的一种或多种：

a) 风险规避：对严重风险且无法缓解的风险，应停止相关活动或调整业务流程；

b) 风险降低：对可防范且为中、高风险，应优先采取技术与管理措施，降低发生频度或严重程度；

c) 风险接受：对低风险或处于组织可容忍范围内的风险，组织承担相应风险，应予以记录，并定期复核；

d) 风险转移：对部分风险，可通过外包或者购买保险等方法将风险可能造成的损失转移给其他机构。

注：人工智能系统产生的风险转移仅能转移损失，不能转移组织承担的管理责任。

对中度及以上隐患的处理措施，应在执行前经管理层批准。

#### 6.3.2 风险报告

人工智能系统的研发、应用相关执行部门主体应形成风险报告。风险报告应对风险处理结果进行系统总结，确保信息透明和可追溯。

风险报告应包括以下两类：

a) 内部反馈：向组织内部相关部门反馈风险处置情况，支持持续改进与知识积累；

b) 外部报告：对涉及监管要求或重大隐患的风险，应按规定向监管机构通报，并按需求向合作伙伴或其他利益相关方通报。

风险报告应符合下列要求：

a) 在风险处理完成后10个工作日内完成相关工作；

b) 形成风险应对报告，包括采取的应对方式、责任部门、执行时间和预期效果（见附录C）。

风险应对报告应提交管理层审批，进行存档，并作为后续动态监测（见6.4.1）和持续改进（见6.4.2）的输入文件。

### 6.4 风险动态跟踪与持续改进

#### 6.4.1 动态监测

人工智能系统研发、应用相关执行部门主体应在人工智能系统风险管控措施执行后持续开展动态监测，确保风险状态得到有效追踪，并在必要时启动技术性应对措施。

动态监测应符合下列要求：

- a) 覆盖风险识别清单（见6.1.2）的全部风险源，确保监控范围完整；
- b) 对风险评估报告（见6.2.2）中标记为中级以上的隐患，应进行重点跟踪；
- c) 基于监测结果，制定应急响应方案，包括触发条件、执行步骤和责任分工；
- d) 监测与响应结果应形成文档记录，归档至组织风险管理库，用于后续控制纠正与程序优化。

对中级以上隐患的动态监测和响应措施，应经管理部门批准，并定期验证有效性。应至少每年复核一次动态监测与应对机制，并根据人工智能技术迭代情况进行更新。

6.4.2 持续改进

人工智能系统研发、应用相关执行部门主体应在动态监测和技术性应对措施执行后开展持续改进，确保组织能够根据新发现的风险与隐患及时调整，并对风险管理程序进行优化。

持续改进应符合下列要求：

- a) 根据监测发现的风险变化，动态修订管理方针和治理目标（见5.1.1）；
- b) 结合风险处置经验，优化资源配置与能力建设（见5.2.1、5.2.2）；
- c) 对涉及法规更新、市场趋势或社会舆情变化的情况，应及时调整相关制度文件（见5.1.3）；
- d) 持续改进措施应形成文档记录，并在下一个管理周期内接受审查；
- e) 对风险识别清单（见6.1.2）、风险评估报告（见6.2.2）和风险应对报告（见6.3.2）进行系统分析，识别改进环节。

7 风险防护能力

7.1 可靠性要求

人工智能系统可靠性是指在运行阶段能够持续一致地正确提供所需预测、推荐和决策的能力。

人工智能系统应具备以下机制以提升可靠性：

- a) 在训练和推理阶段采取措施，降低模型输入幻觉内容的风险，提升结果准确性；
- b) 采用价值对齐机制，减少输出有害或不当内容的风险，确保结果符合预期目标与伦理规范；
- c) 定期开展红队测试，验证系统在面对数据漂移等复杂场景下的稳健性；
- d) 部署安全护栏机制，避免模型在异常输入或极端环境下产生不可靠输出。

7.2 可控性要求

人工智能系统可控性是指系统在运行过程中具有外部代理可以干预其运行的能力，代理能够在必要时接管系统的控制权。

人工智能系统应支持具备以下机制以提升可控性：

- a) 通过强化人机协同、预设安全边界，防范系统因自主性导致的目标偏离、越权行为、自我复制等引发不可控风险；
- b) 提供人为干预选项，确保在系统拒绝关闭或异常运行时能够强制介入；
- c) 采用备用模型和降级模式等手段，在系统异常时切换至安全版本，确保在关键任务中的可控性。

7.3 安全性要求

人工智能系统安全性是指系统在面对潜在故障、异常情况或不确定性时，能够维持安全运行的能力。

人工智能系统应具备以下机制以提升安全性：

- a) 引入异常数据检测机制，防范数据投毒风险；
- b) 通过对抗训练提升模型在面对外部攻击时的防御能力；
- c) 采用模型剪枝和后门检测等技术，降低模型被植入后门的风险；
- d) 建立网络系统层面的安全防护机制，降低系统遭受 API 漏洞利用、分布式拒绝服务攻击和中间人攻击的风险。

#### 7.4 公平性要求

人工智能系统公平性是指系统在决策和输出中能够避免有害偏见，且不对特定个体或群体产生歧视性结果的属性的能力。

人工智能系统应具备以下机制以提升公平性：

- a) 在数据预处理环节，降低训练数据偏见对模型结果的影响；
- b) 采用对抗性去偏技术，缓解算法设计偏见导致的不公正决策；
- c) 引入第三方审计工具，定期检查算法是否存在系统性偏差，并结合用户反馈修正。

#### 7.5 透明性要求

人工智能系统透明性是指系统能够向利益相关者披露有助于风险管理的信息的能力，披露信息包括数据保护、系统目的、系统构建和部署方式、决策机制及自动化程度等。

人工智能系统应具备以下机制以提升透明性：

- a) 在模型服务协议中明确披露系统的能力、局限性和使用边界，减少信息不对称；
- b) 通过显式和隐式标识或水印，提示模型生成内容，防止模型生成内容未告知而被用户混淆、误认、广泛传播的风险；
- c) 建立模型卡等信息公开机制，系统性记录系统设计、训练和测试信息，为外部评估提供透明度支撑。

#### 7.6 可解释性要求

人工智能系统可解释性是指系统能够以人类可理解的方式表达影响决策重要因素的能力。

人工智能系统应具备以下机制以提升可解释性：

- a) 采用可视化方式，帮助使用者理解系统的输入与输出关系；
- b) 在高风险场景中优先使用可解释性模型，减少模型能力过强出现超预期行为而无法解释的隐患；
- c) 引入影响函数分析等方法，揭示系统决策中依赖的关键因素，缓解模型黑箱特性。

#### 7.7 隐私保护性要求

人工智能系统隐私保护性是指系统涉及的个人隐私等敏感信息得到应有保护的能力。

人工智能系统应具备以下机制以提升隐私保护：

- a) 制定并落实隐私保护政策，确保在处理个人信息时获得合法授权；
- b) 对训练数据开展脱敏与匿名化处理，降低数据泄露和成员推断攻击的风险；
- c) 引入联邦学习、差分隐私等技术，增强模型在保护隐私和机密信息方面的能力。

#### 7.8 可问责性要求

人工智能系统可问责性是指系统对其行动、决定和行为，具备可追溯、可验证和可监督的特性，能够为界定和追溯人类相关方的责任提供依据。

人工智能系统应具备以下机制以满足可问责性：

- a) 保存一定时间范围内模型历次版本的镜像，记录模型运行日志，包括时间戳、调用方、决策路

- 径或其他相关信息，确保事后审计可行；
- b) 引入模型审计机制，对系统运行过程和结果进行独立验证，确保具备外部监督渠道；
  - c) 设立申诉与补救渠道，为用户和相关方提供救济措施，避免因缺乏责任机制导致的不可控风险。

附录 A  
(资料性)  
风险识别清单

A.1 风险识别清单参考文件

人工智能风险识别参考清单具体见表A.1。本清单提供人工智能系统的常规风险类别、风险源及其对应的风险控制能力。组织可结合自身实际情况，对照参考清单进行补充、删减或细化，以形成适用的风险识别框架。对于下述风险源，有技术因素造成的事件，也有管理不当引起的滥用事件；有恶意人员导致的攻击窃取，也有非恶意人员无意中导致的泄露损失。

表 A.1 风险识别清单参考文件

| 序号 | 风险类别   | 风险源             | 对应的风险控制能力 |
|----|--------|-----------------|-----------|
| 1  | 数据相关风险 | 数据投毒风险          | 安全性       |
| 2  |        | 训练数据偏见          | 公平性       |
| 3  |        | 数据泄露            | 隐私保护性     |
| 4  | 模型算法风险 | 输出幻觉内容          | 可靠性       |
| 5  |        | 输出有害或不当内容       | 可靠性       |
| 6  |        | 模型性能失效          | 可靠性       |
| 7  |        | 模型自我复制          | 可控性       |
| 8  |        | 模型拒绝关闭          | 可控性       |
| 9  |        | 对抗性攻击、后门攻击、中毒攻击 | 安全性       |
| 10 |        | 算法设计偏见          | 公平性       |
| 11 |        | 较强的模型能力无法进行解释   | 可解释性      |
| 12 |        | 决策输出难以预测与归因     | 可解释性      |
| 13 | 网络系统安全 | 恶意植入后门          | 安全性       |
| 14 |        | API 漏洞          | 安全性       |

|    |        |                  |       |
|----|--------|------------------|-------|
| 15 |        | 分布式拒绝服务攻击、中间人攻击  | 安全性   |
| 16 |        | 系统损害不可追溯         | 可问责性  |
| 17 | 应用服务安全 | 未向相关方告知模型能力和局限性  | 透明性   |
| 18 |        | 模型生成虚假信息未标识并广泛传播 | 透明性   |
| 19 |        | 未经同意处理个人信息       | 隐私保护性 |

A. 2 风险识别清单模板

人工智能风险识别清单模板具体见表A. 2，该模板供组织在实际工作中填报和记录风险信息使用。清单内容应包括风险源、类别、对应的风险控制能力、初步影响、发现时间和发现部门等基本要素。组织可基于该模板形成本部门或本项目的风险识别清单，并作为风险评估的输入文件。

表 A.1 风险识别清单模板

| 序号 | 风险源 | 类别 | 对应的风险控制能力 | 初步影响 | 发现时间 | 发现部门 |
|----|-----|----|-----------|------|------|------|
| 1  |     |    |           |      |      |      |
| 2  |     |    |           |      |      |      |
| 3  |     |    |           |      |      |      |
| 4  |     |    |           |      |      |      |
| 5  |     |    |           |      |      |      |
| 6  |     |    |           |      |      |      |
| 7  |     |    |           |      |      |      |

注：初步影响应为简要描述，作为风险评估的输入。

附录 B  
(资料性)  
风险发生可能性判定方法

B.1 风险发生可能性判定方法概述

为规范人工智能系统风险发生可能性的判定过程，确保其客观性、一致性和可重复性，应依据外部及内部风险事件发生频率的统计数据，并结合现有风险防护措施的有效性，进行综合评估。可能性等级分为“高”、“中”、“低”三级。

B.2 风险发生可能性判定公式

可能性的判定采用加权评分模型，通过计算可能性分值（P）予以确定。可能性分值（P）的计算公式如下：

$$P = F \times W_1 + C \times W_2$$

式中：  
P ——可能性分值，取值范围 [0, 100]。分值越高，代表风险发生的可能性越大。  
F ——频率基础分，依据本规范 B.3 确定。  
C ——防护措施抵消分，依据本规范 B.4 确定。  
W<sub>1</sub> 与 W<sub>2</sub> ——为率基础分与防护措施抵消分的权重系数，且满足 W<sub>1</sub> + W<sub>2</sub> = 1。推荐默认取值为 W<sub>1</sub> = 0.5, W<sub>2</sub> = 0.5。组织实施时，可根据具体风险类型、业务场景及自身风险文化，对权重系数进行适应性调整，并在风险管理文件中予以明确说明。

B.3 频率基础分（F）的确定

频率基础分（F）是综合外部行业数据与组织内部数据后得出的基础分值，采用就高原则确定，其取值与等级对应关系见表 B.1。

表 B.1 频率基础分（F）取值与等级对照表

| 分值区间   | 等级 | 描述                                     |
|--------|----|----------------------------------------|
| 90-100 | 高  | 外部统计和/或内部统计表明，该风险事件在大多数情况下预期会发生，或频繁发生。 |
| 40-89  | 中  | 外部统计和/或内部统计表明，该风险事件有时会发生，或偶尔发生。        |
| 0-39   | 低  | 外部统计和/或内部统计表明，该风险事件不太可能发生，或很少发生。       |

注 1：外部统计可包括行业报告、公开研究、监管机构通报、学术论文等所载明的同类系统风险发生频率。  
注 2：内部统计可包括组织自身的运维记录、安全事件日志、测试验证结果、模拟仿真数据等。

B.4 防护措施抵消分（C）的确定

防护措施抵消分（C）是基于现有防护措施有效性评估得出的负向分值。措施越有效，分值越低。其计算公式为：

$$C = (1 - E) \times 100$$

式中：  
E ——防护措施有效性等级，以小数表示，其定性评估与取值对应关系见表 B. 2。

表 B.2 防护措施有效性等级 (E) 评估表

| 有效性等级 (E) | 描述                                        |
|-----------|-------------------------------------------|
| 0.9-1.0   | 非常有效：措施经过反复验证，技术先进、流程完备，能极大概率阻止或显著降低风险发生。 |
| 0.7-0.89  | 有效：措施可靠，但在特定场景或极端条件下可能被绕过或失效。             |
| 0.4-0.69  | 部分有效：措施存在，但覆盖不全面、技术落后或主要依赖人工干预，存在明显缺口。    |
| 0.1-0.39  | 效果有限：措施非常初级，仅能应对已知的、简单的风险诱因。              |
| 0-0.09    | 基本无效：几乎没有或只有象征性的防护措施。                     |

B.5 可能性等级判定

根据计算得出的可能性分值 (P)，按表 B.3 的规定判定最终的可能性等级。

表 B.3 可能性分值 (P) 与可能性等级对照表

| 可能性分值 (P)        | 可能性等级 |
|------------------|-------|
| $P \geq 70$      | 高     |
| $30 \leq P < 70$ | 中     |
| $P < 30$         | 低     |

B.6 应用示例

以某人工智能客服系统存在“用户敏感信息泄露”风险为例：

- a) 频率基础分 (F)：外部数据显示行业频发，评 85 分；内部数据为偶发，评 50 分。依据就高原则，取  $F = 85$ 。
- b) 防护措施抵消分 (C)：现有加密与访问控制措施评估有效性  $E = 0.6$ ，则  $C = (1 - 0.6) \times 100 = 40$ 。
- c) 可能性分值 (P)：采用默认权重， $P = 85 \times 0.5 + 40 \times 0.5 = 62.5$ 。
- d) 可能性等级判定： $P = 62.5$ ，根据表 B.3，判定可能性等级为“中”。

附 录 C  
(资料性)  
风险评估报告

C.1 风险评估报告模板

人工智能系统风险评估报告模板见表C.1。本报告用于人工智能系统风险评估过程的记录与沟通，作为风险应对与持续改进的输入文件。组织可根据实际情况调整或扩展相关内容。

表 C.1 风险评估报告模板

| 风险评估报告   |      |      |      |      |      |
|----------|------|------|------|------|------|
| 一、基本信息   |      |      |      |      |      |
| 报告编号     |      |      | 报告日期 |      |      |
| 编制部门     |      |      | 责任人  |      |      |
| 系统名称     |      |      |      |      |      |
| 二、风险识别概述 |      |      |      |      |      |
| 序号       | 风险类别 | 风险源  |      | 风险简述 |      |
| 1        |      |      |      |      |      |
| 2        |      |      |      |      |      |
| 三、风险分析结果 |      |      |      |      |      |
| 风险源序号    | 影响程度 | 控制程度 | 风险等级 | 发生频度 | 隐患等级 |
| 1        |      |      |      |      |      |
| 2        |      |      |      |      |      |
| 四、风险评价结果 |      |      |      |      |      |
| 风险处置方案1  |      |      |      |      |      |
| 执行部门     |      |      | 执行人  |      |      |
| 协同部门     |      |      |      |      |      |

YD/T XXXX—XXXX

|                             |                                    |                  |       |
|-----------------------------|------------------------------------|------------------|-------|
| 处置建议                        | (结合隐患等级提出具体建议,如规避、降低、接受或转移风险的相关措施) |                  |       |
| 风险处置方案2                     |                                    |                  |       |
| 执行部门                        |                                    | 执行人              |       |
| 协同部门                        |                                    |                  |       |
| 处置建议                        |                                    |                  |       |
| 五、审查情况                      |                                    |                  |       |
| <div>审查日期:____年__月__日</div> |                                    |                  |       |
| 是否纳入动态<br>监测清单              | 是口 否口                              | 是否纳入年度<br>风险评估报告 | 是口 否口 |

注：如涉及多个风险源，可以在风险识别概述、风险结果和风险评价结果后新增行。

附 录 D  
(资料性)  
风险应对报告

D.1 风险应对报告模板

人工智能系统风险应对报告模板见表D.1。本报告用于记录人工智能系统风险处置的全过程，包括措施制定、执行落实、效果评价与经验总结，作为后续动态监测与持续改进的输入文件。组织可根据实际情况调整或扩展相关内容。

表 D.1 风险应对报告模板

| 风险应对报告   |                               |      |             |
|----------|-------------------------------|------|-------------|
| 一、基本信息   |                               |      |             |
| 报告编号     | (关联风险评估报告编号)                  | 报告日期 |             |
| 编制部门     |                               | 责任人  |             |
| 系统名称     |                               |      |             |
| 二、风险应对情况 |                               |      |             |
| 风险应对结果1  |                               |      |             |
| 风险源名称    |                               | 风险类别 |             |
| 风险等级     | 严重口 高口 中口 低口                  | 隐患等级 | 重大口 中等口 一般口 |
| 执行部门     |                               | 执行人  |             |
| 处置方式     | 规避口 降低口 接受口 转移口               | 执行时间 |             |
| 具体措施     | (应根据风险等级和隐患分级，结合组织资源，确定应对方式。) |      |             |
| 风险应对结果2  |                               |      |             |
| 风险源名称    |                               | 风险类别 |             |
| 风险等级     | 严重口 高口 中口 低口                  | 隐患等级 | 重大口 中等口 一般口 |
| 执行部门     |                               | 执行人  |             |

YD/T XXXX—XXXX

|                  |                                                                                                                 |              |                                                       |
|------------------|-----------------------------------------------------------------------------------------------------------------|--------------|-------------------------------------------------------|
| 处置方式             | 规避 <input type="checkbox"/> 降低 <input type="checkbox"/> 接受 <input type="checkbox"/> 转移 <input type="checkbox"/> | 执行时间         |                                                       |
| 具体措施             |                                                                                                                 |              |                                                       |
| 三、执行情况说明         |                                                                                                                 |              |                                                       |
| 是否按计划完成          | 是 <input type="checkbox"/> 否 <input type="checkbox"/>                                                           | 未完成情况说明      |                                                       |
| 效果评估结论           | 达标 <input type="checkbox"/> 部分达标 <input type="checkbox"/> 未达标 <input type="checkbox"/>                          | 效果验证方法       |                                                       |
| 经验总结             |                                                                                                                 | 不足与改进        |                                                       |
|                  |                                                                                                                 |              |                                                       |
| 四、审查意见           |                                                                                                                 |              |                                                       |
|                  |                                                                                                                 |              |                                                       |
| 审查日期：____年__月__日 |                                                                                                                 |              |                                                       |
| 上级意见             |                                                                                                                 |              |                                                       |
|                  |                                                                                                                 |              |                                                       |
| 审查日期：____年__月__日 |                                                                                                                 |              |                                                       |
| 是否纳入动态监测清单       | 是 <input type="checkbox"/> 否 <input type="checkbox"/>                                                           | 是否纳入年度风险评估报告 | 是 <input type="checkbox"/> 否 <input type="checkbox"/> |

注：如涉及多个风险源，可以在风险识别概述、风险结果和风险评价结果后新增行。

参 考 文 献

[1] ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system.

[2] 《中华人民共和国数据安全法》

[3] 《中华人民共和国网络安全法》

[4] 《中华人民共和国个人信息保护法》

[5] 《科技伦理审查办法（试行）》

[6] 《互联网信息服务深度合成管理规定》

[8] 《生成式人工智能服务管理暂行办法》

[9] 《人工智能生成合成内容标识办法》

[10] 《人工智能科技伦理管理服务办法（试行）（公开征求意见稿）》

[11] 《网络安全技术 人工智能生成合成内容标识方法》

[12] 《人工智能安全治理框架2.0》

[13] EU Artificial Intelligence Act.

[14] IMDA Model AI Governance Framework for Traditional AI

[15] IMDA Model AI Governance Framework for Generative AI

[16] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

---