



中华人民共和国国家标准

GB/T XXXXX—XXXX

人工智能 风险管理能力评估

Artificial intelligence - Service capability maturity evaluation

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX—XX—XX 发布

XXXX—XX—XX 实施

国家市场监督管理总局
国家标准化管理委员会

发布

目 次

前言 III

引言 V

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 人工智能风险管理能力评级 2

 4.1 人工智能风险管理能力等级 2

 4.2 人工智能风险管理能力等级划分规则 3

 4.3 人工智能风险管理过程 4

 4.4 人工智能风险管理能力框架 5

 4.5 人工智能风险管理能力子项 6

5 人工智能风险管理能力评估 9

 5.1 人工智能风险管理策划能力评估 9

 5.2 人工智能风险沟通能力评估 12

 5.3 人工智能风险评估能力评估 13

 5.4 人工智能风险处理能力评估 17

 5.5 人工智能风险监控能力评估 18

 5.6 人工智能风险回顾能力评估 20

附录 A（资料性）推荐权重设置 22

附录 B（资料性）风险源 24

附录 C（资料性）风险管理方法、工具与技术 27

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。

本文件由全国信息技术标准化技术委员会（SAC/TC28）提出并归口。

本文件起草单位：中国电子技术标准化研究院、中国科学院软件研究所。

本文件主要起草人：×××

引 言

随着技术的快速发展和广泛应用，人工智能系统面临的数据风险、算法风险及模型风险日益凸显，可能引发个人隐私泄露、经济安全受损等多重问题。同时，技术滥用、伦理道德争议等潜在风险也不容忽视。因此，加强人工智能风险管理，建立健全的风险防控机制，能够提升人工智能技术的可靠性和安全性，还能为技术的可持续发展奠定坚实基础，对于保障数据安全、维护社会稳定、促进经济健康发展具有重要意义。

本文件适用于对组织人工智能风险管理能力的评估。其中，组织是指人工智能系统在全生命周期、全产业链中利益相关的公司、集团公司、企事业单位、行业协会、权力机构或研究机构等。

“人工智能风险管理”是一个广泛的概念，它涵盖了与人工智能相关的所有风险的管理，包括但不限于技术风险（如算法偏差、数据泄露等）、伦理风险（如算法歧视、意识形态侵蚀等）、应用风险（如系统被黑客攻击、算法被用于欺骗等）和法律风险（如算法合规、知识侵权等）。

本文件主要关注组织对人工智能系统本身的风险管理能力，这些风险可能涉及系统的稳定性、安全性、性能以及与其他系统的兼容性等方面。人工智能系统的风险管理是为了确保系统能够按照预期运行，并避免由于系统故障或漏洞而导致的潜在问题。

需要注意的是，在谈及人工智能风险管理能力评估时，涉及到技术风险、安全风险等概念，需要结合人工智能系统的风险管理进行讨论。这些组织级风险管理能力是人工智能风险管理在具体人工智能系统中的具体实现。

人工智能 风险管理能力评估

1 范围

本文件提出了组织的人工智能风险管理能力评级与划分规则，规定了组织对于人工智能风险管理的能力框架和评估方法。

本文件适用于对组织级利益相关方所提供的人工智能风险管理能力的评估。

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 24353-2022 风险管理指南

GB/T 23694-2013 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 概念和术语 (Information technology — Artificial intelligence — Artificial intelligence concepts and terminology)

ISO/IEC 23894:2023 信息技术 人工智能 风险管理指南 (Information technology — Artificial intelligence — Guidance on risk management)

ISO/IEC 31000:2018 风险管理指南 (Risk management — Guidelines)

ISO/IEC 42001:2023 信息技术 人工智能 管理体系 (Information technology — Artificial intelligence — Management system)

3 术语和定义

GB/T 41867-2022与ISO/IEC 22989:2022、ISO/IEC 31000:2018界定的以及下列术语和定义适用于本文件。

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

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。

注1：该工程系统使用人工智能相关的多种技术和方法，开发表征数据、知识、过程等的模型，用于执行任务。

注2：人工智能系统具备不同的自动化级别。

[来源：GB/T 41867-2022，3.1.8]

3.3

风险 risk

不确定性对目标的影响。

[来源：GB/T 23694-2023，2.1]

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 23694-2023，3.1]

3.5

风险源 risk resource

单独或组合有可能产生风险的元素。

[来源：GB/T 23694-2023，4.5.1.2]

3.6

能力主域 capability domain

风险管理过程框架中相关过程的集合。

3.7

能力子域 capability sub-domain

能力主域中相关风险管理活动的集合。

3.8

能力子项 capability sub-item

能力子域中构成能力模型的若干特性指标单项。

3.9

能力子项权重 factor of capability sub-item

对应能力子项在所述能力子域评估中的权重因子。

4 人工智能风险管理能力评级

4.1 人工智能风险管理能力等级

人工智能风险管理能力等级分为初始级、规范级、全面级、量化级、卓越级共5级，见图1。

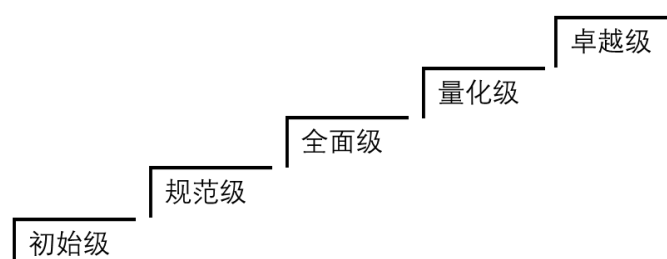


图1 人工智能风险管理能力等级

每个能力等级表明人工智能风险管理能力所达到的水平。较高的能力等级涵盖了低于其能力等级的全部要求。

4.2 人工智能风险管理能力等级划分规则

人工智能风险管理能力等级划分可采用评分量化的方式进行，风险管理能力等级的评分可参考如下公式。其中，各能力主域和能力子域的评分可分别适用相关公式进行评估：

能力等级得分的计算公式如公式（1）所示：

$$\text{Capa_level} = \sum_k (\text{capa_domain}_k \times \text{factor}_k) \quad \dots\dots\dots (1)$$

式中：

Capa_level ——能力等级得分；
capa_domain_k ——第k个能力主域的得分；
factor_k ——第k个能力主域的权重。

能力主域得分的计算公式如公式（2）所示：

$$\text{Capa_domain} = \sum_j (\text{sub_domain}_j \times \text{factor}_j) \quad \dots\dots\dots (2)$$

式中：

Capa_domain ——能力主域得分；
sub_domain_j ——第j个能力子域的得分；
factor_j ——第j个能力子域的权重。

能力子域得分的计算公式如公式（3）所示：

$$\text{Sub_domain} = \sum_i (\text{sub_item}_i \times \text{factor}_i) \quad \dots\dots\dots (3)$$

式中：

Sub_domain ——能力子域得分；
sub_item_i ——第i个能力子项的得分；
factor_i ——第i个能力子项的权重。

公式（1）、（2）、（3）中的权重设置宜参考附录A。

人工智能风险管理能力、各能力主域、能力子域的能力等级划分规则见表1表5，可根据风险管理能力评估得分确定组织的风险管理能力等级。

表1 人工智能风险管理能力等级划分参考规则

等 级	取值范围	等级说明
初始级	得分<1.5	组织具备基本的人工智能风险管理能力。组织会应用一些风险管理活动，但未建立系统的人工智能风险管理机制。风险管理往往以被动和临时的方式进行，

		风险管理结果不可预测且难以复制，通常更多地依赖于组织中人员的能力。
规范级	$1.5 \leq \text{得分} < 2.5$	组织具备可拓展的人工智能风险管理能力。组织意识到人工智能风险管理的重要性，建立了基本的风险管理流程和机制。风险管理活动被分配给有能力、职责明确、有足够资源的人员，但多数风险管理活动还是基于过去经验的重复而缺乏体系。
全面级	$2.5 \leq \text{得分} < 3.5$	组织具备稳定的人工智能风险管理能力。组织建立了全面的风险管理流程和机制，风险管理活动可以被标准化、协调和一致地推动。组织将人工智能风险管理视为组织运作的关键要素，在决策和资源分配中给予重视，并反映在绩效管理流程中。
量化级	$3.5 \leq \text{得分} < 4.5$	组织具备优秀的人工智能风险管理能力。组织建立了完备的风险管理流程和机制，且持续根据内外环境变化进行更新。组织应用定量和统计方法来管理、测量和评估风险管理过程，风险管理被纳入到组织的战略规划和业绩管理活动，风险偏好和政策得到了明确的阐述。
卓越级	$4.5 \leq \text{得分} \leq 5$	组织具备卓越的人工智能风险管理能力。组织建立了先进的、完备的风险管理流程和机制，且持续根据内外环境变化进行更新。在识别、监测和应对风险和过程中使用先进的技术工具，风险管理完全嵌入到组织的管理流程中。组织的人工智能风险管理能力已成为行业标杆，可以在整个行业内分享最佳实践的先进经验。
注：风险管理能力等级是对组织的人工智能风险管理策划能力、风险沟通能力、风险评估能力、风险处理能力、风险监控能力以及风险回顾能力等各种人工智能风险管理能力的分级评估结果。		

4.3 人工智能风险管理过程

面向人工智能的风险管理过程包括风险管理策划、风险沟通、风险评估、风险处理、风险监控、风险回顾等六个过程域，见图2。其中：

- 风险管理策划，是风险管理的起始环节，涉及确定风险管理的范围、环境和准则，旨在有针对性地设计风险管理过程。风险管理策划包括确定领导作用、制定风险政策等过程。
- 风险评估，是在风险管理策划的基础上进行，是对风险进行定性或量化分析的过程，旨在确定风险发生的可能性和影响程度。风险评估包括风险识别、风险分析和风险评价三个步骤，为风险处理提供决策依据。
- 风险处理，是组织根据风险评估的结果，制定并执行风险应对策略和措施。风险处理旨在降低风险发生的可能性或减轻风险造成的损失。
- 风险沟通，贯穿于风险管理的全过程，涉及组织内部各级别之间以及组织与外部利益相关者之间的信息交流。
- 风险监控，是对风险管理活动进行持续跟踪和监测的过程，旨在及时发现并应对新的风险或原有风险的变化。风险监控贯穿于风险管理的全过程。
- 风险回顾，是对风险管理策划、风险评估、风险处理、风险沟通与风险监控等其他五个过程域的风险管理活动与结果进行传达，从而为组织提供决策信息，以改进其风险管理活动。

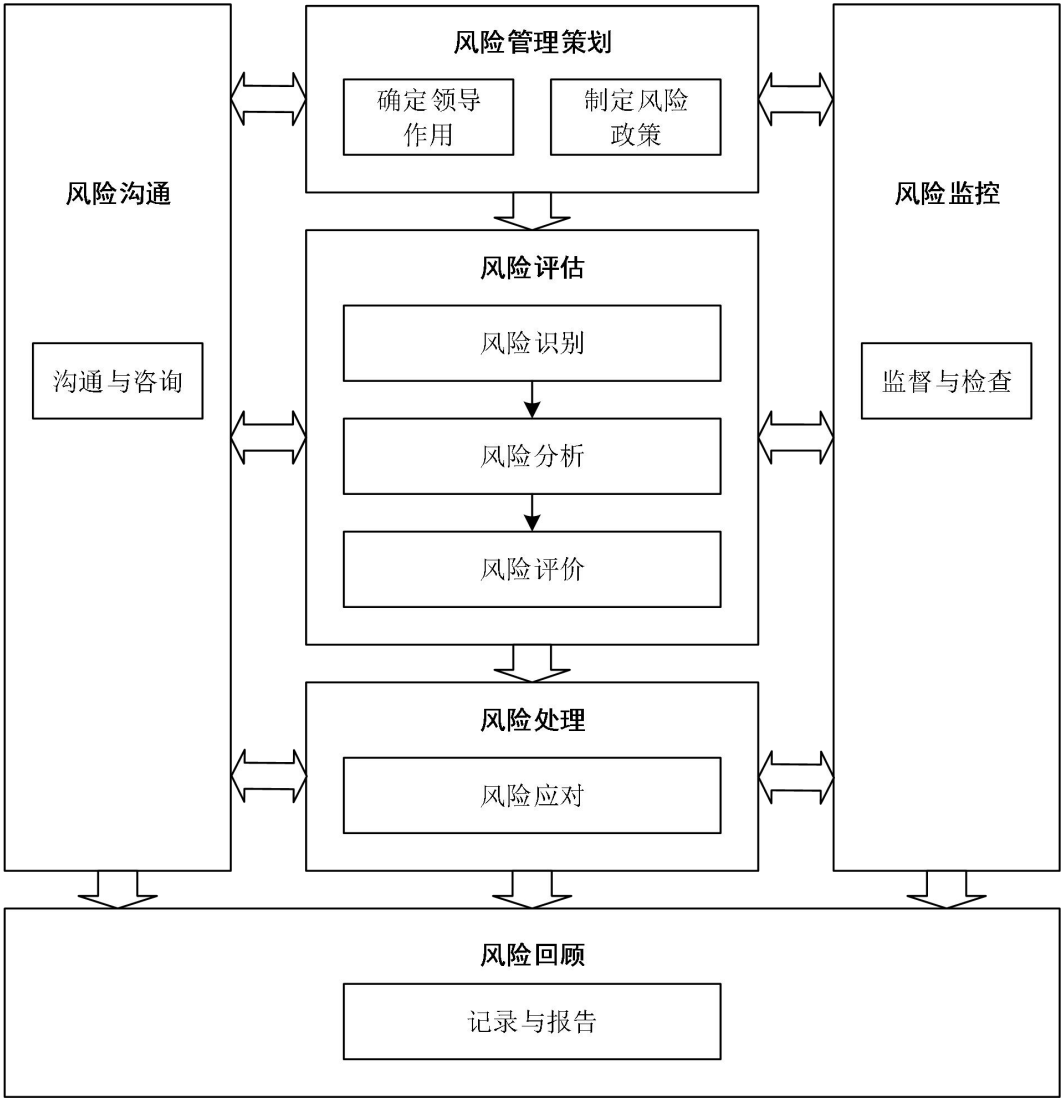


图2 人工智能风险管理过程

4.4 人工智能风险管理能力框架

人工智能风险管理能力框架的能力主域分为风险管理策划能力、风险沟通能力、风险评估能力、风险处理能力、风险监控能力以及风险回顾能力。

各能力主域的能力子域和能力子项按照衡量相关能力所需维度进行规定，其框架见表2。

表 2 风险管理能力框架

能力主域	能力子域	能力子项
风险管理策划	确定领导作用	确定组织方针
		确定人工智能风险管理目标
		进行资源协调
		指派职责
	制定风险政策	定义人工智能风险管理活动范围
		定义风险参数
		定义风险准则
		确定风险管理策略
		确定风险储备
风险沟通	沟通与咨询	风险沟通
		风险咨询
风险评估	风险识别	选择风险识别工具/技术/方法
		识别人工智能风险源
		识别风险后果
	风险分析	进行人工智能风险分类
		进行风险概率分析
		进行风险影响分析
	风险评价	建立风险概率影响矩阵
		确定风险优先级
风险处理	风险应对	选择风险应对方案
		制定和实施风险处理计划
风险监控	监督与检查	风险监督
		风险检查
风险回顾	记录与报告	风险记录
		风险报告

4.5 人工智能风险管理能力子项

4.5.1 风险管理策划能力

4.5.1.1 确定领导作用

确定领导作用能力包括：

- a) 确定组织方针：组织需要了解其内部和外部环境，包括其在人工智能系统中的角色，以及这些角色如何影响其实现人工智能管理体系目标的能力，以确定风险管理框架的设计原则；
- b) 确定人工智能风险管理目标：组织确定人工智能风险管理目标，确保风险管理目标与组织的战略目标一致；
- c) 进行资源协调：领导确保提供所需的资源，建立并维护过程所需的组织机构和岗位的职责，包括人力、技能、经验和能力，以及风险管理所需的过程、方法和工具，以支撑风险管理的实施与改进；

- d) 指派职责：领导确保负责风险管理的相关角色的权力、责任、决策流程和问责机制，确保风险管理的有效实施和持续改进。

4.5.1.2 制定风险政策

制定风险政策能力包括：

- a) 定义人工智能风险管理活动范围：组织确定其风险管理活动的范围，并与组织的战略目标和运营需求相一致；
- b) 定义风险参数：定义用于风险管理的参数，包括用于风险分析、分类和排序的参数，以及用于控制和管理风险的参数。如：风险发生的可能性、风险发生的时间段、影响程度等度量参数；
- c) 定义风险准则：组织规定其可能承担或不承担的风险的数量和类型，以及评估风险重要性和支持决策过程的标准；
- d) 确定风险管理策略：确定人工智能风险管理活动的时机、方法和管理要求；
- e) 确定风险储备：确定人工智能相关风险的应急储备和管理储备。

4.5.2 风险沟通能力

4.5.2.1 沟通与咨询

沟通与咨询能力包括：

- a) 风险沟通
 - 1) 透明沟通：确保与人工智能相关的决策过程、风险评估结果和风险管理策略对内外部利益相关者保持透明。包括清晰地传达人工智能系统的预期用途、潜在风险、已采取的缓解措施以及在发生风险时的应对计划。
 - 2) 及时沟通：信息应在风险管理过程的关键时刻及时传达，以确保利益相关者能够做出基于最新信息的决策。
 - 3) 全面沟通：确保向利益相关者传达信息同时收集和整合利益相关者的反馈和意见。沟通内容包括风险识别、评估、处理策略和结果，以及任何更新或变更。
- b) 风险咨询
 - 1) 利益相关方识别：识别所有可能受人工智能系统影响的内外部利益相关方，确保这些利益相关方的观点和利益在后续的风险管理流程中得到考虑。
 - 2) 反馈机制：确保具备系统化的反馈流程，能够收集利益相关者的意见及其对风险管理的建议。
 - 3) 多元化视角：确保咨询过程中包含多样化的观点，增加风险管理的深度和广度，组织能够从多角度审视风险。

4.5.3 风险评估能力

4.5.3.1 风险识别

风险识别能力包括：

- a) 选择风险识别工具/技术/方法：组织应根据人工智能系统的特点和组织管理需求，建立风险识别的工具/技术/方法库，并选择适合的工具/技术/方法来识别风险；
- b) 识别人工智能风险源：组织识别与人工智能的开发或使用相关的风险源；
- c) 识别风险后果：组织识别风险可能带来的直接和间接后果，评估这些后果对组织目标、利益相关者以及社会环境的潜在影响。

4.5.3.2 风险分析

风险分析能力包括：

- a) 进行人工智能风险分类：组织根据不同维度明确风险分类准则，并进行分类。如：按内部/外部分类，按技术、资源、管理等方面分类；
- b) 进行风险概率分析：基于定性或定量方法，进行风险概率分析，如：使用历史数据、专家判断、模拟等手段；
- c) 进行风险影响分析：基于定性或定量方法，进行风险影响分析，如：财务、声誉、运营等方面的影响。

4.5.4 风险处理能力

4.5.4.1 风险应对

风险应对能力包括：

- a) 选择风险应对方案：基于风险评估结果和风险管理目标，根据风险优先等级，选择适合的风险应对方案，如避免、转移、缓解、接受等；
- b) 制定和实施风险处理计划：
 - 1) 制定风险处理计划：计划内容可包括风险应对方案的理由、预期收益、涉及的实施及批准人员、具体活动、所需资源（包括财务预算和人力资源）、绩效指标、时间限制、监控和审查机制等。
 - 2) 实施风险处理计划：按照风险处理计划执行风险应对措施，确保计划的实施与组织目标和风险管理策略一致。定期监控风险处理措施的效果，评估是否达到预期的降低风险的目标。基于监控结果和业务环境的变化，定期审查，在必要时调整风险处理计划。

4.5.5 风险监控能力

4.5.5.1 风险评价

风险评价能力包括：

- a) 建立风险概率影响矩阵：根据风险概率分析和风险影响分析的结果，基于定性或定量方法，建立风险概率影响矩阵；
- b) 确定风险优先级：根据风险参数划分等级（如高、中、低）并根据参数进行风险排序。

4.5.5.2 监督与检查

风险监督与检查能力包括：

- a) 风险监督：收集风险管理信息、监督风险处理措施的实施情况、记录监督过程中发生的问题和结果，并提供有利于促进风险管理过程持续改进的反馈；
- b) 风险检查：对风险管理效果进行检查，将检查结果纳入整个组织的绩效管理中。

4.5.6 风险回顾能力

4.5.6.1 记录与报告

风险记录与报告能力包括：

- a) 风险记录：确定风险记录的信息需求、要求、方法、频率，为风险管理决策提供信息；

- b) 风险报告：确定报告的方法、成本、频率和及时性，传达整个组织的风险管理活动和结果，从而改进风险管理活动。

5 人工智能风险管理能力评估

5.1 人工智能风险管理策划能力评估

5.1.1 确定领导作用

确定领导作用能力子域的能力子项能力评估分值见表3。

表 3 确定领导作用能力评估描述及分值

能力子项	评估描述	分值
确定组织方针	高： 组织将人工智能风险管理作为整体业务发展战略的一部分，能够牵头或参与人工智能风险管理相关的国际标准、国家标准及行业标准的制定，能够评估现有人工智能风险管理策略方针和管理制度的适用性，对人工智能系统全生命周期制定完整的风险管理策略和风险应对方案，并明确组织的最高管理者对风险管理的承诺。 注：本文件涉及的人工智能系统生命周期的各个阶段应符合 ISO/IEC 42001:2023《信息技术-人工智能-管理系统》(Information technology — Artificial intelligence — Management system) 等标准文件的相关规定。	5
	中： 组织将人工智能风险管理纳入到组织整体风险管理领域，建立较规范的人工智能风险管理框架与流程。	3
	低： 组织仅将人工智能风险管理纳入组织整体风险管理领域，建立基本的人工智能风险管理框架与流程。	1
确定人工智能风险管理目标	高： 具备卓越的风险管理能力，能够对新型风险进行识别，建立完善的风险应对措施，确保组织战略目标达成。	5
	中： 具备发展的风险管理能力，能够对关键的、常见的人工智能风险进行识别和应对，确保风险管理正常运行。	3
	低： 具备基本的风险管理能力，能够识别和应对重大风险，以符合相关法律法规的要求、避免业务运营中止等。	1
进行资源协调	高： (1) 人员：设置具备丰富风险管理经验的专职中高层风险管理人员，能够实现实时、动态的风险管理，建立人工智能相关人才激励政策，将人工智能风险管理与考核机制结合。 (2) 工具：使用专用软件系统工具实现完善的智能化管理。	5
	中：	3

	(1) 人员：设置具备一般风险管理能力的专职风险管理人员，能够实现基本的风险管理。	
	(2) 工具：使用通用软件系统工具支持基本的风险管理活动。	
指派职责	低： (1) 人员：未设专职风险管理人员，由其他职能的人员兼职管理，能够完成最低限度的风险管理活动。	1
	(2) 工具：使用通用软件系统工具支持最低限度的风险管理活动。	
	高： (1) 组织建立专门的风险管理委员会，定期向组织高级管理层汇报人工智能发展与风险管理情况，确保人工智能风险管理活动具有高优先级。	5
	(2) 组织设置至少2名具备独立性的专职风险管理人员，明确其权力、责任、决策流程、问责与奖励机制，以保障人工智能风险管理活动的全面、有效开展和持续改进。	
	中： 组织设置至少1名具备独立性的专职风险管理人员，明确其权力、责任、决策流程与问责机制，以保障人工智能风险管理活动的开展。	3
	低： 组织未设置专职风险管理人员，一般由不具备独立性的兼职人员承担人工智能风险管理职责，其权力、责任、决策流程或问责机制未明确。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.1.2 制定风险政策

制定风险政策能力子域的能力子项评估分值见表4。

表 4 制定风险政策能力评估描述及分值

能力子项	评估描述	分值
定义人工智能风险管理活动范围	高： (1) 领导参与：在定义人工智能风险管理活动范围时，包含高层管理者的全程参与和决策支持，以确保所有活动所需资源。 (2) 一致性：在定义风险管理活动时，充分考虑了各项活动与组织风险管理目标以及组织战略目标的一致性。 (3) 全面性：充分考虑到风险管理活动需要贯穿组织的各个层面和部门，充分考虑到所有相关部门的协作和配合，确保全面覆盖和有效应对。 (4) 目标与结果：明确风险管理活动的目标与预期结果。	5
	中： (1) 领导参与：在定义人工智能风险管理活动范围时，由中层管理者负责，能够保障重要活动所需资源。 (2) 一致性：在定义风险管理活动时，大多数活动与组织风险管理目标以及组织战略目标的一致性基本符合，少量活动的一致性存在偏离。 (3) 全面性：考虑了部分重要相关部门之间的协作和配合，确保重点覆盖和有效应对。	3

	(4) 目标与结果：基本明确风险管理活动的目标与预期结果。	
	低： (1) 领导参与：在定义人工智能风险管理活动范围时，仅由项目经理及项目团队负责，缺乏重要活动所需资源的支持。 (2) 一致性：在定义风险管理活动时，忽略了各项活动与组织战略目标以及风险管理目标的一致性，甚至一致性存在显著偏离。 (3) 全面性：仅在部门内部进行，未考虑跨部门的协作和配合，未做到全面覆盖，无法有效应对。 (4) 目标与结果：未明确或无法明确风险管理活动的目标或预期结果。	1
定义风险参数	高： (1) 风险参数类别：风险参数涵盖技术、财务、人力、设备、环境等5个或以上的维度。 (2) 风险参数边界划分：风险参数边界划分清晰。 (3) 风险参数可用性：风险参数从技术和经济上来说，完全能够获得和可用。 (4) 定义了共同且一致的准则，用以比较需要管理的各种风险。	5
	中： (1) 风险参数类别：风险参数涵盖3-5（含）种维度。 (2) 风险参数边界划分：风险参数边界划分基本清晰。 (3) 风险参数可用性：风险参数从技术和经济上来说，基本能够获得和可用。 (4) 准则定义部分一致，可以比较需要管理的部分风险。	3
	低： (1) 风险参数类别：风险参数仅涵盖1-2种维度。 (2) 风险参数边界划分：风险参数边界划分不清晰。 (3) 风险参数可用性：风险参数从技术和经济上来说，难以获得或可用。 (4) 无准则或准则定义完全不一致，无法比较需要管理的不同风险。	1
定义风险准则	高： (1) 风险数量与类型：明确组织能够接受或不可接受的风险的数量与类型（参考附录B），并给出清晰的风险定义与阈值范围。 (2) 风险组合：能够考虑到多种风险的组合、关联以及顺序。 (3) 明确风险评估：清晰地确定了评估风险重要性和支持决策过程的标准。 (4) 定义风险准则的频度：在完成初始化之后，能够按需、动态地进行风险准则的审查与修订。	5
	中： (1) 风险数量与类型：基本明确组织能够接受或不可接受的风险的数量与类型（参考附录B），能够给出部分风险定义与阈值范围。 (2) 风险组合：能够考虑到少数重要风险的组合、关联以及顺序。 (3) 明确风险评估：部分确定了评估风险重要性和支持决策过程的标准。 (4) 定义风险准则的频度：在完成初始化之后，定期进行风险准则的审查与修订。	3
	低： (1) 风险数量与类型：未明确组织能够接受或不可接受的风险的数量与类型（参考附录B），风险定义与阈值范围不清晰或无法给出。	1

	(2) 风险组合：未考虑到多种风险的组合、关联以及顺序。 (3) 明确风险评估：未确定评估风险重要性和支持决策过程的标准。 (4) 定义风险准则的频度：在完成初始化之后，无法定期进行风险准则的审查与修订。	
确定风险管理策略	高： (1) 风险管理的主动性：倾向于主动预测和管理风险，而不是仅当风险发生时才采取行动。主动定义不可接受的风险阈值，并制定风险接近阈值的相应应对措施。 (2) 风险管理全面性：考虑组织整体的风险状况，覆盖所有关键业务领域和潜在风险源。 (3) 风险管理的深度：对于每种风险，都进行深入的分析 and 评估，以确定其潜在影响、可能性、紧急性和阈值。 (4) 风险管理的演进性：对风险管理工具、技术与方法不断演进、创新，从而不断提升组织的风险管理能力等级。	5
	中： (1) 风险管理的主动性：当风险发生或接近阈值时，能够迅速做出反应，并采取相应的措施来减轻风险的影响。 (2) 风险管理全面性：针对组织内的重要业务领域和主要风险源。 (3) 风险管理的深度：对风险进行适度的分析和评估，以确定其潜在影响、可能性和阈值。 (4) 风险管理的演进性：能够适应一定程度的环境变化，并调整风险管理工具、技术与方法，以适应新的风险状况。	3
	低： (1) 风险管理的主动性：仅在风险发生后才采取行动，而不是主动预测和管理风险。 (2) 风险管理全面性：仅关注组织内的某些局部领域或特定风险源。 (3) 风险管理的深度：对风险进行基本的分析和评估，但不够深入。 (4) 风险管理的演进性：采用相对稳定的风险管理策略，应用固定的风险管理工具、技术与方法，不进行相应的调整。	1
确定风险储备	高： 对于人工智能可能造成的风险，在关注度和资源上付出与其可能造成的影响相称的投入，或具备项目总额占比3%以上的风险储备预算。	5
	中： 具备项目总额占比1%~3%（含）的风险储备预算。	3
	低： 风险储备预算不为0，但占比低于项目总额的1%。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.2 人工智能风险沟通能力评估

沟通与咨询能力子域的能力子项能力评估分值见表 5。

表 5 沟通与咨询能力评估描述及分值

能力子项	评估描述	分值
风险沟通	高： 确保所有相关方及时、准确地理解人工智能相关风险信息。包括使用清晰、易懂的语言，根据不同受众的理解水平和需求，采用适当的沟通方式和工具（如会议、报告、在线平台）进行沟通等。在传递过程中，能有效地传达风险的重要性和紧迫性、风险管理措施的成果。	5
	中： 具备基本沟通能力，能够向相关方提供大部分必要的风险信息。在沟通内容的精准度和深度上存在一定的限制。	3
	低： 沟通能力较弱，难以有效地向利益相关者传达人工智能相关风险信息，可能导致信息不完整、不准确。	1
风险咨询	高： 能够整合来自内外部利益相关方的反馈和建议为组织和项目团队提供全面的人工智能风险管理建议和解决方案。能够通过深入分析和评估，为决策者提供具体的战略建议，帮助他们理解和应对人工智能系统可能面临的各类风险。	5
	中： 具备基本的咨询能力，能够为组织和项目团队提供一般性的人工智能风险管理建议。能够识别并联系一些关键的内外利益相关方，但在整合多样性观点和反馈方面可能存在局限，咨询过程不够系统和全面。	3
	低： 咨询能力有限，很少或没有主动征求内外利益相关方意见，仅能就人工智能风险的基本问题提供简单的建议，可能无法满足复杂决策者的需求。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.3 人工智能风险评估能力评估

5.3.1 风险识别

风险识别能力子域的能力子项能力评估分值见表6。

表 6 风险识别能力评估描述及分值

能力子项	评估描述	分值
选择风险识别工具/技术/方法	高： （1）工具清单：拥有完备且前沿的人工智能风险识别工具/技术/方法清单（参考附录C），以确保涵盖所有相关领域。能够在现有工具/技术/方法的基础上进行整合与创新，提出新的风险识别方法。 （2）运用能力：能够根据实际场景，运用已经掌握的人工智能风险识别工具/技术/方法，对每种风险提出2套或者以上的完整解决方案。 （3）匹配度和有效性：选择的工具/技术/方法准确、适用，能够高效地识别项目、产品或过程中的潜在风险。	5
	中： （1）工具清单：拥有较完备的人工智能风险识别工具/技术/方法清单。	3

	(2) 运用能力：能够在实际场景中，运用已有的人工智能风险识别工具/技术/方法，对每种风险提出至少1套完整的解决方案。	
	(3) 匹配度和有效性：选择的工具/技术/方法部分准确、适用，能够识别项目、产品或过程中的主要潜在风险。	
识别人工智能风险源	低： (1) 工具清单：拥有初步的人工智能风险识别工具/技术/方法清单，或无清单。 (2) 运用能力：能够简单了解人工智能风险识别工具/技术/方法并进行运用。 (3) 匹配度和有效性：选择的工具/技术/方法不准确或不适用，能够识别一些常见风险，但可能忽略一些潜在风险。	1
	高： 对人工智能系统、人工智能技术原理、应用场景以及潜在风险源有深入的理解和广泛的知识，能够识别出各种风险源，并且能够动态更新和完善风险源列表（参考附录B）。	5
	中： 对人工智能系统及其风险源比较熟悉，能够识别出常见的风险源，能够定期更新重要风险源。	3
	低： 对人工智能系统及其风险源有初步的认识和了解，但缺乏深入的理解，无法全面识别人工智能风险源。	1
识别风险后果	高： (1) 风险后果影响对象：能够全面识别风险对系统、业务、个人、组织、社会等利益相关方的后果影响。 (2) 风险后果影响维度：能够全面识别风险的财务影响（组织财务状况的潜在影响）、运营影响（组织日常运营的影响）、法务影响（是否可能导致组织违反法律法规或面临法律诉讼）、声誉影响（对组织声誉和品牌形象的潜在影响）、其他影响（社会影响、伦理影响）等。	5
	中： (1) 风险后果影响对象：能够基本识别风险对系统、业务、组织等重要利益相关方的后果影响。 (2) 风险后果影响维度：能够重点识别风险的财务影响（组织财务状况的潜在影响）、运营影响（组织日常运营的影响）、法务影响（是否可能导致组织违反法律法规或面临法律诉讼）等。	3
	低： (1) 风险后果影响对象：仅能够识别风险对系统、业务的后果影响。 (2) 风险后果影响维度：能够基本识别风险的财务影响（组织财务状况的潜在影响）、运营影响（组织日常运营的影响）等。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.3.2 风险分析

风险分析能力子域的能力子项评估分值见表7。

表 7 风险分析能力评估描述及分值

能力子项	评估描述	分值
进行人工智能风险分类	高： 出众的分类能力：能够将面临的风险进行详细的分类，类别达到5个维度或以上，包括与传统软件系统相近的风险、人工智能系统数据相关风险、人工智能系统硬件相关风险、人工智能系统框架相关风险、人工智能特性相关风险等，分类结果能够明确风险源和具体风险类型，直观准确。	5
	中： 一般的分类能力：能够将面临的风险进行基本的分类，类别达到3-4个维度，包括与传统软件系统相近的风险、人工智能系统相关风险、环境复杂性相关风险等，分类结果能够明确风险源，但缺乏对具体风险类型的分类能力。	3
	低： 有限的分类能力：能够将面临的风险进行大致的分类，类别达到1-2个维度，包括人工智能系统相关风险、非人工智能系统相关风险等，分类结果不足以详细区分不同的风险源或类型，且在风险多样性和准确性方面表现不足。	1
进行风险概率分析	高： (1) 分析方法：以定量分析为主，能够在复杂数据环境下基于大量数据，熟练运用先进的统计方法和数据分析算法，对风险概率进行量化。对于不可量化的不确定性，即使缺乏精确的数据或统计方法，也能够运用定性分析和专家判断，对潜在的风险概率进行综合评估和分析，从而为决策提供依据。 (2) 分析因素：能够深入分析单个风险因素，同时理解多个风险源之间的相互关联和影响，进行多因素综合分析，增强概率预测的全面性、准确性和可靠性。 (3) 分析结果：得到量化的风险事件概率，并且具有基于最新数据和环境变化进行风险概率动态更新的能力，能够及时调整预测模型，确保概率预测始终基于当前的条件和最新的信息。	5
	中： (1) 分析方法：定性与定量方法相结合，能够通过定性分析评估风险事件的可能性，并结合定量数据进行简单的概率估计。 (2) 分析因素：能够理解并分析关键、重要的风险因素，并针对单一风险源进行有针对性的分析。 (3) 分析结果：能够将风险事件的概率分类为高、中、低三个等级。	3
	低： (1) 分析方法：能够进行基本的定性分析（直观判断或非结构化的方法），缺乏深入的数据支持和详细的量化分析。 (2) 分析因素：能够对直接和表面的风险因素进行分析，缺乏深入分析单个风险因素或其相互作用的能力。 (3) 分析结果：仅能够提供风险概率的大致估计，判断风险事件发生的概率是否高于或低于某个预设值。	1
进行风险影响分析	高： (1) 分析方法：以定量分析为主，能够熟练运用先进的统计方法和数据分析算法，对风险事件可能带来的负面影响进行精确的量化。对于不可量化的风险影响，能够运用定性分析和专家判断，对潜在的风险影响进行综合评估和分析，从而为决策提供依据。	5

	(2) 分析维度：能够综合考虑财务、声誉、运营、法律、环境等多个维度的负面影响，能够分析风险事件的影响传播链条，从而全面评估风险事件的总体负面影响。 (3) 分析结果：能够得到量化的风险事件概率，全面、准确地预测风险对组织和系统的综合影响。	
	中： (1) 分析方法：定性与定量方法相结合，能够结合定性描述和定量数据分析，对风险影响进行初步评估，使用案例分析、专家意见和历史数据来进行风险的负面影响分析。 (2) 分析维度：能够分析部分影响维度，大致估计风险事件可能影响的业务和相关职能，但在分析的维度、层次、精度方面有所欠缺。 (3) 分析结果：能够将风险事件的影响分类为高、中、低三个等级。	3
	低： (1) 分析方法：能够依赖基本的定性分析（直观判断或非结构化的方法），将风险事件的负面影响分类为高、中、低三个等级，缺乏深入的数据支持和详细的量化分析。 (2) 分析维度：能够分析有限的影响维度，对风险事件的负面影响的理解通常局限于最直接和显著的影响。缺乏对可能发生的连锁反应或更深层次影响的探讨，不能充分理解风险事件在不同业务和操作层面上的潜在后果。 (3) 分析结果：仅能够关注最显而易见的风险影响，判断风险事件发生的概率是否高于或低于某个预设值。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.3.3 风险评价

风险评价能力子域的能力子项评估分值见表8。

表 8 风险评价能力评估描述及分值

能力子项	评估描述	分值
建立风险概率影响矩阵	高： (1) 矩阵：能够建立二维、9个象限的风险概率影响矩阵，利用横轴与纵轴的标识刻度在矩阵中定位各风险的对应位置，支持对风险概率影响的定量分析。 (2) 更新：能够实时、动态地更新风险概率影响矩阵，并根据新的评估结果调整风险应对策略。	5
	中： (1) 矩阵：能够建立二维、9个象限的风险概率影响矩阵，能够对风险概率影响进行定性-定量相结合的分析。 (2) 更新：能够按月/周的频度，对风险概率影响矩阵进行定期更新。	3
	低： (1) 矩阵：能够建立二维、4个象限的风险概率影响矩阵，能够对风险概率影响进行基本的定性分析。 (2) 更新：在完成风险概率影响矩阵初始化后，不再定期更新。	1
确定风险优先级	高： 概率影响评估：同时具备定量分析和定性分析的能力，对风险概率影响进行优先级排	5

	序。	
	中： 概率影响评估：基于定性-定量分析，对风险概率影响进行优先级排序。	3
	低： 概率影响评估：基于定性分析，对风险概率影响进行优先级排序。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.4 人工智能风险处理能力评估

风险应对能力子域的能力子项能力评估分值见表9。

表 9 风险应对能力评估描述及分值

能力子项	评估描述	分值
选择风险应对方案	高： （1）多样化且主动的应对方案：针对高风险，能够提出3种或以上风险应对方案，包括风险规避、风险减轻、风险转移和风险接受等，主动预测和管理风险的方案明确、具体，并能根据实际情况灵活选择，并考虑了应对方案失败后的措施。 （2）成本效益分析：在选择应对方案时，会综合考虑成本、时间和资源投入，以及预期收益和潜在风险，确保选择的方案在符合风险管理策略的前提下具有最佳的成本效益。 （3）决策质量：决策过程严谨、科学、透明、可追溯，能够避免重大遗漏或误判，确保决策符合法律法规和组织的价值观。在符合法律和社会规范下，决策结果符合组织的长远利益最大化原则。	5
	中： （1）基本的应对方案：针对高风险，能够提出2种应对方案，方案基本明确、具体。 （2）基本成本效益分析：在选择应对方案时，会考虑基本的成本效益因素，但可能缺乏深入的分析和比较。 （3）决策质量：决策过程基本合理，但可能存在一些遗漏或误判，决策结果基本符合组织的利益。	3
	低： （1）有限的应对方案：针对高风险，没有应对方案，或只能够提出1种风险应对方案，但缺乏多样性和灵活性，方案不明确、具体。 （2）成本效益分析：在选择应对方案时，较少考虑成本效益因素，可能导致显著的资源浪费或收益不足。 （3）决策质量：决策过程可能存在较多遗漏或误判，决策结果可能不符合组织的利益最大化原则。	1
制定和实施风险处理计划	高： （1）制定风险处理计划 1) 明确了风险处理责任人和主要的备援人员，以及各项参与相关活动的利益相关者。 2) 明确了具体的时间节点、活动优先级排序、活动依赖关系以及风险储备情况，能够保障所有活动所需资源。 （2）实施过程 1) 团队成员对风险处理计划有深入的理解和认识。	5

	2) 执行流程清晰、高效，监控机制完善，能够及时发现和处理问题。 3) 应对措施能够按时、高效执行，并能够在必要时进行灵活调整，风险得到有效应对。	
	中： （1）制定风险处理计划 1) 明确了风险处理责任人。 2) 明确了风险处理的具体时间节点，能够保障重要活动所需资源。 （2）实施过程 1) 有部分团队成员对风险处理计划理解不够深入。 2) 有明确的执行流程，但执行中可能存在小范围的延误或不足。 3) 应对措施基本合理，能够按计划执行并基本有效应对。	3
	低： （1）制定风险处理计划 1) 指定了各项参与相关活动的利益相关者，但无明确的风险处理责任人。 2) 确定了风险处理可接受的时间范围。 （2）实施过程 1) 团队成员对风险处理计划缺乏基本理解，认知存在较大偏差。 2) 实施过程混乱，缺乏明确的执行流程。 3) 应对措施失效或执行不及时，导致风险未得到有效应对。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.5 人工智能风险监控能力评估

监督与检查能力子域的能力子项评估分值见表 10。

表 10 监督与检查能力评估描述及分值

能力子项	评估描述	分值
风险监督	高： （1）监督主体：有明确的责任机制，有专职部门或多名专职人员负责持续监控和应急响应。 （2）监督范围：监督大部分关键业务领域和潜在风险点，能够涵盖已有清单上 80%或以上的人工智能风险源，并能够动态、实时地扩充或更新人工智能风险源清单。 （3）监督频度：实时或准实时风险监督，确保对潜在风险变化的快速响应。 （4）监督工具：建立符合法律法规的监控系统，监督风险处理措施的实施情况，实时监测人工智能系统性能、可用性和安全状况，以及对输入、输出数据、模型算法等进行异常检测，并自动化告警。监督工具定期维护和升级，能适应不断变化的风险环境和法规要求。 （5）响应速度：设有快速响应机制，能够在发现风险后迅速采取控制措施（如8小时内完成风险处理）。	5
	中：	3

风险检查	(1) 监督主体：有至少一名专职人员负责监控和应急响应。 (2) 监督范围：监督重要关键业务领域和潜在风险点，能够涵盖已有清单上50%-80%的人工智能风险源，能够定期扩充或更新人工智能风险源清单。 (3) 监督频度：定期（如每日、每周）进行风险监督。 (4) 监督工具：使用分析工具和模型（参考附录C）进行半自动的风险监督，监督风险处理措施的实施情况，定期对监督模型进行审查和更新。 (5) 响应速度：设有一定的响应机制，能够在发现风险后采取相应措施（如8-24小时之间）。	
	低： (1) 监督主体：有兼职人员或无人员负责监控或应急响应。 (2) 监督范围：仅监督少量关键业务领域和有限的风险点，仅能关注已有清单50%以下的人工智能风险源。 (3) 监督频度：不定期或偶尔进行风险监督。 (4) 监督工具：使用基本的分析方法（参考附录C）进行全人工的风险监督，监督工具或模型更新不频繁，或仅在发现问题时进行调整。 (5) 响应速度：响应速度较慢，可能需要较长时间才能采取控制措施（如超过24小时）。	1
	高： (1) 风险检查的覆盖范围：检查深入、全面，涵盖了80%或以上潜在的风险源（见附录B）以及所有风险管理过程或活动。 (2) 风险检查的及时性：检查及时，能够在风险应对后24小时内完成。 (3) 风险检查的准确性：检查结果准确、可靠，基于充分的数据和分析。 (4) 风险检查的实用性：检查结果对风险管理和控制有显著的贡献，提供了有价值的建议和改进措施。 (5) 风险检查的可靠性：检查结果是基于以下4项内容： 1) 对人工智能系统使用的数据进行全面审计。 2) 对人工智能算法进行审查。 3) 对人工智能系统进行安全测试。 4) 评估人工智能系统在不同场景下的表现和风险。 (6) 风险检查绩效管理：风险检查的结果纳入中高层领导、部门负责人以及团队成员的绩效考核。	5
	中： (1) 风险检查的覆盖范围：检查较为全面，涵盖了50-80%的潜在风险源（见附录B）以及关键风险管理过程或活动。 (2) 风险检查的及时性：检查及时性尚可，能够在风险应对后24-72小时内完成。 (3) 风险检查的准确性：检查结果基本准确，但可能存在一些误差或不足。 (4) 风险检查的实用性：检查结果对风险管理和控制有一定的贡献，提供了一些建议和改进措施。 (5) 风险检查的可靠性：检查结果是基于以下任意2-3项： 1) 对人工智能系统使用的数据进行全面审计。 2) 对人工智能算法进行审查。 3) 对人工智能系统进行安全测试。	3

	4) 评估人工智能系统在不同场景下的表现和风险。 (6) 风险检查绩效管理: 风险检查的结果纳入部门负责人与团队成员的绩效考核。	
	低: (1) 风险检查的覆盖范围: 检查全面性一般, 仅涵盖50%以下的潜在风险源(见附录B)及少量风险管理过程或活动。 (2) 风险检查的及时性: 检查及时性一般, 通常在72小时以上完成。 (3) 风险检查的准确性: 检查结果可能存在较大的误差或不足, 无法准确反映实际情况。 (4) 风险检查的实用性: 检查结果对风险管理和控制的贡献有限, 缺乏有价值的建议和改进措施。 (5) 风险检查的可靠性: 检查结果是基于以下0-1项: 1) 对人工智能系统使用的数据进行全面审计。 2) 对人工智能算法进行审查。 3) 对人工智能系统进行安全测试。 4) 评估人工智能系统在不同场景下的表现和风险。 (6) 风险检查的绩效考核: 风险检查的结果仅纳入团队成员的绩效考核, 或未纳入绩效考核。	1
注: 以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

5.6 人工智能风险回顾能力评估

记录与报告能力子域的能力子项评估分值见表11。

表 11 记录与报告能力评估描述及分值

能力子项	评估描述	分值
风险记录	高: (1) 记录工具: 通过智能化项目管理系统建立、记录和维护一个风险清单或风险跟踪表, 制定统一的风险记录格式, 确保所有风险记录的一致性和可比性。 (2) 记录内容: 包括并不限于所分析人工智能系统的预期用途说明、风险评估组织、风险评估人员、风险描述、风险识别过程、风险概率影响、风险应对预案等信息, 能够根据项目/产品的特点进行扩展, 能够保证已识别风险在所有风险管理过程中的可追踪性。 (3) 记录频度: 基于事件驱动, 及时、动态地记录风险。	5
	中: (1) 记录工具: 通过基础的项目管理工具建立、记录和维护一个风险清单或风险跟踪表。 (2) 记录内容: 包括所分析人工智能系统的预期用途说明、风险评估组织、风险评估人员、风险描述、风险识别过程、风险概率影响、风险应对预案等全部内容。 (3) 记录频度: 每月定期记录风险。	3
	低:	1

	(1) 记录工具：通过看板、文档等工具建立、记录和维护一个风险清单或风险跟踪表。 (2) 记录内容：包括所分析人工智能系统的预期用途说明、风险评估组织、风险评估人员、风险描述、风险识别过程、风险概率影响、风险应对预案等部分内容。 (3) 记录频度：仅在项目启动或里程碑节点，不定期地记录风险。	
风险报告	高： (1) 报告内容：根据项目需求，动态形成重大风险管理情况报告，清晰、准确地描述风险，根据风险源（参考附录B）、风险管理全过程（治理、沟通、评估、处理、监控、回顾等）、风险级别（高/中/低）、风险影响范围等因素进行详细分析。 (2) 报告频度：根据项目需求，动态、实时地进行报告。 (3) 风险管理复盘：由项目负责人组织，风险利益相关者及中高层领导参加，根据风险报告，组织详细的复盘与分析，形成书面经验与教训，使得风险管理的全过程纳入组织过程资源中。	5
	中： (1) 报告内容：比较全面地描述风险处理全过程。 (2) 报告频度：在项目例会（如每日例会、周例会、双周会等频度）进行报告。 (3) 风险管理复盘：团队重要利益相关者参加，根据风险报告进行比较全面的复盘与分析，能够形成会议纪要。	3
	低： (1) 报告内容：概要地描述风险处理过程。 (2) 报告频度：按月度以上频度进行风险报告，缺乏时效性。 (3) 风险管理复盘：团队少数重要成员仅进行简单的口头复盘。	1
注：以上能力未达到最低评分要求或完全不具备该能力项功能的应评为0分。		

附 录 A
(资料性)

推荐权重设置

表A. 1给出了能力主域、能力子域、能力子项的推荐权重值设置。

表 A. 1 能力子项权重

能力主域	能力子域	能力子项	能力子项权重
风险管理策划 (权重 20%)	确定领导作用 (权重 50%)	确定组织方针	25%
		确定人工智能风险管理目标	25%
		进行资源协调	25%
		指派职责	25%
	制定风险政策 (权重 50%)	定义人工智能风险管理活动范围	20%
		定义风险参数	20%
		定义风险准则	20%
		确定风险管理策略	20%
		确定风险储备	20%
风险沟通 (权重10%)	沟通与咨询 (权重100%)	风险沟通	50%
		风险咨询	50%
风险评估 (权重 25%)	风险识别 (权重 40%)	选择风险识别工具/技术/方法	30%
		识别人工智能风险源	40%
		识别风险后果	30%
	风险分析 (权重 30%)	进行人工智能风险分类	26%
		进行风险概率分析	36%
		进行风险影响分析	38%
	风险评价 (权重 30%)	建立风险概率影响矩阵	50%
		确定风险优先级	50%
风险处理	风险应对	选择风险应对方案	50%

（权重 20%）	（权重 100%）	制定和实施风险处理计划	50%
风险监控 （权重 15%）	风险监督与检查 （权重 100%）	风险监督	50%
		风险检查	50%
风险回顾 （权重 10%）	风险记录与报告 （权重 100%）	风险记录	50%
		风险报告	50%

附 录 B
(资料性)
风险源

表 B. 1 给出了人工智能风险源的分类。

表 B. 1 人工智能-风险源

风险源	风险
风险源1：与传统软件系统相近的风险	需求风险
	成本风险
	进度风险
	质量风险
	技术风险
	人员风险
	法律风险（合规性）
风险源2：人工智能系统数据相关风险	数据质量风险（数据不完备、数据选择偏差、数据投毒等）
	数据安全风险（数据泄露、数据篡改、数据删除等）
	人工智能数据处理过程风险（数据采集、数据生成、数据预处理、特征提取、数据传输、数据存储风险等）
风险源3：人工智能系统硬件相关风险	硬件系统缺陷
	硬件系统软错误
	硬件系统性能限制
	硬件系统网络风险
	硬件系统兼容性风险
	硬件维护和更新风险
风险源4：人工智能系统框架相关风险	安全可控性风险
	硬件兼容性风险
	软件兼容性风险
风险源5：人工智能属	有效性

性相关风险	可靠性
	（物理）安全性
	（信息）安全性
	韧性
	可问责性
	透明性
	可解释性
	隐私保护性
	公平性
风险源6：人工智能系统生命周期相关风险	计划与设计
	收集与处理数据
	构建与使用模型
	验证与确认
	部署与应用
	运行与监测
风险源7：环境复杂性相关风险	自然环境（天气环境、地理环境、电磁环境、气候变化、地理孤立性）
	宏观环境（政治环境、军事环境、经济环境、社会文化环境）
	实施环境（软件环境、硬件环境、网络环境）
	基础服务环境（存储、计算、安全与隐私、机器学习、数据治理）
	业务场景环境（智能无人系统、自动驾驶、智能语音、医疗健康、金融服务、工业自动化等）
风险源8：恶意使用风险（滥用风险）	应用于生成虚假内容对个人造成伤害（欺诈、钓鱼攻击、深度伪造色情视频等）
	应用于虚假信息和操纵公众舆论（试图影响政治、选举）
	应用于网络攻击（扩大和部分自动化某些类型的网络操作）
	应用于双重用途科学风险（例如生物、化学、放射、核威胁等武器的开发和恶意使用）

风险源9：人工智能内生（自身）安全风险	模型算法安全风险（可解释性差的风险、偏见与歧视风险、鲁棒性弱的风险、被窃取或篡改的风险、输出不可靠的风险、对抗攻击风险）
	数据安全风险（违规收集使用数据风险、训练数据含不当内容或被“投毒”风险、训练数据标注不规范风险、数据泄露风险）
	系统安全风险（缺陷或后门被攻击利用风险、算力安全风险、供应链安全风险）
风险源10：人工智能应用安全风险	网络域风险（信息内容安全风险、混淆事实/误导用户/绕过鉴权的风险、不当使用引发信息泄露风险、滥用于网络攻击的风险、模型复用的缺陷传导风险）
	现实域风险（诱发传统经济社会安全风险、用于违法犯罪活动的风险、两用物项和技术滥用风险）
	认知域风险（加剧“信息茧房”效应风险、用于开展认知战的风险）
	伦理域风险（加剧社会歧视偏见/扩大智能鸿沟的风险、挑战传统社会秩序的风险、未来脱离控制的风险）

附 录 C

(资料性)

风险管理方法、工具与技术

表C.1给出了部分人工智能风险管理方法、工具与技术。

表 C.1 人工智能-风险管理方法、工具与技术

序号	方法/工具/技术名称	方法/工具/技术说明	所属能力子域
1	规划会议和分析	项目经理、团队成员等组织内与风险管理相关人员进行规划和 分析	制定风险政策
2	专家判断	组建由项目内外部专家组成的小组，利用其专业知识和经验对 信息进行分析和判断，以决定是否批准某项决议或采取措施	风险识别
3	情景分析	通过构建不同场景中人工智能技术及其应用可能出现的问题 和结果，对人工智能技术及其应用（可能）面临的风险进行模 拟和推演	风险识别
4	访谈	与干系人直接交流，通常是一对一	风险识别
5	调查问卷法	设计并发放问卷，调查利益相关者对人工智能技术及其应用的 风险看法和态度	风险识别
6	头脑风暴法	通过面对面的群体讨论快速生成创意；属于群体创新技术	风险识别
7	德尔菲法	通过匿名和多轮次的问卷调查，汇集专家意见，达到较为客观 的结论；属于群体创新技术	风险识别
8	文档审查	对项目文档（合同、计划、项目档案等）进行评估审查	风险识别
9	根本原因分析	一项结构化的问题处理法，用以逐步找出问题的根本原因并加 以解决，而不是仅仅关注问题的表征	风险识别
10	核对表分析	根据编制的风险识别清单来进行风险识别，优点是快，缺点不 够全面	风险识别
11	假设分析	检验之前的假设是否还成立，如果不成立就是风险	风险识别
12	图解技术	使用因果图、流程图、影响图等进行风险识别	风险识别
13	SWOT分析	从内部优势、劣势；外部机会和威胁等不同角度进行分析	风险识别
14	风险清单	作为一个动态的数据库，记录和跟踪识别出的所有风险，包括 并不限于所分析人工智能系统的预期用途说明、风险评估组 织、风险评估人员、风险描述、风险识别过程、风险概率影响、 风险应对预案等信息	风险识别
15	概率与影响评估	分析风险发生可能性和后果	风险分析
16	概率影响矩阵	通过矩阵对比不同风险的严重程度	风险分析
17	案例研究法	通过选择和研究与人工智能技术及其应用相关或类似的案例， 理解和揭示人工智能技术及其应用的规律和特点，进而为治理	风险分析

		决策提供丰富的实践信息	
18	风险数据质量评估	评估风险识别和分析所依据数据的可信度	风险分析
19	风险分类	根据工作分解结构（WBS）、项目阶段、干系人或风险分解结构（RBS）对风险进行分类	风险分析
20	风险紧迫性评估	评估哪些风险需要立即应对，哪些可以延后处理	风险分析
21	数据收集与表现	包括访谈、概率分布（贝塔分布、三角分布）	风险分析
22	敏感性分析	确定对项目影响最大的风险，用龙卷风图表示	风险分析
23	预期货币价值（EMV）	把各种可能的后果和概率相乘，常和决策树一起使用	风险分析
24	蒙特卡罗模拟	通过计算机模拟进行大量随机试验，预测项目风险的可能结果和其概率分布	风险分析
25	偏差分析	根据基准分析实际绩效偏差，并采取纠正或预防措施	监督与检查
26	储备分析	考虑应急储备（如缓冲时间或储备金）	监督与检查
27	风险再评估	对现有评估进行复查，识别新风险，删除已过风险，团队成员共同完成	监督与检查
28	风险审计	评估风险管理过程的有效性，通常由外部人员来做	监督与检查
29	技术绩效衡量	比较技术成果与项目计划的符合度	监督与检查
30	状态审查会	全面的评估会议，识别和控制项目过程中面临的风险和不确定性	监督与检查
31	消极风险和威胁应对	采取规避、转移、减轻或接受的措施	风险应对
32	积极风险和机会应对	采取开拓、增强、分享或接受的策略	风险应对
33	应急应对策略	制定应急计划，在风险发生时执行	风险应对
34	风险演练	对预估的风险进行实际演练、模拟，以评估和提高组织对潜在风险和危机情况的准备和响应能力	风险应对
35	数据可用性评估工具	监测工具是否带毒及后门移除，并实现数据去毒	风险应对
36	数据质量评估工具	对数据的准确性、代表性、完整性、一致性等进行评估	风险应对
37	数据安全评估	对显著偏差和干扰样本进行监测和处理	风险应对
38	数据隐私评估	对敏感内容、隐私信息等进行评估及处理	风险应对
39	数据资产流通及保护	对数字沙箱、主动投毒等的防护	风险应对
40	模型体检工具	对模型的鲁棒性、安全性、可解释性、公平性等进行评测	风险应对

41	模型加固工具	对模型进行鲁棒训练，增加AI防火墙等	风险应对
42	深度伪造监测	包括防伪监测及防伪溯源等	风险应对
43	内容合规审核	包括对涉黄、涉恐、涉政的检测	风险应对