ICS 35.020
CCS L01

# 团　　体　　标　　准

T/ISC 0028—2023

# 人工智能模型风险管理能力成熟度模型

Capability maturity model for risk management of artificial intelligence model

（发布稿）

2023－06－12 发布　　2023－09－01 实施

中　国　互　联　网　协　会　　发　布

# 目　次

# 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。

本文件由中国互联网协会归口。

本文件主要起草单位：

中国信息通信研究院、蚂蚁科技集团股份有限公司、北京京东世纪贸易有限公司、度小满（北京）科技有限公司、同盾科技有限公司、易车公司、中国联合网络通信集团有限公司、百度在线网络技术（北京）有限公司、北京字节跳动科技有限公司、西安恩耐博人工智能科技有限公司、北京快手科技有限公司、三六零安全科技股份有限公司、腾讯云计算（北京）有限责任公司、北京京东科技有限公司、北京中科闻歌科技股份有限公司、北京小桔科技有限公司、杭州安恒技术股份有限公司、OPPO广东移动通信有限公司、北京榕树科技有限公司等。

本文件主要起草人：杨玲玲、陈杨、王阳、梁叶、应叶、董纪伟、小宇、梅亮、彭晋、林冠辰、薛峰、周杨、张天翼、杨舟、彭鸿涛、赵磊、唐佳伟、于跃、李鹏、魏芳、于城、张立彤、赵皓星、周少雄、姚一楠、王永霞、王春兴、何佳、李剑锋、梁伟、王璋盛、杨大伟、王蓬华等。

# 引　言

近年来，人工智能技术的持续发展为企业的生产运营和管理模式带来了巨大的变革，已成为带动经济增长的重要引擎。与此同时出现了新的技术风险、伦理风险等威胁和挑战，如模型数据来源、准确性和及时度不当产生的数据风险，模型缺陷造成的决策失衡和结果偏差，模型安全和监控缺失造成的失控事件，以及模型使用不当造成的伦理挑战等。

本文件针对组织在开发、实施和使用人工智能模型过程中面临的主要风险挑战，根据法律法规和行业特性，明确人工智能模型风险管理的原则、目标，厘清主体责任，建立覆盖围绕人工智能模型的风险管理能力成熟度模型，规范模型在需求分析、数据准备、模型构建、检验验证、模型部署、持续验证、模型修正、模型下线等环节中的风险管理活动，从策略制度、组织架构、资源配置、技术手段等方面提供保障，提出切实有效的管理措施，快速灵活应对模型风险，确保组织健康有序发展，推动科技向善。

对本文件中的具体事项，法律法规另有规定的，需遵照其规定执行。

# 人工智能模型风险管理能力成熟度模型

## 1 范围

本文件提供了组织在开展人工智能模型需求分析、数据准备、模型构建、检验验证、模型部署、运行监控、持续验证、模型修正、模型下线等关键活动过程中进行风险管理的建议。

本文件适用于管理、研发、供应、使用人工智能模型的相关机构。

## 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 25069—2022 信息安全技术 术语

GB/T 24353—2022 风险管理 指南（ISO 31000—2018,IDT）

## 3 术语和定义

GB/T 25069—2022和GB/T 24353—2022界定的以及下列术语和定义适用于本文件。

3.1

**模型 model**

系统、实体、现象或过程的物理、数学或其他逻辑表示。

[来源：ISO/IEC 18023-1:2006,3.1.11]

3.2

**人工智能模型 artificial intelligence model**

使用一种或多种人工智能技术和方法构建的模型（3.1）。

注：人工智能模型可以针对一组给定的人类定义的目标生成输出，例如内容、预测、建议或影响他们互动环境的决策。

3.3

**风险 risk**

不确定性对目标的影响。

注1：影响是指偏离预期，偏离可以是正面的和/或负面的，可能带来机会和威胁。

注2：目标可有不同维度和类型，可应用在不同层级。

注3：通常风险可以用风险源、潜在事件及其后果和可能性来描述。

[来源：GB/T 24353—2022,3.1]

3.4

**模型风险 model risk**

基于不正确的模型（3.1）、乱用模型输出和报告进行决策可能产生的风险（3.3）。

3.5

**能力成熟度模型 capability maturity model**

一种模型，该模型包含一个或多个学科的有效过程的基本要素，并描述了从临时的、不成熟的过程到具有改进的质量和有效性的、受到训练的、成熟的过程的进化改进路径。

[来源：ISO/IEC/IEEE 24765:2017，3.472]

## 4 人工智能模型风险管理

### 4.1 风险管理目标

a） 安全可靠：应具备与所面临的安全风险相匹配的安全能力，并采取足够的管理措施和技术手段，保障人工智能模型的完整性和鲁棒性。

b） 透明可解释：应能够解释人工智能模型如何工作以及如何达到特定的预测，或如何在决策过程中发挥作用，建立理解和信任。

c） 公平公正：应保障利益相关者的权益，促进机会均等。

d） 尊重隐私：应尊重和保护个人隐私，充分保障个人的知情权和选择权。

e） 可监督、可问责、可审计：应保障人工智能模型生命周期中利益相关方职责清晰，并具备足够的监督、控制能力，留存足够的记录以支持审计和取证活动。

### 4.2 风险管理框架

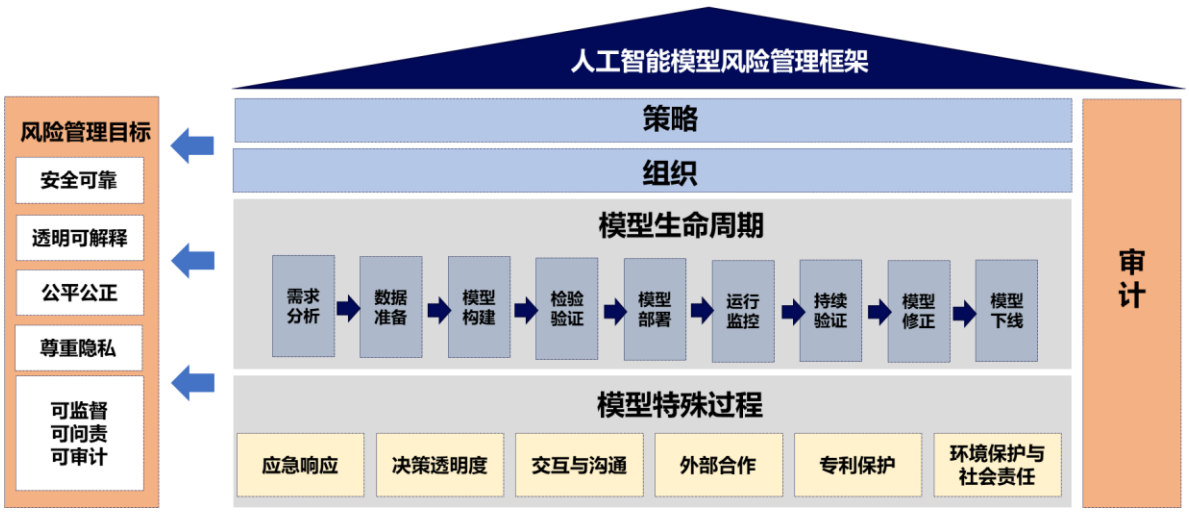


图1 人工智能模型风险管理框架

人工智能模型风险管理框架包括以下模块：

a） 策略：指组织为实现人工智能模型风险管理而采取的策略，涵盖规划、执行及控制过程，主要关注组织策略方针和管理制度。

b）组织：指管理层对人工智能模型风险管理的职责，确保组织业务战略和重大策略的一致性，评估和控制人工智能模型风险管理效率和效果，确定组织的风险偏好等。

c）模型生命周期：指对模型全生命周期进行风险管理，涵盖需求分析、数据准备、模型构建、检验验证、模型部署、运行监控、持续验证、模型修正、模型下线九个关键环节。

d）应急响应：指对开发、实施和使用人工智能模型过程中的风险事件进行及时响应、跟踪和处置，及时消减风险事件的影响，保证业务连续性。

e）决策透明度：指对人工智能模型如何工作和决策的过程进行解释。

f）交互与沟通：指与监管机构、合作方、用户等相关方建立交互沟通机制，以及采取措施保障利益相关方的权益。

g）外部合作：指在开发、实施和使用人工智能模型的过程中涉及与合作伙伴、第三方服务商等外部机构交互，对由于外部因素导致的业务中断、结果不可靠、责任不清、安全事件等风险进行管理。

h）专利保护：指对开发、实施和使用人工智能模型可能涉及的知识产权侵权风险进行管理，以及对自主知识产权进行保护。

i）环境保护与社会责任：指在开发、实施和使用人工智能模型的过程中对环境和社会产生的影响，并采取相应解决方案。

j）管理层监控和审计：指人工智能模型开发、实施和使用相关部门对现有制度流程的自我监控评估和独立审计。

## 5 能力成熟度模型

### 5.1 能力成熟度等级

| 能力成熟度等级 | 特征 | 详细说明 |
| --- | --- | --- |
| 第一级<br>（基础级） | 基本运行，各部门/业务线/项目为了经营管理正常运行，对人工智能模型风险活动进行自发管理。 | 各部门/业务线/项目为了经营管理正常运行，对该项活动进行自发管理。<br>（1）管理范围仅限于项目/部门/业务线，由各业务条线/各项目/部门各自管理。<br>（2）有基本的管理活动（部分环节、零散的），但一般依赖于约定俗成、个人经验。管理活动不一定可重复。<br>（3）管理要求可能落于纸面，但缺少规范化的、正式发布的、组织级的制度。<br>（4）管理活动以符合法律要求、避免影响业务运营的重大风险为主目标。<br>（5）一般情况下没有专岗人员，由其他职能的人员自发进行。<br>（6）基于风险事件的被动应对，仅基于个人经验和临时反应。<br>（7）没有系统工具支持管理活动。 |

| | | |
|---|---|---|
| 第二级<br>（增强级） | 安全合规，在第一级的基础上，各部门/业务线/项目为了人工智能模型风险管理活动满足外部行业监管或自身增强性的风险管理要求。 | 在第一级的基础上，各部门/业务线/项目为了该项活动满足外部行业监管或自身增强性的风险管理要求。<br>（1）管理范围仅限于项目/部门/业务线，各业务条线/各项目/部门各自管理，或通过其他现有的管理体系和机制来实施人工智能模型风险管理，并非制定了专门针对人工智能模型风险管理的机制。<br>（2）有较成熟的管理活动，但主要以符合行业监管要求和组织自身提出的增强性风险管理要求为目标。<br>（3）管理活动已在项目/部门/业务线内形成规范、流程。<br>（4）一般情况下没有专岗人员，但已明确其他职能或兼岗人员的职责。<br>（5）基于风险事件的被动应对，但具备一定的事件响应能力。<br>（6）使用简单的工具支持管理活动。 |
| 第三级<br>（优秀级） | 规范有序，在第二级的基础上，实现组织级、标准化的管理活动，管理活动覆盖关键环节，以达到行业优秀水平为目标。 | 在第二级的基础上，实现组织级、标准化的管理活动，达到行业优秀水平。<br>（1）组织范围内实现统一管理。<br>（2）建立了标准化的管理制度和流程，且管理质量可控，能够对关键的、常见的人工智能风险进行应对。<br>（3）管理活动覆盖关键环节，以达到行业优秀水平为目标。<br>（4）一般情况下设置了专岗人员。<br>（5）主动进行风险管理，可以对已知风险进行识别、评估和处置，但管理手段较为单一。<br>（6）使用系统工具支撑关键的管理活动。<br>（7）参与国家、行业相关标准制定。 |
| 第四级<br>（卓越级） | 全面高效，在第三级的基础上，实现高效的、平衡的管理，管理活动具有全面性，以达到业界最佳实践为目标，且能被充分借鉴。 | 在第三级的基础上，实现高效的、平衡的管理，达到业界领先水平。<br>（1）较第三级管理效率大幅提升，风险管理与运营效率、人工智能技术应用与相关方利益均达到较好的平衡。<br>（2）管理活动具有全面性，以达到业界最佳实践为目标，且能被充分借鉴。<br>（3）使用系统工具实现全面的管理自动化。<br>（4）建立指标体系，实现可量化的管理。<br>（5）主动进行风险管理，可以对已知风险进行识别、评估和处置，并预先建立了对可预见风险的应对措施。<br>（6）风险管理活动能够持续改进和优化。<br>（7）牵头/主导国家、行业相关标准制定，或参与国际相关标准制定。<br>（8）在相应领域部署了研发资源，探索、尝试该领域前沿技术和理念。 |

| | | |
|---|---|---|
| 第五级<br>（引领级） | 前瞻引领，在第四级的基础上，实现智能化的管理，能够应对未知风险，前沿性地探索，引领行业。 | 在第四级的基础上，实现智能化的管理，前沿性地探索，引领行业。<br>（1）使用系统工具实现自洽的智能化管理。<br>（2）实现实时、动态的管理。<br>（3）实现前瞻性的风险管理部署，可以对未知风险进行识别，并预先建立了相应的应对措施。<br>（4）将人工智能风险管理作为企业主要战略或核心竞争力，持续探索、尝试该领域前沿技术和理念并具备引领性和话语权。<br>（5）牵头/主导国际相关标准制定。 |

## 5.2 风险管理能力

### 5.2.1 策略

#### 5.2.1.1 第一级

本级能力如下：

a） 具备基本的人工智能模型风险管理活动，管理活动可依赖于约定俗成、个人经验。

b） 人工智能模型风险管理要求可落于纸面，但可能缺少规范化的、正式发布的、组织级的制度。

#### 5.2.1.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 组织高级管理层分配适当的资源以支持人工智能模型风险管理工作。

c） 管理活动能够通过其他现有的管理体系和机制来保障实施。

d） 业务团队制定人工智能模型相关管理规范和流程。

#### 5.2.1.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 组织高级管理层制定、审批、发布人工智能模型风险管理策略方针。

c） 组织高级管理层基于策略方针，制定、审批和发布人工智能模型风险管理制度。管理制度宜明确组织涉及人工智能模型开发、实施和使用相关的管理目标、职责分工、具体流程、过程文档、风险管理活动应覆盖的关键环节等内容。

d） 组织高级管理层对策略方针和管理制度进行充分宣贯，并与外部机构和利益相关方进行适当沟通。

e） 组织高级管理层定期或在组织战略、外部环境、相关技术等发生重大变化时，对策略方针和管理制度进行审阅和更新。

#### 5.2.1.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 策略方针涵盖可预见的人工智能模型风险，并且与外部机构和相关利益方的沟通宜更加充分。

c） 管理制度全面涵盖人工智能模型开发、实施和使用的各个环节。

d） 风险管理活动在系统工具的支持下实现自动化和可量化。

e） 组织宜积极收集相关信息和最佳实践，不断完善现有的策略方阵、制度和流程。

#### 5.2.1.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 组织高级管理层前瞻性地部署策略方针，实现对未知风险的管理。

c） 组织高级管理层将人工智能模型风险管理作为组织主要战略或核心竞争力，持续探索、尝试风险管理前沿技术和理念并具备引领性和话语权。

d） 管理活动实现自洽的智能化，能够对人工智能模型风险进行实时、动态的管理。

### 5.2.2 组织

#### 5.2.2.1 第一级

本级能力如下：

a） 尚未将人工智能纳入整体业务战略，尚未将人工智能相关风险纳入组织整体风险管理和管理层职责范围。

#### 5.2.2.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 组织设置人员或团队对人工智能模型的开发、实施和使用进行管理。

c） 组织根据自身业务战略和技术要求，基于国家相关法律法规及行业标准等监管的要求将与人工智能模型相关风险（如操作风险、科技风险等）纳入组织风险管理体系。

d） 组织明确了管理层以及执行层面在上述风险领域的权责范围。

e） 制定与人工智能模型相关的风险管理要求，并在组织风险管理相关职能部门实际工作中有效落实相关制度和流程（如数据安全部门、内控合规部门等）。

#### 5.2.2.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 组织将人工智能作为组织整体业务发展战略的一部分。

c） 组织明确了管理层，人工智能开发、实施和使用相关部门和团队在人工智能模型风险管理方面的职责职能与授权。从组织架构上保障人工智能模型风险管理活动具备独立性，确保适当的人员对人工智能模型具有要求解释和拒绝使用的权力。

d） 管理层需对人工智能模型风险管理框架进行审批。

e） 组织建立与人工智能模型风险管理框架配套的风险管理机制，包括风险识别、风险评估、风险处置、监控与审计，形成完整的风险管理闭环。

#### 5.2.2.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 组织建立专门的治理委员会并定期实现向董事会和组织高级管理层汇报和披露人工智能发展与风险管理情况。如组织内部已经形成了如风险管理委员会、信息科技委员会等机构，且将与该人工智能模型相关的风险管理纳入到组织整体风险管理领域并开展专门的业务决策、风险评估与处置管理，则也可以满足该要求。

c） 通过量化手段实现风险计量，实现监控与审计的部分自动化。

d） 建立人工智能相关人才激励政策，并将人工智能风险管理与考核机制结合。

#### 5.2.2.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 组织实现对组织不同层面及不同人工智能模型面对的风险的实时监控，以帮助组织高级管理层对组织人工智能模型风险进行及时的了解。

c） 组织实现风险应对决策的自动化，以提高改进计划制定的效率。

d） 组织借助系统支持改进计划的执行和跟踪，以达到更有效的持续管理改善。

e） 组织实现对未知风险的识别，并据此评估现有人工智能模型风险管理策略方针和管理制度的适用性，制定相应的管理策略调整和风险应对方案，以确保人工智能模型风险管理具有前瞻性。

### 5.2.3 模型生命周期

#### 5.2.3.1 需求分析

##### 5.2.3.1.1 第一级

本级能力如下：

a） 相关人员和人员可基于经验开展基本的需求分析活动。

b） 需求分析活动明确了模型构建的业务背景、业务目标和应用场景，一般宜涵盖问题输入、目标拆解、资源投入等相关内容，

c） 具备模型需求提出、记录、流转、跟踪的渠道。

##### 5.2.3.1.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 制定模型需求管理流程规范。

c） 涵盖行业监管要求、组织风险管理与安全相关要求。

d） 建立需求评审机制，需求评审应包括业务可行性和技术可行性、模型应用目标和预期效果与组织战略的符合性、潜在的业务价值。需求评审相关方宜至少包括业务人员、模型开发人员、安全人员、合规人员、风险管理人员。
e） 根据需求分析内容形成需求说明文档，并经过需求评审相关方确认。

#### 5.2.3.1.3 第三级

本级能力如下：
a） 前两级的全部内容。
b） 制定模型需求管理制度，一般宜涵盖需求评审、需求变更、紧急需求等关键的管理活动及资源保障。
c） 开展模型风险评估，通常宜关注模型业务应用场景风险以及威胁和脆弱性、与人工智能风险管理原则的符合性、与风险管理策略方针的符合性、安全控制措施的有效性、风险事件应急响应机制、对利益相关方可能产生的不利影响等。在模型存在的高风险缺陷或问题的情况下，应制定风险缓释措施降低风险。
d） 需求说明文档具体描述人工智能模型业务目标，模型构建和应用成功与否的衡量标准，模型构建及应用所需的资源和时间，模型风险评估。
e） 具备自动化工具支持模型需求管理。

#### 5.2.3.1.4 第四级

本级能力如下：
a） 前三级的全部内容。
b） 针对需求的执行情况进行周期性评估，并检查需求执行效果。
c） 需求评估具备可量化的指标。为进一步提升服务效能、质量、安全等目的，宜建设需求执行验证平台或工具。

#### 5.2.3.1.5 第五级

本级能力如下：
a） 前四级的全部内容。
b） 能够对模型需求分析过程进行自适应调整，根据业务需求、监管环境、安全态势、技术发展等情况对需求分析的范围、目标、内容进行自动调整。
c） 针对需求分析变化能够自动调整资源变更的流程，并完成对于执行效果的完整性检查。

### 5.2.3.2 数据准备

#### 5.2.3.2.1 第一级

本级能力如下：
a） 相关部门和人员可基于经验进行基本的数据准备活动，产出模型构建所需的数据集。

#### 5.2.3.2.2 第二级

本级能力如下：
a） 前一级的全部内容。

b） 制定数据采集、存储、使用、传输、销毁等生命周期的安全要求，且应符合组织总体的安全合规要求。
c） 建立数据开发管理流程，规范数据准备处理操作，一般宜包括数据标记、标注、清洗和聚合等。
d） 数据开发人员实际有效地落实数据开发相关管理要求和流程。
e） 对数据集的数据质量进行管理，保障数据的准确性、完整性、有效性、时效性。
f） 具备良好的团队协作机制，参与数据准备活动的各职能人员应分工明确、沟通顺畅。

#### 5.2.3.2.3 第三级

本级能力如下：
a） 前两级的全部内容。
b） 建立数据开发管理制度，规范数据处理流程，管控数据处理各环节相关风险。
c） 针对模型的预期用途进行专有的数据开发，考虑特定的应用环境所特有的特征或要素。
d） 记录数据准备操作过程，包括数据获取时间、训练数据来源、训练数据量、数据存储介质标识、采样方法。
e） 利用角色、流程、权限设置对训练数据集和测试数据集进行管控，保证数据集的安全性、完整性、一致性及隐私保护。
**注：**涉及生物识别、医疗健康、金融账户、行踪轨迹等敏感个人信息的数据处理活动应符合《中华人民共和国个人信息保护法》的要求。
f） 确保数据开发环境中使用的数据与生产环境的一致性。
g） 具备数据开发工具，通常宜支持数据集成、数据清洗和数据加工等操作。
h） 保留和备份数据准备过程中产生的关键日志信息，日志内容和保存时长应满足组织审计需要。

#### 5.2.3.2.4 第四级

本级能力如下：
a） 前三级的全部内容。
b） 对数据准备活动生成的数据集进行全面评估，避免存在数据污染、数据投毒攻击、数据偏差、用户歧视、用户隐私侵犯等情况，针对评估所识别出的风险应及时制定相应的解决方案。
c） 宜采用数据安全领域先进技术，在保证数据安全及用户权益的前提下，多方位引进模型内外 部数据，扩展数据维度。

#### 5.2.3.2.5 第五级

本级能力如下：
a） 前四级的全部内容。
b） 可借助自动化检测及管理工具，智能化识别数据准备活动中的现有风险，对可能发生的未知风险进行预测，制定有效的风险应对方案，实现数据准备活动风险管理的前瞻性部署。

### 5.2.3.3 模型构建

#### 5.2.3.3.1 第一级

本级能力如下：

a） 相关部门和人员可基于经验或团队约定俗成的方式开展算法选择、模型开发训练、测试等模型构建活动。

#### 5.2.3.3.2 第二级

本级能力如下：

a） 前两级的全部内容。

b） 制定模型构建管理流程，明确代码安全等相关要求。

c） 模型开发人员实际有效地落实模型构建相关管理要求和流程。

d） 模型开发人员对模型的有效性、准确性、鲁棒性进行自测，确保模型满足业务需求。

e） 模型开发人员通过对模型参数的合理适当性进行复核、避免选择偶发性参数等措施保证模型构建的科学合理性。

f） 需记录模型构建过程，一般宜包括建模脚本、建模过程中的软硬件环境、模型训练方式、特征选择和决策结果、建模人员、建模时间、版本迭代等。

#### 5.2.3.3.3 第三级

本级能力如下：

a） 前三级的全部内容。

b） 制定模型构建管理制度。

c） 制定模型构建方案。模型开发人员应尝试使用不同的算法以找到最适合的模型方案，并了解所选算法的局限性，测试模型相关假设（如有）并提供合理的证明，进行不同时间样本的模型性能比较以验证模型稳定性，提供和基准模型(如有)性能的比较以说明新模型的优越性。

d） 对目标函数进行说明，目标函数设计上不应存在针对特殊群体的偏见歧视。

e） 对模型重现能力进行评估，确保模型决策结果在相同场景下的一致性。

f） 对深度学习模型使用过程中产生的相关计算数据包括输出向量、模型参数、模型梯度等可能会泄露训练数据的敏感信息或者模型自身的属性参数，制定保护策略，保护策略宜采取限制恶意访问次数、引入随机性、添加模型水印等措施。

g） 使用代码扫描工具和人工代码检查方式对代码进行评审，发现安全缺陷并修复。

h） 形成模型开发文档，包括算法选择、模型训练、参数选择、模型测试等模型开发过程关键内容。

i） 建立模型评审机制，成立模型评审委员会或组织合适的相关方负责模型评审工作。评审人员应独立于模型开发人员，评估模型效果、模型性能等影响模型应用的关键内容。

j） 建立模型构建的环境或平台，并设置完善的访问控制、安全控制等措施。

k） 保留和备份模型构建过程中产生的关键日志信息，日志内容和保存时长应满足组织审计需要。

#### 5.2.3.3.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 建立模型构建安全规范，明确模型构建过程中应遵循的安全原则和安全要求，一般宜包括模型威胁建模和攻击面分析，模型开源风险，模型安全漏洞等。

c） 模型构建的环境或平台具备适用于多种用户角色的建模工具，一般宜包括给业务人员提供图形化的建模工具、给数据科学家提供专业的交互式建模工具等，建模工具提供的算法库能够支持多种模型的构建需求，包括有监督、无监督，以及主流的神经网络、深度学习、自动化建模等。
d） 模型构建的环境或平台具备模型文件校验等技术能力，对模型文件格式、大小、参数范围、网络拓扑、节点名称、数据维度等关键信息进行检测校验，避免加载恶意模型文件。
e） 保障模型可靠性和可重现性，若发生不可避免的偶发性不准确预测，应当能计算出错误发生的概率。
f） 保障模型的差分隐私性，确保无法根据模型结果反向计算出用户的敏感特征。
g） 针对模型潜在的恶意攻击风险，采用对抗训练、对抗样本检测等防御手段提升模型鲁棒性和安全性。

#### 5.2.3.3.5 第五级

本级能力如下：
a） 前四级的全部内容。
b） 实时评估、监控模型构建过程的可预见风险，并自动化地实施风险处置措施。
c） 对模型构建过程可能存在的未知风险设置前瞻性的风险预防措施。

### 5.2.3.4 检验验证

#### 5.2.3.4.1 第一级

本级能力如下：
a） 未建立模型检验验证管理机制，未设置检验验证人员，相关部门和人员可基于经验或团队约定俗成的方式开展检验验证活动。
b） 在检验验证发现模型有重大缺陷且未解决之前，需要在非常严格的限制条件下使用模型或应拒绝使用该模型。

#### 5.2.3.4.2 第二级

本级能力如下：
a） 前两级的全部内容。
b） 制定模型检验验证管理流程，通常宜包括检验验证的标准执行流程、检验验证目标和内容、数据安全管理要求等，确保模型与行业监管要求及业务应用方面的符合程度。
c） 设置负责模型检验验证的人员和岗位。
d） 确保人工智能模型的相关特征及特征工程过程与风险策略相吻合，一般宜包括但不限于模型输入的正确转换、特征选择的标准、特征对应的业务逻辑等。
e） 对数据输入及数据处理结果进行风险评估,避免数据采集及处理结果中包含可恢复或者涉及安全隐私的敏感数据。
f） 进行模型结果有效性验证和分析，即对模型输出与相应的实际结果进行比较。
g） 在相同场景下采取不同的数据集来对模型进行多次校验，并对在相同条件下出现的差异化结果进行进一步分析。
h） 对检验验证过程中发现的问题制定解决方案并持续跟进。

i） 对模型检验验证过程进行记录，通常宜包括验证方式、验证人员、验证结果等。

#### 5.2.3.4.3 第三级

本级能力如下：

a） 前三级的全部内容。
b） 制定模型检验验证管理制度。
c） 设置专职负责模型检验验证的人员和岗位，检验验证人员岗位须具备适当的权限和独立性。
d） 模型检验验证范围宜包括模型的所有组件，即输入、处理、输出、报告等。
e） 模型检验验证内容宜包括数据稳定性、模型的鲁棒性和模型效果检验等。
f） 对于模型的合理性及风险可控程度进行全面的核查，通常宜包括使用数据项的适当性、合理性、权重比例的科学性等问题。
g） 采取必要措施以保证模型决策结果公平公正，确定公平评判标准，一般宜包含人群均等、机会均等、几率均等、人群无关性等标准，并对其进行公平性测试，必要时进行修正。
h） 针对验证过程中发现的缺陷或风险，分析对应缺陷引发的原因、导致的问题、是否有解决方案及最大化风险敞口的可能。记录留档并在缺陷无法解决的情况下，并按组织风险汇报流程升级至适当的管理层进行风险决策。
i） 制定检验验证报告。

#### 5.2.3.4.4 第四级

本级能力如下：

a） 前三级的全部内容。
b） 采取人工及系统评估双重交叉验证的方式进行检验验证，通过多人复评多个系统交叉组合验证的方式以核验模型的完整性和准确性。
c） 建立检验验证平台或工具，宜采用模拟数据窃取、成员推理攻击、数据逆向还原等方法模拟对模型数据的窃取行为。
d） 实现模型验证关键环节可量化。
e） 在模型结构或技术发生重大变化，或模型进行重新开发的情况下，模型重新实施之前应接受恰当范围和严格程度的验证。
f） 在适用的情况下，可选择在模型开发和验证中采用回测或敏感性分析来完成自洽检验，以检查输入和参数值的微小变化对模型输出的影响，确保结果在预期范围内。
g） 在数据特征交互作用复杂且不直观的情况下，通过同时改变多个输入来发现意外的交互作用。在缺乏数据或数据质量不稳定的情况，需更加注意模型结果的局限性。在使用模型结果进行决策时，宜将这些局限性充分告知管理层及利益相关方。
h） 通过对抗攻击等方式验证关键模型的鲁棒性，以确知关键模型性能保持稳定所允许的扰动范围。
i） 检测模型在特征各个分段上的性能差异，对于性能偏差的特征分段应采取该特征分布持续监控等措施。

#### 5.2.3.4.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 支持智能化地评估检验验证模型关键假设和变量选择，并分析其对模型输出地影响。
c） 针对未知风险应支持智能化地做出最优决策，并对于模型调整及验证环节进行智能化地评估、检查。
d） 在模型考虑新的数据或技术，或由于性能下降进行定期调整的情况下，可采用平行结果分析对模型调整进行重要分析。原始模型和调整后模型的预测都要与实际的结果相比较，如存在调整后的模型没有优于原始模型，需对于模型进行额外的更改或重新设计。

#### 5.2.3.5 模型部署

##### 5.2.3.5.1 第一级

本级能力如下：
a） 相关部门和人员可依据通用部署机制或基于经验进行模型部署。
b） 制定模型部署回退方案。

##### 5.2.3.5.2 第二级

本级能力如下：
a） 前一级的全部内容。
b） 制定模型部署管理流程。
c） 制定模型部署实施计划和方案，确定部署策略和步骤。
d） 模型部署活动一般宜涵盖生产数据接入、模型校验、上线审核和后评估等。
e） 对模型部署过程进行记录，一般宜包括模型部署的操作人员、部署环境、部署步骤、部署时间和部署结果等。
f） 模型部署前需开展模型上线评审，评审应包括适当的利益相关方。
g） 根据模型部署影响范围，提前将模型部署可能造成的影响告知相关方。

##### 5.2.3.5.3 第三级

本级能力如下：
a） 前两级的全部内容。
b） 制定模型部署管理制度。
c） 对准备部署的模型文件进行完整性校验。
d） 通过模型灰度发布对模型的业务效果进行评价，确保模型性能符合业务需求。
e) 模型部署前需开展评审，评审内容应包含对模型需求分析报告、模型评审报告、模型验证报告、模型技术文档等，并经过利益相关方确认。针对评审发现的高风险缺陷或问题，应整改完成后才能上线。
f） 模型部署模型上线需经过恰当的审批和授权，未完成审批的模型不应部署上线。
g） 具备修改模型保存路径和保存方式等内容的能力。
h） 建立模型部署和运行的安全环境或平台，并设置完善的安全检测和访问控制等措施。
i） 保留和备份模型部署过程中产生的关键日志信息，日志内容和保存时长应满足组织审计需要。

##### 5.2.3.5.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 模型部署的环境或平台能够支持自动化、精细化、多样化的模型部署方式。

c） 模型部署的环境或平台能够实时监控模型部署过程产生的风险并可视化展示。

d） 通过系统自动防御模型部署过程中的操作风险。

e） 模型的存储数据库应能够支持审计和安全风险的自动告警。

#### 5.2.3.5.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 实现模型动态迭代部署，直到生成最合适结果的模型为止，将该模型及其结果纳入应用程序中并能够智能地应对部署过程风险，提供预测、决策、解决问题和触发操作。

c） 具备应对各种安全风险问题的能力，保证组织部署的模型、配套的训练数据以及权重参数不会被攻击者影响而改变结果或泄漏数据。

### 5.2.3.6 运行监控

#### 5.2.3.6.1 第一级

本级能力如下：

a） 相关部门和人员可基于经验或通用管理机制对模型进行评估和监控。

b） 模型监控和评价主要基于线下方式进行，定期发布模型监控评价信息。

c） 能够对模型运行过程中的重大风险、决策异常进行识别和风险提示。

#### 5.2.3.6.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 制定模型监控管理流程。

c） 明确的模型应用的业务目标和应用风险，建立模型运行监控体系和评价方法。不同职能人员根据各自的目标和职责设置模型运行监控和评价指标。

d） 模型监控和评价维度包含数据监控和模型运行监控。

**注：**数据监控主要监控模型所需数据的提取、加工和数据分布的变化；模型运行监控主要监控模型运行在正常与否、模型运行的资源消耗、模型运行响应的性能等。

e） 定期识别和量化模型效果或风险，能够对由模型效果降低带来的风险进行预警。

f） 定期出具模型监控报告，报告涵盖的信息应足够支持做出模型修正、模型下线等决策。

g） 建立适当的自动化工具或平台对模型的运行情况进行监控，并自动化触发阈值告警。

h） 定期开展监控策略的合理性评估工作，按照需要及时调整运行监控策略。

#### 5.2.3.6.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 制定模型监控管理制度，明确相关方岗位职责，模型监控、模型使用、模型开发、应急处置等相关部门和人员应形成完善的风险处置协同机制。
c） 根据模型决策重要性和风险影响制定差异化的监控和响应策略。高风险模型决策结果应用前经过人工审核，中风险模型决策结果突破风险阈值后由人工进行处置。
d） 模型监控工具或平台能够通过对监控指标的灵活配置支持不同模型类型的不同性能监控，如模型稳定性、区分性能、准确性、同质异质性等指标。
注：PSI、IV、WOE、KS、ROC、AUC、GINI、卡方检验、F 检验、T 检验、秩和检验等常见模型指标监控。
e） 对相关方反馈进行持续监测。
f） 模型监控工具或平台支持模型监控处置策略，通常宜包括熔断、降级、隔离、标记、下线、模型（自动）更新等。模型监控工具或平台能够及时输出模型监测评价报告。
g） 模型监控工具或平台支持系统自动处置，也支持人工进行部分或全部干预和处置。
h） 持续地进行模型监控策略的评估和改进，优化模型风险管理方式。
i） 保留和备份运行监控过程中产生的关键日志信息，日志内容和保存时长应满足组织审计需要。

#### 5.2.3.6.4 第四级

本级能力如下：
a） 前三级的全部内容。
b） 模型使用人员需了解模型业务价值的监控结果并做出适当的决策。模型管理人员需了解和掌握模型资产的整体价值和收益、风险等信息。
c） 构建适当的工具或平台对模型业务成效进行监控。
d） 形成模型运行监控的指标体系，基于不同业务和模型应用场景采用不同的监控指标进行监测和评估。
e） 基于模型验证的结果，对于模型敏感特征进行监控，并制定敏感特征分布发生显著变化时的相关处置措施。
f） 基于运行监控结果做出数据资产运营的决策，一般宜包括持续维护建设、下线等，并据此调整相应的资源投入。

#### 5.2.3.6.5 第五级

本级能力如下：
a） 前四级的全部内容。
b） 通过模型监控和评价信息实现自动化的模型审计。
c） 实时监控模型相关的监管环境、风险态势、外部舆情、市场变动、合作方经营、用户行为和社会价值观的变化，智能地识别潜在风险并执行风险处置。

### 5.2.3.7 持续验证

#### 5.2.3.7.1 第一级

本级能力如下：
a） 相关人员可基于经验或团队约定俗成的方式开展持续验证活动。持续验证一般宜包括确认以前的验证活动、对以前的验证活动进行更新、或者要求额外的验证活动。
b） 模型持续验证设定一定的频次或验证周期。

c） 在持续验证发现模型有重大缺陷且未解决之前，须在非常严格的限制条件下使用模型或应进行模型下线。

#### 5.2.3.7.2 第二级

本级能力如下：

a） 前一级的全部内容。
b） 制定模型持续验证管理流程，一般宜包括持续验证的标准执行流程，持续验证目标和内容等。
c） 设置负责模型持续验证的人员和岗位。
d） 模型持续验证周期符合行业监管要求和组织风险管理要求。
e） 持续验证需结合模型运行监控内容进行验证检查。
f） 对持续验证过程进行记录。

#### 5.2.3.7.3 第三级

本级能力如下：

a） 前两级的全部内容。
b） 制定模型持续验证管理制度。
c） 持续验证评估是否需要根据产品、风险敞口、活动、客户或市场状况的变化，决定对模型是否进行阶段性调整、重新开发或更换，以及验证超出模型原始范围的任何扩展是否有效。
d） 持续验证的频次和周期保证能够应对及时性风险。
e） 持续验证的范围包括模型表现的合理性、安全性、鲁棒性、一致性、稳定性、公平性，是否持续使用最新数据，以及前期阶段所发现的模型局限性。
f） 在监管政策环境及社会普适性道德价值观均未发生变化的情况下，能够通过外部舆情监测等形式对外部信息进行收集整理，并采取相应的模型迭代调整策略。
g） 保留和备份持续验证过程中产生的关键日志信息，日志内容和保存时长应满足组织审计需要。

#### 5.2.3.7.4 第四级

本级能力如下：

a） 前三级的全部内容。
b） 评估在不可重复的情况下模型出现异常的风险，以及采取相应措施来识别和处理模型异常风险。
c） 在适用的情况下，采用回测或敏感性分析进行评估，以检查输入和参数值的微小变化对模型输出的影响。

#### 5.2.3.7.5 第五级

本级能力如下：

a） 前四级的全部内容。
b） 基于持续验证的量化评估，设立明确的规划来持续验证模型效果，并满足未来可能存在的新需求和新变化。
c） 能够自动执行模型验证，且模型在持续验证中能够自动化、智能化的进行调整和优化。

#### 5.2.3.8 模型修正

##### 5.2.3.8.1 第一级

本级能力如下：

a） 相关部门和人员可基于经验或团队约定俗成的方式开展模型修正活动。

b） 在模型部署或运行监控过程中发现模型存在异常，及时开模型修正活动。

c） 模型修正后同步开展检验验证及运行监控活动。

##### 5.2.3.8.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 制定模型修正管理流程。

c） 设置负责模型修正的人员和岗位。

d） 定期进行模型修正，保证模型的准确性。

e） 跟踪国内外监管要求的变化并及时进行模型修正，避免模型因未及时采取适当的调整而引发的侵害用户权益、违反监管要求及社会价值观等方面的情况。

f） 在模型修正过程中，避免引入新的安全合规问题。

##### 5.2.3.8.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 制定模型修正管理制度。

c） 设置明确的模型修正的触发条件。在组织业务战略和目标、风险策略或业务价值发生变化的情况下，应开展模型修正。

d） 确保部署的模型满足随着时间推移的需求变换，根据更新的训练数据集更新模型。

e） 保留模型修正前后的相关文档。

f） 对模型修正过程中的模型版本进行管理，记录修正过程中应用的模型参数、入模数据以及不同模型版本的效果差异。

g） 对模型越控及进行重点分析和记录，分析模型越控原因，评估越控影响，采取适当的应对措施。

注：模型越控指模型输出结果基于模型使用者的专家判断被忽略、更改或被反转的情况。

h） 保留和备份模型修正过程中产生的关键日志信息，日志内容和保存时长应满足组织审计需要。

##### 5.2.3.8.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 具备自动化工具或平台应支持模型的重新训练、重新上线；支持模型自动修正，包括自查、自重建和自刷新；支持模型自动修正后的检验验证、自动部署和自动监控。

c） 通过系统自动防御模型变更过程中的操作风险。

d） 对模型修正进行定期审查。

##### 5.2.3.8.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 能够智能化地对模型修正进行执行和改进。

c） 持续改进模型修正流程，以应对未来可能发生的未知风险。

#### 5.2.3.9 模型下线

##### 5.2.3.9.1 第一级

本级能力如下：

a） 相关部门和人员可依据通用系统下线管理机制或基于经验开展模型下线活动。

b） 在模型新版本迭代或目前线上版本存在重大问题无法使用时，执行模型下线处置。

c） 模型下线前评估其影响范围。

##### 5.2.3.9.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 制定模型下线管理流程，一般宜涵盖模型下线影响分析、下线确认、下线后评价等。

c） 模型下线前制定模型下线计划和方案，确定下线策略和步骤。

d） 对模型下线过程进行记录，通常宜包括模型下线的操作人员、环境、步骤、时间和结果等。

e） 模型下线后将结果告知相关方。

##### 5.2.3.9.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 制定模型下线管理制度。

c） 模型下线评估流程明确模型下线条件、模型下线时机、模型下线需求确认、模型下线影响评估、模型回滚和下线后评估，并形成相应的报告文档。

d） 根据模型下线的影响程度采用适当的下线方式，对业务影响较大的模型应通过模型灰度发布下线，确保模型下线结果符合预期。

e） 模型下线后进行验证，确保下线模型与验证结果一致。并持续监测和评估对业务的影响，确保模型下线符合预期。

f） 定期排查、下线不符合业务预期的、无效的或存在重大问题的模型。

g） 对下线的模型进行归档，明确模型归档的部门和职责，明确对已归档模型及相关文档的安全管理要求。

##### 5.2.3.9.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 通过自动化的方式管理模型下线，能够具备自动化、精细化、多样化的模型下线方式。

c） 实时监控模型下线过程产生的风险并能够可视化展示，记录模型下线过程中的关键信息并保存满足审计需要。

d） 通过模型下游血缘穿透，自动将下线信息通知到下游使用方，根据下线反馈推进下线流程。

#### 5.2.3.9.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 实现智能化地识别模型下线风险，进行模型下线决策，并自动化地下线模型并进行后续归档。

### 5.2.4 应急响应

#### 5.2.4.1 第一级

本级能力如下：

a） 相关部门和人员可基于个人经验和临时反应对模型风险事件进行应急处理。可能建立了通用的应急响应机制，但未建立人工智能模型风险事件应急响应机制。

b） 能够执行模型下线等应急处理措施。

#### 5.2.4.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 在相关制度中设置有模型相关应急响应条款。

c） 相关部门约定基本的应急协同机制。若处置动作对组织或用户存在重大影响，需明确决策机制。

d） 未设置人工智能模型应急响应专人专岗，但相关人员具备模型风险识别和应急处置的基本知识和能力。

e） 明确各相关方在模型应急处理方面的责任。

f） 能够执行精细化的应急处理措施，一般宜包括模型降级、隔离、标记等。

g） 定期开展培训和应急演练。

#### 5.2.4.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 制定针对人工智能模型的应急响应制度。

c） 明确模型应急响应专人专岗，负责统筹模型应急响应工作。

d） 梳理组织中涉及人工智能模型的业务场景，明确相关的业务指标、模型性能指标和对应责任人。根据梳理的人工智能模型相关业务指标和性能指标，应建立模型应急事件分级标准。

e） 根据梳理的模型应急事件分级标准，设置适当的决策机制。

f） 根据模型应急事件分级标准细化应急响应流程，如明确响应时效、同步时效、止血时效等。

g） 梳理关键的、常见的人工智能模型应急场景和事件，制定应急预案，宜提前设置人工介入流程及熔断机制、提前准备常规解释说明文件等。

h） 设置应急复盘机制，对模型生命周期的各个环节进行回溯，确认模型应急事件根本原因和整改方案。

i） 对模型应急响应相关的人员进行针对性的培训，提升应急处置能力。

#### 5.2.4.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 建立人工智能模型应急响应组，至少包含负责应急管理组、数据科学家组、数据质量组以及技术团队相关人员。

c） 通过红蓝攻防、白帽测试、模型入参干扰等方式进行专项或实战演练。

d） 宜在复盘明确人工智能模型应急事件根本原因的基础上进行定责追责。

e） 应急响应工具或平台应实现模型应急监控、定级、预警、处置、复盘、预案演练等功能。

f） 对人工智能模型应急响应组人员进行针对性培训和认证，并对相关领域的全员进行应急响应宣传，提升应急处置能力。

g） 在既定的风险定级的和风险处置的链路里，实现历史风险的自动化分析能力，并定期对定级方法、处置效能进行迭代。

#### 5.2.4.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 基于智能化的风险分类、定级能力，实现自动化的模型应急人员管理，缩短应急成员协调的时限，优化应急响应时效。

c） 实现风险智能分类、预估风险影响结果、分析风险发生原因，并自动与模型应急预案体系互相关联，基于智能监控的结果实现风险应急自动化。

d） 具备智能化的人工智能模型应急响应工具或平台，实现模型异常自动诊断、预案自动调用、处置进展自动同步、决策自动升级。

e） 牵头建立人工智能模型应急响应标准体系，推动在行业内复用。

### 5.2.5 决策透明度

#### 5.2.5.1 第一级

本级能力如下：

a） 对于提取的每一单一维度特征实现自我解释。能够给出每一单一特征的具体物理含义。

#### 5.2.5.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 将模型的输入特征与模型的输出的结果建立联系。能够实现对于给定的模型输出，关联到与该输出最为相关的输入特征中。

#### 5.2.5.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 在建立模型输出与输入特征的关联基础上，解释结果应捕捉到特征与特征之间的联系，并展示有关联的特征如何共同作用于模型输出结果。

#### 5.2.5.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 能够实现给定具体模型输出，将模型输出的结果与一部分主要训练数据相关联。除利用输入特征作为解释外，解释结果应针对每一个输出的结果给出对应的训练数据。

#### 5.2.5.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 解释模型在决策过程中的因果关系逻辑，该因果关系逻辑应能够使模型使用者理解（符合使用人的认知逻辑）。

### 5.2.6 交互与沟通

#### 5.2.6.1 第一级

本级能力如下：

a） 未在任何部门中设立负责人工智能模型交互与沟通相关的人员和团队，可由业务部门临时承担交互与沟通工作，基于经验应对监管机构或用户等利益相关方偶发性的交互与沟通需求。

#### 5.2.6.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 建立与外部利益相关方（如监管部门、合作方、用户等）的沟通渠道，宜披露联系人、投诉电话或邮箱、信息反馈功能等。

#### 5.2.6.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 设置具体人员负责人工智能交互与沟通工作。

c） 制定标准化的说明文件模板对人工智能模型进行信息披露，内容应包括但不限于：

1） 人工智能模型的责任主体和联系方式；

2） 人工智能模型的预期用途；

3） 可能对相关方构成的危害或产生的不利影响；

4） 为了确保人工智能模型的公平性、安全性和可靠性，以及数据的安全性，责任主体所采取的措施。

d） 以非技术的通俗语言进行说明。

#### 5.2.6.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 设置专职负责与人工智能模型交互与沟通的人员和岗位，专职人员具有相关的专业能力。

c） 建立统一的管理流程，规范与外部利益相关方的交互与沟通，明确组织内部沟通协作机制。

d） 在第三级 c)的基础上包括但不限于：

1) 模型所涉及的人工智能技术的介绍，在包含“黑箱”技术的情况下宜补充说明；

2) 人工智能模型所使用的数据以及具体使用方式；

3) 人工智能模型的决策机制和逻辑；

4) 人工智能模型决策面对的对象（是否涉及特殊人群）。

e） 说明文件应简明完整、准确清晰、便于获取。

#### 5.2.6.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 设置专门的人工智能信任管理部门或岗位，负责组织整体的人工智能模型的交互与沟通，并与外部相关方对接。

c） 依据不同的人工智能模型的性质和风险等级，以及相关方的信息获取需求、获取目的和背景，采用适当的交互方式：

1) 当相关方需要获取人工智能系统或模型相关信息时，能够提供说明文件；

2) 为用户提供选择退出或服务中止的渠道，渠道应便捷易触达；

3) 当利益相关方对人工智能决策机制产生质疑时，宜采用人工讲解、演示、可视化媒体、图形展示等手段进行说明。

d） 重视持续改进机制，通过积极收集外部及不同渠道的信息，不断地完善现有与人工智能交互与沟通机制。

### 5.2.7 外部合作

#### 5.2.7.1 第一级

本级能力如下：

a） 相关部门和人员可基于个人经验评估外部合作可能引入的相关风险，并进行风险应对。

b） 与外部合作机构通过合同方式，明确双方关键的风险责任与义务。

#### 5.2.7.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 制定模型外部合作管理制度和流程，并由相关职能部门在人工智能模型外部合作中落实要求。

c） 对外部合作方开展尽职调查和定期的风险评估。

d） 要求外部合作方就责任范围内的相关内容出具评估报告，证明其在业务需求、人工智能模型风险管理、行业监管要求及组织要求方面的符合程度。

e） 针对重要的人工智能模型外部合作，制定业务中断、信息泄露等方面的应急预案，并定期进行双方联动演练。
f） 定期评价模型相关的外部合作。

#### 5.2.7.3 第三级

本级能力如下：
a） 前两级的全部内容。
b） 建立针对人工智能模型外部合作的评估机制，并设置准入条件。
c） 在部署前对涉及外部合作的人工智能模型进行风险评估，或要求外部合作方提供由独立的第三方机构出具的评估报告。
d） 与外部合作方通过合同等形式明确双方应实施的安全措施，建立在发生风险事件、合作中止或提前结束、用户权益纠纷等情况时双方的沟通渠道和响应机制。

#### 5.2.7.4 第四级

本级能力如下：
a） 前三级的全部内容。
b） 设立岗位和人员对外部合作方进行定期的风险评估，确保外部合作方的组织架构、风险管理机制、模型开发流程、质量管理体系等方面符合组织的风险管理要求。
c） 对外部合作方提供的人工智能模型进行实时监控。

#### 5.2.7.5 第五级

本级能力如下：
a） 前四级的全部内容。
b） 对外部合作方的财务、运营、安全管理、人力资源、技术发展，以及所处的市场、监管和生态环境进行持续监控，关注变化可能给外部合作方带来的影响，及时调整合作处理方案。

### 5.2.8 专利保护

#### 5.2.8.1 第一级

本级能力如下：
a） 在相关制度中设置有知识产权风险管理条款，确保模型的开发、实施和使用不侵犯他人知识产权，进而引发诉讼及公司声誉风险。

#### 5.2.8.2 第二级

本级能力如下：
a） 前一级的全部内容。
b） 相关人员具备识别人工智能模型知识产权的知识和能力。
c） 采取侵权风险宣贯、外部技术和物料法规审核、盗版软件监控等措施，避免或降低人工智能模型开发、实施和使用相关人员在工作中使用的软件、技术、物料侵犯他人知识产权的风险。

d） 在进行人工智能模型委托开发或合作开发时，签订书面合同，约定知识产权权属、许可及利益分配、后续改进的权属和使用。

e） 及时发现和监控知识产权被侵犯的情况，适时运用行政和司法途径保护知识产权。

#### 5.2.8.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 对知识产权风险进行识别和评测，并采取相应风险控制措施。

c） 定期监控人工智能模型涉及他人知识产权的状况，分析可能发生的纠纷及其对组织的损害程度，提出防范预案。

d） 通过专利、软著、商标申请等方式，对核心人工智能模型开发成果的知识产权进行保护。

e） 宣明确员工知识产权创造、保护和运用的奖励和报酬，明确员工造成知识产权损失的责任。

#### 5.2.8.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 通过日常检索预警方式监控人工智能模型可能涉及他人知识产权的状况。

c） 通过专利、软著、商标申请等方式，对组织内全部人工智能模型开发成果的知识产权，进行全面布局和保护。

d） 通过劳动合同、竞业协议、保密协议等方式对员工进行管理，约定模型知识产权权属、保密条款。

e） 对新入职员工进行适当的知识产权背景调查，以避免存在侵犯他人知识产权的情形。对离职的员工进行相应的知识产权保护事项提醒。

#### 5.2.8.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 通过知识产权分析及市场调研，明确人工智能模型相关潜在的合作伙伴和竞争对手，并预测行业发展方向和技术热点。

c） 利用产生的人工智能模型知识产权成果，实现知识产权运营，通常宜包括实施、许可、转让、投融资等。

d） 进行人工智能模型开发成果自身知识产权成果的再产出，明确知识产权归属。

### 5.2.9 环境保护与社会责任

#### 5.2.9.1 第一级

本级能力如下：

a） 相关部门和人员从业务角度出发开发、实施和使用人工智能模型，未考虑模型对环境和社会的影响，未从对人工智能模型相关的环境和社会议题进行管理。

#### 5.2.9.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 主动管理部分重要的人工智能模型在开发、实施和使用过程中所涉及的部分重要的环境和社会议题。

c） 识别人工智能模型可能带来的环境和社会方面的正负面影响。

d） 明确禁止涉足的不负责任人工智能领域，并与国内外主流机构标准保持一致。

#### 5.2.9.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 以风险为导向，系统性管理全部重要的人工智能模型在开发、实施和使用过程中所涉及的全部重要环境和社会议题。

c） 建立模型相关的环境和社会议题管理工作流程，宜包括信息收集汇报、识别改善空间并采取改进措施。

d） 评估现有人工智能模型的能源消耗总体水平，制定并采取措施提高模型的能源使用效率。宜研究使用可再生能源的可能性，提出能源替代方案。

e） 追踪评估人工智能模型的碳足迹，实施碳盘查，制定并采取措施减少碳排放。

f） 评估人工智能模型的应用所产生和排放的电子废弃物种类，制定并采取措施减少电子废弃物产生和排放。

g） 评估人工智能模型的应用对就业的影响，包括评估受影响的岗位和员工以及影响程度，并建立应对方案。

h） 系统化地识别并书面明确人工智能模型开发、实施和使用中，对员工造成的安全隐患并制定应对方案。

#### 5.2.9.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 在满足业务发展需求同时，发挥人工智能模型对环境和社会的积极影响，致力于联合国可持续发展目标的达成。

c） 结合国内外标准、趋势、政策以及自身经营情况，针对人工智能模型各项重要环境和社会议题制定量化的短、中、长期管理目标，并监督目标完成情况。

#### 5.2.9.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 量化人工智能模型主要的环境和社会影响价值，以指导投资决策，持续致力于最大化人工智能模型为环境和社会带来的积极影响价值。

c） 组织宜发挥组织自身影响力，持续引领人工智能行业负责任发展。

### 5.2.10 管理层监控和审计

#### 5.2.10.1 第一级

本级能力如下：

a） 未建立管理层监控机制流程，未借助独立审计对人工智能模型风险管理活动进行评估。

#### 5.2.10.2 第二级

本级能力如下：

a） 前一级的全部内容。

b） 管理层和审计人员参与风险管理过程，能够涵盖部分与人工智能模型风险相关的监控和审计。

#### 5.2.10.3 第三级

本级能力如下：

a） 前两级的全部内容。

b） 将人工智能模型风险纳入现有组织风险管理体系，或者单独为人工智能模型风险制定风险管理制度。

c） 将人工智能模型风险管理纳入审计计划，对模型本身、模型风险管理的有效性进行审计。审计计划应涵盖重要的模型，通过文档审阅、流程审阅、数据审阅、代码审阅等审计作业方法识别模型生命周期各阶段中的相关风险。

#### 5.2.10.4 第四级

本级能力如下：

a） 前三级的全部内容。

b） 配置适当资源和具备适当技术能力的人员开展模型审计相关工作，审计人员对模型的基本原理、相关编码语言、模型生命周期等有较好的理解，以及熟练使用相关自动化工具。

c） 审计计划涵盖大部分人工智能模型，除文档审阅、流程审阅、数据审阅和代码审阅外，采用白盒对比测试等技术验证方式对模型有效性进行独立验证。

#### 5.2.10.5 第五级

本级能力如下：

a） 前四级的全部内容。

b） 高级管理层定期审阅人工智能模型风险监控结果报告，及时了解人工智能模型存在的风险。

c） 定期评估现有人工智能模型风险管理策略方针和管理制度的适用性，制定相应的管理策略调整和风险应对方案，以确保人工智能风险管理具有前瞻性，更好的应对未知风险。

# 参考文献

[1] ISO/IEC 18023-1:2006 Information technology—SEDRIS—Part 1:Functional specification