国际标准

ISO/IEC
23894
第1版
2023-02

信息技术 —人工智能— 风险管理指南

Information technology - Artificial intelligence - Guidance on risk management

（雷泽佳译，2024-08）

目　次

前　言

国际标准化组织（ISO）是由各国标准化团体（ISO 成员团体）组成的世界性的联合会。制定国际标准工作通常由 ISO 的技术委员会完成。各成员团体若对某技术委员会确定的项目感兴趣，均有权参加该委员会的工作。与 ISO 保持联系的各国际组织（官方的或非官方的）也可参加有关工作。ISO 与国际电工委员会（IEC）在电工技术标准化方面保持密切合作的关系。

制定本标准及其后续标准维护的程序在 ISO/IEC 指引 第 1 部分均有描述。应特别注意用于各不同类别 ISO 文件批准准则。本标准根据 ISO/IEC 导则第 2 部分的规则起草（见 www. iso. org/directives 或 www. iec. ch/members_experts/refdocs）。

本标准中的某些内容有可能涉及一些专利权问题，对此应引起注意。ISO 不负责识别任何这样的专利权问题。在标准制定期间识别的专利权细节将出现在引言/或收到的 ISO 专利权声明清单中（www.iso.org/patents）。

本标准中使用的任何商品名称仅为方便用户而提供的信息， 并不构成认可。

ISO 与合格评定相关的特定术语和表述含义的解释以及 ISO 遵循的世界贸易组织（WTO）贸易技术壁垒（TBT）原则相关信息访问以下 URLL：www.iso.org/iso/foreword.html.。在 IEC 中， 请参见 www. iec. ch/understanding-standards

本标准由 ISO/IEC JTC 1，联合技术委员会，信息技术 SC 42 人工智能分委员会编写。

关于本标准的任何反馈或者疑问都应直接向用户的国家标准机构提出。 完整的国家标准机构列表可访问 www.iso.org/members.html 以及 www.iec.ch/national-committees 获取。

引　言

风险管理的目的是创造和保护价值。它能够提高绩效、鼓励创新并支持实现目标。

本文件旨在与ISO 31000:2018结合使用。当本文件扩展ISO 31000:2018中给出的指南时，会适当引用ISO 31000:2018的条款，并在适用的情况下提供特定于人工智能（AI）的指导。为了使本文件与ISO 31000:2018之间的关系更加明确，本文件采用了ISO 31000:2018的条款结构，并在需要时通过子条款进行了修订。

本文件分为三个主要部分：

第4章：原则——本条描述了风险管理的基本原则。如ISO 31000:2018第4章所述，使用人工智能需要对其中一些原则进行特别考虑。

第5章：框架——风险管理框架的目的是帮助组织将风险管理整合到重要活动和职能中。ISO 31000:2018第5章描述了人工智能体系的开发、提供或使用的具体方面。

第6章：过程——风险管理过程涉及将方针、程序和实践系统地应用于沟通、协商、确定环境、评估、应对、监视、评审、记录和报告风险的活动。ISO 31000:2018第6章描述了此类过程在人工智能方面的专业化应用。

附录A和附录B提供了与人工智能相关的常见目标和风险来源。附录C提供了风险管理过程与人工智能体系生命周期之间的示例映射。

信息技术 —人工智能— 风险管理指南

1 范围

本文件为开发、生产、部署或使用利用人工智能（AI）的产品、系统和服务的组织提供了关于如何管理特定于AI的风险的指导。该指导还旨在帮助组织将风险管理整合至其与AI相关的活动和职能中。此外，它还描述了有效实施和整合AI风险管理过程。

本指导的应用可根据任何组织及其环境进行定制。

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，仅注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。

ISO 31000：2018，风险管理——指南

ISO Guide 73:2009 风险管理——词汇

ISO/IEC 22989:2022 信息技术 人工智能 人工智能概念和术语

3 术语和定义

本文件采用ISO 31000：2018、ISO/IEC 22989：2022和ISO 导则 73:2009中给出的术语和定义。

ISO和IEC在以下地址维护用于标准化的术语数据库：

ISO在线浏览平台： https://www.iso.org/obr

IEC在线电工术语库： http://www.electropedia.org/

4 人工智能风险管理原则

风险管理应采用整合、结构化和全面的方法来解决组织的需求。指导原则使组织能够确定优先级，并就如何管理不确定性对目标的影响做出决策。这些原则适用于所有组织层级和目标，无论是战略性的还是操作性的。

系统和过程通常在各种环境中部署各种技术和功能组合，以用于特定的用例。风险管理应考虑整个系统及其所有技术和功能，以及其对环境和相关方的影响。

人工智能系统可能为组织引入新的或新兴的风险，对目标产生积极或消极的后果，或改变现有风险的可能性。它们也可能需要组织进行特定的考虑。本文件为组织可以实施的风险管理原则、框架和过程提供了额外的指导。

注：不同的国际标准对“风险”一词的定义存在显著差异。在ISO 31000：2018及相关国际标准中，“风险”涉及偏离目标的情况，可能是负面的或正面的。而在其他一些国际标准中，“风险”仅涉及潜在的负面结果，例如与安全相关的问题。这种关注点的差异在尝试理解和正确实施符合标准的风险管理过程时常常会造成混淆。

ISO 31000:2018的第4章规定了风险管理的几个通用原则。除了ISO 31000：2018第4章中的指导外，表1还提供了在必要时如何应用这些原则的进一步指导。

表1：应用于人工智能的风险管理原则

原则	ISO 31000:2018 中的描述（第 4 章）	对人工智能开发和使用的影响
a）整合	风险管理是组织所有活动的有机组成部分。	超出 ISO31000:2018 的具体指导无。
b）结构化和全面性	采用结构化和全面性的方法开展风险管理，有助于获得一致的和可比较的结果。	超出 ISO31000:2018 的具体指导无。
c）定制化	组织根据自身目标所对应的内外部环境，定制设计风险管理框架和过程。	超出 ISO31000:2018 的具体指导无。
d）包容性	相关方适当、及时的参与，可以使他们的知识、观点和认知得到充分考虑。这样有助于提高组织的风险意识，并促进风险管理信息的充分沟通。	由于人工智能可能对相关方产生深远的影响，组织应寻求与多样化的内部和外部团体进行对话，既沟通人工智能的利弊，又将反馈和意识纳入风险管理过程。组织还应意识到，人工智能系统的使用可能会引入额外的相关方。相关方知识、观点和看法有益的领域包括但不限于： - 机器学习（ML）特别依赖于一组适合实现其目标的数据。相关方可以帮助识别与数据收集、处理操作、数据来源和类型以及数据在特定情况或数据主体可能为异常值时的使用相关的风险。 - 人工智能技术的复杂性带来了与人工智能系统透明度和可解释性相关的挑战。由于多种数据类型、人工智能

原则	ISO 31000:2018 中的描述（第 4 章）	对人工智能开发和使用的影响
		模型拓扑结构以及应根据相关方的需求选择的透明度和报告机制等特点，人工智能技术的多样性进一步加剧了这些挑战。相关方可以帮助确定目标并描述增强人工智能系统透明度和可解释性的手段。在某些情况下，这些目标和手段可以在用例和涉及的不同相关方之间推广。在其他情况下，可以根据用例为相关角色（例如“监管机构”“业务所有者”“模型风险评估者”）定制透明度框架和报告机制的相关方细分。 - 使用人工智能系统进行自动化决策可以直接影响内部和外部相关方。这些相关方可以提供他们对例如何时需要人工监督的看法。相关方可以帮助定义公平性标准，并帮助识别人工智能系统工作中构成偏见的内容。
e) 动态性	随着组织内外部环境的变化，组织面临的风险可能会出现、变化或消失。风险管理以适当、及时的方式预测、发现、确认和应对这些变化和事件。	要实施 ISO31000:2018 提供的指导，组织应建立组织结构和措施，以识别与人工智能系统相关的新兴风险、趋势、技术、用途和参与者相关的问题和机遇。〈br〉动态风险管理对人工智能系统尤为重要，因为： <- 人工智能系统本身具有动态性，因为需要不断进行学习、提炼、评估和验证。 - 一些人工智能系统具有根据此循环进行适应和优化的能力，从而自行产生动态变化。 - 顾客对人工智能系统的期望很高，并且可能会随着系统本身的快速变化而变化。 - 与人工智能相关的法律和监管要求经常发生变化并不断更新。 此外，还可以考虑与质量、环境足迹、安全、医疗保健、法律或企业责任等管理体系的整合，以进一步了解和管理对组织、个人和社会的人工智能相关风险。
f) 最佳可用信息	风险管理的信息输入是基于历史信息、当前信息和未来预期的。在风险管理过程中宜明确考虑与这些信息和预期相关的限制条件和不确定性。信	考虑到人工智能影响个人与技术交互和反应方式的预期，从事人工智能系统开发的组织应跟踪其开发的人工智能系统进一步使用的相关信息，而人工智能系统的用户可以在整个人工智能系统生命周期内保持对这些系统

原则	ISO 31000:2018 中的描述（第 4 章）	对人工智能开发和使用的影响
	息宜及时、清晰，并且是有关的相关方可获得的。	使用的记录。〈br〉由于人工智能是一项新兴技术并不断发展，历史信息可能有限，未来预期可能会迅速变化。组织应考虑这一点。〈br〉还应考虑人工智能系统的内部使用（如果有的话）。由于知识产权、合同或特定市场限制，跟踪顾客和外部用户对人工智能系统的使用可能受到限制。这些限制应在人工智能风险管理过程中得到体现，并在业务条件需要时进行更新。
g) 人和文化因素	人的行为和文化在各个层级和阶段显著影响着风险管理的各个方面。	从事人工智能系统设计、开发或部署，或这些活动的任何组合的组织，应监视其所处的人文和文化环境。组织应重点关注识别人工智能系统或组件如何与现有的社会模式相互作用，这些模式可能导致对公平结果、隐私、言论自由、公平性、安全、就业、环境和广泛的人权产生影响。
h) 持续改进	通过不断学习和实践，持续改进风险管理。	在持续改进过程中，应考虑与人工智能系统使用相关的先前未知风险的识别。从事人工智能系统或系统组件的设计、开发或部署，或这些活动的任何组合的组织，应监视人工智能生态系统的性能成功、不足和吸取的教训，并保持对新的人工智能研究成果和技术的了解（改进机会）。

5 框架

5.1 总则

风险管理框架的目的是帮助组织将风险管理整合至关重要的活动和职能中。ISO 31000:2018中5.1提供的指导适用。

风险管理涉及收集与组织做出决策和应对风险相关的相关信息。虽然治理机构规定了总体风险承受能力和组织目标，但它将识别、评价和应对风险的决策过程委托给组织内部的管理层。

ISO/IEC 38507描述了组织在开发、购买或使用人工智能系统时需要考虑的额外治理因素。这些因素包括新的机遇、风险承受能力的潜在变化以及新的治理方针，以确保组织负责任地使用人工智能。它可以与本文件中描述的风险管理过程结合使用，以帮助指导ISO 31000:2018中5.2描述的组织动态和迭代整合。

5.2 领导作用和承诺

ISO 31000:2018中5.2提供的指导适用。

除了ISO 31000:2018中5.2提供的指导外，以下也适用：

鉴于人工智能的开发和使用与信任和问责制的特别重要性，最高管理者应考虑如何将与人工智能风险和风险管理相关的方针和声明传达给相关方。证实这一层级的领导作用和承诺对于确保相关方相信人工智能正在以负责任的方式被开发和使用至关重要。

因此，组织应考虑发布与其对人工智能风险管理的承诺相关的声明，以增加相关方对其使用人工智能的信心。

最高管理者还应意识到可能需要专门资源来管理人工智能风险，并适当分配这些资源。

5.3 整合

ISO 31000:2018中5.3提供的指导适用。

5.4 设计

5.4.1 理解组织及其环境

ISO 31000:2018中5.4.1提供的指导适用。

除了ISO 31000:2018中5.4.1提供的指导外，表2列出了在理解组织外部环境时应考虑的额外因素。

表2：建立组织外部环境时应考虑的因素

通用指导（ISO31000:2018, 5.4.1） 组织还应考虑其外部环境的以下要素：	人工智能相关组织的额外指导 组织还应考虑但不限于以下要素：
国际、国内、区域或地方的社会、文化、政治、法律、监管、金融、技术、经济、自然环境	- 相关的法律要求，包括与人工智能特别相关的法律要求。 - 政府相关部门、监管机构、标准化机构、民间社会、学术界和行业协会发布的关于人工智能和自动化系统道德使用和设计的指南。 - 与人工智能相关的特定领域指南和框架。
对组织目标产生影响的关键驱动因素和趋势	- 人工智能各领域的技术趋势和进步。 - 人工智能系统部署的社会和政治影响，包括来自社会科学的指导。
与外部相关方的关系，以及他们的认知、价值取向、需求和期望	- 相关方可能受到诸如人工智能系统缺乏透明度（也称为不透明性）或人工智能系统存在偏见等因素的影响。 - 相关方对特定基于人工智能的解决方案的可用性的期望，以及人工智能模型如何提供（例如，通过用户界面、软件

	开发工具包）的方式。
合同关系和承诺；	- 人工智能的使用，尤其是使用持续学习的人工智能系统，如何影响组织履行合同义务和保证的能力。因此，组织应仔细考虑相关合同的范围。 - 设计和生产人工智能系统和服务过程中的合同关系。例如，当第三方提供测试和训练数据时，应考虑其所有权和使用权。
组织所处关系网络的复杂性及依赖关系	- 人工智能的使用可能会增加网络和依赖关系的复杂性。
（超越 ISO31000:2018 的指导）	- 当一个人工智能系统取代现有系统时，可以对人工智能系统与现有系统的风险收益和风险转移进行评估，同时考虑与人工智能系统实施相关的安全、环境、社会、技术和财务问题。

除了ISO 31000:2018中5.4.1提供的指导外，表3还列出了在了解组织内部环境时需要考虑的其他因素。

表3：确定组织内部环境时需要考虑的因素

通用指导（ISO31000:2018，5.4.1） 组织还应考虑其内部环境的以下要素：	**人工智能相关组织的额外指导 组织还应考虑但不限于以下要素：**
愿景、使命和价值观；	- 除 ISO 31000:2018 外无具体指导
治理方式、组织结构、职能、责任和绩效考核；	- 除 ISO 31000:2018 外无具体指导
战略、目标和方针；	- 除 ISO 31000:2018 外无具体指导
组织文化；	- 人工智能系统通过转变和引入新的职责、角色和任务，对组织文化产生的影响。
组织采用的标准、指南和模型；	- 使用人工智能系统所带来的任何额外的国际、地区、国家和地方标准和准则。

组织在资源和知识方面所具备的能力（即资本、时间、人力、知识产权、程序、系统和技术等）；	- 种能力所需的人力资源数量发生变化，或者所需资源类型发生变化，例如，由于人工智能系统越来越多地支持人类决策，可能会导致技能降低或专业知识流失。 - 开发和使用人工智能系统所需的人工智能技术和数据科学方面的专业知识。 - 人工智能工具、平台和库的可用性可能使得在没有完全理解技术、其局限性和潜在陷阱的情况下就能开发人工智能系统。 - 人工智能可能为特定的人工智能系统带来与知识产权相关的问题和机遇。组织应考虑自己在这一领域的知识产权，以及知识产权如何影响透明度、安全性和与相关方合作的能力，以确定是否应采取任何措施。
数据、信息系统和信息流；	- 人工智能系统可用于自动化、优化和增强数据处理。 - 作为数据的消费者，人工智能系统可能会对数据和信息施加额外的质量和完整性约束。
与内部相关方的关系，充分考虑其认知和价值取向；	- 相关方的认知可能会受到人工智能系统缺乏透明度或存在偏见等问题的影响。 - 特定的人工智能系统可以在更大程度上满足相关方的需求和期望。 - 需要教育相关方了解人工智能系统的能力、故障模式和故障管理。
合同关系和承诺；	- 相关方的认知可能会受到与人工智能系统相关的不同挑战的影响，如潜在的缺乏透明度和不公平性。 - 特定的人工智能系统可以满足相关方的需求和期望。 - 需要教育相关方了解人工智能系统的能力、故障模式和故障管理。 - 相关方对隐私的期望，以及个人和集体的基本权利和自由。
相互依赖性和相互关联性。	- 使用人工智能系统可能会增加相互依赖和相互联系的复杂性。

除ISO 31000:2018中5.4.1提供的指导外，组织还应考虑使用人工智能系统可能会增加对专业培训的需求。

5.4.2 明确表达风险管理承诺

适用ISO 31000:2018中5.4.2提供的指导。

5.4.3 确组织角色、权限、职责和责任

适用ISO 31000:2018中5.4.3提供的指导。

除ISO 31000:2018，5.4.3中的指导外，最高管理者和监督机构（如适用）应配置资源并确定人员：

——具有应对人工智能风险的权限；

——负责建立和监视应对人工智能风险过程。

5.4.4 资源配置

适用ISO 31000:2018中5.4.4提供的指导。

5.4.5 沟通和协商

适用ISO 31000:2018中5.4.5提供的指导。

5.5 实施

适用ISO 31000:2018中5.5提供的指导。

5.6 评价

适用ISO 31000:2018中5.6提供的指导。

5.7 改进

5.7.1 调整

适用ISO 31000:2018中5.7.1提供的指导。

5.7.2 持续改进

适用ISO 31000:2018中5.7.2中提供的指导。

6 风险管理过程

6.1 总则

遵循ISO 31000:2018中6.1中提供的指导。

组织应采用基于风险的方法来识别、评价和理解其所面临的AI风险，并根据风险水平采取相应的应对措施。组织整体AI风险管理过程的成功依赖于在战略、运行、计划和项目层面识别、建立并成功实施范围明确的风险管理过程。由于一些基于AI的技术具有潜在的复杂性、缺乏透明度和不可预测性等令人担忧的

问题，因此应特别关注AI系统项目层面的风险管理过程。这些系统项目层面的过程应与组织的目标相一致，并应受到其他层面风险管理的影响，同时也应对其他层面风险管理产生影响。例如，应将AI项目层面的升级和吸取的经验教训纳入更高级别的风险管理，如战略、运行和计划层面，以及其他适用的层面。

项目层面风险管理过程的范围、环境和准则直接受到项目范围内AI系统生命周期阶段的影响。附录C展示了项目层面风险管理过程与AI系统生命周期（如ISO/IEC 22989:2022中所定义）之间可能的关系。

6.2 沟通和协商

遵循ISO 31000:2018中6.2中提供的指导。

受AI系统影响的相关方群体可能比最初预想的更大，可能包括未考虑到的外部相关方，并可能延伸至社会的其他部分。

用AI系统的领域。此类AI开发和使用的清单应形成文件，并纳入组织的风险管理过程。

6.3 范围、环境、准则

6.3.1 总则

遵循ISO 31000:2018中6.3.1提供的指导。

除了ISO 31000:2018中6.3.1提供的指导外，对于使用AI的组织而言，AI风险管理的范围、AI风险管理过程的环境以及用于评价风险重要性以支持决策过程的准则，应扩展到识别组织内正在开发或使用AI系统的领域。此类AI开发和使用的清单应形成文件，并纳入组织的风险管理过程。

6.3.2 界定范围

遵循ISO 31000:2018中6.3.2提供的指导。

确定范围时应考虑组织不同层级的具体任务和职责。此外，还应考虑组织开发或使用的AI系统的目标和目的。

6.3.3 内外部环境

遵循ISO 31000:2018中6.3.3中提供的指导。

由于AI系统的潜在影响巨大，组织在形成和建立风险管理过程的环境时，应特别关注其相关方的环境。

应谨慎考虑一系列相关方，包括但不限于：

——组织（本身）；

——顾客、合作伙伴和第三方；

——供方；

——最终用户；

——监管机构；

——民间组织；

——个人；

——受影响的社区；

——社会。

对于外部和内部环境的其他一些考虑因素包括：

AI系统是否会伤害人类、拒绝提供基本服务（若中断将危及生命、健康或个人安全）或侵犯人权（例如，通过不公平和有偏见的自动化决策）或导致环境损害；

——对组织社会责任的外部和内部期望；

——对组织环境责任的外部和内部期望。

ISO 26000:2010[2]中概述社会责任方面的指南应作为理解和处理风险的框架，特别是在组织治理、人权、劳动实践、环境、公平运营实践、消费者问题和社区参与及发展等核心主题上。

注：关于可信度的更多环境信息，可在ISO/IEC TR 24028:2020中找到。

6.3.4 界定风险准则

遵循ISO 31000:2018中6.3.4中提供的指导。

除了ISO 31000:2018中6.3.4中提供的指导外，表4还提供了在界定风险准则时需要考虑的因素的额外指南：

表4：界定风险准则时的额外指南

ISO 31000:2018 中 6.3.4 规定的界定风险准则的考虑因素：	人工智能系统开发和使用环境下的其他考虑因素
可能影响结果和目标的不确定因素的性质和类型（包括有形的和无形的）；	- 组织应采取合理的步骤，以了解人工智能（AI）系统中所有部分的不确定性，包括使用的数据、软件、数学模型、物理扩展以及系统中的人为因素（如在数据收集和标注过程中的任何相关人为活动）。
如何界定和度量后果（包括正面的和负面的）和可能性；	
时间相关因素；	- 除 ISO31000:2018 外，无其他具体指导。
采用度量标准的一致性；	- 组织应意识到，人工智能（AI）是一个快速发展的技术领域。应根据其有效性和对正在使用的人工智能系统的适用性，持续评估测量方法。
如何确定风险等级；	- 组织应建立一致的方法来确定风险水平。该方法应反映人工智能系统对不同人工智能相关目标（见附录 A）的潜在影响。
如何考虑多项风险的组合及顺序；	- 除 ISO 31000:2018 外，无其他具体指导。
组织的风险容量。	- 在决定人工智能（AI）风险承受度时，应考虑组织的 AI 能力、知识水平以及缓解已识别 AI 风险的能力。

6.4 风险评估

6.4.1 总则

ISO 31000:2018,6.4.1中提供的指导适用。

应识别人工智能（AI）风险，对其进行量化或定性描述，并根据与组织相关的风险准则和目标对其进行优先级排序。附录B提供了与AI相关的风险来源的示例目录。此类示例目录不能被视为全面的。然而，经验表明，对于首次进行风险评估或将AI风险管理整合到现有管理结构中的任何组织而言，使用此类目录作为基础具有价值。该目录可作为这些组织的文件化基线。

因此，从事AI系统开发、供应或应用的组织应将其风险评估活动与系统生命周期保持一致。系统生命周期的不同阶段可采用不同的风险评估方法。

6.4.2 风险识别

6.4.2.1 总则

ISO 31000:2018,6.4.2中提供的指导适用。

6.4.2.2 资产及其价值的识别

组织应识别与AI的设计和使用相关的资产，这些资产属于6.3.2中定义的风险管理过程的范围。了解哪些资产属于该范围以及这些资产的相对关键性或价值，对于评估影响至关重要。应同时考虑资产的价值和资产的性质（有形或无形）。此外，在AI的开发和使用方面，应在包括但不限于以下要素的环境下考虑资产：

——组织资产及其价值：

- 有形资产可包括数据、模型和 AI 系统本身。
- 无形资产可包括声誉和信任。

——个人资产及其价值：

- 有形资产可包括个人的个人数据。
- 无形资产可包括个人的隐私、健康和安全。

——社区和社会资产及其价值：

- 有形资产可包括环境。
- 无形资产可能更多基于价值，如社会文化信仰、社区知识、教育机会和公平。

关于资产的估值及其与影响的关系，见6.4.2.6和6.4.3.2。

注：本条款中使用“资产”一词及其示例不具有任何法律含义。

6.4.2.3 风险源的识别

组织应在所规定的范围内，识别与AI的开发或使用或两者相关的风险源列表。

可在但不限于以下领域识别风险源：

——组织；

——过程和程序；

——管理例程；

——人员；

——物理环境；

——数据；

——AI 系统配置；

——部署环境；

——硬件、软件、网络资源和服务；

——对外部方的依赖。

附录B中提供了与AI相关的风险源的示例。

6.4.2.4 潜在事件和结果的识别

组织应识别与AI的开发或使用相关的潜在事件，这些事件可能导致各种有形或无形的后果。

可通过以下一种或多种方法和来源识别事件：

——已发布的标准；

——已发布的技术规范；

——已发布的技术报告；

——已发布的科学论文；

——有关已在使用的类似系统或应用的市场数据；

——有关已在使用的类似系统或应用的事件报告；

——现场试验；

——可用性研究；

——适当的调查结果；

——相关方报告；

——与内部或外部专家的面谈及其报告；

——模拟。

6.4.2.5 控制措施的识别

组织应识别与AI的开发或使用或两者相关的控制措施。应在风险管理活动中识别控制措施，并形成文件（记录在内部系统、程序、审核报告等中）。

控制措施可用于通过缓解风险源、事件和结果来积极影响总体风险。

还应考虑所识别的控制措施的运行有效性，特别是控制失效的情况。

6.4.2.6 后果的识别

作为AI风险评估的一部分，组织应识别可能导致风险的风险源、事件或结果。它还应识别对组织本身、个人、社区、群体和社会造成的任何后果。组织应特别注意识别从技术中受益的群体和经历负面后果的群体之间的任何差异。

对组织的后果必然不同于对个人和社会的后果。对组织的后果可以包括但不限于：

——调查和修复时间；

——（工作）时间的得失；

——机会的得失；

——对个人健康或安全的威胁；

——修复损害所需特定技能的财务成本；

——员工的招聘、满意度和留存率；

——形象、声誉和商誉；

——处罚和罚款；

——顾客诉讼。

根据具体情况，对个人和社会的后果可能更为普遍，在这种情况下，组织可能无法准确估计对每个人或社会的影响。

与其具体说明每种影响类型，不如将其视为影响的关键要素，即影响的严重程度（例如，对个人的隐私、公平性、人权等方面的影响，或对社会环境的影响）。

确切的影响可能取决于组织运营的环境以及AI系统的开发和使用领域。

注 1：后果可以是积极的或消极的。组织在评估对组织、个人和社会的影响时，可以考虑两者。

注 2：对个人和社会的影响通常也可能导致对组织的影响。例如，组织产品的用户发生安全事故可能会导致组织承担责任，并对其声誉和产品销售产生负面影响。

6.4.3 风险分析

6.4.3.1 总则

应用ISO 31000:2018标准中6.4.3部分提供的指导。

分析方法应与作为建立环境（见6.3）一部分而制定的风险标准相一致。

6.4.3.2 后果评估

在评估风险评估中确定的后果时，组织应区分业务影响评估、个人影响评估和社会影响评估。

业务影响分析应确定组织受影响的程度，并考虑包括但不限于以下要素：

——影响的严重程度；

——有形和无形影响；

——用于确定总体影响的标准（如 6.3.4 中所确定）。

个人影响分析应确定个人受组织开发或使用AI，或两者共同影响的程度。它们应考虑包括但不限于以下要素：

——从个人那里使用的数据类型；

——开发或使用 AI 的预期影响；

——对个人的潜在偏见影响；

——可能对个人造成物质和非物质损害的潜在基本权利影响；

——对个人的潜在公平性影响；

——个人的安全；

——围绕不期望的偏见和不公平性的保护措施和缓解控制；

——个人的司法管辖区和文化环境（这可能影响如何确定相对影响）。

社会影响分析应确定社会受组织开发或使用AI，或两者共同影响的程度。它们应考虑包括但不限于以下要素：

——社会影响的范围（AI 系统对不同人群的覆盖程度有多广），包括谁在使用或设计该系统（例如，政府使用可能对社会的影响大于私人使用）；

——AI 系统如何影响各种受影响群体所持的社会和文化价值观（包括 AI 系统以特定方式放大或减少对不同社会群体已存在的伤害模式）。

6.4.3.3 可能性评估

在适用的情况下，组织应评估导致风险的事件和结果发生的可能性。可能性可以在定性或定量尺度上确定，并应与6.3.4部分制定的标准相一致。可能性的确定和受影响可由（但不限于）以下因素：

——风险来源的类型、重要性和数量；

——威胁的频率、严重性和普遍性；

——内部因素，如政策和程序的运行成功度以及内部行为者的动机；

——外部因素，如地理位置以及其他社会、经济和环境因素；

——控制措施的成功（缓解）或失败（见 6.4.2.5）。

组织仅应在可能性的计算和应用对于确定风险应对措施的应用位置是有用和适用的情况下，才纳入可能性计算。基于可能性的决策可能存在重大的技术、经济和启发式问题，特别是当可能性无法计算或计算结果存在较大误差时。

6.4.4 风险评价

应用ISO 31000:2018中6.4.4提供的指导。

6.5 风险应对

6.5.1 总则

应用ISO 31000:2018中6.5.1提供的指导。

6.5.2 选择风险应对方案

应用ISO 31000:2018中6.5.2提供的指导。

组织定义的风险应对方案应旨在将风险的负面后果降低到可接受的水平，并提高实现积极结果的可能性。如果通过应用不同的风险应对方案无法实现负面结果的必要减少，组织应对剩余风险进行风险－效益分析。

根据ISO 31000:2018中6.5.2，组织应考虑：

——决定不开始或退出会导致风险的活动，来规避风险；

——承担或增加风险，以寻求机会；

——消除风险源；

——改变可能性；

——改变后果；

——分担风险（如通过签订合同，购买保险）；

——慎重考虑后决定保留风险。

6.5.3 编制和实施风险应对计划

应用ISO 31000:2018中6.5.3提供的指导。

一旦风险应对计划得到记录，就应实施6.5.2中选择的风险应对措施。

应根据6.7验证并记录每项风险应对措施的实施及其有效性。

6.6 监视和评审

应用ISO 31000:2018中6.6提供的指导。

6.7 记录和报告

应用ISO 31000:2018标准中6.7部分提供的指导。

组织应建立、记录并保持系统，用于收集和验证实施阶段及实施后阶段的产品或类似产品的信息。组织还应收集和评审市场上类似系统的公开可用信息。

然后，应评估这些信息对人工智能系统可信度的可能相关性。特别是，评估应判断是否存在之前未检测到的风险，或之前评估过的风险是否已不再可接受。这些信息可以作为组织人工智能风险管理过程中的目标调整、用例或经验教训等因素进行输入和考虑。

如果满足其中任何一项条件，组织应执行以下操作：

评估对之前风险管理活动的影响，并将此评估的结果反馈到风险管理过程中。

对人工智能系统的风险管理活动进行评审。如果残余风险或其可接受性有可能发生变化，则应评估对现有风险控制措施的影响。

应记录此评估的结果。风险管理记录应允许追踪每个已识别风险在所有风险管理过程中的情况。记录可以利用组织商定的通用模板。

除了记录范围、环境和准则（见6.3）风险评估（见6.4）和风险应对（见6.5）之外，记录还应至少包括以下信息：

——已分析系统的描述和标识；

——应用的方法论；

——人工智能系统预期用途的描述；

——进行风险评估的人员和组织的身份；

——风险评估的参考条款和日期；

——风险评估的发布状态；

——目标是否以及在多大程度上得到了满足。

附 录 A（资料性）：目标

A.1 总则

在识别人工智能系统的风险时，应根据所考虑系统的性质及其应用环境，考虑各种与人工智能相关的目标。与人工智能相关的目标包括但不限于条款A.2至A.12中所述的目标。

A.2 责任

责任既指组织的特性，也指系统的属性：

组织责任意味着组织通过解释其决策和行动，并对治理机构、法律机构以及更广泛的相关方负责，来承担其决策和行动的责任。

系统责任与能够追溯实体的决策和行动至该实体相关。

人工智能的使用可以改变现有的责任框架。在此之前，人员执行他们将被追究责任的行动，而现在这些行动可以完全或部分由人工智能系统执行。在这种情况下，谁应负责是监管机构正在考虑的问题。人工智能系统的开发者和使用者应了解该系统上市和使用所在国家的相关法律。

A.3 AI专业知识

AI系统及其开发不同于非AI软件解决方案。需要选择一批具备跨学科技能和专业知识的专家，以评估、开发和部署AI系统。组织应确保具有此类专业知识的人员参与AI系统的开发和规范制定。

AI的专业知识应扩展到AI系统的最终用户。用户应充分了解AI系统的功能，并有权检测和纠正错误的决策或输出。

A.4 训练和测试数据的可用性和质量

基于机器学习（ML）的AI系统需要训练和测试数据来训练和验证系统的预期行为。部署的AI系统在生产数据上运行。训练、测试和生产数据在数据类型和质量方面应符合预期行为。

应验证训练和测试数据的时效性和相关性，以符合其预期目的。所需的训练和测试数据的数量可能因预期功能和环境的复杂性而异。训练和测试数据应具有足够多样的特征，以便为AI系统提供强大的预测能力。此外，应确保训练和测试数据之间的一致性，并在适用时使用独立的数据集。

训练和测试数据可能在公司内部不可用，而是从外部获取。在这种情况下，也应确保数据质量。

A.5 环境影响

使用AI可能会对环境产生影响。AI的使用可能对环境产生积极影响。例如，AI系统可用于减少燃气涡轮中的氮氧化物。然而，由于大量使用资源，AI的使用也可能对环境产生负面影响。例如，一些AI系统的训练阶段需要计算资源，并可能消耗大量电力。应考虑这些对环境的影响。

A.6 公平性

使用AI系统进行自动化决策可能对特定个人或群体不公平。不公平的结果有多种原因，如目标函数中的偏见、数据集不平衡以及训练数据和向系统提供反馈中的人为偏见。产品概念中的偏见问题、问题表述或关于何时何地部署AI系统的选择也可能导致不公平。

有关AI系统中的偏见和公平性的更多信息，请参见ISO/IEC TR 24027。

A.7 可维护性

可维护性与组织处理AI系统修改以纠正缺陷或适应新要求的能力有关。由于基于机器学习的AI系统是通过训练实现的，并不遵循基于规则的方法，因此应调查AI系统的可维护性及其影响。

A.8 隐私

隐私涉及个人控制或影响与其相关的信息（如收集、存储和处理的信息）的能力，以及谁可以披露这些信息。如ISO/IEC TR 24028:2020[31]所述，“许多AI技术（如深度学习）高度依赖于大数据，因为它们的准确性取决于所使用的数据量。一些数据（尤其是个人和敏感数据，如健康记录）的滥用或披露可能对数据主体产生有害影响。因此，在大数据分析和AI中，隐私保护已成为一个主要问题。”

应考虑确定AI系统是否能够推断出敏感的个人数据。对于AI系统，保护隐私包括保护用于构建和运行AI系统的数据，确保AI系统不会被用于非法访问其数据，以及保护为个性化模型或可用于推断类似个人的信息或特征的模型的访问。

不当收集、使用和披露个人信息也可能对基本人权（如歧视、言论自由和信息自由）产生直接影响。还应考虑对尊重人类价值和人类尊严的道德原则的影响。

注：数据保护影响评估（见ISO/IEC 29134:2017[51]），通常称为隐私影响评估，是管理在数据收集、AI系统训练和AI系统使用过程中使用个人数据的相关风险的有用工具。

A.9 稳健性

稳健性与系统在各种使用情况下保持其性能水平的能力有关。应考虑AI系统或相关组件在存在无效输入或压力环境条件下正确运行的程度，以及再现措施和结果的能力。

在AI系统的环境中，稳健性带来了新的挑战。神经网络架构代表了一个特定的挑战，因为它们既难以解释，又由于其非线性性质而有时表现出意外的行为。表征神经网络的稳健性是一个开放的研究领域，测试和验证方法都存在局限性。

有关神经网络稳健性的更多信息，请参见ISO/IEC TR 24029-1。

A.10 安全性

使用AI系统可能会引入新的安全威胁。安全性与系统在定义条件下不会导致人类生命、健康、财产或环境处于危险状态的预期有关。在自动驾驶车辆、制造设备和机器人中使用AI系统可能会引入与安全相关的风险。对于这些领域的AI系统，应考虑特定应用领域（如机械设计、交通、医疗设备）的特定标准。

有关AI系统中功能安全的更多信息，请参见ISO/IEC TR 5469。

A.11 （保安）安全性

信息安全风险管理在ISO/IEC 27005:2022中有所定义。在AI的环境中，特别是基于机器学习方法的AI系统，除了传统的信息和系统安全问题外，还应考虑ISO/IEC TR 24028:2020中描述的数据投毒、对抗性攻击和模型窃取等新问题。

A.12 透明度和可解释性

透明度既与运营AI系统的组织的特征有关，也与这些系统本身的特征有关。组织有时会在如何应用这些系统、如何使用收集的数据（如消费者和用户数据、公共数据、其他收集的数据集）、他们采取了哪些措施来管理AI系统、理解和控制其风险等方面保持透明。AI系统的透明度是指向相关方提供关于系统的适当信息（如能力和限制），以使他们能够根据自己的目标评估AI系统的开发、运行和使用。AI系统的可解释性是指能够合理化和帮助理解特定系统如何产生其结果的能力。

附　录　B（资料性）：风险源

B.1　总则

在识别人工智能（AI）系统的风险时，应根据所考虑系统的性质及其应用环境，考虑各种风险来源。需要考虑的风险来源包括但不限于B.2至B.8条款中所述的因素和机遇。

B.2　环境的复杂性

AI系统所处环境的复杂性决定了该系统在其运行环境中预期要支持的各种潜在情境的范围。

某些AI技术，如机器学习（ML），特别适合处理复杂环境，因此常用于自动驾驶等复杂环境中的应用系统。然而，一个巨大的挑战是在设计和开发过程中识别出系统预期要处理的所有相关情境，并确保训练和测试数据涵盖所有这些情境。

因此，与简单环境相比，复杂环境可能导致额外的风险。应特别考虑确定AI系统环境被理解的程度：

——对于简单、可预测或受控的环境，可以完全理解环境，这样 AI 系统就能为可能遇到的所有环境状态做好准备，从而更好地控制风险。

——在由于环境的高度复杂性或不确定性而导致只能部分理解的情况下（例如，自动驾驶），AI 系统无法预测环境的所有可能状态，因此不能假定已经考虑了所有相关情境。这可能导致一定程度的不确定性，这是风险的来源，在设计此类系统时应考虑这一点。

B.3　缺乏透明度和可解释性

透明度指向有关相关方传达组织的适当活动和决策（如政策、流程）以及关于AI系统的适当信息（如能力、性能、限制、设计选择、算法、训练和测试数据、验证和确认过程及结果）。这可以使相关方能够根据他们的期望评估AI系统的开发、运行和使用。适当的信息类型和水平在很大程度上取决于相关方、用例、系统类型和法规要求。如果组织无法向有关相关方提供适当的信息，这可能会对组织和AI系统的可信度和可问责性产生负面影响。

可解释性指AI系统的重要决策影响因素能够以人类可以理解的方式表达出来的特性。一个机器学习模型的行为可能很难通过检查模型或用于训练它的算法来理解，尤其是在深度学习的情况下。如果无法表达这些重要因素，那么AI系统的验证和人类对系统的信任都会受到负面影响，因为不清楚系统为什么会做出某个决策，以及它是否能在所有情况下都做出正确的决策。这种不确定性可能导致许多风险，并严重影响总体目标（如可信度和可问责性）和具体目标（如安全性、安保性、公平性和稳健性）。因此，可解释性不仅对于相关方作为AI系统透明度的一部分很重要，而且对于组织本身对AI系统的验证和确认也很重要。

过度的透明度和可解释性也可能导致与隐私、安保性、保密要求和知识产权相关的风险。

B.4　自动化水平

AI系统可以以不同的自动化水平运行。它们可以从无自动化（操作员完全控制系统）到完全自动化系统不等。AI系统通常是自动化系统。根据特定的用例，这类系统的自动化决策可能对安全性、公平性或安保性等各个关注领域产生影响。

对于需要外部代理在必要时准备就绪的自动化水平，从系统到代理的交接可能是一个风险源（例如时间限制、代理的注意力）。

有关自动化水平的更多信息，请参见ISO/IEC 22989:2022中的5.2。

B.5 与机器学习相关的风险源

AI的许多进展都与机器学习（ML）及其子领域（如深度学习）有关。机器学习系统的行为不仅取决于所使用的算法，还取决于用于训练机器学习模型的数据。因此，可能对AI特性产生的影响包括：

——数据质量：训练和测试数据的质量直接影响系统的功能。数据质量不足可能影响各种目标，如公平性、安全性和稳健性。

- 对于利用机器学习的 AI 系统，用于收集数据的过程是风险源，尤其难以诊断和检测。例如：
- 数据可能无法代表应用领域，从而给业务目标带来风险。

——数据来源和存储可能带来重大的伦理和法律风险。无法确保数据收集过程的安全可能导致来自对抗性攻击、数据投毒或其他操纵的风险。

——持续学习的 AI 系统旨在根据不断发展的生产数据改进系统，但同时也可能加剧风险，因为它们在使用过程中可能会以未预期的方式改变其行为。

B.6 系统硬件因素

与硬件因素相关的风险源包括但不限于：

——基于缺陷组件的硬件错误。例如，单个或多个存储单元的短路或中断、缺陷总线、漂移振荡器、集成电路输入或输出端的固定故障或寄生振荡。

——软错误，如存储单元或逻辑组件的不希望发生的临时状态变化，主要由高能辐射引起。

——由于不同系统在处理能力、内存和专用 AI 硬件加速器方面的差异，训练好的机器学习模型在不同系统之间的迁移可能会受到限制。

——当 AI 系统需要远程处理和存储时，由于网络资源的有限性和共享性，可能会出现网络错误、带宽限制和延迟增加。

B.7 系统生命周期因素

在AI系统的生命周期中，使用不当或不足的方法、过程以及使用方式都可能导致风险。此类风险的例子包括：

——设计和开发：有缺陷的设计过程可能无法预见 AI 系统使用的环境，导致在这些环境中使用时意外失败。

——验证和确认：发布 AI 系统更新版本的验证和确认过程不充分，可能导致意外的回归或质量、可靠性或安全性的无意恶化或降低。

——部署：不充分的部署配置可能导致与内存、计算、网络、存储、冗余或负载均衡相关的资源问题。

——维护、更新和修订：开发者不再支持或维护但仍在使用的 AI 系统，可能给开发组织带来长期风险或责任。

——重用：一个功能正常的 AI 系统可能被用于其最初未设计的环境，由于设计和实际使用之间的要求差异而导致问题。例如，一个为识别社交网络上共享的照片中的人脸而设计的系统，可能被用于尝试识别监

视录像中的犯罪嫌疑人的人脸，这是一个比原始用例要求更高精度的应用。

——退役：组织终止使用基于 AI 技术的某个 AI 系统或组件时，可能会失去退役系统提供的信息或决策专业知识。此外，如果使用另一个系统替换退役系统，组织处理信息和做出决策的方式可能会发生变化。

B.8 技术成熟度

技术成熟度表明给定技术在给定应用环境中的成熟程度。在AI系统的开发和应用中，使用较不成熟的技术可能会给组织带来未知或难以评估的风险。对于成熟的技术，可以获得更多种类的经验数据，从而使风险更容易识别和评估。然而，如果技术成熟，也存在自满和技术债务的风险。

附　录　C（资料性）：风险管理与 AI 系统生命周期

表C.1展示了根据ISO/IEC 22989:2022定义的风险管理过程与AI系统生命周期之间的映射关系示例。

表C.1：风险管理与AI系统生命周期

风险管理		风险管理过程（第 6 章）				
AI 系统生命周期	风险管理框架（第 5 章）	范围、环境与准则	风险评估	风险应对	监视与评审	记录与报告
组织层面的风险管理活动	治理机构设定 AI 风险管理方向。最高管理者作出承诺。 建立高级别的风险管理偏好和一般准则。	接收并处理来自 AI 系统风险管理过程的反馈报告。 因此，通过扩展和完善组织的风险管理工具来改进组织的风险管理框架				
		风险准则目录	潜在风险源目录、风险源评估与测量技术目录	已知或已实施的 AI 系统监视与控制技术目录	已知或已实施的缓解措施目录	为追踪、记录、报告和与内部和外部相关方共享 AI 系统信息而建立的方

风险管理	风险管理框架（第 5 章）	风险管理过程（第 6 章）				
AI 系统生命周期		范围、环境与准则	风险评估	风险应对	监视与评审	记录与报告
						法和定义的格式目录。
启动	治理机构在组织和相关方的原则和价值观的环境中评审 AI 系统目标。基于（通常是多层的）分析，确定 AI 系统是否可行并解决了组织寻求解决的问题。	通过定制组织的风险管理框架，建立 AI 系统风险管理过程和体系的风险准则。	识别（可能以多层方式）并详细描述特定于 AI 系统的风险源。	制定详细的风险应对计划。可能定义概念验证方法。	实施、测试和评估必要的"概念验证"方法。	分析和其结果以及建议被记录并传达给最高管理者。
设计与开发	治理机构根据收到的反馈报告不断重新评估目标、效	可能由于反馈报告而修改 AI 系统风险准则。	持续进行风险评估（可能在多	实施风险应对计划。风险应对和	在测试、验证和确认过程中，评	记录结果并将其反馈到相关的风险管

风险管理	风险管理框架（第 5 章）	风险管理过程（第 6 章）				
AI 系统生命周期		范围、环境与准则	风险评估	风险应对	监视与评审	记录与报告
验证与确认	能和系统的可行性。		个层面上）。	（剩余）风险评估继续进行，直到满足既定的风险准则。	估和调整系统组件以及整个系统的风险应对计划。	理过程活动中。如有必要，将结论传达给管理层和治理机构。
部署	治理机构根据收到的反馈报告不断重新评估目标和系统的可行性。	针对必要的"配置"更改，调整 AI 系统风险准则和风险管理过程。	持续进行风险评估（可能在多个层面上）。	由于"配置"更改，可能更新并实施风险应对计划。风险应对和（剩余）风险评估继续进行，直到满足既	重新评估 AI 系统的风险应对计划，以便进行必要的调整。	

风险管理	风险管理框架（第 5 章）	风险管理过程（第 6 章）				
AI 系统生命周期		范围、环境与准则	风险评估	风险应对	监视与评审	记录与报告
				定的风险准则。		
运营、监视 持续验证	治理机构根据收到的反馈报告不断重新评估目标和系统的可行性。	可能由于反馈报告而修改 AI 系统风险准则。	可能针对风险准则的变化调整系统的风险评估计划。	可能针对风险评估结果中的风险变化调整系统的风险应对计划。	评估和调整系统组件的风险应对计划。	
再评价	治理机构重新评审 AI 系统目标及其与组织和相关方原则和价值观的关系。 基于分析，确定 AI 系统是	针对 AI 系统的特定目的和范围、运营监视结果和新监管要求的任何潜在变化，对 AI 系统风险	评审现有特定于 AI 系统的风险源列表的相关性和任何可	可能更新风险应对计划。风险应对和（剩余）风险评估继续进行，直到	重新评估 AI 系统的风险应对计划，以便进行必要的调整。	

风险管理	风险管理框架（第 5 章）	风险管理过程（第 6 章）				
AI 系统生命周期		范围、环境与准则	风险评估	风险应对	监视与评审	记录与报告
	否可行。	管理过程和系统的风险准则进行再评估。	能的差距。	满足既定的风险准则。		
退役或替换	触发具有新目标、新风险及其缓解措施的新风险管理过程。	治理机构基于分析重新评审 AI 系统目标，确定 AI 系统退役或替换是否可行。	建立 AI 系统风险管理退役流程和系统的退役风险准则。	识别和详细描述特定于 AI 系统退役的风险源。	制定详细的风险应对计划。实施、测试和评估必要的"概念验证"方法。	

参 考 文 献

[1] ISO/IEC 38507:2022，信息技术—信息技术治理—组织使用人工智能的治理影响

[2] ISO 26000:2010，社会责任指南

[3] ISO/IEC TR 24028:2020，信息技术—人工智能—人工智能可信性概述

[4] ISO/IEC TR 24027:2021，信息技术—人工智能（AI）—AI系统中的偏见和AI辅助决策

[5] ISO/IEC 29134:2017，信息技术—安全技术—隐私影响评估指南

[6] ISO/IEC TR 24029-1:2021，人工智能（AI）—神经网络鲁棒性评估—第1部分：概述

[7] ISO/IEC TR 54692，人工智能—功能安全和AI系统

[8] ISO/IEC 27005:2022，信息安全、网络安全和隐私保护—信息安全风险管理指南

[9] RUSSELL S. J.，NORVIG P.，《人工智能：一种现代的方法》，第3版，Upper Saddle River，NJ：Prentice Hall，2010

华信金泰检验认证有限公司

Huaxin Jintai Inspection and Certification Co., Ltd.

人工智能风险管理体系要求

文件编号：CTS HXJT/YQMS-14-2026

文件版次：A/0

编　　制：文件编制小组

审　　核：

批　　准：

受控状态：受控

发布日期：2026 年 03 月 18 日　　实施日期：2026 年 03 月 18 日

文件修改记录

修订说明	修订页数	修订日期	批准

目录