|  |  |
| --- | --- |
| ICS  | 点击此处添加ICS号 |
| CCS  | 点击此处添加CCS号 |

|  |
| --- |
|  31 |

上海市地方标准

DB 31/T XXXX—XXXX

信息技术 人工智能系统生命周期治理指南

Information Technology - guidelines for life cycle governance of artificial intelligence system

XXXX - XX - XX发布

XXXX - XX - XX实施

       发布

目次

[前言 II](#_Toc168653002)

[引言 III](#_Toc168653003)

[1 范围 1](#_Toc168653004)

[2 规范性引用文件 1](#_Toc168653005)

[3 术语和定义 1](#_Toc168653006)

[4 人工智能系统生命周期治理总则 2](#_Toc168653007)

[4.1 概述 2](#_Toc168653008)

[4.2 目标 2](#_Toc168653009)

[4.3 任务 2](#_Toc168653010)

[5 人工智能系统生命周期治理框架 2](#_Toc168653011)

[6 人工智能系统生命周期治理流程 3](#_Toc168653012)

[6.1 概述 3](#_Toc168653013)

[6.2 统筹和规划 4](#_Toc168653014)

[6.3 构建和运行 4](#_Toc168653015)

[6.4 监控和评价 4](#_Toc168653016)

[6.5 改进和优化 5](#_Toc168653017)

[6.6 多主体协同敏捷治理 5](#_Toc168653018)

[7 人工智能系统生命周期治理措施 5](#_Toc168653019)

[7.1 生命周期通用的治理措施 5](#_Toc168653020)

[7.2 初始阶段的治理措施 7](#_Toc168653021)

[7.3 设计开发阶段的治理措施 8](#_Toc168653022)

[7.4 测试验证阶段的治理措施 8](#_Toc168653023)

[7.5 部署阶段的治理措施 10](#_Toc168653024)

[7.6 运行监测阶段的治理措施 11](#_Toc168653025)

[7.7 重新评估阶段的治理措施 12](#_Toc168653026)

[7.8 退出使用阶段的治理措施 13](#_Toc168653027)

[参考文献 15](#_Toc168653028)

[附录A（资料性） 人工智能系统生命周期治理监督检查表 16](#_Toc168653029)

1. 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。

本文件由上海市经济和信息化委员会提出并组织实施。

本文件由上海市人工智能标准化技术委员会归口。

本文件起草单位：上海计算机软件技术开发中心、上海市人工智能行业协会、上海依图科技有限公司、上海商汤智能科技有限公司、公安部第三研究所、上海人工智能实验室、上海交通大学、上海人工智能研究院有限公司、华东师范大学等。

本文件主要起草人：

1. 引言

当前，人工智能已成为新一轮科技革命和产业变革的重要驱动力量，上海已将人工智能定位为重点培育的三大先导产业之一。目前，全市正在加快建设具有全球影响力的人工智能高地，全行业呈现技术越做越精、场景日益做深、价值红利不断释放的趋势。与此同时，人们在享受人工智能技术进步带来的红利和价值的同时，也面临着其引发的伦理和技术风险，因此迫切需要技术性指南来指导人工智能系统生命周期的治理，确保人工智能产业健康发展和对社会的正面影响。2017年7月，国务院印发的《新一代人工智能发展规划》，明确提出在大力发展人工智能的同时，必须高度重视安全风险挑战，确保人工智能安全、可靠、可控发展。2019年6月，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，提出公平公正、安全可控、开放协作、敏捷治理等治理原则。

目前人工智能企业、社会、政府对人工智能系统治理的重要性和必要性已达成基本共识。比如ISO/IEC SC42发布了ISO/IEC TR 24368:2022 《Information technology — Artificial intelligence — Overview of ethical and societal concerns》（信息技术 人工智能 伦理和社会关注概述）与 ISO/IEC 38507:2022 《Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations》（信息技术 IT治理 组织使用人工智能的治理影响）。然而，上述人工智能治理的标准侧重于建立人工智能治理的基本原则，而非形成技术指导文件。总体上，目前国际与国内缺乏标准化、可应用落地的人工智能系统生命周期治理的技术性指南。因此，制定人工智能系统生命周期治理指南，形成人工智能系统治理、评估和管控方法，加强前瞻预防与约束引导，有利于实现安全可信、负责任的人工智能应用，最大程度地释放人工智能在社会、经济、环境等各方面的正向价值。

本文件参考了ISO/IEC 22989、ISO/IEC TR 24368、ISO/IEC 38507等人工智能国际标准中人工智能生命周期、人工智能治理相关条文，提出了人工智能系统生命周期治理的总则、治理框架和治理流程，提供了人工智能系统生命周期治理的建议，包括初始阶段、设计开发、测试验证、部署、运行监测、重新评估、退出使用等全过程治理措施。本文件适用于上海市企业内部在研发人工智能系统时治理的实施和自我评价，以及政府监管机构、第三方机构对企业人工智能系统研发、应用过程中治理能力的评价。

信息技术 人工智能系统生命周期治理指南

* 1. 范围

本文件提出了人工智能系统生命周期治理总则、治理框架和治理流程，提供了人工智能系统管理者对人工智能系统各生命周期治理的建议，包括初始阶段、设计开发、测试验证、部署、运行监测、重新评估、退出使用等全生命周期的治理措施。

本文件适用于：

1. 指导本市企业内部人工智能系统生命周期治理的实施和自我评价。
2. 指导本市政府监管机构、第三方评估机构对企业的人工智能治理能力的评价。
	1. 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 41867-2022 信息技术 人工智能 术语

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 35273-2020 信息安全技术 个人信息安全规范

GB/T 25000.24-2017 系统与软件工程 系统与软件质量要求和评价（SQuaRE） 第24部分：数据质量测量

TC260-003 生成式人工智能服务安全基本要求

* 1. 术语和定义

GB/T 41867-2022界定的以及下列术语和定义适用于本文件。

人工智能治理 AI Governance

对人工智能研究、开发、生产和应用中出现的公共安全、道德伦理等问题进行协调、处理、监管和规范的过程。

人工智能系统生命周期 AI System Life Cycle

人工智能系统从初始阶段到退出使用的演变过程。

决策治理 Decision Governance

对人工智能系统作出的决策、人机协作生成的决策进行治理，确保其满足功能、性能要求。

数据治理 Data Governance

对数据来源、数据质量、和数据安全进行治理。

伦理治理 Ethic Governance

对人工智能系统的价值观、公平性、隐私性、透明度等进行治理，确保其符合人类的道德规范。

合规治理 Compliance Governance

对人工智能系统输出内容的法律法规符合性进行治理，确保其符合相关法律法规。

风险治理 Risk Governance

对人工智能系统的风险进行治理，包括风险的识别、分析和应对。

* 1. 人工智能系统生命周期治理总则
		1. 概述

人工智能系统生命周期治理源于组织的外部监管、内部管理、市场及利益相关者的需求，主要包括：

1. 法律法规、行业监管和内部管控等对人工智能系统的安全、合规的要求；
2. 人工智能系统生命周期过程中治理体系建设的要求；
3. 市场及利益相关者对人工智能系统的要求。
	* 1. 目标

人工智能系统生命周期治理的目标是保障人工智能系统的安全、合规，建立人工智能系统的治理体系，满足市场及利益相关者的期望。主要包括：

1. 安全：确保人工智能系统的数据、模型、系统、流程、物理和网络机制及服务得到内部或外部保护，以防止意外和未经授权的访问、使用、更改或破坏；确保人工智能系统不会对利益相关者的生命、健康、财产、数据以及物理环境造成不良后果。
2. 合规：确保人工智能系统符合法律法规和行业监管的规定，遵循社会普适的道德规范和伦理。
3. 治理体系建设：构建人工智能系统决策治理、数据治理、伦理治理、合规治理、风险治理体系。
	* 1. 任务

组织宜通过评估、指导和监督的方法，按照统筹和规划、构建和运行、监控和评价以及改进和优化的过程，实施人工智能系统生命周期治理的任务，主要包括：

1. 评估人工智能系统生命周期治理的现状及需求、系统内数据管理和系统运行情况；
2. 指导人工智能系统生命周期治理体系的构建和人工智能系统生命周期治理的实施；
3. 制定合理的评价体系与审计规范，监督人工智能系统生命周期治理内控、合规和绩效。
	1. 人工智能系统生命周期治理框架

人工智能系统生命周期治理框架包含人工智能系统在初始阶段、设计开发、测试验证、部署、运行监测、重新评估、退出使用等全生命周期中的治理，见图1。



1. 人工智能系统生命周期治理框架

人工智能系统生命周期过程中的治理包括决策治理、数据治理、伦理治理、合规治理、风险治理。

决策治理主要考虑人工智能系统的功能是否符合需求、系统性能等。

数据治理主要考虑数据的管理、质量、安全等。

伦理治理主要考虑人工智能的伦理、公平性、隐私性、透明度、可解释性等。

合规治理主要考虑人工智能的滥用、法律法规符合性等。

风险治理主要考虑人工智能的技术风险、应用风险、管理风险等的识别和应对。

人工智能系统生命周期过程中的治理受到管理者、人工智能系统特点、市场需求、利益相关者的期望等因素影响。人工智能系统相比传统信息系统，具有决策自动化、输出概率化、数据驱动、持续学习等特点，对治理提出新的挑战。此外市场对人工智能系统的合规性、伦理、实时性等需求、不同利益相关者对系统性能的不同要求，以及管理者对人工智能系统的期望也对人工智能系统治理产生影响。

1. 管理者是指在最高层指挥并控制人工智能系统的一个人或一组人
	1. 人工智能系统生命周期治理流程
		1. 概述

人工智能系统生命周期治理流程包含统筹和规划、构建和运行、监控和评价以及改进和优化，同时考虑多主体协同敏捷治理，如图2所示。



1. 人工智能系统生命周期治理流程
	* 1. 统筹和规划

明确人工智能系统生命周期治理目标和任务，营造必要的治理环境，做好人工智能系统生命周期治理实施的准备，包括：

1. 评估人工智能系统生命周期治理的资源、环境和人员能力等现状，分析与法律法规、行业监管、业务发展以及利益相关方需求等方面的差距，为人工智能系统生命周期治理方案的制定提供依据；
2. 指导人工智能系统生命周期治理方案的制定，包括组织机构和责任权利的规划、治理范围和任务的明确以及实施策略和流程的设计；
3. 监督人工智能系统生命周期治理的统筹和规划过程，保证现状评估的客观、组织机构设计的合理以及人工智能系统生命周期治理方案的可行。
	* 1. 构建和运行

构建人工智能系统生命周期治理实施的机制和路径，确保人工智能系统生命周期治理实施的有序运行，包括：

1. 评估人工智能系统生命周期治理方案与现有资源、环境和能力的匹配程度，为人工智能系统生命周期治理的实施提供指导；
2. 制定人工智能系统生命周期治理实施的方案，包括组织机构和团队的构建，责任权利的划分、实施路线图的制定、实施方法的选择以及管理制度的建立和运行等；
3. 监督人工智能系统生命周期治理的构建和运行过程，保证人工智能系统生命周期治理实施过程与方案的符合、治理资源的可用和治理活动的可持续。人工智能系统各生命周期的治理监督检查表可参考附录A。
	* 1. 监控和评价

监控人工智能系统生命周期治理的过程，评价人工智能系统生命周期治理的绩效、风险与合规，保障人工智能系统生命周期治理目标的实现，包括：

1. 构建必要的绩效评估体系、内控体系或审计体系，制定评价机制、流程和制度；
2. 评估人工智能系统生命周期治理成效与目标的符合性，必要时可聘请外部机构进行评估，为人工智能系统生命周期治理方案的改进和优化提供参考；
3. 定期评价人工智能系统生命周期治理实施的有效性、合规性，确保数据及其系统应用符合法律法规和行业监管要求。
	* 1. 改进和优化

改进人工智能系统生命周期治理方案，优化人工智能系统生命周期治理实施策略、方法和流程，促进人工智能系统生命周期治理体系的完善，包括：

1. 持续评估人工智能系统生命周期治理相关的资源、环境、能力、实施和绩效等，支撑人工智能系统生命周期治理体系的建设；
2. 指导人工智能系统生命周期治理方案的改进，优化人工智能系统生命周期治理的实施策略、方法、流程和制度，促进人工智能系统生命周期治理体系的完善；
3. 监督人工智能系统生命周期治理的改进和优化过程，为人工智能系统的合规、安全提供保障。
	* 1. 多主体协同敏捷治理

人工智能系统的治理宜采用政府、产业、学术研究、公众用户等多权利相关主体的协同共治，统筹安全与发展，保持持续创新能力，探索敏捷的治理方式。敏捷治理宜考虑以下因素：

1. 治理价值的平衡性：一方面要预先客观评估重大风险点，有效管控人工智能技术进步带来的安全隐患；另一方面也要拥抱新技术，优先保护人工智能技术发展的需要，以技术进步来缓解治理难题；
2. 治理主体的协同性：集聚政产学研用等多方力量参与到治理之中，积极在对话和合作中寻求社会共识，全面地、包容地优化治理方案，实现他律和自律的有效结合；
3. 治理手段的灵活性：根据新技术发展情况动态调整治理原则，及时明确治理目标与要求，积极应对、快速调整、有效管控，以降低技术路径和商业策略转变带来的风险。
	1. 人工智能系统生命周期治理措施
		1. 生命周期通用的治理措施
			1. 概述

组织在人工智能系统的整个生命周期中宜对系统的决策、数据、伦理、合规、风险开展治理，确保人工智能系统符合安全、合规的要求。

* + - 1. 决策治理

在决策治理方面，宜包括以下措施：

1. 建立人工智能系统生命周期中的日志记录机制，包括人工智能系统在生命周期过程中关键阶段点的系统情况，包括系统状态、人员及与建设、维护项目有关的必要信息；
2. 建立人工智能系统运行过程中的运行数据记录机制，对于涉及模型逻辑执行或更新的过程，需按用户及特定场景的要求进行记录。记录内容包括系统输入、输出、中间结果、运行环境上下文及改变情况、部署活动、配置及其更变等；
3. 建立数据记录的分层实现机制，从计算设备、加速库、软件框架、系统本身视角记录关键信息，各层次部件或组件所记录的信息能在语义上相互印证和说明；
4. 建立数据记录的防篡改、访问控制、防物理损坏等机制。
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 确保人工智能系统所在信息系统满足GB/T 22239-2019的要求，并对其进行测试；
2. 检查人工智能系统的控制流数据、信息流数据（包括特征流数据、模型流数据、结果流数据）及数据传输过程中的信息安全；
3. 对人工智能系统的数据、模型、代码进行安全扫描，并制定相应的安全事件处理机制。
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 可由人工智能系统管理者成立“伦理委员会”，监督和审查人工智能系统项目生命周期内的伦理符合性。
2. 按照GB/T 35273-2020的要求，制定人工智能系统所需数据的采集、标注、存储、传输、使用、删除的规范；
3. 评估人工智能模型的隐私保护机制，制定人工智能模型在训练、推理、部署过程中的隐私数据保护机制。
4. 从伦理技术角度，如公平性、鲁棒性、可解释性、隐私性等，可采用不同的技术和工具，应用于人工智能系统生命周期的不同阶段，使其满足相应指标要求；
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 对人工智能系统输出进行管理，制定管理规范，防止人工智能技术生成的虚假合成信息被用于不法活动，并开发相应的虚假信息甄别技术；
2. 确保人工智能系统在生成内容安全、准确性、可靠性方面满足《生成式人工智能服务安全基本要求》的规定；
3. 制定人工智能系统各利益相关方的权责划分和活动约束规范、系统生命周期各阶段的管理要求。并对各方活动进行指导、检视、评估和报告，使人工智能系统的提供方或使用方按照系统设计及使用说明中的承诺来提供或使用服务。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 管理层和监督机构应确保将风险管理纳入所有组织活动，并且要有针对性地设计和实现框架的所有要素，发布风险管理的方法、计划或行动方针的声明或政策，确保为管理风险分配必要的资源，在组织内相应的级别分配权限和职责。
2. 确定组织内的风险管理责任和监督角色是组织治理的组成部分。将风险管理集成到组织中是一个动态的迭代过程，应该根据组织的需求和文化进行定制。
3. 风险识别：识别系统的技术风险、应用风险、管理风险等。包括对资产及其价值、风险源、潜在事件和结果、控制措施和后果的识别。
4. 风险分析：风险分析涉及对不确定性、风险来源、后果、可能性、事件、场景、控制及其有效性的详细考虑。风险分析包括后果评估（包括业务影响评估、个人影响评估和社会影响评估）和可能性评估。
5. 风险应对的目的是选择和实施应对风险的方案。风险应对涉及以下迭代过程：制定和选择风险应对方案；计划和实施风险应对方案；评估应对的有效性；确定剩余风险是否可接受——如果不能接受，采取进一步的应对。
	* 1. 初始阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 确保人工智能系统的概念设计与系统目标保持一致。
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 根据人工智能系统的应用场景制定数据质量的需求；
2. 对数据的来源进行记录和管理；
3. 确保训练数据来源、数据内容、数据标注符合《生成式人工智能服务安全基本要求》等要求。
4. 对数据的版权合法合规性进行审查；
5. 参考GB/T 25000.24—2017评估数据集的质量，数据有标注的，对数据的标注进行审核；
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 评估人工智能系统对社会价值观方面产生的影响和风险；
2. 评估人工智能系统对人的基本权利，包括人身安全、隐私、财产等造成的影响和风险；
3. 评估人工智能系统对人类特定群体的主观或客观偏见。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 调研相关的法律法规，审核人工智能系统的概念设计以确保符合法律法规的最新要求；
2. 建立人工智能系统合规性开发、管理规范以及监督审查机制；
3. 明确定义人工智能系统开发、维护过程中的责任划分，并将其分配给具有适当权限和工具的人员；
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 制定人工智能系统风险评估的流程规范，包括风险识别、风险分析、风险应对、风险监控和风险记录的规范；
2. 组织风险评估小组开展人工智能系统的风险识别，包括系统层面、技术层面、管理层面的风险；
3. 制定风险分析评估量表，对识别的风险进行分析；
4. 确保人工智能系统的概念设计能够在分配的资源、定义的风险范围内实施；
5. 根据风险分析的结果制定风险应对措施，同时制定风险应对措施的实施和监控规范；
6. 制定风险监控规范，对已识别的风险及人工智能系统生命周期中新出现的风险进行监控和处理；
7. 对风险评估各阶段的活动进行记录。
	* 1. 设计开发阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 明确系统开发的硬件环境、软件环境、开发语言要求，并对开发环境的变更制定管理规范；
2. 制定统一的代码规范，并在系统开发过程中对代码进行审核，监督代码规范的执行情况；
3. 对系统开发的软件环境进行安全扫描，包括操作系统、人工智能框架、第三方库等，及时为操作系统、人工智能框架、第三方库更新安全补丁；
4. 开展算法模型机制机理审核，对系统采用的技术、使用的数据、干预的方式等重要因素开展审核，对不合理的部分提出具体可实施的整改要求，并监督有关部门落实；
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 对数据进行清洗、安全过滤，剔除含有风险的数据。
2. 根据需要对数据集进行扩充和增强。
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 可通过技术手段将人类价值观量化并嵌入系统模型，令人工智能系统对齐人类的价值；
2. 采用技术手段降低个人信息在训练数据中的比例和真实性。例如采取“关键词+正则表达式匹配”的方式检测身份证、电话号码等个人信息，并参考GB/T 37964-2019《信息安全技术 个人信息去标识化指南》对个人信息去标识化。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 开展算法模型合规自评估，对算法模型的目的、使用的数据、模型、训练方法、评测过程、干预策略等进行评审。
2. 对系统开发过程中数据的收集、处理、存储、使用等环节进行合规性评估，确保符合《网络安全法》、《个人信息保护法》、《数据安全法》、《个人信息安全规范》、《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理暂行办法》等要求。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 识别系统模型层面的风险，包括但不限于数据安全、模型安全等。制定相应技术和管理手段对系统进行安全加固；
2. 对系统设计、开发、模型训练、模型推理过程进行记录，并制定相关记录的保护措施。
	* 1. 测试验证阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 采用基准测试、机器与人工结合的评测等方式，对系统做全面深入的评测。
2. 评估人工智能系统所依赖的云侧、端侧和边缘侧的硬件设备的故障率，并制定故障发生时的应急管理机制；
3. 根据应用场景评估人工智能系统的功能有效性指标，如准确率、召回率等；
4. 测试模型输入的扰动对输出结果的影响，包括自然条件下的扰动和人为对抗性的扰动，并制定相应的防御方法；
5. 在软件架构和算法模型上制定针对系统故障的预防、监测和追溯机制；
6. 评估训练数据和测试数据的分布差异，并相应制定模型的完善机制；
7. 通过增加训练数据类型、调整模型参数等技术手段重新训练模型，并监控模型可泛化性的变化；
8. 建设人工智能模型/系统的评测体系，在发生模型迭代、服务功能变更时，日常化执行评测，持续全面地跟踪系统性能情况。
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 确保测试数据集标注信息准确、完备，多样，体现真实应用场景，类别覆盖全面，评估测试数据集是否包含需求中的规定的使用场景；
2. 确保测试数据集中不应包含训练数据集中的数据；
3. 测试数据集中宜包含干扰样本、对抗性样本、业务不相关样本等异常样本，用以评估系统的鲁棒性、安全性等方面。
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 评估训练数据的“数据特征”可解释性，如对训练数据进行统计学分析；
2. 评估人工智能模型的可解释性是否符合应用需求；
3. 评估模型预测结果的可解释性，包括全局可解释和单个数据样本可解释；
4. 采用技术手段提升模型的可解释性；
5. 评估人工智能系统的模型处理逻辑的可解释性，分析推理输入、过程与结果之间的关系；
6. 在数据收集和预处理阶段，检查不同类别、不同群体的数据分布，确保不同群体间数据分布没有明显偏向性，同时对数据的标注进行审核，避免标注带有主观偏见；
7. 在算法模型构建和测试阶段，制定公平性的评估方法并将其作为模型的考核指标之一；
8. 在系统运行阶段，评估系统的个性化服务是否符合社会伦理道德。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 评估系统输出的合规性，例如对于违法不当请求、生成虚假个人信息请求、获取个人敏感信息请求应予以拒绝。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 测试验证系统时，实时监测人工智能系统的运行状态，建立系统状态异常时人为控制的介入机制，并测试其有效性；
2. 测试验证系统前对系统进行备份，评估人工智能系统备份系统与原系统的一致性。
	* 1. 部署阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 评估人工智能系统在相同的数据、算法模型、参数配置、操作顺序前提下，在不同运行环境下的输出差异，评估差异是否可接受，分析差异产生的原因，提出改进方案；
2. 评估人工智能系统的需求描述准确性和适当性；
3. 评估人工智能系统设计是否符合需求的规定；
4. 对于超出需求限制的操作，评估人工智能系统的性能下降是否可预测，以及人工智能系统恢复到可预测状态的能力；
5. 制定包括部署准备、代码编译、镜像打包、容器编排等阶段相应的自动化流程，明确每个环节的职责和操作指南，确保部署流程的可重复性和可审查性；
6. 制定部署后的监控机制和紧急故障处理机制，及时发现和解决运行过程中的问题，并针对重要的系统故障实施快速响应和恢复，保证系统的正常运行。
7. 确认系统的硬件、软件和网络环境是否满足系统的要求，例如计算能力、存储容量、网络带宽等。
8. 确认系统的稳定性和兼容性，包括与其他系统的兼容性、不同操作系统和浏览器的兼容性等。
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 制定人工智能系统部署后所涉及数据的备份管理规范，包括备份的数量、时间间隔、地点、方式等，并定期检查数据备份的执行情况。
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 建立人工智能系统的权限管理制度，防止滥用、误用人工智能系统；
2. 实施安全措施，包括网络安全、数据安全、系统安全等，防止黑客攻击、数据泄露等安全事件。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 根据主管部门的管理办法要求，技术提供者和服务提供者需要向主管机关做相关的算法模型备案、向用户提供用户协议、公示算法模型机制机理等。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 评估人工智能系统的硬件可靠性，包括计算、存储等硬件系统的平均失效间隔工作时间、电磁兼容性、环境适应性等；
2. 评估人工智能系统的软件可靠性：包括算法框架、算法库、系统环境的安全性，系统的功能有效性，以及数据集的均衡性、规模性和标注质量，并在系统开发、运行层面制定容错和易恢复的机制；
3. 制定人工智能系统冗余的管理规范，评估备份系统和原系统的一致性和切换逻辑，定期检查备份系统的运行情况。
	* 1. 运行监测阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 测试人工智能系统接到输入至完整返回结果的时间间隔；
2. 测试人工智能系统对实时任务的完成时延受到正在处理的任务数量的影响的程度；
3. 测试人工智能系统保存或暂停特定已有任务的计算，并转而接收、计算并输出另一个任务的延时；
4. 测试一组特定的人工智能系统任务完成延时与数学期望之间的偏离程度。
5. 评估人工智能系统的输出结果是否符合人类的预期和预测；
6. 评估人工智能系统对未知输入数据的反应和处理能力；
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 对人工智能系统运行阶段的新增训练数据进行监控，防止新数据的扰动对后续的训练效果产生不良影响。
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 评估人工智能系统对社会价值方面产生的影响和风险；
2. 评估人工智能系统对人的基本权利，包括人身、隐私、财产等造成的影响和风险；
3. 评估人工智能系统对人类特定群体的主观或客观偏见；
4. 明确人工智能系统相关利益方的责任界定。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 按照相关法律法规，做好对系统使用者的身份核验；
2. 系统运行时收集个人信息的，按照相关法律法规，应当取得个人同意。服务提供者不得收集非必要个人信息，不得非法留存能够识别使用者身份的输入信息和使用记录，不得非法向他人提供使用者的输入信息和使用记录。提供者应当依法及时受理和处理个人关于查阅、复制、更正、补充、删除其个人信息等的请求；
3. 建立系统输出内容的审核机制，对用户输入信息、模型输出信息进行安全审核。对于涉及生物特征，包含特殊含义物体，新闻等生成内容进行更严格的审核，以检测可能的违法不良信息、个人信息、错误价值观与歧视偏见等；
4. 对于生成式人工智能系统，针对底线及原则问题，需要建设标准答案库。对于模型生成不适宜内容的情况，制定应急处理的技术手段；
5. 对于生成式人工智能系统，按照相关法律法规对生成内容添加明确标识。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 制定人工智能系统运行中的日志记录监控、分析、回溯流程，以在系统发生故障时定位故障原因；
2. 评估模型迭代更新的时间周期及模型训练时间；
3. 评估模型迭代后系统的功能有效性、鲁棒性等变化趋势；
4. 评估系统发生故障时，恢复功能所需要的时间；
5. 测试系统的容错能力和抗干扰能力，评估相应预案的容错、容灾机制的稳定性和可靠性；
6. 对系统进行定期的备份和恢复测试，以确保备份数据的可用性和恢复能力。
7. 评估人工智能系统进行不可控迭代、执行未知命令的风险性，确保其输出结果不会威胁到人类和环境的安全。
	* 1. 重新评估阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 根据使用者的反馈，重新评估人工智能系统对利益相关者的生命、健康、财产、数据以及物理环境造成的影响；
2. 根据系统运行情况，重新评估人工智能系统所依赖的云侧、端侧和边缘侧的硬件设备及外接设备的故障率；
3. 根据更新后的数据和模型，重新评估人工智能系统的功能有效性指标，如准确率、召回率等；
4. 根据系统运行中实际输入的数据情况，测试模型输出结果与测试验证阶段时模型输出结果的差异，并制定相应的处理方法；
5. 根据人工智能系统运行情况评估系统的需求描述准确性和适当性，以及设计是否符合需求的规定；
6. 对于实际运行时遇到的超出需求限制的操作，评估人工智能系统的性能下降程度，以及人工智能系统恢复到可预测状态的能力；
7. 评估人工智能系统的实际运行输入是否包含需求中的规定的使用场景；
8. 可以通过用户调查、市场分析等方式获取用户反馈和需求，以进一步提高系统的用户体验和满意度。
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 对人工智能系统的训练数据进行增强，加入系统运行阶段的新数据以提高系统的鲁棒性，加入扰动数据以提高系统的抗扰动效果；
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 重新评估系统的可解释性、公平性、隐私安全性等伦理问题，确保模型、数据的更新迭代符合社会伦理道德；
2. 对系统运行过程中出现的伦理问题进行总结和复盘分析，制定常态化的解决伦理问题方案。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 重新评估系统输出的合规性，确保模型、数据的更新迭代符合法律法规要求；
2. 对系统运行过程中出现的合规问题进行总结和复盘分析，制定常态化的解决合规问题方案。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 重新评估人工智能系统的硬件可靠性，包括计算、存储等硬件系统的平均失效间隔工作时间、电磁兼容性、环境适应性等；
2. 总结系统运行情况及故障原因，并在系统开发、运行层面制定进一步的容错和易恢复的机制；
3. 总结人工智能系统的权限管理应用情况，改进权限管理制度；
4. 考虑技术变革对系统的影响，并制定相应的风险管理措施，以保证系统的运行安全和可靠性；
5. 在重新评估阶段需要考虑系统的运行成本、资源利用率以及运行环境对系统的影响，以制定相应的可持续性策略。
	* 1. 退出使用阶段的治理措施
			1. 决策治理

在决策治理方面，宜包括以下措施：

1. 制定人工智能系统下线的操作流程规范，并监督系统下线的操作过程；
2. 人工智能系统停止运行后，检查系统资源是否还能访问、系统开放的端口是否关闭、系统相关进程是否终止、系统占用的资源是否被释放。以上情况如有异常宜制定相应的操作规范解决。
	* + 1. 数据治理

在数据治理方面，宜包括以下措施：

1. 制定相关规范对系统的代码、程序、模型、数据进行归档或销毁；
2. 在确保数据安全和隐私保护的条件下，对其系统所处理的数据通过硬盘、云端或其他存储设备进行备份和转移，以便后续的分析或者处理；
3. 采用版本控制系统等方式，对工智能系统的参数进行备份和转移，以便在需要时进行恢复。
	* + 1. 伦理治理

在伦理治理方面，宜包括以下措施：

1. 在处理数据和参数时，需要考虑其对社会的影响，例如是否会产生偏见或歧视。需要采取适当的措施来确保公平和公正，并避免对某些群体造成负面影响。
	* + 1. 合规治理

在合规治理方面，宜包括以下措施：

1. 在备份、转移、继承和更新数据和参数时，需要考虑相关法律和合规性要求。例如，需要遵守数据保护法规和隐私法规，以保护用户的个人信息。
	* + 1. 风险治理

在风险治理方面，宜包括以下措施：

1. 总结人工智能系统的生命周期的运行情况，形成总结报告。

参考文献

1. 新一代人工智能发展规划（国发〔2017〕35号）
2. 新一代人工智能治理原则——发展负责任的人工智能（国家新一代人工智能治理专业委员会）
3. 中华人民共和国个人信息保护法（2021年8月20日第十三届全国人民代表大会常务委员会第三十次会议通过）
4. 生成式人工智能服务管理暂行办法
5. 可信赖人工智能标准化白皮书（2022版）
6. ISO/IEC 38507—2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations
7. ISO/IEC 23894—2023 Information technology—Artificial intelligence—Guidance on risk management
8. ISO TR 24368 Information technology—Artificial intelligence—Overview of ethical and societal concerns
9. ISO/IEC 22989 Artificial intelligence—Concepts and terminology
10. 人工智能伦理治理标准化指南（2023版）
11. GB/T 34960.1—2017 信息技术服务 治理 第1部分：通用要求
12. GB/T 34960.5—2018 信息技术服务 治理 第5部分：数据治理规范
13. 生成式人工智能治理与实践白皮书
14.
15. （资料性）
人工智能系统生命周期治理监督检查表

人工智能系统生命周期治理监督检查表见表A.1，表中列出了人工智能系统各生命周期阶段中治理实施的输出产物。

* 1. 人工智能系统生命周期治理监督检查表

| 生命周期阶段 | 决策治理 | 数据治理 | 伦理治理 | 合规治理 | 风险治理 |
| --- | --- | --- | --- | --- | --- |
| 初始阶段 | 人工智能系统初始阶段审查结果 | 数据来源、质量、合规性审查结果 | 人工智能系统伦理影响审查结果 | 人工智能系统法律法规符合性审查结果人工智能系统合规性管理规范 | 人工智能系统风险评估流程规范 |
| 设计开发 | 人工智能系统开发管理规范开发环境安全报告算法模型机制审核结果 | 数据预处理结果报告 | 人工智能系统对齐开发方案 | 算法模型合规自评估报告 | 人工智能系统模型风险记录 |
| 测试验证 | 人工智能系统单元测试报告人工智能系统改进及回归测试报告 | 测试数据集构造方案 | 人工智能系统可解释性、公平性、伦理性测试报告 | 人工智能系统合规性测试报告 | 人工智能系统异常处理报告 |
| 部署 | 人工智能系统集成测试报告人工智能系统紧急故障处理方案 | 数据的备份管理规范 | 人工智能系统的权限管理制度 | 人工智能系统备案材料 | 人工智能系统可靠性管理规范 |
| 运行监测 | 人工智能系统性能测试报告 | 在线数据监控报告 | 人工智能系统在线伦理性报告 | 人工智能系统收集人工信息方案人工智能系统输出内容审核方案生成式人工智能标准答案库 | 人工智能系统日志监控报告人工智能系统在线风险评估报告人工智能系统灾备方案 |
| 重新评估 | 人工智能系统重新评估及改进报告 | 训练数据增强方案 | 系统运行伦理问题总结报告 | 系统运行合规问题总结报告 | 系统故障总结及管理措施报告 |
| 退出使用 | 人工智能系统下线流程规范 | 人工智能系统数据归档方案 | 人工智能系统下线伦理评估报告 | 人工智能系统下线合规评估报告 | 系统运行风险总结报告 |

