ICS 35.020;03.100.70
CCS L 01

中华人民共和国国家标准

GB/T 45081—2024/ISO/IEC 42001:2023

人工智能　管理体系

Artificial intelligence—Managment system

（ISO/IEC 42001:2023，Information technology—Artificial intelligence—Management system，IDT）

2024-11-28 发布　　2024-11-28 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
1　范围……1
2　规范性引用文件……1
3　术语和定义……1
4　组织环境……4
4.1　理解组织及其环境……4
4.2　理解相关方的需求和期望……5
4.3　确定人工智能管理体系的范围……5
4.4　人工智能管理体系……5
5　领导作用……5
5.1　领导作用和承诺……5
5.2　人工智能方针……6
5.3　岗位、职责和权限……6
6　策划……6
6.1　应对风险和机会的措施……6
6.2　人工智能目标及其实现的策划……8
6.3　变更的策划……9
7　支持……9
7.1　资源……9
7.2　能力……9
7.3　意识……9
7.4　沟通……9
7.5　文件化信息……10
8　运行……10
8.1　运行的策划和控制……10
8.2　人工智能风险评估……11
8.3　人工智能风险应对……11
8.4　人工智能系统影响评估……11
9　绩效评价……11
9.1　监视、测量、分析和评价……11
9.2　内部审核……11
9.3　管理评审……12
10　改进……12

10.1 持续改进……12
10.2 不符合和纠正措施……12
附录 A(规范性) 参考控制目标与控制……14
附录 B(规范性) 人工智能控制的实施指南……17
B.1 概述……17
B.2 与人工智能相关的方针……17
B.3 内部组织……18
B.4 人工智能系统的资源……19
B.5 评估人工智能系统的影响……21
B.6 人工智能系统生存周期……24
B.7 人工智能系统的数据……29
B.8 人工智能系统相关方的信息……30
B.9 人工智能系统的使用……32
B.10 第三方及客户关系……34
附录 C(资料性) 潜在的与人工智能相关的组织目标和风险源……36
C.1 概述……36
C.2 目标……36
C.3 风险源……37
附录 D(资料性) 跨领域或跨行业使用人工智能管理体系……38
D.1 概述……38
D.2 人工智能管理体系与其他管理体系的集成……38
参考文献……40

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。

本文件等同采用 ISO/IEC 42001:2023《信息技术　人工智能　管理体系》。

本文件做了下列最小限度的编辑性改动：

——为与现有标准体系保持一致，将标准名称改为《人工智能　管理体系》。

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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、广州赛西标准检测研究院有限公司、蚂蚁科技集团股份有限公司、华为技术有限公司、深圳云天励飞技术股份有限公司、中国科学院软件研究所、阿里云计算有限公司、罗克佳华科技集团股份有限公司、广东粤电信息科技有限公司、中国南方电网有限责任公司超高压输电公司、OPPO 广东移动通信有限公司、深圳市优必选科技股份有限公司、上海燧原科技股份有限公司、北京赛西认证有限责任公司、科大讯飞股份有限公司、浪潮软件科技有限公司、上海商汤智能科技有限公司、上海市人工智能行业协会、上海计算机软件技术开发中心、山东省计算中心(国家超级计算济南中心)、万达信息股份有限公司。

本文件主要起草人：孙宁、黄胜华、杨舟、马万钟、林冠辰、郑文先、孟令中、徐浩、饶雪、李玮、薛学琴、贾一君、沈芷月、张万里、刘张宇、王宁、李根、梁乔玲、梅敬青、蒋燕、孙佩、纪元隆、乔玉平、杨彤晖、钟俊浩、林一伟、蒋慧、吴庚、陈敏刚、高永超、童庆。

人工智能　管理体系

1　范围

本文件为在组织范围内建立、实施、维护和持续改进人工智能管理体系规定了要求并提供了指南。

本文件适用于提供或使用人工智能系统的产品或服务的组织。本文件旨在帮助组织负责任地开发、提供或使用人工智能系统，以实现其目标，并满足适用的法规要求，以及相关方的义务和期望。

本文件适用于各种规模、类型和性质的提供或使用人工智能系统产品或服务的组织。

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。

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

ISO/IEC 22989:2022 界定的以及下列术语和定义适用于本文件。

3.1

组织　organization

为实现目标(3.6)，由职责、权限和相互关系构成自身功能的一个人或一组人。

注1：组织的概念包括但不限于个体经营者、公司、集团公司、商行、企事业单位、监管机构、合伙企业、慈善机构或研究机构，或上述组织的部分或组合，无论是否具有法人资格，公有或私有。

注2：如果组织是大型实体的某个组成部分，那么术语“组织”仅指在人工智能管理体系(3.4)范围内的这个组成部分。

3.2

相关方　interested party

能够影响决策或活动、受决策或活动影响或自认为受决策或活动影响的个人或组织(3.1)。

注：ISO/IEC 22989:2022的5.19中提供了人工智能相关方的概述。

3.3

最高管理者　top management

在最高层指挥和控制组织(3.1)的一个人或一组人。

注1：最高管理者有权在组织内部授权和提供资源。

注2：如果管理体系(3.4)的范围仅覆盖组织的某个组织部分，那么最高管理者是指指挥和控制该部分的一个人或一组人。

3.4

管理体系　management system

组织(3.1)为确立方针(3.5)和目标(3.6)以及实现这些目标的过程(3.8)所形成的相互关联或相互作用的一组要件。

注1：一个管理体系可能针对一个或几个主题。

注2：管理体系要件包括组织的结构、岗位和职责、策划和运行。

3.5

方针　policy

由最高管理者(3.3)正式表述的组织(3.1)的意图和方向。

3.6

目标　objective

要实现的结果。

注1：目标可能是战略的、战术的或运行的。

注2：目标可能涉及不同的主题(如财务、健康和安全、环境)。它们可能存在于不同层面，诸如组织整体层面或项目、产品或过程(3.8)层面。

注3：目标能用其他方式表述，如：预期的结果、宗旨、运行准则，人工智能目标或使用其他有类似含义的词(如：终点或指标)。

注4：在人工智能管理体系(3.4)中，组织(3.1)设定的人工智能目标与人工智能方针(3.5)保持一致，以实现特定的结果。

3.7

风险　risk

不确定性的影响。

注1：影响是对预期的偏离——正面的或负面的。

注2：不确定性是一种状态，是指对某个事件、事件的后果或可能性缺乏甚至部分缺乏相关信息、理解或知识。

注3：通常，风险以潜在事件(见GB/T 23694—2013中4.5.1.3的定义)和后果(见GB/T 23694—2013中4.6.1.3的定义)或二者的组合来描述其特性。

注4：通常，风险以某个事件的后果(包括情况的变化)及其发生的可能性(见GB/T 23694—2013中4.6.1.1的定义)的组合来表述。

3.8

过程　process

使用或转化输入以实现结果的一组相互关联或相互作用的活动。

注：某个过程的结果是称为输出，还是称为产品或服务，取决于相关语境。

3.9

能力　competence

应用知识和技能实现预期结果的本领。

3.10

文件化信息　documented information

组织(3.1)需要控制和维护的信息及其载体。

注1：文件化信息能够以任何形式和载体存在，且来源不限。

注2：文件化信息可能涉及：

——管理体系(3.4)，包括相关过程(3.8)；

——为组织运行而创建的信息(文件)；

——实现的结果的证据(记录)。

3.11

绩效　performance

可测量的结果。

注1：绩效可能涉及定量的或定性的结果。

注2：绩效可能与活动、过程(3.8)、产品、服务、体系或组织(3.1)的管理有关。

注3：在本文件中，绩效既指使用人工智能系统取得的结果，也指与人工智能管理体系(3.4)相关的结果。该术语的正确解释从其使用的环境中得出。

3.12

持续改进　continual improvement

提高绩效(3.11)的循环活动。

3.13

有效性　effectiveness

完成策划的活动和实现策划的结果的程度。

3.14

要求/需求　requirement

规定的、不言而喻的或有义务履行的需求或期望。

注1：不言而喻的或有义务履行的需要或期望是指需求。其中，“不言而喻”是指组织(3.1)和相关方(3.2)的惯例或一般做法，不言而喻的需求或期望是不用说就明白的。

注2：规定的需要或期望是指要求，也就是符合GB/T 1.1中定义的要求，即表达声明符合该文件需要满足的客观可证实的准则。

3.15

符合　conformity

满足要求(3.14)。

3.16

不符合　nonconformity

未满足要求(3.14)。

3.17

纠正措施　corrective action

为消除不符合(3.16)的原因并防止再次发生而采取的措施。

3.18

审核　audit

获取审核证据并对其进行客观评价，以确定审核准则满足程度所进行的系统的、独立的过程(3.8)。

注1：审核可能为内部(第一方)审核或外部(第二方或第三方)审核，也可能为多体系审核(合并两个或多个主题)。

注2：内部审核由组织(3.1)自行实施或代表组织的外部机构实施。

注3：“审核证据”和“审核准则”的定义见ISO 19011。

3.19

测量　measurement

确定数值的过程(3.8)。

3.20

监视　monitoring

确定体系、过程(3.8)或活动的状态。

注：确定状态可能需要检查、监督或严格观察。

3.21

控制　control

〈风险〉保持和/或改变风险(3.7)的措施。

注1：控制包括但不限于保持和/或改变风险的任何过程、策略、措施、操作或其他条件和/或行动。

注2：控制并非总能发挥预期或假定的改变效果。

[来源：GB/T 24353—2022，3.8，有修改]

3.22

治理层　governing body

对组织的绩效和符合性负责的一个人或一组人。

注1：并非所有组织，特别是小型组织，都有一个独立于最高管理者的治理层。

注2：治理层包括但不限于董事会、董事会委员会、监事会、受托人或监督人。

[来源：ISO/IEC 38500:2015，2.9，有修改]

3.23

信息安全　information security

对信息的保密性、完整性和可用性的保全。

注：另外，还能包括诸如真实性、可问责性、抗抵赖性和可靠性等其他特性。

[来源：GB/T 29246　2023，3.28]

3.24

人工智能系统影响评估　AI system impact assessment

由开发、提供或使用人工智能产品或服务的组织识别、评估和解决对个人和(或)群体和社会的影响的正式的、文件化的过程。

3.25

数据质量　data quality

满足组织在特定情况下数据要求的数据特征。

[来源：ISO/IEC 5259-1:2024，3.4]

3.26

适用性声明　statement of applicability

所有必要控制(3.21)，以及包括或排除控制的理由的文件。

注1：组织可能不需要附录A中列出的所有控制，甚至可能超出附录A中所列的控制，并由组织自己制定额外的控制。

注2：所有已识别的风险，组织按本文件的要求形成文件。所有已识别的风险和为解决这些风险而制定的风险管理措施(控制)反映在适用性声明中。

4　组织环境

4.1　理解组织及其环境

组织应确定与其宗旨相关的，并影响其实现人工智能管理体系预期结果的能力的内部和外部事项。

组织应确定气候变化是否是一个相关因素。

组织应顾及组织开发、提供或使用人工智能系统的预期目的。组织应确定其在人工智能系统中的角色。

注1：组织确定其相对于人工智能系统的角色有助于理解组织及其环境。这些角色包括但不限于以下一种或多种：

——人工智能提供方，包括人工智能平台提供方、人工智能产品或服务提供方；

——人工智能生产方，包括人工智能开发方、人工智能设计方、人工智能运营方、人工智能测试和评估机构、人工智能部署方、人工智能人因研究机构、领域专家、人工智能影响评估机构、采购方、人工智能治理和监督专业机构；

——人工智能客户，包括人工智能用户；

——人工智能合作伙伴，包括人工智能系统集成商和数据提供商；

——人工智能主体，包括数据主体和其他主体；

——相关监管机构，包括方针制定者和监管方。

ISO/IEC 22989:2022 提供了这些角色的详细描述。此外,角色类型及其与人工智能系统生存周期的关系也在美国国家标准与技术研究院(NIST)发布的《人工智能风险管理框架》中进行了描述。组织的角色能确定本文件中的要求和控制的适用性和适用性程度。

注2:根据本条要解决的外部和内部因素可能因组织的角色和管辖权及其对实现人工智能管理体系预期结果的能力的影响而有所不同。这些包括但不限于如下内容。

a) 外部环境相关的考虑,如:
   1) 适用的法律要求,包括禁止使用人工智能;
   2) 监管机构的方针、指南和决定对人工智能系统开发和使用中法律要求的解释或执行产生影响;
   3) 与人工智能系统的预期目的和使用相关的激励或后果;
   4) 人工智能开发和使用方面的文化、传统、价值观、规范和伦理;
   5) 使用人工智能系统的新产品和服务的竞争格局和趋势。
b) 内部环境相关的考虑,如:
   1) 组织环境、治理、目标(见 6.2)、方针和程序;
   2) 合同义务;
   3) 拟开发或使用人工智能系统的预期目的。

注3:角色的确定能通过与组织处理的数据类别相关的义务来确定[例如,在处理个人可识别信息时,个人可识别信息处理者或个人可识别信息控制者]。有关个人可识别信息和相关角色,见ISO/IEC 29100。角色也能通过特定于人工智能系统的法律要求来了解。

4.2 理解相关方的需求和期望

组织应确定:

——与人工智能管理体系有关的相关方;

——这些相关方的有关需求;

——哪些需求将通过人工智能管理体系予以解决。

注:相关方能对气候变化提出需求。

4.3 确定人工智能管理体系的范围

组织应确定人工智能管理体系的边界和适用性,以确定其范围。

组织应根据以下内容确定人工智能管理体系的范围:

——4.1 提及的内部和外部因素;

——4.2 提及的需求。

范围应作为文件化信息可获取。

人工智能管理体系的范围应根据本文件对人工智能管理体系、领导、策划、支持、运行、绩效、评价、改进、控制和目标的要求确定组织的活动。

4.4 人工智能管理体系

组织应按本文件的要求,建立、实施、维护和持续改进人工智能管理体系,包括所需的过程及其相互作用。

5 领导作用

5.1 领导作用和承诺

最高管理者应通过以下方面证实其对人工智能管理体系的领导作用和承诺:

——确保制定人工智能方针(见 5.2)和人工智能目标(见 6.2),并与组织的战略方向一致;

——确保将人工智能管理体系要求融入组织的业务过程；
——确保人工智能管理体系所需的资源可获取；
——就有效的人工智能管理的重要性以及符合人工智能管理体系要求的重要性进行沟通；
——确保人工智能管理体系实现其预期结果；
——指导和支持人员为人工智能管理体系的有效性作出贡献；
——促进持续改进；
——支持其他相关岗位在职责范围内证实其领导作用。

注1：本文件中提及的“业务”能作广义解释，指那些与组织存在目的相关的核心活动。

注2：在组织内部建立、鼓励和构建一种文化，以负责任的方式使用、开发和治理人工智能系统，这是最高管理层承诺和领导力的重要体现。确保意识到并遵守这种负责任的方法，并通过领导支持有助于人工智能管理体系的成功。

5.2 人工智能方针

最高管理者应确立人工智能方针，该方针：

a） 适合于组织的宗旨；
b） 为设定人工智能目标提供框架（见6.2）；
c） 包括满足适用需求的承诺；
d） 包括持续改进人工智能管理体系的承诺。

人工智能方针应：

——作为文件化信息可获取；
——参考其他相关的组织方针；
——在组织内予以沟通；
——视情况，可被相关方获取。

表A.1中A.2提供了制定人工智能方针的控制目标和控制。这些控制的实施指南见附录B的B.2。

注：ISO/IEC 38507提供了组织在制定人工智能方针时的注意事项。

5.3 岗位、职责和权限

最高管理者应确保在组织内部分配并沟通相关岗位的职责和权限。

最高管理者应分配职责和权限，以便：

a） 确保人工智能管理体系符合本文件的要求；
b） 向最高管理者报告人工智能管理体系的绩效。

注：表A.1中A.3.2提供了规定和分配岗位与职责的控制。B.3.2提供了该控制的实施指南。

6 策划

6.1 应对风险和机会的措施

6.1.1 通则

在策划人工智能管理体系时，组织应根据4.1中提及的事项和4.2中提及的需求，并确定需要应对的风险和机会以便：

——确保人工智能管理体系能够实现预期结果；
——预防或减少不利影响；
——实现持续改进。

组织应建立和维护人工智能风险准则，以支持以下工作：

——区分可接受与不可接受的风险；

——进行人工智能风险评估；

——实施人工智能风险应对；

——评估人工智能风险的影响。

注1：ISO/IEC 38507和ISO/IEC 23894中提供了确定组织愿意追求或保留的风险数量和类型的考虑因素。

组织应根据以下因素确定风险和机会：

——人工智能系统的领域和应用环境；

——预期用途；

——4.1中描述的外部和内部环境。

注2：在人工智能管理体系的范围内能考虑不止一个人工智能系统。在这种情况下，针对每个人工智能系统或人工智能系统组确定机会和用途。

组织应计划：

a） 应对这些风险和机会的措施；

b） 如何做：

1） 将这些措施纳入其人工智能管理体系过程并加以实施；

2） 评价这些措施的有效性。

组织应保留为识别和应对人工智能风险和机会而采取的措施的文件化信息。

注3：关于如何为开发、提供或使用人工智能产品、系统和服务的组织实施风险管理的指导见ISO/IEC 23894。

注4：组织及其活动的环境会对组织的风险管理活动产生影响。

注5：不同部门和行业对风险的规定以及风险管理的设想可能有所不同。3.7中对风险的规范性规定允许对风险有一个广泛的认识，以适应任何部门，如附录D中D.1中提到的部门。在任何情况下，作为风险评估的一部分，组织的职责是首先采用适合其环境的风险观。这能包括通过人工智能系统为之开发和使用的部门所使用的规定来看待风险，见ISO/IEC Guide 51中的规定。

6.1.2 人工智能风险评估

组织应规定并建立人工智能风险评估过程，该过程应：

a） 参考并符合人工智能方针（见5.2）和人工智能目标（见6.2）；

注：在评估作为6.1.2d)1)中的后果时，组织能利用6.1.4所述的人工智能系统影响评估。

b） 在策划上使重复的人工智能风险评估能够产生一致、有效和可比较的结果；

c） 识别有助于或妨碍实现人工智能目标的风险；

d） 分析人工智能风险，以便：

1） 评估如果确定的风险成为现实，将对组织、个人和社会造成的潜在后果；

2） 酌情评估已识别风险的现实可能性；

3） 确定风险等级。

e） 评价人工智能风险，以便：

1） 将风险分析结果与风险准则（见6.1.1）进行比较；

2） 对评估的风险进行优先排序，以便进行风险应对。

组织应保留有关人工智能风险评估过程的文件化信息。

6.1.3 人工智能风险应对

考虑到风险评估结果，组织应确定人工智能风险应对过程，以便：

a） 选择适当的人工智能风险应对方案；

b） 确定实施所选人工智能风险应对方案所必需的所有控制，并将这些控制与附录 A 中的控制进行比较，以核实没有遗漏任何必要的控制；

注1：附录A提供了实现组织目标和处理与人工智能系统的设计和使用有关的风险的参考控制。

c） 考虑附录 A 中与实施人工智能风险应对方案相关的控制；

d） 确定除了附录 A 中的控制外，是否还需要其他控制，以实施所有风险应对备选方案；

e） 参考附录 B，了解 b)和 c)中确定的控制的实施指南；

注2：控制目标隐含在所选择的控制中。组织能根据需要选择附录A中列出的控制目标和控制。附录A中的控制并非详尽无遗，可能还需要额外的控制目标和控制。如果需要附录A以外的不同或额外控制，组织能策划此类控制或从现有来源获取。如适用，人工智能风险管理可纳入其他管理系统。

f） 编制一份适用性声明，其中包含必要的控制[见 b)、c)和 d)]，并说明纳入和排除控制的理由；排除的理由能包括风险评估认为不需要的控制，以及适用的外部要求不需要(或属于例外情况)的控制；

注3：组织能提供文件证明排除一般或特定人工智能系统控制目标的理由，不论是附录A中列出的还是组织自己制定的。

g） 制定人工智能风险应对办法。

获得指定管理层对人工智能风险应对计划和接受残余人工智能风险的批准。必要的控制应：

——与 6.2 中的目标相一致；

——作为文件化信息可获取；

——在组织内予以沟通；

——视情况，可被相关方获取。

组织应保留有关人工智能风险应对过程的文件化信息。

6.1.4 人工智能系统影响评估

组织应制定一个过程，用于评估开发、提供或使用人工智能系统可能对个人和(或)群体和社会造成的潜在影响。

人工智能系统影响评估应确定人工智能系统的部署、预期使用和可预见的滥用对个人和(或)群体和社会造成的潜在影响。

影响评估应顾及人工智能系统的具体技术和社会环境以便人工智能系统可在管辖范围内部署和适用。

人工智能系统影响评估的结果应记录在案。在适当情况下，能将人工智能系统影响评估的结果提供给组织规定的相关方。

组织应在风险评估中考虑人工智能系统影响评估结果(见 6.1.2)。表 A.1 中 A.5 提供了评估人工智能系统影响的控制。

注：在某些环境下(如对安全或隐私至关重要的人工智能系统)，组织能要求进行特定学科的人工智能系统影响评估(如物理安全、隐私或信息安全影响)，作为组织整体风险管理活动的一部分。

6.2 人工智能目标及其实现的策划

组织应在相关职能和层级上确立人工智能目标。

人工智能目标应：

a） 与人工智能方针一致(见 5.2)；

b） 可测量(如果可行)；

c） 体现适用的需求；

d） 予以监视；

e） 予以沟通；

f） 视情况予以更新；

g） 作为文件化信息可获取。

策划如何实现人工智能目标时，组织应确定：

——要做什么；

——需要什么资源；

——由谁负责；

——何时完成；

——如何评价结果。

注：附录C提供了与风险管理有关的人工智能目标的非排他性清单。表A.1中A.6.1和A.9.3提供了确定负责任地开发和使用人工智能系统的控制目标和控制。B.6.1和B.9.3提供了这些控制的实施指南定状态可能需要检查、监督或严格观察。

6.3 变更的策划

当组织确定需要变更人工智能管理体系时，应对这些变更的实施进行策划。

7 支持

7.1 资源

为建立、实施、保持和持续改进人工智能管理体系，组织应确定并提供所需的资源。

注：人工智能资源的控制目标和控制见表A.1中A.4。B.4提供了这些控制的实施指南。

7.2 能力

组织应：

——确定在其控制下工作、影响人工智能绩效的人员所需的能力；

——确保这些人员在适当的教育、培训或经验的基础上胜任工作；

——适用时，采取措施获得所需的能力，并评价所采取措施的有效性。

适当的文件化信息应作为能力证据可获取。

注1：B.4.6提供了关于人力资源的实施指南，包括考虑所需的专业知识。

注2：适用的措施可能包括，例如：向现有员工提供培训、指导或重新分配工作；聘用或劳务雇用能够胜任的人员。

7.3 意识

在组织控制下工作的人员应知道：

——人工智能方针(见5.2)；

——他们对人工智能管理体系有效性的贡献，包括改善人工智能绩效带来的效益；

——不符合人工智能管理体系要求的后果。

7.4 沟通

组织应确定与人工智能管理体系有关的内部和外部沟通，包括：

——沟通什么；

——何时沟通；

——与谁沟通；

——如何沟通。

7.5 文件化信息

7.5.1 通则

组织的人工智能管理体系应包括：

a） 本文件要求的文件化信息；

b） 组织确定的，对于人工智能管理体系有效性所必需的文件化信息。

注：不同组织的人工智能管理体系文件化信息的程度可能不同，取决于：

——组织的规模及其活动、过程、产品和服务的类型；

——过程及其相互作用的复杂度；

——人员的能力。

7.5.2 文件化信息的创建和更新

在创建和更新文件化信息时，组织应确保适当的：

——标记和说明（例如，标题、日期、作者或文件编号）；

——形式（例如，语言文字、软件版本、图形）和载体（例如，纸质的、电子的）；

——针对适宜性和充分性的评审和批准。

7.5.3 文件化信息的控制

应控制人工智能管理体系和本文件要求的文件化信息，以确保其：

a） 在需要的场所和时间均可获得并适于使用；

b） 得到充分保护（例如，防止泄密、不当使用或完整性受损）。

为了控制文件化信息，组织应开展以下适用的活动：

——分发、访问、检索和使用；

——存储和防护，包括保持易读性；

——对变更的控制（例如，版本控制）；

——保留和处置。

对于组织确定的，策划和运行人工智能管理体系必要的、来自外部的文件化信息，应视情况进行识别，并予以控制。

注：访问可能意味着只允许查看文件化信息的权限，或者允许并授权查看和变更文件化信息的权限。

8 运行

8.1 运行的策划和控制

为满足要求和实施第 6 章确定的措施，组织应通过以下方式策划、实施和控制所需的过程：

——对过程确立准则；

——按准则对过程实施控制。

组织应实施根据 6.1.3 确定的与人工智能管理体系运行相关的控制（如与人工智能系统开发和使用生存周期相关的控制）。

应监视这些控制的有效性，如果未能达到预期结果，应顾及采取纠正措施。附录 A 列出了参考控制，附录 B 提供了控制的实施指南。

文件化信息应根据必要程度可获取，以便确认过程已按策划得到实施。

组织应控制策划的变更，并评审非预期变更的后果，必要时采取措施减轻不利影响。

组织应确保与人工智能管理体系相关的,由外部提供的产品、过程或服务受控。

8.2 人工智能风险评估

组织应按计划的时间间隔或在提出重大变更时,根据 6.1.2 进行人工智能风险评估。

组织应保留所有人工智能风险评估结果的文件化信息。

8.3 人工智能风险应对

组织应按 6.1.3 实施人工智能风险应对计划,并验证其有效性。

当风险评估识别出需要处理的新风险时,应对这些风险执行 6.1.3 的风险应对过程。

当风险应对计划确定的风险应对方案无效时,应按 6.1.3 的风险应对过程对这些风险应对方案进行评审和重新验证,并更新风险应对计划。

组织应保留所有人工智能风险应对结果的文件化信息。

8.4 人工智能系统影响评估

组织应按计划的时间间隔或在提出重大变更时,根据 6.1.4 进行人工智能系统影响评估。

组织应保留所有人工智能系统影响评估结果的文件化信息。

9 绩效评价

9.1 监视、测量、分析和评价

组织应确定:

——需要监视和测量什么;

——适用的监视、测量、分析和评价方法,以确保有效的结果;

——何时实施监视和测量;

——何时对监视和测量的结果进行分析和评价。

文件化信息应作为可获取的结果证据。

组织应评价人工智能管理体系的绩效和有效性。

9.2 内部审核

9.2.1 通则

组织应在策划的时间间隔内实施内部审核,以便为人工智能管理体系提供以下信息。

a) 是否符合:

1) 组织自身对人工智能管理体系的要求;

2) 本文件的要求。

b) 是否得到了有效地实施和维护。

9.2.2 内部审核程序

组织应策划、确立、实施和维护审核方案,包括频次、方法、职责、策划要求和报告。

组织应根据相关过程的重要性和以往审核的结果,确立内部审核方案。

组织应:

a) 界定每次审核的目标、准则和范围;

b) 选择审核员并实施审核,以确保审核过程的客观性和公正性;

c） 确保向相关管理者报告审核结果。

文件化信息应作为实施审核方案和审核结果的证据可获取。

9.3 管理评审

9.3.1 通则

最高管理者应在策划的时间间隔内对组织的人工智能管理体系进行评审，以确保人工智能管理体系持续的适宜性、充分性和有效性。

9.3.2 管理评审输入

管理评审应包括：

a） 以往管理评审所采取措施的状况。

b） 与人工智能管理体系有关的外部和内部事项的变化。

c） 与人工智能管理体系有关的相关方的需要和期望的变化。

d） 关于人工智能管理体系绩效的信息，包括以下方面的趋势：

1） 不符合与纠正措施；

2） 监视和测量的结果；

3） 审核结果。

e） 持续改进的机会。

9.3.3 管理评审结果

管理评审的结果应包括持续改进的机会，以及变更人工智能管理体系的任何需要的决定。

文件化信息应作为管理评审结果证据可获取。

10 改进

10.1 持续改进

组织应持续改进人工智能管理体系的适宜性、充分性和有效性。

10.2 不符合和纠正措施

发生不符合时，组织应：

a） 对不符合做出反应，并且如适用：

1） 采取措施和纠正措施；

2） 处置后果。

b） 通过以下活动评价采取措施的需要，以消除产生不符合的原因，避免其再次发生或在其他地方发生：

1） 评审不符合；

2） 确定产生不符合的原因；

3） 确定是否存在或可能发生类似的不符合。

c） 实施任何所需的措施。

d） 评审所采取的任何纠正措施的有效性。

e） 如必要，变更人工智能管理体系。

纠正措施应与不符合产生的影响相适应。

文件化信息应作为以下事项的证据可获取：

——不符合的性质和所采取的任何后续措施；

——任何纠正措施的结果。

附 录 A
（规范性）
参考控制目标与控制

A.1 概述

表 A.1 详细列出的控制为组织提供了一个参考，以实现组织目标和应对与人工智能系统的设计和运行有关的风险。表 A.1 中列出的所有控制目标和控制并非都必须使用，组织能设计和实施自己的控制（见 6.1.3）。

附录 B 提供了表 A.1 列出的所有控制的实施指南。

表 A.1 控制目标和控制

A.2 与人工智能相关的方针		
目标：根据业务需求，为人工智能系统提供管理指导和支持		
	主题	控制
A.2.2	人工智能方针	组织应记录有关开发或使用人工智能系统的方针
A.2.3	与其他组织方针保持一致	组织应确定在人工智能系统方面的目标是否可能影响或适用于其他方针
A.2.4	人工智能方针的评审	人工智能方针应按计划的时间间隔进行评审，或根据需要进行额外评审，以确保其持续的适宜性、充分性和有效性
A.3 内部组织		
目标：在组织内部建立问责制，坚持以负责任的方式实施、运行和管理人工智能系统		
	主题	控制
A.3.2	人工智能的岗位和职责	应根据组织的需要确定和分配人工智能的岗位和职责
A.3.3	报告关切	组织应制定并实施一个过程，以报告与人工智能系统全生存周期有关的组织中的岗位的关切
A.4 人工智能系统的资源		
目标：确保组织考虑人工智能系统的资源（包括人工智能系统组件和资产）进行核算，以充分理解并应对风险和影响		
	主题	控制
A.4.2	资源记录	组织应识别并记录在人工智能系统生存周期特定阶段开展的活动，以及与组织相关的其他人工智能相关活动所需的相关资源
A.4.3	数据资源	作为资源识别的一部分，组织应记录有关人工智能系统所使用的数据资源的信息
A.4.4	工具资源	作为资源识别的一部分，组织应记录有关人工智能系统所使用的工具资源的信息
A.4.5	系统和计算资源	作为资源识别的一部分，组织应记录有关人工智能系统所使用的系统和计算资源的信息
A.4.6	人力资源	作为资源识别的一部分，组织应记录开发、部署、运行、变更管理、维护、转让和退役以及验证和集成人工智能系统所使用的人力资源及其能力的相关信息

表 A.1　控制目标和控制（续）

A.5　评估人工智能系统的影响		
目标：评估人工智能系统在全生存周期内对个人和（或）群体，以及社会的影响		
	主题	控制
A.5.2	人工智能系统影响评估过程	组织应建立一个过程，以评估人工智能系统在其全生存周期内可能对个人和（或）群体，以及对社会造成的潜在后果
A.5.3	人工智能系统影响评估记录	组织应记录人工智能系统影响评估的结果，并在规定期限内保留结果
A.5.4	评估人工智能系统对个人或群体的影响	组织应评估并记录人工智能系统在全系统生存周期内对个人或群体的潜在影响
A.5.5	评估人工智能系统的社会影响	组织应评估并记录其人工智能系统在全生存周期中可能产生的社会影响
A.6　人工智能系统生存周期		
A.6.1　人工智能系统开发管理的指南		
目标：确保组织识别和记录目标并实施负责任的人工智能系统设计和开发过程		
	主题	控制
A.6.1.2	负责任地开发人工智能系统的目标	组织应识别并记录目标，以指导负责任地开发人工智能系统，并在开发生存周期中考虑这些目标，并纳入实现这些目标的措施
A.6.1.3	负责任地设计和开发人工智能系统的过程	组织应制定并记录负责任地设计和开发人工智能系统的具体过程
A.6.2　人工智能系统生存周期		
目标：规定人工智能系统生存周期各阶段的准则和要求		
	主题	控制
A.6.2.2	人工智能系统的要求和规范	组织应规定并记录对新的人工智能系统或现有系统的重大改进的要求
A.6.2.3	人工智能系统的设计和开发记录	组织应根据组织目标、文件化的要求和规范准则，记录人工智能系统的设计和开发
A.6.2.4	人工智能系统的验证和确认	组织应规定并记录人工智能系统的验证和确认措施，并规定其使用准则
A.6.2.5	人工智能系统的部署	组织应记录部署计划，并确保在部署前满足适当的要求
A.6.2.6	人工智能系统的运行和监视	组织应规定并记录人工智能系统持续运行的必要要素。至少宜包括系统和绩效监视、维修、更新和支持
A.6.2.7	人工智能系统的技术文件化信息	组织应确定用户、合作伙伴、监管机构等各类相关方需要的人工智能系统技术记录，并以适当形式向他们提供技术文件化信息
A.6.2.8	人工智能系统的事件日志记录	组织应确定在人工智能系统生存周期的哪些阶段宜启用事件日志记录，但至少应在人工智能系统使用时启用

表 A.1　控制目标和控制（续）

A.7　人工智能系统的数据		
目标：确保组织理解人工智能系统中的数据在人工智能系统生存周期中的应用、开发、提供或使用中的作用和影响		
	主题	控制
A.7.2	用于开发和增强人工智能系统的数据	组织应规定、记录和实施与人工智能系统开发相关的数据管理过程
A.7.3	数据采集	组织应确定并记录人工智能系统所用数据的采集和选择细节
A.7.4	人工智能系统的数据质量	组织应规定和记录数据质量要求，并确保用于开发和运行人工智能系统的数据符合这些要求
A.7.5	数据来源	组织应规定并记录一个过程，用于在数据和人工智能系统的生存周期中记录其人工智能系统中使用的数据的来源
A.7.6	数据准备	组织应规定并记录其选择数据准备的准则和要使用的数据准备的方法
A.8　人工智能系统相关方的信息		
目标：确保相关方掌握必要的信息，以了解和评估风险及其影响（正面和负面）		
	主题	控制
A.8.2	为用户提供的系统文件和信息	组织应确定并向人工智能系统用户提供必要的信息
A.8.3	外部报告	组织应具有为相关方报告人工智能系统负面影响的能力
A.8.4	事故通报	组织应确定并记录向人工智能系统用户通报事故的计划
A.8.5	向相关方提供信息	组织应确定并记录向相关方报告人工智能系统信息的义务
A.9　人工智能系统的使用		
目标：确保组织负责任地并按组织的方针使用人工智能系统		
	主题	控制
A.9.2	负责任地使用人工智能系统的过程	组织应规定并记录负责任地使用人工智能系统的过程
A.9.3	负责任地使用人工智能系统的目标	组织应识别并记录用于指导负责任地使用人工智能系统的目标
A.9.4	人工智能系统的预期用途	组织应确保按人工智能系统及其附带文件的预期用途使用人工智能系统
A.10　第三方及客户关系		
目标：确保组织了解其责任并可问责，并在人工智能系统生存周期的任何阶段涉及第三方时适当分摊风险		
	主题	控制
A.10.2	分配职责	组织应确保在其人工智能系统生存周期内的责任被分配在组织、其合作伙伴、供应商、客户和第三方之间
A.10.3	供应商	组织应建立一个过程，确保其对供应商提供的服务、产品或材料的使用符合组织负责任地开发和使用人工智能系统的方针
A.10.4	客户	组织应确保其开发和使用人工智能系统的方法负责任并考虑到客户的期望和需求

附 录 B
（规范性）
人工智能控制的实施指南

B.1 概述

本附录中记录的实施指南与表 A.1 中列出的控制有关。它提供了支持实施表 A.1 中列出的控制和实现控制目标的信息，但组织不必在适用性声明（见 6.1.3）中记录或证明纳入或不纳入实施指南。

实施指南并非适用于或充分适用于所有情况，也并非总能满足组织的具体控制要求。组织能根据其具体要求和风险应对需要，扩展或修改实施指南，或自行规定控制的实施。

本附录作为在本文件规定的人工智能管理体系中确定和实施人工智能风险应对控制的指南。除本附录所列控制外，还能确定其他组织上和技术上的控制（见 6.1.3 中的人工智能系统管理风险应对）。本附录能视作制定组织特定控制实施的起点。

B.2 与人工智能相关的方针

B.2.1 目标

根据业务需求，为人工智能系统提供管理指导和支持。

B.2.2 人工智能方针

控制

组织宜记录有关开发或使用人工智能系统的方针。

实施指南

人工智能方针宜包括：

——业务战略；

——组织的价值观和文化，以及组织愿意承担或保留的风险程度；

——人工智能系统带来的风险程度；

——法律要求，包括合同；

——组织的风险环境；

——对相关方的影响（见 6.1.4）。

人工智能方针宜包括（除 5.2 规定的要求外）：

——指导组织与人工智能有关的所有活动的原则；

——处理偏离方针和例外情况的过程。

必要时，人工智能方针宜考虑特定主题方面，以提供更多指导或提供与涉及这些方面的其他方针的交叉参考。此类主题的示例包括：

——人工智能资源和资产；

——人工智能系统影响评估（见 6.1.4）；

——人工智能系统开发。

相关方针宜指导人工智能系统的开发、购买、运行和使用。

B.2.3 与其他组织方针保持一致

控制

组织宜确定在人工智能系统方面的目标是否可能影响或适用于其他方针。

实施指南

人工智能与多个领域有交叉，包括质量、信息安全、物理安全和隐私。组织宜考虑进行全面分析，以确定当前方针是否以及在哪些方面存在必然交叉，并在需要更新时更新这些方针，或在人工智能方针中纳入相关规定。

其他信息

治理机构代表组织制定的方针宜为人工智能方针提供参考。ISO/IEC 38507 为治理管理机构的成员提供了指导，以便在人工智能系统的全生存周期内启用和治理人工智能系统。

B.2.4 人工智能方针的评审

控制

人工智能方针宜按计划的时间间隔进行评审，或根据需要进行额外评审，以确保其持续的适宜性、充分性和有效性。

实施指南

宜由管理层批准的一个岗位负责人工智能方针或其组成部分的制定、评审和评价。评审宜包括评估改进组织管理人工智能系统的方针和方法的机会，以应对组织环境、业务情况、法律条件或技术环境的变化。

人工智能方针评审宜将管理评审结果作为考虑因素。

B.3 内部组织

B.3.1 目标

在组织内部建立问责制，坚持以负责任的方式实施、运行和管理人工智能系统。

B.3.2 人工智能的岗位和职责

控制

宜根据组织的需求确定和分配人工智能的岗位和职责。

实施指南

规定岗位和职责对于确保全组织中与人工智能系统全生存周期有关的岗位可问责至关重要。在分配岗位和责任时，组织宜考虑人工智能政策、人工智能目标和已识别的风险，以确保涵盖所有相关领域。组织能对岗位和职责分配进行优先处理。需要明确岗位和职责的领域包括：

——风险管理；
——人工智能系统影响评估；
——资产和资源管理；
——信息安全；
——物理安全；
——隐私；
——开发；
——绩效；
——人员监督；
——供应商关系；
——证明有一直符合法律要求的能力；
——数据质量管理(全生存周期)。

各种岗位的职责宜界定到适合个人履行其职责的程度。

B.3.3 报告关切

控制

组织宜制定并实施一套过程，以报告与人工智能系统全生存周期有关的组织中的岗位的关切。

实施指南

报告机制宜具备以下功能：

a) 保密或匿名或两者兼有的选项；

b) 可供并鼓励受雇人员和合同人员使用；

c) 配备合格的工作人员；

d) 为 c)中提到的人员规定适当的调查权和决议权；

e) 提供及时向管理层报告和上报的机制；

f) 为报告和调查相关人员提供有效保护，使其免遭报复(如允许匿名和保密报告)；

g) 根据 4.4 和 e)(适用时)提供报告，同时保持 a)中的保密性和匿名性，并尊重一般的商业机密考虑因素；

h) 提供适当时限内的反馈机制。

注：组织能利用现有的报告机制作为该过程的一部分。

其他信息

除本条提供的实施指南外，组织还宜进一步考虑 ISO 37002。

B.4 人工智能系统的资源

B.4.1 目标

确保组织考虑对人工智能系统的资源(包括人工智能系统组件和资产)进行核算，以充分了解和应对风险和影响。

B.4.2 资源记录

控制

组织宜识别并记录在人工智能系统生存周期特定阶段开展的活动，以及与组织相关的其他人工智能相关活动所需的相关资源。

实施指南

记录人工智能系统的资源对于了解风险以及人工智能系统对个人和(或)群体，以及对社会的潜在影响(包括正面和负面影响)至关重要。记录此类资源(利用数据流图或系统架构图)能为人工智能系统影响评估(见 B.5)提供信息。

资源包括但不限于：

——人工智能系统组件；

——数据资源，即在人工智能系统生存周期内任何阶段使用的数据；

——工具资源(如人工智能算法、模型或工具)；

——系统和计算资源(如开发和运行人工智能模型的硬件、数据存储和工具资源)；

——人力资源，即与组织在人工智能系统生存周期中的岗位相关的、具有必要专业知识的人员(如开发、销售、培训、运行和维护人工智能系统)。

资源由组织本身、客户或第三方提供。

其他信息

记录资源也有助于确定是否有可用资源，如果没有可用资源，组织宜修改人工智能系统的设计规

范或其部署要求。

B.4.3 数据资源

控制

作为资源识别的一部分,组织宜记录有关人工智能系统所使用的数据资源信息。

实施指南

数据文档宜包括但不限于以下内容:

——数据的出处;

——数据最后更新或修改的日期(如元数据中的日期标签);

——对于机器学习,数据类别(如训练、验证、测试和生产数据);

——数据类别(如 ISO/IEC 19944-1 中规定的类别);

——标注数据的过程;

——数据的预期用途;

——数据的质量[如 ISO/IEC 5259(所有部分)中的描述];

——适用的数据保留和处置方针;

——数据中已知或潜在的偏见问题;

——数据准备。

B.4.4 工具资源

控制

作为资源识别的一部分,组织宜记录有关人工智能系统所使用的工具资源信息。

实施指南

人工智能系统特别是机器学习的工具资源包括但不限于:

——算法类型和机器学习模型;

——数据调节工具或过程;

——优化方法;

——评价方法;

——资源配置工具;

——辅助模型开发的工具;

——用于人工智能系统设计、开发和部署的软件和硬件。

其他信息

ISO/IEC 23053 为机器学习各种工具资源的类型、方法和途径提供了详细指导。

B.4.5 系统和计算资源

控制

作为资源识别的一部分,组织宜记录有关人工智能系统所使用的系统和计算资源信息。

实施指南

人工智能系统的系统和计算资源信息包括但不限于:

——人工智能系统的资源要求(即帮助确保系统能在资源有限的设备上运行);

——系统和计算资源的位置(如本地、云计算或边缘计算);

——处理资源(包括网络和存储);

——用于运行人工智能系统工作负载的硬件造成的影响(例如,通过使用或制造硬件或使用硬件

的成本对环境造成的影响)。

组织宜考虑持续改进人工智能系统可能需要不同的资源。系统的开发、部署和运行可能有不同的系统需求和要求。

注:ISO/IEC 22989:2022描述了各种系统资源考虑因素。

B.4.6 人力资源

控制

作为资源识别的一部分,组织宜记录开发、部署、运行、变更管理、维护、转让和退役以及验证和集成人工智能系统所使用的人力资源及其能力的相关信息。

实施指南

组织宜考虑对不同专业知识的需求,并纳入系统所需的岗位类型。例如,如果与用于训练机器学习模型的数据集相关的特定人群是系统设计的必要组成部分,则组织能将其纳入其中。必要的人力资源包括但不限于:

——数据科学家;

——与人工智能系统人工监督相关的岗位;

——物理安全、信息安全和隐私等可信赖领域的专家;

——人工智能研究人员和专家,以及与人工智能系统相关领域的专家。

在人工智能系统生存周期的不同阶段需要不同的资源。

B.5 评估人工智能系统的影响

B.5.1 目标

评估人工智能系统在全生存周期内对个人和(或)群体,以及社会的影响。

B.5.2 人工智能系统影响评估过程

控制

组织宜建立一个过程,以评估人工智能系统在其全生存周期内可能对个人和(或)群体,以及对社会造成的潜在后果。

实施指南

由于人工智能系统可能对个人和(或)群体,以及对社会产生重大影响,提供和使用此类系统的组织宜根据这些系统的预期目的和用途,评估这些系统对这些群体的潜在影响。

组织宜考虑人工智能系统是否会影响以下方面:

——个人的法律地位或生活机会;

——个人的身心健康;

——普遍人权;

——社会。

组织的过程宜包括但不限于:

a) 宜进行人工智能系统影响评估的情况,包括但不限于:
   1) 使用人工智能系统的预期目的和环境的关键性,或这些方面的任何重大变化;
   2) 人工智能技术的复杂性和人工智能系统的自动化程度,或这方面的任何重大变化;
   3) 人工智能系统处理的数据类型和来源的敏感性或任何重大变化。
b) 作为人工智能系统影响评估过程一部分的要素,可能包括:
   1) 识别(如来源、事件和结果);

2) 分析(如后果和可能性);
3) 评价(如接受决定和优先级);
4) 处置(如缓解措施);
5) 记录、报告和沟通(见 7.4、7.5 和 B.3.3)。

c) 由谁进行人工智能系统影响评估。
d) 如何利用人工智能系统影响评估[例如,如何为系统的设计或使用提供信息(见 B.6 和 B.9),是否可能触发评审和批准]。
e) 根据系统的预期目的、用途和特点,可能受到影响的个人和社会(例如,针对个人、个人群体或社会的评估)。

影响评估宜考虑到人工智能系统的各个方面,包括用于开发人工智能系统的数据、使用的人工智能技术和全系统的功能。

根据组织的角色、人工智能应用领域以及影响评估的具体学科(如信息安全、隐私和物理安全),过程可能会有所不同。

其他信息

对于某些学科或组织来说,详细考虑对个人和(或)群体,以及对社会的影响是风险管理的一部分,特别是在信息安全、物理安全和环境管理等学科。组织宜确定,作为此类风险管理过程一部分的特定学科影响评估是否充分纳入了对这些特定方面(如隐私)的人工智能考虑。

注:ISO/IEC 23894 描述了组织如何对组织本身以及个人和(或)群体,以及对社会进行影响分析,作为整体风险管理过程的一部分。

B.5.3 人工智能系统影响评估记录

控制

组织宜记录人工智能系统影响评估的结果,并在规定期限内保留结果。

实施指南

这些文件有助于确定宜传达给用户和其他相关方的信息。

人工智能系统影响评估宜按 B.5.2 中记录的人工智能系统影响评估要素保留并在需要时更新。保留期限能遵循组织保留时间表,或根据法律要求或其他要求确定。

组织宜考虑记录的项目包括但不限于:

——人工智能系统的预期用途和可合理预见的滥用;
——人工智能系统对相关个人和(或)群体,以及社会的积极和消极影响;
——可预测的故障、其潜在影响以及为减轻影响而采取的措施;
——系统适用的相关人群;
——系统的复杂性;
——人在避免系统产生负面影响的作用,包括人的监督能力、过程和工具;
——就业和员工技能。

B.5.4 评估人工智能系统对个人或群体的影响

控制

组织宜评估并记录人工智能系统在全系统生存周期内对个人或群体的潜在影响。

实施指南

在评估对个人和(或)群体,以及对社会的影响时,组织宜考虑其治理原则、人工智能方针和目标。使用人工智能系统的个人或其个人可识别信息被人工智能系统处理的个人,可能对人工智能系统的可信赖抱有期望。宜考虑到儿童、残疾人、老年人和工人等群体的特殊保护需求。作为系统影响评估的

一部分,组织宜评估这些期望,并考虑应对这些期望的方法。

根据人工智能系统的目的和使用范围,作为评估一部分需要考虑的影响领域可能包括但不限于:

——公平性;

——可问责;

——透明性和可解释性;

——信息安全和隐私;

——物理安全与健康;

——财务后果;

——可访问性;

——人权。

其他信息

必要时,组织宜咨询专家(如研究人员、主题专家和用户),以充分了解人工智能系统对个人和(或)群体,以及对社会的潜在影响。

B.5.5 评估人工智能系统的社会影响

控制

组织宜评估并记录其人工智能系统在全生存周期中可能产生的社会影响。

实施指南

社会影响因组织环境和人工智能系统类型的不同而有所差异。人工智能系统的社会影响既可能是有益的,也可能是有害的。这些潜在的社会影响领域包括:

——环境可持续性(包括对自然资源和温室气体排放的影响);

——经济(包括获得金融服务、就业机会、税收、贸易和商业);

——政府(包括立法程序、为政治利益提供错误信息、国家安全和刑事司法系统);

——健康与安全(包括获得医疗保障、医疗诊断和治疗,以及潜在的身心伤害);

——行为规范、习俗、文化和价值观[包括导致偏见或对个人和(或)群体,以及对社会造成伤害的错误信息]。

其他信息

人工智能系统的开发和使用可能是计算密集型的,会对环境可持续性产生相关影响(例如,因用电量增加而产生的温室气体排放,对水、土地、动植物的影响)。同样,人工智能系统也能用于改善其他系统的环境可持续性(如减少与建筑和运输有关的温室气体排放)。组织宜结合其总体环境可持续性目标和战略,考虑其人工智能系统的影响。

组织宜考虑人工智能系统会如何被滥用从而造成社会危害,以及如何利用人工智能系统解决历史伤害。例如,人工智能系统是否会阻碍贷款、赠款、保险和投资等金融服务的获取?同样,人工智能系统能否改善对这些金融服务的获取?

人工智能系统已被用于影响选举结果和制造错误信息(如数字媒体中的深度伪造),从而可能导致政治和社会动荡。政府将人工智能系统用于刑事司法目的,暴露出对个人和(或)群体,以及对社会的偏见风险。组织宜分析不良行为者如何滥用人工智能系统,以及人工智能系统如何强化不受欢迎的历史性社会偏见。

人工智能系统能用于诊断和治疗疾病,并确定享受医疗福利的资格。人工智能系统还能用于发生故障可能导致人员伤亡的场景(如自动驾驶汽车、人机协作)。组织在使用人工智能系统时,如在与健康和安全相关的场景中,宜同时考虑正面和负面的结果。

注:ISO/IEC TR 24368提供了与人工智能系统和应用有关的伦理和社会问题的高级概述。

B.6 人工智能系统生存周期

B.6.1 人工智能系统开发管理指南

B.6.1.1 目标

确保组织识别和记录目标并实施负责任的人工智能系统设计和开发过程。

B.6.1.2 负责任地开发人工智能系统的目标

控制

组织宜识别并记录目标，以指导负责任地开发人工智能系统，并在开发生存周期中考虑这些目标，并纳入实现这些目标的措施。

实施指南

组织宜确定影响人工智能系统设计和开发过程的目标(见 6.2)。这些目标宜在设计和开发过程中加以考虑。例如，如果一个组织将“公平性”规定为目标之一，则宜将其纳入需求说明、数据采集、数据调节、模型训练、验证和确认等过程中。组织宜提供必要的要求和指南，以确保在各个阶段都有相应的措施(例如，要求使用特定的测试工具或方法来解决不公平或不必要的偏见)，从而实现这些目标。

其他信息

人工智能技术正被用于增强安全措施，如威胁预测、检测和安全攻击预防。这是人工智能技术的一种应用，能用于加强安全措施，保护人工智能系统和传统的非人工智能软件系统。附录 C 提供了管理风险的组织目标示例，有助于确定人工智能系统开发的目标。

B.6.1.3 负责任地设计和开发人工智能系统的过程

控制

组织宜规定并记录负责任地设计和开发人工智能系统的具体过程。

实施指南

负责任的人工智能系统过程开发宜考虑包括但不限于：

——生存周期阶段(ISO/IEC 22989:2022 提供了通用的人工智能系统生存周期模型，但组织能指定自己的生存周期阶段)；

——测试要求和计划测试手段；

——人工监督要求，包括过程和工具，特别是当人工智能系统可能对自然人产生影响时；

——宜在哪些阶段进行人工智能系统影响评估；

——训练数据预期和规则(如能使用哪些数据、经批准的数据供应商和标签)；

——人工智能系统开发人员所需的专业知识(主题领域或其他)或培训，或两者兼而有之；

——发布准则；

——各阶段所需的批准和签核；

——变更控制；

——可用性和可控性；

——相关方的参与。

具体的设计和开发过程取决于人工智能系统打算使用的功能和人工智能技术。

B.6.2 人工智能系统生存周期

B.6.2.1 目标

规定人工智能系统生存周期各阶段的准则和要求。

B.6.2.2 人工智能系统的要求和规范

控制

组织宜规定并记录新的人工智能系统或对现有系统的重大改进的要求。

实施指南

组织宜记录开发人工智能系统的理由及其目标。宜考虑、记录和理解的一些因素包括：

a) 为什么要开发人工智能系统，例如，是受业务案例、客户要求还是政府方针的驱动；

b) 如何训练模型以及如何实现数据要求。

宜明确人工智能系统的要求，并宜贯穿全人工智能系统的生存周期。如果所开发的人工智能系统无法按预期运行，或出现了可能用于更改和改进要求的新信息，则宜重新审查这些要求。例如，从财务角度看，开发人工智能系统可能变得不可行。

其他信息

ISO/IEC 5338 提供了描述人工智能系统生存周期的过程。有关交互系统以人为本设计的更多信息，见 ISO 9241-210。

B.6.2.3 人工智能系统的设计和开发记录

控制

组织宜根据组织目标、文件化的要求和规范准则，记录人工智能系统的设计和开发。

实施指南

人工智能系统有许多必要的设计选择，包括但不限于：

——机器学习方法(如监督式与非监督式)；

——所使用的机器学习模型的学习算法和类型；

——模型的训练方式和数据质量(见 B.7)；

——模型的评价和改进；

——硬件和软件组件；

——在全人工智能系统生存周期中考虑的安全威胁；人工智能系统特有的安全威胁包括数据投毒、模型窃取或模型反转攻击；

——界面和输出展示；

——人类如何与系统互动；

——互操作性和可移植性方面的考虑。

在设计和开发之间可能会有多次迭代，但宜保留各阶段的文档，并提供最终的系统架构文档。

其他信息

有关交互系统的以人为本设计的更多信息，见 ISO 9241-210。

B.6.2.4 人工智能系统的验证和确认

控制

组织宜规定并记录人工智能系统的验证和确认措施，并规定其使用准则。

实施指南

验证和确认措施可能包括但不限于：

——测试方法和工具；

——测试数据的选择及其在预期使用领域的代表性；

——发布准则要求。

组织宜规定和记录评价准则如下，但不限于：

——评价人工智能系统组件和全人工智能系统对个人和(或)群体，以及对社会影响风险的计划。

——评价计划能基于以下因素，例如：

- 人工智能系统的可靠性和安全性要求，包括人工智能系统绩效的可接受误差率；
- 负责任的人工智能系统开发和使用目标，如 B.6.1.2 和 B.9.3 中的目标；
- 操作因素，如数据质量、预期用途，包括每个操作因素的可接受范围；
- 任何可能需要规定更严格操作要素的预期用途，包括不同的操作要素可接受范围或较低的误差率。

——用于评价根据人工智能系统输出结果作出决定或受制于决定的相关利益方是否能够充分解释人工智能系统输出结果的方法、指导或衡量准则。宜根据人工智能系统影响评估的结果确定评价频率。

——任何可接受的因素，这些因素能导致无法达到目标的最低绩效水平，特别是在评价人工智能系统对个人和社会的影响时(例如，计算机视觉系统图像分辨率低或背景噪声影响语音识别系统)。还宜记录处理这些因素导致的人工智能系统绩效低下的机制。

人工智能系统宜根据记录的评价准则进行评价。

如果人工智能系统不能满足记录在案的评价准则，特别是不能满足负责任的人工智能系统开发和使用目标(见 B.6.1.2，B.9.3)，组织宜重新考虑或管理人工智能系统预期用途的缺陷、其绩效要求以及组织如何有效处理对个人和社会的影响。

注：有关如何处理神经网络鲁棒性的更多信息，见ISO/IEC TR 24029-1。

B.6.2.5 人工智能系统的部署

控制

组织宜记录部署计划，并确保在部署前满足适当的要求。

实施指南

人工智能系统可能在不同的环境中开发，也可能在其他环境中部署(如在本地开发，使用云计算部署)，组织在制定部署计划时宜考虑到这些差异。组织还宜考虑组件是否分开部署(例如，软件和模型可能独立部署)。此外，组织在发布和部署之前宜满足一组要求(有时称为“发布准则”)。这可能包括需要通过的验证和确认措施、达到的绩效指标、完成的用户测试，以及获得的管理审批和签批。部署计划宜考虑到相关利益方的观点和影响。

B.6.2.6 人工智能系统的运行和监视

控制

组织宜规定并记录人工智能系统持续运行的必要要素。至少宜包括系统和绩效监视、维修、更新和支持。

实施指南

运行和监测的每项最低活动都可能考虑各种因素。例如：

——系统和绩效监视可能包括对一般错误和故障的监视，以及对系统是否按预期运行生产数据的监视。技术绩效准则可能包括解决问题或完成任务的成功率或信任率。其他准则可能与满足相关方的承诺或期望和需求有关，例如包括持续监视，以确保符合客户要求或适用的法律要求。

——一些已部署的人工智能系统通过机器学习来提高其绩效，其中生产数据和输出数据被用于进一步训练机器学习模型。在使用持续学习的情况下，组织宜监视人工智能系统的绩效，以确保其持续满足设计目标，并按预期在生产数据上运行。

——有些人工智能系统即使不使用持续学习，其绩效也会发生变化，这通常是由于生产数据中的概念或数据漂移造成的。在这种情况下，监视能确定是否需要重新训练，以确保人工智能系统继续实现其设计目标，并按预期在生产数据上运行。更多信息见 ISO/IEC 23053。

——维修可能包括对系统中的错误和故障做出响应。组织宜制定应对和修复这些问题的过程。此外，随着系统的发展，或随着关键问题的发现，或由于外部发现的问题（如不符合客户期望或法律要求），可能有必要进行更新。宜制定更新系统的过程，包括受影响的组件、更新时间表、向用户提供的关于更新内容的信息。

——系统更新还可能包括系统操作的变更、新的或修改后的预期用途，或系统功能的其他变更。组织宜制定程序来处理操作变更，包括与用户沟通。

——对系统的支持可能是内部的、外部的或两者兼有，这取决于组织的需求和系统的获取方式。支持过程宜考虑用户如何联系适当的帮助、如何报告问题和事件、支持服务级别协议和指标。

——如果人工智能系统的使用目的与设计目的不同，或使用的方式未在预料之中，则宜考虑此类使用的适当性。

——宜识别与组织应用和开发的人工智能系统相关的人工智能特有的信息安全威胁。人工智能特有的信息安全威胁包括但不限于数据投毒、模型窃取和模型反转攻击。

其他信息

组织宜考虑可能影响相关方的运行绩效，并在设计和确定绩效准则时考虑这一点。

运行中的人工智能系统的绩效准则宜根据所考虑的任务来确定，如分类、回归、排序、聚类或降维。

绩效准则可能包括错误率和处理持续时间等统计方面。对于每项准则，组织宜确定所有相关指标以及指标之间的相互依赖关系。对于每个指标，组织宜根据领域专家的建议和对相关方相对于现有非人工智能实践的期望分析等，考虑可接受的值。

例如，如 ISO/IEC TS 4213 所述，组织能根据其对假阳性和假阴性影响的评估，确定 F_1 分数是一个合适的绩效指标。然后，组织能确定人工智能系统宜达到的 F_1 值。宜评估这些问题是否能通过现有措施来解决。如果不能，则宜考虑更改现有措施，或规定其他措施来检测和处理这些问题。

组织宜考虑运行中的非人工智能系统或过程的绩效，并在制定绩效准则时将其作为潜在的相关背景。

组织宜确保用于评估人工智能系统的手段和过程，包括在适用的情况下选择和管理评价数据，以提高根据规定准则评估其绩效的完整性和可靠性。

绩效评估方法的制定可能以准则、指标和价值为基础。这些准则、指标和价值宜能为评估中使用的数据量和过程类型以及评估人员的作用和专业技能提供依据。

绩效评估方法宜尽可能地反映运行和使用的属性和特征，以确保评估结果的有用性和相关性。绩效评估的某些方面可能需要有控制地引入错误或虚假数据或过程，以评估对绩效的影响。

ISO/IEC 25059 中的质量模型能用于规定绩效准则。

B.6.2.7 人工智能系统的技术文件化信息

控制

组织宜确定用户、合作伙伴、监管机构等各类相关方需要的人工智能系统技术记录，并以适当形式向他们提供技术文件化信息。

实施指南

人工智能系统技术文件包括但不限于：

——人工智能系统的一般说明，包括其预期目的；

——使用说明；

——关于其部署和运行的技术假设(运行环境、相关软件和硬件能力、对数据的假设等);
——技术限制(如可接受的错误率、准确性、可靠性、鲁棒性);
——允许用户或操作员影响系统运行的监视能力和功能。

与所有人工智能系统生存周期阶段(见 ISO/IEC 22989:2022)有关的文档要素可能包括但不限于:

——设计和系统架构规范;
——在系统开发过程中做出的设计选择和采取的质量措施;
——系统开发过程中使用的数据信息;
——对数据质量所作的假设和采取的质量措施(如假设的统计分布);
——人工智能系统开发或运行过程中开展的管理活动(如风险管理);
——验证和确认记录;
——人工智能系统运行时所作的改动;
——B.5 所述的影响评估文件。

组织宜记录与负责任地运行人工智能系统有关的技术信息。这可能包括但不限于:

——记录管理故障的计划。例如,这可能包括需要说明人工智能系统的回滚计划、关闭人工智能系统的功能、更新过程或将人工智能系统的变更通知客户、用户等的计划、系统故障的最新信息以及如何缓解这些故障。
——记录监测人工智能系统健康状况的过程(即人工智能系统按预期并在正常运行范围内运行,也称为可观察性)和处理人工智能系统故障的过程。
——记录人工智能系统的标准操作程序,包括宜监视哪些事件以及如何优先处理和审查事件日志。还可能包括如何调查故障和预防故障。
——记录负责人工智能系统操作的人员和负责系统使用的人员的职责,特别是在处理人工智能系统故障的影响或管理人工智能系统更新方面。
——记录系统更新,如系统操作的变化、新的或修改后的预期用途,或系统功能的其他变化。

组织宜制定适当的过程来解决操作上的变更,包括与用户的沟通和对变更类型的内部评价。

文件宜及时更新并准确无误。文件宜得到组织内相关管理层的批准。

作为用户文档的一部分提供时,宜考虑表 A.1 中提供的控制。

B.6.2.8 人工智能系统的事件记录日志

控制

组织宜确定在人工智能系统生存周期的哪些阶段宜启用事件日志记录,但至少在人工智能系统使用时启用。

实施指南

组织宜确保为其部署的人工智能系统记录日志,以自动收集和记录与操作期间发生的某些事件相关的事件日志。此类日志记录包括但不限于:

——人工智能系统功能的可追溯性,以确保人工智能系统按预期运行;
——通过监视人工智能系统的运行,检测人工智能系统在预期运行条件之外的绩效,这可能会导致生产数据的不良绩效或对相关方造成影响。

人工智能系统事件日志包括一些信息,如每次使用人工智能系统的时间和日期、人工智能系统运行的生产数据、超出人工智能系统预期运行范围的输出等。

事件日志的保存时间宜根据人工智能系统的预期用途以及组织的数据保存方针和与数据保存相关的法律要求而定。

其他信息

某些人工智能系统(如生物识别系统)可能会根据管辖范围而有额外的日志记录要求。组织宜了解这些要求。

B.7 人工智能系统的数据

B.7.1 目标

确保组织理解人工智能系统中的数据在人工智能系统生存周期中的应用、开发、提供或使用中的作用和影响。

B.7.2 用于开发和增强人工智能系统的数据

控制

组织宜规定、记录和实施与人工智能系统开发相关的数据管理过程。

实施指南

数据管理包括各种控制,包括但不限于:

——由于使用数据而影响到隐私和信息安全,其中一些数据在本质上可能是敏感的;

——依赖于数据的人工智能系统开发可能产生的信息安全和物理安全威胁;

——透明性和可解释性方面,包括数据来源,以及如果系统需要透明性和可解释性,则提供解释数据如何用于确定人工智能系统的输出的能力;

——训练数据与运行使用领域相比的代表性;

——数据的准确性和完整性。

注:人工智能系统生存周期和数据管理概念的详细信息见ISO/IEC 22989:2022。

B.7.3 数据采集

控制

组织宜确定并记录人工智能系统所用数据的采集和选择细节。

实施指南

根据人工智能系统的范围和使用情况,组织可能需要来自不同来源的不同类别的数据。关于数据采集的详细信息包括:

——人工智能系统所需的数据类别;

——所需数据量;

——数据源(如内部、购买、共享、开放数据、合成);

——数据源的特性(如静态、流式、收集、机器生成);

——数据主体的人口统计学特征(如已知或潜在的偏见或其他系统性错误);

——事先处理数据(如以前的使用、符合隐私和信息安全要求);

——数据权利(如个人可识别信息、版权);

——相关的元数据(如数据标签和增强的细节);

——数据的来源。

其他信息

ISO/IEC 19944-1 中的数据类别和数据使用的结构能用于记录有关数据采集和使用的细节。

B.7.4 人工智能系统的数据质量

控制

组织宜规定和记录数据质量要求,并确保用于开发和运行人工智能系统的数据符合这些要求。

实施指南

用于开发和操作人工智能系统的数据质量可能会对系统输出的有效性产生重大影响。ISO/IEC 25024 将数据质量规定为在规定条件下使用时,数据特征满足规定和隐含需求的程度。对于使用监督式或半监督式机器学习的人工智能系统,尽可能对训练、验证、测试和生产数据的质量进行规定、测量和改进是很重要的,组织宜确保数据能达到其预期目的。组织宜考虑偏见对系统绩效和系统公平性的影响,并对用于提高绩效和公平性的模型和数据做出必要的调整,以便它们能够被用例所接受。

其他信息

关于数据质量的其他信息,见 ISO/IEC 5259(所有部分)中的关于分析和机器学习数据质量。关于在人工智能系统中使用的数据中不同形式的偏见的其他信息见 ISO/IEC TR 24027。

B.7.5 数据来源

控制

组织宜规定并记录一个过程,用于在数据和人工智能系统的生存周期中记录其人工智能系统中使用的数据的来源。

实施指南

根据 ISO 8000-2,数据来源的记录可能包括关于数据控制的创建、更新、转录、抽象、验证和传输的信息。此外,还可能在数据来源下考虑数据共享(不转移控制)和数据转换。根据数据来源、数据内容和使用环境等因素,组织宜考虑是否需要采取验证数据来源的措施。

B.7.6 数据准备

控制

组织宜规定并记录其选择数据准备的准则和要使用的数据准备方法。

实施指南

在人工智能系统中使用的数据通常需要做好准备,以使其能用于给定的人工智能任务。例如,机器学习算法有时不能容忍缺失或不正确的条目、非正态分布和广泛变化的尺度。准备方法和转换能用来提高数据的质量。如果未能正确准备好数据,可能会导致人工智能系统错误。在人工智能系统中使用的数据的常用准备方法和转换包括:

——对数据的统计探索(例如分布、平均值、中位数、标准差、范围、分层、抽样)和统计元数据(例如数据文档计划规范);

——数据清洗(即纠正条目,处理缺失的条目);

——估算(即用于填写缺失条目的方法);

——规范化;

——缩放比例;

——添加目标变量的标签;

——编码(例如,将分类变量转换为数字)。

对于给定的人工智能任务,组织宜记录其选择特定数据准备方法和转换的准则,以及在人工智能任务中使用的特定方法和转换。

注:关于机器学习特有的数据准备的更多信息,见ISO/IEC 5259(所有部分)和ISO/IEC 23053。

B.8 人工智能系统相关方的信息

B.8.1 目标

确保相关方掌握必要的信息,以了解和评估风险及其影响(正面和负面)。

B.8.2 为用户提供的系统文件和信息

控制

组织宜确定并向人工智能系统用户提供必要的信息。

实施指南

关于人工智能系统的信息可能包括技术细节和说明，以及对用户正在与人工智能系统交互的一般通知，这取决于环境。这还可能包括系统本身，以及系统的潜在输出(例如，通知用户一个图像是由人工智能创建的)。

虽然人工智能系统可能很复杂，但用户能够理解他们何时与人工智能系统交互，以及该系统的工作原理是至关重要的。用户还需要了解其预期目的和预期用途，以及其对用户造成伤害或受益的可能性。一些系统文档能用于更多的技术用途(例如系统管理员)，组织宜了解不同感兴趣方的需求以及可理解性对他们意味着什么。这些信息宜是可访问的，无论是在查找它的易用性方面，还是对于那些可能需要额外的可访问性功能的用户方面。

可能提供给用户的信息包括但不限于：

——系统的用途；

——用户正在与一个人工智能系统进行交互；

——如何与系统进行交互；

——如何以及何时覆盖该系统；

——系统运行的技术要求，包括所需的计算资源，以及系统的限制及其预期寿命；

——对人类监督的需求；

——关于准确性和绩效的信息；

——来自影响评估的相关信息，包括潜在的好处和危害，特别是如果它们适用于特定的情况或某些人群(见 B.5.2 和 B.5.4)；

——对系统的好处的修正；

——更新和更改系统的工作方式，以及任何必要的维护措施，包括其频率；

——联系方式；

——供系统使用的教育材料。

组织用来决定是否提供什么信息的准则宜被记录下来。相关准则包括但不限于人工智能系统的预期用途和合理可预见的误用、用户的专业知识和人工智能系统的具体影响。

信息可能以多种方式提供给用户，包括形成文件的使用说明、系统本身内置的警报和其他通知、网页上的信息等。根据组织使用的提供信息的方法，宜验证用户是否可访问这些信息，并且所提供的信息是否是完整的、最新的和准确的。

B.8.3 外部报告

控制

组织宜具有为相关方报告人工智能系统负面影响的能力。

实施指南

监测系统运行是否存在报告的问题和故障的同时，组织宜为用户或其他外部各方提供报告不利影响(例如，不公平)的能力。

B.8.4 事故通报

控制

组织宜确定并记录向人工智能系统用户通报事故的计划。

实施指南

与人工智能系统有关的事件可能是人工智能系统本身的特定事件，也可能是与信息安全或隐私有关的事件(如数据泄露)。组织宜根据系统运行的具体情况，了解其在通知用户和其他相关方有关事件方面的义务。例如，作为影响安全的产品一部分的人工智能组件发生事故时，其通知要求可能与其他类型的系统不同。法律要求(如合同)和监管活动可能适用，它们能明确规定以下要求：

——必须通报的事件类型；

——通知的时限；

——是否必须通知有关机构以及通知哪些机构；

——必须通报的详细信息。

组织可能将人工智能的事件响应和报告活动整合到更广泛的组织事件管理活动中，但宜注意与人工智能系统或人工智能系统单个组件相关的独特要求(例如，系统训练数据中的个人可识别信息数据泄露可能会产生与隐私相关的不同报告要求)。

其他信息

ISO/IEC 27001 和 ISO/IEC 27701 分别提供了有关安全和隐私事件管理的更多详细信息。

B.8.5 向相关方提供信息

控制

组织宜确定并记录其向相关各方报告有关人工智能系统的信息的义务。

实施指南

在某些情况下，一个司法管辖区可能会要求与监管方等监管机构共享有关该系统的信息。信息可能在适当的时间框架内报告给相关方，如客户或监管机构。共享的信息能包括，例如：

——技术系统文件，包括但不限于，用于训练、确认和测试的数据集，以及算法选择的理由和验证和确认记录；

——与本系统相关的风险；

——影响评估结果；

——日志和其他系统记录。

组织宜了解其在这方面的义务，并确保与正确的监管方分享适当的信息。此外，假定组织了解与执法监管方共享的信息有关的司法要求。

B.9 人工智能系统的使用

B.9.1 目标

确保组织负责任地并按组织的方针使用人工智能系统。

B.9.2 负责任地使用人工智能系统的过程

控制

组织宜规定并记录负责任地使用人工智能系统的过程。

实施指南

根据其环境，组织在决定是否使用特定的人工智能系统时可能有许多考虑因素。无论人工智能系统是由组织本身开发的还是来自第三方，组织都宜清楚这些考虑是什么，并制定方针来解决这些问题。例如：

——需求的批准；

——成本(包括持续的监测和维护费用)；

——批准的采购需求；

——适用于组织的法律要求。

如果组织已经接受了使用其他系统、资产等的方针，那么如果需要，可能合并这些方针。

B.9.3 负责任地使用人工智能系统的目标

控制

组织宜识别并记录用于指导负责任地使用人工智能系统的目标。

实施指南

在不同的环境下运作的组织可能对人工智能系统的负责任地开发有不同的期望和目标。根据其环境，组织宜确定其与负责任地使用相关的目标。目标包括：

——公平性；

——可问责；

——透明性；

——可解释性；

——可靠性；

——物理安全；

——鲁棒性和冗余性；

——隐私和信息安全；

——可访问性。

一旦规定，组织宜在组织内实施机制来实现其目标。这可能包括确定第三方解决方案是否满足了组织的目标，或者内部开发的解决方案是否适用于预期的用途。组织宜确定在人工智能系统生存周期的哪个阶段宜纳入有意义的人类监督目标。这可能包括：

——让人类审查员检查人工智能系统的输出，包括有权覆盖人工智能系统做出的决策；

——如果需要根据与预期部署人工智能系统相关的指示或其他文件使用人工智能系统，确保包括人工监督；

——监视人工智能系统的绩效，包括人工智能系统输出的准确性；

——报告有关人工智能系统输出的关切及其对相关方的影响；

——报告人工智能系统正确输出生产数据的绩效或能力的变化；

——考虑自动决策是否适合于负责任的方法来使用人工智能系统和人工智能系统的预期使用。

通过人工智能系统影响评估(见 B.5)能得知人工监督的必要性。参与与人工智能系统相关的人工监督活动的人员宜被告知、培训和理解有关人工智能系统的说明和其他文件，以及它们为满足人工监督目标所执行的职责。在报告绩效问题时，人工监督可能增强自动监视。

其他信息

附录 C 提供了管理风险的组织目标的示例，这有助于确定人工智能系统使用的目标。

B.9.4 工智能系统的预期用途

控制

组织宜确保按人工智能系统及其附带文件的预期用途使用人工智能系统。

实施指南

人工智能系统宜根据说明和其他与人工智能系统相关的文档(见 B.8.2)进行部署。部署可能需要特定的资源来支持部署，包括确保根据需要应用人工监督的需要(见 B.9.3)。为了可接受地使用人工智能系统，人工智能系统使用的数据需要与人工智能系统相关的文件一致，以确保人工智能系统绩效

准确。

宜监视人工智能系统的运行情况(见 B.6.2.6)。如果根据相关说明正确部署的人工智能系统对相关方或组织的法律要求造成影响,组织宜将其对影响的关切传达给组织内部的相关人员以及人工智能系统的任何第三方供应商。

组织宜保存与人工智能系统的部署和操作相关的事件日志或其他文档,这些文档能用于证明人工智能系统正在按预期使用,或帮助沟通与预期使用相关的问题。事件日志和其他文档的保存时间取决于人工智能系统的预期用途、组织的数据保留方针以及数据保留的相关法律要求。

B.10 第三方及客户关系

B.10.1 目标

确保组织了解其责任并可问责,并在人工智能系统生存周期的任何阶段涉及第三方时适当分摊风险。

B.10.2 分配责任

控制

组织宜确保在其人工智能系统生存周期内的责任被分配在组织、其合作伙伴、供应商、客户和第三方之间。

实施指南

在人工智能系统的生存周期中,责任能在提供数据的各方、提供算法和模型的各方、开发或使用人工智能系统的各方之间分配,并对部分或所有相关方负责。组织宜记录干预人工智能系统生存周期的所有各方及其角色,并确定其责任。

当组织向第三方提供人工智能系统时,组织宜确保其采取负责任的方法来开发人工智能系统,见 B.6 中的控制和实施指南。组织宜能够向相关方及组织供应人工智能系统的第三方提供人工智能系统所需的文件(见 B.6.2.7 和 B.8.2)。

当处理后的数据包括个人可识别信息时,职责通常在个人可识别信息处理者和控制者之间分配。ISO/IEC 29100 提供了关于个人可识别信息控制者和个人可识别信息处理者的进一步信息。如果要保留个人可识别信息的隐私,则宜考虑诸如 ISO/IEC 27701 中所描述的控制。根据组织和人工智能系统对个人可识别信息的数据处理活动以及组织在人工智能系统全生存周期的应用和开发中的角色,组织可能承担个人可识别信息控制者(或联合个人可识别信息控制者)、个人可识别信息处理者,或两者均承担。

B.10.3 供应商

控制

组织宜建立一套过程,确保其对供应商提供的服务、产品或材料的使用符合组织负责任地开发和使用人工智能系统的方针。

实施指南

组织开发或使用人工智能系统能借助多种供应商,从采购数据集,机器学习算法或模型,或系统的其他组件,如软件库,到整个人工智能系统本身,供其单独使用或作为另一个产品(如车辆)的一部分使用。

组织在选择供应商,确定对这些供应商的要求以及确定对这些供应商持续监视和评价的级别时,宜考虑不同类型的供应商、其提供的产品以及其可能给全系统和组织带来的不同程度的风险。

组织宜记录这些人工智能系统和人工智能系统组件如何集成到组织开发或使用的人工智能系

统中。

如果组织认为供应商提供的人工智能系统或人工智能系统组件未按预期运行，或其对个人和社会造成的影响，与组织对人工智能系统采取的负责任的方法不一致，则组织宜要求供应商采取纠正措施。组织能决定与供应商合作以实现这一目标。

组织宜确保人工智能系统的供应商提供与人工智能系统相关的适当且充分的文件(见 B.6.2.7 和 B.8.2)。

B.10.4 客户

控制

组织宜确保其开发和使用人工智能系统的方法负责任并考虑到客户的期望和需求。

实施指南

当组织提供与人工智能系统相关的产品或服务时(即当它本身是一个供应商时)，组织宜了解客户的期望和需求。这些可能以在设计或工程阶段对产品或服务本身的要求的形式出现，也可能以合同要求或一般使用协议的形式出现。一个组织可能有许多不同类型的客户关系，而这些客户关系都可能有不同的需求和期望。

组织宜特别了解供应商和客户关系的复杂性，并了解人工智能系统的供应商的职责，以及客户的职责，同时需要满足需求和期望。

例如，组织能识别与客户使用其人工智能产品和服务相关的风险，并能通过向其客户提供适当的信息来决定处理已识别的风险，这样客户就能处理相应的风险。

作为适当信息的示例，当一个人工智能系统对某个使用域有效时，该域的限制宜传达给客户(见 B.6.2.7 和 B.8.2)。

附 录 C
（资料性）
潜在的与人工智能相关的组织目标和风险源

C.1 概述

本附录概述了组织在管理风险时可能考虑的潜在组织目标、风险源及描述。本附录并非详尽无遗，也并非适用于每个组织。组织宜确定相关的目标和风险源。ISO/IEC 23894 就这些目标和风险源及其与风险管理的关系提供了更详细的信息。对人工智能系统的初始评价、定期评价以及必要时的评价，需给出正在根据组织目标进行评估的佐证依据。

C.2 目标

C.2.1 可问责性

人工智能的使用可能改变现有的可问责框架。过去，人们为其行为负责，而现在，他们的行为可能借助于人工智能系统的支持或基于人工智能系统的使用。

C.2.2 人工智能专业知识

需要挑选出在评估、开发和部署人工智能系统方面具有跨学科技能和专业知识的专家。

C.2.3 训练和测试数据的可用性和质量

基于机器学习的人工智能系统需要训练、验证和测试数据，以便训练和验证系统的预期行为。

C.2.4 环境影响

人工智能的使用可能会对环境产生积极的和消极的影响。

C.2.5 公平性

人工智能系统在自动决策方面的不当应用可能对特定的个人或人群不公平。

C.2.6 可维护性

可维护性是指组织为了纠正缺陷或适应新需求而修改人工智能系统的能力。

C.2.7 隐私性

滥用或披露个人和敏感数据（如健康记录）可能对数据主体产生有害影响。

C.2.8 鲁棒性

在人工智能中，鲁棒性特性表明系统在新数据上能否具备与经过训练的数据或典型操作数据相似的绩效。

C.2.9 物理安全

物理安全涉及期望系统在规定的条件下不会导致人类的生命、健康、财产或环境受到威胁的状态。

C.2.10 信息安全

在人工智能的环境下，特别是基于机器学习方法的人工智能系统，宜考虑超过传统信息和系统安

全的全新信息安全问题。

C.2.11 透明性和可解释性

透明性既涉及操作人工智能系统的组织的特征，也涉及这些系统本身。可解释性与对影响人工智能系统结果的重要因素的解释有关，这些因素以一种人类可理解的方式提供给相关方。

C.3 风险源

C.3.1 环境复杂性

当人工智能系统在复杂的环境中运行时，情境的范围很广，绩效可能存在不确定性，因此成为一种风险源（例如，自动驾驶的复杂环境）。

C.3.2 缺乏透明性和可解释性

无法向相关方提供适当的信息可能是一种风险源（例如，在组织的可信赖和可问责性方面）。

C.3.3 自动化程度

自动化程度可能对各种相关的领域产生影响，如物理安全、公平性或信息安全。

C.3.4 与机器学习相关的风险源

用于机器学习的数据质量和用于收集数据的过程可能是一种风险源，因为它可能影响诸如物理安全和鲁棒性等目标（例如，由于数据质量或数据投毒问题）。

C.3.5 系统硬件问题

与硬件相关的风险源包括基于有缺陷组件的硬件错误，或在不同系统之间转移训练过的机器学习模型。

C.3.6 系统生存周期问题

风险源可能出现在整个人工智能系统的生存周期中（例如，设计缺陷、部署不足、缺乏维护、退役问题）。

C.3.7 技术准备

风险可能源于使用了未知因素（如系统限制和边界条件、绩效漂移）导致的不成熟技术有关，也可能源于因技术自满而使用了更成熟的技术有关。

附 录 D
（资料性）
跨领域或跨行业使用人工智能管理体系

D.1 概述

该管理体系适用于任何开发、提供或使用人工智能系统产品或服务的组织。因此，它可能适用于不同行业的各种产品和服务，这些产品和服务与各相关方的义务、良好实践、期望或合同承诺相符合。各行业示例：

——健康，
——国防，
——交通，
——金融，
——就业，
——能源。

为了负责任地开发和使用人工智能系统，考虑各种组织目标（可能的目标见附录 C）。本文件提供了来自人工智能技术特定视角的要求和指导。对于一些潜在的目标，已经存在通用或行业特定的管理体系标准。这些管理体系标准通常从技术中立的角度来考虑目标，而人工智能管理体系则为人工智能技术的特定考量提供了具体考虑因素。

人工智能系统不仅由使用人工智能技术的组件组成，而且能使用各种技术和组件。因此，对于负责任的人工智能系统开发与使用来说，不仅需要考虑特定于人工智能的因素，还需要考虑将该系统视为一个整体，包括所使用的所有技术和组件。即使是人工智能技术的特定部分，除了人工智能特定的考虑外，还宜考虑其他方面。例如，由于人工智能是一种信息处理技术，信息安全一般也适用于它。诸如物理安全、信息安全、隐私和环境影响等目标宜对人工智能和系统的其他组件进行全面管理，而不是单独管理。因此，将人工智能管理体系与相关主题的通用或行业特定的管理体系标准相结合，对于负责任地开发和使用人工智能系统至关重要。

D.2 人工智能管理体系与其他管理体系的集成

当提供或使用人工智能系统时，组织可能会有与其他管理体系标准相关的目标或义务。这些包括信息安全、隐私、质量等主题，分别对应 ISO/IEC 27001、ISO/IEC 27701 和 ISO 9001。

在提供、使用或开发人工智能系统时，这些潜在的相关通用管理体系标准包括但不限于：

——ISO/IEC 27001：在大多数情况下，信息安全是通过人工智能系统实现组织目标的关键。一个组织追求信息安全目标的方式取决于其所处的环境以及其自身的和其自己的方针。如果某组织确定需要实施一个人工智能管理体系，它能部署一个符合 ISO/IEC 27001 的信息安全管理体系，并以类似的彻底和系统的方式来实现其信息安全目标。鉴于 ISO/IEC 27001 和人工智能管理体系都使用了高层级架构，它们易于集成使用，并为组织提供了很高的收益。在这种情况下，本文件（见 B.6.1.2）中与信息安全相关的（部分）控制的实施方式能与组织实施 ISO/IEC 27001 相结合。

——ISO/IEC 27701：在许多行业环境和应用领域中，个人可识别信息是由人工智能系统处理的。然后，组织能遵守适用的隐私义务和它自身的方针和目标。类似地，对于 ISO/IEC 27001，组织能从 ISO/IEC 27701 与人工智能管理体系的集成中获益。人工智能管理体系的隐私相关目标和控制（见 B.2.3 和 B.5.4）能与组织实施的 ISO/IEC 27701 相集成。

——ISO 9001：对于许多组织来说，符合 ISO 9001 是表明他们以客户为导向、真正关注内部有效性的关键标志。对 ISO 9001 进行的独立符合性评估促进了跨组织的业务，并激发了客户对产品或服务的信心。当涉及人工智能技术时，在人工智能管理体系与 ISO 9001 联合实施时，客户对组织或人工智能系统的信心水平能得到高度增强。在帮助组织实现其目标方面，人工智能管理体系能作为 ISO 9001 要求的补充（风险管理、软件开发、供应链一致性等）。

除了上面提到的通用管理体系标准外，人工智能管理体系还能与专门用于某个行业的管理体系联合使用。例如，ISO 22000 与人工智能管理体系都与用于食品生产、准备和物流的人工智能系统相关。又如 ISO 13485，与人工智能管理体系的集成能支持 ISO 13485 中与医疗器械软件相关的需求，或来自医疗行业的其他标准的要求，如 IEC 62304。

参 考 文 献

[1] GB/T 1.1 标准化工作导则 第1部分:标准化文件的结构和起草规则

[2] GB/T 23694—2013 风险管理 术语(ISO Guide 73:2009,IDT)

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

[4] GB/T 29246—2023 信息安全技术 信息安全管理体系 概述和词汇(ISO/IEC 27000:2018,IDT)

[5] ISO 8000-2 Data quality—Part 2: Vocabulary

[6] ISO 9001 Quality management systems—Requirements

[7] ISO 9241-210 Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems

[8] ISO 13485 Medical devices—Quality management systems—Requirements for regulatory purposes

[9] ISO 19011 Guidelines for auditing management systems

[10] ISO 22000 Food safety management systems—Requirements for any organization in the food chain

[11] ISO 37002 Whistleblowing management systems—Guidelines

[12] ISO/IEC Guide 51 Safety aspects—Guidelines for their inclusion in standards

[13] ISO/IEC TS 4213 Information technology—Artificial intelligence—Assessment of machine learning classification performance

[14] ISO/IEC 5259 (all parts) Artificial intelligence—Data quality for analytics and Machine Learning(ML)

[15] ISO/IEC 5338 Information technology—Artificial intelligence—AI system life cycle process

[16] ISO/IEC 17065 Conformity assessment—Requirements for bodies certifying products, processes and services

[17] ISO/IEC 19944-1 Cloud computing and distributed platforms—Data flow, data categories and data use—Part 1: Fundamentals

[18] ISO/IEC 23053 Framework for Artificial Intelligence (AI)Systems Using Machine Learning (ML)

[19] ISO/IEC 23894 Information technology—Artificial intelligence—Guidance on risk management

[20] ISO/IEC TR 24027 Information technology—Artificial intelligence (AI)—Bias in AI systems and AI aided decision making

[21] ISO/IEC TR 24029-1 Artificial Intelligence (AI)—Assessment of the robustness of neural networks—Part 1: Overview

[22] ISO/IEC TR 24368 Information technology—Artificial intelligence—Overview of ethical and societal concerns

[23] ISO/IEC 25024 Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—Measurement of data quality

[24] ISO/IEC 25059 Software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—Quality model for AI systems

[25] ISO/IEC 27001 Information security, cybersecurity and privacy protection—Information security management systems—Requirements

[26] ISO/IEC 27701 Security techniques—Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management—Requirements and guidelines

[27] ISO/IEC 29100 Information technology—Security techniques—Privacy framework

[28] ISO/IEC 38500:2015 Information technology—Governance of IT for the organization

[29] ISO/IEC 38507 Information technology—Governance of IT—Governance implications of the use of artificial intelligence by organizations

[30] IEC 62304 Medical device software—Software life cycle processes

[31] Lifecycle D.D.I. 3.3, 2020-04-15. Data Documentation Initiative (DDI)Alliance. [viewed on 202202-19]. Available at: https:// ddialliance .org/ Specification/ DDI-Lifecycle/ 3 .3/

[32] Risk Framework N.I.S.T.-A.I. 1.0, 2023-01-26. National Institute of Technology (NIST) [viewed on 2023-04-17] https:// www .nist .gov/ itl/ ai -risk -management -framework