国际的
标准

国际标准化组织/国际电工委员会
42001

第一版
2023年12月

信息技术人工智能管理系统

信息技术　智能技术　管理系统

参考编号
ISO/IEC 42001:2023(E)

受版权保护的文档

ISO 版权局 CP 401 ·
第 1 章de Blandonnet 8 CH-1214
Vernier, 日内瓦 电话：+41 22
749 01 11 电子邮件：
copyright@iso.org 网站：
www.iso.org
瑞士出版

内容

页

前言

ISO（国际标准化组织)和 IEC（国际电工委员会)构成了全球标准化的专业体系。作为 ISO 或 IEC 成员的国家机构通过各自组织设立的技术委员会参与国际标准的制定,以处理特定的技术活动领域。 ISO 和 IEC 技术委员会在共同感兴趣的领域开展合作。其他政府和非政府国际组织也与 ISO 和 IEC 合作参与了这项工作。

用于制定本文件的程序及其进一步维护的程序在 ISO/IEC 指令第 1 部分中进行了描述。特别是,应注意不同类型文件所需的不同批准标准。本文件是根据 ISO/IEC 指令第 2 部分的编辑规则起草的（参见www.iso.org/directives或www.iec.ch/members_experts/refdocs）。

ISO 和 IEC 提请注意本文件的实施可能涉及使用专利的可能性。 ISO 和 IEC 对任何所主张的专利权的证据、有效性或适用性不采取任何立场。截至本文件发布之日,ISO 和 IEC 尚未收到实施本文件可能需要的专利通知。然而,实施者请注意,这可能并不代表最新信息,最新信息可以从www.iso.org/patents和https://patents.iec.ch上提供的专利数据库中获得。 ISO 和 IEC 不负责识别任何或所有此类专利权。

本文档中使用的任何商品名称都是为了方便用户而提供的信息,并不构成认可。

有关标准自愿性的解释、与合格评定相关的 ISO 特定术语和表达方式的含义,以及有关 ISO 遵守世界贸易组织 (WTO) 贸易技术壁垒 (TBT) 原则的信息,请参见www .iso.org/iso/foreword.html。在 IEC 中,请参阅www.iec.ch/understanding-standards。

本文件由联合技术委员会 ISO/IEC JTC 1 （信息技术）、小组委员会 SC 42（人工智能)编写。

有关本文件的任何反馈或问题应直接提交给用户的国家标准机构。这些机构的完整列表可以在www.iso.org/members.html和www.iec.ch/national-committees上找到。

介绍

人工智能（AI)越来越多地应用于利用信息技术的各个领域,预计将成为主要的经济驱动力之一。这一趋势的结果是,某些应用可能会在未来几年引发社会挑战。

本文件旨在帮助组织负责任地履行其在人工智能系统方面的角色（例如使用、开发、监控或提供利用人工智能的产品或服务）。人工智能可能会提出具体的考虑因素,例如:

使用人工智能进行自动决策,有时以不透明和不可解释的方式进行,可能需要超出传统 IT 系统管理的具体管理。

使用数据分析、洞察力和机器学习,而不是人类编码的逻辑来设计系统,既增加了人工智能系统的应用机会,又改变了此类系统的开发、论证和部署方式。

执行持续学习的人工智能系统会在使用过程中改变其行为。他们需要特别考虑,以确保他们负责任地使用随着行为的改变而继续。

本文件提供了在组织范围内建立、实施、维护和持续改进人工智能管理系统的要求。组织应将其要求的应用重点放在人工智能独有的功能上。人工智能的某些特征,例如持续学习和改进的能力或缺乏透明度或可解释性,如果与传统的任务执行方式相比引起了额外的担忧,则可以采取不同的保护措施。采用人工智能管理系统来扩展现有的管理结构是组织的战略决策。

组织的需求和目标、流程、规模和结构以及各利益相关方的期望影响人工智能管理体系的建立和实施。影响人工智能管理体系建立和实施的另一组因素是人工智能的众多用例以及需要在治理机制和创新之间取得适当的平衡。组织可以选择使用基于风险的方法来应用这些要求,以确保对组织范围内的特定人工智能用例、服务或产品应用适当级别的控制。所有这些影响因素预计都会发生变化并不时进行审查。

人工智能管理系统应与组织的流程和整体管理结构相集成。在流程、信息系统和控制的设计中应考虑与人工智能相关的具体问题。此类管理流程的重要示例包括:

组织目标的确定、相关方的参与和组织政策;

风险和机遇的管理;

管理与人工智能系统可信度相关的问题的流程,例如人工智能系统整个生命周期的安全性、安全性、公平性、透明度、数据质量和质量;

为组织提供或开发人工智能系统的供应商、合作伙伴和第三方的管理流程。

本文件提供了部署适用控制措施以支持此类流程的指南。

本文件避免了对管理流程的具体指导。组织可以结合普遍接受的框架、其他国际标准和自身经验来实施适合范围内特定人工智能用例、产品或服务的关键流程,例如风险管理、生命周期管理和数据质量管理。

符合本文件要求的组织可以生成其在人工智能系统方面的角色和责任的证据。

本文档中提出要求的顺序并不反映其重要性或暗示其实施顺序。列举的列表项仅供参考。

与其他管理体系标准的兼容性

本文件应用了为加强管理体系标准 (MSS) 之间的一致性而开发的统一结构（相同的条款编号、条款标题、文本和通用术语以及核心定义）。人工智能管理系统提供了专门用于管理组织中使用人工智能所产生的问题和风险的要求。这种通用方法有利于实施并与其他管理体系标准保持一致，例如与质量、安全、安保和隐私相关的标准。

信息技术　人工智能 管理系统

1 范围

本文件规定了在组织范围内建立、实施、维护和持续改进 AI（人工智能)管理系统的要求并提供指导。

本文档供提供或使用利用人工智能系统的产品或服务的组织使用。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要求、与利益相关方相关的义务以及他们的期望。

本文件适用于提供或使用利用人工智能系统的产品或服务的任何组织，无论规模、类型和性质如何。

2 规范性引用文件

正文中引用下列文件时，其部分或全部内容构成本文件的要求。对于注明日期的参考文献，仅引用的版本适用。对于未注明日期的参考文献，适用参考文件的最新版本（包括任何修订）。

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

就本文件而言，ISO/IEC 22989 中给出的术语和定义以及以下内容适用。

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/obp

IEC Electropedia：可在https://www. electropedia.org/获取

3.1
组织
具有自己的职能、责任、权力和关系以实现其目标的个人或群体(3.6)

注 1：组织的概念包括但不限于个体工商户、公司、公司、事务所、企业、当局、合伙企业、慈善机构或机构或其部分或组合，无论是否为法人组织、公共或私人。

注 2：如果组织是较大实体的一部分，则术语“组织”仅指人工智能管理系统(3.4) 范围内的较大实体的部分。

3.2
利害关系方
能够影响决策或活动、受决策或活动影响或认为自己受决策或活动影响的个人或组织(3.1)

注 1：ISO/IEC 22989:2022, 5.19 中提供了人工智能相关方的概述。

3.3

高层管理人员

在最高级别指导和控制组织（3.1)的个人或团体

注 1:最高管理层有权在组织内授予权力并提供资源。

注 2:如果管理体系(3.4)的范围仅涵盖组织的一部分,则最高管理层是指指导和控制组织该部分的人员。

3.4

管理系统

组织(3.1)的一组相互关联或相互作用的要素,用于建立政策(3.5)和目标(3.6),以及实现这些目标的流程(3.8)

注 1:管理体系可以针对单个学科或多个学科。

注 2:管理体系要素包括组织的结构、角色和职责、规划和运营。

3.5

政策

由最高管理层正式表达的组织(3.1)意图和方向(3.3)

3.6

客观的

所要达到的结果

注 1:目标可以是战略目标、战术目标或操作目标。

注 2:目标可能涉及不同学科（例如金融、健康与安全以及环境）。
例如,它们可以是组织范围内的,也可以是特定于项目、产品或流程的(3.8)。

注 3:目标可以用其他方式表达,例如作为预期结果、作为目的、作为操作标准、作为 AI 目标或使用具有类似含义的其他词语（例如目的、目标或目标）。

注 4:在人工智能管理系统(3.4) 的背景下,人工智能目标由组织(3.1)设定,与人工智能政策(3.5) 一致,以实现特定结果。

3.7

风险

不确定性的影响

注 1:效应是指与预期的偏差　正的或负的。

注 2:不确定性是指与事件、其后果或可能性相关的信息缺乏、理解或知识缺乏的状态,甚至是部分缺乏。

注 3:风险通常通过潜在事件（如 ISO 指南 73 中的定义)和后果（如 ISO 指南 73 中的定义)或这些的组合来表征。

注 4:风险通常以事件后果（包括情况变化)和相关发生可能性（如 ISO 指南 73 中定义)的组合来表示。

3.8

过程

使用或转换输入来交付结果的一组相互关联或相互作用的活动

注 1:流程的结果是否被称为输出、产品或服务取决于引用的上下文。

3.9
能力应用知识和
技能以实现预期结果的能力

3.10
文件化信息组织（3.1)及其包含
的介质需要控制和维护的信息。

注 1:记录信息可以采用任何格式、任何媒体以及来自任何来源。

注2:记录信息可以参考:

管理体系（3.4），包括相关过程（3.8）；

为组织运作而创建的信息（文件）；

取得成果的证据（记录）。

3.11
绩效可测量的结
果

注 1:绩效可以与定量或定性发现相关。

注 2:绩效可以与管理活动、流程(3.8)、产品、服务、系统或组织(3.1) 相关。

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系统（3.4)相关的结果。从该术语的使用上下文可以清楚地看出该术语的正确解释。

3.12
持续改进提高绩效的经常性活
动（3.11）

3.13
有效性实现计划
活动和实现计划结果的程度

3.14
要求明示的、一般
暗示的或强制性的需要或期望

注 1:“一般暗示”是指组织(3.1)和相关方(3.2)的习惯或惯例是暗示所考虑的需要或期望。

注 2:特定要求是指在文件化信息(3.10) 等中规定的要求。

3.15
要求（3.14)的
符合性满足

3.16
nonconformity不
符合要求（3.14）

3.17
纠正措施消除不合格
（3.16)原因并防止再次发生的措施

3.18
审核系
统且独立的过程（3.8），用于获取证据并对其进行客观评估，以确定审核标准的满足程度

注 1：审核可以是内部审核（第一方）或外部审核（第二方或第三方），也可以是组合审核（结合两个或多个学科）。

注 2：内部审核由组织(3.1)本身或由外部方代表其进行。

注 3："审核证据"和"审核标准"在 ISO 19011 中定义。

3.19
测量
过程（3.8）确定一个值

3.20
监控确定系统、过
程（3.8）或活动的状态

注 1：为了确定状态，可能需要检查、监督或严格观察。

3.21
控制
<risk> 维持和/或修改风险的措施(3.7)

注 1：控制措施包括但不限于维持和/或改变风险的任何流程、政策、设备、实践或其他条件和/或行动。

注 2：控制措施可能并不总是能发挥预期或假设的改变效果。

[来源：ISO 31000:2018, 3.8，已修改 - 添加 <risk> 作为应用程序域]

3.22
管理机构对组织的绩效
和一致性负责的个人或团体

注 1：并非所有组织，尤其是小型组织，都会有一个独立于最高管理层的治理机构。

注 2：治理机构可以包括但不限于董事会、董事会委员会、监事会、受托人或监督者。

[来源：ISO/IEC 38500:2015, 2.9，已修改　添加了条目注释。]

3.23
信息安全 information security
信息机密性、完整性和可用性的保存

注 1：还可能涉及其他属性，例如真实性、责任性、不可否认性和可靠性。

[资料来源：ISO/IEC 27000:2018, 3.28]

3.24
人工智能系统影响评估正式、记录的过程，开
发、提供或使用利用人工智能的产品或服务的组织通过该过程来识别、评估和解决对个人、个人群体或两者以及社会的影响

3.25
数据质量
数据的特征,即数据满足组织对特定上下文的数据要求

[资料来源:ISO/IEC 5259-1:1), 3.4]

3.26
适用性声明
记录所有必要的控制措施(3.23)以及包含或排除控制措施的理由

注 1:组织可能不需要附件 A中列出的所有控制措施,甚至可能会超出附件 A中列出的由组织本身建立的额外控制措施。

注 2:组织应根据本文件的要求记录所有已识别的风险。所有已识别的风险以及为解决这些风险而建立的风险管理措施（控制)均应反映在适用性声明中。

4 组织背景

4.1 了解组织及其背景

组织应确定与其目的相关并影响其实现人工智能管理系统预期结果的能力的外部和内部问题。

组织应确定气候变化是否是一个相关问题。

组织应考虑组织开发、提供或使用的人工智能系统的预期目的。组织应确定其在这些人工智能系统中的角色。

注 1:为了了解组织及其背景,组织确定其相对于人工智能系统的角色可能会有所帮助。这些角色可以包括但不限于以下一项或多项:

人工智能提供商,包括人工智能平台提供商、人工智能产品或服务提供商;

人工智能生产者,包括人工智能开发者、人工智能设计者、人工智能运营者、人工智能测试者和评估者、人工智能部署者、人工智能人为因素专业人员、领域专家、人工智能影响评估者、采购者、人工智能治理和监督专业人员;

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;

相关当局,包括政策制定者和监管者。

ISO/IEC 22989 提供了这些角色的详细描述。此外,NIST AI 风险管理框架中还描述了角色的类型及其与 AI 系统生命周期的关系。 [29]
组织的角色可以确定本文件中的要求和控制的适用性和适用性范围。

注2:本条款下要解决的外部和内部问题可能会根据组织的角色和管辖范围及其对其实现人工智能管理系统预期结果的能力的影响而有所不同。这些可以包括但不限于:

a) 与外部环境相关的考虑因素,例如:

1) 适用的法律要求,包括禁止使用人工智能;

2)监管机构对人工智能系统开发和使用中法律要求的解释或执行有影响的政策、指南和决定;

1)正在准备中。 ISO/IEC DIS 5259-1:2023 发布时的阶段。

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)与人工智能开发和使用相关的文化、传统、价值观、规范和道德;

5)利用人工智能系统的新产品和服务的竞争格局和趋势;

b) 内部背景相关考虑因素,例如:

1) 组织背景、治理、目标（见6.2）、政策和程序;

2)合同义务;

3)要开发或使用的人工智能系统的预期目的。

注3:角色确定可以通过与组织处理的数据类别相关的义务来形成（例如,个人身份信息（PII)处理者或处理PII 时的PII 控制者）。有关 PII 和相关角色,请参阅 ISO/IEC 29100。角色还可以根据人工智能系统特定的法律要求来确定。

4.2 了解相关方的需求和期望

组织应确定:

与人工智能管理系统相关的利益相关方;

这些利益相关方的相关要求;

这些要求中哪些将通过人工智能管理系统得到解决。

注:相关利益方可以提出与气候变化相关的要求。

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。

在确定该范围时,组织应考虑:

4.1中提到的外部和内部问题;

4.2中提到的要求。

范围应作为文件化信息提供。

人工智能管理体系的范围应确定组织根据本文件对人工智能管理体系的要求进行的活动、领导、规划、支持、操作、绩效、评估、改进、控制和目标。

4.4 人工智能管理系统

组织应根据本文件的要求建立、实施、维护、持续改进并记录人工智能管理体系,包括所需的过程及其相互作用。

5 领导力

5.1 领导力和承诺

最高管理层应通过以下方式展示对人工智能管理系统的领导力和承诺：

确保制定人工智能政策（见5.2)和人工智能目标（见6.2)并与组织的战略方向相一致；

确保将人工智能管理系统要求整合到组织的管理体系中
业务流程；

确保人工智能管理系统所需的资源可用；

传达有效人工智能管理和符合人工智能管理体系要求的重要性；

确保人工智能管理系统实现其预期结果；

指导和支持人员为人工智能管理系统的有效性做出贡献；

促进持续改进；

- 支持其他相关角色展示其在其领域的领导力
责任。

注1:本文件中提到的“业务”可以广义地解释为那些对组织存在的目的至关重要的活动。

注 2:在组织内建立、鼓励和塑造一种文化,采取负责任的方法来使用、开发和管理人工智能系统,可以是高层管理人员承诺和领导力的重要体现。确保认识并遵守这种负责任的方法,并通过领导力支持人工智能管理系统,可以帮助人工智能管理系统取得成功。

5.2 人工智能政策

最高管理层应制定人工智能政策：

a) 适合组织的目的；

b) 提供设定人工智能目标的框架（见6.2）；

c) 包括满足适用要求的承诺；

d) 包括对持续改进人工智能管理系统的承诺。

人工智能政策应：

作为文件化信息提供；

参考与其他组织政策相关的内容；

在组织内进行沟通；

酌情向感兴趣的各方提供。

表 A.1中的 A.2 提供了建立人工智能政策的控制目标和控制措施。
B.2中提供了这些控制措施的实施指南。

注 ISO/IEC 38507 中提供了组织制定人工智能政策时的注意事项。

5.3 角色、责任和权限

最高管理层应确保在组织内分配并传达相关角色的职责和权限。

最高管理层应分配以下职责和权力：

a) 确保人工智能管理体系符合本文件的要求；

b) 向最高管理层报告人工智能管理系统的绩效。

注：表 A.1 中的 A.3.2 提供了定义和分配角色和职责的控制。
B.3.2中提供了该控制的实施指南。

6 规划

6.1 应对风险和机遇的行动

6.1.1 概述

在规划人工智能管理系统时，组织应考虑4.1中提到的问题和4.2中提到的要求，并确定需要解决的风险和机遇：

保证人工智能管理系统能够实现其预期结果；

防止或减少不良影响；

实现持续改进。

组织应建立并维护人工智能风险标准，以支持：

区分可接受和不可接受的风险；

进行人工智能风险评估；

进行人工智能风险处理；

评估人工智能风险影响。

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

组织应根据以下方面确定风险和机遇：

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：人工智能管理系统的范围可以考虑多个人工智能系统。在这种情况下，为每个人工智能系统或人工智能系统组执行机会和用途的确定。

组织应计划：

a) 应对这些风险和机遇的行动；

b) 如何：

1) 将这些行动整合并实施到其人工智能管理系统流程中；

2) 评估这些行动的有效性。

组织应保留有关识别和解决人工智能风险和人工智能机遇所采取行动的书面信息。

注 3:ISO/IEC 23894 提供了如何对开发、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南。

注4:组织的背景及其活动可能对组织的风险管理活动产生影响。

注 5:定义风险的方式以及设想风险管理的方式可能因部门和行业而异。 3.7中的风险定义允许适用于任何部门的广泛风险愿景,例如附件D 中提到的部门。在任何情况下,作为风险评估的一部分,组织的作用是首先采用以下愿景:适应其背景的风险。这可以包括通过人工智能系统开发和使用的部门中使用的定义来处理风险,例如 ISO/IEC 指南 51 中的定义。

6.1.2 人工智能风险评估

组织应定义并建立人工智能风险评估流程,该流程:

a) 遵循人工智能政策（见5.2)和人工智能目标（见6.2)并与之保持一致;

注:在评估6.1.2 d) 1)的后果时,组织可以利用6.1.4 中所示的人工智能系统影响评估。

b) 旨在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
结果;

c) 识别有助于或阻碍实现人工智能目标的风险;

d) 分析人工智能风险:

1)评估对组织、个人和社会的潜在后果
如果已识别的风险成为现实,将会产生什么结果;

2) 在适用的情况下,评估已识别风险的现实可能性;

3)确定风险级别;

e) 评估人工智能风险:

1) 将风险分析结果与风险标准进行比较（见6.1.1）;

2)对评估的风险进行优先级排序,进行风险处理。

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人工智能风险处理

考虑到风险评估结果,组织应定义人工智能风险处理流程,以:

a) 选择适当的人工智能风险处理方案;

b) 确定实施所选人工智能风险处理方案所需的所有控制措施,并将这些控制措施与附件 A中的控制措施进行比较,以验证没有遗漏任何必要的控制措施;

注 1:附件 A提供了用于满足组织目标并解决与人工智能系统的设计和使用相关的风险的参考控制。

c) 考虑附件 A中与实施人工智能风险处理相关的控制措施
选项;

d) 确定除了附件 A之外是否还需要额外的控制措施,以实施所有风险
治疗方案;

e) 考虑附件 B中关于实施 b) 和 c) 中确定的控制措施的指南；

注 2:控制目标隐含地包含在所选择的控制中。组织可以从附录A中选择一组适当的控制目标和控制措施。附录 A 的控制措施并不详尽,可能需要额外的控制目标和控制措施。如果除了附件A 之外还需要不同或额外的控制措施,组织可以设计此类控制措施或从现有来源获取这些控制措施。如果适用,人工智能风险管理可以集成到其他管理系统中。

f) 生成包含必要控制措施的适用性声明[见 b)、c) 和 d)],并提供包含和排除控制措施的理由。排除的理由可以包括风险评估认为没有必要采取控制措施以及适用的外部要求不要求采取控制措施（或存在例外情况）。

注3:组织可以提供排除任何一般控制目标或特定人工智能系统控制目标的书面理由,无论是附件A中列出的控制目标还是组织本身建立的控制目标。

g) 制定人工智能风险处置方案。

组织的人工智能风险处理计划和剩余人工智能风险的接受应获得指定管理层的批准。必要的控制措施应是：

与 6.2 中的目标保持一致；

作为文件化信息提供；

在组织内进行沟通；

酌情提供给感兴趣的各方。

组织应保留有关人工智能风险处理过程的文件化信息。

6.1.4 人工智能系统影响评估

组织应定义一个流程,用于评估人工智能系统的开发、提供或使用可能对个人或个人群体或两者以及社会造成的潜在后果。

人工智能系统影响评估应确定人工智能系统的部署、预期用途和可预见的滥用对个人或个人群体或两者和社会的潜在后果。

人工智能系统影响评估应考虑人工智能系统部署的具体技术和社会背景以及适用的司法管辖区。

人工智能系统影响评估的结果应记录在案。在适当的情况下,系统影响评估的结果可以提供给组织定义的相关利益方。

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

注:在某些情况下（例如安全或隐私关键的人工智能系统）,组织可以要求将特定学科的人工智能系统影响评估（例如安全、隐私或安保影响)作为组织整体风险管理活动的一部分进行。

6.2 人工智能目标和实现目标的规划

组织应在相关职能和级别制定人工智能目标。

人工智能的目标应：

a) 与人工智能政策一致（见5.2）；

b) 可测量（如果可行）；

c) 考虑适用的要求；

d) 受到监控；

e) 进行沟通；

f) 适当更新；

g) 可作为文件化信息提供。

在规划如何实现其人工智能目标时，组织应确定：

将要做什么；

需要什么资源；

谁将负责；

何时完成；

如何评估结果。

注：附件C中提供了与风险管理相关的人工智能目标的非排他性列表。用于确定负责任地开发和使用人工智能系统的目标的控制目标和控制措施以及实现这些目标的措施在附录 C 中的 A.6.1 和 A.9.3 中提供。表 A.1。 B.6.1和B.9.3中提供了这些控制措施的实施指南。

6.3 变更计划

当组织确定需要对人工智能管理体系进行变更时，变更应有计划地进行。

7 支持

7.1 资源

组织应确定并提供建立、实施、维护和持续改进人工智能管理体系所需的资源。

注：表 A.1 中的 A.4 提供了人工智能资源的控制目标和控制。第 B.4 条中提供了这些控制措施的实施指南。

7.2 能力

该组织应：

确定在其控制下从事影响人工智能的工作的人员的必要能力
表现；

确保这些人员在适当的教育、培训或培训的基础上具备能力
经验；

在适用的情况下，采取行动以获得必要的能力，并评估有效性
所采取的行动。

应提供适当的文件化信息作为能力的证据。

注 1：B.4.6 中提供了人力资源实施指南，包括考虑必要的专业知识。

注 2:适用的行动可包括,例如:为当前雇员提供培训、指导或重新分配;或雇用或承包有能力的人员。

7.3 意识

在组织控制下工作的人员应了解:

人工智能政策（见5.2）;

他们对人工智能管理系统有效性的贡献,包括以下好处:

提高人工智能性能;

不符合人工智能管理系统要求的影响。

7.4 通讯

组织应确定与人工智能管理系统相关的内部和外部沟通,包括:

它将传达什么信息;

何时沟通;

与谁沟通;

如何沟通。

7.5 记录信息

7.5.1 概述

组织的人工智能管理系统应包括:

a) 本文件要求的文件化信息;

b) 组织确定的人工智能管理体系有效性所必需的文件化信息。

注:人工智能管理系统的记录信息范围可能因组织而异,原因如下:

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性;

人员的能力。

7.5.2 创建和更新文件化信息

创建和更新文件化信息时,组织应确保适当:

标识和描述（例如标题、日期、作者或参考号）;

格式（如语言、软件版本、图形)和媒体（如纸质、电子）;

适当性和充分性的审查和批准。

7.5.3 文件化信息的控制

人工智能管理系统和本文件要求的文件化信息应
受控以确保：

a) 在需要的地方和时间可用且适合使用；

b) 得到充分保护（例如，防止机密性丧失、使用不当或完整性丧失）。

为了控制文件化信息，组织应酌情开展以下活动：

分发、访问、检索和使用；

储存和保存，包括保持可读性；

变更控制（例如版本控制）；

保留和处置。

组织确定的人工智能管理系统的规划和运行所必需的外部来源的文件化信息应被适当地识别和控制。

注：访问可能意味着有关仅查看文件化信息的许可，或查看和更改文件化信息的许可和授权的决定。

8 操作

8.1 运行规划和控制

组织应通过以下方式计划、实施和控制满足要求所需的过程，并实施第6 条中确定的行动：

建立过程标准；

按照标准实施过程控制。

组织应实施根据6.1.3确定的与人工智能管理系统运行相关的控制（例如人工智能系统开发和使用生命周期相关控制）。

应监控这些控制措施的有效性，如果未达到预期结果，应考虑采取纠正措施。附件 A列出了参考控制措施，附件 B提供了它们的实施指南。

应在必要的范围内提供文件化信息，以确保流程已按计划进行。

组织应控制计划的变更并审查非预期变更的后果，必要时采取措施减轻任何不利影响。

组织应确保外部提供的与人工智能管理体系相关的过程、产品或服务受到控制。

8.2 人工智能风险评估

组织应按计划的时间间隔或在提议或发生重大变更时根据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 人工智能风险处理

组织应根据6.1.3实施人工智能风险处理计划并验证其有效性。

当风险评估发现需要处理的新风险时,应对这些风险执行6.1.3规定的风险处理流程。

当风险处理计划定义的风险处理方案无效时,应按照6.1.3的风险处理流程对这些处理方案进行审查和重新验证,并更新风险处理计划。

组织应保留所有人工智能风险处理结果的文件化信息。

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拟发生重大变更时根据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 绩效评估

9.1 监测、测量、分析和评价

组织应确定:

需要监测和测量什么;

适用时的监测、测量、分析和评价方法,以确保有效结果;

何时进行监视和测量;

何时对监测和测量的结果进行分析和评估。

形成文件的信息应作为结果的证据。

组织应评价人工智能管理体系的绩效和有效性。

9.2 内部审计

9.2.1 概述

组织应按计划的时间间隔进行内部审核,以提供有关人工智能管理系统是否:

a) 符合:

1)组织自身对其人工智能管理系统的要求;

2)本文件的要求;

b) 得到有效实施和维护。

9.2.2 内部审核计划

组织应规划、建立、实施和维护审核方案,包括频率、方法、职责、规划要求和报告。

在制定内部审核方案时,组织应考虑相关过程的重要性以及以往审核的结果。

该组织应:

a) 确定每次审核的审核目标、标准和范围;

b) 选择审核员并进行审核,以确保审核过程的客观性和公正性;

c) 确保审核结果报告给相关管理人员。

应提供书面信息作为审核计划实施情况和审核结果的证据。

9.3 管理评审

9.3.1 概述

最高管理层应按计划的时间间隔审查组织的人工智能管理系统,以确保其持续适用性、充分性和有效性。

9.3.2 管理评审输入

管理评审应包括:

a) 先前管理评审的行动状态;

b) 与人工智能管理体系相关的外部和内部问题的变化;

c) 与人工智能管理相关的利益相关方的需求和期望的变化
系统;

d) 有关人工智能管理系统绩效的信息,包括以下方面的趋势:

1) 不合格及纠正措施;

2)监测和测量结果;

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与持续改进机会和人工智能管理系统变更需求相关的决策。

应提供文件化信息作为管理评审结果的证据。

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 不合格和纠正措施

当发生不合格时,组织应:

a) 对不符合项做出反应,并在适用时:

1)采取行动加以控制和纠正;

2)处理后果;

b) 通过以下方式评估是否需要采取行动消除不合格的原因,以使其不会在其他地方再次发生或发生:

1)不符合项的审查;

2)确定不合格的原因;

3)确定是否存在或可能发生类似的不合格;

c) 实施任何所需的行动;

d) 审查所采取的任何纠正措施的有效性;

e) 如有必要,对人工智能管理系统进行更改。

纠正措施应适合所遇到的不合格的影响。

形成文件的信息应作为以下证据:

不合格的性质以及随后采取的任何措施;

任何纠正措施的结果。

附录A
（规范）

参考控制目标和控制

A.1 概述

表A.1中详述的控制措施为组织提供了满足组织目标和解决与人工智能系统设计和运营相关的风险的参考。并非表A.1中列出的所有控制目标和控制措施都需要使用，组织可以设计和实施自己的控制措施（见6.1.3）。

附录 B提供了表 A.1 中列出的所有控制措施的实施指南。

表 A.1　控制目标和控制措施

A.2 人工智能相关政策		
目标：根据业务需求，为AI系统提供管理指导和支持。		
	话题	控制
A.2.2	人工智能政策	组织应记录人工智能系统开发或使用的政策。
A.2.3	与其他组织政策保持一致	组织应确定其他政策可能在哪些方面受到组织的人工智能系统目标的影响或适用于该目标。
A.2.4	人工智能政策检讨	应按计划或根据需要额外审查人工智能政策，以确保其持续适用性、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，以坚持对人工智能系统的实施、运营和管理采取负责任的方法。		
	话题	控制
A.3.2	人工智能的角色和职责	人工智能的角色和职责应根据组织的需要进行定义和分配。
A.3.3	报告疑虑	组织应定义并实施一个流程，以报告组织在人工智能系统整个生命周期中的角色的担忧。
A.4 人工智能系统的资源		
目标：确保组织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，以充分理解和应对风险和影响。		
	话题	控制
A.4.2	资源文档	组织应识别并记录给定人工智能系统生命周期阶段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使用的工具资源的信息。

表 A.1 （续）

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,组织应记录有关人力资源及其用于人工智能系统的开发、部署、运营、变更管理、维护、转移和退役以及验证和集成的能力的信息。
A.5 评估人工智能系统的影响		
目标:评估人工智能系统对个人或个人群体（或两者)以及受人工智能系统整个生命周期影响的社会的影响。		
	话题	控制
A.5.2	AI系统影响评估流程	组织应建立一个流程来评估人工智能系统在其整个生命周期中可能对个人或个人群体（或两者)以及社会产生的潜在后果。
A.5.3	人工智能系统影响评估的文档	组织应记录人工智能系统影响评估的结果,并在规定的期限内保留结果。
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整个系统生命周期中对个人或个人群体的潜在影响。
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其整个生命周期中的潜在社会影响。
A.6 AI系统生命周期		
A.6.1 人工智能系统开发管理指南		
目标:确保组织确定并记录目标并实施负责设计和开发人工智能系统的流程。		
	话题	控制
A.6.1.2	负责任地开发人工智能系统的目标	组织应确定并记录目标以指导负责任的开发人工智能系统,并考虑这些目标并整合措施以在开发生命周期中实现这些目标。
A.6.1.3	负责任的人工智能系统设计和开发流程	组织应定义并记录人工智能系统负责任设计和开发的具体流程。
A.6.2 AI系统生命周期		
目标:定义人工智能系统生命周期每个阶段的标准和要求。		
	话题	控制
A.6.2.2	AI系统要求和规范	组织应指定并记录新人工智能系统或现有系统的材料增强的要求。
A.6.2.3	人工智能系统设计的文档组织应记录人工智能系统的设计和开发	基于组织目标、文件化要求和规范标准的开发。
A.6.2.4	AI系统验证和确认	组织应定义并记录人工智能系统的验证和确认措施,并指定其使用标准。
A.6.2.5	人工智能系统部署	组织应记录部署计划并确保在部署之前满足适当的要求。

表 A.1 （续）

A.6.2.6	AI系统运行与监控	组织应定义并记录人工智能系统持续运行的必要元素。至少,这应包括系统和性能监控、维修、更新和支持。
A.6.2.7	人工智能系统技术文件 组织应确定人工智能系统技术文件	每个相关类别的利益相关方（例如用户、合作伙伴、监管机构)都需要 Cal 文件,并以适当的形式向他们提供技术文件。
A.6.2.8	事件日志的人工智能系统记录 组织应确定事件发生在哪些阶段	人工智能系统生命周期、事件日志的记录保存应启用,但至少在人工智能系统使用时启用。
A.7 人工智能系统的数据		
目标:确保组织了解人工智能系统中数据在人工智能系统整个生命周期的应用和开发、提供或使用中的作用和影响。		
	用于开	控制
A.7.2	发和增强人工智能系统的主题数据	组织应定义、记录和实施与人工智能系统开发相关的数据管理流程。
A.7.3	数据采集	组织应确定并记录有关人工智能系统中使用的数据的获取和选择的详细信息。
A.7.4	人工智能系统的数据质量	组织应定义并记录数据质量要求,并确保用于开发和运行人工智能系统的数据满足这些要求。
A.7.5	数据来源	组织应定义并记录一个流程,用于在数据和人工智能系统的生命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	组织应定义并记录其选择数据准备和要使用的数据准备方法的标准。
A.8 人工智能系统相关方的信息		
目标:确保相关利益方拥有必要的信息来了解和评估风险及其影响（积极和消极）。		
	话题	控制
A.8.2	系统文档和用户信息	组织应确定并向人工智能系统的用户提供必要的信息。
A.8.3	外部报告	组织应为利益相关方提供报告人工智能系统不利影响的能力。
A.8.4	事件通报	组织应确定并记录向人工智能系统用户传达事件的计划。
A.8.5	相关方信息 组织应确定并记录其	向利益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:确保组织按照组织政策负责任地使用人工智能系统。		
	主题 负	控制
A.9.2	责任地使用人工智能的流程 组织应定义并记录负责任地使用人工智能系统的流程。系统	
A.9.3	负责任地使用人工智能的目标 组织应确定并记录目标,以指导负责任地使用人工智能系统。系统	

表 A.1 （续）

A.9.4	AI系统的预期用途	组织应确保按照人工智能系统及其随附文件的预期用途使用人工智能系统。
A.10 第三方和客户关系		
目标:确保组织了解其责任并承担责任,并在人工智能系统生命周期的任何阶段涉及第三方时适当分摊风险。		
	话题	控制
A.10.2	职责分配	组织应确保其人工智能系统生命周期内的责任在组织、其合作伙伴、供应商、客户和第三方之间分配。
A.10.3	供应商	组织应建立一个流程,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法。
A.10.4	顾客	组织应确保其对人工智能系统的开发和使用负责任的方法考虑到客户的期望和需求。

附录B
（规范）

AI控制实施指南

B.1 概述

本附件中记录的实施指南涉及表 A.1 中列出的控制措施。它提供了支持实施表 A.1中列出的控制措施并满足控制目标的信息，但组织不必在适用性声明中记录或证明包含或排除实施指南的合理性（见6.1.3）。

实施指南并不总是适合或充分适应所有情况，并且并不总是满足组织的特定控制要求。组织可以根据其具体要求和风险处理需要扩展或修改实施指南或定义自己的控制实施。

本附件将用作在本文件定义的人工智能管理系统中确定和实施人工智能风险处理控制措施的指南。可以确定除本附件中包含的控制之外的其他组织和技术控制（参见6.1.3 中的人工智能系统管理风险处理）。本附件可被视为制定组织特定控制实施的起点。

B.2 人工智能相关政策

B.2.1 目的

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 人工智能政策

控制

组织应记录人工智能系统开发或使用的政策。

实施指导

人工智能政策应通过以下方式了解：

- 经营策略；

  组织价值观和文化以及组织愿意承担的风险量或保持；

  人工智能系统带来的风险水平；

  法律要求，包括合同；

  组织的风险环境；

  对相关利益方的影响（见6.1.4）。

人工智能政策应包括（除5.2 中的要求外）：

  指导组织所有与人工智能相关的活动的原则；

处理政策偏差和例外的流程。

人工智能政策应在必要时考虑特定主题的方面，以提供额外的指导或提供与处理这些方面的其他政策的交叉引用。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。

相关政策应指导人工智能系统的开发、购买、运营和使用。

B.2.3 与其他组织政策保持一致

控制

组织应确定其他政策可能会受到组织有关人工智能系统的目标的影响或适用于哪些方面。

实施指导

许多领域与人工智能都有交叉，包括质量、安全、安全和隐私。组织应考虑进行彻底分析，以确定当前政策是否以及在何处必然相交，并在需要更新时更新这些政策，或者在人工智能政策中纳入规定。

其他信息

管理机构代表组织制定的政策应为人工智能政策提供信息。
ISO/IEC 38507 为组织管理机构成员提供指导，以在人工智能系统的整个生命周期中启用和管理人工智能系统。

B.2.4 人工智能政策审查

控制

应按计划或根据需要额外审查人工智能政策，以确保其持续适用性、充分性和有效性。

实施指导

经管理层批准的角色应负责人工智能政策或其组成部分的制定、审查和评估。审查应包括评估改进组织管理人工智能系统的政策和方法的机会，以应对组织环境、业务环境、法律条件或技术环境的变化。

人工智能政策的审查应考虑管理审查的结果。

B.3 内部组织

B.3.1 目的

在组织内部建立问责制，以坚持对人工智能系统的实施、运营和管理负责的方法。

B.3.2 人工智能的角色和职责

控制

人工智能的角色和职责应根据组织的需求进行定义和分配。

实施指导

定义角色和职责对于确保整个组织在人工智能系统整个生命周期中的角色负责至关重要。组织在分配角色和职责时应考虑人工智能政策、人工智能目标和已识别的风险，以确保覆盖所有相关领域。组织可以优先考虑如何分配角色和职责。需要明确角色和职责的领域示例包括：

- 风险管理；

人工智能系统影响评估；

资产和资源管理；

- 安全；

- 安全；

- 隐私；

- 发展；

- 表现；

人类监督；

供应商关系；

证明其有能力持续满足法律要求；

数据质量管理（整个生命周期）。

应将各种角色的职责定义为适合个人履行其职责的级别。

B.3.3 报告问题

控制

组织应定义并实施一个流程，以报告组织在人工智能系统整个生命周期中所扮演角色的担忧。

实施指导

报告机制应当履行以下职能：

a) 保密或匿名或两者兼有的选项；

b) 可供雇员和合同工使用并晋升；

c) 配备合格人员；

d) 为 c) 中提到的人员规定适当的调查和解决权力；

e) 提供及时报告和上报给管理层的机制；

f) 为举报和调查相关人员提供有效保护，使其免受报复（例如允许匿名和保密地举报）；

g) 根据4.4提供报告，如果适用，还提供 e)；同时在 a) 中保持机密性和匿名性，并尊重一般商业机密性考虑因素；

h) 在适当的时间范围内提供响应机制。

注:组织可以利用现有的报告机制作为此过程的一部分。

其他信息

除了本条款中提供的实施指南外,组织还应进一步考虑 ISO 37002。

B.4 人工智能系统的资源

B.4.1 目的

确保组织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,以充分理解和应对风险和影响。

B.4.2 资源文档

控制

组织应识别并记录给定人工智能系统生命周期阶段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。

实施指导

人工智能系统资源的记录对于了解风险以及人工智能系统对个人或个人群体或两者和社会的潜在影响（积极和消极)至关重要。此类资源的文档（例如,可以利用数据流图或系统架构图)可以为人工智能系统影响评估提供信息（参见B.5）。

资源可以包括但不限于:

人工智能系统组件;

数据资源,即人工智能系统生命周期任何阶段使用的数据;

工具资源（例如人工智能算法、模型或工具）;

- 系统和计算资源（例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,即具有与组织在整个人工智能系统生命周期中的角色相关的必要专业知识（例如人工智能系统的开发、销售、培训、运营和维护)的人员。

资源可以由组织本身、其客户或第三方提供。

其他信息

资源文档还可以帮助确定资源是否可用,如果资源不可用,组织应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,组织应记录有关人工智能系统所使用的数据资源的信息。

实施指导

数据文档应包括但不限于以下主题：

数据的来源；

数据最后更新或修改的日期（例如元数据中的日期标签）；

对于机器学习，数据的类别（例如训练、验证、测试和生产数据）；

数据类别（例如ISO/IEC 19944-1中定义的）；

标记数据的过程；

数据的预期用途；

数据质量（例如ISO/IEC 5259系列2中所描述的）；

适用的数据保留和处置政策；

数据中已知或潜在的偏差问题；

数据准备。

B.4.4 工具资源

控制

作为资源识别的一部分，组织应记录有关人工智能系统使用的工具资源的信息。

实施指导

人工智能系统（特别是机器学习）的工具资源可以包括但不限于：

算法类型和机器学习模型；

数据调理工具或流程；

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 为机器学习的各种工具资源的类型、方法和途径提供了详细的指导。

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。

---

2)正在准备中。发布时的阶段：ISO/IEC DIS 5259-1:2023、ISO/IEC DIS 5259-2:2023、ISO/IEC DIS 5259-3:2023、ISO/IEC DIS 5259-4:2023、ISO/IEC CD 5259-5:2023。

实施指导

有关人工智能系统的系统和计算资源的信息可以包括但不限于：

人工智能系统的资源需求（即帮助确保系统可以在受限的环境下运行）
资源设备）；

系统和计算资源所在的位置（例如本地、云计算或边缘计算）
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的影响（例如，通过硬件的使用或制造或使用硬件的成本对环境的影响）。

组织应考虑到持续改进人工智能系统可能需要不同的资源。系统的开发、部署和运营可能有不同的系统需求和要求。

注 ISO/IEC 22989 描述了各种系统资源注意事项。

B.4.6 人力资源

控制

作为资源识别的一部分，组织应记录有关人力资源及其用于人工智能系统的开发、部署、运营、变更管理、维护、转移和退役以及验证和集成的能力的信息。

实施指导

组织应考虑对不同专业知识的需求，并包括系统所需的角色类型。例如，组织可以包括与用于训练机器学习模型的数据集相关的特定人口统计群体，如果它们的包含是系统设计的必要组成部分。必要的人力资源可以包括但不限于：

数据科学家；

与人类对人工智能系统的监督相关的角色；

安全、安保和隐私等可信主题专家；

人工智能研究人员和专家，以及与人工智能系统相关的领域专家。

人工智能系统生命周期的不同阶段可能需要不同的资源。

B.5 评估人工智能系统的影响

B.5.1 目的

评估人工智能系统对个人或个人群体（或两者）以及受人工智能系统整个生命周期影响的社会的影响。

B.5.2 人工智能系统影响评估流程

控制

组织应建立一个流程来评估人工智能系统在其整个生命周期中可能对个人或个人群体或两者以及社会产生的潜在后果。

实施指导

由于人工智能系统可能对个人、个人群体和社会产生重大影响,因此提供和使用此类系统的组织应根据这些系统的预期目的和用途,评估这些系统 ，或者对这些群体的潜在影响。

组织应考虑人工智能系统是否影响:

个人的法律地位或生活机会;

个人的身体或心理健康;

普遍人权;

社团。

该组织的程序应包括但不限于:

a) 应当进行人工智能系统影响评估的情况,包括但不限于:

1)人工智能系统的预期目的和使用环境的重要性或任何重要的
对这些进行更改;

2)人工智能技术的复杂性和人工智能系统的自动化水平或任何重要的
对此进行更改;

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
那;

b) 人工智能系统影响评估流程的要素,其中可以包括:

1) 识别（例如来源、事件和结果）;

2)分析（例如后果和可能性）;

3) 评估（例如接受决定和优先顺序）;

4) 治疗（例如缓解措施）;

5) 文件、报告和沟通（见7.4、7.5和B.3.3）;

c) 谁执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,它如何为系统的设计或使用提供信息（见B.6和B.9）,是否可以触发审查和批准];

e) 根据系统的预期目的、用途和特征（例如对个人、个人群体或社会的评估)而可能受到影响的个人和社会。

影响评估应考虑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以及整个系统的功能。

这些流程可能会根据组织的角色和人工智能应用领域以及评估影响的具体学科（例如安全、隐私和安全)而有所不同。

其他信息

对于某些学科或组织来说,详细考虑对个人或个人群体或两者和社会的影响是风险管理的一部分,特别是在信息安全、安全和环境管理等学科中。组织应确定

作为此类风险管理流程的一部分进行的特定学科影响评估是否充分整合了人工智能对这些特定方面（例如隐私）的考虑。

笔记 ISO/IEC 23894 描述了组织如何对组织本身、个人或个人团体或两者以及社会进行影响分析，作为整体风险管理的一部分过程。

B.5.3 人工智能系统影响评估的记录

控制

组织应记录人工智能系统影响评估的结果，并将结果保留一段规定的时间。

实施指导

该文档有助于确定应传达给用户和其他相关利益方的信息。

应根据 B.5.2 中记录的人工智能系统影响评估的要素保留并根据需要更新人工智能系统影响评估。保留期限可以遵循组织保留时间表，或者根据法律要求或其他要求确定。

组织应考虑记录的项目包括但不限于：

人工智能系统的预期用途以及人工智能系统的任何合理可预见的误用；

人工智能系统对相关个人或个人群体或两者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；

系统适用的相关人口群体；

系统的复杂性；

人类在与系统的关系中的作用，包括可用于避免负面影响的人类监督能力、流程和工具；

就业和员工技能。

B.5.4 评估人工智能系统对个人或个人群体的影响

控制

组织应评估并记录人工智能系统在整个系统生命周期中对个人或个人群体的潜在影响。

实施指导

在评估对个人或个人群体或两者和社会的影响时，组织应考虑其治理原则、人工智能政策和目标。使用人工智能系统或其 PII 由人工智能系统处理的个人可以对人工智能系统的可信度抱有期望。应考虑儿童、残疾人、老年人和工人等群体的具体保护需求。组织应评估这些期望并考虑解决这些期望的方法，作为系统影响评估的一部分。

根据人工智能系统目的和使用的范围，作为评估一部分考虑的影响领域可以包括但不限于：

- 公平；

问责制；

透明度和可解释性；

安全和隐私；

安全与健康；

财务后果；

可达性；

人权。

其他信息

必要时,组织应咨询专家（例如研究人员、主题专家和用户）,以充分了解人工智能系统对个人或个人群体或两者和社会的潜在影响。

B.5.5 评估人工智能系统的社会影响

控制

组织应评估并记录其人工智能系统在整个生命周期中的潜在社会影响。

实施指导

根据组织的背景和人工智能系统的类型,社会影响可能会有很大差异。
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。这些潜在社会影响的例子包括：

环境可持续性（包括对自然资源和温室气体的影响
排放）；

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、政治利益错误信息、国家安全和刑事司法系统）；

健康和安全（包括获得医疗保健、医疗诊断和治疗以及潜在的
身体和心理伤害）；

规范、传统、文化和价值观（包括导致偏见或伤害的错误信息）
个人或个人团体,或两者兼而有之,以及社会）。

其他信息

人工智能系统的开发和使用可能需要大量计算,并对环境可持续性产生相关影响（例如,由于用电量增加而导致的温室气体排放,对水、土地、动植物的影响）。同样,人工智能系统可用于改善其他系统的环境可持续性（例如减少与建筑物和交通相关的温室气体排放）。组织应在其总体环境可持续性目标和战略的背景下考虑人工智能系统的影响。

组织应考虑其人工智能系统如何被滥用来造成社会危害,以及如何使用它们来解决历史危害。例如,人工智能系统是否可以阻止获得贷款、赠款、保险和投资等金融服务,同样,人工智能系统是否可以改善对这些工具的获取？

人工智能系统已被用来影响选举结果并制造可能导致政治和社会动荡的错误信息（例如数字媒体中的深度伪造）。政府将人工智能系统用于刑事司法目的暴露了社会、个人或团体存在偏见的风险

个人。组织应该分析参与者如何滥用人工智能系统，以及人工智能系统如何强化不必要的历史社会偏见。

人工智能系统可用于诊断和治疗疾病，并确定健康福利的资格。人工智能系统还部署在故障可能导致人类死亡或受伤的场景中（例如自动驾驶汽车、人机协作）。组织应考虑使用人工智能系统时的积极和消极结果，例如在健康和安全相关场景中。

注 ISO/IEC TR 24368 提供了与人工智能系统和应用相关的道德和社会问题的高级概述。

B.6 人工智能系统生命周期

B.6.1 人工智能系统开发管理指南

B.6.1.1 目的

确保组织确定并记录目标并实施负责设计和开发人工智能系统的流程。

B.6.1.2 人工智能系统负责任开发的目标

控制

组织应确定并记录目标以指导人工智能系统的负责任开发，并考虑这些目标并整合措施以在开发生命周期中实现这些目标。

实施指导

组织应确定影响人工智能系统设计和开发过程的目标（见6.2）。在设计和开发过程中应考虑这些目标。

例如，如果组织将“公平”定义为一个目标，则应将其纳入需求规范、数据采集、数据调节、模型训练、验证和确认等中。组织应提供必要的要求和指南，以确保措施被整合到各个阶段（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）以实现这些目标。

其他信息

人工智能技术被用来增强安全措施，例如威胁预测检测和预防安全攻击。这是人工智能技术的应用，可用于加强安全措施，以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的示例，这对于确定人工智能系统开发的目标很有用。

B.6.1.3 人工智能系统负责任的设计和开发流程

控制

组织应定义并记录人工智能系统负责任设计和开发的具体流程。

实施指导

人工智能系统流程的负责任开发应包括但不限于以下考虑：

生命周期阶段（ISO/IEC 22989 提供了通用的人工智能系统生命周期模型，但
组织可以指定自己的生命周期阶段）；

测试要求和计划的测试方法；

- 人类监督要求，包括流程和工具，特别是当人工智能系统可以
影响自然人；

应在什么阶段进行人工智能系统影响评估；

培训数据期望和规则（例如可以使用哪些数据、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
两个都；

发布标准；

各个阶段所需的批准和签字；

- 切换控制；

可用性和可控性；

利益相关方的参与。

具体的设计和开发过程取决于人工智能系统的功能和人工智能技术。

B.6.2 AI系统生命周期

B.6.2.1 目标

定义人工智能系统生命周期每个阶段的标准和要求。

B.6.2.2 AI系统要求和规范

控制

组织应指定并记录新人工智能系统的要求或对现有系统的材料增强。

实施指导

组织应记录开发人工智能系统的基本原理及其目标。应考虑、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，这是由业务案例、客户驱动的吗？
要求或政府政策；

b) 如何训练模型以及如何实现数据要求。

应明确人工智能系统需求，并应涵盖整个人工智能系统生命周期。如果开发的人工智能系统无法按预期运行或出现可用于更改和改进要求的新信息，则应重新审视此类要求。例如，从财务角度来看，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命周期的流程。有关交互式系统以人为本的设计的更多信息，请参阅 ISO 9241-210。

B.6.2.3 AI系统设计和开发的文档

控制

组织应根据组织目标、记录的要求和规范标准记录人工智能系统的设计和开发。

实施指导

人工智能系统需要许多设计选择，包括但不限于：

机器学习方法（例如监督与无监督）；

学习算法和所使用的机器学习模型的类型；

模型的训练方式和数据质量（见B.7）；

模型的评估和完善；

硬件和软件组件；

在整个人工智能系统生命周期中考虑的安全威胁； AI 特有的安全威胁系统包括数据中毒、模型窃取或模型反转攻击；

输出的界面和呈现；

人类如何与系统交互；

互操作性和可移植性考虑。

设计和开发之间可以有多次迭代，但应该维护阶段的文档，并且应该提供最终的系统架构文档。

其他信息

有关交互式系统以人为本的设计的更多信息，请参阅 ISO 9241-210。

B.6.2.4 AI系统验证和确认

控制

组织应定义并记录人工智能系统的验证和确认措施，并指定其使用标准。

实施指导

验证和确认措施可以包括但不限于：

测试方法和工具；

测试数据的选择及其对预期使用领域的表示；

发布标准要求。

组织应定义并记录评估标准，例如但不限于：

评估人工智能系统组件和整个人工智能系统对个人或个人群体或两者和社会影响相关风险的计划；

评估计划可以基于,例如:

人工智能系统的可靠性和安全性要求,包括人工智能系统性能可接受的错误率;

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,例如数据质量、预期用途,包括每个因素的可接受范围
操作因素;

任何可能需要定义更严格的操作因素的预期用途,包括操作因素的不同可接受范围或更低的错误率;

用于评估基于人工智能系统输出做出决策或接受决策的相关利益方是否能够充分解释人工智能系统输出的方法、指南或指标。应根据人工智能系统影响评估的结果确定评估频率;

- 任何可以解释无法满足目标最低性能水平的可接受因素,特别是当评估人工智能系统对个人或个人群体或两者以及社会的影响时(例如计算机视觉系统或背景的图像分辨率较差)影响语音识别系统的噪声)。还应记录处理因这些因素而导致的人工智能系统性能不佳的机制。

应根据记录的评估标准对人工智能系统进行评估。

如果人工智能系统无法满足记录的评估标准,特别是针对负责任的人工智能系统开发和使用目标(见B.6.1.2和B.9.3),组织应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地解决对个人或个人群体或两者和社会的影响。

注:有关如何处理神经网络鲁棒性的更多信息可以在 ISO/IEC TR 24029-1 中找到。

B.6.2.5 人工智能系统部署

控制

组织应记录部署计划并确保在部署之前满足适当的要求。

实施指导

人工智能系统可以在各种环境中开发并部署在其他环境中(例如本地开发和使用云计算部署),组织应在部署计划中考虑这些差异。组织还应考虑组件是否单独部署(例如软件和模型可以独立部署)。此外,组织在发布和部署之前应该满足一组要求(有时称为“发布标准”)。这可以包括要通过的验证和确认措施、要满足的性能指标、要完成的用户测试以及要获得的管理层批准和签核。部署计划应考虑相关利益相关方的观点和影响。

B.6.2.6 AI系统运行与监控

控制

组织应定义并记录人工智能系统持续运行的必要元素。至少应包括系统和性能监控、维修、更新和支持。

实施指导

用于操作和监视的每个最小活动可以考虑各种考虑因素。例如：

系统和性能监控可以包括监控一般错误和故障，以及系统是否按照生产数据的预期运行。技术绩效标准可以包括解决问题或完成任务的成功率或置信度。

其他标准可能与满足相关方的承诺或期望和需求有关，包括例如持续监控以确保符合客户要求或适用的法律要求。

一些部署的人工智能系统通过机器学习来提高其性能，其中生产数据和输出数据用于进一步训练机器学习模型。在使用持续学习的情况下，组织应监控人工智能系统的性能，以确保其继续满足其设计目标并按预期运行生产数据。

即使某些人工智能系统不使用持续学习，其性能也可能会发生变化，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，监控可以确定是否需要重新培训，以确保人工智能系统继续满足其设计目标并按预期运行生产数据。更多信息请参阅 ISO/IEC 23053。

修复可以包括对系统中的错误和故障的响应。组织应该制定适当的流程来响应和修复这些问题。此外，随着系统的发展或关键问题的发现，或外部发现的问题（例如不符合客户期望或法律要求）的结果，更新可能是必要的。应该有适当的流程来更新系统，包括受影响的组件、更新时间表、向用户提供有关更新内容的信息。

系统更新还可以包括系统操作的更改、新的或修改的预期用途或系统功能的其他更改。组织应制定适当的程序来解决运营变更，包括与用户的沟通。

对系统的支持可以是内部的、外部的或两者兼而有之，具体取决于组织的需求以及系统的获取方式。支持流程应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。

如果人工智能系统被用于其设计目的以外的目的或以未预期的方式使用，则应考虑此类用途的适当性。

应识别与组织应用和开发的人工智能系统相关的特定于人工智能的信息安全威胁。人工智能特有的信息安全威胁包括但不限于数据中毒、模型窃取和模型反转攻击。

其他信息

组织应考虑可能影响相关方的运营绩效，并在设计和确定绩效标准时考虑这一点。

运行中的人工智能系统的性能标准应根据所考虑的任务来确定，例如分类、回归、排序、聚类或降维。

性能标准可以包括统计方面，例如错误率和处理持续时间。
对于每个标准，组织应确定所有相关指标以及指标之间的相互依赖性。对于每个指标，组织应根据领域专家的建议以及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，组织可以根据对误报和漏报影响的评估来确定F1分数是适当的性能指标，如中所述

ISO/IEC TS 4213。然后，组织可以建立AI 系统预期满足的F1值。应评估这些问题是否可以通过现有措施解决。如果情况并非如此，则应考虑更改现有措施或制定额外措施来检测和处理这些问题。

组织应考虑运行中的非人工智能系统或流程的性能，并在建立性能标准时将其用作潜在的相关背景。

组织还应确保用于评估人工智能系统的手段和流程，包括（如适用）评估数据的选择和管理，提高其绩效评估相对于既定标准的完整性和可靠性。

绩效评估方法的开发可以基于标准、指标和价值观。
这些应告知评估中使用的数据量和流程类型以及进行评估的人员的角色和专业知识。

绩效评估方法应尽可能反映操作和使用的属性和特征，以确保评估结果有用且相关。绩效评估的某些方面可能需要受控地引入错误或虚假数据或流程来评估对绩效的影响。

ISO/IEC 25059 中的质量模型可用于定义性能标准。

B.6.2.7 AI系统技术文档

控制

组织应确定每个相关类别的相关方（例如用户、合作伙伴、监管机构）需要哪些人工智能系统技术文档，并以适当的形式向他们提供技术文档。

实施指导

AI系统技术文档可以包括但不限于以下内容：

人工智能系统的一般描述，包括其预期目的；

使用说明；

关于其部署和操作的技术假设（运行时环境、相关软件
以及硬件功能、对数据所做的假设等）；

技术限制（例如可接受的错误率、准确性、可靠性、稳健性）；

允许用户或操作员影响系统的监控能力和功能
手术。

与所有人工智能系统生命周期阶段（如 ISO/IEC 22989 中定义）相关的文档元素可以包括但不限于：

设计和系统架构规范；

系统开发过程中做出的设计选择和采取的质量措施；

有关系统开发期间使用的数据的信息；

对数据质量所做的假设和采取的质量措施（例如假设的统计分布）；

人工智能开发或运行期间采取的管理活动（例如风险管理）
系统；

验证和确认记录；

人工智能系统运行时发生的变化；

如B.5 中所述的影响评估文件。

组织应记录与人工智能系统负责任运行相关的技术信息。这可以包括但不限于：

记录管理失败的计划。例如，这可以包括需要描述人工智能系统的回滚计划、关闭人工智能系统的功能、通知客户、用户等人工智能系统变化的更新过程或计划、更新的信息系统故障以及如何缓解这些故障；

记录监控人工智能系统健康状况的流程（即人工智能系统按预期运行并在其正常运行裕度内，也称为可观察性）以及解决人工智能系统故障的流程；

记录人工智能系统的标准操作程序，包括应监控哪些事件以及如何对事件日志进行优先级排序和审查。它还可以包括如何调查故障和预防故障；

记录负责人工智能系统运行的人员以及负责系统使用责任的人员的角色，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的更新方面；

记录系统更新，例如系统操作的更改、新的或修改的预期用途或系统功能的其他更改。

组织应制定适当的程序来解决运营变更，包括与用户的沟通和对变更类型的内部评估。

文件应是最新且准确的。文件应得到组织内相关管理层的批准。

当作为用户文档的一部分提供时，应考虑表 A.1中提供的控制措施。

B.6.2.8 AI系统记录事件日志

控制

组织应确定在人工智能系统生命周期的哪些阶段应启用事件日志的记录保存，但至少在人工智能系统使用时启用。

实施指导

组织应确保其部署的人工智能系统进行日志记录，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。此类记录可以包括但不限于：

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；

通过监控人工智能系统的运行，检测人工智能系统的性能超出了人工智能系统的预期运行条件，这可能导致生产数据出现不良性能或对相关利益方产生影响。

人工智能系统事件日志可以包括信息，例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、 ETC。

事件日志的保存时间应符合人工智能系统的预期用途并符合组织的数据保留政策。与数据保留相关的法律要求可能适用。

其他信息

一些人工智能系统,例如生物特征识别系统,根据管辖范围可能有额外的日志记录要求。组织应该了解这些要求。

B.7 人工智能系统的数据

B.7.1 目标

确保组织了解人工智能系统中数据在人工智能系统整个生命周期的应用和开发、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

组织应定义、记录和实施与人工智能系统开发相关的数据管理流程。

实施指导

数据管理可以包括各种主题,例如但不限于:

由于使用数据而产生的隐私和安全影响,其中一些数据可能本质上是敏感的;

依赖数据的人工智能系统开发可能产生的安全威胁;

透明度和可解释性方面,包括数据来源以及在系统需要透明度和可解释性的情况下解释如何使用数据来确定人工智能系统的输出的能力;

与使用的操作领域相比,训练数据的代表性;

数据的准确性和完整性。

注:AI系统生命周期和数据管理概念的详细信息由ISO/IEC 22989提供。

B.7.3 数据采集

控制

组织应确定并记录有关人工智能系统中使用的数据的获取和选择的详细信息。

实施指导

根据人工智能系统的范围和用途,组织可能需要来自不同来源的不同类别的数据。数据采集的详细信息可包括:

人工智能系统所需的数据类别;

所需数据的数量;

数据源(例如内部、购买、共享、开放数据、合成);

数据源的特征(例如静态的、流式的、聚集的、机器生成的);

数据主体的人口统计和特征(例如已知或潜在的偏见或其他系统性偏见)
错误);

数据的事先处理(例如以前的使用、符合隐私和安全要求);

数据权利（例如 PII、版权）；

相关元数据（例如数据标记和增强的细节）；

数据的来源。

其他信息

ISO/IEC 19944-1 中的数据类别和数据使用结构可用于记录有关数据采集和使用的详细信息。

B.7.4 人工智能系统的数据质量

控制

组织应定义并记录数据质量要求，并确保用于开发和运行人工智能系统的数据满足这些要求。

实施指导

用于开发和运行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 将数据质量定义为在指定条件下使用时数据特性满足明示和暗示需求的程度。对于使用监督或半监督机器学习的人工智能系统，重要的是尽可能地定义、测量和改进训练、验证、测试和生产数据的质量，并且组织应确保数据适合其预期目的。组织应考虑偏差对系统性能和系统公平性的影响，并对用于提高性能和公平性的模型和数据进行必要的调整，以便它们对于用例来说是可以接受的。

其他信息

有关数据质量的更多信息，请参阅有关分析和机器学习数据质量的 ISO/IEC 5259 系列[2]。 ISO/IEC TR 24027 中提供了有关人工智能系统中使用的数据中不同形式的偏差的更多信息。

B.7.5 数据来源

控制

组织应定义并记录一个流程，用于在数据和人工智能系统的生命周期中记录其人工智能系统中使用的数据的来源。

实施指导

根据 ISO 8000-2，数据来源记录可以包括有关数据控制的创建、更新、转录、抽象、验证和传输的信息。此外，数据共享（不转移控制）和数据转换可以在数据来源下考虑。根据数据来源、数据内容及其使用环境等因素，组织应考虑是否需要采取措施来验证数据的来源。

B.7.6 数据准备

控制

组织应定义并记录其选择数据准备和要使用的数据准备方法的标准。

实施指导

人工智能系统中使用的数据通常需要准备才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丢失或不正确的条目，非

正态分布和广泛变化的尺度。准备方法和转换可用于提高数据质量。未能正确准备数据可能会导致人工智能系统错误。人工智能系统中使用的数据的常见准备方法和转换包括：

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（例如数据文档倡议（DDI）规范[28]）；

清理（即更正条目、处理缺失条目）；

插补（即填写缺失条目的方法）；

标准化；

缩放；

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，组织应记录其选择特定数据准备方法和转换的标准以及人工智能任务中使用的特定方法和转换。

注：有关机器学习特定数据准备的更多信息，请参阅 ISO/IEC 5259 系列[2]和 ISO/IEC 23053。

B.8 相关方信息

B.8.1 目的

确保相关利益方拥有必要的信息来了解和评估风险及其影响（积极和消极）。

B.8.2 系统文档和用户信息

控制

组织应确定并向系统用户提供必要的信息。

实施指导

有关人工智能系统的信息可以包括技术细节和说明，以及向用户发出的有关他们正在与人工智能系统交互的一般通知，具体取决于上下文。这还可以包括系统本身，以及系统的潜在输出（例如，通知用户图像是由人工智能创建的）。

尽管人工智能系统可能很复杂，但用户在与人工智能系统交互时能够理解系统的工作原理至关重要。用户还需要了解其预期目的和预期用途、其对用户造成伤害或受益的可能性。某些系统文档必然会针对更多技术用途（例如系统管理员），并且组织应该了解不同相关方的需求以及可理解性对他们意味着什么。这些信息还应该是可访问的，无论是在查找信息的易用性方面，还是对于可能需要其他辅助功能的用户而言。

可以向用户提供的信息包括但不限于：

系统的目的；

用户正在与人工智能系统交互；

如何与系统交互；

如何以及何时覆盖系统；

系统运行的技术要求，包括所需的计算资源，以及
系统的局限性及其预期寿命；

需要人工监督；

有关准确性和性能的信息；

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情况或某些人口群体（见B.5.2和B.5.4）；

对有关系统好处的声明进行修改；

系统工作方式的更新和变化，以及任何必要的维护措施，
包括它们的频率；

- 联系信息；

系统使用的教育材料。

组织用来确定是否提供信息以及提供哪些信息的标准应记录在案。相关标准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、用户的专业知识以及人工智能系统的具体影响。

信息可以通过多种方式提供给用户，包括记录的使用说明、系统本身内置的警报和其他通知、网页上的信息等。根据组织用于提供信息的方法，它应该验证用户可以访问这些信息，并且所提供的信息是完整的、最新的和准确的。

B.8.3 外部报告

控制

组织应为相关方提供报告系统不利影响的能力。

实施指导

虽然应监控系统运行是否有报告的问题和故障，但组织还应为用户或其他外部方提供报告不利影响（例如不公平）的能力。

B.8.4 事件通报

控制

组织应确定并记录向系统用户传达事件的计划。

实施指导

与人工智能系统相关的事件可以特定于人工智能系统本身，也可以与信息安全或隐私相关（例如数据泄露）。组织应了解其向用户和其他相关方通知事件的义务，具体取决于系统运行的上下文。例如，涉及影响安全的产品一部分的人工智能组件的事件可能与其他类型的系统有不同的通知要求。法律要求（例如合同）和监管活动可以适用，其中可以指定以下要求：

必须通报的事件类型；

通知的时间表；

是否必须通知以及哪些当局必须被通知；

需要沟通的细节。

组织可以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，但应了解与人工智能系统或人工智能系统的各个组件相关的独特要求（例如，系统培训数据中的 PII 数据泄露可能会导致有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 相关方信息

控制

组织应确定并记录其向利益相关方报告人工智能系统信息的义务。

实施指导

在某些情况下，司法管辖区可能要求与监管机构等当局共享有关系统的信息。可以在适当的时间内向相关方（例如客户或监管机构）报告信息。共享的信息可以包括，例如：

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

与系统相关的风险；

影响评估的结果；

日志和其他系统记录。

组织应了解其在这方面的义务，并确保与正确的当局共享适当的信息。此外，还假定该组织了解与执法机构共享信息相关的管辖要求。

B.9 人工智能系统的使用

B.9.1 目的

确保组织按照组织政策负责任地使用人工智能系统。

B.9.2 负责任地使用人工智能系统的流程

控制

组织应定义并记录负责任地使用人工智能系统的流程。

实施指导

根据具体情况，组织在确定是否使用特定人工智能系统时可能会考虑很多因素。无论人工智能系统是由组织本身开发还是来自第三方，组织都应该清楚这些考虑因素是什么，并制定政策来解决这些问题。一些例子是：

所需的批准；

成本（包括持续监控和维护）；

批准的采购要求；

适用于组织的法律要求。

如果组织已接受使用其他系统、资产等的策略，则可以根据需要合并这些策略。

B.9.3 负责任地使用人工智能系统的目标

控制

组织应确定并记录目标，以指导负责任地使用人工智能系统。

实施指导

在不同环境下运营的组织对于人工智能系统的负责任开发可能有不同的期望和目标。根据其背景，组织应确定与负责任使用相关的目标。一些目标包括：

- 公平；

问责制；

- 透明度；

可解释性；

可靠性；

- 安全；

鲁棒性和冗余性；

隐私和安全；

可达性。

一旦定义，组织应实施机制以在组织内实现其目标。这可以包括确定第三方解决方案是否满足组织的目标，或者内部开发的解决方案是否适用于预期用途。组织应确定在人工智能系统生命周期的哪些阶段应纳入有意义的人类监督目标。这可以包括：

让人类审查员检查人工智能系统的输出，包括有权推翻人工智能系统做出的决定；

根据与人工智能系统预期部署相关的说明或其他文件，如果需要人工智能系统的可接受使用，确保包括人工监督；

监控人工智能系统的性能，包括人工智能系统输出的准确性；

- 报告与人工智能系统的输出相关的问题及其对相关利益相关者的影响
派对；

- 报告对人工智能系统的性能或能力变化的担忧，以做出正确的决定
生产数据的输出；

- 考虑自动化决策是否适合采取负责任的方法
人工智能系统的使用以及人工智能系统的预期用途。

人工智能系统影响评估可以告知是否需要人工监督（见B.5）。应告知参与与人工智能系统相关的人类监督活动的人员并对其进行培训

并了解人工智能系统的指令和其他文件以及它们为满足人类监督目标而履行的职责。报告性能问题时，人工监督可以增强自动监控。

其他信息

附录 C提供了管理风险的组织目标的示例，这对于确定人工智能系统的使用目标很有用。

B.9.4 AI 系统的预期用途

控制

组织应确保按照人工智能系统及其随附文件的预期用途使用人工智能系统。

实施指导

应根据与 AI 系统相关的说明和其他文档部署 AI 系统（见B.8.2）。部署可能需要特定资源来支持部署，包括需要确保根据需要应用人工监督（参见B.9.3）。为了使 AI 系统得到可接受的使用，AI 系统所使用的数据必须与 AI 系统相关的文档保持一致，以确保 AI 系统性能准确。

应监控人工智能系统的运行（见B.6.2.6）。如果根据相关指令正确部署人工智能系统引起对相关利益方或组织法律要求的影响的担忧，组织应将其担忧传达给组织内部的相关人员以及任何第三方供应商的人工智能系统。

组织应保留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，这些文件可用于证明人工智能系统正在按预期使用，或帮助沟通与人工智能系统的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、组织的数据保留政策以及数据保留的相关法律要求。

B.10第三方和客户关系

B.10.1 目标

确保组织了解其责任并承担责任，并在人工智能系统生命周期的任何阶段涉及第三方时适当分配风险。

B.10.2 职责分配

控制

组织应确保其人工智能系统生命周期内的责任在组织、其合作伙伴、供应商、客户和第三方之间分配。

实施指导

在人工智能系统生命周期中，责任可以在提供数据的各方、提供算法和模型的各方、开发或使用人工智能系统的各方以及对部分或所有利益相关方负责的各方之间划分。组织应记录参与人工智能系统生命周期的所有各方及其角色，并确定他们的责任。

如果组织向第三方提供人工智能系统，组织应确保采取负责任的方法来开发人工智能系统。请参阅 B.6 中的控制和指导。组织应能够为A提供必要的文件（见B.6.2.7和B.8.2）

系统向相关利益方以及组织提供人工智能系统的第三方提供。

当处理的数据包括 PII 时,责任通常在 PII 处理者和控制者之间划分。 ISO/IEC 29100 提供了有关 PII 控制器和 PII 处理器的更多信息。

如果要保护 PII 的隐私,应考虑 ISO/IEC 27701 中描述的控制措施。根据组织和人工智能系统对 PII 的数据处理活动以及组织在人工智能系统整个生命周期的应用和开发中的角色,组织可以承担 PII 控制者（或联合 PII 控制者）、PII 处理者或两个都。

B.10.3 供应商

控制

组织应建立一个流程,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法。

实施指导

开发或使用人工智能系统的组织可以通过多种方式利用供应商,从采购数据集、机器学习算法或模型,或系统的其他组件（例如软件库）,到整个人工智能系统本身单独使用或作为另一种产品（例如车辆)的一部分。

组织在确定供应商的选择、对这些供应商的要求以及持续监控和管理的级别时,应考虑不同类型的供应商、他们提供的产品以及这可能给整个系统和组织带来的不同程度的风险。需要对供应商进行评估。

组织应记录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的人工智能系统或人工智能系统组件未按预期运行或可能对个人或个人群体或两者以及与组织所采取的人工智能系统负责任方法不一致的社会造成影响组织应要求供应商采取纠正措施。组织可以决定与供应商合作以实现这一目标。

组织应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 顾客

控制

组织应确保其负责任的人工智能系统开发和使用方法考虑到客户的期望和需求。

实施指导

当组织提供与人工智能系统相关的产品或服务时（即,当它本身是供应商时）,组织应该了解客户的期望和需求。这些可以以设计或工程阶段对产品或服务本身的要求的形式出现,也可以以合同要求或一般使用协议的形式出现。一个组织可以拥有多种不同类型的客户关系,并且这些关系都可以有不同的需求和期望。

组织应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提供商的责任在哪里,客户的责任在哪里,同时仍然满足需求和期望。

例如,组织可以识别客户使用其人工智能产品和服务相关的风险,并可以决定通过向客户提供适当的信息来处理已识别的风险,以便客户可以处理相应的风险。

作为适当信息的示例,当人工智能系统对于特定使用领域有效时,应将该领域的限制传达给客户。参见B.6.2.7和B.8.2。

附录C
（资料性）

潜在的人工智能相关组织目标和风险来源

C.1 概述

本附件概述了组织在管理风险时可以考虑的潜在组织目标、风险来源和描述。本附件并非详尽无遗或适用于每个组织。组织应确定相关的目标和风险源。 ISO/IEC 23894 提供了有关这些目标和风险源及其与风险管理的关系的更详细信息。对人工智能系统进行初步、定期和必要时的评估，可以提供证据表明正在根据组织目标评估人工智能系统。

C.2 目标

C.2.1 问责制

人工智能的使用可以改变现有的问责框架。以前人们要对其行为负责，而现在他们的行为可以得到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

需要选择一批具有跨学科技能和评估、开发和部署人工智能系统专业知识的专门专家。

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

人工智能的使用会对环境产生积极和消极的影响。

C.2.5 公平性

人工智能系统在自动化决策中的不当应用可能对特定个人或群体不公平。

C.2.6 可维护性

可维护性与组织处理人工智能系统修改以纠正缺陷或适应新要求的能力有关。

C.2.7 隐私

滥用或披露个人和敏感数据（例如健康记录）可能会对数据主体产生有害影响。

C.2.8 稳健性

在人工智能中,鲁棒性属性表明系统在新数据上具有与在其训练数据或典型操作数据上相当的性能的能力（或无能力）。

C.2.9 安全

安全是指系统在规定条件下不会导致人类生命、健康、财产或环境受到危害的状态。

C.2.10 安全

在人工智能背景下,特别是基于机器学习方法的人工智能系统,应考虑传统信息和系统安全问题之外的新安全问题。

C.2.11 透明度和可解释性

透明度既与运行人工智能系统的组织的特征有关,也与这些系统本身有关。可解释性涉及对影响人工智能系统结果的重要因素的解释,这些因素以人类可以理解的方式提供给感兴趣的各方。

C.3 风险来源

C.3.1 环境的复杂性

当人工智能系统在复杂环境中运行时,情况范围广泛,性能可能存在不确定性,因此成为风险来源（例如自动驾驶的复杂环境）。

C.3.2 缺乏透明度和可解释性

无法向相关方提供适当的信息可能是风险的来源（即在组织的可信度和责任方面）。

C.3.3 自动化水平

自动化水平可能会对各个关注领域产生影响,例如安全、公平或安保。

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收集数据的过程可能是风险来源,因为它们可能会影响安全性和稳健性等目标（例如,由于数据质量或数据中毒问题）。

C.3.5 系统硬件问题

与硬件相关的风险源包括基于缺陷组件的硬件错误或在不同系统之间传输经过训练的机器学习模型。

C.3.6 系统生命周期问题

风险源可能出现在整个人工智能系统生命周期中（例如设计缺陷、部署不充分、缺乏维护、退役问题）。

C.3.7 技术准备情况

风险源可能与由于未知因素（例如系统限制和边界条件、性能漂移）而导致的不太成熟的技术有关，也可能与由于技术自满而导致技术更加成熟有关。

附录D
（资料性）

跨领域或部门使用人工智能管理系统

D.1 概述

该管理体系适用于任何开发、提供或使用利用人工智能系统的产品或服务的组织。因此，它可能适用于不同行业的各种产品和服务，这些产品和服务受到对相关方的义务、良好实践、期望或合同承诺的约束。部门的例子有：

- 健康；
- 防御；
- 运输；
- 金融；
- 就业；
- 活力。

为了负责任地开发和使用人工智能系统，可以考虑各种组织目标（有关可能的目标，请参阅附录 C）。本文件从人工智能技术特定角度提供了要求和指导。对于一些潜在目标，存在通用或特定部门的管理体系标准。这些管理体系标准通常从技术中立的角度考虑目标，而人工智能管理体系则提供人工智能技术的具体考虑因素。

人工智能系统不仅由使用人工智能技术的组件组成，而且可以使用多种技术和组件。因此，负责任地开发和使用人工智能系统不仅需要考虑人工智能特定的考虑因素，还需要考虑整个系统以及所使用的所有技术和组件。即使对于人工智能技术特定部分，也应该考虑人工智能特定因素之外的其他方面。例如，人工智能是一种信息处理技术，信息安全也普遍适用于它。对于人工智能和系统的其他组成部分，安全、安保、隐私和环境影响等目标应该进行整体管理，而不是分开管理。因此，将人工智能管理系统与相关主题的通用或特定部门管理系统标准相集成对于负责任地开发和使用人工智能系统至关重要。

D.2 AI管理系统与其他管理系统标准的集成

在提供或使用人工智能系统时，组织可以具有与其他管理体系标准主题相关的目标或义务。例如，这些主题可以包括安全、隐私、质量以及 ISO/IEC 27001、ISO/IEC 27701 和 ISO 9001 中分别涵盖的主题。

在提供、使用或开发人工智能系统时，潜在的相关通用管理系统标准（但不限于此）包括：

ISO/IEC 27001：在大多数情况下，安全性是利用人工智能系统实现组织目标的关键。组织追求安全目标的方式取决于其环境和自身的策略。如果组织确定需要实施人工智能管理系统

为了以类似彻底和系统的方式解决安全目标,可以实施符合ISO/IEC 27001的信息安全管理体系。鉴于ISO/IEC 27001和人工智能管理体系都采用高层结构,它们的集成使用起来很方便,并且对组织有很大好处。在这种情况下,实施本文件中（部分)与信息安全相关的控制的方式（参见B.6.1.2)可以与组织实施 ISO/IEC 27001 相结合。

ISO/IEC 27701:在许多上下文和应用领域,PII 由人工智能系统处理。然后,组织可以遵守适用的隐私义务以及自己的政策和目标。同样,对于 ISO/IEC 27001,组织可以从 ISO/IEC 27701 与人工智能管理系统的集成中受益。 AI 管理系统的隐私相关目标和控制（参见B.2.3和B.5.4）可以与 ISO/IEC 27701 的组织实施相集成。

ISO 9001:对于许多组织来说,遵守 ISO 9001 是他们以客户为导向并真正关心内部有效性的关键标志。 ISO9001 独立合格评定可促进跨组织的业务并激发客户对产品或服务的信心。当涉及人工智能技术的人工智能管理系统与 ISO 9001 联合实施时,可以极大地增强客户对组织或人工智能系统的信心水平。 AI 管理系统可以补充 ISO 9001 要求（风险管理、软件开发、供应链一致性等）,帮助组织实现其目标。

除了上述通用管理系统标准外,人工智能管理系统还可以与行业专用管理系统联合使用。例如,ISO 22000 和人工智能管理系统都与用于食品生产、准备和物流的人工智能系统相关。另一个例子是 ISO 13485。人工智能管理系统的实施可以支持 ISO 13485 中与医疗设备软件相关的要求或其他要求

医疗领域的国际标准,例如 IEC 62304。

参考书目

[1] ISO 8000-2,数据质量 第 2 部分:词汇

[2] ISO 9001,质量管理体系 要求

[3] ISO 9241-210,人机交互的人体工程学 - 第 210 部分:交互系统的以人为本的设计

[4] ISO 13485,医疗器械 质量管理体系 监管要求

[5] ISO 22000,食品安全管理体系 对食品链中任何组织的要求

[6] IEC 62304,医疗设备软件 软件生命周期流程

[7] ISO/IEC 指南 51,安全方面 将其纳入标准的指南

[8] ISO/IEC TS 4213,信息技术 人工智能 机器学习分类性能评估

[9] ISO/IEC 5259（所有部分2）,分析和机器学习 (ML) 的数据质量

[10] ISO/IEC 5338,信息技术 人工智能 人工智能系统生命周期过程

[11] ISO/IEC 17065,合格评定 对产品、流程认证机构的要求
和服务

[12] ISO/IEC 19944-1,云计算和分布式平台 — 数据流、数据类别和数据
使用 第 1 部分:基础知识

[13] ISO/IEC 23053,使用机器学习 (ML) 的人工智能 (AI) 系统框架

[14] ISO/IEC 23894,信息技术 人工智能 风险管理指南

[15] ISO/IEC TR 24027,信息技术 人工智能 (AI) AI 系统中的偏差和
人工智能辅助决策

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

[17] ISO/IEC TR 24368,信息技术 人工智能 道德和道德概述
社会关注

[18] ISO/IEC 25024,系统和软件工程 系统和软件质量要求
和评估 (SQuaRE) 数据质量测量

[19] ISO/IEC 25059,软件工程 系统和软件质量要求和
评估 (SQuaRE) AI 系统的质量模型

[20] ISO/IEC 27000:2018,信息技术 安全技术 信息安全
管理系统 概述和词汇

[21] ISO/IEC 27701,安全技术 ISO/IEC 27001 和 ISO/IEC 27002 隐私扩展
信息管理 要求和指南

[22] ISO/IEC 27001,信息安全、网络安全和隐私保护 信息安全管理体系 要求

[23] ISO/IEC 29100,信息技术 安全技术 隐私框架

[24] ISO 31000:2018,风险管理　指南

[25] ISO 37002,举报管理体系　指南

[26] ISO/IEC 38500:2015,信息技术　组织的 IT 治理

[27] ISO/IEC 38507,信息技术　IT 治理　组织使用人工智能的治理影响

[28] 生命周期 DDI 3.3,2020-04-15。数据文档倡议 (DDI) 联盟。 [浏览时间:2022-02-19]。网址： https: //ddialliance.org/Specification/DDI-Lifecycle/3.3/

[29] 风险框架 NIST-AI 1.0,2023-01-26。美国国家技术研究所 (NIST) [查看于 2023 年 4 月 17 日] https://www.nist.gov/itl/ai-risk-management-framework

ICS 03.100.70； 35.020价
格基于 51 页