Check for updates

NIST AI 100-1



人工智能风险管理框架（AI RMF 1.0）

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

本译本由机器翻译生成，供快速理解条文与内部培训使用；
涉及合规判断或法律适用时，请咨询专业顾问，并以官方英文文本为准。
翻译您自己的文件，请选择桐鸾译馆 · https://www.ivyivy.com.cn

扫码关注服务号

NIST AI 100-1

人工智能风险管理框架(AI RMF 1.0)

本出版物可从以下网址免费获取：
https://doi.org/10.6028/NIST.AI.100-1

2023年1月

美国商务部吉娜·*M*·雷蒙多，部长

美国国家标准与技术研究院
*Laurie E.Locascio,NIST*主任兼商务部标准与技术副部长

本文件中可能会指明某些商业实体、设备或材料，以便充分描述实验程序或概念。此类指明并非意图暗示国家标准化与技术研究院的推荐或认可，也非意图暗示这些实体、材料或设备必然是用于该目的的最佳选择。

本出版物可从以下网址免费获取：https://doi.org/10.6028/NIST.AI.100-1

更新计划与版本

人工智能风险管理框架(AI RMF)旨在成为一份动态更新的文件。

NIST 将定期审查框架的内容和实用性，以确定是否适合更新；预计不迟于 2028 年进行一项包含 AI 社区正式意见的审查。框架将采用双数字版本系统来跟踪和识别重大和次要变更。第一个数字将代表 AI RMF 及其配套文档的世代(例如 1.0)，并且仅在重大修订时更改。次要修订将使用世代号后的“.n”进行跟踪(例如 1.1)。所有变更将使用版本控制表进行跟踪，该表标识历史记录，包括版本号、变更日期和变更描述。NIST 计划频繁更新 AI RMF Playbook。有关 AI RMF Playbook 的意见可随时通过电子邮件发送至 AIframework@nist.gov，并将每半年审查和整合一次。

目录

执行摘要 1

第1部分：基础信息 4

1 **1 风险界定** 4

1.1 理解并应对风险、影响和危害 4

1.2 人工智能风险管理的挑战 5

1.2.1 风险度量 5

1.2.2 风险容忍度 7

1.2.3 风险优先级排序 7

1.2.4 组织整合与风险管理 8

2 **2 Audience** 9

3 **3 AI 风险与可信度** 12

3.1 有效且可靠 13

3.2 安全 14

3.3 安全与韧性 15

3.4 可问责且透明 15

3.5 可解释性与可解读性 16

3.6 隐私增强 17

3.7 公平--有害偏差得到管理 17

4 **4 AI RMF 的有效性** 19

第2部分：核心与概要 20

5 **5 AI RMF 核心** 20

5.1 治理 21

5.2 映射 24

5.3 度量 28

5.4 管理 31

6 **6 AI RMF 概况** 33

附录 A：图 2 和图 3 中 AI 行为者任务的描述 35

附录 B:AI 风险与传统软件风险有何不同 38

附录 C:AI 风险管理与人机交互 40

附录 D:AI RMF 的属性 42

表目录

表 1 GOVERN 功能的类别和子类别。 22

表 2 **MAP**功能的类别和子类别。 26

表 3 **MEASURE**功能的类别和子类别。 29

表4 **管理**功能的类别和子类别。 32

图目录

图 1 与 AI 系统相关的潜在危害示例。可信人工智能系统及其负责任的使用可以减轻负面风险，并有助于为人、组织和生态系统带来益处。 5

图2 AI系统的生命周期与关键维度。修改自OECD (2022) OECD人工智能系统分类框架--OECD数字经济论文。两个内圈显示人工智能系统的关键维度，外圈显示人工智能生命周期阶段。理想情况下，风险管理工作从应用情境中的规划与设计功能开始，并在人工智能系统生命周期中贯穿执行。代表性人工智能参与者见图3。 10

图3 人工智能生命周期各阶段的人工智能参与者。详细描述见附录A 人工智能参与者任务的描述，包括关于测试、评估、验证和确认任务的细节。请注意，人工智能模型维度(图2)中的人工智能参与者作为最佳实践被分开，构建和使用模型的人员与验证和确认模型的人员相分离。 11

图4 可信人工智能系统的特征。有效与可靠是可信性的必要条件，并作为其他可信性特征的基础。可问责与透明以垂直框显示，因为它与其他所有特征都相关。 12

图5 功能在最高层面组织人工智能风险管理活动，以治理、映射、度量和管理人工智能风险。治理被设计为一项跨领域功能，以告知并融入其他三项功能之中。 20

执行摘要

人工智能(AI)技术具有改变社会和人们生活的巨大潜力--从商业和健康到交通和网络安全，再到环境和我们的星球。AI 技术可以推动包容性经济增长，并支持改善我们世界状况的科学进步。然而，AI 技术也带来风险，可能对个人、群体、组织、社区、社会、环境和地球产生负面影响。与其他类型技术的风险一样，AI 风险可以以多种方式出现，并且可以表征为长期或短期、高概率或低概率、系统性或局部性，以及高影响或低影响。

> AI RMF将人工智能系统称为一种工程化或基于机器的系统，对于给定的一组目标，该系统可以生成影响真实或虚拟环境的预测、建议或决策等输出。人工智能系统被设计为以不同程度的自主性运行(改编自：OECD人工智能建议书：2019;ISO / IEC 22989:2022)。

尽管有无数标准和最佳实践可帮助组织减轻传统软件或基于信息的系统的风险，但 AI 系统带来的风险在许多方面是独特的(见附录 B)。例如，AI 系统可能基于随时间变化的数据进行训练，有时变化显著且出乎意料，以难以理解的方式影响系统功能和可信度。AI 系统及其部署环境通常很复杂，使得在发生故障时难以检测和响应。AI 系统本质上是社会技术性的，这意味着它们受到社会动态和人类行为的影响。AI 风险--以及收益--可能来自技术方面与以下社会因素的相互作用：系统的使用方式、与其他 AI 系统的交互、操作者以及部署的社会背景。

这些风险使人工智能成为对组织和社会而言部署和利用都极具挑战性的技术。如果没有适当的控制措施，人工智能系统可能放大、延续或加剧对个人和社区的不公平或不良后果。有了适当的控制措施，人工智能系统可以减轻和管理不公平的后果。

AI 风险管理是负责任地开发和使用 AI 系统的关键组成部分。负责任的 AI 实践有助于使有关 AI 系统设计、开发和使用的决策与预期目标和价值观保持一致。负责任的 AI 核心概念强调以人为本、社会责任和可持续性。AI 风险管理可以通过促使设计、开发和部署 AI 的组织及其内部团队更批判性地思考背景以及潜在或意外的负面和正面影响，来推动负责任的使用和实践。理解和管理 AI 系统的风险将有助于增强可信度，进而培养公众信任。

社会责任可以指组织“通过透明和道德的行为对其决策和活动对社会和环境的影响”的责任(ISO 26000:2010)。可持续性是指“全球系统(包括环境、社会和经济方面)的状态，在这种状态下，满足当代人的需求而不损害后代人满足其自身需求的能力”(ISO / IEC TR 24368:2022)。负责任的 AI 旨在产生公平且可问责的技术。期望组织实践按照“专业责任”进行，ISO 将其定义为一种方法，“旨在确保设计、开发或部署 AI 系统和应用或基于 AI 的产品或系统的专业人员认识到他们在影响人、社会和 AI 未来方面的独特地位”(ISO / IEC TR 24368:2022)。

根据2020年《国家人工智能倡议法案》(P.L.116-283)的指示，AI RMF的目标是为设计、开发、部署或使用人工智能系统的组织提供一种资源，以帮助管理人工智能的诸多风险，并促进可信且负责任的人工智能系统开发与使用。该框架旨在成为**自愿性**、权利保护性、非特定行业且与用例无关的框架，为所有规模、所有行业以及全社会中的组织提供灵活性，以实施框架中的方法。

该框架旨在为组织和个人--此处称为人工智能参与者--提供增强人工智能系统可信度的方法，并帮助促进人工智能系统在长期内的负责任设计、开发、部署和使用。人工智能参与者被经济合作与发展组织(OECD)定义为“在人工智能系统生命周期中发挥积极作用的各方，包括部署或运营人工智能的组织和个人”[OECD(2019) Artificial Intelligence in Society-OECD iLibrary](见附录A)。

AI RMF旨在具有实用性，随着人工智能技术的持续发展而适应人工智能格局，并由组织以不同程度的实施和能力加以操作化，使社会既能从人工智能中受益，又能免受其潜在危害。

框架及支持资源将根据不断发展的技术、世界各地的标准格局以及人工智能社区的经验和反馈进行更新、扩展和改进。NIST将继续使AI RMF及相关指南与适用的国际标准、指南和实践保持一致。随着AI RMF的投入使用，将汲取更多经验教训，为未来的更新和额外资源提供参考。

框架分为两部分。第1部分讨论组织如何构建与人工智能相关的风险框架，并描述目标受众。接下来，分析人工智能风险和可信性，概述可信人工智能系统的特征，包括

有效且可靠、安全、安保且有韧性、可问责且透明、可解释且可理解、隐私增强，以及公平且其有害偏见得到管理。

第2部分构成本框架的“核心”。它描述了四个具体功能，以帮助组织在实践中应对人工智能系统的风险。这些功能--**治理**、**映射**、**度量**和**管理**--进一步细分为类别和子类别。虽然**治理**适用于组织人工智能风险管理流程和程序的所有阶段，但**映射**、**度量**和**管理**功能可应用于特定于人工智能系统的情境以及人工智能生命周期的特定阶段。

与框架相关的额外资源包含在AI RMF Playbook中，可通过NIST AI RMF网站获取：https://www.nist.gov/itl/ai-risk-management-framework.

NIST与私营和公共部门合作开发AI RMF，是受2020年《国家人工智能倡议法案》、国家人工智能安全委员会建议以及《联邦参与制定技术标准及相关工具计划》的指示并与其所呼吁的更广泛人工智能工作保持一致。在本框架开发过程中与人工智能社区的互动--通过对正式信息请求的回应、三次广泛参与的研讨会、对概念文件和两版框架草案的公众评论、多个公共论坛上的讨论以及许多小组会议--为AI RMF 1.0的开发以及NIST和其他方开展的人工智能研发与评估提供了信息。将增强本框架的优先研究和附加指南将被纳入相关的AI Risk Management Framework Roadmap,NIST和更广泛的社区均可为其作出贡献。

第1部分：基础信息

1. 风险界定

人工智能风险管理提供了一条途径，以最大限度地减少人工智能系统的潜在负面影响，例如对公民自由和权利的威胁，同时提供最大化积极影响的机会。有效应对、记录和管理人工智能风险及潜在负面影响，可以带来更可信的人工智能系统。

1.1 理解并应对风险、影响与危害

在AI RMF的语境中，风险是指事件发生概率与相应事件后果的严重程度或程度的综合度量。人工智能系统的影响或后果可以是积极的、消极的，或两者兼有，并可能带来机遇或威胁(改编自：ISO 31000:2018)。在考虑潜在事件的负面影响时，风险是以下两者的函数：1)如果情况或事件发生将产生的负面影响或伤害程度，以及2)发生的可能性(改编自：OMB Circular A-130:2016)。负面影响或伤害可由个人、群体、社区、组织、社会、环境和地球承受。

> “风险管理是指为针对风险指导和掌控组织而进行的协调活动”(来源：ISO 31000:2018)。

虽然风险管理流程通常针对负面影响，但本框架提供了最小化人工智能系统预期负面影响并识别最大化积极影响机会的方法。有效管理潜在危害风险可以带来更可信的人工智能系统，并释放对个人(个体、社区和社会)、组织和系统/生态系统的潜在益处。风险管理能够使人工智能开发者和用户理解影响，并对其模型和系统中固有的局限性和不确定性负责，这反过来可以提高整体系统性能和可信度，以及人工智能技术以有益方式被使用的可能性。

AI RMF旨在在新风险出现时加以应对。这种灵活性在影响不易预见且应用不断演变的情况下尤为重要。虽然某些人工智能风险和益处是众所周知的，但评估负面影响和危害程度可能具有挑战性。图1提供了可能与人工智能系统相关的潜在危害示例。

人工智能风险管理工作应考虑，人类可能假定人工智能系统在所有环境中都能工作--并且工作良好。例如，无论正确与否，人工智能系统常被认为比人类更客观，或比一般软件提供更强的能力。



图1.与人工智能系统相关的潜在危害示例。可信人工智能系统及其负责任的使用可以减轻负面风险，并有助于为人、组织和生态系统带来益处。

1.2 人工智能风险管理的挑战

下文描述了若干挑战。在管理风险以追求人工智能可信性时，应将这些挑战纳入考虑。

1.2.1 风险度量

未明确定义或未充分理解的人工智能风险或故障难以定量或定性度量。无法适当度量人工智能风险并不意味着人工智能系统必然构成高风险或低风险。一些风险度量挑战包括：

与第三方软件、硬件和数据相关的风险：第三方数据或系统可以加速研发并促进技术转化。它们也可能使风险度量复杂化。风险可能来自第三方数据、软件或硬件本身及其使用方式。开发人工智能系统的组织所使用的风险指标或方法可能与部署或运营该系统的组织所使用的风险指标或方法不一致。此外，开发人工智能系统的组织可能对其使用的风险指标或方法不够透明。客户如何将第三方数据或系统使用或集成到人工智能产品或服务中，尤其是在缺乏足够的内部治理结构和技术保障措施的情况下，可能使风险度量和管理复杂化。无论如何，各方和人工智能参与者应管理其作为独立或集成组件开发、部署或使用的人工智能系统中的风险。

跟踪新兴风险：通过识别和跟踪新兴风险并考虑度量这些风险的技术，组织的风险管理工作将得到加强。

人工智能系统影响评估方法可以帮助人工智能参与者理解特定情境中的潜在影响或危害。

可靠指标的可用性：当前对于风险和可信度的稳健且可验证的度量方法及其对不同人工智能用例的适用性缺乏共识，这是一项人工智能风险度量挑战。在寻求度量负面风险或危害时，潜在的陷阱包括：指标的开发往往是一项制度性工作，可能无意中反映与底层影响无关的因素。此外，度量方法可能被过度简化、被操纵、缺乏关键细微差别、以意想不到的方式被依赖，或未能考虑受影响群体和情境的差异。

对人群影响的度量方法，如果能够认识到情境很重要、危害可能对不同群体或子群体产生不同影响，以及可能受到伤害的社区或其他子群体并不总是系统的直接用户，则效果最佳。

人工智能生命周期不同阶段的风险：在人工智能生命周期较早阶段度量风险可能产生与较晚阶段度量风险不同的结果；某些风险在特定时间点可能是潜在的，并可能随着人工智能系统的适应和演化而增加。此外，人工智能生命周期中不同的人工智能参与者可能具有不同的风险视角。例如，提供人工智能软件(如预训练模型)的人工智能开发者可能具有与负责在特定用例中部署该预训练模型的人工智能参与者不同的风险视角。此类部署者可能未意识到其特定用途可能带来与初始开发者所感知的不同风险。所有相关的人工智能参与者共同承担设计、开发和部署适合用途的可信人工智能系统的责任。

真实环境中的风险：虽然在实验室或受控环境中度量人工智能风险可能在部署前产生重要见解，但这些度量可能与在操作性的真实环境中出现的风险不同。

不可测性：不可测的人工智能系统会使风险度量复杂化。不可测性可能是人工智能系统不透明的本质(可解释性或可理解性有限)、人工智能系统开发或部署中缺乏透明度或文档，或人工智能系统固有的不确定性所导致的结果。

人类基线：对旨在增强或替代人类活动(例如决策)的人工智能系统进行风险管理，需要某种形式的基线度量以进行比较。这难以系统化，因为人工智能系统执行的任务与人类不同--且执行任务的方式也与人类不同。

1.2.2 风险容忍度

虽然 AI RMF 可用于确定风险优先级，但它并不规定风险容忍度。风险容忍度是指组织或人工智能参与者(见附录 A)为实现其目标而承担风险的准备程度。风险容忍度可能受到法律或监管要求的影响(改编自：ISO GUIDE 73)。风险容忍度以及组织或社会可接受的风险水平高度依赖于情境和具体应用及用例。风险容忍度可能受到人工智能系统所有者、组织、行业、社区或政策制定者所制定的政策和规范的影响。风险容忍度可能随着人工智能系统、政策和规范的演进而随时间变化。不同组织可能因其特定的组织优先事项和资源考量而具有不同的风险容忍度。

用于更好地为危害/成本效益权衡提供信息的新兴知识和方法将继续由企业、政府、学术界和公民社会开发和辩论。在明确人工智能风险容忍度方面的挑战仍未解决的范围内，可能存在某些情境，其中风险管理框架尚不能 readily 适用于缓解负面人工智能风险。

该框架旨在保持灵活性，并增强现有的风险实践，这些实践应与适用的法律、法规和规范保持一致。组织应遵循由组织、领域、学科、行业或专业要求所确立的现有法规和指南，涉及风险标准、容忍度和响应。某些行业或产业可能已确立损害的定义或既定的文档、报告和披露要求。在行业内，风险管理可能取决于针对特定应用和用例场景的现有指南。在不存在既定指南的情况下，组织应定义合理的风险容忍度。一旦定义了容忍度，即可使用本 AI RMF 来管理风险并记录风险管理过程。

1.2.3 风险优先级排序

试图完全消除负面风险在实践中可能适得其反，因为并非所有事件和故障都能被消除。对风险的不切实际期望可能导致组织以低效或不切实际的方式进行风险分类，或浪费稀缺资源。风险管理文化可以帮助组织认识到并非所有 AI 风险都相同，从而可以有目的地分配资源。可操作的风险管理工作为评估组织开发或部署的每个人工智能系统的可信度制定了明确的指南。政策和资源应根据所评估的风险水平和人工智能系统的潜在影响来确定优先级。人工智能系统可由 AI 部署者根据特定使用情境进行定制或调整的程度，可能是一个促成因素。

在应用 AI RMF 时，组织确定为特定使用情境下人工智能系统最高风险的风险，需要最紧迫的优先级排序和最彻底的风险管理过程。在人工智能系统呈现不可接受的负面风险水平的情况下--例如重大负面影响即将发生、严重损害正在实际发生或存在灾难性风险--应以安全方式停止开发和部署，直至风险能够得到充分管理。如果人工智能系统的开发、部署和用例在特定情境中被发现为低风险，这可能表明潜在的优先级较低。

设计或部署用于直接与人类交互的人工智能系统与不直接与人类交互的人工智能系统之间，风险优先级排序可能有所不同。在人工智能系统使用由敏感或受保护数据(如个人可识别信息)组成的大型数据集进行训练的情况下，或人工智能系统的输出对人类产生直接或间接影响的情况下，可能需要更高的初始优先级。设计为仅与计算系统交互并使用非敏感数据集(例如从物理环境收集的数据)训练的人工智能系统，可能需要较低的初始优先级。尽管如此，基于情境定期评估和优先排序风险仍然很重要，因为非面向人类的人工智能系统可能具有下游安全或社会影响。

剩余风险--定义为风险处理后剩余的风险(来源：ISO GUIDE 73)--直接影响最终用户或受影响的个人和社区。记录剩余风险将要求系统提供者充分考虑部署人工智能产品的风险，并将告知最终用户与系统交互的潜在负面影响。

1.2.4 组织整合与风险管理

AI 风险不应被孤立地考虑。不同的人工智能参与者根据其在生命周期中的角色承担不同的责任和认知。例如，开发人工智能系统的组织通常不会掌握该系统可能如何被使用的信息。AI 风险管理应被整合并纳入更广泛的企业风险管理策略和流程中。将 AI 风险与网络安全和隐私等其他关键风险一并处理，将产生更整合的成果和组织效率。

AI RMF 可与相关指南和框架一起使用，以管理人工智能系统风险或更广泛的企业风险。与人工智能系统相关的一些风险在其他类型的软件开发和部署中也很常见。重叠风险的示例包括：与使用底层数据训练人工智能系统相关的隐私问题；与资源密集型计算需求相关的能源和环境影响；与系统及其训练和输出数据的机密性、完整性和可用性相关的安全问题；以及人工智能系统底层软件和硬件的一般安全性。

组织需要建立并维护适当的问责机制、角色与职责、文化以及激励结构，才能使风险管理有效。仅使用 AI RMF 不会带来这些变化，也不会提供适当的激励。有效的风险管理通过高层领导的组织承诺来实现，并且可能需要在组织或行业内部进行文化变革。此外，与大型组织相比，管理人工智能风险或实施 AI RMF 的中小型组织可能因其能力和资源不同而面临不同的挑战。

2. Audience

识别和管理人工智能风险及潜在影响--包括正面和负面--需要人工智能生命周期中广泛的视角和参与者。理想情况下，人工智能参与者应代表多样化的经验、专业知识和背景，并组成人口统计和学科多样化的团队。AI RMF 旨在供人工智能生命周期和各个维度中的人工智能参与者使用。

OECD 制定了一个框架，根据五个关键社会技术维度对人工智能生命周期活动进行分类，每个维度都具有与人工智能政策和治理(包括风险管理)相关的属性[OECD(2022) OECD Framework for the Classification of AI systems - OECD Digital Economy Papers]。图 2 显示了这些维度，NIST 为本框架的目的稍作修改。NIST 的修改强调了测试、评估、验证与确认(TEVV)过程在整个人工智能生命周期中的重要性，并概括了人工智能系统的运行背景。

图 2 中显示的人工智能维度是应用情境、数据和输入、人工智能模型，以及任务和输出。参与这些维度、执行或管理人工智能系统的设计、开发、部署、评估和使用并推动人工智能风险管理努力的人工智能参与者是主要 AI RMF 受众。

图 3 列出了生命周期维度中具有代表性的人工智能参与者，并在附录 A 中详细描述。在 AI RMF 中，所有人工智能参与者共同努力管理风险，实现可信和负责任的人工智能目标。具备 TEVV 专业知识的人工智能参与者贯穿整个人工智能生命周期，尤其可能从本框架中受益。定期执行 TEVV 任务可以提供与技术、社会、法律和伦理标准或规范相关的见解，并有助于预测影响以及评估和跟踪新兴风险。作为人工智能生命周期内的常规过程，TEVV 允许进行中途补救和事后风险管理。

图 2 中心的人与地球维度代表人权以及社会和地球更广泛的福祉。该维度中的人工智能参与者构成一个单独的 AI RMF 受众，他们告知主要受众。这些人工智能参与者可能包括行业协会、标准制定组织、研究人员、倡导团体、

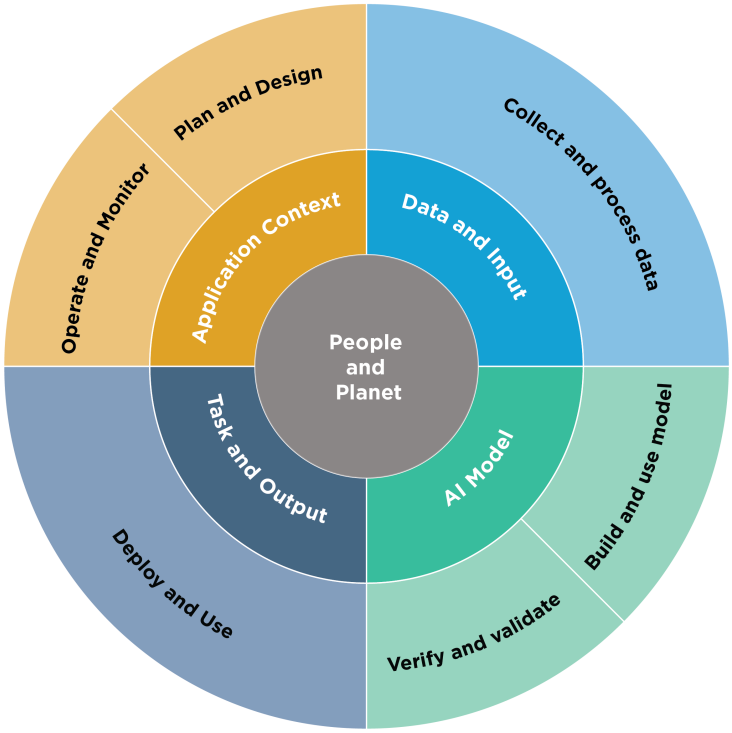


图 2.人工智能系统的生命周期和关键维度。修改自 OECD(2022)OECD 人工智能系统分类框架--OECD 数字经济论文。两个内圈显示人工智能系统的关键维度，外圈显示人工智能生命周期阶段。理想情况下，风险管理努力从应用情境中的规划和设计功能开始，并在人工智能系统生命周期中持续进行。有关代表性人工智能参与者，见图 3。

环境团体、公民社会组织、最终用户以及可能受影响的个人和社区。这些参与者可以：

- 协助提供背景信息并理解潜在和实际影响；
- 成为人工智能风险管理正式或准正式规范和指导的来源；
- 为人工智能运行指定边界(技术、社会、法律和伦理方面)；以及
- 促进关于平衡与社会价值观和优先事项(涉及公民自由和权利、公平、环境和地球以及经济)所需权衡的讨论。

成功的风险管理取决于图 3 所示人工智能参与者之间的集体责任感。第 5 节描述的 AI RMF 功能需要不同的视角、学科、专业和经验。多元化的团队有助于更开放地分享关于技术目的和功能的观点和假设--使这些隐性方面更加明确。这种更广泛的集体视角为发现问题以及识别现有和新兴风险创造了机会。

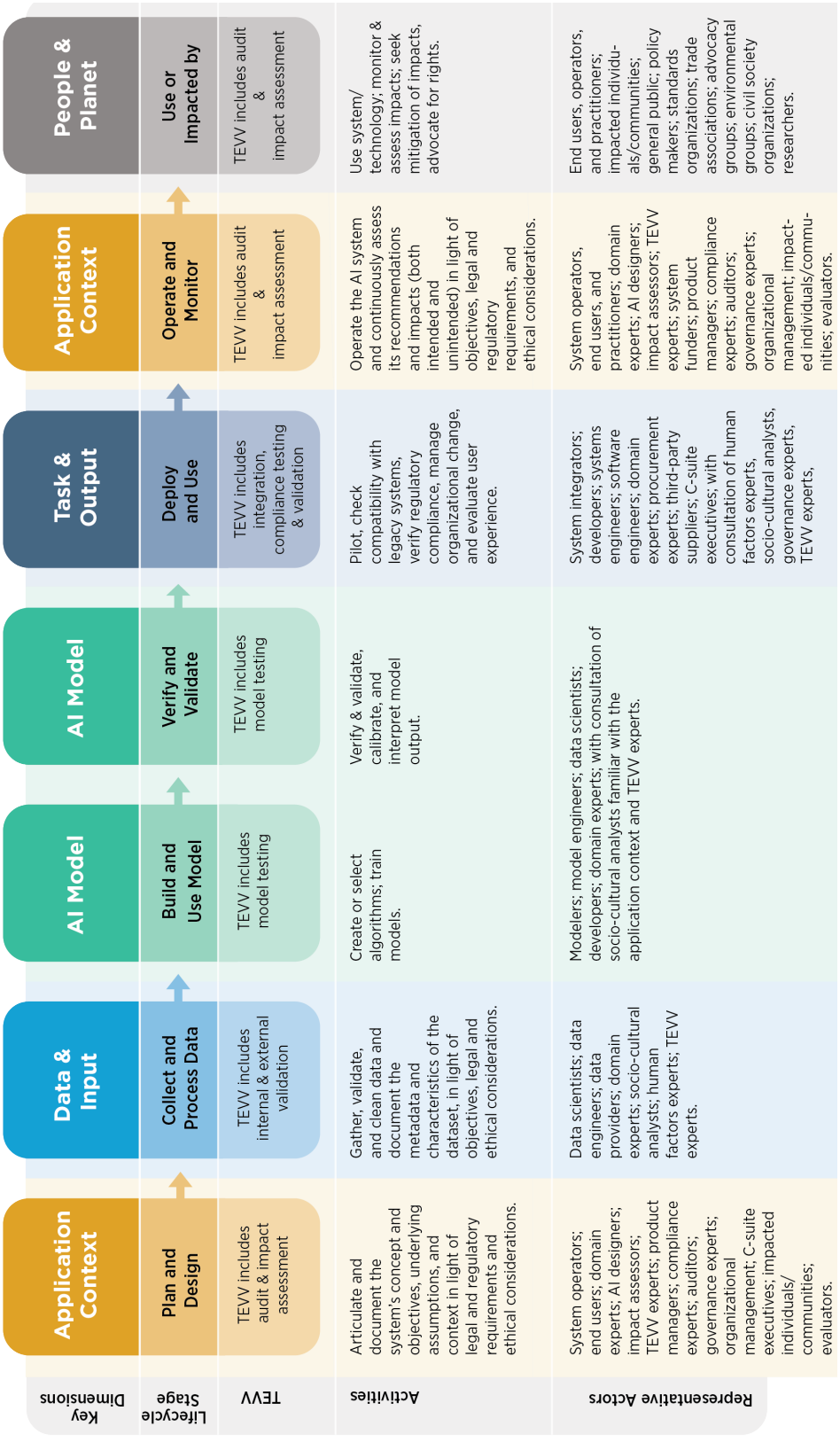


图 3.跨人工智能生命周期阶段的人工智能参与者。有关人工智能参与者任务的详细描述，包括关于测试的详细信息，见附录 A，评估、验证和确认任务。请注意，人工智能模型维度(图 2)中的人工智能参与者被分开作为最佳实践，其中构建和使用模型的人员与验证和确认模型的人员分开。

3. AI 风险与可信度

为使人工智能系统可信，它们通常需要对利益相关方所重视的多重标准作出响应。增强人工智能可信性的方法可以减少负面的人工智能风险。本框架阐述了可信人工智能的以下**特性**，并提供了应对这些特性的指导。可信人工智能系统的特性包括：**有效且可靠、安全、可靠且有韧性、可问责且透明、可解释且可理解、隐私增强，以及公平且有害偏见得到管理。**创建可信人工智能需要根据人工智能系统的使用情境平衡上述每一项特性。虽然所有特性都是社会技术系统属性，但可问责性和透明度还与人工智能系统内部及其外部环境中的过程和活动相关。忽视这些特性会增加负面后果的概率和严重程度。

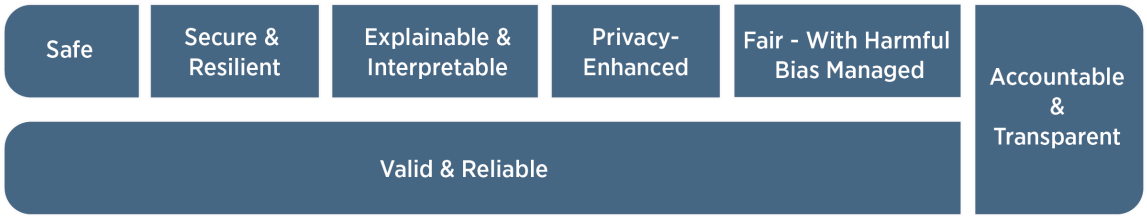


图4.可信人工智能系统的特征。有效且可靠是可信性的必要条件，并作为其他可信性特征的基础。可问责且透明被显示为一个垂直框，因为它与其他所有特征都相关。

可信性特征(如图4所示)与社会和组织行为、人工智能系统所使用的数据集、人工智能模型和算法的选择以及构建者所做的决策，以及与提供洞察和监督此类系统的人类之间的互动密不可分。在决定与人工智能可信性特征相关的具体指标以及这些指标的精确阈值时，应当运用人类判断。

单独处理人工智能可信性特征并不能确保人工智能系统的可信性；通常涉及权衡，很少所有特征都适用于每种情境，而在任何给定情况下某些特征会或多或少地重要。最终，可信性是一个社会概念，跨越一个谱系，其强度仅取决于其最弱的特征。

在管理人工智能风险时，组织在平衡这些特征方面可能面临艰难决策。例如，在某些场景中，优化可解释性与实现隐私之间可能出现权衡。在其他情况下，组织可能面临预测准确性与可解释性之间的权衡。或者，在某些条件下，如数据稀疏时，隐私增强技术可能导致准确性损失，影响决策

关于某些领域中的公平性和其他价值观。处理权衡需要考虑决策背景。这些分析可以突显不同度量之间权衡的存在和程度，但它们不能回答如何驾驭权衡的问题。这些取决于相关背景中起作用的价值观，并应以既透明又适当合理的方式加以解决。

在人工智能生命周期中增强情境意识有多种方法。例如，主题专家可以协助评估TEVV发现结果，并与产品和部署团队合作，使TEVV参数与需求和部署条件保持一致。在资源充足的情况下，在整个人工智能生命周期中增加来自利益相关方和相关人工智能参与者的输入的广度和多样性，可以增加为情境敏感评估提供信息以及识别人工智能系统效益和积极影响的机会。这些做法可以提高适当管理社会情境中产生的风险的可能性。

对可信性特征的理解和处理取决于人工智能参与者在人工智能生命周期中的特定角色。对于任何给定的人工智能系统，人工智能设计者或开发者对特征的看法可能与部署者不同。

本文件中解释的可信性特性相互影响。高度安全但不公平的系统、准确但不透明且不可解释的系统，以及不准确但安全、隐私增强且透明的系统，都是不可取的。全面的风险管理方法要求平衡可信性特性之间的权衡。所有人工智能参与者共同负责确定人工智能技术对于特定情境或目的是否是适当或必要的工具，以及如何负责任地使用它。委托或部署人工智能系统的决定应基于对可信性特性以及相对风险、影响、成本和效益的情境评估，并应听取广泛利益相关方的意见。

3.1 有效且可靠

确认是“通过提供客观证据，确认特定预期用途或应用的要求已得到满足”(来源：ISO 9000:2015)。部署不准确、不可靠或对其训练数据及训练之外的环境泛化能力差的人工智能系统会创造并增加负面人工智能风险，并降低可信性。

可靠性在同一标准中定义为“项目在给定条件下、在给定时间间隔内无故障地按要求执行的能力”(来源：ISO / IEC TS 5723:2022)。可靠性是人工智能系统在预期使用条件下和给定时间段内(包括系统的整个生命周期)整体正确运行的目标。

准确性和稳健性有助于人工智能系统的有效性和可信性，并且在人工智能系统中可能相互紧张。

准确性由ISO / IEC TS 5723:2022定义为“观测、计算或估计结果与真实值或被接受为真实的值之间的接近程度”。准确性度量应考虑以计算为中心的度量(例如，假阳性和假阴性率)、人机协作，并展示外部有效性(可推广到训练条件之外)。准确性测量应始终与明确定义且具有代表性的测试集(代表预期使用条件)以及测试方法细节配对；这些应包含在相关文档中。准确性测量可以包括对不同数据段的结果进行分解。

稳健性或泛化性被定义为“系统在多种情况下保持其性能水平的能力”(来源：ISO / IEC TS 5723:2022)。稳健性是在广泛条件和情况下实现适当系统功能的目标，包括对人工智能系统最初未预料到的使用。稳健性不仅要求系统在预期使用下完全按其所为运行，还要求如果系统在意外环境中运行，其运行方式应尽量减少对人的潜在伤害。

已部署人工智能系统的有效性和可靠性通常通过持续测试或监测来评估，以确认系统按预期运行。对有效性、准确性、鲁棒性和可靠性的度量有助于提升可信度，并应考虑某些类型的故障可能造成更大危害。人工智能风险管理工作应优先考虑最小化潜在负面影响，并且在人工智能系统无法检测或纠正错误的情况下，可能需要纳入人工干预。

3.2 安全

人工智能系统应“在规定的条件下，不会导致人类生命、健康、财产或环境受到危害的状态”(来源：ISO / IEC TS 5723:2022)。人工智能系统的安全运行通过以下方式得到改善：

- 负责任的设计、开发与部署实践；
- 向部署者提供关于系统负责任使用的明确信息；
- 部署方和最终用户做出负责任的决策；以及
- 基于事件经验证据的风险解释和文档记录。

不同类型的安全风险可能需要基于背景和潜在风险的严重程度采取定制的人工智能风险管理方法。构成严重伤害或死亡潜在风险的安全风险需要最紧急的优先处理和最彻底的风险管理过程。

在生命周期中考虑安全因素并尽早从规划和设计开始，可以防止可能导致系统危险的故障或状况。人工智能安全的其他实用方法通常涉及严格的模拟和域内测试、实时监控，以及对偏离预期或期望功能的系统进行关闭、修改或人工干预的能力。

人工智能安全风险管理方法应从交通和医疗等领域的安全工作和指南中汲取经验，并与现有的行业或应用特定指南或标准保持一致。

3.3 安全与韧性

人工智能系统及其部署所在的生态系统，如果能够承受意外的不利事件或其环境或用途的意外变化--或者如果能够在面对内部和外部变化时维持其功能和结构，并在必要时安全、平稳地降级--则可被称为韧性(改编自：ISO / IEC TS 5723:2022)。常见的安全关切涉及对抗性样本、数据投毒，以及通过人工智能系统端点泄露模型、训练数据或其他知识产权。能够通过防止未经授权访问和使用的保护机制维持机密性、完整性和可用性的人工智能系统，可被称为安全。NIST网络安全框架和风险管理框架中的指南属于此处适用的指南之列。

安全性和韧性是相关但不同的特征。韧性是在意外不利事件后恢复正常功能的能力，而安全性包括韧性，但也涵盖避免、防护、响应或从攻击中恢复的协议。韧性与稳健性相关，并超越数据来源，涵盖对模型或数据的意外或对抗性使用(或滥用或误用)。

3.4 可问责且透明

可信人工智能依赖于问责制。问责制以透明度为前提。透明度反映了关于人工智能系统及其输出的信息可供与该系统交互的个人获取的程度--无论他们是否意识到自己正在这样做。有意义的透明度根据人工智能生命周期的阶段提供适当层级的信息获取渠道，并针对人工智能参与者或与人工智能系统交互或使用该系统的个人的角色或知识进行调整。通过促进更高水平的理解，透明度增强了对人工智能系统的信心。

该特征的范围从设计决策和训练数据到模型训练、模型结构、其预期用例，以及部署、部署后或最终用户决策的方式、时间和决策者。透明度通常对于与不正确或以其他方式导致负面影响的人工智能系统输出相关的可操作补救是必要的。透明度应考虑人机互动：例如，

当检测到由人工智能系统引起的潜在或实际不利结果时，如何通知人类操作员或用户。透明的系统不一定是准确的、隐私增强的、安全的或公平的系统。然而，很难确定一个不透明的系统是否具备这些特征，并且随着复杂系统的演变而长期确定这一点也很困难。

在寻求人工智能系统结果的问责时，应考虑人工智能参与者的角色。与人工智能及更广泛的技术系统相关的风险与问责之间的关系因文化、法律、行业和社会背景而异。当后果严重时，例如当生命和自由受到威胁时，人工智能开发者和部署者应考虑按比例且主动地调整其透明度和问责实践。维持减少危害的组织实践和治理结构(如风险管理)有助于形成更负责任的系统。

增强透明度和可问责性的措施还应考虑这些努力对实施实体的影响，包括所需资源水平和保护专有信息的需要。

维护训练数据的来源并支持将人工智能系统的决策归因于训练数据子集，可以同时有助于透明度和可问责性。训练数据也可能受版权约束，并应遵循适用的知识产权法律。

随着人工智能系统透明度工具和相关文档的持续发展，鼓励人工智能系统的开发者与人工智能部署者合作测试不同类型的透明度工具，以确保人工智能系统按预期使用。

3.5 可解释性与可理解性

可解释性指的是对人工智能系统运行背后机制的表示，而可解读性指的是人工智能系统输出在其设计功能目的背景下的含义。可解释性和可解读性共同帮助操作或监督人工智能系统的人员以及人工智能系统的用户更深入地了解系统的功能性和可信度，包括其输出。其基本假设是，对负面风险的感知源于缺乏适当理解或情境化系统输出的能力。可解释和可解读的人工智能系统提供的信息将帮助最终用户理解人工智能系统的目的和潜在影响。

缺乏可解释性带来的风险可以通过描述人工智能系统如何运作来管理，描述应针对个体差异(如用户的角色、知识和技能水平)进行定制。可解释的系统可以更容易地调试和监控，并且有助于更彻底的文档记录、审计和治理。

可解释性风险通常可以通过传达关于人工智能系统为何做出特定预测或建议的描述来解决。(参见“可解释人工智能的四项原则”和“人工智能中可解释性和可理解性的心理学基础”，见此处。)

透明度、可解释性和可理解性是相互支持的 distinct 特征。透明度可以回答系统中“发生了什么”的问题。可解释性可以回答系统中决策是“如何”做出的问题。可理解性可以回答系统为何做出决策及其对用户的意义或背景的问题。

3.6 隐私增强

隐私通常指有助于保障人类自主性、身份和尊严的规范和做法。这些规范和做法通常涉及免受侵扰、限制观察，或个人同意披露或控制其身份各方面(如身体、数据、声誉)的自主权。(参见NIST隐私框架：通过企业风险管理改善隐私的工具。)

隐私价值观，如匿名性、机密性和控制权，通常应指导人工智能系统设计、开发和部署的选择。与隐私相关的风险可能影响安全性、偏见和透明度，并与这些其他特性存在权衡。与安全性和保障性一样，人工智能系统的特定技术特征可能促进或降低隐私。人工智能系统还可能通过允许推断识别个人或个人的先前私密信息，对隐私构成新的风险。

面向人工智能的隐私增强技术(“PETs”)，以及针对某些模型输出的数据最小化方法(如去标识化和聚合)，可以支持隐私增强人工智能系统的设计。在某些条件下，如数据稀疏时，隐私增强技术可能导致准确性损失，影响关于某些领域中公平性和其他价值观的决策。

3.7 公平--有害偏差受控

公平性在人工智能中包括对平等和公正的关注，通过解决有害偏见和歧视等问题。公平的标准可能复杂且难以定义，因为对公平的认知因文化而异，并可能因应用而发生变化。组织的风险管理工作将通过认识和考虑这些差异而得到加强。有害偏见得到缓解的系统不一定公平。例如，在人口统计群体之间预测 somewhat 平衡的系统可能仍然无法为残疾人所访问，或受到数字鸿沟的影响，或可能加剧现有的差距或系统性偏见。

偏见比人口统计平衡和数据代表性更广泛。NIST 已识别出三类主要的人工智能偏见，需要考虑和管理：系统性、计算和统计性以及人类认知性。这些偏见中的每一种都可能在不存在偏见、偏袒或歧视意图的情况下发生。系统性偏见可能存在于人工智能数据集、人工智能生命周期中的组织规范、实践和流程，以及使用人工智能系统的更广泛社会中。计算和统计偏见可能存在于人工智能数据集和算法过程中，并且通常源于由于非代表性样本导致的系统性错误。人类认知偏见涉及个人或群体如何感知人工智能系统信息以做出决策或填补缺失信息，或人类如何思考人工智能系统的目的和功能。人类认知偏见在人工智能生命周期和系统使用的决策过程中无处不在，包括人工智能的设计、实施、操作和维护。

偏见以多种形式存在，并可能根植于帮助做出关于我们生活的决策的自动化系统中。虽然偏见并不总是负面现象，但人工智能系统可能潜在地增加偏见的速度和规模，并延续和放大对个人、群体、社区、组织和社会的影响。偏见与社会中的透明度以及公平性概念密切相关。(有关偏见(包括三个类别)的更多信息，请参见 NIST 特别出版物 1270,迈向识别和管理人工智能偏见的标准。)

4. AI RMF的有效性

对AI RMF有效性的评估--包括衡量人工智能系统可信性底线改进的方法--将与人工智能社区合作，成为NIST未来活动的一部分。

鼓励组织和其他框架用户定期评估 AI RMF 是否提高了他们管理人工智能风险的能力，包括但不限于他们的政策、流程、实践、实施计划、指标、测量和预期成果。NIST 打算与其他人合作开发评估 AI RMF 有效性的指标、方法和目标，并广泛分享结果和支持信息。框架用户预期将受益于：

- 增强治理、映射、度量和管理人工智能风险的流程，并清晰记录结果；
- 提高对可信性特征、社会技术方法和人工智能风险之间关系和权衡的认识；
- 明确的系统调试和部署进行/不进行决策流程；
- 建立政策、流程、实践和程序，以改善与人工智能系统风险相关的组织可问责性工作；
- 增强的组织文化，优先识别和管理人工智能系统风险及其对个人、社区、组织和社会的影响；
- 在组织内部和跨组织之间更好地共享关于风险、决策过程、责任、常见陷阱、TEVV(测试、评估、验证与确认)实践以及持续改进方法的信息；
- 更丰富的上下文知识，以提高对下游风险的认识；
- 加强与利益相关方及相关人工智能参与者的互动；以及
- 增强人工智能系统及相关风险的TEVV能力。

第2部分：核心与概要

5. AI RMF 核心

AI RMF 核心提供了成果和行动，使对话、理解和活动能够管理 AI 风险并负责任地开发可信人工智能系统。如图 5 所示，核心由四个功能组成：**治理、映射、度量、**和**管理**。每个高层级功能被分解为类别和子类别。类别和子类别被细分为具体行动和成果。行动不构成检查清单，也不一定是按顺序排列的步骤集。

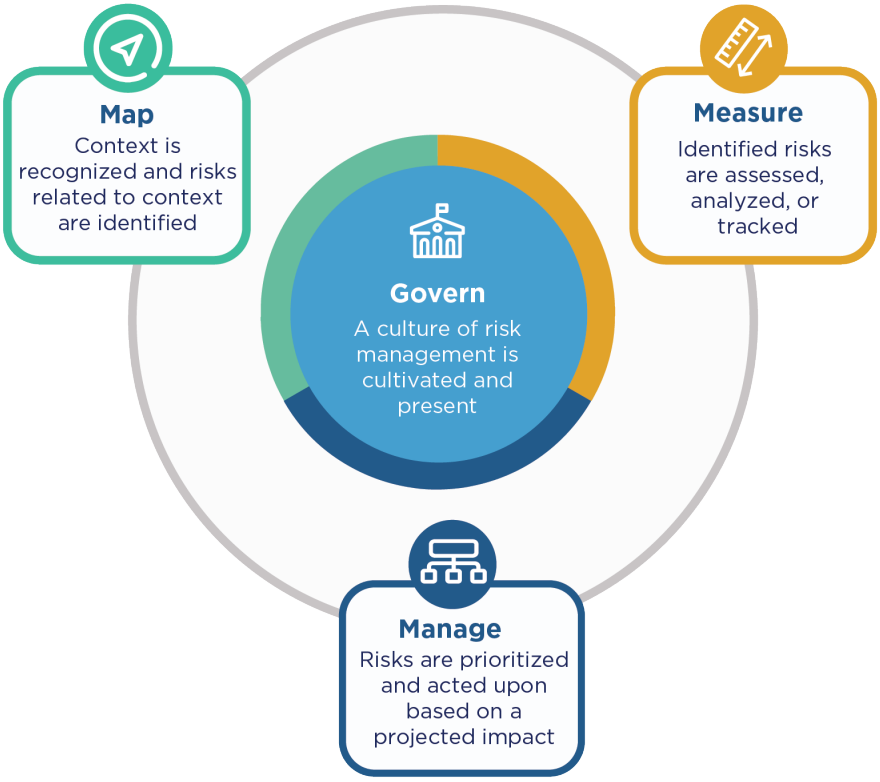


图 5.功能在最高层面组织人工智能风险管理活动，以治理、映射、度量和管理人工智能风险。治理被设计为一项跨领域功能，为其他三项功能提供信息并贯穿其中。

风险管理应当是持续的、及时的，并在人工智能系统生命周期的各个维度中贯穿执行。AI RMF 核心功能应以反映多元化和多学科视角的方式实施，可能包括组织外部人工智能参与者的观点。拥有多元化团队有助于更开放地分享关于所设计、开发、

部署或评估的技术的用途和功能的观点和假设--这可以创造机会来发现问题并识别现有和新兴风险。

NIST AI RMF Playbook 是 AI RMF 的在线配套资源，可帮助组织理解 AI RMF 并通过建议的战术行动实现其成果，这些行动可在自身情境中应用。与 AI RMF 一样，Playbook 是自愿性的，组织可根据自身需求和兴趣使用这些建议。Playbook 用户可从建议材料中创建定制指南供自己使用，并贡献建议与更广泛的社区分享。与 AI RMF 一样，Playbook 是 NIST 可信与负责任人工智能资源中心的一部分。

框架用户可根据自身资源和能力，以最适合其管理 AI 风险需求的方式应用这些功能。一些组织可能选择从类别和子类别中进行选取；其他组织可能选择并有能力应用所有类别和子类别。假设治理结构已到位，功能可在人工智能生命周期中以任何顺序执行，只要框架用户认为能增加价值。在实施**治理**中的成果后，大多数 AI RMF 用户会从**映射**功能开始，然后继续到**度量**或**管理**。无论用户如何整合这些功能，过程应是迭代的，必要时在功能之间交叉参照。同样，有些类别和子类别包含适用于多个功能的要素，或逻辑上应在某些子类别决策之前进行。

5.1 治理

治理功能：

- 在涉及设计、开发、部署、评估或获取人工智能系统的组织内培育和实施风险管理文化；
- 概述用于预测、识别和管理系统可能带来的风险(包括对用户和社会其他成员的风险)的流程、文件和组织方案--以及实现这些成果的程序；
- 纳入评估潜在影响的过程；
- 提供一种结构，使人工智能风险管理功能能够与组织原则、政策和战略优先事项保持一致；
- 将人工智能系统设计和开发的技术方面与组织价值观和原则相连接，并为参与获取、训练、部署和监控此类系统的个人提供组织实践和能力；

和

- 涵盖完整的产品生命周期及相关流程，包括与使用第三方软件或硬件系统和数据有关的法律问题及其他问题。

治理是一项跨领域功能，贯穿于整个人工智能风险管理过程，并赋能该过程中的其他功能。**治理**的各个方面，尤其是与合规或评估相关的方面，应整合到其他各项功能中。对治理的关注是人工智能系统生命周期内有效人工智能风险管理的持续且内在的要求，也贯穿于组织的层级结构。

强有力的治理可以推动和增强内部实践和规范，以促进组织风险文化。治理机构可以确定指导组织使命、目标、价值观、文化和风险容忍度的总体政策。高层领导为组织内的风险管理定下基调，并由此塑造组织文化。管理层将 AI 风险管理的技术方面与政策和运营对齐。文档可以增强透明度、改进人工审查流程，并加强 AI 系统团队的问责制。

在建立**治理**功能中所述的结构、系统、流程和团队之后，组织应受益于一种以风险理解和管理为重点的使命驱动型文化。随着知识、文化以及人工智能参与者的需求或期望随时间演变，框架用户有责任继续执行**治理**功能。

与治理人工智能风险相关的实践在 NIST AI RMF Playbook 中有所描述。表 1 列出了**治理**功能的类别和子类别。

表1:GOVERN功能的类别与子类别。

类别	子类别
GOVERN 1: 政策、流程、程序，以及实践贯穿整个与组织相关对映射而言，度量与管理人工智能已落实风险，透明且已实施有效地。	**GOVERN 1.1:**涉及人工智能的法律和监管要求已被理解、管理并记录在案。 **GOVERN 1.2:**可信人工智能的特征已整合到组织政策、流程、程序以及 实践。 **GOVERN 1.3:**已建立流程、程序和做法，以根据组织的风险容忍度确定所需的风险管理活动水平。 **GOVERN 1.4:**风险管理过程及其结果通过基于组织风险优先事项的透明政策、程序和其他控制措施予以确立。

续下页

表 1:**治理**功能的类别和子类别。(续)

类别	子类别
	GOVERN 1.5:已规划对风险管理过程及其结果的持续监控和定期审查，并明确界定组织角色和职责，包括确定定期审查的频率。 **GOVERN 1.6:**已建立机制来清点人工智能系统，并根据组织风险优先事项配置资源。**GOVERN 1.7:**已建立流程和程序，以安全的方式停用和逐步淘汰人工智能系统，且不增加风险或降低组织的 可信度。
GOVERN 2: 问责制 结构在于 以便 合适的团队 并且个人是 被赋权， 负责，并且 为映射而训练， 度量与 管理人工智能风险。	**GOVERN 2.1:**与映射、度量和管人工智能风险相关的角色、职责和沟通渠道已被记录，并向整个组织中的个人和团队明确传达。 **GOVERN 2.2:**组织的人员和合作伙伴接受人工智能风险管理培训，使其能够按照相关政策、程序和协议履行职责和责任。 **GOVERN 2.3:**组织的行政领导层对与人工智能系统开发和部署相关风险的决策承担责任。
GOVERN 3: 劳动力多样性， 公平、包容、 和可访问性 流程是 在.....中优先考虑 映射， 度量与 管理人工智能 贯穿整个 生命周期。	**GOVERN 3.1:**在整个生命周期中与映射、度量和管人工智能风险相关的决策由多元化团队(例如，人口统计特征、学科、经验、专业知识和背景的多样性)提供信息。 **GOVERN 3.2:**已建立政策和程序，以界定和区分人机配置的角色和职责以及对人工智能系统的监督。
GOVERN 4: 组织 团队已承诺 一种文化	**GOVERN 4.1:**组织已制定相关政策和实践，在人工智能系统的设计、开发、部署和使用中培养批判性思维和安全优先的心态，以最大限度减少潜在的负面影响。

续下页

表 1:**治理**功能的类别和子类别。(续)

类别	子类别
考虑并 传达人工智能 风险。	**GOVERN 4.2:**组织团队记录其设计、开发、部署、评估和使用的人工智能技术的风险和潜在影响，并就影响进行更多沟通 广泛地。 **GOVERN 4.3:**组织已制定相关实践，以支持人工智能测试、事件识别和信息共享。**GOVERN 5.1:**组织已制定相关政策和实践，以收
GOVERN 5: 流程处于 稳健之地 与.....的接触 相关的人工智能行为体。	集、考虑、优先排序并整合来自开发或部署人工智能系统的团队之外的人员关于潜在个人和社会影响的反馈 人工智能风险 。 **GOVERN 5.2:**已建立相关机制，使开发或部署人工智能系统的团队能够定期将相关人工智能参与者经裁定的反馈纳入系统设计和实施。
治理 6:政策 和程序是 已到位以应对 AI 风险与收益 源于 第三方软件 以及数据和其他供应链问题。	**GOVERN 6.1:**已制定相关政策和程序，处理与第三方实体相关的人工智能风险，包括侵犯第三方知识产权或其他权利的风险。 **GOVERN 6.2:**已制定应急流程，以处理第三方数据或人工智能系统的故障或事件 属于高风险。

5.2 Map

映射功能建立背景，以框定与人工智能系统相关的风险。人工智能生命周期由许多相互依赖的活动组成，涉及多样化的参与者(见图 3)。在实践中，负责流程某一部分的 AI 参与者往往对其他部分及其相关背景没有完全的可见性或控制权。这些活动之间以及相关 AI 参与者之间的相互依赖关系，可能使可靠预测人工智能系统的影响变得困难。例如，识别人工智能系统目的和目标的早期决策可以改变其行为和能力，而部署环境的动态(如最终用户或受影响个人)可以塑造人工智能系统决策的影响。因此，人工智能生命周期某一维度内的最佳意图可能因与其他后续活动中的决策和条件的交互而受到削弱。

这种复杂性以及不同程度的可见性可能给风险管理实践带来不确定性。预判、评估并以其他方式应对潜在的负面风险来源，可以减轻这种不确定性并增强决策过程的完整性。

在执行**映射**功能时收集的信息有助于预防负面风险，并为模型管理等流程的决策提供依据，同时也有助于就人工智能解决方案的适当性或必要性做出初步决定。**映射**功能的成果是**度量**和**管理**功能的基础。如果缺乏上下文知识以及对已识别情境中风险的认识，风险管理将难以实施。**映射**功能旨在增强组织识别风险和更广泛影响因素的能力。

通过纳入多元化内部团队的视角，以及与开发或部署该人工智能系统团队之外的人员进行接触，可以增强本功能的实施效果。与外部合作者、最终用户、可能受影响的社区及其他方的接触方式，可能因特定人工智能系统的风险水平、内部团队的构成以及组织政策而有所不同。汇集如此广泛的视角有助于组织主动预防负面风险，并通过以下方式开发更可信的人工智能系统：

- 提高其理解上下文的能力；
- 检查其关于使用情境的假设；
- 使人们能够识别系统在其预期情境之内或之外何时无法正常运作；

- 识别其现有AI系统的积极和有益用途；
- 增进对 AI 和 ML 流程局限性的理解；
- 识别现实应用中可能导致负面影响的约束条件；
- 识别与人工智能系统预期用途相关的已知和可预见的负面影响；以及

- 预见人工智能系统超出预期用途使用时的风险。

完成**映射**功能后，框架使用者应已对人工智能系统的影响具备充分的上下文知识，从而就是否设计、开发或部署人工智能系统做出初步的继续/停止决定。如果决定继续推进，组织应利用**度量**和**管理**功能，以及**治理**功能中建立的政策和程序，来协助人工智能风险管理工作。框架使用者有责任随着情境、能力、风险、收益和潜在影响随时间演变，持续对人工智能系统应用**映射**功能。

与映射人工智能风险相关的实践在 NIST AI RMF Playbook 中有所描述。表 2 列出了**MAP**功能的类别和子类别。

表 2:**MAP**功能的类别与子类别。

类别	子类别
映射 1:上下文是已建立并被理解。	**MAP 1.1:**预期目的、潜在有益用途、背景--特定的法律、规范和期望，以及前瞻性的环境 人工智能系统将被部署的情境已被理解并记录在案。考虑因素包括：特定的用户群体或用户类型及其期望；系统使用对个人、社区、组织、社会和地球的潜在正面和负面影响；关于人工智能系统在开发或产品人工智能生命周期中的目的、用途和风险的假设及相关局限性；以及相关的TEVV和系统度量指标。**MAP 1.2:**用于确立情境的跨学科人工智能参与者、能力、技能和才干体现了人口多样性以及广泛的领域和用户体验专业知识，并且他们的参与已被记录。跨学科合作的机会被优先考虑。 **MAP 1.3:**组织对人工智能技术的使命和相关目标已被理解并记录在案。 **MAP 1.4:**商业价值或商业用途背景已被明确定义，或在评估现有人工智能系统的情况下已被重新评估。 **MAP 1.5:**组织风险容忍度已确定，并且被记录。 **映射 1.6:**系统需求(例如“系统应尊重用户的隐私”)由相关人工智能参与者引出并理解。设计决策考虑社会技术影响，以应对人工智能风险。
MAP 2: 分类AI系统是被执行。	**映射 2.1:**用于实施该任务的具体任务和方法AI系统将支持的任务已定义(例如，分类器，生成模型、推荐系统)。 **映射 2.2:**关于人工智能系统知识局限以及系统输出如何被人类利用和监督的信息被记录在案。文档提供充分信息，以协助相关人工智能参与者在做出决策和采取后续行动时参考。

续下页

表 2:**MAP**功能的类别与子类别。(续)

类别	子类别
	映射 2.3:科学诚信和TEVV(测试、评估、验证与确认)考量被识别并记录，包括与实验设计、数据收集与选择(例如可用性、代表性、适用性)、系统可信度以及构念相关的考量。 验证。
映射 3: AI能力，有针对性的使用、目标和预期收益与成本相比与适当的基准是被理解。	**映射 3.1:**预期人工智能系统功能和性能的潜在效益被审查并记录。 **映射 3.2:**由预期或已实现的人工智能错误或系统功能与可信度所导致的潜在成本(包括非货币成本)--与组织风险容忍度相关联--被审查并记录。 **映射 3.3:**基于系统能力、既定背景和人工智能系统分类，明确并记录目标应用范围。 **映射 3.4:**操作人员和从业者在人工智能系统性能和可信度方面的熟练程度--以及相关技术标准和认证--的流程被定义、评估并记录。 被记录。 **映射 3.5:**人类监督流程按照**治理**功能的组织政策被定义、评估并记录。
映射 4:风险与收益已映射对于所有组件人工智能系统包括第三方软件和数据。	**映射 4.1:**映射人工智能技术及其组件法律风险的方法--包括第三方数据或软件的使用--已到位、被遵循并记录，侵犯第三方知识产权或其他权利的风险亦然。**映射 4.2:**人工智能系统组件(包括第三方人工智能技术)的内部风险控制被识别并记录。 被记录。
映射 5:对以下方面的影响个人、群体、社区，组织和社会是被描述。	**映射 5.1:**基于预期用途、人工智能系统在类似情境中的过往使用、公开事件报告、开发或部署人工智能系统的团队外部人员的反馈或其他数据，每个已识别影响(包括潜在有益和有害影响)的可能性和严重程度被识别并记录。 被记录。

续下页

表 2:**MAP**功能的类别与子类别。(续)

类别	子类别
	映射 5.2:支持与相关人工智能参与者定期互动并整合关于正面、负面和意外影响的反馈的实践和人员已到位，并且 被记录。

5.3 Measure

度量功能采用定量、定性或混合方法的工具、技术和方法论来分析、评估、基准测试和监测人工智能风险及相关影响。它使用在**映射**功能中识别的与人工智能风险相关的知识，并为**管理**功能提供信息。人工智能系统应在部署前进行测试，并在运行期间定期测试。人工智能风险度量包括记录系统功能性和可信度的各个方面。

度量人工智能风险包括跟踪可信特征、社会影响和人机配置的指标。在**度量**功能中开发或采用的流程应包括严格的软件测试和性能评估方法，并附有相关的不确定性度量、与性能基准的比较，以及结果的正式报告和记录。独立审查流程可以提高测试的有效性，并可以减少内部偏见和潜在的利益冲突。

建立。

当可信特征之间出现权衡时，度量提供可追溯的基础以支持管理决策。选项可能包括重新校准、影响缓解或将系统从设计、开发、生产或使用中移除，以及一系列补偿性、检测性、威慑性、指导性和恢复性控制措施。

完成**度量**功能后，客观、可重复或可扩展的测试、评估、验证与确认(TEVV)流程(包括指标、方法和方法论)已到位、得到遵循并记录在案。指标和测量方法应遵循科学、法律和伦理规范，并在公开透明的过程中进行。可能需要开发新型测量方法，包括定性和定量方法。应考虑每种测量类型为人工智能风险评估提供独特且有价值信息的程度。框架用户将增强其全面评估系统可信度、识别和跟踪现有及新兴风险以及验证指标有效性的能力。测量结果将在**管理**功能中用于协助风险监测和响应工作。框架用户有责任随着知识、方法论、风险和影响随时间演变，继续对人工智能系统应用**度量**功能。

与度量人工智能风险相关的实践在NIST AI RMF Playbook中描述。表3列出了**度量**功能的类别和子类别。

表3:**MEASURE**功能的类别与子类别。

类别	子类别
MEASURE 1: 适当的方法和指标被识别并应用。	**度量 1.1:**在**映射**功能中列举的用于度量人工智能风险的方法和指标被选择用于实施，从最重要的人工智能风险开始。将不会--或无法--度量的风险或可信特征被适当记录。 **度量 1.2:**定期评估和更新人工智能指标的适当性和现有控制措施的有效性，包括错误报告和对受影响社区的潜在影响。**度量 1.3:**未担任该系统一线开发人员的内部专家和/或独立评估人员参与定期评估和更新。根据组织风险容忍度，酌情咨询领域专家、用户、开发或部署人工智能系统团队之外的人工智能参与者以及受影响社区以支持评估。**度量 2.1:**记录测试集、指标以及TEVV期间所用工具的详细信息。
度量 2: AI 系统是评估可信特征。	**度量 2.2:**涉及人类受试者的评估满足适用要求(包括人类受试者保护)，并且代表相关人群。 **度量 2.3:**人工智能系统性能或保证标准被定性或定量度量，并在与部署环境相似的条件下得到证明。度量是 被记录。 **度量 2.4:**人工智能系统及其组件的功能和行为--如**映射**功能中所识别--在生产中进行监测。 **度量 2.5:**待部署的人工智能系统被证明是有效和可靠的。超出技术开发条件的泛化局限性被记录。

续下页

表3:**度量**功能的类别和子类别。(续)

类别	子类别
	度量 2.6:人工智能系统定期评估安全风险--如**映射**功能中所识别。待部署的人工智能系统被证明是安全的，其残余负面风险不超过风险容忍度，并且能够安全失败，特别是在被迫超出其知识限制运行的情况下。安全指标反映系统可靠性和稳健性、实时监测以及人工智能系统故障的响应时间。
	度量 2.7:人工智能系统安全性和韧性--如**映射**功能中所识别--被评估和记录。
	度量 2.8:与透明度和问责制相关的风险--如**映射**功能中所识别--被审查并 被记录。
	度量 2.9:人工智能模型被解释、验证和记录，人工智能系统输出在其上下文中被解释--如**映射**功能中所识别--以支持负责任的使用和 治理。
	度量 2.10:人工智能系统的隐私风险--如**映射**功能中所识别--被审查和记录。
	度量 2.11:公平性和偏见--如**映射**功能中所识别--被评估，结果被记录。
	度量 2.12:人工智能模型训练和管理活动的环境影响和可持续性--如**映射**功能中所识别--被评估和记录。
	度量 2.13:所采用的TEVV(测试、评估、验证与确认)指标和流程在度量功能中的有效性被评估并 被记录。
MEASURE 3: 机制，用于 跟踪已识别的 人工智能风险随时间演变 已就位。	**度量 3.1:**方法、人员和文档已 已建立机制，定期识别和跟踪现有的、未预料到的、 以及基于预期用途等因素的新兴人工智能风险，以及 部署环境中的实际性能。
	度量 3.2:对于使用当前可用度量技术难以评估人工智能风险或指标尚不可用的环境，考虑风险跟踪方法 可用。

续下页

表3:**度量**功能的类别和子类别。(续)

类别	子类别
MEASURE 4: 关于.....的反馈 效力 测量是 收集并 评估。	**度量 3.3:**最终用户和受影响社区报告问题和申诉系统结果的反馈流程被建立并整合到人工智能系统评估指标中。**度量 4.1:**用于识别人工智能风险的度量方法与部署环境相关联，并通过与领域专家和其他最终用户的咨询获得信息。方法被记录。 **度量 4.2:**关于人工智能系统在部署环境中以及跨人工智能生命周期的可信度的度量结果，通过领域专家和相关人工智能参与者的输入获得信息，以验证系统是否按预期一致运行。结果被记录。 **度量 4.3:**基于与相关人工智能参与者(包括受影响社区)的咨询以及关于情境相关风险和可信特征的现场数据，可测量的性能改进或下降被识别和记录。

5.4 Manage

管理功能需要定期并按照**治理**功能的定义，为已映射和已度量的风险分配风险资源。风险处理包括应对、恢复和沟通事件或事故的计划。

从专家咨询中获得的背景信息以及来自相关人工智能参与者的输入--在**治理**中确立并在**映射**中执行--在此功能中用于降低系统故障和负面影响的可能。在**治理**中确立并在**映射**和**度量**中使用的系统化记录实践，可加强人工智能风险管理工作并提高透明度和问责制。评估新兴风险的流程已到位，并配有持续改进机制。

完成**管理**功能后，将制定风险优先级排序以及定期监测和改进的计划。框架用户将增强管理已部署人工智能系统风险的能力，并根据评估和优先级排序的风险分配风险管理资源。框架用户有责任随着方法、情境、风险以及相关人工智能参与者的需求或期望随时间演变，继续对已部署的人工智能系统应用**管理**功能。

与管理人工智能风险相关的实践在NIST AI RMF Playbook中描述。表4列出了**管理**功能的类别和子类别。

表 4:**管理**功能的类别与子类别。

类别	子类别
管理 1: AI 基于风险评估和其他分析性来自**映射**和**度量**功能包括优先排序，已回应，并且管理。	**管理 1.1:**确定人工智能系统是否实现其预期目的和既定目标，以及其开发或部署是否应继续进行。 **管理 1.2:**对已记录的人工智能风险的处理基于影响、可能性和可用资源或方法进行优先排序。**管理 1.3:**对**映射**功能所识别的被视为高优先级的人工智能风险的响应被制定、规划和记录。风险响应选项可包括缓解、转移、避免或接受。 **管理 1.4:**对人工智能系统下游获取者和最终用户的负面残余风险(定义为所有未缓解风险的总和)被记录。
MANAGE 2: 制定策略以最大化人工智能益处和尽量减少负面影响影响是经过规划的，准备，实施，已记录，且基于输入信息来自相关的AI行为者。	**管理 2.1:**管理人工智能风险所需的资源--连同可行的非人工智能替代系统、方法或途径--被纳入考虑，以减少潜在影响的严重程度或可能性。 **管理 2.2:**已建立并应用相关机制，以维持已部署人工智能系统的价值。 **管理 2.3:**遵循相关程序，以在识别出先前未知风险时进行响应和恢复。 **管理 2.4:**已建立并应用相关机制，且已分配并理解相关责任，以取代、脱离或停用表现出与预期用途不一致的性能或结果的人工智能系统。
管理 3: AI 风险和收益来自第三方实体是管理。	**管理 3.1:**定期监测来自第三方资源的人工智能风险和收益，并应用风险控制措施，且被记录。 **管理 3.2:**用于开发的预训练模型作为人工智能系统定期监测的一部分进行监测，且维护。

续下页

表 4:**管理**功能的类别和子类别。(续)

类别	子类别
管理 4:风险治疗，包括响应以及恢复，并且沟通针对.....的计划已识别并已度量的 AI 风险已记录并定期监测。	**管理 4.1:**实施部署后人工智能系统监测计划，包括用于捕获和评估来自用户及其他相关人工智能参与者的输入、申诉与推翻、退役、事件响应、恢复和变更管理的机制。 **管理 4.2:**将持续改进的可衡量活动整合到人工智能系统更新中，并包括与利益相关方(包括相关人工智能参与者)的定期互动。 **管理 4.3:**将事件和错误传达给相关人工智能参与者，包括受影响的社区。遵循并记录用于跟踪、响应和从事件和错误中恢复的流程。

6. AI RMF 概况

AI RMF 用例配置文件是AI RMF功能、类别和子类别的实现，针对特定设置或应用，基于框架用户的需求、风险容忍度和资源：例如，AI RMF 招聘配置文件或AI RMF 公平住房配置文件。配置文件可以说明并提供见解，关于如何在人工智能生命周期的各个阶段或特定部门、技术或最终用途应用中管理风险。AI RMF配置文件帮助组织决定如何最佳管理AI风险，与其目标良好对齐，考虑法律/监管要求和最佳实践，并反映风险管理优先级。

AI RMF 时间轮廓描述特定部门、行业、组织或应用背景下特定人工智能风险管理活动的当前状态或期望的目标状态。AI RMF 当前轮廓指示人工智能当前如何被管理以及就当前结果而言的相关风险。目标轮廓指示实现期望或目标人工智能风险管理目标所需的结果。

比较当前配置文件和目标配置文件可能揭示需要解决的差距，以满足AI风险管理目标。可以制定行动计划来解决这些差距，以实现给定类别或子类别中的成果。差距缓解的优先级由用户的需求和风险管理过程驱动。这种基于风险的方法还使框架用户能够将他们的方法与其他方法进行比较，并评估实现AI风险管理目标所需的资源(例如，人员配备、资金)，以成本有效、优先化的方式。

AI RMF 跨部门轮廓涵盖可跨用例或部门使用的模型或应用的风险。跨部门轮廓还可涵盖如何治理、映射、度量和管理跨部门常见活动或业务流程的风险，例如使用大型语言模型、基于云的服务或采购。

本框架不规定轮廓模板，允许实施中的灵活性。

附录 A：图 2 和图 3 中 AI 行为者任务的描述

AI设计任务在图2中的人工智能生命周期的应用上下文和数据与输入阶段执行。AI设计参与者创建AI系统的概念和目标，并负责AI系统的规划、设计和数据收集与处理任务，以使AI系统合法且适合目的。任务包括阐述和记录系统的概念和目标、基本假设、上下文和要求；收集和清理数据；以及记录数据集的元数据和特征。此类别中的AI参与者包括数据科学家、领域专家、社会文化分析师、多样性、公平、包容和可及性领域的专家、受影响社区的成员、人为因素专家(例如，UX/UI设计)、治理专家、数据工程师、数据提供者、系统资助者、产品经理、第三方实体、评估者以及法律和隐私治理。

AI开发任务在图2中生命周期的AI模型阶段执行。AI开发参与者提供AI系统的初始基础设施，并负责模型构建和解释任务，这些任务涉及模型或算法的创建、选择、校准、训练和/或测试。此类别中的AI参与者包括机器学习专家、数据科学家、开发者、第三方实体、法律和隐私治理专家，以及与部署设置相关的社会文化和上下文因素的专家。

AI部署任务在图2中生命周期的任务和输出阶段执行。AI部署参与者负责与如何使用AI系统相关的上下文决策，以确保系统部署到生产环境。相关任务包括试点系统、检查与遗留系统的兼容性、确保监管合规、管理组织变革以及评估用户体验。此类别中的AI参与者包括系统集成商、软件开发者、最终用户、操作员和从业者、评估者，以及具有人为因素、社会文化分析和治理专业知识的领域专家。

运行与监控任务在图2所示生命周期中的应用场景/运行与监控阶段执行。这些任务由负责人工智能系统运行并与其他人员合作定期评估系统输出和影响的人工智能参与者执行。此类别中的人工智能参与者包括系统操作员、领域专家、人工智能设计者、解释或采纳人工智能系统输出的用户、产品开发者、评估者和审计者、合规专家、组织管理层以及研究界成员。

测试、评估、验证与确认(TEVV)任务在整个人工智能生命周期中执行。这些任务由人工智能参与者执行，他们检查人工智能系统或其组件，或检测并修复问题。理想情况下，执行验证

和确认任务的人工智能参与者与执行测试和评估行动的人员不同。任务可尽早纳入设计阶段，在该阶段根据设计要求规划测试。

- 针对设计、规划和数据的 TEVV 任务可集中于系统设计、数据收集和测量相对于预期部署或应用背景的假设的内部和外部验证。
- 针对开发(即模型构建)的 TEVV 任务包括模型验证和评估。
- 针对部署的 TEVV 任务包括生产中的系统验证和集成，以及针对系统和流程集成、用户体验以及遵守现有法律、监管和伦理规范的测试和重新校准。
- 针对运营的 TEVV 任务涉及持续监测以进行定期更新、测试和主题专家(SME)对模型的重新校准、跟踪报告的事件或错误及其管理、检测涌现属性及相关影响，以及补救和响应流程。

人为因素任务和活动贯穿人工智能生命周期的各个维度。它们包括以人为本的设计实践和方法、促进最终用户和其他利益相关方及相关人工智能参与者的积极参与、在系统设计中纳入特定情境的规范和价值观、评估和调整最终用户体验，以及在人工智能生命周期的所有阶段广泛整合人类和人类动态。人为因素专业人员提供多学科技能和视角，以理解使用情境、促进跨学科和人口多样性、参与协商过程、设计和评估用户体验、执行以人为本的评估和测试，并为影响评估提供信息。

领域专家任务涉及来自多学科从业者或学者的输入，他们提供关于人工智能系统所使用行业部门、经济部门、背景或应用领域的知识和专业知识。作为领域专家的人工智能参与者可为人工智能系统设计和开发提供重要指导，并解释输出以支持 TEVV 和人工智能影响评估团队执行的工作。

人工智能影响评估任务包括评估和评价人工智能系统问责制、打击有害偏见、审查人工智能系统影响、产品安全、责任和安全等方面的要求。影响评估者和评价者等人工智能参与者提供技术、人为因素、社会文化和法律专业知识。

采购任务由具有财务、法律或政策管理权限的人工智能参与者执行，以从第三方开发者、供应商或承包商处获取人工智能模型、产品或服务。

治理与监督任务由具有管理、受托和法律权限并对设计、开发和/或部署人工智能系统的组织负有责任的人工智能参与者承担。

负责人工智能治理的关键人工智能参与者包括组织管理层、高级领导和董事会。这些参与者是关注整个组织影响和可持续性的各方。

其他人工智能行为者

第三方实体包括为另一组织或该组织的客户或顾客提供数据、算法、模型和/或系统及相关服务的提供商、开发者、供应商和评估者。第三方实体全部或部分负责人工智能设计和开发任务。根据定义，它们外在于获取其技术或服务的组织的设计、开发或部署团队。从第三方实体获取的技术可能复杂或不透明，风险容忍度可能与部署或运营组织不一致。

人工智能系统的**最终用户**是为特定目的使用系统的个人或群体。这些个人或群体在特定背景下与人工智能系统交互。最终用户的能力范围可从人工智能专家到首次使用技术的最终用户。

受影响的个人/社区涵盖直接或间接受人工智能系统或基于人工智能系统输出的决策影响的所有个人、群体、社区或组织。这些个人不一定与已部署的系统或应用交互。

其他人工智能参与者可为规定和管理人工智能风险提供正式或准正式规范或指导。他们可包括**行业协会、标准制定组织、倡导团体、研究人员、环境团体和民间社会组织**。

公众最有可能直接体验人工智能技术的积极和消极影响。他们可为人工智能参与者采取的行动提供动机。该群体可包括与人工智能系统开发或部署背景相关的个人、社区和消费者。

附录 B:AI 风险与传统软件风险有何不同

与传统软件一样，基于人工智能的技术的风险可能超出企业范围、跨越多个组织并导致社会影响。人工智能系统还带来一系列当前风险框架和方法未能全面解决的风险。一些带来风险的人工智能系统特征也可能是有益的。例如，与其他模型和方法相比，预训练模型和迁移学习可以推进研究并提高准确性和韧性。在**映射**功能中识别情境因素将有助于人工智能参与者确定风险水平和潜在的管理工作。

与传统软件相比，新增或增加的人工智能特定风险包括以下内容：

- 用于构建人工智能系统的数据可能不是人工智能系统背景或预期用途的真实或适当表示，且真实基准可能不存在或不可用。此外，有害偏见和其他数据质量问题可能影响人工智能系统的可信度，从而可能导致负面影响。

- 人工智能系统对训练任务数据的依赖，以及此类数据通常伴随的数据量和复杂性的增加。
- 训练过程中有意或无意的变化可能从根本上改变人工智能系统的性能。

- 用于训练人工智能系统的数据集可能脱离其原始和预期语境，或相对于部署语境变得陈旧或过时。
- 人工智能系统的规模和复杂性(许多系统包含数十亿甚至数万亿个决策点)嵌入在更为传统的软件应用中。
- 使用预训练模型可以推进研究并提高性能，但也可能增加统计不确定性水平，并导致偏差管理、科学有效性和可复现性方面的问题。

- 预测大规模预训练模型涌现特性的失效模式难度更高。

- 由于人工智能系统数据聚合能力增强而带来的隐私风险。
- 由于数据、模型或概念漂移，人工智能系统可能需要更频繁的维护和触发纠正性维护。

- 不透明度增加以及对可复现性的担忧。
- 软件测试标准不完善，且除最简单的情况外，无法按照传统工程化软件所期望的标准记录基于人工智能的实践。

- 难以执行定期的基于人工智能的软件测试，或确定测试内容，因为人工智能系统不受与传统代码开发相同的控制。

- 开发人工智能系统的计算成本及其对环境和地球的影响。

- 无法在统计度量之外预测或检测基于人工智能的系统的副作用。

隐私和网络安全风险管理考量和方法适用于人工智能系统的设计、开发、部署、评估和使用。隐私和网络安全风险也被视为更广泛的企业风险管理考量的一部分，其中可能纳入人工智能风险。作为解决“安全与韧性”和“隐私增强”等人工智能可信特征工作的一部分，组织可考虑利用现有标准和指南，这些标准和指南为组织提供广泛指导以降低安全和隐私风险，例如包括但不限于NIST网络安全框架、NIST隐私框架、NIST风险管理框架和安全软件开发框架。这些框架与AI RMF有一些共同特征。与大多数风险管理方法一样，它们基于结果而非规定性，并且通常围绕一组核心功能、类别和子类别构建。尽管这些框架因所涉领域不同而存在显著差异--并且因为人工智能风险管理需要解决许多其他类型的风险--上述框架可为AI RMF的**映射**、**度量**和管理功能中的安全和隐私考量提供参考。

与此同时，在本 AI RMF 发布之前可获得的指南并未全面涵盖许多人工智能系统风险。例如，现有框架和指南无法：

- 充分管理人工智能系统中的有害偏见问题；
- 应对与生成式人工智能相关的挑战性风险；
- 全面解决与规避、模型提取、成员推断、可用性或其他机器学习攻击相关的安全问题；

- 考虑人工智能系统的复杂攻击面或由人工智能系统促成的其他安全滥用；以及

- 考虑与第三方人工智能技术、迁移学习和标签外使用相关的风险，即人工智能系统可能在组织安全控制之外进行决策训练，或在一个领域训练后针对另一领域进行“微调”。

人工智能和传统软件技术与系统都处于快速创新之中。应监测技术进展并加以部署，以利用这些发展，朝着既可信又负责任的人工智能未来努力。

附录 C:AI 风险管理与人机交互

为在运营环境中使用而设计、开发或部署人工智能系统的组织，可以通过理解人机交互的当前局限性来增强其人工智能风险管理。AI RMF 提供了机会，以明确界定和区分在使用、交互或管理人工智能系统时各种人类角色和职责。

人工智能系统所依赖的许多数据驱动方法试图将个人和社会的观察与决策实践转换或表示为可测量的量。用数学模型表示复杂的人类现象可能以去除必要语境为代价。这种语境的丧失反过来可能使理解对人工智能风险管理工作至关重要的个人和社会影响变得困难。

值得进一步考虑和研究的问题包括：

1. **在决策和监督人工智能系统方面，人类的角色和职责需要被清晰地定义和区分。**人机配置的范围可以从完全自主到完全手动。人工智能系统可以自主做出决策，将决策权交给人类专家，或者被人类决策者用作额外的意见。某些人工智能系统可能不需要人类监督，例如用于改进视频压缩的模型。其他系统则可能特别需要人类监督。

2. **人工智能系统的设计、开发、部署、评估和使用中所涉及的决策反映了系统性和人类认知偏差。**人工智能参与者将其个体和群体的认知偏差带入过程中。偏差可能源于最终用户的决策任务，并通过人类在设计 and 建模任务中的假设、期望和决策在整个人工智能生命周期中被引入。这些偏差不一定总是有害的，但可能因人工智能系统的不透明性以及由此产生的透明度缺乏而加剧。组织层面的系统性偏差可能影响团队的结构方式以及谁在整个人工智能生命周期中控制决策过程。这些偏差还可能影响最终用户、决策者和政策制定者的下游决策，并可能导致负面影响。

3. **人机交互结果各不相同。**在某些条件下--例如在基于感知的判断任务中--人机交互中的人工智能部分可能放大人类偏差，导致比单独的人工智能或人类更具偏差的决策。然而，当在组织人机团队时审慎考虑这些差异，它们可以产生互补性并改善整体性能。

4. **向人类呈现人工智能系统信息是复杂的。**人类以不同方式感知人工智能系统输出和解释并从中得出意义，反映了不同的个人偏好、特质和技能。

治理功能为组织提供了机会，以澄清和界定人机团队配置中的人类以及监督人工智能系统性能的人员的角色和职责。**治理**功能还为组织创建了机制，使其决策过程更加明确，以帮助对抗系统性偏差。

映射功能提出了定义和记录操作员和从业者在人工智能系统性能和可信度概念方面的熟练程度流程的机会，并定义相关技术标准和认证。实施**映射**功能的类别和子类别可能有助于组织提高其内部能力，以分析上下文、识别程序和系统限制、探索和检查基于人工智能的系统在现实世界中的影响，以及评估整个人工智能生命周期中的决策过程。

治理和**映射**功能描述了跨学科性和人口多样性团队的重要性，以及利用可能受影响的个人和社区的反馈。AI RMF 中提到的执行人为因素任务和活动的人工智能参与者可以通过将设计和开发实践锚定于用户意图、更广泛的人工智能社区代表以及社会价值观来协助技术团队。这些参与者进一步帮助将特定情境的规范和价值观纳入系统设计，并与人工智能系统一起评估最终用户体验。

针对人机配置的人工智能风险管理方法将通过持续的研究和评估得到增强。例如，人类被赋权和激励去质疑人工智能系统输出的程度需要进一步研究。关于人类在已部署系统中否决人工智能系统输出的频率和理由的数据可能有助于收集和分析。

附录 D:AI RMF 的属性

NIST 在框架工作最初开始时描述了 AI RMF 的几个关键属性。这些属性一直保持不变，并被用于指导 AI RMF 的开发。此处提供作为参考。

AI RMF 致力于：

1. 以风险为基础、资源高效、支持创新且自愿。
2. 以共识为驱动，并通过开放、透明的过程制定和定期更新。所有利益相关方都应有机会为 AI RMF 的制定做出贡献。

3. 使用清晰平实的语言，使广泛受众能够理解，包括高级管理人员、政府官员、非政府组织领导层以及非 AI 专业人士--同时仍具备足够的技术深度以对从业者有用。AI RMF 应允许在组织内部、组织之间、与客户以及与广大公众之间沟通 AI 风险。

4. 提供管理 AI 风险的共同语言和理解。AI RMF 应提供 AI 风险的分类法、术语、定义、度量指标和特征描述。

5. 易于使用，并与风险管理的其他方面良好契合。框架的使用应直观，并易于作为组织更广泛风险管理战略和流程的一部分进行调整。它应与其他管理 AI 风险的方法保持一致或对齐。
6. 对广泛的视角、行业和技术领域有用。AI RMF 应普遍适用于任何 AI 技术和特定情境的用例。

7. 以结果为导向且非规定性。框架应提供结果和方法的目录，而不是规定一刀切的要求。

8. 利用并促进对现有管理 AI 风险的标准、指南、最佳实践、方法和工具的更大认识--同时说明需要额外、改进的资源。

9. 与法律和法规无关。框架应支持组织在适用的国内和国际法律或监管制度下运营的能力。

10. 作为一份活文件。AI RMF 应随着技术、对 AI 可信度的理解和方法的改变以及 AI 使用的变化，以及利益相关方从实施 AI 风险管理(总体上)和本框架(特别是)中学习而随时更新。

本出版物可从以下网址免费获取：
https://doi.org/10.6028/NIST.AI.100-1



这份文件是怎么做出来的

本译本是《人工智能风险管理框架（AI RMF 1.0）》的中文全译本，共 48 页。表格、图注、页眉页脚、编号层级都按原文位置一一还原，未做任何删节。

你手上如果也有英文文件要处理

- 合同、标书、标准、专利、报告：翻译后保留原版式，0.15 元/页
- PDF 转 Word：100 页内免费，超出 0.1 元/页
- 扫描件、图片版 PDF：先 OCR 再转 Word，免费
- 不用注册、不用登录，传上去直接下载

打开下面这个地址，把你自己的文件传上去试试：

https://www.ivyivy.com.cn/nist-ai-rmf-zh

支持私有化部署（需按贵方模型与合规要求评估）

扫码关注服务号
咨询定制交付

本译本由机器翻译生成，仅供阅读参考；专业技术判断与合规引用，请以英文原文为准。

本页为译制说明，不属于原文件内容。