NIST INFORMATION TECHNOLOGY LABORATORY

SP 800-53 安全人工智能系统控制叠加概念文件

采用人工智能（AI）技术的进步和潜在用例既带来了新的机遇，也带来了新的网络安全风险。虽然现代人工智能系统主要是软件，但它们引入了与传统软件不同的安全挑战和风险。人工智能系统的安全性与其运行和操作的基础IT基础设施的安全性密切相关。

美国国家标准与技术研究院提供了一系列关于设计和实施安全、可信赖人工智能的指南，包括 人工智能风险管理框架 (RMF)， 管理高级人工智能的滥用风险指南 (草稿)，以及 人工智能攻击的分类和缓解措施 . 在此期间收到的反馈
网络安全与人工智能简介工作坊 2025年4月以及与人工智能和网络安全利益相关者的持续互动表明，各方强烈要求NIST提供更多以实施为重点的指导方针，这些指导方针应基于现有的资源和框架，以提高人工智能系统的网络安全。

为了补充正在进行的创建工作的努力 人工智能安全框架配置文件 [1] 并支持采用人工智能风险管理框架（AI RMF），美国国家标准与技术研究院（NIST）提议开发一系列 ***Securing AI Systems Control Overlays (COSAIS)*** 使用 美国国家标准与技术研究院特别出版物（sp）800-53控制 .

为什么使用 SP 800-53 安全控制措施和叠加层？

控件 [2] 需要对人工智能系统和组件进行风险管理的要求，将与任何类型软件所需的要求相似。许多组织已经熟悉 [SP 800-53] 控制措施，并建立了机构流程来规划评估其实施情况。这些控制措施既提供了满足独特需求的灵活性，又提供了一种特定程度，允许一致的技术实施。

控制覆盖层使组织和利益相关社群能够针对特定技术、系统、任务空间和运行环境定制控制项（或控制基线）。控制项可从 [SP 800-53] 控制目录中进行选择，可进行修改以应对独特风险或应用场景，并可进行补充以提供针对实施者的应用特定指导。分配和选择操作的参数值也可以设置。

使用 sp 800-53 控制措施为识别网络安全结果提供了共同的技术基础，而开发叠加层可以实现在人工智能系统中对最关键的控 制措施进行定制和优先考虑。

[1] 随着人工智能网络安全框架概况工作进展，NIST 将与该 利益共同体 为确定控制叠加层如何影响配置文件的网络安全结果，反之亦然，以确保配置文件和叠加层都为组织提供互补的指导方针。

[2] 根据[OMB循环A-130]，“信息系统的安全保护措施或针对组织为保护系统及其信息的机密性、完整性和可用性，或符合相关隐私要求和管理隐私风险所规定的内容。”

表1。 NIST AI 与网络安全工作之间的关系

	控制覆盖层 保障人工智能系统	**网络安全框架 人工智能简介**	**人工智能风险管理框架**	**滥用风险 双重用途基金会 模型**
目的	提供一系列 指南 实现 SP 800- 53个安全控制措施 用于不同的AI用途	提供一个战略 组织路线图 适应并响应 新增或修改 网络安全风险 由于进步 在 AI	提供一个方法 建立和 保持 人工智能的可信度 被使用的系统 用于网络安全	提供指南以 识别，减轻，和 管理犯罪风险 人工智能模型的误用
范围	**安全控制** 到 管理独特风险 为用户 人工智能开发者 特定系统 场景	优先级和影响 为保障人工智能系统 和他们的组件， 使用人工智能进行网络安全防御 和防御人工智能- 已启用网络攻击	风险识别 和机会 使用AI和建议 措施 缓解/实现那些 风险与机遇	组织和发展 具体建议 to manage risks from 化学，生物 放射性的、核的和 爆炸威胁（核生化） 网络攻击和其他 与国家安全相关的 人工智能能力
观众	网络安全 修行者，人工智能 用户，AI开发者 (取决于使用 case)	网络安全高级 领导力，首席信息安全官，风险 官员，网络安全 实践者，AI提供者 和用户	AI 演员在人工智能 生命周期，包括 网络安全 实践者	人工智能开发者，尤其是 边界模型
假设	组织/企业 提升网络安全 政策，程序， 以及技术控制 就地管理 风险	组织/企业 政策，程序和 已部署技术控制 管理网络安全 风险	企业风险 管理 功能已经 现场（例如，网络， 隐私)	企业风险 管理能力 已就位

[3] 包含 AI RMF 配置文件

用于保障人工智能系统的控制叠加层将是一系列以实施为重点的指南，涵盖不同类型的人工智能系统、特定的组件（例如，训练和测试数据、模型权重、配置设置）以及不同的用例。这些叠加层将专注于保护每个用例的信息机密性、完整性和可用性。 [4]

除了 sp 800-53 控制措施外，nist 还有三项基础的人工智能和网络安全出版物：

- SP 800-218A *生成式人工智能和双用途基础模型的软件开发实践：一个SSDF社区概况* ，作为识别特定信息资产和安全实践，并将其转化为人工智能开发人员控制措施的起点。 NIST 可信和负责任的AI（AI）100-2e2025 *对抗机器学习：攻击和缓解方法的分类和术语* ，提供了一套人工智能用例、攻击和缓解措施。• 起草 NIST AI 800-1，管理双用途基础模型的滥用风险 提供组织和特定模型相关的实践，以管理 AI 系统被犯罪利用的风险。

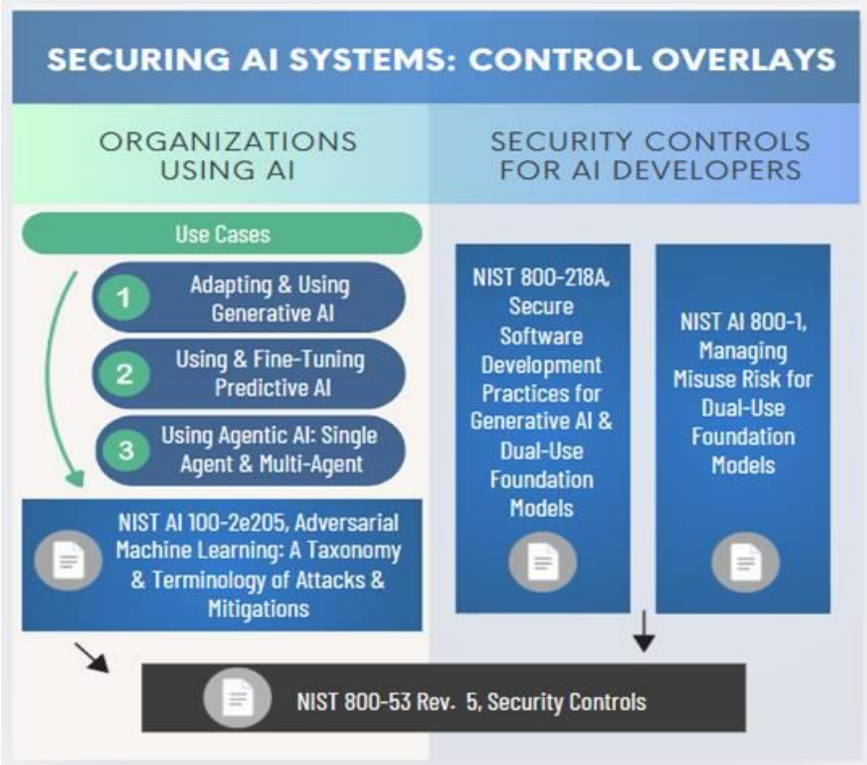

图1. 控制覆盖层与NIST出版物之间的关系

[4] 人工智能风险管理框架（AI RMF）解决了超出 sp 800-53 安全和隐私控制范围的额外风险和可信度考虑因素（例如，有效性、可靠性、安全性、问责制、透明度、可解释性、可解释性、隐私增强、公平性）。一系列控制叠加可与人工智能风险管理框架（AI RMF）结合使用，以实现人工智能风险管理的整体方法。

如图1所示，用于保障人工智能系统的控制措施将利用AI 100-2e2025、草案AI 800-1、SP 800-218A和SP 800-53。AI 100-2e2025中的攻击和缓解措施将指导针对为使用人工智能的组织设计的四种用例选择和定制SP 800-53控制措施。SP 800-218A和草案AI 800-1将指导针对为人工智能开发者设计的用例选择和定制SP 800-53控制措施。

用于安全控制叠加的 AI 应用案例建议

预期输出是一个组织可单独使用或组合使用的覆盖层库，用于更好地管理人工智能使用和开发中的网络安全风险。这些覆盖层将设计为结合组织的网络安全风险管理计划和现有控制措施，解决不同类型人工智能使用所面临的特定风险。这些覆盖层不会是一个保障企业的完整控制措施集合，并将假定某些控制措施已到位（例如，全组织的政策、程序以及数据集和服务、账户管理、身份识别和认证、配置管理、事件响应的访问控制实施）。

NIST 为使用人工智能的开发者和组织提出了一个初步的五个用例集合。每个用例代表一个常见的基本场景类别，并针对特定的网络安全风险。每个用例将根据需要解决模型、模型工件、部署基础设施内相关信息资产的保护，以及输出安全的问题。

用例1：适应和使用生成式人工智能——助手/大型语言模型（LLM）	
观众	组织有兴趣使用人工智能创建新内容（例如，文本，images）。
目的	生成式AI基于...创建新内容（例如，文本、图像、音频、视频） 从大规模数据集中学习并识别模式来获取用户提示 数据集。
用例 描述	本用例将涵盖用于内部业务的生成式AI示例 内部用户进行的增强（例如，创建摘要、分析数据）。 A. 本地托管的LLM及其支撑基础设施：检索- 基于专有本地数据源的增强生成（RAG），混合 本地和网页来源，且仅限外部网页来源。 B. 第三方托管的 LLM 及其支持基础设施：具有专有技术的 RAG 本地数据源，混合本地和网页数据源，以及外部网页 仅限来源。

用例 2：使用和微调预测人工智能	
观众	组织使用预测人工智能系统分析历史数据以提供信息 决策（例如，用于商业和服务增强）。
目的	预测性人工智能使用统计分析与机器学习来分析 历史数据并预测未来结果、趋势或行为。示例 可能包括推荐服务、分类服务以及商务

	通过自动化决策提高工作流效率（例如） 招聘中的简历审核、信贷审批)
用例 描述	这些用例将在三个阶段解决网络安全风险 预测式人工智能生命周期：(i)模型训练，(ii)模型部署，以及(iii)模型 维护。每个场景将针对不同的业务工作流 一个专注于使用人工智能系统所带来的独特网络安全风险（例如，在-）的示例 场或第三方托管的AI模型，使用专有数据或公开 已获得的数据)。 A. 使用第三方预测人工智能自动化业务流程 专有输入数据的主机模型：确定一个合法需求 授权用户进行物理访问并更新模型以使其精细化 观察和误报/误报后的改进数据 汇率。 B. 使用第三方预测人工智能自动化业务流程 提供面向公众的输入数据的托管模型：识别“恶意行为者” 在观察后，使用精细数据更新模型以进行改进 假阳性/假阴性率 C. 使用本地化部署的预测式人工智能自动化业务流程 仅使用专有数据的模型：评估贷款信用价值 申请者并更新模型，使用经过细化的数据进行改进后 成功或失败的贷款分配。 D. 使用本地化部署的预测式人工智能自动化业务流程 包含面向公众的输入数据的模型：用于招聘目的（例如） 根据组织识别面试/招聘候选人的简历 已定义标准)并在之后用精炼数据更新以改进 成功或失败的面试/招聘。

用例 3：使用 AI 代理系统（AI 代理）– 单个代理	
观众	对使用人工智能代理系统自动化业务任务感兴趣的机构 以及工作流。
目的	人工智能代理系统具有自主决策的能力 采取行动以在有限的人工监督下运行以实现复杂 目标。人工智能代理系统的特点包括理解能力 背景，原因，计划，适应，并执行任务。
用例 描述	本用例将涵盖人工智能代理系统使用的示例。 A. 企业副驾 - 连接到用户的个人企业 环境，包括电子邮件、文件、日历或内部企业 系统（例如，客户关系管理，it请求）。此外，企业副驾驶可以协助 具有常见任务的用户，例如创建日历事件 简化工作流，并提供上下文洞察。使用 MCP 到 访问数据源。这种架构的一个示例可在 OWASP代理式人工智能威胁与缓解. B. 编码助手 — 理解企业代码库并自动化 通过自然语言命令开发软件。使用 MCP用于访问数据源。此场景将包括：

	a. 连接到源代码库以启用编辑 文件； b. 与代码仓库交互以创建和提交拉取 请求（PRs），解决合并冲突，以及类似操作； c. 开发、执行和修复单元和集成测试; d. 浏览专有和网页资源； e. 协助软件部署。

用例 4：使用人工智能代理系统（人工智能代理）——多代理	
这个用例在采用方面目前是最不成熟的，但它预计将发展 并且随着理解的加深而完善，并且当实施挑战被识别时 并且回应了。	
观众	组织有兴趣使用人工智能代理系统来自动化复杂 包含多个交互工作流的企业任务和工作流程。
目的	多智能体AI系统具备自主决策的能力 并且有多个代理协同工作采取行动来操作 在有限的 human 监督下协同工作以实现复杂目标。 多智能体ai系统的特点包括理解能力 背景、原因、计划、适应、协调行动，并执行任务。
用例 描述	多个智能体协同工作以提取数据并在其基础上采取行动 其中的数据，利用如MCP等标准化协议 A2A for agent- 与代理通信。例如，基于参考架构 OWASP多代理系统威胁建模指南: 机器人流程 自动化费用报销代理。多个代理人负责 从费用报销单中提取信息（例如，提交的收据、表格）， 利用 RAG 验证公司政策层面的主张，并进行路由 批准支付的费用以自动化员工报销 企业中的工作流程。

用例5：人工智能开发者的安全控制	
观众	人工智能开发者
目的	NIST 开发了一个 SSDF 社区简介：安全软件开发 生成式人工智能和 dual-use 基础模型的实践 () SP 800-218A), 识别安全关键模型工件和良好实践 保护它们。NIST CAISI 发布了《管理双重用途风险》。 基础模型 (草稿AI 800-1)作为人工智能开发者的资源。一个映射 从 SP 800-53 中的安全控制到 SP 中的这些工件和实践 800-218A以及在草稿AI 800-1中，资源待定，可惠及社区 人工智能开发人员，并允许基于现有进行有效的风险管理 组织实践。

NIST叠加层保护人工智能Slack频道

安全控制叠加层对于人工智能系统项目将利用一个新启动的 NIST 用于保障人工智能叠加层 Slack 渠道 (#nist-overlays-securing-ai)，这是一个网络安全和人工智能社区讨论这些叠加层发展的中心。

加入 NIST人工智能叠加频道 要获取更新，请与NIST首席研究员和其他小组成员参与促进性讨论，分享想法，提供实时反馈，并为叠加开发做出贡献。所有感兴趣方都欢迎。

下一步与 NIST 合作

nist希望您对提议的高级用例和潜在的未来工作进行反馈。具体而言，nist寻求对以下方面的反馈：

- 用例如何捕捉用户社区中具有代表性的AI应用类型，有哪些潜在差距领域需要解决？ - 架构和示例用例在多大程度上反映了现实世界的应用模式，可能在哪些地方存在差距或问题？ - NIST应该如何优先考虑用例的叠加开发？ - 是否有额外的领域或用例需要考虑用于未来的工作？

关于上述问题的反馈以及对概念性文件的意见应发送至 overlays-securing- ai@list.nist.gov ，也可以在Slack频道中分享。了解更多关于AI项目的控制覆盖层、Slack空间以及如何参与的信息。 https://csrc.nist.gov/preview/projects/cosais .

根据反馈，NIST将启动第一个用例，目标是在2026财年早期发布公开征求意见稿，一次一个用例。在公开征求意见期间，还将举办一场公开研讨会，以持续与利益相关者互动。