人工智能系统安全指南

新加坡网络安全局 (Cyber Security Agency of Singapore, CSA)

2024年10月

中文翻译版 · 仅供参考，以英文原文为准

原文来源：https://www.csa.gov.sg/

目录

---

1. 引言

人工智能（Artificial Intelligence，AI）为经济、社会和国家安全带来了诸多益处。它有潜力在几乎每个行业——从商业和医疗保健到交通和网络安全——推动效率提升和创新。

要充分利用AI的益处，用户必须相信AI将按照设计运行，且结果是安全可靠的。然而，除安全性风险外，AI系统还可能受到对抗性攻击（adversarial attacks）的威胁，即恶意行为者蓄意操纵或欺骗AI系统。AI的采用可能引入或加剧企业系统面临的现有网络安全风险。这些风险可能导致数据泄露或数据违规，或产生有害的或不期望的模型输出。

因此，作为一项关键原则，AI应当和所有软件系统一样，**在设计上安全、在默认状态下安全**（secure by design and secure by default）。这将使系统所有者能够从上游管理安全风险。这将补充系统所有者为解决AI安全性以及公平性或透明度等其他相关考量而可能采取的其他控制措施和缓解策略（这些不在本文件讨论范围内）。

新加坡网络安全局（Cyber Security Agency of Singapore，CSA）制定了《人工智能系统安全指南》，供系统所有者在AI的整个生命周期中保障其使用安全。随着AI日益融入企业系统，应从系统层面整体考虑安全问题。因此，这些指南应与现有的IT环境安全最佳实践和要求配合使用。虽然这些指南不是强制性的，但我们强烈鼓励系统所有者考虑这些关键原则，以便就AI的采用及其潜在风险做出明智的决定。

AI安全是一个不断发展的工作领域，缓解控制措施也在持续演进。为此，CSA还与AI和网络安全从业者合作编制《人工智能系统安全配套指南》（Companion Guide on Securing AI Systems）。该配套指南旨在作为社区驱动的资源，补充本指南，为系统所有者在采用本指南时可根据其使用场景考虑的实用措施和控制提供有用参考。配套指南不是强制性的、规定性的或详尽的。由于AI安全领域持续快速发展，配套指南将不断更新以反映该领域的重大发展。

1.1 本文件的目的和范围

目的

本指南旨在支持正在采用或考虑采用AI系统的系统所有者。它识别了与AI使用相关的潜在安全风险，并为在AI生命周期的每个阶段缓解安全风险提供了指南。

本文件可与《人工智能系统安全配套指南》配合阅读，该配套指南提供了系统所有者在实施本指南时可考虑的实用安全控制措施的信息汇编。

范围

本指南解决AI系统面临的网络安全风险。它不寻求解决AI安全性（safety），或AI的其他常见相关考量（如公平性、透明度或包容性），也不解决AI系统引入的网络安全风险，尽管一些建议的措施可能存在重叠。它也不涉及AI在网络攻击中的滥用（AI赋能的恶意软件）、虚假/误导信息和诈骗（深度伪造）。

2. 了解AI威胁

AI是一种软件系统，本身容易受到网络威胁的影响，同时也为其所集成或对接的更广泛企业系统带来了新的攻击面。因此，保障AI安全是在实践良好"传统"网络安全卫生之外的额外要求。

保障AI系统安全引入了在传统IT系统中可能不常见的新挑战。除传统网络安全风险外，AI本身还容易受到对抗性机器学习（Adversarial Machine Learning，ML）等新型攻击的威胁，这些攻击旨在扭曲模型的行为。有关AI安全威胁的更多详情，请参阅附录A。

传统网络安全风险对AI系统的影响

AI系统需要大量数据进行训练；有些还需要导入外部模型和库。如果安全保障不充分，AI系统可能被供应链攻击所破坏，或可能因AI模型或底层IT基础设施中的漏洞而遭受入侵或未经授权的访问。此外，如果云服务、数据中心运营或其他数字基础设施中断（例如通过拒绝服务攻击），组织和用户将面临无法访问和使用AI工具的风险，这反过来可能导致依赖AI工具运行的系统瘫痪。

对抗性机器学习（Adversarial Machine Learning）

恶意行为者可能使用新型对抗性机器学习技术攻击AI模型和数据，影响机器学习模型产生不准确、有偏见或有害的输出；和/或泄露机密信息。对抗性机器学习攻击包括：数据投毒（data poisoning，向训练数据集注入恶意或损坏的数据）或规避攻击（evasion attacks，对已训练模型的攻击）以扭曲结果，推断攻击（inference attacks）或提取攻击（extraction attacks，探测模型）以暴露敏感或受限数据，或窃取模型。

3. 保障AI安全

AI安全是一个被广泛关注的问题，但该工作领域仍相对新兴。虽然从业者持续扩大关于AI安全威胁的研究和资源，但本指南列出了系统所有者应采取的关键考量，以支持AI的安全采用。鉴于AI发展的快速步伐，系统所有者应持续了解AI安全的最新发展，并相应更新其风险管理策略。

3.1 采用生命周期方法

AI系统生命周期有五个关键阶段——规划与设计、开发、部署、运营与维护、以及退役（End of Life）。

与良好的网络安全实践一样，CSA建议系统所有者采用生命周期方法来考虑安全风险。仅加固AI模型不足以确保对AI相关威胁进行全面防御。参与AI系统生命周期的所有利益相关者都应寻求更好地了解安全威胁及其对AI系统预期结果的潜在影响，以及需要做出哪些决策或权衡。

AI生命周期代表了设计AI解决方案以满足业务或运营需求的迭代过程。因此，系统所有者在交付AI解决方案的过程中可能会多次重新审视生命周期中的规划与设计、开发和部署步骤。

一些组织可能已实施机器学习运维（Machine Learning Operations, ML Ops）流水线，其可能与AI系统开发生命周期（AI SDLC）不完全对应。尽管如此，运行包含ML设计、开发和运营阶段的开发运维流水线的ML Ops团队会发现，AI SDLC中规划与设计、开发、部署和运营阶段的指南同样适用。

3.2 从风险评估开始

鉴于AI使用场景的多样性，实施安全没有万能的解决方案。因此，有效的网络安全始于进行风险评估。这将使组织能够识别潜在风险、确定优先事项，进而制定适当的风险管理策略。

AI与传统软件的根本区别在于：传统软件依赖静态规则和显式编程，而AI使用机器学习和神经网络来自主学习和决策，无需为每项任务提供详细指令。因此，组织应考虑比传统系统更频繁地进行风险评估，即使其风险评估方法总体上基于现有的治理和政策。这些评估还可以通过持续监控和强有力的反馈循环来补充。

我们建议以下四个步骤来量身定制系统化的防御计划，以最佳方式应对组织的最高优先级风险——保护您最关心的资产。

步骤1：进行风险评估，聚焦AI系统的安全风险

进行风险评估，聚焦与AI系统相关的安全风险，可基于最佳实践或组织现有的企业风险评估/管理框架。

风险评估可参考CSA发布的指南（如适用）： - 《网络威胁建模指南》（Guide To Cyber Threat Modelling） - 《关键信息基础设施网络安全风险评估指南》（Guide To Conducting Cybersecurity Risk Assessment for Critical Information Infrastructure）

步骤2：根据风险/影响/资源确定优先处理领域

根据风险水平、影响和可用资源，确定优先处理哪些风险。

步骤3：识别并实施相关措施以保障AI系统安全

识别相关措施和控制手段以保障AI系统安全，例如参考《人工智能系统安全配套指南》中列出的措施，并在AI生命周期中实施。

步骤4：评估残余风险以进行缓解或接受

在为AI系统实施安全措施后评估残余风险，为接受或处理残余风险的决策提供依据。

3.3 人工智能系统安全指南

这些指南适用于AI系统生命周期的各个阶段。系统所有者应将其视为在保障AI采用安全方面需要考虑的关键问题。鉴于使用场景的多样性和AI安全的发展，这些指南不提供规定性的控制措施或要求。

系统所有者应将这些指南应用于其具体情境，并可参考《人工智能系统安全配套指南》了解潜在的控制措施。

1. 规划与设计

1.1 提高对安全风险的认知和能力

组织应了解AI带来的潜在安全风险，以便就AI的采用做出明智的决定。为所有人员（包括开发人员、系统所有者和高级管理人员）提供关于AI安全风险的充分培训和指导。

1.2 进行安全风险评估

风险管理策略应以安全风险评估为依据，这将有助于确定关键风险和优先事项。根据相关行业标准/最佳实践，采用全面的流程对AI系统的威胁和风险进行建模。

2. 开发

2.1 保障供应链安全

AI供应链包括（但不限于）训练数据、模型、API和软件库。每个组件都可能引入新的漏洞（例如，模型可能携带编码为模型参数的恶意软件，使攻击者能够提取和注入恶意软件到用户机器上）。在AI系统的整个生命周期中评估和监控其供应链的潜在安全风险。确保供应商遵守安全政策和国际公认标准，或以其他方式适当管理风险。考虑评估供应链组件（例如通过软件物料清单[SBOM]、代码检查或对照漏洞数据库）。

2.2 在选择适当模型时考虑安全收益和权衡

不同的AI模型（如机器学习、深度学习、生成式）具有独特的特征和风险（例如，大语言模型可能容易受到输入操纵攻击），因此需要不同的安全措施。在为您的系统开发或选择适当的AI模型时，考虑可能影响其安全性的因素（如复杂性、可解释性、可理解性以及训练数据的敏感性）。

2.3 识别、追踪和保护AI相关资产

随着AI系统日益融入业务运营，它们将成为组织战略资产的一部分，应相应地加以保护。否则，敏感数据、知识产权和组织资产将面临潜在威胁和违规的风险。了解AI相关资产的价值，包括模型、数据、提示、日志和评估。建立流程来追踪、验证、版本控制和保护资产。

2.4 保障AI开发环境安全

AI模型需要访问大量训练数据，不安全的开发环境可能引入数据违规风险（例如个人身份信息或机密商业信息的暴露）。不安全的开发还可能使AI模型容易受到攻击（如投毒攻击），导致模型行为被破坏，或使模型和其他知识产权面临被盗、未经授权复制或滥用的风险。应用标准基础设施安全原则，如实施适当的访问控制和日志/监控、环境隔离以及默认安全配置。

3. 部署

3.1 保障AI系统的部署基础设施和环境安全

与2.4"保障AI开发环境安全"类似的考量。应用标准基础设施安全原则，如访问控制和日志/监控、环境隔离、默认安全配置以及防火墙。

3.2 建立事件管理程序

AI系统是复杂且自适应的，这有时可能导致不可预测的行为。鉴于AI使用场景的多样性，事件范围可从聊天机器人故障等小问题到关键基础设施运营中断等重大后果。系统所有者应制定适当的事件响应、升级和补救计划。

3.3 负责任地发布AI系统

AI系统可能面临上述风险，包括滥用、数据违规和模型操纵。这些风险影响用户的信任和信心，并可能对组织的声誉产生影响。良好的做法是仅在经过适当有效的安全检查和评估后才发布模型、应用程序或系统。

4. 运营与维护

4.1 监控AI系统输入

AI系统是动态的且对输入具有自适应性。现实中已有事件表明，用户/攻击者故意构造输入以欺骗AI系统做出错误或非预期的决定。AI系统所有者可能希望监控和记录AI系统的输入，如查询、提示和请求，因为第三方提供商可能出于隐私原因不这样做。适当的日志记录有助于合规、审计、调查和补救。

4.2 监控AI系统输出和行为

AI系统可能在生产阶段出现故障或性能下降。部署后监控模型将确保其按预期运行，并在出现潜在问题时（无论是由对抗性攻击还是其他原因造成）向系统所有者发出警报。运营人员应监控可能表明入侵、被破坏或数据漂移（data drift）的异常行为。

4.3 对更新和持续学习采用安全设计方法

数据和模型的变化可能导致行为变化。系统所有者应确保与模型更新相关的风险已被考虑并得到适当管理。

4.4 建立漏洞披露流程

即使有监控机制，AI的自适应特性也可能使检测攻击和非预期行为变得具有挑战性。应有反馈流程供用户分享任何令人担忧的发现，这可能揭示系统的潜在漏洞。

5. 退役

5.1 确保数据和模型的妥善处置

由于模型基于大量训练数据（包括可能的机密信息）进行训练，不当处置可能导致数据违规等事件。应根据相关行业标准或法规对数据和模型进行适当和安全的处置/销毁。

术语表

术语	简要说明
AI系统（AI system）	人工智能。一种基于机器的系统，针对明确或隐含的目标，从其接收的输入中推断如何生成预测、内容、建议或决策等输出，这些输出可以影响物理或虚拟环境。不同的AI系统在部署后的自主性和适应性水平各不相同。
对抗性机器学习（Adversarial Machine Learning）	提取关于ML系统行为和特征信息的过程，和/或学习如何操纵ML系统的输入以获得期望结果。
异常检测（Anomaly Detection）	识别偏离通常、标准或预期的观察、事件或数据点，使其与其余数据不一致。
API（应用程序编程接口）	Application Programming Interface。一组确定两个软件应用程序如何相互交互的协议。
后门攻击（Backdoor attack）	攻击者在训练过程中巧妙地修改AI模型，导致在某些触发条件下出现非预期行为。
聊天机器人（Chatbot）	一种旨在通过文本或语音命令模拟人类对话的软件应用程序。
计算机视觉（Computer Vision）	一个跨学科的科学技术领域，专注于计算机如何从图像和视频中获得理解。
数据违规（Data Breach）	当威胁行为者获得对敏感/机密数据的未授权访问时发生的事件。
数据完整性（Data Integrity）	数据未被以未授权方式更改的属性。数据完整性涵盖存储中的数据、处理中的数据和传输中的数据。
数据泄露（Data Leakage）	敏感、受保护或机密信息在其预期环境之外的无意暴露。
数据丢失防护（Data Loss Prevention）	系统识别、监控和保护使用中的数据（如终端操作）、传输中的数据（如网络操作）和静态数据（如数据存储）的能力，通过深度包内容检查和交易的上下文安全分析在集中管理框架内实现。
数据投毒（Data Poisoning）	通过修改训练数据来控制模型。
数据科学（Data Science）	一个跨学科的技术领域，使用算法和流程收集和分析大量数据，以发现为商业决策提供信息的模式和见解。

术语	简要说明
深度学习（Deep Learning）	AI的一项功能，通过学习人脑如何构建和处理信息来做出决策来模仿人脑。这种机器学习的子集可以在无监督的情况下从非结构化数据中学习，而不是依赖只能执行一项特定任务的算法。
纵深防御（Defence-in-Depth）	纵深防御是一种利用多层安全措施保护组织资产的策略。其理念是，如果一道防线被突破，还有额外的层作为后备，确保威胁在途中被阻止。
规避攻击（Evasion attack）	构造AI的输入以误导其错误执行任务。
提取攻击（Extraction attack）	通过适当采样输入空间并观察输出来复制或窃取AI模型，以构建行为类似的替代模型。
生成式AI（Generative AI）	一种专注于创建新数据（包括文本、视频、代码和图像）的机器学习类型。生成式AI系统使用大量数据进行训练，以便找到生成新内容的模式。
护栏（Guardrails）	对AI系统施加的限制和规则，以确保其适当处理数据且不生成不道德的内容。
幻觉（Hallucination）	AI系统的错误回应，或输出中以事实信息形式呈现的虚假信息。
图像识别（Image Recognition）	图像识别是在图像或视频中识别对象、人物、地点或文本的过程。
LLM（大语言模型）	Large Language Model。一种处理和生成类人文本的AI模型。LLM专门在大量自然语言数据集上训练，以生成类人输出。
ML（机器学习）	Machine Learning。AI的一个子集，融合了计算机科学、数学和编程的各个方面。机器学习专注于开发能够从数据中学习、并对新数据做出预测和决策的算法和模型。
成员推断攻击（Membership Inference attack）	数据隐私攻击，用于确定某个数据样本是否属于机器学习模型训练集的一部分。
NLP（自然语言处理）	Natural Language Processing。AI的一个子集，使计算机能够理解口语和书面人类语言。NLP支持设备上的文本和语音识别等功能。
神经网络（Neural Network）	一种旨在模拟人脑结构的深度学习技术。神经网络需要大量数据集来进行计算和创建输出，从而支持语音和视觉识别等功能。

术语	简要说明
过拟合（Overfitting）	在机器学习训练中，当算法只能处理训练数据中的特定示例时发生。正常运行的AI模型应能够泛化数据中的模式以处理新任务。
提示（Prompt）	提示是用户向AI系统提供的自然语言输入，以获取结果或输出。
强化学习（Reinforcement Learning）	一种机器学习类型，算法通过与环境交互来学习，然后根据其行动获得奖励或惩罚。
SDLC（软件开发生命周期）	Software Development Life Cycle。将安全考量和实践整合到软件开发各个阶段的过程。这种整合对于确保软件从设计阶段到部署和维护都是安全的至关重要。
训练数据（Training data）	训练数据是提供给AI系统的信息或示例，使其能够学习、发现模式和创建新内容。

附录A：了解AI威胁

对抗性威胁是由蓄意造成损害的威胁行为者引起的。通常，这些威胁行为者被称为攻击者或对手。

要了解这些威胁，系统所有者可以参考OWASP大语言模型应用十大安全风险（OWASP Top 10 for Large Language Model Applications）、OWASP机器学习安全十大风险（OWASP Machine Learning Security Top 10），或MITRE ATLAS™（人工智能系统对抗性威胁态势，Adversarial Threat Landscape for Artificial-Intelligence Systems）等资源。MITRE ATLAS特别提供了一个结构化的知识库，供AI和网络安全专业人员理解和防御AI网络威胁。它基于真实世界的观察、ML红队和安全团队的演示以及学术研究中的前沿可能性，汇编了AI系统的对手战术、技术和案例研究。

任何保障AI系统安全的尝试都应建立在"传统"良好网络安全卫生的基础之上，如实施最小权限原则、多因素认证、持续安全监控和审计。

ATLAS矩阵涵盖两类对抗性"技术":

- **AI/ML系统特有的技术**（在原文中以橙色方框标示），以及
- **传统网络安全攻击技术**，适用于AI和非AI系统，直接来自MITRE Enterprise ATT&CK矩阵（在原文中以白色方框标示）。

系统所有者应继续利用这些资源建立对安全威胁的认知，以更好地了解可能对其AI采用产生影响的新兴风险。随着该领域的持续发展，这些资源将帮助AI和网络安全团队进行安全风险评估和管理活动。

MITRE ATLAS矩阵

以下列出MITRE ATLAS矩阵中的主要战术类别及其相关技术：

侦察（Reconnaissance）： 搜索受害者公开可用的研究材料、搜索公开可用的对抗性漏洞分析、搜索受害者拥有的网站、搜索应用程序存储库、主动扫描。

资源开发（Resource Development）： 获取公共ML工件、获取能力、开发能力、获取基础设施、发布投毒数据集、投毒训练数据、建立账户。

初始访问（Initial Access）： ML供应链入侵、有效账户、规避ML模型、利用面向公众的应用程序、LLM提示注入、网络钓鱼。

ML模型访问（ML Model Access）： ML模型推理API访问、ML赋能的产品或服务、物理环境访问、完全ML模型访问。

执行（Execution）： 用户执行、命令和脚本解释器、LLM插件入侵。

持久化（Persistence）： 投毒训练数据、植入ML模型后门、LLM提示注入。

权限提升（Privilege Escalation）： LLM提示注入、LLM插件入侵、LLM越狱。

防御规避（Defence Evasion）： 规避ML模型、LLM提示注入、LLM越狱。

凭证访问（Credential Access）： 不安全的凭证。

发现（Discovery）： 发现ML模型本体、发现ML模型系列、发现ML工件、LLM元提示提取。

收集（Collection）： ML工件收集、来自信息库的数据、来自本地系统的数据。

ML攻击准备（ML Attack Staging）： 创建代理ML模型、植入ML模型后门、验证攻击、构造对抗性数据。

数据窃取（Exfiltration）： 通过ML推理API窃取、通过网络手段窃取、LLM元提示提取、LLM数据泄露。

影响（Impact）： 规避ML模型、拒绝ML服务、用干扰数据对ML系统进行垃圾攻击、侵蚀ML模型完整性、成本收割、外部危害。