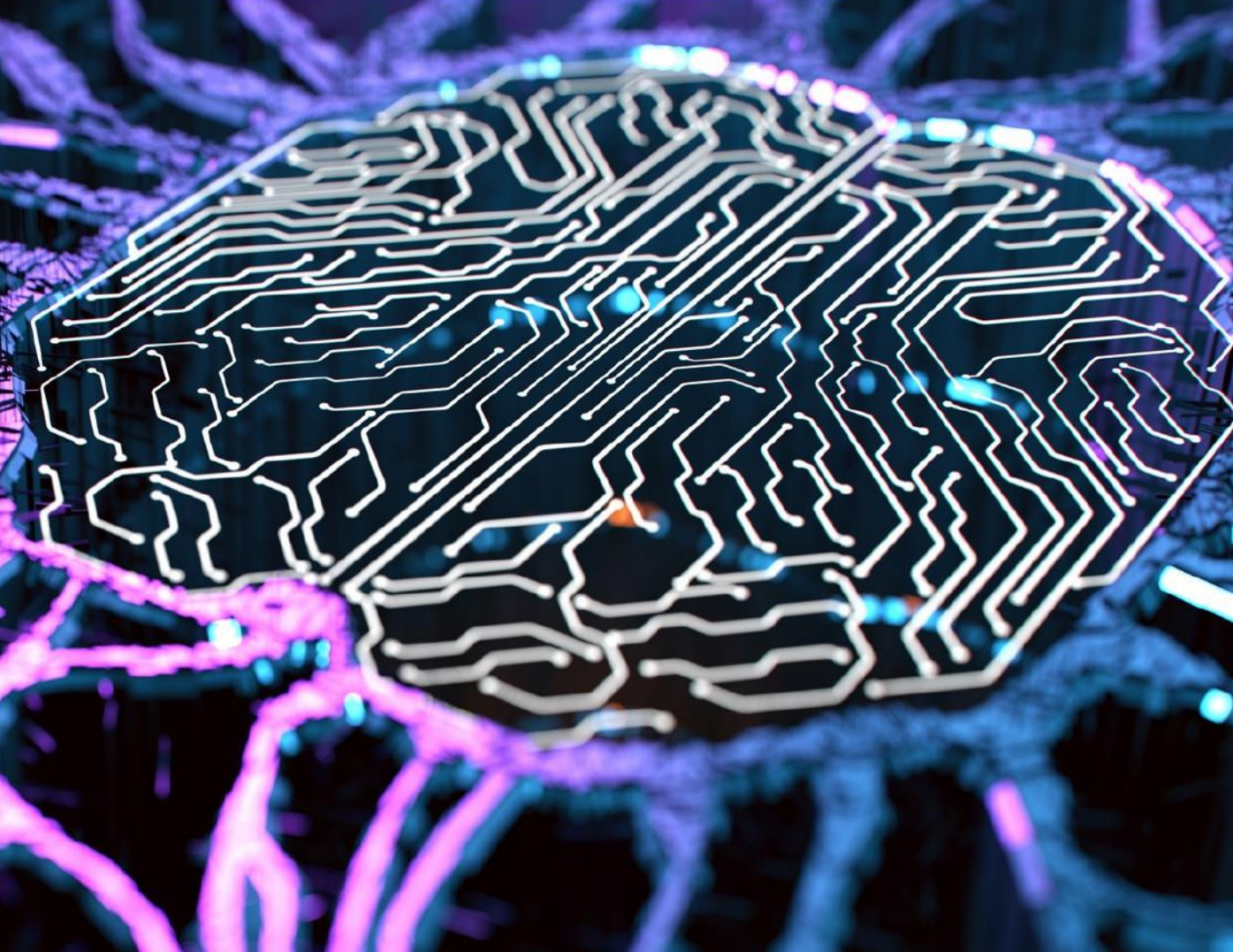
**附件：**

**安全人工智能系统开发指南**



本文系北京比瓴科技有限公司员工翻译，并在securityGPT社区公开分享。GPTSecurity（https://github.com/mo-xiaoxi/GPTSecurity）是一个涵盖了前沿学术研究和实践经验分享的社区，集成了生成预训练Transformer（GPT）、人工智能生成内容（AIGC）以及大型语言模型（LLM）等安全领域应用的知识。

翻译者：高梓茹、刘舒骐、赵忠祥

**原文地址：**

<https://www.cisa.gov/news-events/news/dhs-cisa-and-uk-ncsc-release-joint-guidelines-secure-ai-system-development>





**关于本文件**

本文件由英国国家网络安全中心(NCSC)、美国网络安全与基础设施安全局(CISA)及以下国际合作伙伴联合发布:

美国国家安全局(NSA)

美国联邦调查局(FBI)

澳大利亚信号局澳大利亚网络安全中心(ACSC)

加拿大网络安全中心

新西兰国家网络安全中心(NCSC-NZ)

智利政府CSIRT

捷克国家网络和信息安全局(NUKIB)爱沙尼亚信息系统管理局(RIA)

爱沙尼亚国家网络安全中心(NCSC-EE)

法国网络安全机构(ANSSI)

德国联邦信息安全局(BSI)

以色列国家网络局(INCD)

意大利国家网络安全机构

日本国家网络安全事件准备与战略中心(NISC)

日本科学、技术和创新政策秘书处、内阁府

尼日利亚国家信息技术开发局(NITDA)

挪威国家网络安全中心(NCSC-NO)

波兰数字事务部

波兰NASK国家研究院(NASK)

大韩民国国家情报院(NIS)

新加坡网络安全局(CSA)

**致谢**

以下组织对这些指南的制定做出了贡献:

AlanTuringInstitute

Amazon

Anthropic

Databricks

GeorgetownUniversity’sCenterforSecurityandEmergingTechnology

Google

GoogleDeepMind

HuggingFace

IBM

Imbue

Inflection

Microsoft

OpenAI

Palantir

RAND

ScaleAI

SoftwareEngineeringInstituteatCarnegieMellonUniversity

StanfordCenterforAISafety

StanfordProgramonGeopolitics,TechnologyandGovernance

**免责声明**

本文件中的信息是由NCSC和撰写机构“按原样”提供的，除了法律规定的情况外，他们不对因使用本文件而造成的任何损失、伤害或损害承担责任。本文档中的信息不构成或暗示NCSC或作者机构对任何第三方组织、产品或服务的认可或推荐。对网站和第三方材料的链接和引用仅供参考，不代表对此类资源的认可或推荐。

本文件以TLP:CLEAR为基础(https://www.first.org/tlp/)提供。

目录

[执行概要 5](#_Toc20481)

[关于本指南 5](#_Toc32693)

[介绍 6](#_Toc23759)

[安全AI系统开发指南 8](#_Toc31439)

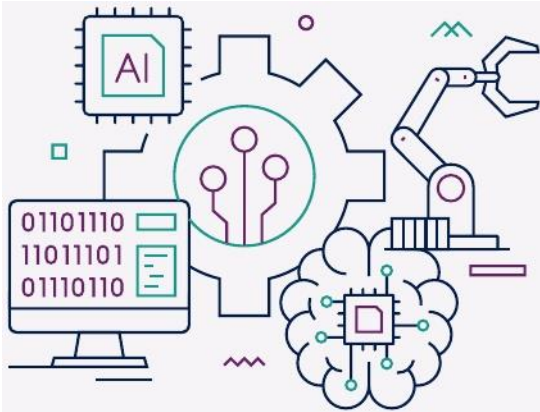
[1.安全的设计 8](#_Toc21269)

[2.安全开发 10](#_Toc31975)

[3.安全的部署 11](#_Toc19248)

[4.安全维护 12](#_Toc4613)

[进一步的阅读 14](#_Toc10667)



**执行概要**

本文件为使用人工智能(AI)任何系统的提供者推荐了指导方针，无论这些系统是从头开始创建的，还是建立在其他人提供的工具和服务之上的。实施这些指导方针将有助于提供商构建按预期运行、在需要时可用并且不向未授权方泄露敏感数据的情况下工作的AI系统。

本文档主要针对使用由组织托管的模型或使用外部应用程序编程接口(API)的AI系统提供商。我们敦促所有利益相关者(包括数据科学家、开发人员、管理人员、决策者和风险所有者)阅读这些指南，以帮助他们在AI系统的设计、开发、部署和运营方面做出明智的决策。

**关于本指南**

AI系统有可能给社会带来许多好处。然而，为了充分利用AI的机会，必须以安全和负责任的方式开发、部署和运营它。

AI系统容易受到新的安全漏洞影响，需要与网络安全威胁一起考虑。当发展速度很高时——就像AI的情况一样——安全往往是次要的考虑因素。安全性是AI开发系统的核心需求，不仅仅是在开发阶段，而是贯穿在系统的整个生命周期中。

因此，指南在AI系统开发生命周期中分为四个关键领域:安全设计、安全开发、安全部署以及安全操作和维护。对于每个部分，我们建议考虑和缓解措施，这将有助于降低组织AI系统开发过程的总体风险。

**1.安全的设计**

本节包含适用于AI系统开发生命周期的设计阶段的指导方针。它涵盖了对风险和威胁建模的理解，以及在系统和模型设计中要考虑的具体主题和权衡。

**2.安全开发**

本节包含适用于AI系统开发生命周期的开发阶段的指导方针，包括供应链安全、文档以及资产和技术债务管理。

**3.安全的部署**

本部分包含适用于AI系统开发生命周期的部署阶段的指导方针，包括保护基础设施和模型免受损害、威胁或损失、开发事件管理流程以及负责任发布。

**4.安全维护**

本节包含适用于AI系统开发生命周期安全运维阶段的指导原则。它提供了系统部署后相关的行动指南，包括日志记录和监控、更新管理和信息共享。

该指南遵循“默认安全”方法，并与NCSC的安全开发和部署指南、NIST的安全软件开发框架以及由CISA、NCSC和国际网络机构发布的“设计原则安全”中定义的实践密切一致。他们优先考虑:

* 为客户掌控安全成果
* 拥抱彻底的透明度和问责制
* 建立组织结构和领导力，确保设计安全是企业的首要任务

**介绍**

人工智能(AI)系统可能给社会带来许多好处。然而，为了充分利用AI的机会，必须以安全和负责任的方式开发、部署和运营它。网络安全是AI系统的安全性、弹性、隐私性、公平性、有效性和可靠性的必要前提。

然而，AI系统受到新的安全漏洞影响，需要与网络安全威胁一起考虑。当发展速度很高时——就像AI的情况一样——安全往往是次要的考虑因素。安全性是AI开发系统的核心需求，不仅仅是在开发阶段，而是贯穿在系统的整个生命周期中。

本文件为任何使用AI系统的提供者**[[1]](#footnote-0)**推荐了指导方针，无论这些系统是从头开始创建的，还是建立在其他人提供的工具和服务之上的。实施这些指导方针将有助于提供商构建按预期运行、在需要时可用，并且不向未授权方泄露敏感数据的情况下工作的AI系统。

这些指导方针应与既定的网络安全、风险管理和事件响应最佳实践结合考虑。我们敦促供应商遵循由美国网络安全和基础设施安全局(CISA)、英国国家网络安全中心(NCSC)和我们所有的国际合作伙伴制定的“设计安全”原则[[2]](#footnote-1)。这些原则优先考虑:

* 为客户掌控安全成果
* 拥抱彻底的透明度和问责制
* 建立组织结构和领导力，确保设计安全是企业的首要任务

遵循“安全设计”原则需要在整个系统生命周期中投入大量资源。开发人员必须在系统设计的每一层以及开发生命周期的所有阶段进行投资，并且对客户工具的功能、机制和实现进行优先级排序，这样做可以防止以后昂贵的重新设计，也可以在短期内保护客户以及其数据。

## 为什么AI安全不同?

在本文中，我们使用“AI”专指机器学习(ML)应用程序[[3]](#footnote-2)。所有类型的机器学习都在此范围内。我们将ML应用定义为:

* 涉及软件组件(模型)，这些组件(模型)允许计算机识别数据中的模式并为其带来上下文，而无需人类显式编程规则
* 基于统计推理生成预测、建议或决策

除了现有的网络安全威胁外，AI系统还面临着新型漏洞的影响。“对抗性机器学习”(AML)用于描述在机器学习组件中，包括硬件、软件、工作流和供应链中，利用基本性漏洞的攻击方式。AML使攻击者能够在机器学习系统中造成意外行为，其中包括:

* 影响模型的分类或回归性能
* 允许用户执行未经授权的操作
* 提取敏感模型信息

实现这些效果的方法有很多，比如在大型语言模型(LLM)领域进行提示注入攻击，或者故意破坏训练数据或用户反馈(称为“数据中毒”)。

## **谁应该阅读这份文件?**

本文件主要针对人工智能系统的提供商，无论是基于组织托管的模型还是利用外部应用程序接口（API）的系统。该指南敦促所有利益相关者，包括数据科学家、开发人员、管理人员、决策者和风险所有者阅读该指南，以帮助他们在 AI 系统的设计、开发、部署和运维方面做出明智的决策。

也就是说，并非所有的指导方针都能直接适用于所有组织。根据攻击AI系统的复杂程度和攻击方法因所面临的对手而有所不同，因此应将指导方针与组织的用例和威胁概况一起考虑。

## **谁负责开发安全的AI?**

在现代AI供应链中，通常有许多参与者。一个简单的方法假设了两个实体:

* 负责数据管理、算法开发、设计、部署和维护的“提供者”
* 提供输入并接收输出的“用户”

虽然这种“提供者-用户”方法在许多应用中使用，但它正变得越来越不常见[[4]](#footnote-3)，因为提供者可能希望将第三方提供的软件、数据、模型和/或远程服务纳入自己的系统。这些复杂的供应链使得最终用户更加难以理解安全AI的责任在哪里。

用户(无论是“最终用户”，还是包含外部AI组件的提供商[[5]](#footnote-4))通常没有足够的可见性和/或专业知识来充分理解、评估或解决与他们正在使用的系统相关的风险。因此，根据“设计安全”原则，AI组件的提供商应该对供应链下游用户的安全结果负责。

供应商应在其模型、管道和/或系统中尽可能实施安全控制和缓解措施，在使用设置的情况下，实施默认最安全的选项。在无法降低风险的情况下，提供商应负责:

* 告知供应链下游的用户：他们和(如果适用)他们自己的用户正在面临的风险
* 建议他们如何安全地使用组件

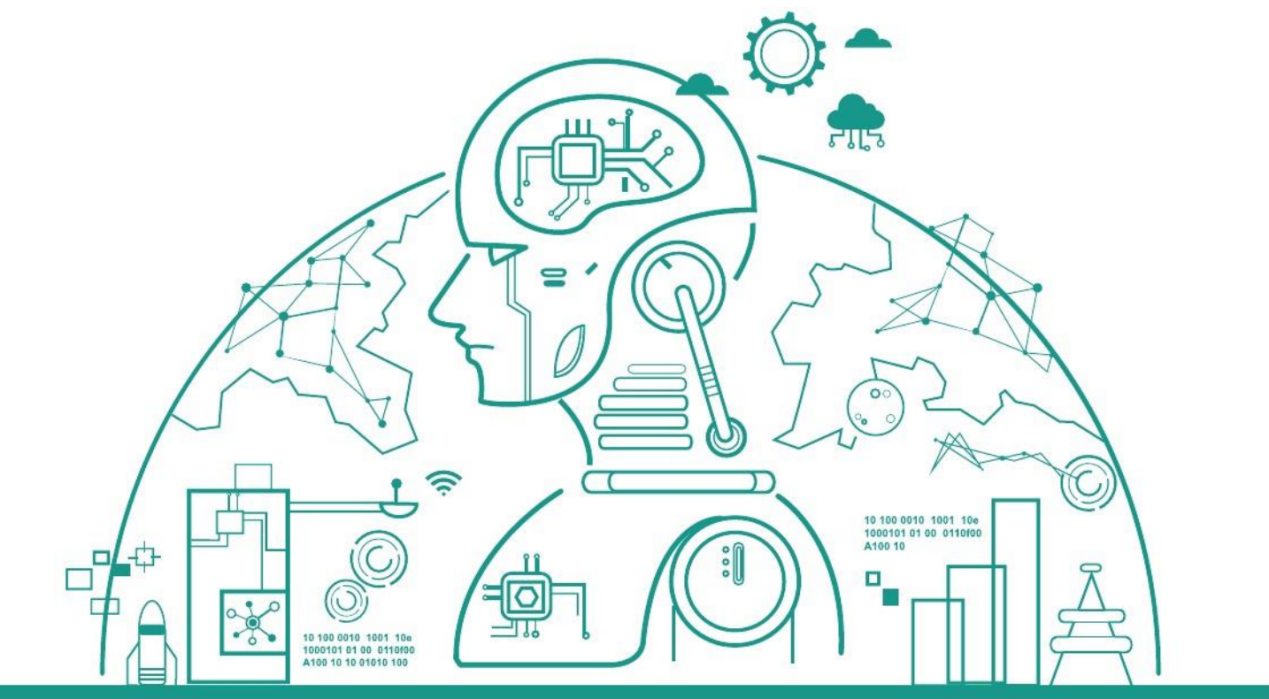
如果系统入侵可能导致有形或广泛的物理或声誉损害、重大业务运营损失、敏感或机密信息泄露和/或法律影响，则应将AI网络安全风险视为重大风险。

**安全AI系统开发指南**

该指南在AI系统开发生命周期中分为四个关键领域:安全设计、安全开发、安全部署、安全运营和维护。对于每个领域，我们建议考虑和缓解措施，这将有助于降低组织AI系统开发过程的总体风险。

本文件中列出的指导方针与软件开发生命周期实践密切一致，定义于:

* NCSC的安全开发和部署指南
* 美国国家标准与技术研究院(NIST)安全软件开发框架(SSDF)[[6]](#footnote-5)



## 1.安全的设计

本节包含适用于人工智能系统开发生命周期的设计阶段的指导方针。它涵盖了对风险和威胁建模的理解，以及在系统和模型设计中要考虑的具体主题和权衡。

**提高员工对威胁和风险的认识**

系统所有者和高级领导者应了解保护人工智能的威胁及其缓解措施。您的数据科学家和开发人员应保持对相关安全威胁和故障模式的认识，并帮助风险所有者做出明智的决策。为用户提供有关人工智能系统面临的独特安全风险的指导(例如，作为标准信息安全培训的一部分)，并对开发人员进行安全编码技术和安全负责任的人工智能实践培训。

**对系统所面临的威胁进行建模**

作为风险管理流程的一部分，您应用整体流程来评估对系统的威胁，其中包括了解如果人工智能组件受到损害或行为意外，对系统、用户、组织和更广泛的社会潜在影响[[7]](#footnote-6)。这个过程包括评估人工智能特定威胁的影响[[8]](#footnote-7)，并记录您的决策。

您意识到系统中使用的数据的敏感性和类型可能会影响其作为攻击目标的价值。您的评估应该考虑到随着人工智能系统越来越被视为高价值目标，一些威胁可能会增长，而且人工智能本身也会带来新的自动化攻击向量。

**在设计系统时，要考虑安全性以及功能和性能**

您确信手头的任务最适合使用人工智能来解决。确定这一点后，您要评估人工智能特定设计选择的适当性。您要考虑您的威胁模型和相关的安全缓解措施，以及功能、用户体验、部署环境、性能、保证、监督、道德和法律要求等因素。例如:

* 在选择是内部开发还是使用外部组件时，您会考虑供应链安全性，例如:
  + 您选择训练新模型，使用现有模型(有或没有微调)或通过外部API访问模型是否适合您的要求
  + 您与外部模型提供商合作的选择包括对该提供商自身安全状况的尽职调查评估
  + 如果使用外部库，您完成尽职调查评估(例如，确保库具有防止系统加载不受信任的模型而不会立即将自己暴露给任意代码执行[[9]](#footnote-8)的控制)
  + 您在导入第三方模型或序列化权重时实现扫描和隔离/沙箱，这些应被视为不受信任的第三方代码，并可能启用远程代码执行
  + 如果使用外部API，则对可以发送到组织控制之外的服务应有适当的控制，例如要求用户在发送潜在敏感信息之前登录并确认
  + 您们对数据和输入应用适当的检查和清理;这包括在将用户反馈或持续学习数据纳入您的模型时，认识到训练数据定义了系统行为
* 将AI软件系统开发集成到现有的安全开发和运营最佳实践中;AI系统的所有元素都是在适当的环境中编写的，使用编码实践和语言，在合理的情况下减少或消除已知的漏洞类别
* 如果AI组件需要触发操作，例如修改文件或将输出定向到外部系统，您可以对可能的操作应做适当的限制(这包括必要的外部AI和非AI故障保护)。
* 围绕用户交互的决策取决于AI特定的风险，例如:
  + 您的系统为用户提供可用的输出，而不会向潜在的攻击者透露不必要的细节
  + 如果有必要，您的系统在模型输出周围提供有效的护栏
  + 如果向外部客户或合作者提供API，则应用适当的控制，以减轻通过API对AI系统的攻击
  + 默认情况下，您应在系统中集成最安全的设置
  + 您应用最小权限原则来限制对系统功能的访问
  + 向用户解释风险较大的功能，并要求用户选择使用它们;您通报禁止使用的功能，并在可能的情况下告知用户替代解决方案

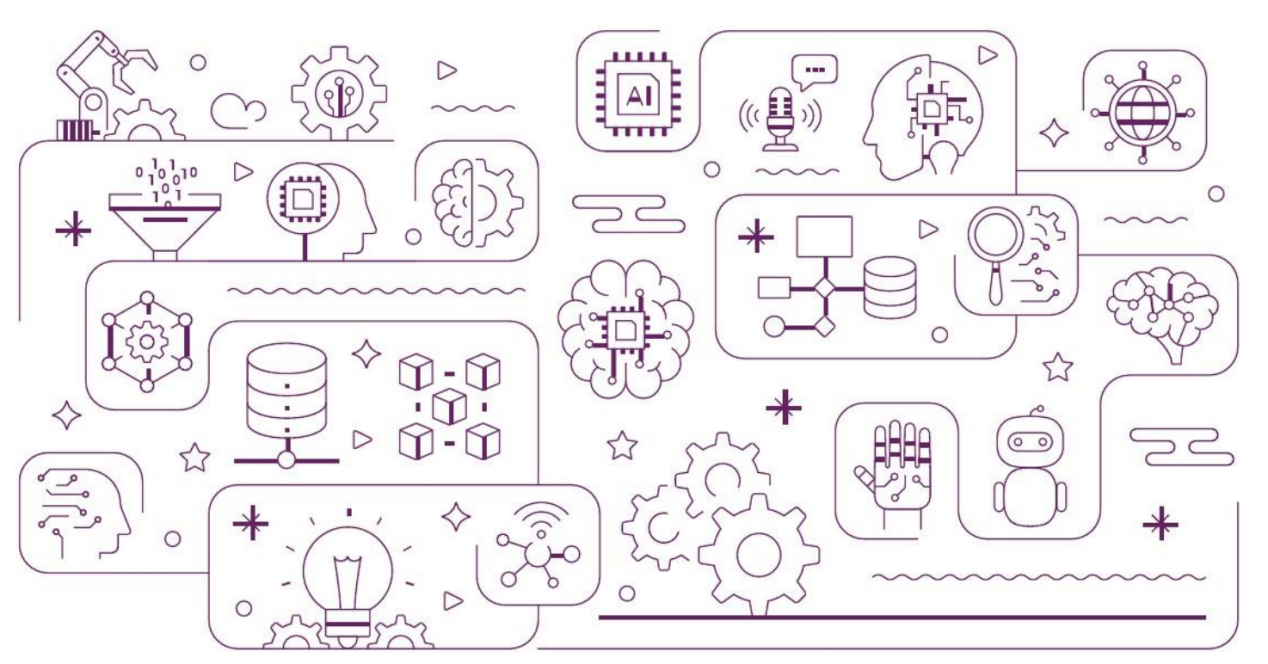
**在选择AI模型时，要考虑安全利益和权衡**

选择AI模型需要平衡一系列需求。这包括模型架构、配置、训练数据、训练算法和超参数的选择。您的决策由您的威胁模型提供信息，并随着AI安全研究的进展和对威胁理解的发展而定期重新评估。

在选择AI模型时，您的考虑因素可能包括但不限于:

* 您所使用的模型的复杂性，即所选择的架构和参数的数量;您的模型所选择的架构和参数数量将影响它需要多少训练数据，以及它在使用时对输入数据变化的鲁棒性
* 模型适合您的用例和/或使其适应您的特定需求的可行性(例如通过微调)
* 调整、解释和说明模型输出的能力(例如用于调试、审计或法规遵从);与难以解释的大型复杂模型相比，使用更简单、更透明的模型可能会有好处
* 训练数据集的特征，包括大小、完整性、质量、灵敏度、年龄、相关性和多样性
* 使用模型强化(如对抗性训练)、规范化和/或隐私增强技术的价值
* 组件的来源和供应链，包括模型或基础模型、训练数据和相关工具

有关这些因素中有多少影响安全结果的更多信息，请参阅NCSC的“机器学习安全原则”，特别是安全设计(模型架构)。



## 2.安全开发

该部分包含适用于AI系统开发生命周期的开发阶段的指导方针，包括供应链安全、文档以及资产和技术债务管理。

**保护您的供应链安全**

您评估和监控整个系统生命周期内AI供应链的安全性，并要求供应商遵守您自己的组织对其他软件的相同标准。如果供应商不能遵守组织的标准，则根据现有的风险管理政策采取行动。

如果不是内部生产，则从经过验证的商业、开源和其他第三方开发人员处获取并维护安全且记录良好的硬件和软件组件(例如，模型、数据、软件库、模块、中间件、框架和外部API)，以确保系统的强大安全性。

如果不符合安全标准，您准备好为关键任务系统切换到替代解决方案。您可以使用诸如NCSC的供应链指南和框架之类的资源，例如用于跟踪供应链和软件开发生命周期的认证的软件工件供应链级别(SLSA)[[10]](#footnote-9)。

**识别、跟踪和保护您的资产**

您了解您的人工智能相关资产对您的组织的价值，包括模型、数据(包括用户反馈)、提示、软件、文档、日志和评估(包括有关潜在不安全功能和故障模式的信息)，并认识到它们在哪些方面代表重大投资，以及在哪些方面使攻击者能够访问它们。您将日志视为敏感数据，并实施控制以保护其机密性、完整性和可用性。

您知道您的资产在哪里，并且已经评估并接受了任何相关的风险。您有流程和工具来跟踪、认证、版本控制和保护您的资产，并且可以在发生泄漏时恢复到已知的良好状态。

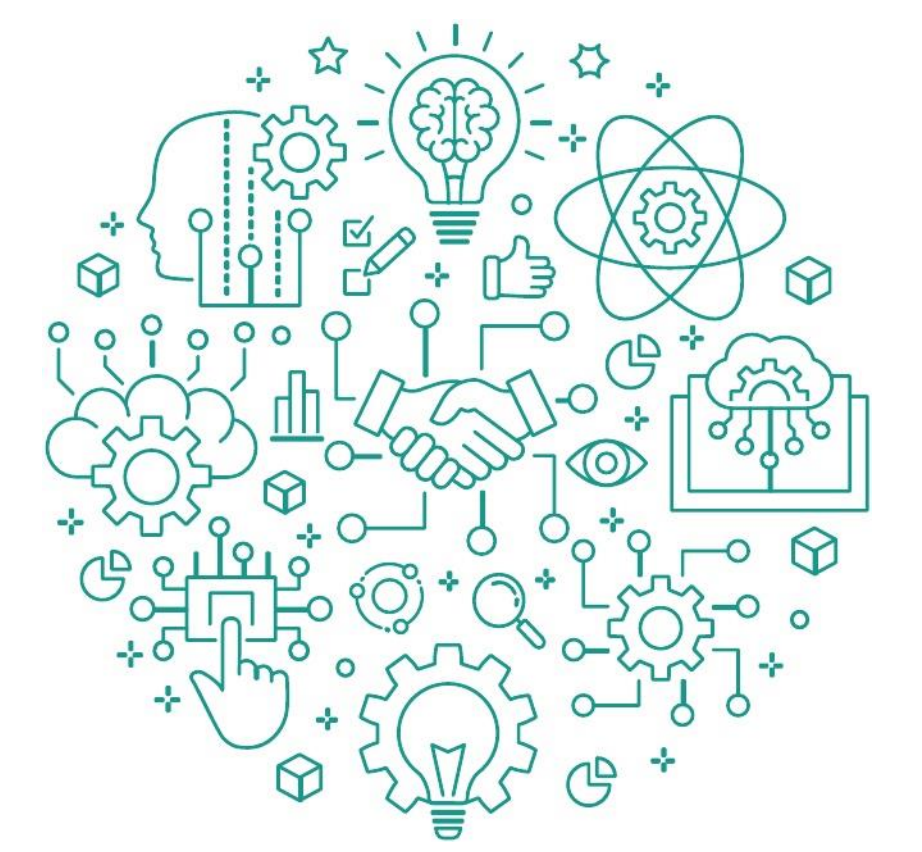
您有适当的流程和控制来管理AI系统可以访问的数据，并根据AI的敏感性(以及生成它的输入的敏感性)管理它生成的内容。

**记录您的数据、模型和提示**

您可以记录任何模型、数据集和元或系统提示符的创建、操作和生命周期管理。您们的文档包括与安全相关的信息，如培训数据的来源(包括微调数据和人工或其他操作反馈)、预期范围和限制、护栏、密码哈希或签名、保留时间、建议的审查频率和潜在的故障模式。有助于做到这一点的结构包括模型卡、数据卡和软件材料清单(SBOMS)。制作全面的文件支持透明度和问责制[[11]](#footnote-10)。

**管理您的技术债务**

与任何软件系统一样，您在整个AI系统的生命周期中识别、跟踪和管理您的“技术债务”(技术债务是指在牺牲长期利益的情况下，为实现短期结果而做出未能达到最佳实践的工程决策)。与金融债务一样，技术债务本身并不是坏事，但应从开发的最早阶段开始管理[[12]](#footnote-11)。您认识到，在AI环境中这样做比在标准软件中更具挑战性，并且由于快速的开发周期和缺乏完善的协议和接口，您的技术债务水平可能会很高。确保您的生命周期计划(包括AI系统退役过程)评估、确认并减轻对未来类似系统的风险。



## 3.安全的部署

本部分包含适用于AI系统开发生命周期的部署阶段的指导方针，包括保护基础设施和模型免受损害、威胁或损失、开发事件管理流程以及负责发布。

**保护您的基础设施**

您将良好的基础设施安全原则应用于系统生命周期中每个部分所使用的基础设施。在研发和部署过程中，对API、模型和数据以及它们的培训和处理管道实施适当的访问控制，这包括对保存敏感代码或数据的环境进行适当隔离，这也将有助于减轻为了窃取模型或损害其性能的标准网络安全攻击。

**持续保护您的模型**

攻击者可能能够通过直接访问模型(通过获取模型权重)或间接访问(通过应用程序或服务查询模型)来重建模型的功能[[13]](#footnote-12)或其被训练的数据[[14]](#footnote-13)。攻击者也可能在训练期间或之后篡改模型、数据或提示，使输出不可信。

您可以分别通过以下方式保护模型和数据不被直接和间接访问:

* 实施标准的网络安全最佳实践
* 在查询接口上实施控制，以检测和防止访问、修改和泄露机密信息的尝试

为了确保消费系统能够验证模型，您需要在模型训练完成后立即计算并共享模型文件(例如模型权重)和数据集(包括检查点)的密码哈希和/或签名。与密码学一样，良好的密钥管理是必不可少的[[15]](#footnote-14)。

您对保密风险缓解的方法会在很大程度上取决于用例和威胁模型。一些应用程序，例如那些涉及非常敏感数据的应用程序，可能需要理论上的保证，这可能很难或成本很高去实现。在适当的情况下，可以使用隐私增强技术(如差分隐私或同态加密)来探索或确保与消费者、用户和攻击者访问模型和输出相关的风险水平。

**制定事件管理程序**

您AI系统受安全事件影响的不可避免性反映在您的事件响应、升级和补救计划中。您的计划反映了不同的场景，并随着系统和更广泛的研究的发展而定期重新评估。您将公司重要的数字资源存储在离线备份中。响应人员已经接受了评估和处理人工智能相关事件的培训。您免费向客户和用户提供高质量的审计日志和其他安全功能或信息，以启用他们的事件响应流程。

**负责任地发布AI**

只有在对模型、应用程序或系统进行适当和有效的安全评估(如基准测试和攻防演练(以及超出本指南范围的其他测试，如安全性或公平性)）之后，您才能发布模型、应用程序或系统，并且您向用户说明已知的限制或者潜在的失效模式。开源安全测试库的详细信息在本文末尾的进一步阅读部分给出。

**让用户更容易做正确的事情**

您认识到，每个新的设置或配置选项都要与它所带来的业务利益以及它所带来的任何安全风险一起进行评估。理想情况下，最安全的设置将作为唯一的选项集成到系统中。当需要配置时，默认选项应该是针对常见威胁的广泛安全(即默认安全)。您可以实施控制来防止以恶意方式使用或部署系统。

您们向用户提供了关于正确使用您们的模型或系统的指导，其中包括强调限制和潜在的故障模式。您们向用户清楚地说明他们负责安全的哪些方面，并透明地说明他们的数据可能在哪里(以及如何)被使用、访问或存储(例如，如果它被用于模型再培训，或由员工或合作伙伴审查)。

## 4.安全维护

本节包含适用于AI系统开发生命周期安全运维阶段的指导原则。它提供了系统部署后特别相关的行动指南，包括日志记录和监控、更新管理和信息共享。

**监控系统的行为**

您测量您的模型和系统的输出和性能，这样您就可以观察到影响安全性的行为突然和逐渐的变化。您可以解释和识别潜在的入侵和妥协，以及自然的数据漂移。

**监控系统的输入**

根据隐私和数据保护要求，您监控并记录系统输入(例如推理请求、查询或提示)，以便在泄露或滥用的情况下实现合规义务、审计、调查和补救。这可能包括明确检测分发外和/或对抗性输入，包括那些旨在利用数据准备步骤(如裁剪和调整图像大小)的输入。

**遵循安全的设计方法进行更新**

在每个产品中默认包含自动更新，并使用安全的模块化更新程序来分发它们。您们的更新流程(包括测试和评估制度)反映了数据、模型或提示的更改可能导致系统行为变化的事实(例如，您们将重大更新视为新版本)。您支持用户评估和响应模型更改(例如，通过提供预览访问和版本化API)。

**收集和分享经验教训**

您参与信息共享社区，在工业、学术界和政府的全球生态系统中开展合作，酌情分享最佳实践。您们保持开放的沟通渠道，以便向组织内部和外部提供有关系统安全的反馈，包括同意安全研究人员研究和报告漏洞。必要时，您们将问题升级到更广泛的社区，例如发布响应漏洞披露的公告，包括详细和完整的常见漏洞枚举。您迅速和适当地采取行动，以减轻和补救问题。

**进一步的阅读**

**人工智能开发**

[机器学习的安全原则](https://www.ncsc.gov.uk/blog-post/introducing-our-new-machine-learning-security-principles)

NCSC关于开发、部署或操作带有ML组件的系统的详细指导。

[设计安全——改变网络安全风险的平衡:设计软件安全的原则和方法](https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf)

该指南由CISA、NCSC和其他机构共同撰写，描述了包括人工智能在内的软件系统制造商应如何采取措施，将安全因素纳入产品开发的设计阶段，并交付安全开箱的产品。

[简而言之，人工智能安全问题](https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_Al-Security_Guide_2023.pdf?__blob=publicationFile&v=5)

该文件由德国联邦信息安全办公室(BSI)制作，介绍了机器学习系统可能受到的攻击以及针对这些攻击的潜在防御措施。

[开发先进人工智能系统组织的广岛进程国际指导原则](https://www.mofa.go.jp/files/100573471.pdf)

作为七国集团广岛人工智能进程的一部分，这些文件为开发最先进的人工智能系统的组织提供指导，包括最先进的基础模型和生成人工智能系统，旨在促进全球安全，可靠和可信赖的人工智能。

[人工智能验证](https://aiverifyfoundation.sg/what-is-ai-verify/)

新加坡的人工智能治理测试框架和软件工具包，通过标准化测试，根据一套国际公认的原则验证人工智能系统的性能。

[人工智能良好网络安全实践的多层框架-ENISA(europa.eu)](https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai)

一个框架，指导国家主管当局和人工智能利益攸关方采取必要步骤，以确保其人工智能系统、运营和流程的安全

[ISO5338:人工智能系统生命周期过程(审查中)](https://www.iso.org/standard/81118.html)

描述基于机器学习和启发式系统的人工智能系统生命周期的一组过程和相关概念。

[人工智能云服务合规标准目录(AIC4)](https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4)

BSI的人工智能云服务合规标准目录提供了特定于人工智能的标准，可以对人工智能服务的整个

生命周期的安全性进行评估。

[NISTIR8269(草案)《对抗性机器学习分类和术语》](https://www.nccoe.nist.gov/ai/adversarial-machine-learning)

描述基于机器学习和启发式系统的人工智能系统生命周期的一组过程和相关概念。

[MITRE地图集](https://atlas.mitre.org/)

一个针对机器学习(ML)系统的对手战术、技术和案例研究的知识库，模仿并链接到MITREATT&CK框架。

[灾难性人工智能风险概述(2023年)](https://www.safe.ai/ai-risk)

该文件由人工智能安全中心制作，列出了人工智能带来的风险领域。

[大型语言模型:行业和当局的机遇和风险](https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/AI-in-lLanguage-processing.pdf?__blob=publicationFile&v=2)

BSI为希望更多地了解开发、部署和/或使用法学硕士的机会和风险的公司、当局和开发人员制作的文档。

[介绍人工智能](https://www.cyber.gov.au/resources-business-and-government/governance-and-user-education/governance/an-introduction-to-artificial-intelligence)

来自澳大利亚网络安全中心的博客，提供了关于人工智能以及如何安全参与的平易近人的指导。

帮助用户安全测试AI模型的开源项目包括:

* [AdversarialRobustnessToolbox](https://adversarial-robustness-toolbox.readthedocs.io/en/latest/)(IBM)
* [CleverHans](https://cleverhans-lab.github.io/)(UniversityofToronto)
* [TextAttack](https://textattack.readthedocs.io/en/latest/0_get_started/basic-Intro.html)(UniversityofVirginia)
* [PromptBench](https://github.com/microsoft/promptbench)(微软)
* [Counterfit](https://github.com/Azure/counterfit)(微软)。
* [AI验证](https://github.com/IMDA-BTG/aiverify)(新加坡信息通信媒体发展管理局)

**网络安全**

[CISA的网络安全绩效目标](https://www.cisa.gov/cpg)

所有关键基础设施实体应实施的一套通用保护措施，以有意义地降低已知风险和对手技术的可能性和影响。

[NCSCCAF框架](https://www.ncsc.gov.uk/collection/caf)

网络评估框架(CAF)为负责至关重要的服务和活动的组织提供指导。

[MITRE的供应链安全框架](https://sot.mitre.org/framework/system_of_trust.html)

用于评估供应链中的供应商和服务提供者的框架。

**风险管理**

[NISTAI风险管理框架(AIRMF)](https://www.nist.gov/itl/ai-risk-management-framework)

AIRMF概述了如何管理与AI相关的个人、组织和社会的社会技术风险。

[ISO27001:信息安全、网络安全和隐私保护](https://www.iso.org/standard/27001)

该标准为组织提供了建立、实施和维护信息安全管理体系的指南

[ISO31000:风险管理](https://www.iso.org/iso-31000-risk-management.html/)

一项国际标准，为组织提供组织内部风险管理的指导方针和原则

[NCSC风险管理指南](https://www.ncsc.gov.uk/collection/risk-management)

该指南帮助网络安全风险从业人员更好地了解和管理影响其组织的网络安全风险。

**笔记**

**1.**这里定义为开发AI系统(或开发AI系统)并将该系统投放市场或以自己的名称或商标投入使用的个人、公共机构、机构或其他机构

2.有关设计安全的更多信息，请参阅CISA的“设计安全”网页和指南“转移网络安全风险的平衡:设计安全软件的原则和方法”

3.与非mlAI方法(如基于规则的系统)相反

4.CEPS在其出版物《协调AI价值链与欧盟人工智能法案》中描述了七种不同类型的AI开发互动。

5.ISO/IEC22989:2022(en)将其定义为“构建AI系统的功能元素”

6.NIST的任务是制定指导方针(并采取其他行动)，以推进人工智能(AI)的安全、可靠和可信的开发和使用。参见2023年10月30日行政命令下的NIST职责

7.有关威胁建模的更多信息可从OWASP基金会获得

8.参见MITREATLAS对抗性机器学习101

9.GitHub:使用恶意Lambda层的Tensorflow的RCEPoC

10.SLSA:“在任何软件供应链中保护工件的完整性”

11.METI(日本经济产业省，2023)，《软件管理软件物料清单(SBOM)引入指南》

12.谷歌研究:机器学习:技术债务的高息信用卡

13.tramtre等人2016，通过预测API窃取机器学习模型

14.Boenisch,2020，针对机器学习隐私的攻击(第1部分):使用IBM-ART框架的模型反转攻击

15.国家网络安全中心，2020，设计和建造私有托管的公钥基础设施

1. 此处定义为：开发人工智能系统（或已开发人工智能系统）并将该系统投放市场或以自己的名义或商标投入使用的个人、公共当局、机构或其他团体 [↑](#footnote-ref-0)
2. 有关安全设计的更多信息，请参阅 CISA 的 [Secure by Design](https://www.cisa.gov/resources-tools/resources/secure-by-design) 网页和指南转移网络安全风险的平衡：[安全设计软件的原则和方法](https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf) [↑](#footnote-ref-1)
3. 与非 ML AI 方法（基于规则的系统等）相反 [↑](#footnote-ref-2)
4. CEPS在其出版物[《将人工智能价值链与欧盟人工智能法案相协调》](https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/)中描述了七种不同类型的人工智能开发互动 [↑](#footnote-ref-3)
5. [ISO/IEC 22989：2022（en）](https://www.iso.org/obp/ui/" \l "iso:std:iso-iec:22989:ed-1:v1:en)将其定义为“构建人工智能系统的功能元素” [↑](#footnote-ref-4)
6. NIST 的任务是制定指南（并采取其他行动），以推进人工智能 （AI） 的安全、可靠和值得信赖的开发和使用。[查看 NIST 在 2023 年 10月 30 日行政命令下的责任](https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence) [↑](#footnote-ref-5)
7. 有关威胁建模的更多信息，请咨询 [OWASP 基金会](https://owasp.org/www-community/Threat_Modeling) [↑](#footnote-ref-6)
8. 请参阅 [MITRE ATLAS 对抗性机器学习 101](https://atlas.mitre.org/resources/adversarial-ml-101/) [↑](#footnote-ref-7)
9. GitHub：[使用恶意 Lambda 层的 Tensorflow RCE PoC](https://github.com/Splinter0/tensorflow-rce) [↑](#footnote-ref-8)
10. SLSA：“[保护任何软件供应链中的工件完整性](https://slsa.dev/)” [↑](#footnote-ref-9)
11. METI（日本经济产业省，2023 年），[“软件管理软件物料清单 （SBOM） 介绍指南”](https://www.meti.go.jp/english/press/2023/0728_001.html" \t "https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development/guidelines/_self) [↑](#footnote-ref-10)
12. 谷歌研究：[机器学习：技术债务的高息信用卡](https://research.google/pubs/pub43146/" \t "https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development/guidelines/_self) [↑](#footnote-ref-11)
13. Tramèr 等人，2016 年，[通过预测 API 窃取机器学习模型](https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer) [↑](#footnote-ref-12)
14. Boenisch，2020 年，[针对机器学习隐私的攻击（第 1 部分）：使用 IBM-ART 框架进行模型反转攻击](https://franziska-boenisch.de/posts/2020/12/model-inversion/) [↑](#footnote-ref-13)
15. 国家网络安全中心，2020 年，[设计和构建私有托管的公钥基础设施](https://www.ncsc.gov.uk/collection/in-house-public-key-infrastructure) [↑](#footnote-ref-14)