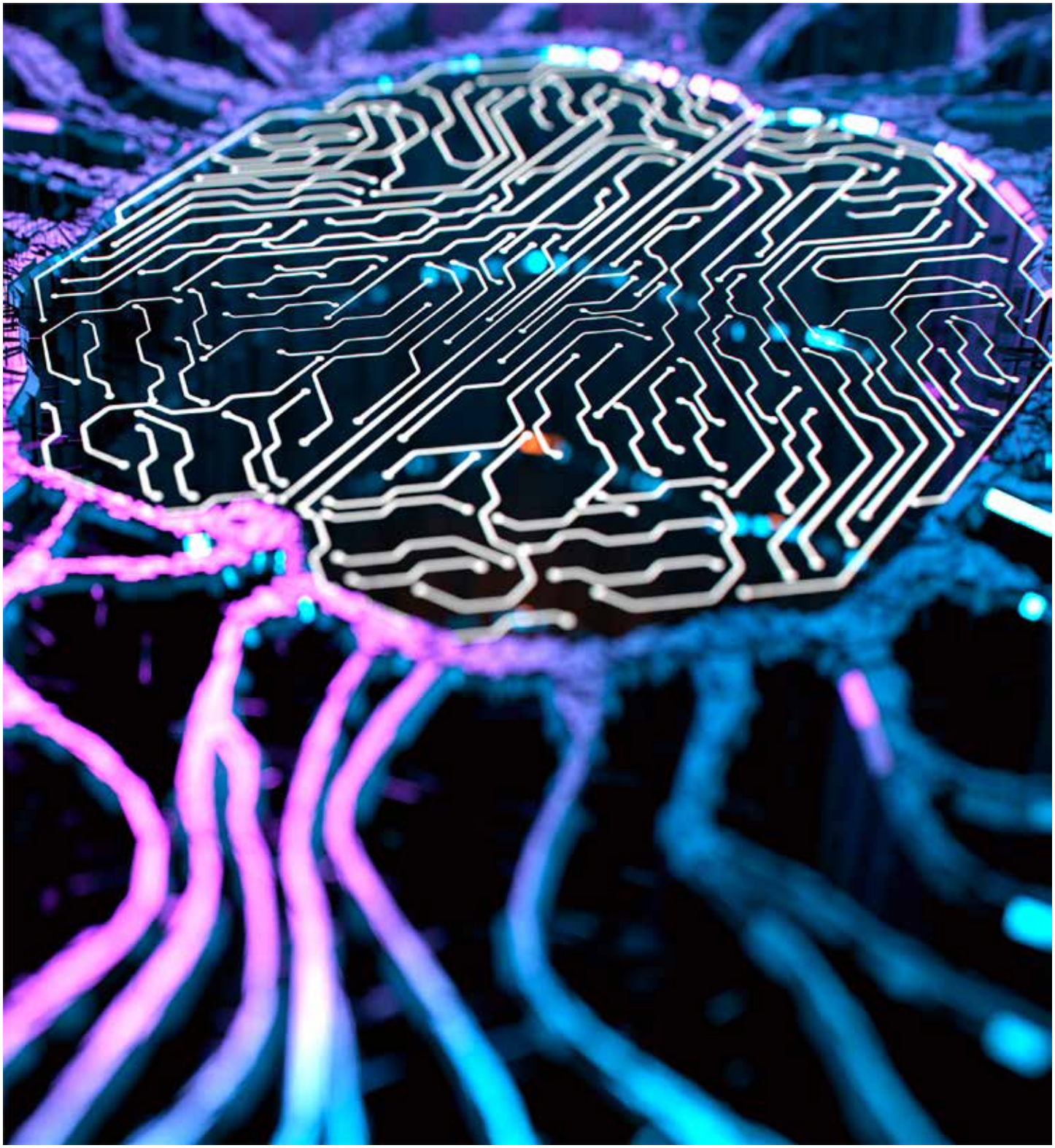
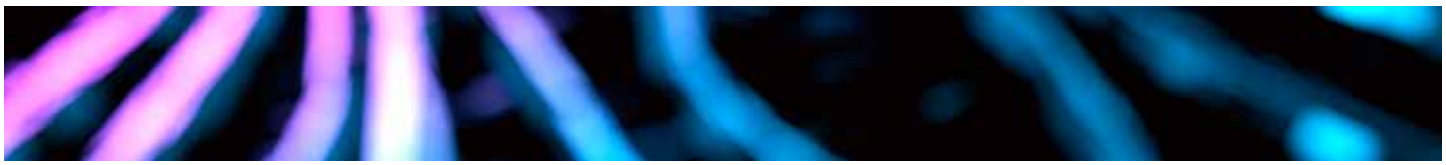
安全AI系统开发指南





National Cyber Security Centre
a part of GCHQ

CISA
CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security
Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

CSIRT
Equipo de Respuesta ante Incidentes de Seguridad Informática

National Cyber and Information Security Agency

Republic of Estonia
Information System Authority

Department of Justice
Federal Bureau of Investigation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Federal Office for Information Security

INCD
Israel National Cyber Directorate

AGENZIA PER LA CYBERSICUREZZA NAZIONALE
ACN
SAPIENTIA ADIUVAT

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

National Cyber Security Centre

NITDA

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

National Security Agency
United States of America

NASK

국가정보원
National Intelligence Service

Ministerstwo Cyfryzacji

CSA SINGAPORE
Cyber Security Agency of Singapore

关于本文档

本文由英国国家网络安全中心（NCSC）、美国网络安全和基础设施安全局（CISA）以及以下国际合作伙伴发布：

- 美国国家安全局（NSA）
- 美国联邦调查局（FBI）
- 澳大利亚信号局的澳大利亚网络安全中心（ACSC）
- 加拿大网络安全中心（CCCS）
- 新西兰国家网络安全中心（NCSC-NZ）
- 智利政府计算机安全事件响应小组（CSIRT）
- 捷克国家网络与信息安全局（NUKIB）
- 爱沙尼亚信息系统管理局（RIA）和爱沙尼亚国家网络安全中心（NCSC-EE）
- 法国网络安全局（ANSSI）
- 德国联邦信息安全办公室（BSI）
- 以色列国家网络安全指导委员会（INCD）
- 意大利国家网络安全局（ACN）
- 日本国家网络安全事件准备和战略中心（NISC）
- 日本内阁府科学技术创新政策秘书处
- 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国家研究所（NASK）
- 韩国国家情报院（NIS）
- 新加坡网络安全局（CSA）

鸣谢

以下组织为制定这些指南做出了贡献：

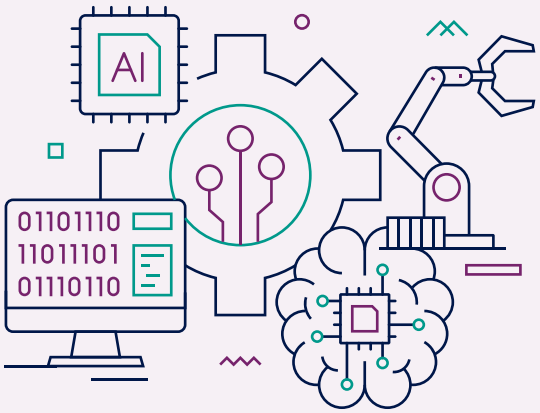
- 艾伦·图灵研究所
- Anthropic
- Databricks
- 乔治敦大学安全与新兴技术中心
- 谷歌
- 谷歌深度思维
- IBM
- ImBue
- 微软
- OpenAI
- Palantir
- 兰德公司
- Scale AI
- 卡内基梅隆大学软件工程学院
- 斯坦福大学人工智能安全中心
- 斯坦福大学地缘政治、科技和治理项目

免责声明

本文档中的信息由NCSC和编写组织“按现状”提供，除非法律要求，否则他们不对因使用本文档而造成的任何损失、伤害或损害承担责任。本文档中的信息不构成或暗示NCSC和编写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提及仅供参考，不代表我们相比于其他资源，更认可或推荐这些资源。

本文档按TLP:CLEAR级别提供（https://www.first.org/tlp/）。

目录



执行摘要

本文档为任何使用人工智能（AI）的系统的提供商提供指南，无论这些系统是从头开始创建的还是构建在其他人提供的工具和服务之上的。实施这些指南将帮助提供商构建AI系统，使其可按预期运行，在需要时可用，而且在运作时不会向未经授权的各方泄露敏感数据。

本文档主要面向那些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**利益相关方（包括数据科学家、开发人员、管理人员、决策者和风险负责人）阅读这些指南，以帮助他们就其AI系统的**设计**、**开发**、**部署**和**运营**做出明智的决策。

关于本指南

AI系统有潜力为社会带来许多好处。然而，为了充分实现AI所带来的机遇，必须以安全和负责任的方式进行开发、部署和运营。

AI系统会面临新型的安全漏洞，需要与标准的网络安全威胁一起考虑。当开发速度很快时——就如AI的情况——安全性往往是次要考虑因素。安全性不仅要在开发阶段，而且必须要在系统的整个生命周期中成为核心要求。

因此，本指南细分为AI系统开发生命周期中的四个关键领域：**安全设计**、**安全开发**、**安全部署**，以及**安全运营和维护**。对于每个部分，我们提出了一些考虑事项和缓解措施，这将有助于降低组织在AI系统开发过程中的总体风险。

1. **安全设计**
   这部分包含适用于AI系统开发生命周期的设计阶段的指南。它涵盖了理解风险和威胁建模，以及在系统和模型设计上需要考虑的特定主题和利弊权衡。

2. **安全开发**
   这部分包含适用于AI系统开发生命周期的开发阶段的指南，包括供应链安全、文档编制以及资产和技术债务管理。

3. **安全部署**
   这部分包含适用于AI系统开发生命周期的部署阶段的指南，包括保护基础设施和模型，使其免遭攻击、威胁或丢失，制定事件管理流程以及负责任的发布。

4. **安全运营和维护**
   这部分包含适用于AI系统开发生命周期的安全运营和维护阶段的指南。它提供了与系统部署后的操作特别相关的指南，包括日志记录和监控、更新管理和信息共享。

本指南遵循"默认安全"的方法，并与NCSC的Secure development and deployment guidance（安全开发和部署指南）、NIST的Secure Software Development Framework（安全软件开发框架）中阐明的实践以及CISA、NCSC和国际网络安全机构发布的"secure by design principles（设计安全原则）"密切契合。这些原则确立了以下优先事项：

- 对客户的安全结果负责
- 奉行完全的透明度和问责制
- 建立组织结构和领导力，使设计安全成为首要业务重点

简介

人工智能（AI）系统有潜力为社会带来许多好处。然而，为了充分实现AI所带来的机遇，必须以安全和负责任的方式进行开发、部署和运营。网络安全是确保AI系统的安全、韧性、隐私性、公平性、有效性和可靠性的必要前提。

然而，AI系统会面临新型的安全漏洞，需要与标准的网络安全威胁一起考虑。当开发速度很快时——就如AI的情况——安全性往往是次要考虑因素。安全性不仅要在开发阶段，而且必须要在系统的整个生命周期中成为核心要求。

本文档为任何使用AI的系统的提供商[1]提供指南，无论这些系统是从头开始创建的还是构建在其他人提供的工具和服务之上的。实施这些指南将帮助提供商构建AI系统，使其可按预期运行，在需要时可用，而且运作时不会向未经授权的各方泄露敏感数据。

这些指南应与已建立的网络安全、风险管理和事件响应最佳实践结合起来考虑。我们特别敦促提供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）以及我们所有的国际合作伙伴共同制定的“设计安全”[2]原则。这些原则确立了以下优先事项：

- 对客户的安全结果负责
- 奉行完全的透明度和问责制
- 建立组织结构和领导力，使设计安全成为首要业务重点。

遵循“设计安全”原则需要在系统的整个生命周期中投入大量资源。这意味着开发人员必须投入资源来优先考虑工具的**功能**、**机制**和**实施**，以便在系统设计的每一层以及开发生命周期的所有阶段都为客户提供保护。这样做可以防止以后进行成本高昂的重新设计，同时在短期内保护客户及其数据。

为什么AI的安全性有所不同？

本文档中，我们用“AI”来特指机器学习（ML）应用[3]。所有类型的ML都在讨论范围内。我们对ML应用的定义如下：

- 涉及软件组件（模型），使计算机能够识别数据中的模式并为其提供背景信息，而无需由人类通过明确编程来制定规则
- 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系统还会面临新型的漏洞。“对抗性机器学习”（adversarial machine learning，简称AML）一词用于描述对ML组件（包括硬件、软件、工作流程和供应链）中基本漏洞的利用。AML使攻击者能够在ML系统中引发意外行为，其中可能包括：

- 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
- 提取敏感的模型信息

实现这些效果的方法有很多，例如大型语言模型（LLM）领域的提示注入攻击，或者故意破坏训练数据或用户反馈（称为“数据投毒”）。

谁应该阅读本文档?

本文档主要面向AI系统的提供商,无论这些系统是基于组织托管的模型还是利用外部应用程序编程接口(API)。但是,我们敦促**所有**利益相关方(包括数据科学家、开发人员、管理人员、决策者和风险负责人)阅读这些指南,以帮助他们就其机器学习AI系统的**设计**、**部署**和**运营**做出明智的决策。

话虽如此,并非所有的指南都直接适用于所有组织。攻击的复杂程度和方法会根据攻击AI系统的对手而有所不同,因此应将这些指南与贵组织的用例和威胁概况一起考虑。

谁负责开发安全的AI?

现代AI供应链中通常涉及许多参与者。一个简单的方法是假定有两个实体:

- 一个是"提供商",负责数据整理、算法开发、设计、部署和维护
- 一个是"用户",提供输入并接收输出

尽管许多应用中使用"提供商-用户"这一方法,但它变得越来越不常见[4],因为提供商可能希望将第三方提供的软件、数据、模型和/或远程服务整合到自己的系统中。这些复杂的供应链使最终用户更难理解应该由谁承担保障AI安全性的责任。

用户(无论是"最终用户"还是将外部AI组件[5]整合进来的提供商)通常缺乏足够的可见性和/或专业知识,因此无法充分理解、评估或处理与他们使用的系统相关的风险。因此,根据"设计安全"原则,**AI组件提供商应对供应链下游用户的安全结果负责。**

提供商应在其模型、管道和/或系统中尽可能实施安全控制和缓解措施,对于使用设置的情况,应默认实施最安全的选项。在无法缓解风险的情况下,提供商应负责:

- 通知供应链下游的用户有关他们及(如适用)他们自己的用户所接受的风险
- 就如何安全使用该组件向他们提供建议

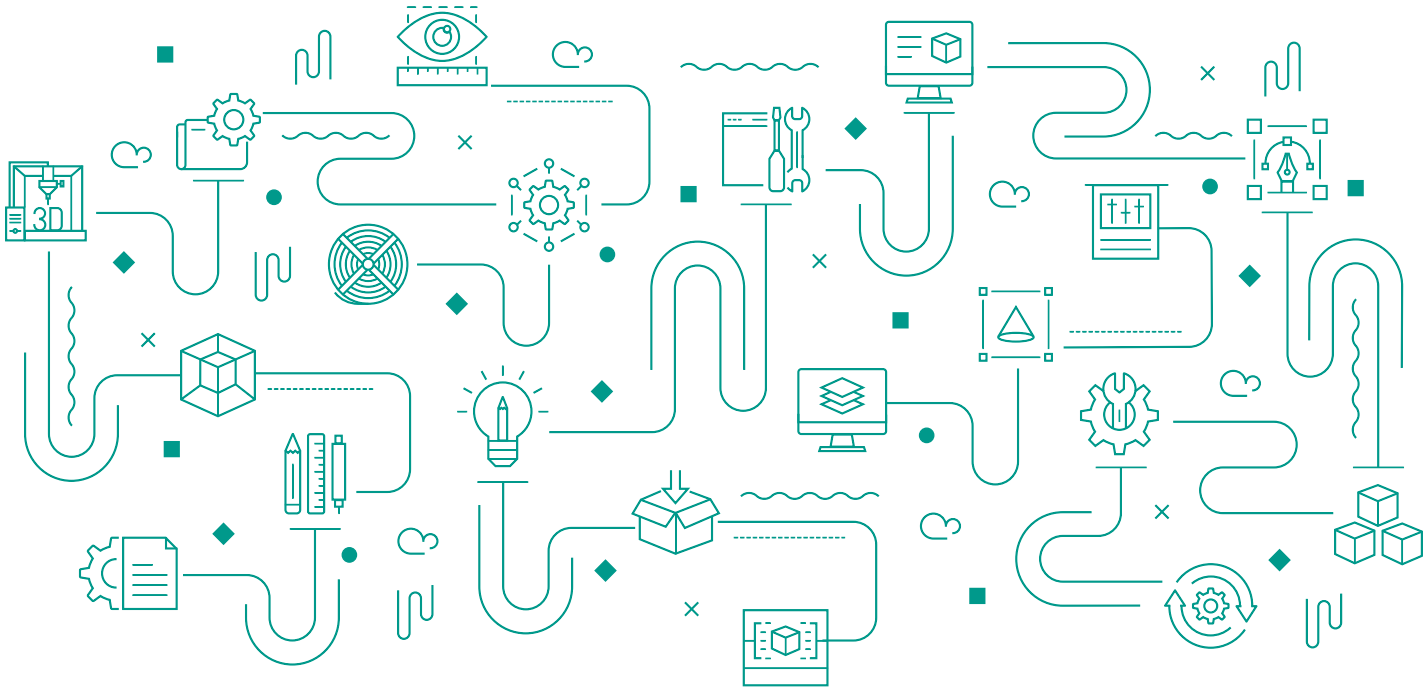
如果系统遭受攻击可能导致有形或广泛的实质或声誉损害、业务运营重大损失、敏感或机密信息泄露和/或法律影响,那么AI的网络安全风险应被视为**严重。**

安全AI系统开发指南

本指南细分为AI系统开发生命周期中的四个关键领域：**安全设计**、**安全开发**、**安全部署**，以及**安全运营和维护**。对于每个领域，我们提出了一些考虑事项和缓解措施，这将有助于降低组织在AI系统开发过程中的总体风险。

本文档中制定的指南与以下文档中阐明的软件开发生命周期实践密切契合：

- NCSC的Secure development and deployment guidance（安全开发和部署指南）
- 美国国家标准与技术研究院（NIST）的Secure Software Development Framework（安全软件开发框架）（SSDF）[6]



1.安全设计

这部分包含适用于AI系统开发生命周期的**设计**阶段的指南。它涵盖了理解风险和威胁建模，以及在系统和模型设计上需要考虑的特定主题和利弊权衡。

提高员工对威胁和风险的意识

系统所有者和高层领导人员需要了解安全的AI所面临的威胁及其缓解措施。您的数据科学家和开发人员需要对相关的安全威胁和故障模式保持认识，并帮助风险负责人做出明智的决策。您为用户提供关于AI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），并培训开发人员掌握安全编码技术以及安全和负责任的AI实践。

对系统面临的威胁进行建模

作为风险管理过程的一部分，应用一种全面的过程来评估对系统的威胁，其中包括了解在AI组件遭受攻击或表现异常的情况下，对系统、用户、组织和更广泛社会有哪些潜在影响[7]。这个过程涉及评估AI特定的威胁的影响[8]，并记录您的决策过程。

您认识到，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者目标的价值。您的评估应考虑到，随着AI系统越来越被视为高价值目标，以及AI本身会促成新的自动化攻击向量，某些威胁可能会增加。

设计系统时考虑安全性、功能性和性能

您确信当前任务最适合使用AI来解决。在确定了这一点之后，可以评估AI特定的设计选择的适当性。您需要考虑威胁模型及相关的安全缓解措施，同时兼顾功能性、用户体验、部署环境、性能、保障、监管、伦理和法律要求等方面的考虑因素。例如：

- 在选择是内部开发还是使用外部组件时，考虑供应链的安全性，例如：
  - 您做出的关于是否训练新模型、使用现有模型（无论是否进行微调）或通过外部API访问模型的选择都要适合您的要求
  - 您做出的与外部模型提供商合作的选择需要包括对该提供商自身的安全状况进行尽职调查评估
  - 如果使用外部库，您需要完成尽职调查评估（例如，确保该库具有控制措施，可防止系统加载不受信任的模型，并且不会立即面临任意代码执行的风险[9]）
  - 在导入第三方模型或序列化权重时实施扫描和隔离/沙箱，这些模型或序列化权重应被视为不受信任的第三方代码并可能启动远程代码执行

- 如果使用外部API，可以对可发送到组织无法控制的服务的数据应用适当的控制，例如要求用户在发送可能的敏感信息之前进行登录和确认
- 可以对数据和输入进行适当的检查和清理；这包括将用户反馈或持续学习数据纳入模型时，认识到训练数据限定了系统的行为

- 可以将AI软件系统的开发整合到现有的安全开发和运营最佳实践中；AI系统的所有元素尽可能在适当的环境中使用能够减少或消除已知类别漏洞的编码实践和语言进行编写

- 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，可以对可能的操作应用适当的限制（必要时包括外部AI和非AI故障保护）

- 有关用户交互的决策利用AI特定的风险的信息，例如：

  - 系统向用户提供可用的输出，同时避免向潜在攻击者透露不必要程度的详细信息
  - 如有必要，系统可为模型输出提供有效的防护措施
  - 如果向外部客户或合作者提供API，可以应用适当的控制措施，以减少通过API对AI系统进行的攻击
  - 默认将最安全的设置整合到系统中
  - 应用最小权限原则来限制对系统功能的访问
  - 向用户解释风险较高的功能，并要求用户通过选择加入来使用它们；向用户传达禁止的用例，并在可能的情况下告知用户替代解决方案

选择AI模型时考虑安全性的益处和利弊权衡

您对AI模型的选择将涉及对一系列要求进行权衡。这包括模型架构、配置、训练数据、训练算法和超参数的选择。决策应利用威胁模型的信息，并随着对AI安全性的研究进步以及对威胁的理解不断发展而定期重新评估。

选择AI模型时，您的考虑因素可能包括但不限于：

- 您使用的模型的复杂性，即所选的架构和参数数量；模型所选的架构和参数数量，还有一些其它因素，会影响它需要多少训练数据以及在使用时对输入数据变化的鲁棒性

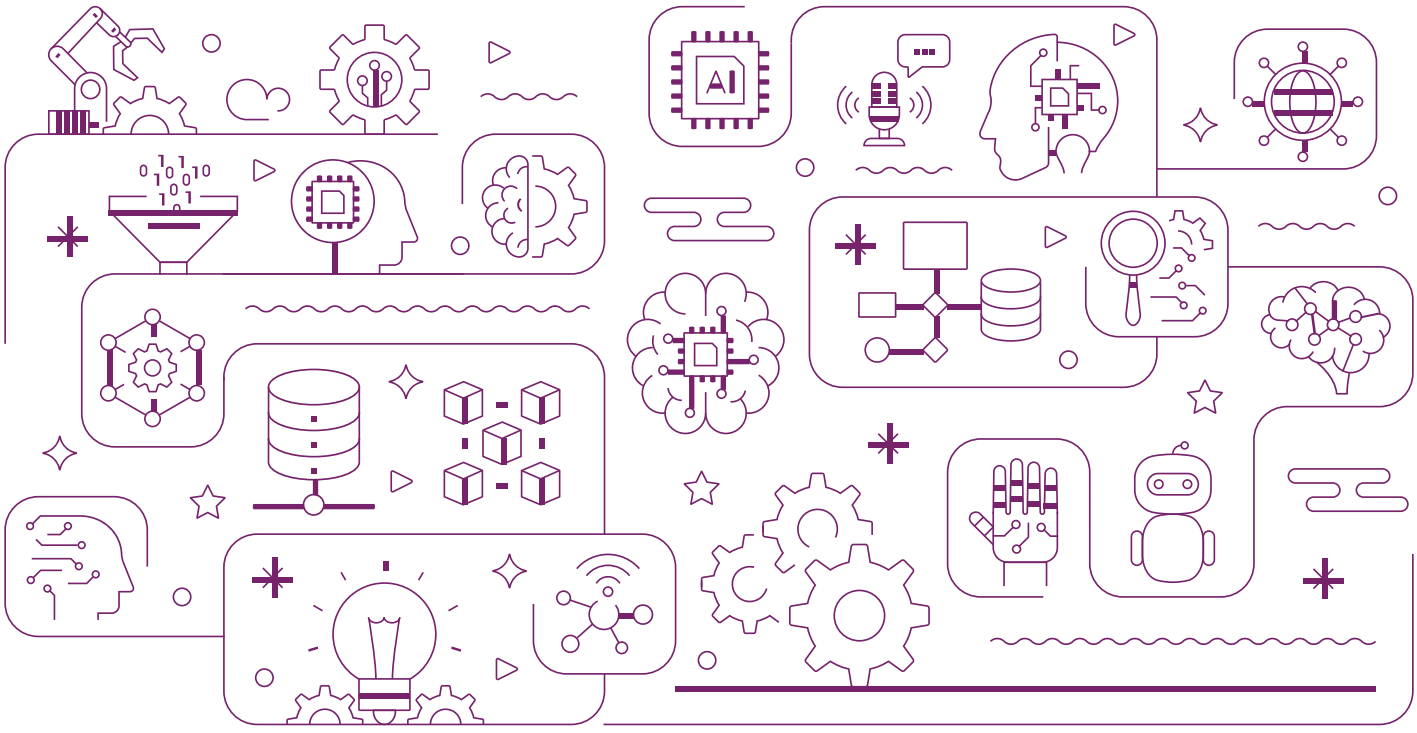
- 模型对于用例的适当性和/或使其适应特定需求的可行性（例如通过微调）

- 对模型输出进行调整、解译和解释的能力（例如用于调试、审计或监管合规性目的）；比起更难解译的大型复杂的模型，使用更简单、更透明的模型可能会带来好处

- 训练数据集的特征，包括大小、完整性、质量、敏感性、年龄、相关性和多样性

- 使用模型强化（例如对抗性训练）、正则化和/或隐私增强技术的价值
- 组件（包括模型或基础模型、培训数据和相关工具）的来源和供应链

如需了解这些因素中有多少因素会影响安全结果的更多信息，请参阅NCSC的Principles for the Security of Machine Learning（机器学习的安全性原则），特别是Design for security (model architecture)（安全设计（模型架构））章节。



2.安全开发

这部分包含适用于AI系统开发生命周期的**开发**阶段的指南，包括供应链安全、文档编制以及资产和技术债务管理。

确保供应链的安全

在系统的整个生命周期中，评估和监控AI供应链的安全性，并要求供应商遵守贵组织对其他软件实施的相同标准。如果供应商无法遵守您组织的标准，则需要按照现有的风险管理政策采取行动。

在非内部生产的情况下，您可以从经过验证的商业、开源和其他第三方开发人员那里获取并维护安全性良好且记录齐全的硬件和软件组件（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。

如果未满足安全标准，您已准备好切换到备用解决方案，以确保关键任务系统的运行。使用NCSC的Supply Chain Guidance（供应链指南）等资源和Supply Chain Levels for Software Artifacts（软件工件供应链级别，简称SLSA）[10]等框架对供应链和软件开发生命周期进行追踪确认。

识别、追踪并保护资产

您了解与AI相关的资产对贵组织的价值，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），并认识到这些资产在哪些方面属于重大投资，在哪些方面对这些资产的访问可能会遭到攻击者的攻击。您需要将日志视为敏感数据并实施控制措施来保护其机密性、完整性和可用性。

您知道您的资产位于何处，并已评估和接受与其相关的任何风险。您具备对资产进行跟踪、认证、版本控制和保护的流程和工具，并且可以在受到攻击的情况下恢复到已知的良好状态。

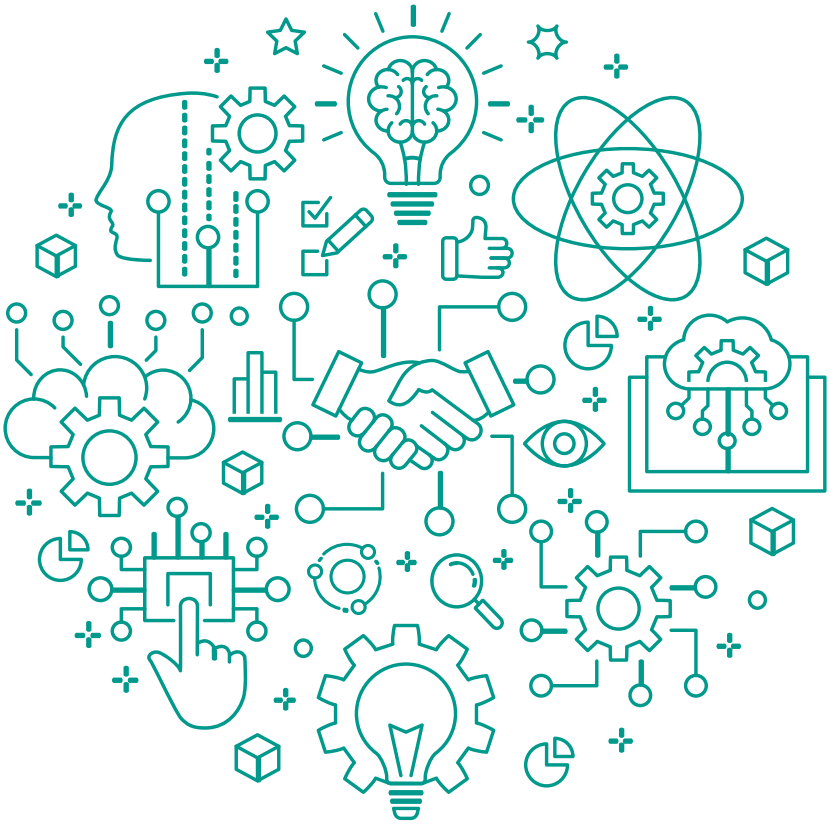
您具备适当的流程和控制措施来管理AI系统可访问的数据，并根据AI生成的内容的敏感性（以及生成它的输入信息的敏感性）来管理生成的内容。

记录数据、模型和提示

记录任何模型、数据集以及元提示或系统提示的创建、操作和生命周期管理。记录的文档需要包括与安全性相关的信息，如训练数据的来源（包括微调数据和人类或其他操作反馈）、预期范围和局限、防护措施、加密哈希或签名、保留时间、建议的审查频率和潜在的故障模式。有助于完成此任务的有用结构包括模型卡、数据卡和软件物料清单（SBOM）。全面的文档编制有助于提供透明度并支持问责制[11]。

管理技术债务

与任何软件系统一样，您可以在AI系统的整个生命周期中识别、追踪和管理“技术债务”（技术债务是指为实现短期结果而做出未达到最佳实践的工程决策，这些工程决策以牺牲长期利益为代价）。与金融债务一样，技术债务本质上并不是坏事，但应该从开发的最早期阶段就开始管理[12]。您认识到，在AI环境中这样做可能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较高。您需要确保AI系统的生命周期计划（包括停用AI系统的流程）会评估、确认和减轻未来类似系统的风险。



3.安全部署

这部分包含适用于AI系统开发生命周期的**部署**阶段的指南，包括保护基础设施和模型，使其免遭攻击、威胁或丢失，制定事件管理流程以及负责任的发布。

确保基础设施的安全

您可以对系统生命周期中每个部分所使用的基础设施都应用良好的基础设施安全原则。在研究、开发和部署阶段对API、模型和数据及其训练和处理管道应用适当的访问控制。这包括对包含敏感代码或数据的环境进行适当隔离。这也有助于缓解旨在窃取模型或损害其性能的标准网络安全攻击。

持续保护模型

攻击者可能能够通过直接访问模型（获取模型权重）或间接访问模型（通过应用程序或服务对模型进行查询）来重建模型的功能[13]或其训练数据[14]。攻击者还可能在训练期间或训练之后篡改模型、数据或提示，使输出变得不可信。

您可以通过以下方式分别保护模型和数据免遭直接和间接访问：

- 实施标准网络安全最佳实践
- 在查询界面上实施控制，以检测并阻止尝试访问、修改和外泄机密信息的行为

为确保消耗系统能够验证模型，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）的加密哈希和/或签名。对于密码学来说，良好的密钥管理始终至关重要[15]。

您对机密性风险的缓解方法很大程度上取决于用例和威胁模型。例如，涉及非常敏感数据的一些应用可能需要理论上的保证，而这可能难以实现或成本较高。如果合适，可以使用隐私增强技术（例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

影响AI系统的安全事件是不可避免的，这应该反映在您的事件响应、升级和纠正计划中。您的计划要反映不同的情境，并且需要随着系统和更广泛的研究的发展而定期重新评估。您可以将公司的关键数字资源存储在离线备份中。响应人员需要接受过培训，能够评估和处理与AI相关的事件。您可以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，以便他们启动事件响应流程，而不额外收费。

负责任地发布AI

您只有在对模型、应用程序或系统进行适当和有效的安全评估，例如基准测试和红队测试（以及这些指南范围之外的其他测试，如安全性或公平性测试）之后才能发布模型、应用程序或系统，并且需要向用户明确说明已知的局限性或潜在的故障模式。有关开源安全测试库的详细信息，请参阅本文档末尾的延伸阅读部分。

让用户更容易采取正确的行动

您认识到每个新设置或配置选项都应结合其带来的业务利益及其引入的任何安全风险进行评估。理想情况下，将最安全的设置作为唯一的选项整合到系统中。当需要配置时，默认选项应该能够广泛抵御常见威胁（即，默认安全）。应用控制措施来防止以恶意方式使用或部署系统。

为用户提供关于适当使用模型或系统的指导，其中包括强调局限性和潜在的故障模式。向用户明确说明他们负责安全的哪些方面，并对他们的数据可能在何处（以及如何）使用、访问或存储保持透明（例如，是否用于模型再训练，或是否会被员工或合作伙伴审查）。

4.安全运营和维护

这部分包含适用于AI系统开发生命周期的**安全运营和维护**阶段的指南。它提供了与系统部署后的操作特别相关的指南，包括日志记录和监控、更新管理和信息共享。

监控系统行为

测量模型和系统的输出和性能，以便您可以观察到影响安全的行为的突然变化和逐渐变化。您可以考虑到并识别潜在的入侵和攻击，以及自然数据漂移。

监控系统输入

根据隐私和数据保护要求，监控并记录系统输入（例如推理请求、查询或提示），以便在遭到攻击或滥用的情况下能够履行合规义务、开展审计、调查并实施补救措施。这可能包括明确检测分布外的输入和/或对抗性输入，包括旨在利用数据准备步骤（例如图像的裁剪和调整大小）的输入。

对更新采取设计安全的方法

默认在每个产品中包含自动更新，并使用安全、模块化的更新程序来进行分发。您的更新流程（包括测试和评估制度）需要反映这样一个事实，即：对数据、模型或提示的更改可能会导致系统行为发生变化（例如，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的API）。

汲取并分享所得到的经验教训

您可以参与信息共享社区，在行业、学术界和政府的全球生态系统中进行合作，酌情分享最佳实践。保持开放的沟通渠道，以获取组织内外有关系统安全性的反馈，包括同意安全方面的研究人员开展研究并报告漏洞。需要时，您可以将问题升级到更广泛的社区，例如针对漏洞披露发布公告，包括详细且完整的常见漏洞枚举。您可以采取行动迅速、适当地缓解和修复问题。

延伸阅读

AI的开发

Principles for the security of machine learning（机器学习的安全性原则）
NCSC关于开发、部署或运营具有ML组件的系统的详细指南。

Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software（设计安全——调整网络安全风险平衡：设计安全软件的原则和方法）
该指南由CISA、NCSC和其他机构共同编写，描述了包括AI在内的软件系统制造商应如何采取措施，将安全性纳入产品开发的设计阶段，并交付开箱即可安全使用的产品。

AI Security Concerns in a Nutshell（AI安全问题简要概述）
这份文档由德国联邦信息安全办公室（BSI）编写，介绍了可能对机器学习系统发起的攻击以及对抗这些攻击的潜在防御手段。

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems（广岛进程先进人工智能系统开发组织国际指导原则）和 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems（广岛进程先进人工智能系统开发组织国际行为准则）
这些文档是G7（七国集团）广岛AI进程的一部分，为开发最先进的AI系统的组织提供指导，包括最先进的基础模型和生成式AI系统，其目的是在全球范围内促进安全、可靠和值得信赖的AI。

AI Verify（AI验证）
新加坡的AI治理测试框架和软件工具包，根据一组国际公认的原则通过标准化测试验证AI系统的性能。

Multilayer Framework for Good Cybersecurity Practices for AI（AI良好网络安全实践的多层框架）— ENISA (europa.eu)
指导国家主管机构和AI利益相关方就他们需要遵循哪些步骤来确保其AI系统、运营和流程安全的框架。

ISO 5338:AI system life cycle processes（AI系统生命周期流程，审核中）
基于机器学习和启发式系统描述AI系统生命周期的一组流程和相关概念。

AI Cloud Service Compliance Criteria Catalogue（AI云服务合规标准目录，简称AIC4）
BSI的AI云服务合规标准目录提供了AI特定的标准，可用于评估AI服务在其生命周期内的安全性。

NIST IR 8269（草稿）A Taxonomy and Terminology of Adversarial Machine Learning（对抗性机器学习的分类和术语）
基于机器学习和启发式系统描述AI系统生命周期的一组流程和相关概念。

MITRE ATLAS
机器学习（ML）系统的对手战术、技术和案例研究的知识库，以 MITRE ATT&CK 框架为模型并链接到该框架。

An Overview of Catastrophic AI Risks（AI灾难性风险概览，2023年）
这份文档由人工智能安全中心（Center for AI Safety）编写，阐述了AI带来的风险领域。

Large Language Models: Opportunities and Risks for Industry and Authorities（大型语言模型：行业与当局的机遇与风险）
BSI为希望详细了解开发、部署和/或使用大型语言模型的机遇和风险的公司、当局和开发人员制作的文档。

帮助用户测试AI模型安全性的开源项目包括：

- Adversarial Robustness Toolbox（对抗鲁棒性工具箱）（IBM）
- CleverHans（多伦多大学）
- TextAttack（弗吉尼亚大学）
- Prompt Bench（微软）
- Counterfit（微软）
- AI Verify（AI验证）（新加坡资讯通信媒体发展局）

网络安全

CISA's Cybersecurity Performance Goals（CISA的网络安全性能目标）
所有关键基础设施实体应实施的一套通用保护措施，以有效降低已知风险和对手技术的可能性和影响。

NCSC CAF Framework（NCSC的CAF框架）
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's Supply Chain Security Framework（MITRE的供应链安全框架）
用于评估供应链内的供应商和服务提供商的框架。

风险管理

NIST AI Risk Management Framework（AI风险管理框架，简称AI RMF）
AI RMF概述了与AI独特相关的个人、组织和社会的社会技术风险管理方法。

ISO 27001: Information security, cybersecurity and privacy protection（ISO 27001：信息安全、网络安全和隐私保护）
该标准为组织提供了关于建立、实施和维护信息安全管理系统的指导。

ISO 31000:Risk management（ISO 31000：风险管理）
为组织提供组织内风险管理指南和原则的一项国际标准。

NCSC Risk Management Guidance（NCSC的风险管理指南）
该指南帮助网络安全风险从业者更好地理解和管理影响其组织的网络安全风险。

注释

1. 这里定义为开发AI系统（或由他人开发AI系统）并将该系统投放市场或以其自己的名称或商标投入使用的个人、公共当局、机构或其他团体
2. 有关设计安全的更多信息，请参阅CISA的Secure by Design（设计安全）网页和指南Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software（调整网络安全风险平衡：设计安全软件的原则和方法）
3. 与非ML的AI方法（例如基于规则的系统）相反
4. CEPS在其出版物中描述了七种不同类型的AI开发交互'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act'（将AI价值链与欧盟的《人工智能法》协调一致）
5. ISO/IEC 22989:2022(en)将其定义为"构建AI系统的功能要素"
6. NIST的任务是制定指南（并采取其他行动），以推动人工智能（AI）的安全、可靠和值得信赖的开发和使用。请参阅NIST's Responsibilities Under the October 30, 2023 Executive Order（NIST在2023年10月30日行政命令下的责任）
7. 有关威胁建模的更多信息可从 OWASP Foundation（OWASP基金会）获取
8. 请参阅MITRE ATLAS Adversarial Machine Learning 101（对抗性机器学习101）
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer（使用恶意Lambda层对Tensorflow进行远程代码执行概念验证）
10. SLSA: 'Safeguarding artifact integrity across any software supply chain'（保护任何软件供应链中的工件完整性）
11. METI（日本经济产业省，2023年），'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'（针对软件管理的软件物料清单（SBOM）引入指南）
12. 谷歌研究院：Machine Learning: The High Interest Credit Card of Technical Debt（机器学习：技术债务的高息信用卡）
13. Tramèr等人，2016年，Stealing Machine Learning Models via Prediction APIs（通过预测API窃取机器学习模型）
14. Boenisch，2020年，Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework（针对机器学习隐私的攻击（第1部分）：使用IBM-ART框架的模型反演攻击）
15. 国家网络安全中心，2020年，Design and build a privately hosted Public Key Infrastructure（设计和构建私人托管的公钥基础设施）