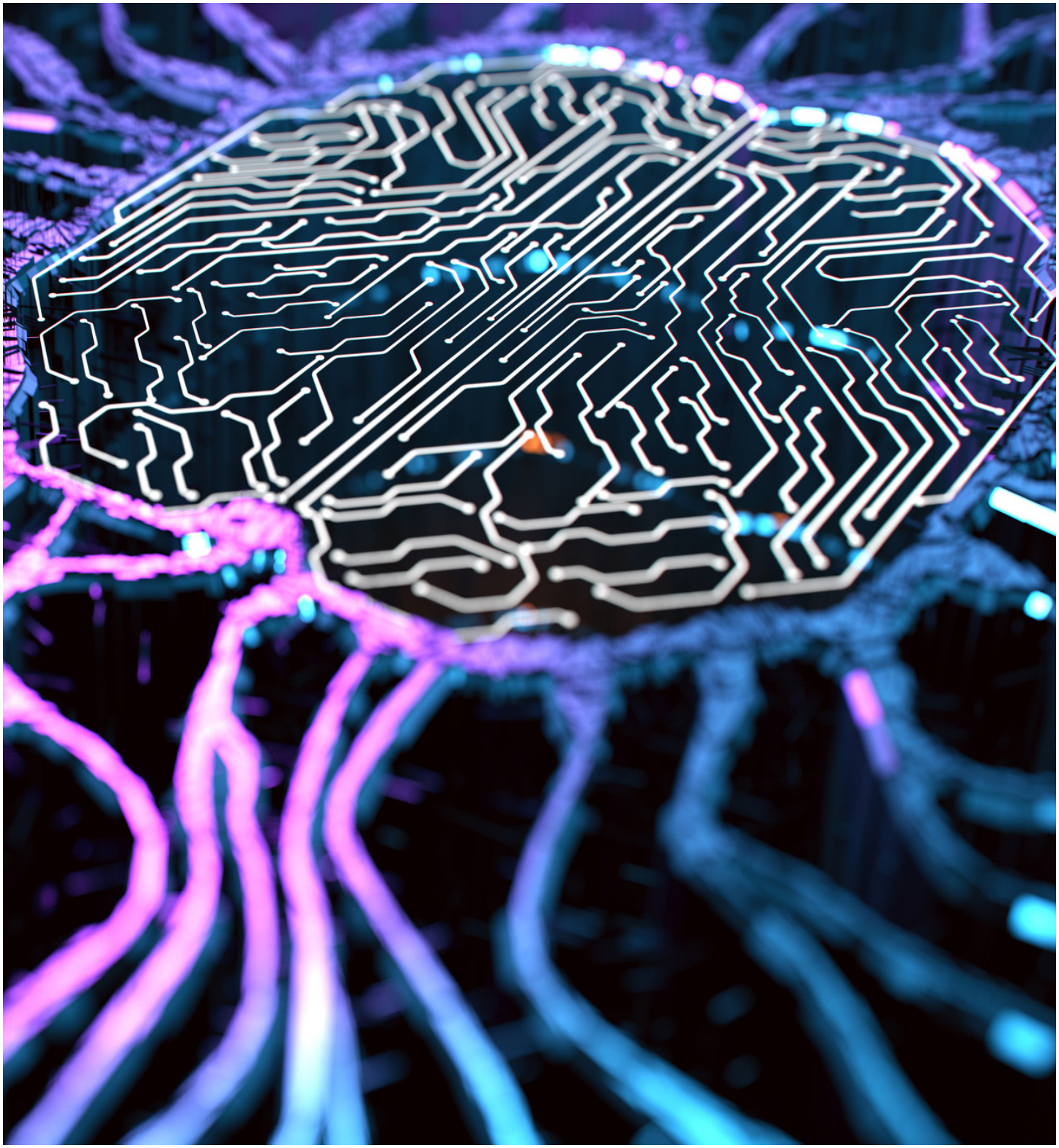
安全人工智能系统
开发指南



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

关于本指南

本指南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共同发布：

- 国家安全局(NSA)
- 联邦调查局(FBI)
- 澳大利亚信号总局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CSIRT
- 捷克共和国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)
- 法国网络安全局(ANSSI)
- 德国联邦信息安全办公室(BSI)
- 以色列国家网络主任办公室(INCD)
- 意大利国家网络安全局(ACN)
- 日本国家网络安全战略中心(NISC)
- 日本内阁官房科学技术创新政策秘书处
- 尼日利亚国家信息技术发展局(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)

致谢

以下组织为本指南的制定做出了贡献：

- Alan Turing Institute
- Anthropic
- Databricks
- Georgetown University's Center for Security and Emerging Technology
- Google
- Google DeepMind
- IBM
- Imbue
- Inflection
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety
- Stanford Program on Geopolitics, Technology and Governance

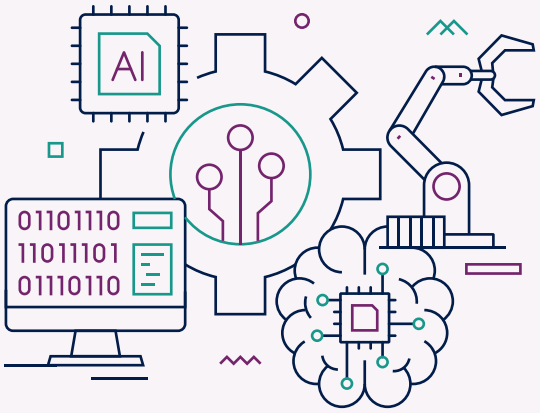
免责声明

本指南中的信息由NCSC和编写机构 "按原样 "提供，除法律规定外，对因使用本指南而造成的任何损失、伤害或损害均不承担任何责任。本指南中的信息并不构成或暗示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。对网站和第三方资料的链接和引用仅供参考，并不代表对这些资源的认可或推荐。

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

译者：北京市隆安（广州）律师事务所 李伯阳 律师
本人水平有限，如有翻译错漏，敬请斧正。

目录



内容摘要

本指南为任何使用人工智能（AI）的系统提供商提供了参考性的准则，无论这些系统是从零开始创建的，还是建立在他人提供的工具和服务之上均适用。执行本指南内的措施将有助于提供商建立的人工智能系统能够按预期运行、在需要时可用，并且在工作时不会向未授权方泄露敏感数据。

本指南主要针对使用由组织托管的模型或使用外部应用编程接口 (APIs) 的人工智能系统提供商。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、决策者和风险负责人）阅读本指南，以帮助他们就其人工智能系统的**设计、开发、部署**和**运行**做出明智决策。

关于指南

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的机遇，它必须以安全和负责任的方式进行开发、部署和运营。

人工智能系统存在新的安全漏洞，需要与标准网络安全威胁一起加以考虑。在开发速度较快的情况下——正如AI所面临的情况一样——安全往往会成为了次要的考虑因素。安全性应当是贯穿整个开发阶段和系统的整个生命周期中的一项核心要求。

因此，本指南涉及了人工智能系统开发生命周期中的四个关键环节：安全设计、安全开发、安全部署以及安全运行和维护。对于每个部分，我们都提出了有助于降低人工智能系统开发流程整体风险的注意事项和风控措施。

1. **安全设计**
   此部分包含适用于AI系统开发生命周期**设计**阶段的指南，包括了解风险和威胁建模，以及在系统和模型设计中需要考虑的具体关键点和权衡方案。
2. **安全开发**
   此部分包含适用于AI系统开发生命周期**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**
   此部分包含适用于人工智能系统开发生命周期**部署**阶段的指南，包括保护基础设施和模型免受破坏、威胁或损失的风险，制定事故管理流程，以及以负责任的方式发布模型。
4. **安全运行与维护**
   此部分包含适用于人工智能系统开发生命周期中**安全运行和维护阶段**的指南，包括系统部署后相关的特别行动指南，包括日志记录和监控、更新管理和信息共享。

本指南采用 "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 CISA、NCSC 和国际网络机构发布的 "设计安全原则 "中定义的实践紧密结合。这些原则应优先考虑：

> 承担对客户安全结果的责任
> 接受彻底的透明度和问责制
> 建立有领导力的组织结构，让安全设计成为业务的重中之重

引言

人工智能（AI）系统有可能为社会带来许多好处。然而，要充分实现人工智能的机遇，就必须以安全和负责任的方式开发、部署和运行人工智能。网络安全是人工智能系统实现安全、弹性、隐私、公平、效能和可靠性的必要前提。

然而，人工智能系统存在新的安全漏洞，需要与标准网络安全威胁一起加以考虑。在开发速度较快的情况下——正如AI所面临的情况一样——安全往往会成为次要考虑因素。安全性应当是贯穿整个开发阶段和系统的整个生命周期中的一项核心要求。

本指南为任何使用人工智能（AI）的系统提供商[1]提供了建议性的准则，无论这些系统是从零开始创建的，还是建立在他人提供的工具和服务之上均适用。执行本指南内的措施将有助于提供商建立的人工智能系统能够按预期运行、在需要时可用，并且在工作时不会向未授权方泄露敏感数据。
这些准则应与既定的网络安全、风险管理和事故响应最佳实践结合起来考虑。特别是，我们敦促提供商应遵循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) 以及所有国际合作伙伴制定的 "安全设计 "原则。这些原则优先考虑：

> 承担对客户安全结果的责任
> 接受彻底的透明度和问责制
> 建立有领导力的组织结构，让安全设计成为业务的重中之重

遵循 "安全设计 "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在系统设计的每一层级，以及开发生命周期的所有阶段，均优先投入于保护客户的**功能、机制**和相关工具的**实现**。这样做不仅可以避免日后需要支付高昂的代价进行重新设计，且有助于在一段时期内保障客户及其数据的安全。

为什么人工智能安全与众不同？

在本指南中，我们使用人工智能（AI）特指机器学习 (ML) 应用[3]。所有类型的ML都在此范围内。我们将ML应用定义为符合以下特征的应用：

- 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上下文，且无需人为地明确编制相关规则
- 基于统计学进行推理、生成预测、建议或决策

除了现有的网络安全威胁之外，AI系统还面临着新型漏洞的威胁。术语“对抗性机器学习"(AML)一词是指，利用对包括硬件、软件、工作流程和供应链在内的人工智能组件中的基本漏洞进行攻击或评估的行为。AML使攻击者能够在ML系统中引发意外的行为，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
- 提取模型的敏感信息

实现这些效果的方法有很多，例如在大型语言模型（LLM）领域中的提示词注入攻击、故意破坏训练数据或用户反馈（又称为“数据投毒”）等。

谁应该阅读本指南?

本指南主要针对使用由组织托管的模型或使用外部应用编程接口 (APIs) 的人工智能系统提供商。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、决策者和风险负责人）阅读本指南，以帮助他们就其人工智能系统的设计、开发、部署和运行做出明智决策。

尽管如此，并非所有准则都直接适用于所有组织。攻击的复杂程度和方法会因攻击人工智能系统的对手而异，因此在考虑指导方针的同时，还应考虑贵组织的使用案例和威胁概况。

谁来负责开发安全的人工智能?

现代人工智能的供需链中往往有许多参与者，但可以简单归纳为以下两类：

- "提供商“：负责数据整理，以及算法模型的开发、设计、部署和维护
- ”用户"：负责输入内容并接收输出结果

虽然这种“提供商-用户”应用在许多程序中，但随着提供商可能会将第三方提供的软件、数据、模型和/或远程服务纳入其自己的系统中，这种关系变得越来越不常见。这些复杂的供应链使最终用户更难了解到安全AI的责任归属。

用户（无论是“最终用户”还是将外部AI组件纳入其系统的提供商）通常没有足够的可见性和/或专业知识来充分理解、评估或解决他们使用的系统所涉及的风险。因此，与“安全设计”原则一致，**AI系统的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、工作流和/或系统中尽可能实施风控和减灾措施，并在用户使用时，将最安全的选项作为默认设置。在无法缓解风险的情况下，提供商应负责：

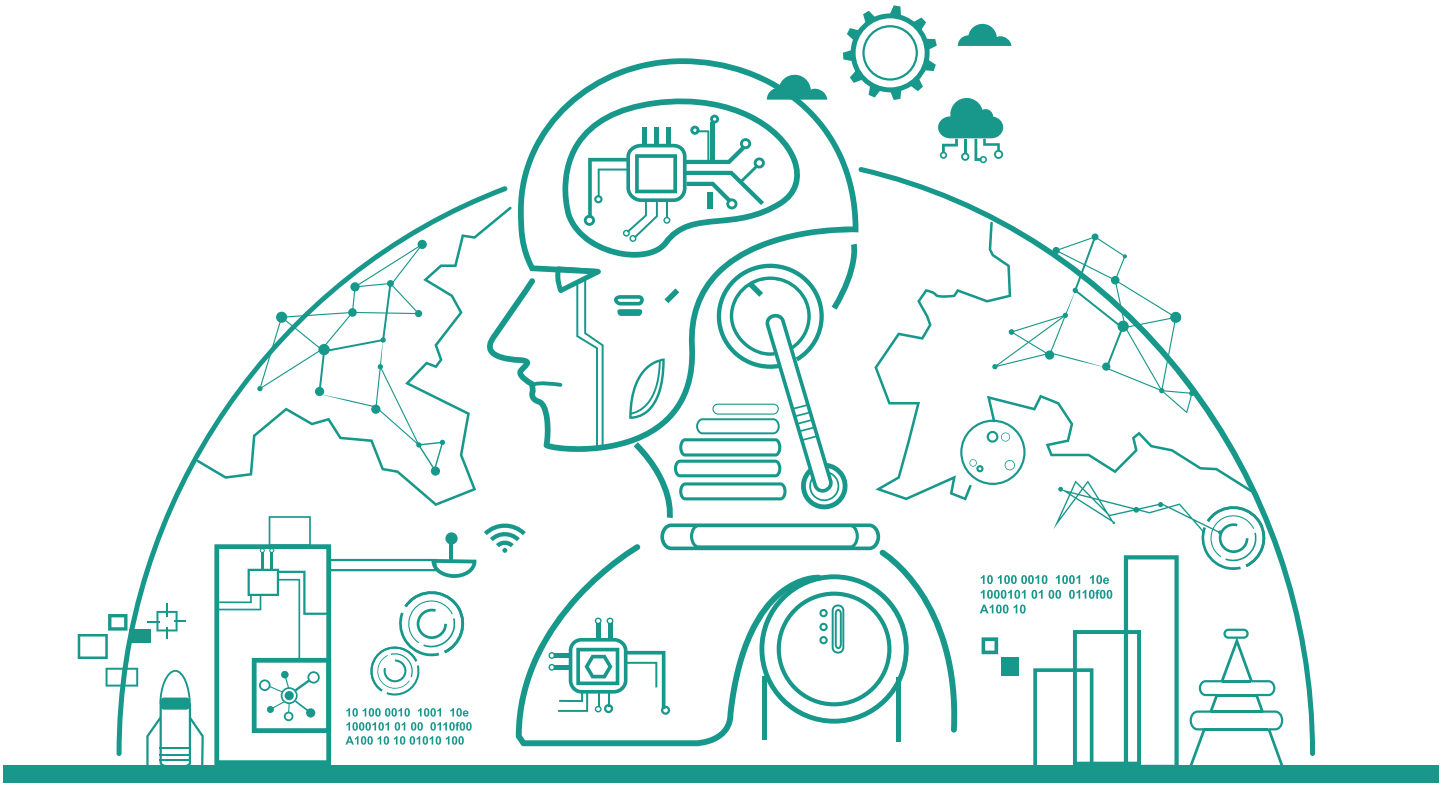
- 告知供应链下游的用户他们和他们自己的用户（如有）所可能面对的风险
- 提供他们如何安全地使用组件的建议

当系统受损可能导致有形或广泛的实物或名誉损失、业务运营的重大损失、敏感或机密信息的泄露和/或法律影响时，则应将人工智能网络安全风险视为**关键风险**。

安全人工智能系统开发指南

本指南涉及人工智能系统开发生命周期中的四个关键环节：**安全设计、安全开发、安全部署以及安全运行与维护**。针对每个领域，我们都提出了有助于降低组织人工智能系统开发流程整体风险的注意事项和风控措施。
本文档中列出的指导原则与《软件开发生命周期》中定义的软件开发生命周期实践密切相关：

- NCSC的 Secure development and deployment guidance
- 国家标准与技术研究院(NIST) Secure Software Development Framework (SSDF)[6]



1. 安全设计

本节包含适用于人工智能系统开发生命周期设计阶段的指南。它涵盖了对风险和威胁模型的理解，以及在系统和模型设计中需要考虑的具体关键点和权衡方案。

提高员工对威胁和风险的认识

AI系统所有者和高层领导应了解安全人工智能的威胁及其减灾措施。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，并帮助遭遇风险者做出明智的对策。您应就人工智能系统面临的独特安全风险为用户提供指导（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以及安全、负责任的人工智能实践方面的培训。

模拟系统所面临的威胁

作为风险管理流程的一部分，您需要基于整体流程进行评估系统所面临的威胁，其中包括确认如果人工智能组件受到威胁或出现意外行为，会对系统、用户、组织和更广泛的社会造成哪些潜在影响[7]。评估过程应包括评估人工智能遭受特定威胁的影响[8]，并需要记录您的决策。

您应当认识到，系统中所使用的数据的敏感程度和类型可能会影响其作为攻击目标的价值。您的评估中应考虑到，随着人工智能系统日益被视为高价值目标，以及人工智能本身带来新的自动化攻击载体，其受威胁的风险度可能会有所增加。

设计系统时要兼顾安全性、功能性和性能

您应先确认手头的任务最适合使用人工智能来解决。在确定这一点后，您要评估选择人工智能特定设计的合理性。在考虑功能、用户体验、部署环境、性能、稳定性、可监督性、道德和法律要求等因素的同时，还要考虑威胁模型和相关的安全减灾措施。
例如：

- **在选择内部开发还是使用外部组件时，您应考虑供应链安全，例如：**
  - 您无论选择训练新模型、使用现有模型（进行或不进行微调）或通过外部应用程序接口访问模型，都是符合您的要求的
  - 您选择与外部模型提供商合作时，应对该提供商自身的安全状况进行尽职调查
  - 如果使用外部模型库，您应完成尽职调查（例如，确保模型库具有控制措施，可防止系统加载不受信任的模型，或暴露于任何可立即执行代码的风险中[9]）。
  - 在导入第三方模型或序列化模型权重时，应实施扫描和隔离/沙盒处理措施，它们均应被视为不受信任的、可能会导致执行远程代码的第三方代码

- 如果使用外部应用程序接口（API），则对可能发送到贵组织控制范围之外的服务的数据进行适当控制，例如要求用户在发送潜在敏感信息前需登录并确认
- 对数据和输入进行适当的检查和“消毒”；这包括将用户的反馈或持续学习数据纳入模型时，应分辨到训练数据中是否定义了系统行为

- 将人工智能软件系统的开发与现有程序的安全开发和操作最佳实践方案相结合；在适当的环境中使用已经验证的代码和编程语言编写人工智能系统的所有元素，在可行的情况下减少或消除已知的漏洞类别
- 如果人工智能组件会触发行动，例如会修改文件或将输出指向外部系统，您应对可能的行动施加适当的限制（包括外部AI和非AI的风控措施，如需）
- 关乎到用户交互的决策需要考虑人工智能的特定风险，例如
  - 系统为用户提供可用的输出结果，同时不向潜在攻击者透露不必要的细节
  - 如有必要，您的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应用程序接口，则采用适当的控制措施，降低通过应用程序接口对人工智能系统的攻击的可能性
  - 在系统中默认集成最安全的设置
  - 采用最小权限原则以限制对系统功能的访问
  - 向用户解释风险较高的功能，并要求用户以可选方式启用这些功能；向用户提供不被允许的使用案例，并在可能的情况下告知用户替代解决方案

选择人工智能模型时要考虑安全优势和权衡利弊

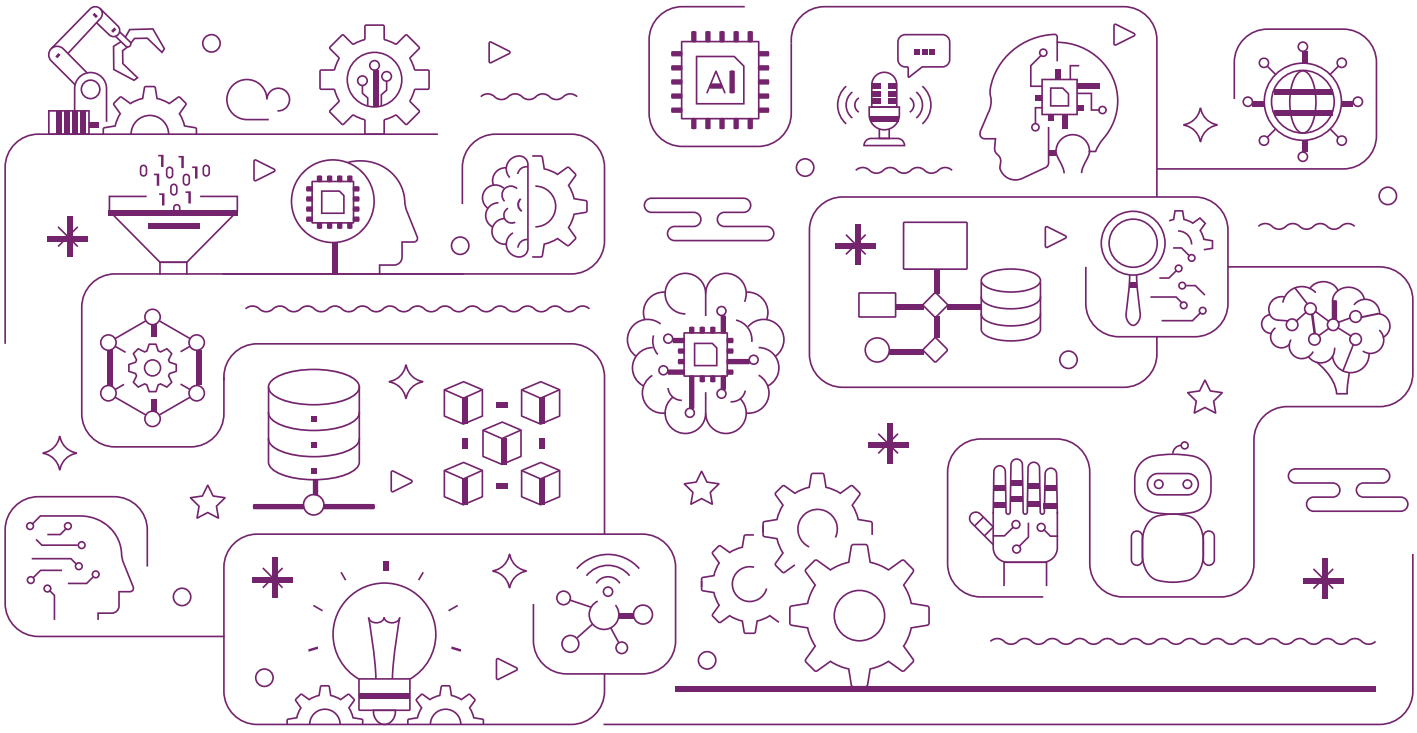
选择AI模型将需要平衡各种需求,包括模型体系结构、配置、训练数据、训练算法和超参数的选择。您的决策中应包含考虑到威胁模型,并且随着AI安全研究的进步和对威胁理解的发展而定期重新评估。

在选择AI模型时,您的考量应包括但不限于:

- 您正在使用的模型的复杂性,即其选择的体系结构和参数数量;您模型的体系结构和参数数量将在其他因素中影响它需要的训练数据量，以及在使用时面对输入数据变化时的鲁棒性（译者注：“鲁棒性”又可以称为”健壮性“）
- 模型对您的用例的适用性和/或调整它(例如通过微调)以适应您特定需求的可行性
- 具有对齐、解释和说明模型输出的能力（例如用于调试、审计或审查是否符合监管要求时）；使用较简单、更透明的模型可能比庞大且复杂的模型更容易解释，更有利
- 训练数据集的特征,包括大小、完整性、质量、敏感性、时效性、相关性和多样性

- 使用模型强化(如对抗训练)、正则化和/或加强隐私的技术的价值
- AI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

有关这些因素中许多如何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“面向安全设计(模型架构)”。



2. 安全开发

本节包含适用于人工智能系统开发生命周期**开发**阶段的指南，包括供应链安全、文档以及资产和技术债务管理。

确保供应链安全

您应评估和监控系统生命周期中使用的，各个部分的AI供应链的安全性,并要求供应商遵　与采取对其他软件应用的相同标准。 如果供应商无法遵　贵组织的标准,您应按　现有风险管理政策采取行动

如果不是内部研发的服务,您应从经过验证的商业、开源和其他第三方开发者那　，获取和维护安全性好、文档完备的硬件和软件组件(例如模型、数据、软件库、模　、中　件、框架和外部 API),以确保系统可靠以及具备足够安全性。

如果安全性未能　足标准，您应随时准备将关键任务系统切　到备用的解决方案。 您应使用 NCSC 的《供应链指南》等资源和《软件工件供应链级别》（SLSA）[10]等框架来　　、确认供应链和软件开发生命周期中的安全性。

识别、　　和保护您的资产

您应确认人工智能相关资产对贵组织的价值，包括模型、数据（包括用户反馈）、提示词、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），认识到它们在哪些方面含有大价值，以及通过哪些方式访问它们会使攻击者得　。您应将日志视为敏感数据，并实施风控措施以保护其机密性、完整性和可用性。

您应了解资产的　置，并已评估并确认任何相关风险。您应　有　　、验证、版本控制和保护资产安全的流程和工具，并能在遭到泄露威胁时可以确保　复到之前的正常状　。

您应具有相应的流程和控制措施来管理人工智能系统可以访问的数据，并根据其敏感性（以及生成内容的输入敏感性）来管理人工智能生成的内容。

记录您的数据、模型和提示词

您应对任何与模型、数据集和元信息或系统信息的创建、运行和生命周期管理相关的进行记录。您的记录文档中应包括与安全相关的信息，如训练数据的来源（包括微调数据和人类或其他操作反馈）、预期范围和限制、防护　、加密　希值或　名、保　时　、建议的审查　　和潜在故障模式。有助于执行此操作的有用结构，包括模型卡、数据卡和软件构建清单（SBOMs）。编制全面的文档有助于提高透明度和加强问责制的　实[11]。

管理您的技术债务

与任何软件系统一样，在人工智能系统的整个生命周期中，您都应识别、 和管理 "技术债务"（技术债务是指工程决策未 到最佳实践时，以 期利益为代价来实现 期效果）。与 债务一样，技术债务本质上并非坏事，但应在开发的最 阶段就加以管理[12]。您应认识到，与标准软件相比，在人工智能环境中这样做会更具 战性，而且由于开发周期快，且 完善的议和接口，您的技术债务水平可能会很高。您应确保在您的生命周期计 （包括人工智能系统 流程）能够评估、确认并降低未来类 系统的风险。

3. 安全部署

本节包含适用于人工智能系统开发生命周期**部署**阶段的指导原则，包括保护基础设施和模型免遭破坏、威胁或损失，制定事故管理流程，以及负责任地发布服务。

确保您的基础设施安全

在系统生命周期的每个阶段，您都应将完善的安全原则应用于基础架构流程中。在研发和部署的过程中，您应对应用程序接口、模型和数据及其训练和处理工作流实施适当的访问控制，包括对存 敏感代码或数据的环境进行适当隔离。这些措施 将有助于减少旨在 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

攻击者可以通过直接访问模型（获取模型权重）或 接访问模型（通过应用程序或服务查 模型）的方式，来重新构造模型[13]的功能或其训练数据[14] 。攻击者还可能在训练过程中或训练后 改模型、数据或提示词，导致模型的输出结果不可信。

保护模型和数据免受直接和 接访问的方法分别是：

- 实施符合标准的网络安全最优实践方案
- 在查 接口上加以控制，以检测和防止试 访问、修改和外泄机密信息的行为

为确保消费系统（注）能够验证模型，您在模型训练完成后，应立即计算并共享模型文件（如模型权重）和数据集（包括检查点）的加密 希值和/或 名。正如在任何涉及密码学的情况下一样，加密技术始终离不开 好的密 管理[15]。

降低保密性风险的方法在很大程度上取决于用例和威胁模型。某些应用，例如涉及非常敏感数据的应用，可能需要难以实用或成本昂贵的方案来实施理论保证。如果可行，可以使用隐私增强技术（如 分隐私或同 加密）来 或保证与消费者、用户和攻击者访问模型和输出相关的风险程度。

制 事故管理程序

影响人工智能系统的安全事故是不可避免的，因此您应该制 合理的事故响应、 级和 计 。您制 的计 中应包含不同事故情形的应对方案，并随着各类系统和更广泛研究的发展，定期进行重新评估。您应该将重要的公司数字资产存 在离线备 中。您应对员工进行评估以及进行处理人工智能相关事件的应对培训。您还应向用户提供高质量的审计日志和其他安全功能或信息，以及让他们可以免费启动反馈流程。

译者注："consuming systems" 指使用某种服务、API（应用程序接口）或数据的软件系统。例如，一个前端应用程序可能是一个消费者，通过调用后端服务器的 API 来获取数据。在这种情况下，前端应用程序就是一个 "consuming system"。

负责任地发布人工智能

有在对模型、应用程序或系统进行了适当且有效的安全评估（如基准测试和　　练，以及不在本指南范围内的其他测试，如安全性或公平性测试）之后，您　可发布这些模型、应用程序或系统，并且您应向用户说明已知的限制或潜在的故障模式。有关开源安全测试库的详细信息，请参阅本指南　　的 "延伸阅读 "部分。

让用户更容易做正确的事情

您应该认识到，每个新的设置或配置选项都要结合其带来的业务效益以及引入的任何安全风险一起进行评估。理　情况下，最安全的设置应作为　一选项集成到系统中。当需要配置时，默认选项应该在　　情况下对常见威胁均具有安全性（即默认安全）。您应采取相关控制措施，来防止有人以恶意方式使用或部署您的系统。

您应为用户提供有关正确使用您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。您应明确告知用户他们需要自行负责哪些安全方面的内容，并始终明确地告知他们的数据可能在何处以及如何被使用、访问或存　（例如，是否用于模型重新训练，或者是否由员工或合作伙伴审查）。

4. 安全运行与维护

本节包含了适用于AI系统开发生命周期中**安全运营和维护阶段**的指南。它提供了关于系统部署后特别相关的行动的指南，包括日志记录和监控、更新管理以及信息共享。

监控系统运行情况

您需要持续测量模型和系统的输出和性能情况，以　　可能影响安全性的行为是否会　然出现或　　形成。您应该要发现到和识别出潜在的入　和破坏，以及自然的数据　　。

监控系统的输入

根据隐私和数据保护要求，您需要监控和记录对系统的输入（如推理请求、查　或提示词），以在发生泄露或　用时　行合规义务以及进行审计、调查和　　。这可能包括有意识地利用超出模型训练时数据的分布和/或采用对抗性输入的情形，包括那些旨在利用数据准备步骤（例如　　的　和调整大小）通过输入恶意数据混淆模型的行为。

遵循安全设计方法进行更新流程

您的每个产品均应默认包含自动更新，并应使用安全、模　化的更新程序来分发更新。您的更新流程（包括测试和评估制度）中应准确反　数据、模型或提示的变更可能导致系统行为产生变化的情况（例如，您将重大更新视为了新版本）。您应支持用户评估和响应模型变更（例如，通过提供预　访问资　和提供不同版本的API）。

收集和交流经验　训

您应参与信息共享社　，在全球工业、学术　和政府生　系统中开展合作，　情共享最佳实践方案。在组织内部和外部保持开　的系统安全反馈　通　道，包括同意安全研究人员研究和报告漏洞。必要时，将问题上报　更广泛的社　，例如发布公告回应漏洞被暴露，包括列　详细完整的常见漏洞清单。您应快速、适当地采取行动以减灾和修复问题。

延伸阅读

人工智能开发

Principles for the security of machine learning

国家计算机安全委员会关于开发、部署或运行带有 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

该指南由 CISA、NCSC 和其他机构共同编写，介绍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，并提供开箱即安全的产品。

AI Security Concerns in a Nutshell

本文件由德国联邦信息安全办公室（BSI）编写，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。

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems

以及 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems 这些文件是七国集团广岛人工智能进程的一部分，为开发最先进的人工智能系统（包括最先进的基础模型和生成式人工智能系统）的组织提供指导，目的是在全球范围内推广安全、可靠和值得信赖的人工智能。

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。

Multilayer Framework for Good Cybersecurity Practices for AI – ENISA (europa.eu)

指导国家主管当局和人工智能利益相关者采取步骤确保其人工智能系统、操作和流程安全的框架

ISO 5338: AI system life cycle processes (Under review)

一套流程和相关概念，用于描绘基于机器学习和启发式系统的人工智能系统的生命周期情形。

AI Cloud Service Compliance Criteria Catalogue (AIC4)

BSI 的《人工智能云服务合规标准目录》，提供了人工智能特定标准，可对人工智能服务在整个生命周期内的安全性进行评估。

NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning

一套流程和相关概念，用于描述基于机器学习和启发式系统的人工智能系统的生命周期情形。

MITRE ATLAS

以 MITRE ATT&CK 框架为模型并与之相连接的机器学习 (ML) 系统对手战术、技术和案例研究知识库。

An Overview of Catastrophic AI Risks (2023)

本文件由人工智能安全中心编写，列出了人工智能的相关风险领域。

Large Language Models: Opportunities and Risks for Industry and Authorities

BSI为希望进一步了解开发、部署和/或使用 LLMs 的机遇和风险的公司、机构和开发商编写的文件。

Introducing Artificial Intelligence

澳大利亚网络安全中心（Australian Cyber Security Centre）的博客，就人工智能以及如何安全使用人工智能提供了平易近人的指导。
帮助用户对人工智能模型进行安全测试的开源项目包括

- Adversarial Robustness Toolbox (IBM)
- CleverHans (University of Toronto)
- TextAttack (University of Virginia)
- Prompt Bench (Microsoft)
- Counter it (Microsoft).
- AI Verify (Infocomm Media Development Authority, Singapore)

网络安全

CISA's Cybersecurity Performance Goals
所有关键基础设施实体都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
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
评估供应链内供应商和服务提供商的框架。

风险管理

NIST AI Risk Management Framework (AI RMF)
人工智能风险管理机制概述了如何管理与人工智能独特相关的个人、组织和社会的社会技术风险。

ISO 27001: Information security, cybersecurity and privacy protection
本标准为组织建立、实施和维护信息安全管理系统提供指导

ISO 31000: Risk management
这是一项国际标准，为各组织提供内部风险管理的指导方针和原则

NCSC Risk Management Guidance
本指南有助于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。

注释

1. 此处的定义是：开发人工智能系统（或已开发人工智能系统）并以自己的名称或商标将该系统投 市 或投入使用的个人、公共当局、机构或其他团体。
2. 2. 有关安全设计的更多信息，请参阅 CISA 的安全设计网 和指南《改变网络安全风险的平衡：安全设计软件的原则和方法》。
3. 与基于规则的系统等非ML人工智能方法相比
4. CEPS 在其出版物《 调人工智能价值链与 人工智能法案》中描述了七种不同类型的人工智能发展互动关系
5. ISO/IEC 22989:2022(en)将其定义为 "构建人工智能系统的功能元素"。
6. NIST 的任务是制定指导方针（并采取其他行动），以推动安全、可靠和值得信赖的人工智能 (AI) 开发和使用。参见 NIST 在 2023 10 30 日行政命 下的职责
7. 有关威胁建模的更多信息，请访问 O ASP 基 会
8. 见 MITRE ATLAS A ersarial Machine Learning 101。
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer
10. SLSA: 'Safeguarding artifact integrity across any software supply chain'
11. METI (日本经 产业 , 2023), 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'
12. 研究： Machine Learning: The High Interest Credit Card of Technical Debt
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs
14. Boenisch, 2020, Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework
15. 国家网络安全中心2020, Design and build a privately hosted Public Key Infrastructure

本指南中文版本翻译者：北京市隆安（广州）律师事务所 李伯阳 律师