机器翻译版本，安天内部讨论使用

安天战略情报中心
信息采编组采编专用

人工智能在网络防御中的机遇

网络安全团队如何运用人工智能



采编说明			
原文题目	Opportunities for AI in cyber defence: Use of AI by cyber security teams		
中译题目	人工智能在网络防御中的机遇：网络安全团队如何运用人工智能		
作者/机构	加拿大网络安全中心（Cyber Centre）、新西兰国家网络安全中心（NCSC-NZ）和英国国家网络安全中心（NCSC-UK）联合发布	**原文发布日期**	2026年9月14日
原文出处	https://www.cyber.gov.au/sites/default/files/2026-05/Opportunities%20for%20AI%20in%20cyber%20defence.pdf		
作者简况			
机构简况	加拿大网络安全中心（Cyber Centre）隶属于加拿大通信安全局（CSE），是该国网络安全技术主管机构，面向政府、关键基础设施运营方、企业与公众统一提供威胁情报、安全指引、事件处置与防御支撑，联动境内多方及国际伙伴，开展事前预警、应急响应与威胁溯源，持续提升加拿大整体网络韧性，是该国统筹网络安全防御与风险治理的核心技术枢纽。 新西兰国家网络安全中心（NCSC-NZ）隶属于新西兰政府通信安全局（GCSB），为新西兰国家级网络安全实操主管机构，面向政府、重点机构、关键基础设施、中小企业及普通民众提供威胁研判、攻击阻断、安全建议与事件响应，同时承担公共部门信息安全统筹、网络科普，并向太平洋岛国输出网络安全能力建设支持。 英国国家网络安全中心（NCSC-UK）隶属于英国政府通信总部（GCHQ），是英国网络安全权威技术机构，2016年成立，衔接政府、产业、学界，目标打造全球安全的线上营商与生活环境，负责网络威胁监测、重大网络事件处置、安全标准制定、网络防御项目运营，面向关键基础设施、政企及公众发布安全指引，依托前沿研究研判新兴风险，同时广泛开展跨境网络安全协作。		
采编方	安天战略情报中心信息采编组	**责编**	卢开思
翻译方式	机翻精校	**机翻平台**	龙译（安天内部平台）
免责声明	本译文为安天战略情报中心信息采编组采集发现，并进行机翻校对。 本文原文来自互联网的公共信源，基于AI平台进行翻译，并进行部分校对，力图符合所获得之电子版本原意进行翻译，但受翻译水平和技术水平所限，不能完全保证译文完全与原文含义一致，同时对所获得原文是否存在臆造、或者是否与其原始版本一致未进行可靠性验证和评价。 本译文对应原文所有观点亦不受本译文中任何打字、排版、印刷或翻译错误的影响。译者与安天不对译文及原文中包含或引用的信息的真实性、准确性、可靠性、或完整性提供任何明示或暗示的保证。译者与安天亦对原文和译文的任何内容不承担任何责任。 安天战略情报中心基于“国家安全是我们的唯一立场和第一视角”的安天企业原则展开工作，采编和翻译本文服务于支撑国家相关工作和企业发展工作需要。原文必然带有相关机构和编写者本身的立场，安天编译本文不代表认同原文表述内容、持有立场和态度。 译者与安天均与原作者与原始发布者没有联系，亦未获得相关的版权授权，鉴于译者及安天出于学习参考之目的翻译本文，而无出版、发售译文等任何商业利益意图，因此亦不对任何可能因此导致的版权问题承担责任。		



Australian Government

Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE

ACSC Australian Cyber Security Centre



Communications Security Establishment Canada

Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications Canada

Centre canadien pour la cybersécurité



National Cyber Security Centre

NEW ZEALAND



National Cyber Security Centre

a part of GCHQ

目录

引言.....4

受众.....4

网络安全形势正在不断演变.....5

将人工智能融入网络安全.....7

安全部署人工智能.....13

结论.....17

更多信息.....17

附录A：面向人工智能供应商的网络安全问题.....18

附录B：面向网络安全供应商的人工智能能力评估问题.....21

引言

人工智能（AI）正在迅速重塑网络安全格局。随着功能强大的AI日益普及，恶意行为者正利用它以更大的规模和更快的速度发起网络威胁。那些不重新评估并加强防御能力的组织，将继续面临这些由AI驱动的网络威胁。

传统上，网络安全主要依赖专业团队和被动的工作流程来管理风险。这些方法依然重要，但现代网络安全环境的规模和复杂性正日益对其构成压力。过度依赖手动流程会使风险优先级排序、潜在威胁调查以及维持一致的防御覆盖变得困难。

人工智能为网络安全防御者提供了重大机遇。如果能够安全、可靠且负责任地使用，人工智能可以：

- 加强对网络风险的优先级管理
- 提升对威胁和漏洞的检测能力
- 支持更快速的响应和恢复
- 减少对重复性手动操作的依赖

本指南由加拿大网络安全中心（Cyber Centre）、新西兰国家网络安全中心（NCSC-NZ）和英国国家网络安全中心（NCSC-UK）联合发布，旨在阐述组织如何利用人工智能（AI）加强组织网络安全，同时管理使用AI带来的风险。指南概述了网络安全形势的发展趋势，并阐述了组织如何结合《信息安全手册》（ISM）中“治理、识别、保护、检测、响应和恢复”这六大网络安全功能来应用人工智能。此外，指南还提出了安全采用人工智能的原则，并列出了网络安全防御人员为确保安全使用而应向人工智能供应商提出的关键问题。

人工监督、治理以及“安全设计”实践仍然至关重要。AI虽能显著提升网络安全水平，但无法取代坚实的网络安全基础。设计不当或治理不善的AI系统可能会引入新的攻击途径。这可能源于过度的系统访问权限、对不可信输入的依赖，或缺乏充分保障措施的自动化操作。对于缺乏健全的身份与访问管理、安全配置、补丁管理、网络分段、监控及事件响应流程的组织而言，仅靠部署AI很难实现更好的安全效果。

受众

本指南旨在为网络安全防御人员（如首席信息安全官和负责网络安全战略、运营及风险管控的高级安全负责人）提供支持。

本指南假设读者已具备网络安全概念的基础知识，并旨在支持就网络安全领域内安全、可靠且负责任地使用人工智能作出明智决策。

网络安全形势正在不断演变

随着恶意行为者越来越多地利用人工智能来提高攻击的速度、规模和复杂程度，网络安全形势正在迅速演变。为了适应这一变化，网络安全防御者可以利用人工智能来加强分析、优先级排序和防御决策。

人工智能在网络安全中的应用范围

各组织正越来越多地探索如何利用人工智能来提升网络安全成效。重点应始终放在改善组织内部的安全性上，同时需注意，仅靠识别漏洞本身并不能增强安全性。如果缺乏适当的背景信息、优先级排序和整改措施，实施不当的人工智能应用可能会带来额外风险，而非降低风险。

组织应根据其目标、风险偏好和可用技术，综合运用各类人工智能能力。这些能力涵盖从先进的前沿AI模型，到通用大型语言模型（LLM），再到现有安全工具中嵌入的人工智能功能。有关支持供应商评估的问题清单，请参阅附录A和B。

作为前沿技术，前沿AI模型能够执行多种多样的任务，体现了当今最先进模型所具备的能力。与更常见的生成式AI系统相比，前沿AI能够实现更复杂的推理、更广泛的任务覆盖，并与工具、数据及运营工作流进行更紧密的集成。使用前沿AI还需考虑成本和可持续性问题，包括代币消耗、计算需求以及持续的运营支出。组织在选择人工智能模型时应综合考量这些因素，确保模型的能力水平与预期用例的要求和价值相匹配。

在网络安全领域，前沿AI应集成到现有工具和流程中以增强其功能，而非作为独立解决方案部署。在网络安全运营中，组织可通过部署现代化且符合特定用途的安全软件来实现更安全、更有效的人工智能应用，这些软件能够限制、测试和管控AI能力，以支持具体的信息安全管理（ISM）功能。“五眼联盟”的联合声明敦促组织领导者利用AI来加强防御能力。

组织采取何种方法，将取决于其目标以及可利用的资源。分层策略使组织能够整合这些能力，在适当情况下采用更先进的人工智能，同时利用现有工具来提高运营效率。

这确保了人工智能被用于强化现有的网络安全实践，而非取代它们，从而支持整个组织安全、可靠且负责任地采用人工智能。

并非所有组织都能以相同的速度采用前沿AI技术，特别是在资源和技能有限的情况下，这会导致网络防御者之间出现能力差距。世界经济论坛指出，小型组织和公共部门组织的网络弹性较低，其中关键基础设施在应对重大事件方面的信心尤其不足。

这些挑战表明，需要加强地方层面的准备工作、建立值得信赖的合作伙伴关系并提供切实可行的支持。

智能体AI

智能体AI指的是能够独立规划、决策并采取行动以实现目标的人工智能系统，而非仅仅对单个提示做出响应。这些系统利用大型语言模型（LLMs）等先进的人工智能模型来理解环境并推演各种选项。它们能够访问各种工具、数据、记忆和工作流，从而以一定程度的自主性进行行动和运作。

与传统人工智能或聊天系统不同，即使目标定义较为宽泛，智能体AI也会积极致力于实现预期结果。它可以在有限的人工持续监督下运行，根据结果调整行为，并在某些情况下创建子任务或子智能体来完成复杂的工作。虽然人类设定目标、约束条件和权限，但智能体AI系统会随着时间的推移执行以目标为导向的行动，而非仅仅提供建议或输出结果。

如需了解更多信息，请参阅我们与美国网络安全与基础设施安全局（CISA）、美国国家安全局（NSA）、加拿大网络安全中心（CCCS）、新西兰国家网络安全中心（NCSC-NZ）以及英国国家网络安全中心（NCSC-UK）联合发布的题为《审慎采用智能体AI服务》的报告。

恶意行为者正在利用人工智能

恶意行为者越来越多地利用人工智能来加速网络攻击的开发并大规模部署。他们将人工智能模型直接嵌入恶意工作流中，以实现侦察的自动化、开发攻击工具、分析遭入侵的数据以及生成量身定制的恶意输出。这加速了漏洞的发现，并缩短了从发现漏洞到利用漏洞之间的时间，导致防御方获得预警和响应的时间更少。

人工智能还降低了技术门槛。技术水平较低的恶意行为者可以利用人工智能执行以往需要专业知识才能完成的活动，例如制作具有隐蔽性的恶意软件、进行大规模数据分析或实施具有说服力的社会工程攻击。

随着人工智能能力的普及，恶意行为者将持续提高恶意网络行动的速度、规模和影响。为了保持领先地位，组织应通过强化网络安全基础来增强防御能力。这些措施包括：最大限度地缩小攻击面、及时修补系统漏洞、实施分层深度防御架构、提高自动化水平以及改进威胁检测能力。如果防御措施未能相应提升，基于人工智能的威胁很可能削弱传统安全方法的有效性。

如需了解更多信息，请参阅英国国家网络安全中心（UK NCSC）发布的《从现在到2027年人工智能对网络威胁的影响》评估报告。

将人工智能融入网络安全

在网络安全领域应用人工智能应符合《信息安全管理（ISM）》及其相关控制措施。ISM将网络安全原则归纳为六大功能：治理、识别、保护、检测、响应和恢复。本节概述了组织如何在每项功能中应用人工智能以支持网络安全，如图1所示。

许多人工智能用例涉及处理敏感的运营信息。组织在应用人工智能时，应按照其安全和数据处理义务，保持适当的控制措施并保护信息。这包括实施最小权限访问原则、限制向人工智能系统暴露的数据范围、在使用前验证输出结果，以及保持人工智能辅助操作的可审计性。

组织还应持续评估成本、效益和风险，并认识到人工智能能力以及恶意行为者对人工智能的利用都将不断演变。



图1. 在六个网络安全功能领域应用人工智能

治理

建立并维护强大且具有韧性的网络安全文化

“治理”功能侧重于整个组织范围内如何指导、理解和管理网络安全及网络风险。它确立了明确的问责制、决策结构和监督机制，将高管层与技术团队紧密连接起来。随着威胁的演变，有效的治理有助于及时审查和调整政策、控制措施及保障安排，以维持组织的韧性。

还应加强针对人工智能的威胁情报共享，利用澳大利亚安全局（ASD）的澳大利亚网络安全中心（ACSC）（包括其合作伙伴计划：https://www.cyber.gov.au/partnershipprogram）、国家网络威胁通报系统（NCTNS）以及美国网络安全与基础设施安全局（CISA）的《人工智能网络安全协作指南》，以支持各合作伙伴之间及时、可信的合作。

随着组织在网络安全领域采用人工智能，治理机制应明确界定对其使用、监督及人类责任的预期。这包括应用“有限自主权”机制，将人工智能限制在既定范围内，特别是在关键系统和安全运营中。此外，还需确保可恢复性，以便能够安全地限制或恢复人工智能功能，从而在系统故障或遭到入侵时保障运营的连续性。

人工智能可协助组织：

- 识别不同业务部门、系统或项目间风险评估中的不一致之处。
- 分析供应链风险，包括软件依赖关系、漏洞暴露情况以及供应商的安全实践。
- 支持编制和使用各类清单，例如软件物料清单（SBOM）和加密物料清单(CBOM)。
- 加强政策解读与合规性，例如利用内部人工智能助手，根据组织政策和监管要求提供基于上下文的指导。
- 根据风险评估结果，对网络安全决策进行优先级排序。

识别

识别资产及相关安全风险

“识别”功能侧重于了解环境中存在哪些资产以及与之相关的安全风险。它通过建立对系统、软件、数据和配置的可见性，支持基于风险的决策。

人工智能可协助组织：

- 利用网络遥测技术识别环境中的未受监控或隐藏资产，从而增强资产发现能力。
- 综合考虑严重性评级、漏洞利用可能性、威胁情报以及漏洞利用对运营的影响等多重因素，对补丁修复决策进行优先级排序。
- 通过将多个低严重性漏洞串联成攻击路径，识别并评估漏洞链。
- 识别不安全的配置并推荐修复措施。
- 审查日志样本，识别攻击者可能在不触发警报的情况下采取的行动，并协助优化检测逻辑。
- 分析软件物料清单（SBOM）和配置物料清单（CBOM），对所有软件组件进行映射，以识别过时的库、隐藏风险和供应链风险。
- 对AI训练和微调数据源进行映射，以评估数据来源、保留期限及数据中毒风险。

红队场景：基于人工智能的攻击路径分析

某组织的红队使用合适的人工智能模型来分析其企业环境，大规模采集系统架构、身份关系和漏洞数据。该模型被配置了相对宽泛的权限，使其能够跨网络、应用程序和访问控制层关联分析结果。

随着时间的推移，AI能够识别出多个低、中严重级别的弱点如何组合成具有实质意义的攻击路径——这些路径若由人工手动检测，往往既困难又耗时。例如，它将一个无害的配置错误、过度的权限以及一个轻微的软件缺陷关联起来，形成一个可能导致权限提升和横向移动的序列。

通过持续生成和优化这些场景，红队利用人工智能揭示出那些在孤立分析问题时难以察觉的高影响链。这不仅提高了测试覆盖率，还帮助组织根据实际攻击的可行性（而非单个漏洞的严重程度）来确定修复优先级。

有关软件物料清单（SBOM）的更多信息，请参阅美国网络安全与基础设施安全局发布的《网络安全软件物料清单（SBOM）的共同愿景》。

保护

实施和维护控制措施以管理安全风险

“保护”功能侧重于实施和维护安全防护措施，以降低网络安全事件发生的概率及其影响。这包括维护安全的配置、管理身份和访问权限、缩小攻击面，并确保在系统和威胁发生变化时，各项控制措施仍能有效发挥作用。

人工智能可协助组织：

- 根据可利用性和环境背景对加固措施进行优先级排序，以缩小攻击面。
- 通过评估复杂环境、信任边界和数据流来模拟真实的攻击路径并检测弱点，从而增强安全架构。
- 分析用户和智能体的身份、角色及行为，以识别违反最小权限原则的情况，检测过高或非预期的权限，并识别权限蔓延或孤立账户。
- 分析实时流量模式以识别潜在安全威胁（包括异常或由自主人工智能驱动的活动），并针对网络分段规则提出改进建议。
- 扫描源代码、基础设施即代码（IaC）以及管道定义，以发现基于模式的扫描器无法察觉的漏洞和业务逻辑缺陷。
- 零信任和现代可防御架构应扩展至AI智能体以及人类用户。这包括用动态令牌取代静态凭证，并对智能体可访问和执行的内容实施明确的权限边界。

代码审查场景：AI识别隐藏漏洞

某组织部署了一款AI工具来审查源代码，该工具会在开发人员提交更改时自动分析源代码。在一次例行更新中，一名开发人员引入了一项功能，该功能处理用户输入并将数据存储到数据库中。虽然代码运行符合预期，但AI发现该功能在存储前并未对输入数据进行正确的验证。

尽管该问题看似微不足道，且由于语法正确且未违反任何明显规则，很可能通过传统检查，但人工智能识别出了其更广泛的安全影响。它指出，恶意行为者可能利用这一漏洞插入恶意数据，从而导致系统遭到入侵。

通过及早标记该问题，该组织能够在发布前修正验证逻辑。这降低了将可被利用的漏洞引入生产环境的风险，并在不延缓开发周期的前提下增强了整体安全性。

有关人工智能安全评估的更多信息，请参阅前沿模型论坛的技术报告：《前沿AI框架中的高级网络风险管理》，该报告概述了业界关于前沿AI模型如何支持网络风险管理（包括识别漏洞和不安全的默认设置）的新兴观点。

检测

检测并分析网络安全事件，以识别网络安全事故

“检测”功能侧重于识别和分析网络安全事件，以便及时发现潜在事故。该功能支持对系统、网络和身份的持续监控，以发现异常活动和新出现的威胁。

人工智能可协助组织：

- 检测网络安全事件和事故，并结合分析师的判断来验证检测结果。
- 利用MITRE ATLAS™、OWASP大语言模型（LLM）应用十大风险、OWASP智能体应用十大风险、NISTAIRMF生成式人工智能配置文件以及MAESTRO威胁模型等框架，为针对人工智能系统的恶意行为提供人工智能辅助检测依据。
- 分析网络遥测数据（例如流量、日志、域名系统和应用程序接口调用），以检测异常行为，例如服务之间的意外通信或与已知恶意端点的连接。
- 通过AI特定的遥测数据（例如模型输入（包括提示词操纵）、决策轨迹、策略检查和置信度信号）检测AI滥用行为，同时保护检测日志的完整性免受篡改。
- 分析行为和配置上下文，以区分合法活动与可疑活动，从而最大限度地减少误报和漏报。
- 建立身份、特权访问和远程访问等高风险活动的基线，以辅助异常检测。

安全运营中心平台场景：基于人工智能的威胁检测与分级处理

某组织拥有一个安全运营中心（SOC）平台，该平台利用人工智能辅助检测，对身份、端点、网络和云遥测数据中的警报进行分级处理。为了帮助分析师应对海量警报，该平台会对域名系统（DNS）、应用程序接口（API）交易以及网络流量之间的数据进行关联分析。此外，它还会追踪特权访问的模式，并将优先级较高的事件呈现给分析师进行审查。

随着时间的推移，该组织对平台抑制误报并突出显示最重要警报的能力越来越有信心。人工智能不断优化其对正常和异常行为的理解，从而提高整个环境的检测准确性。

在此场景中，SOC 平台使分析师能够更快地识别并应对最关键的威胁，在保持强大安全可视性的同时，缩短响应时间并缓解运营负担。

响应

应对网络安全事件

“响应”功能侧重于在网络安全事件发生期间采取有效、及时且协调一致的行动。它确保组织能够在维持关键运营的同时，遏制和缓解事件影响。

人工智能可协助组织：

- 通过将警报、日志和取证数据关联起来，形成对事件的连贯解释，从而协助分析师。
- 结合组织特定的系统、架构和风险环境，解读警报以及观察到的系统或用户行为。
- 在调查过程中减少对人工搜索的依赖。
- 跨身份、终端和网络控制措施协调响应行动的顺序。
- 起草事件更新报告，通过符合既定的沟通预期，在响应人员与高管之间架起沟通桥梁；发布前由响应人员验证准确性。
- 通过自动化分级处理，并在处理多起并发事件时并行准备或启动预先批准且可逆的响应措施，从而增强应急处理能力；对于高影响的遏制、中断、恢复或沟通决策，仍需人工审批，这在人工智能驱动的攻击场景中尤为关键。
- 针对不确定情况，根据组织的事件响应手册提出遏制和补救措施，供人工事件响应人员审核。

恢复

在网络安全事件发生后恢复正常业务运营

“恢复”功能侧重于在网络安全事件发生后恢复系统和服务，以便安全地恢复正常业务运营。该功能基于已验证的系统完整性和可接受的残余风险，对恢复活动进行优先级排序和控制。

人工智能可协助组织：

- 分析重建和恢复路径，以支持恢复规划、顺序安排和保障活动。
- 触发自动恢复操作和应急预案，以修复受影响的资产并使其恢复到安全、可信的状态；对于系统恢复或数据回滚等具有破坏性或不可逆的变更，需经人工批准。
- 根据已知基线验证系统和服务恢复情况，以确认其完整性后再恢复运行。
- 当怀疑存在安全漏洞、数据中毒或意外模型漂移时，将AI模型回滚至先前版本。
- 在重新启用AI增强或自主功能之前，通过针对AI的专用验证检查，核实AI模型及相关数据的完整性与预期行为。
- 通过识别互联系统中潜在的连锁故障并降低其发生风险，实现关键业务服务的快速、安全恢复。
- 针对超出预定义自动化或安全编排、自动化与响应（SOAR）手册范围的新型或复杂事件类型，生成并评估恢复流程。
- 在实际事件中遇到之前，主动识别恢复方案、依赖关系或假设中的薄弱环节。

安全部署人工智能

在采用人工智能时，组织应优先考虑实施严格的安全措施。尽管没有任何缓解策略能提供完全的保护，但组织仍应实施符合澳大利亚安全局（ASD）《信息安全管理框架》（ISM）和“八大关键措施”（Essential Eight）的强有力网络安全基准，以实质性地降低网络安全风险。

随着人工智能能力及网络威胁环境的演变，组织应定期审查和更新其控制措施，以确保人工智能继续以安全、可靠、负责任的方式使用，并符合组织的风险容忍度。鉴于关键基础设施（CI）面临的风险和影响更大，应特别注重维持具有韧性且持续受保障的控制措施。

数据准备是采用人工智能的核心先决条件。人工智能支持的网络防御的可靠性取决于其所依赖的基础数据，包括资产清单、日志覆盖范围、身份和特权数据、漏洞记录以及配置数据。组织应在采用人工智能之前及过程中，评估并提升此类数据的质量、覆盖范围和时效性，同时认识到数据中的缺口或不一致之处将直接限制人工智能生成的洞察的准确性和价值。

各组织还应确保用于网络安全的人工智能系统以及投入实际运营的人工智能系统，能够免受提示词注入、模型规避和模型提取等对抗性技术的影响，并遵循《安全人工智能系统开发指南》、美国国家标准与技术研究院（NIST）《对抗性机器学习：攻击与缓解措施的分类与术语》（NISTAI100-2 E2025）以及MITRE ATLAS™的要求，确保用于网络安全及运营领域的人工智能系统免受此类对抗性技术（如提示注入、模型规避和模型提取）的侵害。

人工监督

人工智能应辅助网络安全防御人员，而非取代人类的判断，特别是在决策可能影响后果严重的网络或安全环境的情况下。随着人工智能系统能力不断提升并开始提出建议或采取行动，强有力的人工监督至关重要。

人工智能的使用会带来风险，若管理不当，这些风险可能会降低可靠性。这些风险包括产生虚构或误导性的输出结果、对抗性操纵（如提示注入或模型规避）、过度依赖人工智能，以及与规模、可用性或成本相关的监督疲劳和运营压力。

尽管人工智能能显著提高速度和规模，但组织仍应根据证据和上下文对人工智能的输出结果进行验证。持续的监控和人类监督有助于确保人工智能驱动的加速能够提高准确性和可靠性，而非引入新的错误。

组织应：

- 确保人工智能支持而非取代人类对网络安全结果的责任。
- 明确界定人工智能驱动的自动化范围，并要求对具有重大安全、运营或安全影响的操作（特别是对访问控制、安全策略、关键系统和服务所做的变更）进行人工审查和批准，同时确保高后果操作的责任仍由授权人员承担。
- 将自主操作限制在范围狭窄、经预先批准且可逆的操作范围内。

- 在采取行动前，根据证据和上下文核查AI生成的结果，特别注意识别幻觉或误导性输出。
- 在运营技术（OT）环境中，对于影响安全、可用性、分段、恢复或流程完整性的AI生成的建议，应在实施前由具备资质的OT、工程和安全人员进行验证。
- 持续监控和审查AI的性能和行为。
- 将AI相关的故障、安全漏洞和幻觉纳入跨IT环境的事件响应计划中。

智能体AI的使用会带来额外风险，因为这些系统可以在减少人工干预的情况下发起并执行操作。组织应审慎采用智能体AI服务，确保其使用受到严格管控、明确限制，并得到强有力的人工监督支持。

系统保护与沙箱隔离

使用人工智能工具进行网络安全防护的组织，应确保在设计和部署这些工具时，即使发生故障、误用或人为操纵，也不会直接对系统、数据或运营造成损害。无论人工智能系统以何种方式或集成在何处，都应通过控制措施限制其访问、影响和执行的范围。

组织应：

- 默认以安全方式部署人工智能工具，采用保守的配置并赋予最低权限。
- 通过沙箱隔离、范围限制和执行限制来约束人工智能的行为和权限。
- 应用技术防护措施，防止人工智能系统在缺乏适当控制的情况下执行高影响或不可逆的操作。
- 确保人工智能辅助操作可追溯且可审查，以支持监督、调查及在必要时进行回滚。
- 保护人工智能组件和接口，防止其被滥用、篡改或未经授权访问。

安全系统集成

将人工智能工具集成到现有系统、工作流和运营流程中的组织，应确保在提升安全性的同时不削弱组织控制力。应将人工智能能力作为企业环境的一部分进行管理，其架构、安全及治理标准应与其他系统保持一致。

组织应：

- 采用经批准的架构集成人工智能工具，并符合企业安全和系统设计标准。
- 使用受控且可审计的集成方法，例如专为安全设计的平台或安全的API集成。
- 避免治理不善的集成，特别是在人工智能系统可在运营环境或高可用性环境中直接触发操作的情况下。
- 确保人工智能辅助产生的输出结果和操作在现有工作流及决策过程中保持可见。
- 使人工智能集成与运营流程保持一致，包括变更管理、监控和事件响应。
- 在将日志和数据集导入其他人工智能系统时实施防护措施，以管控提示注入和不可信输入的风险。

治理

人工智能的采用应有计划、经授权并具备问责机制，特别是在涉及敏感信息或关键系统时。虽然人工智能可以辅助分析和决策，但对结果的责任仍应由人类网络安全防御人员承担。

组织应：

- 记录并设计人工智能用例，确保其与业务目标、网络安全优先事项及组织风险承受能力保持一致。
- 将访问和修改权限限制在授权用户账户范围内，并将管理员权限限定在经过核实的人员范围内。
- 制定并执行相关政策，明确人工智能工具可访问哪些数据、如何使用数据以及输出结果可传输至何处。
- 防止向未经批准的人工智能工具（包括未受管理或面向消费者的AI服务）共享敏感信息。
- 确保透明度和可审计性，包括能够解释输入数据、授权依据以及输出结果如何影响决策。
- 将人工智能监督纳入现有的治理流程，包括风险接受、变更管理、保障和事件响应。

供应链

采用人工智能系统会带来独特的供应链风险，如果管理不当，这些风险可能会威胁到组织的网络安全。虽然使用预训练模型、第三方数据集和外部服务可以加快能力交付，但也可能引入固有的安全隐患、隐藏的依赖关系以及对系统行为的可见性降低。

组织应：

- 通过人工智能物料清单（AIBOM）、软件物料清单（SBOM）或同等文档，确保对人工智能供应链各环节的可见性，包括模型、数据源、第三方服务及托管安排。
- 评估人工智能特有的供应链风险，包括数据敏感性、组件控制以及对抗性机器学习威胁。
- 将人工智能供应链风险纳入现有的网络安全风险管理和事件响应流程。
- 对人工智能供应商和供货商进行尽职调查，并明确安全要求，包括漏洞和事件报告机制。
- 了解组织数据在人工智能供应链中的访问、使用和处理方式，以确保符合数据主权要求，并确定最适合的AI解决方案类型。
- 提升组织对人工智能供应链风险的认知，并确保参与AI开发、部署和运维的工作人员了解相关威胁及安全实践。

有关人工智能供应链风险及缓解措施的更多信息，请参阅：

- 《人工智能与机器学习：供应链风险及缓解措施》
- 《七国集团（G7）网络安全工作组关于人工智能软件物料清单的共同愿景》
- 美国国家标准与技术研究院（NIST）的《保障人工智能生态系统安全：人工智能物料清单（AIBOM）在缓解软件供应链风险中的关键作用》

人工智能工具的测试与保障

组织应测试用于网络安全的人工智能工具，以确认其按预期运行，并能产生准确、可靠且在实践中实用的输出结果。供应商的宣传、演示和有限的测试不足以证明该工具在真实环境中能表现良好。

测试应侧重于该工具在组织自身环境中的表现。随着系统、数据和威胁环境的变化，组织需要持续监测该工具

是否仍保持有效性。如果输出结果变得不可靠，或者该工具不再满足运营需求，组织需要重新评估该工具的作用，并做出必要的调整。

组织应：

- 在投入生产使用前，在自身环境中测试人工智能工具。
- 验证人工智能输出的准确性、相关性，并确保其与观察到的证据和上下文相符。
- 评估人工智能建议是否可操作，并能支持员工做出有把握的决策。
- 评估人工智能工具在现实条件下的表现，包括噪声数据、信息不完整和恶意活动。
- 测试当AI工具发生故障、遭到入侵或产生不可靠输出时，网络防御活动如何持续进行，包括备用方案和升级处理机制。
- 持续监控性能，以识别随着数据和威胁环境的变化而出现的性能下降、偏差或准确性变化。
- 当AI工具的输出结果变得不可靠或不再满足运营需求时，审查其作用和配置。

按需安全的人工智能

组织在选择、实施和运营人工智能系统时，应从一开始就将安全性融入其中，而非事后补救。这就是“按需安全”原则，该原则支持“安全设计”的基础理念，并适用于整个人工智能供应链。这意味着人工智能供应商（及中间商）有责任开发符合“安全设计”原则的产品。“按需安全”意味着客户（及中间商）应通过仅采购符合“安全设计”原则的产品，来督促供应商履行这一责任。

如需了解更多信息，请参阅ASD的：

- “安全设计”基础原则
- ISM的软件开发指南
- 选择安全且可验证的技术

结论

人工智能通过减少对人工流程的依赖、优化风险优先级排序，以及实现对威胁的早期检测和响应，为加强网络防御提供了宝贵机遇。当组织将人工智能整合到符合特定用途的安全工具中，并使其与既定的网络安全基本原则保持一致时，人工智能将帮助它们适应日益复杂且瞬息万变的威胁环境。

然而，在网络安全领域采用人工智能应审慎行事并实施有效治理。实施不当的人工智能系统可能会产生新的漏洞并引发不可预测的故障行为，从而增加组织风险。若审慎使用，人工智能可增强网络安全；若使用不当，则可能使组织面临更大的风险。

更多信息

ASD的资源

- 人工智能
- 安全部署人工智能系统
- 人工智能在运营技术中的安全集成原则
- 网络供应链管理
- 选择安全且可验证的技术
- 安全设计

其他资源

- 英国国家网络安全中心（NCSC）关于支持人工智能应用于英国网络防御的指南
- 欧洲电信标准协会（ETSI）——全球首创：ETSI人工智能全球网络安全标准
- 前沿AI（ITSAP.10.050）——加拿大网络安全中心

附录A：面向人工智能供应商的网络安全问题

附录A提供了一系列问题，旨在帮助组织评估供应商所提供的人工智能能力在安全性、可靠性和成熟度方面的表现。这些问题旨在区分经过精心设计、可投入生产的解决方案，与那些依赖自动化、假设或新兴功能的解决方案。

可量化的安全成效

- 该人工智能在哪些具体安全指标上实现了可量化的提升（例如平均检测时间、平均响应时间、误报/漏报率）？
  - 要求提供基于实际运维的量化证据，而非架构或营销宣传。
- 您如何在真实的安全运维环境中（而非仅限实验室场景）验证和测试AI性能？
  - 确认供应商是否开展持续评估、红队演练、客户端验证及生产环境指标监测。
- 当前AI能力中哪些部分已达到生产级水平，哪些属于路线图规划中的项目？
  - 这有助于区分经过验证的能力与前瞻性或实验性功能。

范围、能力与局限性

- 该人工智能是否仅受限于预定义的规则和工作流，还是能够超越设计好的用例进行推理？
  - 这有助于明确该产品是仅适用于特定狭窄任务的工具，还是具备更通用的分析能力。
- 人工智能模型如何应对新型或前所未见的攻击技术？
  - 应关注其对泛化能力、适应性或模式学习的说明，而非仅依赖静态特征码。

人工监督与决策

- 如何防止使用该人工智能的分析师产生自动化偏见？
  - 应关注那些鼓励审查、质疑和批判性思维的设计特征，而非盲目信任。
- 当人工智能出现错误时，由谁承担责任？错误如何被识别、调查和纠正？
  - 这有助于厘清责任归属、上报流程及持续改进机制。

透明度、可解释性和可审计性

- AI能否以网络防御人员能够审计和防御的方式，解释其推理过程和建议？
  - 透明度对于事件响应、监管审查和高管保障至关重要。
- 您能否证明在发生事件后，AI的输入、决策和输出是如何被记录并重构的？
  - 这有助于取证审查、合规义务以及事件后的经验总结。

AI系统本身的安全性

- AI如何防范篡改、数据中毒或恶意输入？
  - 评估其对抗提示注入、训练数据中毒及反馈循环滥用的韧性。
- 有哪些访问控制和监控措施来保护AI模型、提示词和调优机制？
  - 在安全架构中，应将AI组件视为高价值资产。

数据保护与主权

- 在AI处理过程中，敏感数据如何被处理、存储和保护？
  - 确保符合内部政策、监管要求及客户数据主权义务。
- 哪些数据会被保留、重复使用或用于模型再训练，以及在何种条件下进行？
  - 这有助于识别意外的数据泄露或长期隐私风险。
- 哪些司法管辖区会处理或存储客户数据（包括通过第三方AI服务）？
  - 这对面临跨境或受监管数据限制的组织至关重要。

集成与运营兼容性

- 该AI增强型产品如何与现有的网络安全工具、工作流和数据源进行集成？
  - 请明确说明集成方式是原生集成、基于API，还是需要大量定制化开发。
- 该AI如何与现有的身份和访问管理、日志记录、监控、变更管理及事件响应流程进行集成？
  - AI应强化现有的安全架构，而非作为独立或不透明的系统运行。

韧性、依赖性与长期风险

- 该AI能力对第三方模型、服务或数据源存在哪些依赖？
  - 这有助于评估供应链风险、韧性以及对上游变化的暴露程度。
- 如果AI组件性能下降、发生故障或产生可信度较低的输出，系统将如何响应？
  - 这有助于了解备用机制，以及安全运营能否安全地继续进行。
- 如果风险容忍度或运行条件发生变化，有哪些控制措施可用于限制、调整范围或禁用AI功能？
  - 这对事件响应、监管变化或组织风险偏好的转变至关重要。
- 如果组织更换供应商或部署模式，AI能力的可移植性如何？
  - 这有助于识别锁定风险，特别是在检测或工作流属于专有技术的情况下。
- 供应商是否提供了符合最低要素要求的、可机读的AI软件物料清单（SBOM）和AI组件物料清单（AIBOM）？
  - 这有助于评估AI供应链的透明度及依赖风险。
- 供应商能否在整个AI生命周期内证明模型的清晰溯源？
  - 这有助于保障、可追溯性及事件调查。

附录B：面向网络安全供应商的人工智能能力评估问题

附录B提供了结构化的问题，旨在帮助组织评估网络安全领域中人工智能能力的有效性、可靠性及运营适配性。其重点在于确保人工智能在保持安全、透明并符合组织风险要求的同时，能够带来可衡量的成果。

安全成果与有效性

- 该人工智能能力能否在可衡量的程度上提升安全团队做出决策的速度或准确性？
  - 能够提升决策速度或质量的人工智能具有运营价值；仅提供额外信息的人工智能则可能增加工作量和风险。
- 与现有的非人工智能方法相比，哪些安全活动在人工智能的辅助下得到了明显改善？
  - 这有助于判断人工智能是否真正改变了结果，而非在增加复杂性的同时简单复制现有能力。

范围、假设与限制

- 该人工智能对数据质量、系统架构或威胁行为作出了哪些假设？
  - 理解这些假设有助于识别在经验证条件之外，输出结果可能变得不可靠的情况。
- 有哪些技术控制措施界定了AI无法运行或无法影响结果的范围？
  - 硬性约束可降低滥用风险，或防止AI超出批准用例范围而发生意外扩展。

应对不确定性和错误

- AI如何表达不确定性、低置信度或信息不完整的情况？
  - 明确标示不确定性对于避免产生虚假自信和采取不当应对措施至关重要。
- 有哪些保障措施能限制错误或误导性AI输出带来的影响？
  - 在规模化或高速运行时，隔离机制可缩小错误的影响范围。

透明度与可审计性

- 安全团队能否追溯特定输入如何导致了某项AI推荐?
  - 可追溯性使分析师能够验证、质疑并捍卫由AI支持的决策。
- AI的操作和推荐是否以支持事后审查的方式进行了日志记录?
  - 符合审计要求的日志对于取证调查和合规性保证至关重要。

AI系统的安全性

- AI如何防范对抗性操纵，包括提示词注入和模型规避?
  - 用于防御的人工智能本身也是攻击目标，必须作为攻击面的一部分加以保护。
- 有哪些监控措施可用于检测滥用、性能下降或异常的人工智能行为?
  - 持续监控有助于在新兴风险影响运营之前及时识别。

数据使用与保护

- 人工智能访问哪些运营数据，如何最大限度地限制访问范围?
  - 限制数据访问可降低风险暴露，并支持“最小权限”原则。
- 是否存在将客户数据用于原始目的之外的情况，包括用于模型调优?
  - 明确界定数据使用边界，可降低长期隐私、保密性和法律风险。

系统集成与运营适配

- AI如何集成到现有安全工作流中，同时避免产生并行流程?
  - 紧密集成可维护治理机制并减少运维摩擦。
- 当集成数据源不可用或性能下降时，AI会如何响应?
  - 在性能下降条件下表现出的可预测行为，有助于确保安全运营的持续进行。

韧性与恢复能力

- 能否在不影响核心安全功能的情况下，对AI组件进行回滚或隔离?
  - 在系统遭入侵、配置错误或模型故障后，回滚能力至关重要。
- 如何对AI系统进行安全故障与恢复场景的测试?
  - 韧性测试可确保安全运维在压力下仍保持可靠性。

免责声明

本指南中的内容为一般性信息，不应被视为法律建议，也不应作为在任何特定情况或紧急状况下的求助依据。对于任何重要事项，您应根据自身情况寻求适当的独立专业意见。

联邦政府对因依赖本指南所含信息而造成的任何损害、损失或费用不承担任何责任或义务。

版权所有

© 澳大利亚联邦 2026

除国徽外，且除非另有说明，本出版物中提供的所有材料均依据知识共享署名4.0国际许可协议提供 https://creativecommons.org/licenses/by/4.0/

为避免歧义，这意味着该许可仅适用于本文档中列出的材料。



相关许可条款的详细信息以及CC BY 4.0许可协议的完整法律文本，均可在知识共享网站上查阅 https://creativecommons.org/licenses/by/4.0/legalcode.en

国徽的使用

国徽的使用条款详见总理及内阁部官网 https://www.pmc.gov.au/resources/commonwealth-coat-arms-information-and-guidelines

如需更多信息或报告网络安全事件，请联系我们：

cyber.gov.au | 1300 CYBER1 (1300 292 371)

ASD AUSTRALIAN SIGNALS DIRECTORATE

ACSC Australian Cyber Security Centre