AI Technology Research

IGF Internet Governance Forum RIYADH 2024

人工智能系统安全开发和部署准则

kaspersky 引领未来

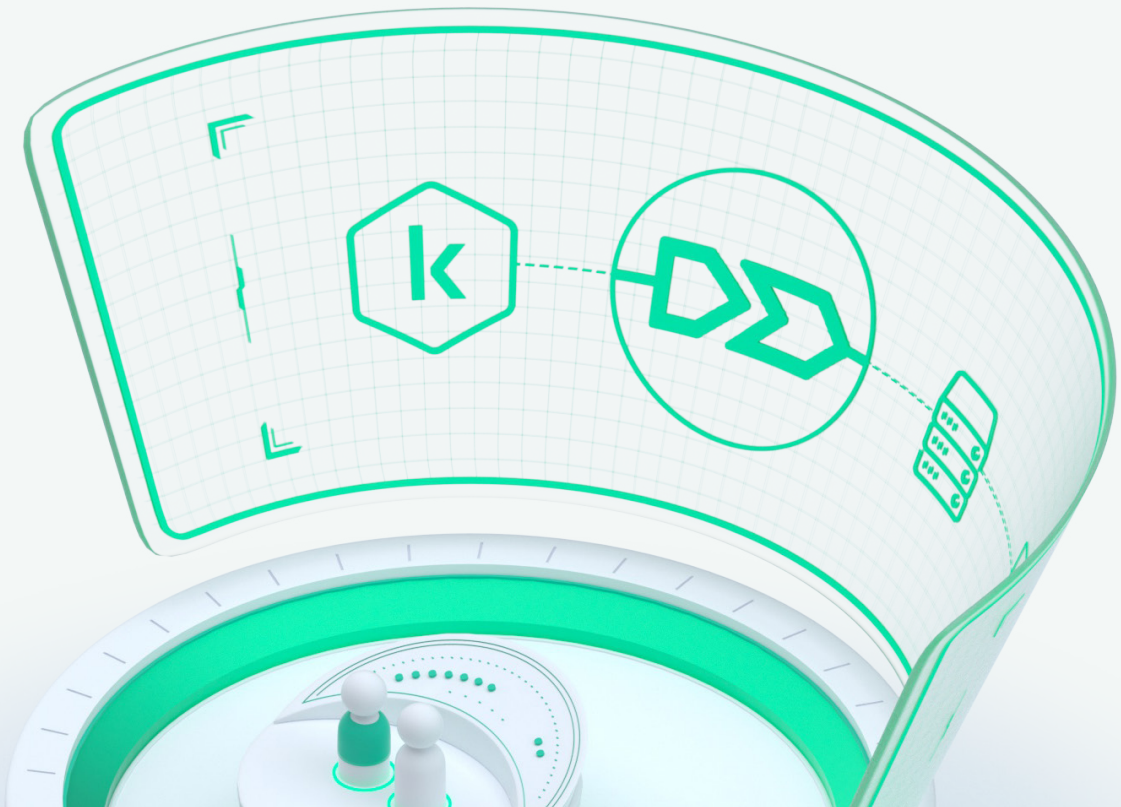
致谢

本论文由卡巴斯基撰写，并在**互联网治理论坛**第 19 届年会（2024 年 12 月 15 日至 19 日）的 **31 号研讨会“人工智能中的网络安全：平衡创新与风险”研讨会**上发表。

以下人士为本文的撰写做出了贡献：

- Yury Shelikhov，卡巴斯基网络安全与数据保护负责人
- Vladislav Tushkanov，卡巴斯基机器学习技术研究事业部、研发事业部经理
- Alexey Antonov，卡巴斯基检测方法分析事业部首席数据科学家
- Allison Wylde，联合国互联网治理论坛人工智能政策网络（互操作性）团队成员
- Melodena Stephens 博士，阿联酋穆罕默德·本·拉希德政府学院创新与技术治理教授
- Sergio Mayo Macías，西班牙阿拉贡技术研究院 (ITA) 创新项目经理
- Igor Kumagin，网络安全专家
- Dmitry Fonarev，卡巴斯基高级公共事务经理

目录



简介

目的

新的数字技术伴随着**新的网络安全风险和攻击途径**。因此，各公司必须确保 AI 的集成免受这些威胁。AI 系统开发中的安全概念已成为各种监管举措（例如欧盟的《人工智能法案》或新加坡的《生成式人工智能模型治理框架》）的重点，以最大限度地降低相关的网络风险。欧盟正在通过《人工智能法案》制定严格的 AI 法规，旨在确保透明度、安全性和道德标准。美国则侧重于制定行业标准和鼓励创新，而不是制定严格的法律。中国则积极制定标准和法规，一方面支持 AI 技术的发展，另一方面也限制其在某些领域的使用。

尽管在监管方面取得了进展，但在一般框架与更技术层面的实际实施之间仍存在重要差距。在本文中，我们探讨了 AI 系统实施中应考虑的**基本网络安全要求**。这些要求应适用于**更广泛的依赖第三方 AI 组件**构建自己解决方案的公司。

为了安全地实施 AI，组织需要技术指导来帮助其在基础设施内开发和部署 AI。在没有适当指导的情况下实施 AI 可能会带来重大风险。本文重点为 AI 系统、MLOps 和 AI DevOps 的开发人员和管理员提供指导，利用现有的基础模型创建通用的 AI 解决方案，并特别强调基于云的 AI 系统。本文探讨了**开发、部署和运行 AI 系统的关键方面**，包括设计、安全最佳实践和集成，但不讨论基础模型开发。

人工智能 (AI) 已发展成为影响全球经济的关键技术，并已融入日常生活。利用 AI，组织能够自动执行日常任务，增强客户服务并让员工更快、更轻松地获取信息。

>50%

卡巴斯基最近的一项研究表明，超过 50% 的公司在基础设施中实施了基于 AI 的解决方案*。

33%

的公司计划在两年内采用这项技术。

* 超过一半的公司在业务流程中使用 AI 和物联网，https://www.kaspersky.com/about/press-releases/more-than-half-of-companies-use-ai-and-iot-in-their-business-processes

随着 AI 技术越来越多地部署于各个组织，AI 系统面临的威胁日益增加。网络攻击会影响 AI 开发的所有阶段，从数据集到算法和模型输出。

AI 威胁态势概述

根据卡巴斯基的研究*，AI 系统面临着独特且不断变化的安全挑战，使系统运行面临风险。恶意行为者利用训练数据中的漏洞，操纵模型以改变其行为，并破坏系统完整性。这凸显了 AI 应用中对全面安全的迫切需求。

模型开发的问题

与传统编程中可以明确理解和测试代码行为不同，机器学习模型，尤其是具有数百万或数十亿参数的深度学习模型，本质上非常复杂，并且通常以“黑箱”方式运行。这种复杂性使得完全预测和解释模型行为变得困难。因此，未被发现的错误产生严重后果（例如影响银行的金融稳定性或患者的生命安全）的风险显著增加。

另一个问题是，AI 模型有时可能基于**无关或不重要的输入数据属性**而不是相关特征来做出决策。例如，图像识别模型可能会学会仅根据动物身上的斑点图案来对猎豹、美洲豹和美洲虎等动物进行分类，而不参考它们的整体解剖结构**。在一个案例中，一个模型错误地将有斑点的沙发分类为美洲豹，因为它将斑点图案与动物相关联。这种分类错误可能导致医疗保健、教育、社会福利、交通、政府部门等领域的关键应用中出现错误输出。

用于训练的数据与部署期间遇到的数据不一致可能会导致模型性能较差。例如，如果一个模型的训练数据来自于一种类型的仪器，但该模型应用于来自不同设备的数据，那么它可能会学会设备特定的特征，而不是本应识别的底层对象。这种不匹配可能导致预测或分类不准确。

从头开始训练模型和微调基础模型都可能面临这样的挑战。虽然微调数据集更小、更易于管理，但也可能引入虚假的相关性，导致生成的模型与预期目标不一致。

* AI 遭受攻击，https://content.kaspersky-labs.com/se/media/en/business-security/enterprise/machine-learning-cybersecurity-whitepaper.pdf

** 一张豹纹沙发突然出现，https://web.archive.org/web/20200208171948/http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

对 AI 模型的攻击

恶意行为者可以通过各种方法将 AI 模型作为攻击目标以下是攻击者如何利用 AI 设计、训练和交互机制中的漏洞的例子：

攻击方式	描述
数据投毒：破坏模型完整性	数据投毒是指攻击者将恶意数据注入训练数据集，以影响模型的行为。通过精心制作和添加带毒样本*，攻击者可以使模型对某些输入做出错误的决策或不正确的分类。这种类型的攻击可能会损害模型的完整性并破坏其可靠性。这也适用于基础模型的微调。
对抗性攻击：AI 的隐形操纵	对抗性攻击会对输入数据进行细微的修改，使 AI 模型将其错误地分类，而人类察觉不到这些变化**。攻击者向输入添加特制的噪声，导致模型产生错误的输出，而输入在人类观察者看来没有发生变化。
AI 数据记忆：无意暴露的风险	现代 AI 模型可能会无意中记住训练数据的某些细节，特别是当数据包含独特或异常的样本时。攻击者可以利用这一点，使用技术手段提取模型无意中存储的敏感信息。这可能导致个人用户数据或机密商业信息泄露。
即时注入：对大型语言模型的威胁	提示注入是特定于 ChatGPT 等大型语言模型 (LLM) 的威胁。开发人员将 LLM 设定为使用自然语言提供初始提示来执行任务。由于用户也使用自然语言与模型交互，因此模型本质上无法区分开发人员指令和用户输入。攻击者可以制作输入来覆盖或操纵模型的行为，使其执行非预期的操作或泄露敏感信息。这些恶意提示可以由用户直接输入，也可以嵌入到模型处理的数据中，例如文档或网页。

这些只是最相关的攻击；对 AI 系统的所有可能攻击的完整描述不在本文讨论范围内。

* 了解数据投毒攻击，https://bdtechtalks.com/2020/10/07/machine-learning-data-poisoning/
针对机器学习的攻击：概述，(https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

** 解释和利用对抗性样本，https://arxiv.org/abs/1412.6572
深度学习中的对抗性攻击与防御，https://arxiv.org/abs/2201.06192
如何混淆反恶意软件神经网络：对抗性攻击与防护，https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

传统安全漏洞

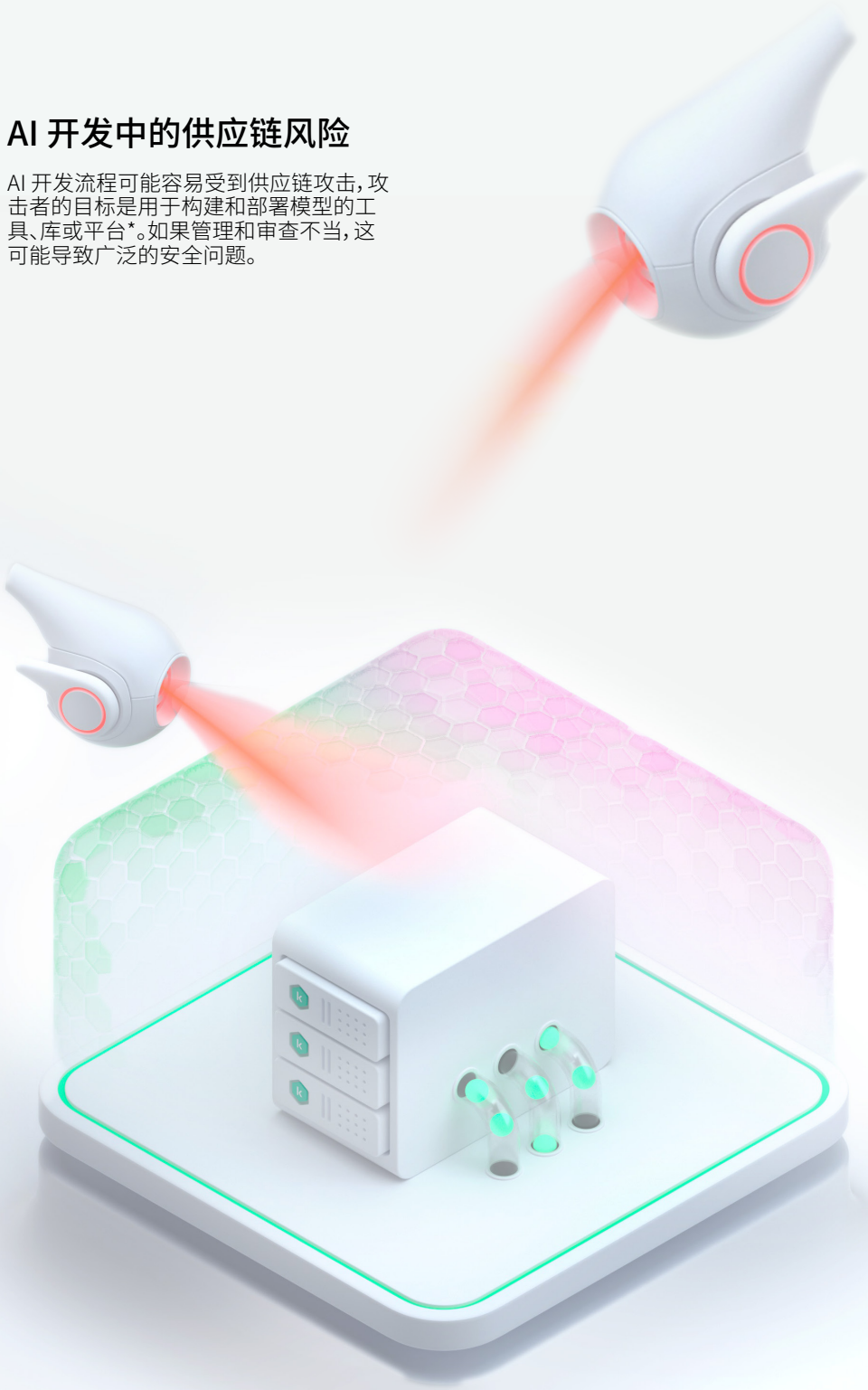
AI 模型可能容易受到传统安全弱点的影响：

来自第三方资源的 AI 漏洞

AI 系统通常依赖于来自开放存储库的第三方模型或数据集。这些资源可能包含无意错误或恶意行为者故意插入的后门。整合此类受损组件可能会将漏洞引入 AI 系统，影响其安全性。

AI 开发中的供应链风险

AI 开发流程可能容易受到供应链攻击，攻击者的目标是用于构建和部署模型的工具、库或平台*。如果管理和审查不当，这可能导致广泛的安全问题。



代码错误暴露 AI 漏洞

AI 访问接口代码中的错误可能会导致漏洞。

AI 组件被盗的风险

AI 系统或其关键组件（例如模型或数据集），如果没有本文中定义的充分保护，可能会被窃取。

* 2022 年 12 月 25 日至 12 月 30 日期间，PyTorch-nightly 依赖链遭到破坏，https://pytorch.org/blog/compromised-nightly-dependency

准则

网络安全意识和培训

AI 等新技术的实施需要**领导层支持、制定内部政策和治理，以及对员工进行与 AI 相关的风险和威胁的专门培训。**由于这项技术发展迅速，后者至关重要。许多供开发人员使用的 AI 资源还不够成熟，或者没有完全采用默认安全和设计安全原则。结果，AI 系统的开发人员要承受额外的负担，以应对需要解释的潜在风险。

为此，组织除了标准安全实践外，还应考虑实施以下措施：

1

组织的领导层需要意识到与使用 AI 服务相关的安全风险，并学习如何管理这些风险。

2

组织的安全政策应进行更新，以满足 AI 服务的特定要求，确保所有员工和承包商都熟悉这些政策。

3

该政策应概述与开发和使用 AI 服务相关的风险，以及根据当地法律法规对使用这些服务所施加的当前限制。

4

该政策应定义与在公司内部使用 AI 服务有关的角色和责任。

5

应开发或购买有关在公司内部安全使用 AI 服务的企业培训课程，确保所有员工和新员工都完成该课程。该计划应涵盖组织政策、现有威胁和事件示例、威胁防护措施、适用法律和其他相关主题，并在适用的情况下包括基于桌面场景的练习。员工在完成课程后应接受测试。课程应定期及时更新。

6

现有的信息安全课程应该进行更新，以包括恶意行为者利用 AI 服务攻击公司所使用的新方法，例如文本生成。这方面的例子包括语音克隆、照片处理或虚假视频生成。

7

应组织监测与安全使用 AI 服务有关的法律法规。内部政策和培训计划应及时更新。

威胁建模有助于在 AI 系统开发的早期阶段识别、理解和减轻潜在的安全风险。

威胁建模/风险评估

此过程对于 AI 系统尤为重要，因为这是一项新兴技术，其风险在不断演变和调整。进行风险评估有助于更好地应对这些挑战。此外，威胁建模可以帮助新接触该领域的开发人员更好地应对与开发 AI 系统相关的挑战。还将使他们能够主动识别并缓解 AI 系统中的弱点，防止这些弱点被利用。

为了确保威胁建模过程有效，组织应遵守**以下建议：**

选择风险评估方法（例如，STRIDE、DREAD、LINDDUN、PASTA、TRIKE）并制定对 AI 服务进行风险评估和威胁建模的程序。评估方法还应包括用于确定风险级别的框架、用于管理组织内风险的策略（包括可接受的风险阈值）、风险监控程序以及负责监督过程的人员分配。

应该对所有现有和新开发的 AI 服务进行风险评估。

风险评估应包括识别潜在威胁行为者或攻击者，以及已识别的威胁和风险。应使用 NIST-AI-600-1、MITRE ATLAS、OWASP Top 10 for LLM Applications、DASF 和 CSA 等参考材料来识别已知的威胁和风险。

管理风险时，考虑将威胁分为以下类别：

- 因不使用 AI 服务而产生的威胁，
- 因不合规而产生的威胁，
- 因用户滥用 AI 服务而产生的威胁，
- 对用于训练的 AI 模型和数据集的威胁，
- AI 模型对服务构成的威胁
- 对相关数据的威胁，以及
- 对环境、社会和治理 (ESG) 的威胁。

有关在 AI 服务中识别到的风险的信息应传达给组织的领导层。

基础设施安全（云）

AI 服务通常以云服务的形式提供，并且通常需要专门的基础设施，例如配备 GPU、FPGA、ASIC 或 TPU 的服务器。鉴于 AI 系统的敏感性，应按照**最先进的网络安全框架**对其加以保护，例如 NIST 网络安全框架或具有类似标准的其他框架。AI 服务通常使用开源或免费软件，如 TensorFlow、PyTorch 或 Keras，以及 Pandas、NumPy 和 SciPy 等库。为确保此环境的安全，应考虑**以下要求**：

1

标识所有资产并维护信息资产清单，例如用于模型训练和测试的数据集、于微调训练的数据、数据库、数据卡、模型和风险、传入和传出服务的数据、模型权重和超参数、来自 LLM 系统的日志数据。

2

在包括网络、操作系统、数据库、软件、数据和模型在内的所有层面控制访问。为管理访问实施双因素身份验证 (2FA)。

3

记录所有事件并确保日志数据受到保护。监控安全事件和潜在漏洞。

4

实施针对恶意软件和其他类型攻击的防护措施。定期对基础设施组件应用安全补丁。

5

对网络进行分段以保护敏感区域。对传输中和静止的数据使用加密。

6

确保关键数据的完整性，并验证正在使用的库和模型的真实性。

7

提供服务器和通信信道冗余。定期执行备份并确保其功能正常。

8

将密钥安全地存储在 KeyVault 中。

9

在整个基础设施中应用最小权限和零信任原则。

根据支持 AI 服务的基础设施，附加要求可能包括：

使用 API 网关管理对模型的访问并通过 API 处理身份验证。

实施特定于 Kubernetes 环境的安全措施。

遵守保护基于云的服务的最佳实践。

确保源代码、训练数据、模型和自动化脚本的完整性。

隔离训练数据、模型和训练环境，以防止数据泄露或污染。

供应链和数据安全

供应链攻击对任何组织的基础设施都构成重大威胁，AI 架构也不例外。已知有专门用于神经网络训练的库成为此类攻击目标的案例。然而，在 AI 领域，人们对服务提供商的安全性和机器学习模型的保护产生了特定的担忧。

对先进 AI 模型（尤其是 LLM）的访问通常依赖于基于云的解决方案。然而，某些模型在某些地区不可用，加之其他限制，可能促使公司员工和开发人员选择通过 API 转售 AI 模型访问权限的第三方服务（代理）来执行日常任务。这种做法引入了重大风险——从在代理服务发生安全事件时开启额外的数据泄露途径，到滥用获得的数据进行转售或训练自己的 LLM 版本等不道德做法。**重要的是要了解这些风险，**仔细审查主要提供商和代理的隐私政策，并在全公司范围内开展关于使用第三方服务处理工作任务的风险意识培训。

为了降低这些风险，组织可以选择部署本地 LLM 服务。这种方法可确保在满足某些指定条件（例如，网络隔离、遥测停用等）的情况下，LLM 处理的机密数据将保留在公司内部。但是，除了与机器学习框架中的漏洞相关的风险，这种方法还涉及与模型后门相关的威胁。在这种情况下，意味着用于分发机器学习模型的数据格式可能具有不同的安全级别，并且某些格式可能允许嵌入任意代码，这些代码可在模型运行时执行。研究表明，公共存储库中存在可以在加载时执行特定代码的模型，尽管这种模型的数量有限。使用第三方资源库中的模型而不是原始模型，可能是由于在某些地区无法下载或受到许可限制。使用安全格式（例如 safetensors）可以解决这个问题，但需要提高开发人员和数据分析师的意识，让他们认识到选择可靠的模型来源和使用安全格式交换模型权重的重要性。

- 确保从可靠、合法的来源获取 AI 模型。避免使用第三方存储库。
- 使用安全格式（如 safetensors）交换模型权重，以避免任意代码执行的风险。
- 实施措施以检测和应对针对 AI 相关组件的供应链攻击。
- 评估和审查用于访问 AI 模型的第三方服务和代理的隐私政策，以确保它们符合安全标准。
- 在确保数据隐私的条件（例如网络隔离和禁用遥测功能）下在本地部署 AI 模型。
- 建立用于安全部署本地模型的协议，以最大限度地降低与机器学习模型中的潜在后门相关的风险。
- 定期更新和修补机器学习框架以解决已知漏洞。
- 实施措施以确保 AI 模型处理的敏感数据不会离开组织的基础设施。
- 在使用第三方 API 时，根据顶级国际标准（如 OWASP API Security Top 10）进行安全审计。

测试和验证

执行评估并确定风险后，了解如何防范在训练和应用模型时出现意外或故意的错误至关重要。为了实现此目标，组织可以考虑实施以下措施：

1

评估系统中意外或故意的错误可能造成的潜在损害。评估用于训练 AI 模型的数据及其处理的数据的价值。

2

确定是否使用开源框架、模型或数据集来构建 AI 系统。

3

确定潜在的用户群：公司员工、客户或开放的公共访问。

4

验证在模型构建中是否遵循机器学习最佳实践。确保根据模型的运作方式正确地将数据集划分为训练集、测试集和验证集。

5

在验证 AI 模型及其指标（假阳性和假阴性）时，检查数据集划分标准是否适合数据的性质（例如，按时间顺序划分时序数据，并避免数据泄露）。

6

评估模型使用哪些特征来做出决策，以及是否与该领域专家的直觉一致。使用模型解释方法（例如 SHAP 载体）了解模型的决策过程。

7

评估模型的真实世界性能，以确保其达到预期结果。持续监控模型，因为输入数据的分布可能会随时间变化，这可能会降低模型的性能。

8

调整测试计划，以检查模型是否容易受到机器学习模型特有漏洞的影响（例如，对抗性攻击和数据投毒）。

漏洞报告

AI 是一个相对较新的技术领域，正在迅速发展。尽管具有显著优势，但许多 AI 系统容易受到这些技术特有的漏洞的影响。

主要问题之一是，某些 AI 系统可能包含漏洞，可被利用来获取对数据的未经授权访问权限。AI 系统漏洞的另一个例子是偏见，即用于训练模型的数据不具代表性或包含隐藏的偏见。例如，当刻板印象和错误的社会假设渗透到算法的数据集时，AI 系统可能会受到偏见的影响，或者因数据不完整而产生测量偏差。结果，这些系统可能会做出不公平或歧视性的决定，对用户产生负面影响并破坏 AI 技术的可信度。

为了解决这些问题，有必要实施一种机制来允许用户报告 AI 系统中**已识别的漏洞**和偏见。这种报告机制将使组织能够快速收到反馈并采取行动：

1. 制定公开的政策，明确 AI 系统中的漏洞并概述用户如何报告这些漏洞。
2. 为用户提供安全的漏洞报告方法，如加密的网络表单或专用的电子邮件地址。
3. 制定快速评估、优先处理和修复所报告漏洞的程序。
4. 与报告漏洞的人员就问题的状态和解决方法进行沟通。
5. 及时告知用户已知漏洞和补救措施，建立信任并展示责任心。
6. 通过漏洞赏金计划与安全研究人员合作。这也有助于了解新出现的威胁和 AI 安全最佳实践。

防御机器学习特定的攻击

鉴于目前 AI 发展的进步，一些 AI 组件可能容易受到机器学习特定的攻击。例如，这些攻击可能通过故意向模型输入畸形数据或隐藏命令来利用漏洞。因此，使用免费的 AI 来构建系统的组织应该意识到这些风险。防范机器学习特定的攻击需要**实施各种安全措施**，例如：

将对抗性样本* 纳入训练数据集，帮助模型学习如何更有效地处理这些输入。

应用蒸馏技术，通过简化模型的决策过程，使模型更加能承受对抗性输入。

考虑使用单调模型**，这可以提高稳定性并降低对对抗性操纵的敏感性。

引入可以检测用户请求中的对抗性或异常输入的系统，使模型能够在处理数据之前检测并拒绝恶意尝试。

为防止数据投毒，AI 系统开发人员应该**分析异常对象的训练样本**，并将新模型的性能与以前的版本进行比较，以确定模型属性的任何急剧变化。

为防范对 LLM 的提示注入攻击，AI 系统开发人员可以实施一套系统，对传入的用户请求或馈入 LLM 输入的其他第三方数据进行分析。另一种方法是分析对此类请求的响应，并评估其是否符合系统的当前任务。

* 如何混淆反恶意软件神经网络。对抗性攻击与防护，https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/

** 用于实时动态检测恶意软件的单调模型，https://arxiv.org/pdf/1804.03643

定期安全更新和维护

AI 领域，特别是 LLM 的应用尚不成熟，代码质量也并不总是能达到高标准。因此，许多用于处理机器学习的框架和工具可能包含**大量漏洞**。幸运的是，我们正处在流行框架积极向生产级标准靠拢的阶段，并且定期发布版本和安全更新。此外，旨在发现机器学习基础设施漏洞的漏洞赏金计划层出不穷，也有助于快速解决这些问题。

这突显了持续监控基础设施状态的重要性，包括监控用于跟踪实验的平台以及旨在与云服务通信的库。对于发展较慢的领域的项目来说，保持基础设施为最新可能需要花费较长的不必要时间。此外，使用最新版本的库可能会导致兼容性问题，需要投入更多资源来开发和维护依赖于这些库的代码功能。在规划 AI 计划时，需要将这些成本考虑在内。

使用基于云的 AI 模型（例如 LLM）的另一个风险是，**每个模型版本的生命周期相对较短**。为项目选择的模型可能在短时间内就被平台提供商替换为新版本。虽然模型的整体质量预期会提高，但其在特定任务中的行为以及对攻击的承受能力可能会发生变化。例如，它可能包括提示注入或试图通过越狱获得未经授权的输出。这需要提前进行规划，以确保模型之间平稳过渡，并且不影响下游任务的质量或安全级别：

1. 确保基础设施及时更新最新的安全补丁和框架更新。
2. 积极参与漏洞赏金计划，并使用漏洞扫描工具检测机器学习框架和 AI 基础设施中的弱点。
3. 定期审查并应用机器学习工具和库的安全更新，以减少已知漏洞的风险。
4. 通过分配开发和测试资源，针对使用最新版本的库和框架时可能出现的兼容性问题做好规划。
5. 实施策略来管理基于云的 AI 模型的生命周期，并制定计划，以便在提供商发布新模型版本时做好过渡。

符合国际标准

随着 AI 监管的快速发展，遵守相关法律和遵循最佳实践变得越来越重要。首先，AI 训练数据可能从不同司法管辖区的多个来源收集，这使得处理和使用这些信息变得困难。此外，模型通常来自**开放存储库**，这位其是否符合特定司法管辖区的监管要求增加了不确定性。

因此，AI 开发人员面临着一项艰巨的任务，即确保遵守系统使用国家的所有法律要求。在这种情况下，最佳策略是遵循中国、欧盟或美国等 **AI 监管领导者的标准**。许多国家已经分享他们的方法并实施类似的要求，让开发人员提前为系统的全球部署做好准备：

1

制定 AI 使用和开发的伦理准则，以确保相关流程的透明度和问责制。

2

确保从不同来源收集的所有数据符合每个司法管辖区的数据隐私法，例如欧洲的《通用数据保护条例》(GDPR) 或美国的《加州消费者隐私法案》(CCPA)。

3

使用来自开放存储库的 AI 模型时，验证它们是否遵守知识产权。

4

遵循领先的监管框架，例如欧盟的《人工智能法案》或美国的《人工智能权利法案》，因为这些框架经常被其他国家用作基准。

5

及时了解全球范围内新出现的和不断变化的 AI 法规。

6

定期审核 AI 模型和系统是否符合国际标准，以识别并降低潜在的法律和道德风险。

结论

与大多数技术创新一样，AI 技术既带来了巨大的机遇，也带来了重大的威胁。与 AI 相关的网络安全风险及其对社会的影响取决于开发人员的行为和意图。为了安全地实施 AI，组织需要遵循在基础设施内开发和部署 AI 的技术准则，在不遵循适当建议的情况下执行此过程可能带来重大风险。对于组织而言，至关重要的是建立贯穿整个 AI 生命周期的安全和责任文化并纳入基本安全控制，从风险评估和系统测试到确保供应链安全和持续维护。成功落实所提出的要求将有助于**缓解**将 AI 系统引入公司运营的相关风险。

了解更多信息

www.kaspersky.com

#kaspersky
#bringonthefuture