OWASP AI Exchange项目介绍与实践

OWASP中国广东分会负责人

肖文棣

OWASP介绍

简介

- Open World Application Security Project 开放全球应用安全项目 (OWASP) 是一个非营利性基金会，致力于提高软件的安全性。

愿景

- 不再有不安全的软件。

使命

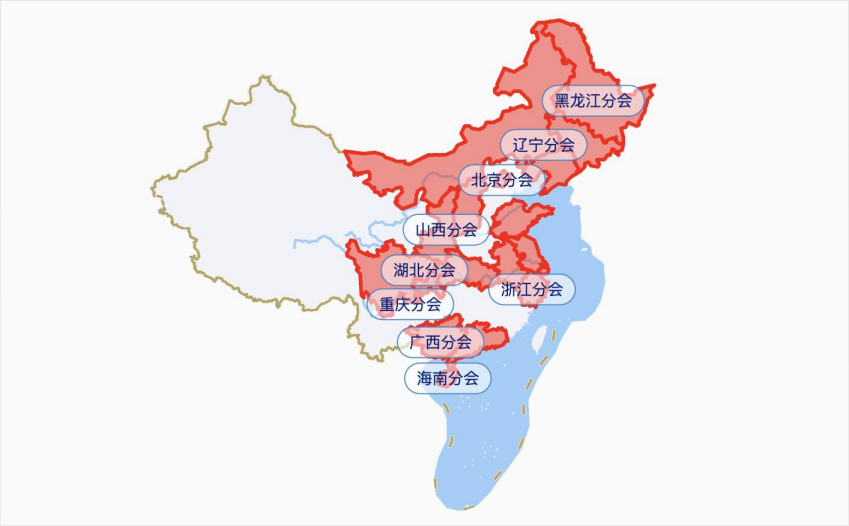
- 成为通过教育、工具和协作为安全软件提供支持的全球开放社区。

https://owasp.org/

OWASP中国介绍

组织性质与使命

- **性质：**OWASP 中国是国际非营利性组织 OWASP 在中国大陆的分支，秉持着开源、共享的理念，致力于推动中国地区应用软件安全的发展。
- **使命：**其使命是提高应用软件的安全性，使软件安全可视化，让个人和组织能够对应用安全风险作出更清晰的决策，从而推动安全标准、安全测试工具、安全指导手册等应用安全技术的进步。



组织架构

- **领导团队：**由主席、副主席、社区经理及各地区负责人等组成。主席是RIP，副主席是王颉。
- **区域：**涵盖了安徽、北京、广东、广西、海南、黑龙江、吉林、江苏、辽宁、内蒙古、陕西、山西、山东、四川、上海、浙江等多个地区，各地区均有相应的负责人。

https://owasp.org/www-chapter-china-mainland/

OWASP中国广东分会

负责人：肖文棣、刘志诚

活动：

- 一年两次的线下会议
- OWASP全球项目的本地化
- 本地项目的全球化 http://www.owasp.org.cn/OWASP-CHINA/owasp-project/

OWASP AI Exchange

AI安全的首选资源共享中心

介绍了如何保护AI和以数据为中心的系统免受安全威胁。

提供了一个社区驱动的环境，让安全专业人员、研究人员和开发人员能够共享知识、工具和最佳实践。

https://owaspai.org/

AI Exchange项目负责人

职位与所属机构：

- Software Improvement Group 的高级主管，同时也是 OWASP AI Exchange 项目的负责人以及 OWASP 集成标准项目的联合负责人.

专业成就：

- 他是多项安全和 AI 标准的作者及合著者， 例如他是 ISO/IEC 5338 标准（关于 AI 工程）的主要作者，还参与了 CEN/CENELEC 针对欧盟 AI 法案的安全工作组以及 ISO/IEC 27090 小组等工作.

对 OWASP 的贡献：

- 在他的带领下，OWASP 推出了 AI Exchange 项目，该项目旨在促进应对 AI 安全及相关监管挑战的专家之间进行开源协作。通过汇集全球专业人士的见解和策略，利用 Software Improvement Group 的威胁模型，为缓解 AI 的安全威胁提供了一个知识共享空间，助力保障 AI 系统的安全，并持续规范 AI 威胁.

推动行业合作与交流：

- 他积极倡导行业内的合作，他呼吁 AI 专家和行业专业人士参与到 OWASP AI Exchange 项目中，鼓励大家访问项目的 GitHub 仓库并为知识库的不断完善贡献力量，强调每个参与者的视角都有助于优化应对 AI 安全问题的方法.

Rob van der Veer

AI Exchange本地化团队

组长

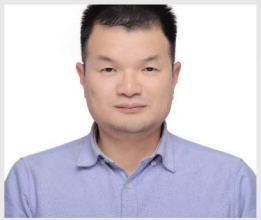
- 肖文棣

成员

- 周乐坤、陈毓灵、严文聪、黄小波、关昕健、唐龙、欧阳宁东、牛承伟、钟英南、刘志诚

AI Exchange本地化团队


钟英南


肖文棣


刘志诚


关昕健


陈毓灵


周乐坤


黄小波

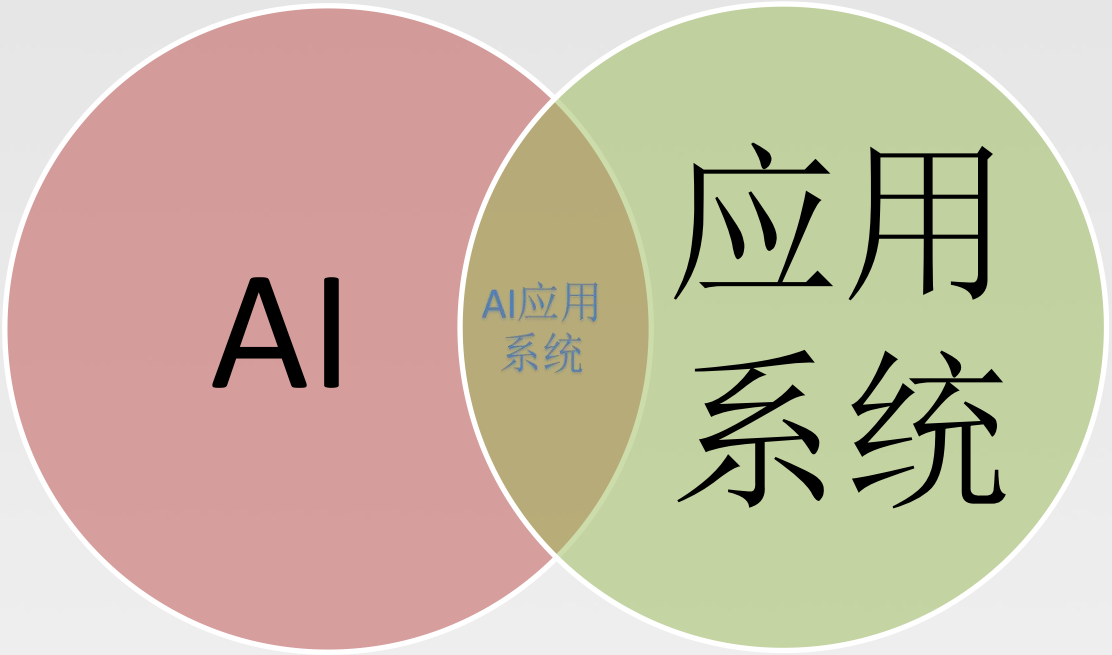

欧阳宁东


唐龙

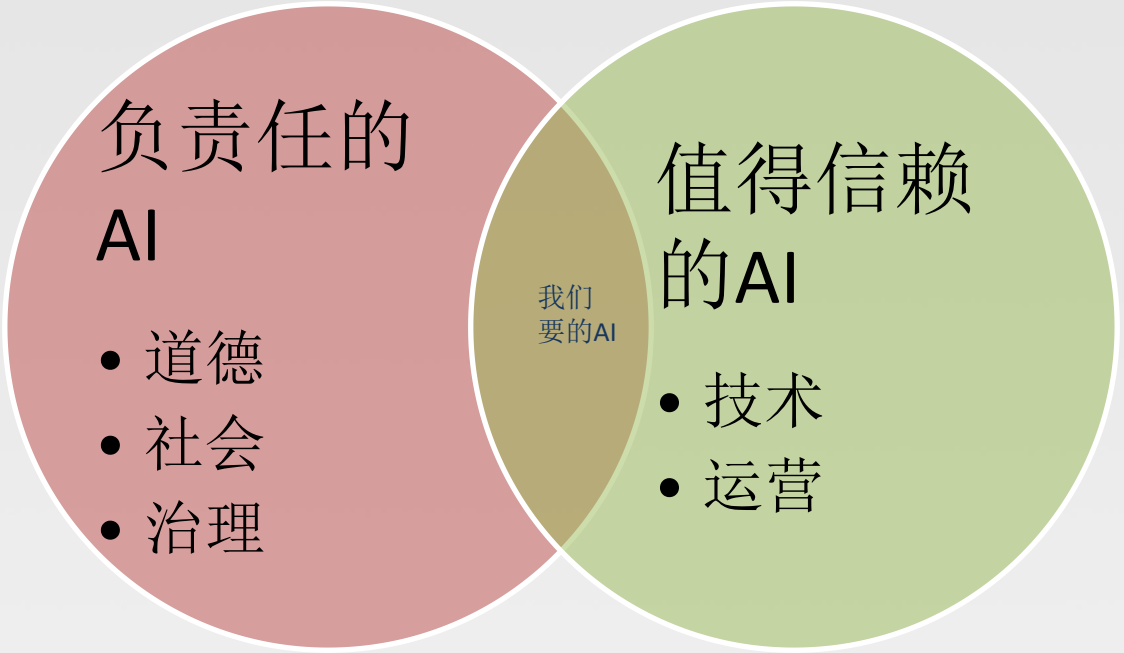

严文聪


牛承伟

AI安全与OWASP



负责任的AI与值得信赖的AI



AI的各个维度

准确性
- AI 模型是否足够正确以执行其“业务功能”
- 安全性的范围仅限于减轻这些攻击的风险

无害性
- 受到保护免受/不太可能造成伤害的条件
- AI 系统的无害性是关于存在伤害风险（通常意味着身体伤害，但不限于此）时的准确性水平

透明度
- 共享有关方法的信息，以警告用户和依赖系统的准确性风险

可解释性
- 共享信息，通过更详细地解释特定结果的产生来帮助用户验证准确性

稳健性
- 预期或意外的输入变化下保持准确性的能力

无歧视
- 没有对受保护属性的不必要偏见
- 没有系统性的不准确

同理心
- 在验证 AI 应用时，应始终考虑可行的安全性水平

问责制
- 问责制与安全的关系是，安全措施应该是可证明的，包括导致这些措施的处理过程

安全性
- AI 的安全方面是 AI Exchange 的中心主题

AI风险分类

战略

运营

财务

合规

声誉

技术

环境、社会与治理（ESG）

AI隐私风险

与安全相关的

- 对训练/测试数据、模型输入或输出中的个人数据进行机密性和完整性保护
- 如果模型行为可能损害个人隐私，则对该行为进行完整性保护

与安全无关的

- 个人的其他权利相关的威胁和控制，如GDPR 等隐私法规所涵盖，包括使用限制、同意、公平性、透明度、数据准确性、更正权/反对权/删除权/请求权

AI版权与安全

AI的输出或生成的内容目前尚未受到版权法的保护。

提供输入内容、文本、训练数据等贡献者可能拥有这些材料的版权。

- 针对 Stability AI、Midjourney 和 DeviantArt 提起的集体诉讼指控，通过使用网络抓取的图像来训练他们的工具，侵犯了数百万艺术家的权利。

在AI训练中使用某些版权材料可能被视为合理使用，但未必合法。

AI风险的治理


识别风险
• 不良模型行为
• 训练数据泄露
• 模型盗窃
• 输入数据泄露
评估风险
• 估计可能性
• 评估影响
• 确定风险优先级
处置风险
• 缓解（控制措施）
• 转移
• 规避
• 接受
风险沟通与监控
• 定期与利益相关者分享信息
• 风险登记册
分配责任
• 责任共担模型
• 确定责任人
验证外部责任
• 定期审核和评估第三方控制措施
选择控制措施
• 首先最高风险
• 从成本最低的控制措施开始
接受剩余风险
进一步管理可选控制措施
• 持续监控
• 文档化
• 报告
• 事件响应
持续监控
• 定期审查和更新风险处理计划以适应新的漏洞

AI Exchange的控制措施

- AI通用控制
- 开发AI时的安全威胁的控制措施
- 使用AI时的安全威胁的控制措施
- AI运行时的安全威胁的控制措施

AI通用控制

AI治理

- 将AI纳入软件生命周期SDLC中进行管理
  - 人工智能纳入信息安全及软件生命周期流程当中：人工智能计划、安全计划、开发计划、安全开发计划、合规检查、安全培训

基于风险的信息技术安全控制措施

- 将标准的控制措施用于AI系统
  - ISO 15408、ASVS、开放 CRE、ISO 27001 标准附录 A、NIST SP800-53
- 调整控制措施适用于AI系统
  - 监控哪些使用模式，如监控使用情况、模型访问控制、速率限制
- 采用新的控制措施
  - 机密计算、模型混淆、提示输入验证、输入隔离

数据科学安全控制措施

- 开发阶段的控制
  - 联邦学习、持续验证、不良偏差测试、抗规避鲁棒模型、抗投毒鲁棒模型、对抗性训练、训练数据失真、对抗鲁棒净化、模型集成、增加训练数据、小型模型、数据质量控制
- 运行阶段的控制
  - 检测异常输入、检测对抗性输入、拒绝服务输入验证、输入失真、过滤敏感模型输出、模糊置信度

最小化数据

- 限制静态和传输中数据的数量，以及数据的存储时间
  - 数据最小化、允许的数据、短期留存、模糊处理训练数据

控制行为影响

- 模型可能会以不当的方式运行 —— 因为失误，或是受操控
  - 监督、最小模型权限、AI透明度、可解释性、持续验证、不良偏差测试

AI通用控制-保密性控制

资产与影响	生命周期内的攻击面	威胁 / 风险类别	控制措施
训练数据保密性	运行时 —— 模型使用	模型输出中的数据泄露	敏感数据限制（最小化数据量、缩短数据留存时间、混淆训练数据），以及：
			监控、速率限制、模型访问控制，以及：
			过滤敏感的模型输出
		模型逆向还原 / 成员推断	敏感数据限制（最小化数据量、缩短数据留存时间、混淆训练数据），以及：
			监控、速率限制、模型访问控制，以及：
			模糊置信度、小型化模型
	开发阶段 —— 工程环境	训练数据泄露	敏感数据限制（最小化数据量、缩短数据留存时间、混淆训练数据），以及：
			开发环境安全、数据隔离、联邦学习
模型保密性	运行时 —— 模型使用	通过使用过程窃取模型（输入输出采集）	监控、速率限制、模型访问控制
	运行时 —— 侵入已部署模型	运行时直接窃取模型	运行时模型保密性、模型混淆
	开发阶段 —— 工程环境	开发阶段窃取模型	开发环境安全、数据隔离、联邦学习
模型输入数据保密性	运行时 —— 所有信息技术环节	模型输入泄露	模型输入保密性

AI通用控制-完整性控制

资产与影响	生命周期内的攻击面	威胁 / 风险类别	控制措施
模型行为完整性	运行时 —— 模型使用（提供输入 / 读取输出）	直接提示注入	限制不良行为、输入验证、在模型内部实施进一步控制措施
		间接提示注入	限制不良行为、输入验证、输入隔离
		规避攻击（例如对抗样本）	限制不良行为、监控、速率限制、模型访问控制，以及： 检测异常输入、检测对抗性输入、构建稳健的规避模型、训练对抗样本、输入失真、对抗鲁棒净化
	运行时 —— 侵入已部署模型	运行时模型投毒（重新编程）	限制不良行为、运行时模型完整性、运行时模型输入 / 输出完整性
	开发阶段 —— 工程环境	开发环境中的模型投毒	限制不良行为、开发环境安全、数据隔离、联邦学习、供应链管理，以及： 模型集成
		训练 / 微调数据的数据投毒	限制不良行为、开发环境安全、数据隔离、联邦学习、供应链管理，以及： 模型集成，以及： 增加训练数据量、数据质量控制、训练数据扭曲、抗投毒模型、训练对抗样本
	开发阶段 —— 供应链	应链模型中毒	限制不良行为， 供应商：开发环境安全、数据隔离、联邦学习 生产商：供应链管理，以及： 模型集成

AI通用控制-可用性控制

资产与影响	生命周期内的攻击面	威胁 / 风险类别	控制措施
模型行为可用性	模型使用	拒绝模型服务（耗尽模型资源）	监控、速率限制、模型访问控制，以及： 拒绝服务输入验证、资源限制

AI通用控制-非特定控制

资产与影响	生命周期内的攻击面	威胁 / 风险类别	控制措施
任何资产，保密性（C）、完整性（I）、可用性（A）	运行时 —— 所有信息技术环节	模型输出包含注入内容	对模型输出进行编码
任何资产，保密性（C）、完整性（I）、可用性（A）	运行时 —— 所有信息技术环节	针对常规资产的常规运行时安全攻击	常规运行时安全控制措施
任何资产，保密性（C）、完整性（I）、可用性（A）	运行时 —— 所有信息技术环节	针对常规供应链的常shekz规攻击	常规供应链管理控制措施

使用AI时的安全威胁的控制措施

逃避攻击

- 检测异常输入
- 检测对抗性输入
- 规避鲁棒性模型
- 对抗训练
- 输入失真或者微调
- 对抗鲁棒性净化

提示注入

- 提示输入验证
- 输入隔离

通过使用而泄露敏感数据

- 过滤敏感的模型输出结果
- 模糊置信度
- 小模式

通过使用而模型盗窃

- 管理控制

人工智能特定元素因使用而发生故障或失灵

- 拒绝服务输入验证
- 限制资源使用

开发AI时的安全威胁的控制措施

广义上的开发时模型中毒

- 模型集成，检测中毒
- 保护训练数据
- 更多的训练数据，数据稀释
- 数据质量控制
- 训练数据失真或者微调
- 抗毒的鲁棒性模型，抵抗有毒数据

开发阶段泄露敏感数据

- 开发阶段的数据保护

AI运行时的安全威胁的控制措施

非AI特定程序的安全威胁

- 安全运营

模型运行时投毒

- 模型运行时的完整性
- 模型运行时输入输出完整性

直接窃取运行时模型

- 运行模型时的保密性
- 模型混淆

不安全的输出处理

- 编码模型输出

泄露敏感的输入数据

- 模型输入保密性

NCSC/CISA指南对于AI安全的指导

安全设计

- 提高员工对风险和威胁的认识
- 对系统面临的威胁进行建模
- 设计系统的安全性以及功能和性能
- 选择 AI 模型时考虑平衡安全优势和全开发周期数据科学控制

安全开发

- 保护你的供应链
- 识别、跟踪和保护你的资产
- 记录你的数据、模型和提示
- 管理你的技术债务

安全部署

- 保护你的基础设施
- 持续保护你的模型
- 制定事件管理程序
- 负责任地发布AI
- 为用户提供便利

安全操作与维护

- 监控系统的行为
- 监控系统的输入
- 遵循安全的设计方法进行更新
- 收集与分享经验教训

Q & A