可 能 的 艺 术
THE ART OF THE POSSIBLE

第16届信息安全高级论坛

美国2024RSA参会热点研讨
INFORMATION SECURITY FORUM 2024

AI安全：从RSA 2024看未来趋势和行动

奇安信集团 邬怡

RSA 2024 AI安全议题概览

1. Responsible Adversarial Attacks on LLMs: 讨论了对大型语言模型（LLMs）进行负责任的对抗性攻击的问题。

2. Avoiding Common Design and Security Mistakes in Cloud AI/ML Environment: 探讨了在云AI/ML环境中避免常见设计和安全错误的策略。

3. What Cloud Has Taught about Securing AI in the Future: 分析了云技术对未来AI安全的影响和教训。

4. Navigating the AI Frontier - The Role of the CISO in GenAI Governance: 讨论了CISO（首席信息安全官）在通用人工智能（GenAI）治理中的角色。

5. Securing AI Apps with the OWASP Top 10 for Large Language Models: 使用OWASP（开放式Web应用程序安全项目）的前10名安全风险列表来保护大型语言模型。

6. AI at the Gates - Combating AI-Driven Assaults on the Customer Experience: 讨论了如何对抗AI驱动的对客户体验的攻击。

7. Securing AI - There Is No Try, Only Do: 强调了在AI安全方面没有尝试，只有行动的重要性。

8. How Large Language Models Are Reshaping the Cybersecurity Landscape: 探讨了大型语言模型如何重塑网络安全领域。

9. AI Law, Policy, and Common Sense Suggestions to Stay Out of Trouble: 提供了关于AI法律、政策以及避免麻烦的常识性建议。

10. Building AI Security in MLSecOps in Practice: 讨论了在实践中构建AI安全的MLSecOps（机器学习安全运维）。

11. Balancing Accessibility, Security, and AI Design - Inclusive Security Tools: 讨论了在可访问性、安全性和AI设计之间找到平衡的包容性安全工具。

12. Advancing AI Security With Insights From The World's Largest AI Red Team: 利用世界上最大AI红队的洞察来推进AI安全。

13. Securing and Governing Generative AI - Learnings from Microsoft: 从微软的经验中学习如何保护和治理生成性AI。

14. Innovate Now, Secure Later - Decisions, Decisions: 讨论了创新与安全之间的决策问题。

15. AI Governance - The Security Perspective: 从安全的角度讨论AI治理。

16. How to Safely Deploy AI Copilots: 探讨了如何安全地部署AI助手。

17. AI in Cyber - Is the Cyber Profession Ready for Its Impact: 讨论了网络安全专业是否准备好应对AI的影响。

18. AI-equipped Threat Actors Versus AI-enhanced Cyber Tools - Who Wins: 分析了装备AI的威胁行为者与增强AI的网络安全工具之间的竞争。

19. IP Protection and Privacy in LLM - Leveraging Fully Homomorphic Encryption: 讨论了在LLM中利用完全同态加密进行知识产权保护和隐私保护。

20. A Step-by-Step Guide to Securing Large Language Models (LLMs): 提供了保护大型语言模型的逐步指南。

21. Lessons Learned from Developing Secure AI Workflows: 从开发安全的AI工作流程中学到的经验。

22. Creating an AI Security and Incident Response Team: 如何创建AI安全和事件响应团队。

23. Securing Software Supply Chain - Problems, Solutions, and AI/ML Challenges: 讨论了保护软件供应链的问题、解决方案以及AI/ML面临的挑战。

24. Secure AI Transformation - What We Can Do Now and in the Future: 讨论了现在和未来可以采取的安全AI转型措施。

RSA2024集中讨论了人工智能（AI）和大型语言模型（LLM）在安全、治理、伦理、法规遵从以及技术创新方面的挑战和机遇，并强调了在这些领域内建立框架、策略和最佳实践的重要性。

AI带来的商业转型与安全挑战

AI is shifting business today

Every **individual**

Every **team**

Every **industry**

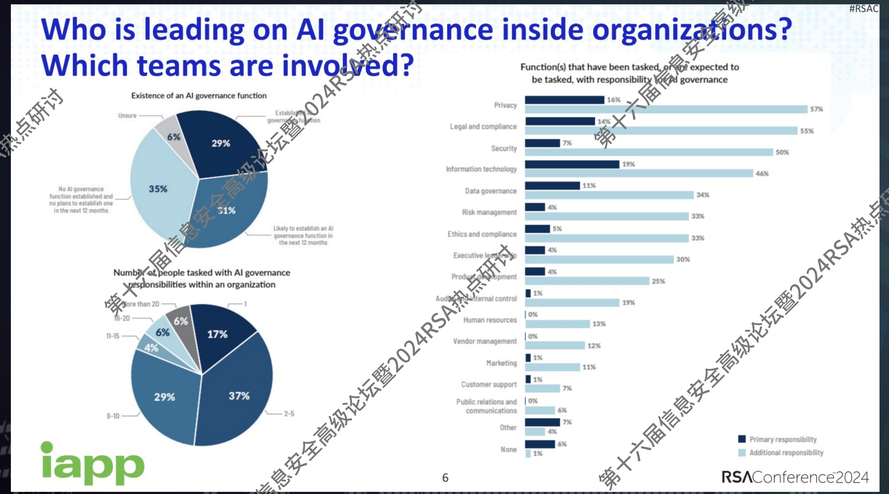
— Efim Hudis, VP Microsoft Security
Secure AI transformation: What we can do now and in the future

AI正在从各个方面重塑商业形态

Generative AI operates in a unique manner

- GenAI apps are a blackbox, versatile, probabilistic, and not deterministic
- Have high connectivity and autonomy
- Use natural language and can be manipulated

AI治理框架的必要性

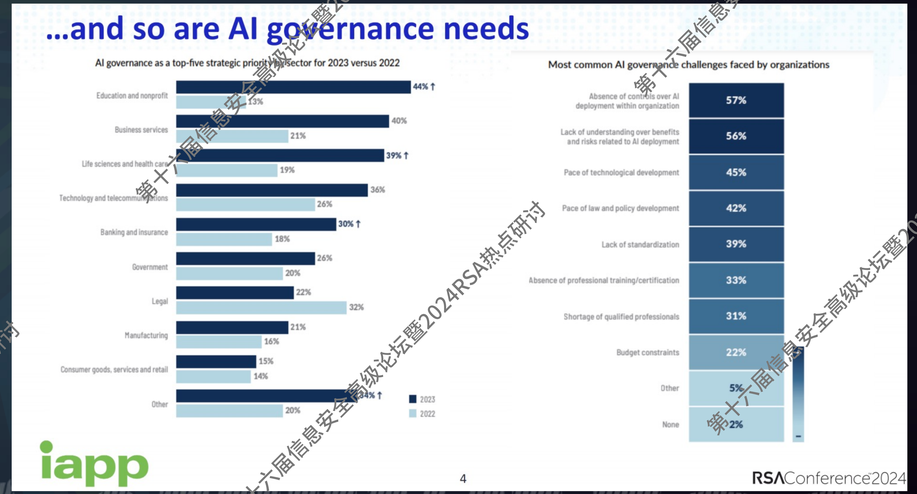


治理优先

AI安全和数据安全、网络安全发展所经历的过程类似，建立治理体系是组织的优先事务。

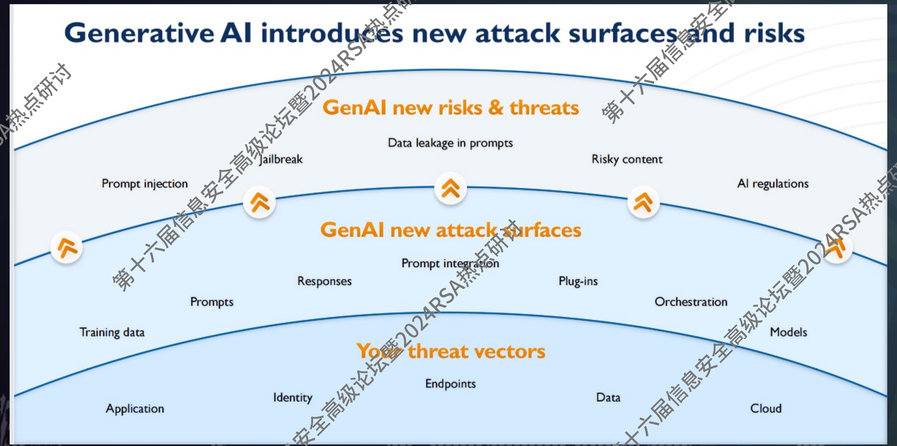
跨职能组织

AI治理需要建立起由隐私、法律和合规、安全及信息技术等部门的跨职能组织来承担相应责任。



— Caitlin Fennessy, VP & Chief Knowledge Officer, International Association of Privacy Professionals (IAPP)
AI Governance: The Security Perspective

新兴的AI攻击面



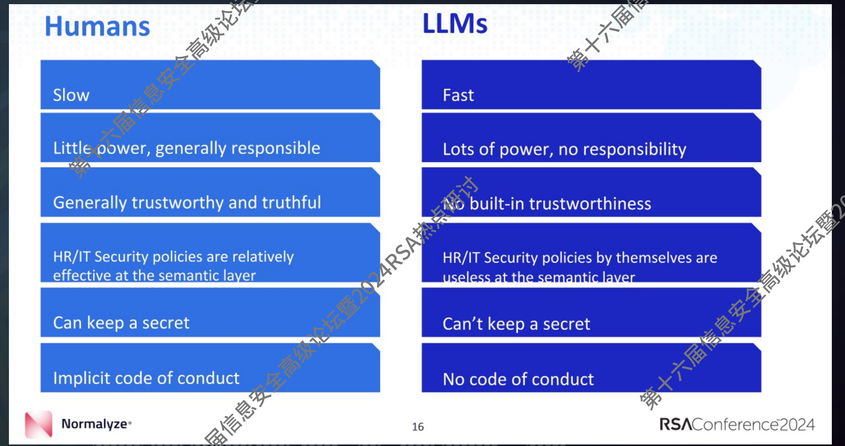
— Efim Hudis, VP Microsoft Security
Secure AI transformation: What we can do now and in the future

不同以往的攻击面

从训练数据到云基础设施，AI系统的每个组成部分都可能成为攻击的目标。

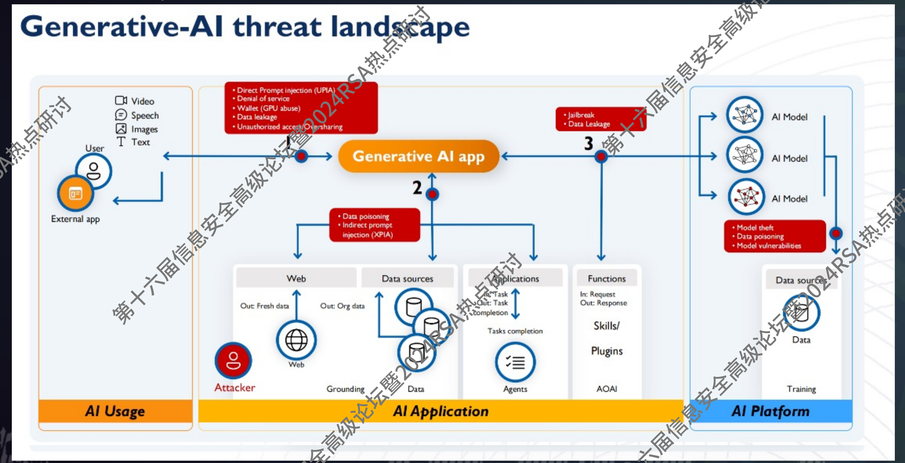
攻击的同构性

LLM在架构分层中更类似人的角色，对人的社工、诈骗手段也能用在对生成式AI攻击中。



Ravi Ithal,Founder and CTO,Normalyze
A Step-by-step Guide to Securing Large Language Models (LLMs)

生成式AI威胁全景

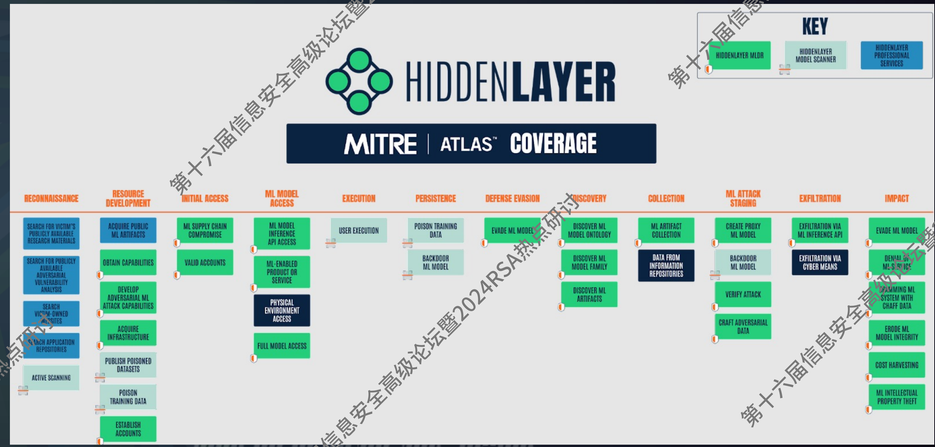


多样性威胁

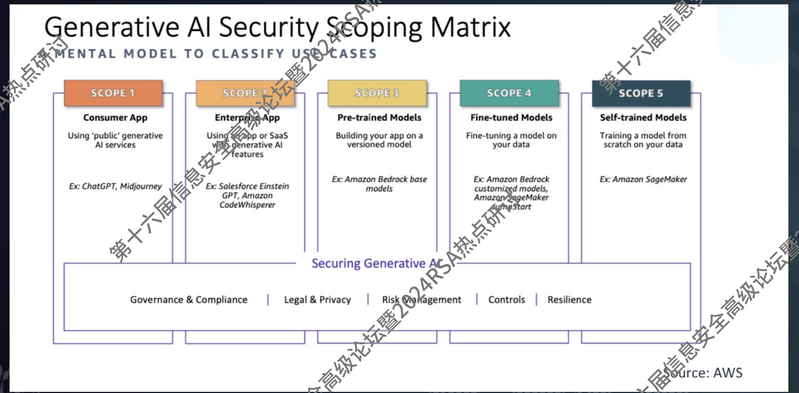
生成式AI安全威胁覆盖交互、应用、平台各层面。

多层次防护

针对生成式AI的复杂的攻击路径和手段需要全面的、多层次的防护措施。



生成式AI生命周期安全

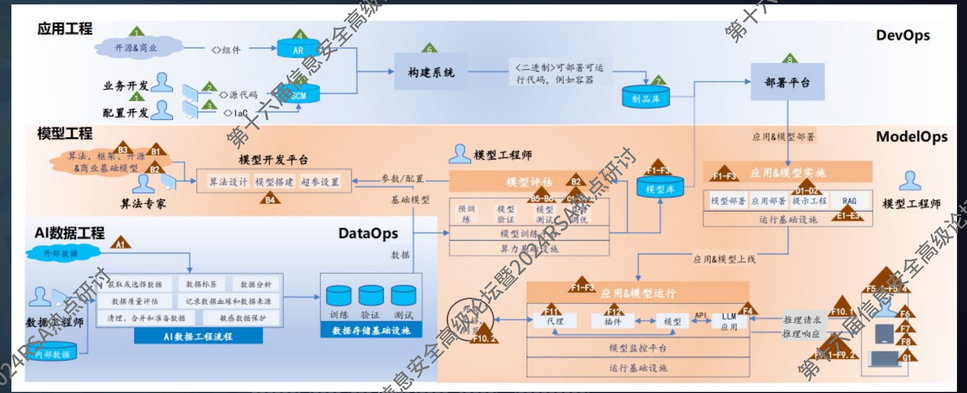


从模型到消费者

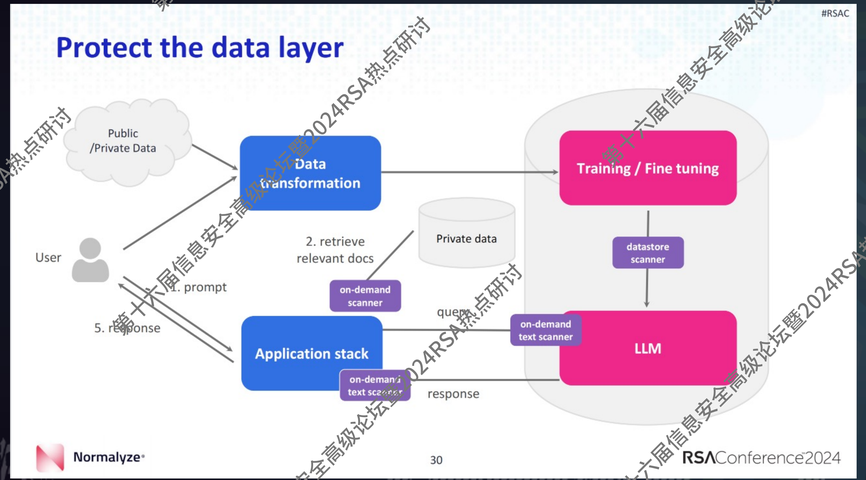
生成式AI的安全需求可以根据应用场景细分为多个层次，包括消费者应用、企业应用、预训练模型、微调模型和自我训练模型。每个层次都有特定的安全需求和风险管理策略。

数据和模型的管理至关重要

数据工程师负责数据的质量与安全，模型工程师则专注于模型的开发和优化。通过DataOps和ModelOps平台，确保生成式AI系统的稳定性和安全性。



生成式AI数据安全性



Ravi Ithal,Founder and CTO,Normalyze
A Step-by-step Guide to Securing Large Language Models (LLMs)

理解和控制数据接口

将大语言模型（LLM）视为黑箱，重点关注其数据接口的安全。识别并保护这些接口，以确保数据在传输和处理过程中的隐私和保密性。

防御对抗性提示和提示注入

LLM的自然语言理解和指令跟随功能使其容易受到对抗性提示和提示注入攻击。这类攻击可能导致敏感数据泄露或模型行为偏差。因此，需要特别关注防御这类攻击的方法。

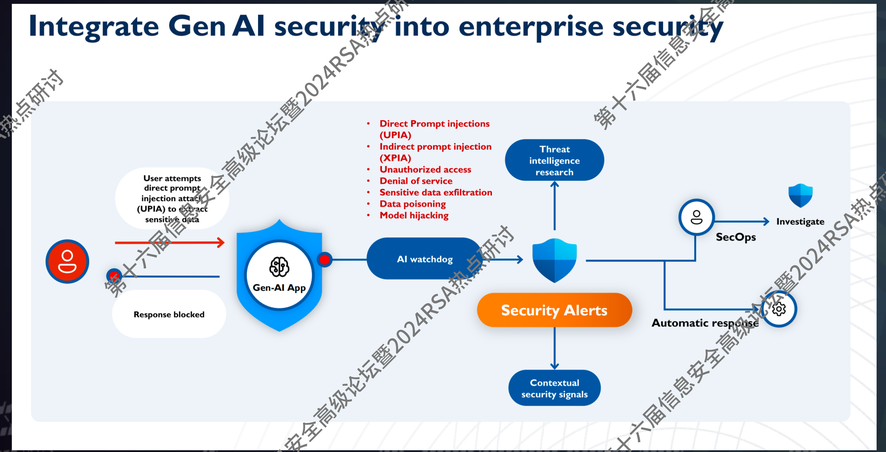
建立LLM责任机制

与人类不同，LLM没有责任感和可信度。因此，需要建立机制来模拟人类的行为规范和责任感，确保LLM在处理敏感数据时能够遵循一定的安全政策和行为准则。

发现影子LLM

发现组织内所有AI服务，追踪资源和数据源，建立LLM应用清单，进行审计和合规性检查。

集成生成式AI安全



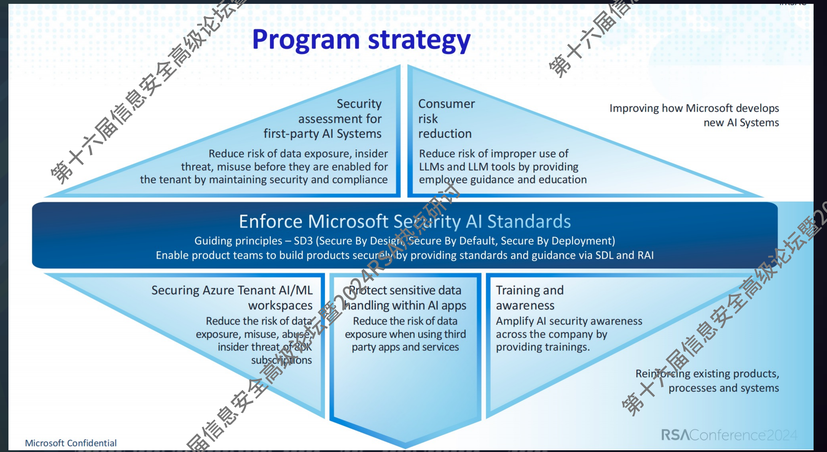
— Efim Hudis, VP Microsoft Security
Secure AI transformation: What we can do now and in the future

全面集成AI安全标准

通过实施SD3（Secure By Design, Secure By Default, Secure By Deployment）指导原则，将AI安全标准融入企业的整体安全策略，确保所有产品团队在开发和部署AI系统时遵循统一的安全标准和流程。

全面集成AI安全运营流程

将生成式AI安全无缝集成到企业安全框架中的策略，通过实时监控和威胁检测、利用威胁情报研究和自动化响应机制，确保AI应用的安全性和企业整体安全。



— Brian Fielder, VPSecurity, Microsoft
Securing and Governing Generative AI: Learnings from Microsoft

AI安全后续行动建议

立即行动

- 评估现有AI系统的安全性
- 建立临时控制措施
- 监控AI活动
- 审查和更新安全政策

90天内

- 建立AI治理团队和计划
- 制定和实施安全控制
- 安全评估和合规性检查
- 员工培训和意识提升

90天后

- 持续监控和评估
- 完善AI安全策略
- 推动AI安全技术的发展
- 建立应急响应计划

可 能 的 艺 术

THE ART OF THE POSSIBLE

第16届信息安全高级论坛

美国2024RSA参会热点研讨

INFORMATION SECURITY FORUM 2024

感谢聆听！