OWASP | TOP 10 LLM APPLICATIONS & GENERATIVE AI

LLM and GenAI Security Center of Excellence Guide

Secure AI Adoption Initiative

Version 1.0
October, 2024

Revision History

Revision	Date	Authors	Description
.1	5/15/2024	Scott Clinton	Initial Outline Draft
.2	7/2/2024	Scott Clinton, Sandy Dunn, Team	Updated with initial comments
.5	7/10/2024	Open Feedback and Comment	Early draft, open for comment and input
1.0rc	10/1/2024	Scott Clinton, Contributing and Review Teams - See Acknowledgements	First Release incorporating all feedback as of 9/27/24

License and Usage

Contents

Introduction

As generative AI technologies evolve and integrate into various aspects of business and society, the need for robust governance, security, and policy management becomes paramount. Establishing a Center of Excellence (COE) for Generative AI Security aims to bring together diverse groups such as security, legal, data science, operations, and end-users to foster collaboration, develop best practices, and ensure safe, efficient deployment of AI capabilities.

Who is this for?

This document is for CISO security teams and cross-functional leadership to gain an understanding and best practices framework to assist in educating teams on implementing their center of excellence for LLM and Generative AI Application security and adoption.

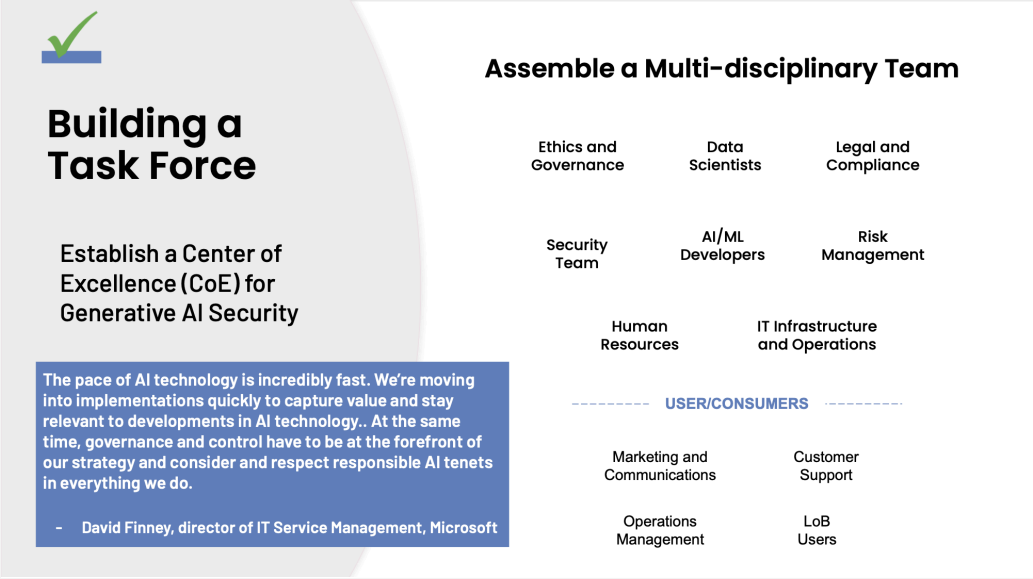
Objective

The primary objective of the COE is to develop and enforce security policies and protocols for generative AI applications, facilitate cross-departmental collaboration to harness expertise from various fields, educate and train teams on the ethical and secure use of generative AI technologies, and serve as an advisory body for AI-related projects and initiatives within the organization.

Creating a COE for Generative AI Security is a critical step toward ensuring that generative AI technologies are developed, deployed, and maintained securely and ethically. Through effective collaboration and governance, the COE will play a pivotal role in shaping the future of AI within the organization.

Creating the COE Structure

Creating an AI Security center of excellence (COE) for managing and securing generative AI applications involves strategic planning and collaboration across multiple departments.



The leadership team is comprised of heads from security, legal, data science, and operations. This team is responsible for making strategic decisions and directing the COE. Specific working groups focused on key areas such as Security and Compliance, Legal and Regulatory Affairs, Data Management and Analytics, Operational Integration, and End-User Engagement.

The COE's structure should be flexible and adaptable, evolving alongside the fast-paced advancements in AI technologies. Regular reviews of roles, responsibilities, and processes are essential to ensure the COE remains effective and aligned with the organization's strategic objectives.

The overall group is tasked with addressing specific challenges and delivering solutions relevant to their expertise, which include:

- Policy Development: Creating security policies tailored to generative AI.
- Risk Assessment and Management: Identify, evaluate and monitor potential risks and develop mitigation strategies.
- Training and Awareness: Conduct regular training sessions and workshops to inform all stakeholders about best practices and emerging threats.
- Research and Development: Stay abreast of the latest developments in AI and security to continuously refine strategies and tools.
- Stakeholder Engagement: Regularly involve end-users and other stakeholders in decision-making to ensure the COE's initiatives align with user needs and organizational goals.

COE Considerations

Leveraging a multidisciplinary team brings together diverse skills and perspectives, which are vital for addressing the complex security challenges of AI. But establishing a multidisciplinary team within a COE presents several challenges that need careful management. In addition to addressing these challenges and enhancing the COE's capabilities, leveraging external expertise can be highly beneficial. With diverse expertise and backgrounds, integrating professionals from security, legal, data science, and operations can lead to conflicts due to differing priorities and perspectives. Aligning these diverse viewpoints towards common objectives is critical.

Some Examples of Internal Challenges

- Communication Barriers: Effective communication among team members with varied professional languages and methodologies can be difficult. Establishing a common language or set of terminologies is essential for seamless collaboration.
- Resistance to Change: Individuals from different departments may resist new workflows or changes that disrupt traditional processes. Managing change effectively and ensuring buy-in from all stakeholders is crucial.
- Resource Allocation: Competing for resources among different departments can create friction. Transparent and equitable resource distribution policies need to be established.
- Skill Gaps: As generative AI is a relatively new and rapidly evolving area, there may be significant gaps in the necessary skills among existing staff, which can hinder the COE's effectiveness.

Utilizing External Expertise

Leveraging external expertise can prove crucial to enhancing the capabilities of the Center of Excellence (COE) and addressing its challenges. Engaging consultants and advisors specializing in AI security, legal regulations related to AI, and data ethics can provide the COE with critical insights and guidance, helping shape effective policies and procedures. Additionally, partnering with external training providers can address skill gaps within the team by offering specialized training in AI and cybersecurity, ensuring that all members are proficient in the latest technologies and industry best practices.

Collaborating with technology partners, such as tech companies and vendors, offers COE access to advanced tools and platforms that boost operational capabilities. Furthermore, forming partnerships with academic and research institutions facilitates continuous learning and helps the COE stay abreast of new developments in the field of generative AI.

Engaging with industry groups and networks also aids in understanding broader trends, gathering insights from similar initiatives, and adopting industry-wide best practices, all of which contribute to the strategic growth and effectiveness of the COE.

Setting Overall Objectives and KPIs - Top 5

Establishing a multi-disciplinary Center of Excellence (COE) for trustworthy and secure generative AI adoption requires clear objectives and measurable Key Performance Indicators (KPIs). Here are top five examples of Objectives and Key Results (OKRs), along with their corresponding KPIs that can serve as a starting point for building an operating plan for your COE. Each objective should be supported by well-defined KPIs that assess compliance, security performance, and their impact on overall business operations. This ensures that AI security initiatives are not only technically sound but also aligned with the organization's strategic priorities.

Objective 1: Enhance Security Frameworks for Generative AI

This objective focuses on developing and implementing comprehensive security policies tailored to the unique needs of generative AI applications. Full compliance with national and international regulations and reducing the incidence of security breaches are critical. By strengthening the security frameworks, the COE aims to protect AI systems against evolving threats and vulnerabilities, ensuring robust defense mechanisms are in place.

Example OKRs	Example KPIs
Develop new (or refine existing) and implement comprehensive security policies specific to generative AI applications.	The number of security policies developed (or refined) and implemented.
Achieve full compliance with national and international regulations concerning AI security and data privacy	The compliance rate with relevant regulations.
Reduce the incidence of security breaches related to AI by X% within the next year.	Frequency and severity of security breaches involving AI technologies.

Objective 2: Foster Collaboration and Knowledge Transfer

The goal here is to enhance synergy among various departments within the COE through regular workshops, training sessions, and a centralized knowledge base. This objective seeks to improve team communication and collaboration, fostering an environment where shared resources and collective expertise contribute to innovative security solutions and efficient problem-solving.

Example OKRs	Example KPIs
Establish regular workshops and training sessions for all departments involved in the COE.	The number of interdepartmental seminars and training sessions conducted.

Create a centralized knowledge base accessible to all COE members, containing up-to-date information on AI security trends and best practices.	Utilization rate of the knowledge base by COE members.
Achieve an X% increase in cross-departmental projects focused on enhancing AI security within the next year.	The number of cross-departmental initiatives launched.

Objective 3: Build Trust and Transparency with Stakeholders

Building trust and maintaining transparency are pivotal in this objective. By developing and releasing quarterly reports and conducting bi-annual stakeholder meetings, the COE aims to engage openly with all stakeholders, ensuring they are well informed about the organization's AI security initiatives. This approach helps to cultivate a relationship of trust and fosters greater acceptance and support for AI technologies.

Example OKRs	**Example KPIs**
Develop and release quarterly reports on the organization's AI security status and initiatives.	The number of quarterly reports published. The total count of new AI security initiatives that are developed and implemented each quarter.
Conduct bi-annual stakeholder meetings to gather feedback and discuss concerns regarding AI security.	The number of stakeholder meetings held. The number of concerns gathered and documented during held meetings.
Implement and track a stakeholder engagement program aimed at improving transparency and trust.	Engagement levels and feedback scores from stakeholders and public surveys. Percentage of invited stakeholders who attend the meetings.

Objective 4: Advance Ethical AI and Security Practices

This objective aims to embed ethical considerations deeply within AI operations, ensuring a solid ethical framework guides all AI deployments. The COE commits to promoting responsible AI usage that aligns with security guidelines and organizational ethics by developing ethical guidelines, conducting annual audits, and increasing awareness of these standards.

Example OKRs	**Example KPIs**
Develop and enforce a set of ethical and security guidelines tailored for generative AI use within the	The number of ethical guidelines developed and implemented.

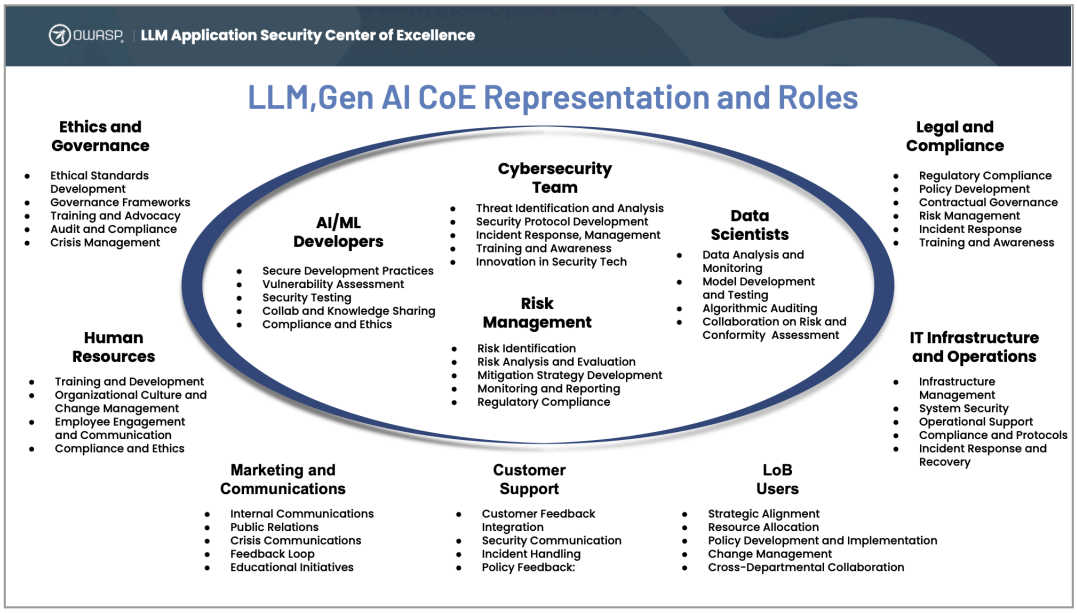
organization.	The percentage of employees who have completed training programs on the proposed/developed ethical and security guidelines.
Conduct an annual audit to ensure adherence to ethical AI and security standards and practices.	Results of the ethics and security audit, detailing compliance and areas for improvement.
Achieve an X% increase in awareness and understanding of ethical AI principles among employees by the end of the year.	Increase in employee scores on AI security and ethical AI awareness assessments.

Objective 5: Optimize AI Performance and Reliability

Optimizing the performance and reliability of AI systems is crucial to this objective. It involves enhancing AI system accuracy, reducing error rates, and ensuring systems can operate effectively under various conditions. The COE aims to deliver high-performing AI applications that stakeholders can consistently rely on by deploying tools and updates that improve system efficiency.

Example OKRs	**Example KPIs**
Reduce error rates in AI systems by 25% over the next 12 months.	Reduction percentage in error rates of AI and AI-based applications.
Implement a system for continuous monitoring and real-time analysis of AI system performance.	Implementation status of the continuous monitoring system.
Develop and deploy at least two new tools or updates that improve the efficiency and reliability of AI operations.	Number and impact of tools or updates deployed for improving AI performance.

Working Group Roles and Responsibilities



Builders and Operating Groups

AI and ML Developers

Including AI and ML Developers in the COE is crucial for bridging the gap between theoretical security measures and practical, actionable AI implementations. Their expertise ensures that security is embedded in the development phase, enhancing the robustness of AI applications from the ground up.

With the adoption of shift-left strategies for application security and ownership applies to AI and ML developers as well. Involving AI and ML developer representation helps to ensure that security is a foundational element of AI applications, leading to more secure solutions. Proactive risk identification and mitigation by developers reduce the likelihood of security breaches and data leaks, thereby protecting the organization and its customers.

Developers also contribute technical innovations that enhance AI security measures, resulting in more robust systems. Ensuring that development practices comply with regulatory requirements prevents legal issues and builds trust with regulators and stakeholders. Additionally, developers integrate security practices across different functional areas of AI projects, aligning and informing all team members. Developers should have access to the latest security tools and training to seamlessly integrate security features into the AI development process. This proactive strategy allows for the early detection and mitigation of potential vulnerabilities throughout the development lifecycle.

COE Responsibilities - AI and ML Developers	
Secure Development Practices	Implement and adhere to secure coding practices tailored explicitly to AI and ML projects. This includes using secure frameworks, regular code reviews, and integrating security at each stage of the development lifecycle.
Vulnerability Assessment	Proactively identify and address potential vulnerabilities within AI algorithms and data processing methods. This includes performing threat modeling and risk assessments during the early stages of development.
Security Testing	Regularly conduct security tests on AI models and applications, including unit testing, to ensure they can withstand attacks and perform reliably under adverse conditions. This may include penetration testing, stress testing, and scenario-based testing.
Collaboration and Knowledge Sharing	Share technical knowledge and insights that can aid non-technical team members. Work closely with other departments within the COE, such as Risk Management and Data Science, to ensure a holistic approach to AI security
Innovation and Research	Participate in research efforts to develop new security features or enhance existing ones.
Compliance and Ethics	Ensure that all AI development is compliant with relevant laws and ethical guidelines. This includes the responsible use of data, transparency in AI operations, and the mitigation of biases in AI models.
Documentation and Reporting	Maintain comprehensive documentation of all AI development processes, security measures, and testing results. This documentation is crucial for audit purposes and sharing COE best practices.

Cybersecurity Team

Incorporating the Cybersecurity team in the COE is essential to ensure that AI technologies are protected from emerging threats and vulnerabilities. Their expertise in digital security forms a critical backbone for developing, deploying, and maintaining AI systems that are secure and resilient against cyber threats including Ai-aided phishing, deepfake, and emerging AI-aided threats. Strong cybersecurity practices are essential for building trust among users and stakeholders, making AI technologies reliable and secure.

The cybersecurity team's expertise is vital in protecting AI systems from sophisticated attacks and potential exploits, ensuring compliance with relevant laws and regulations, safeguarding the organization from legal issues and enhancing its reputation.

Additionally, effective incident response and management maintain the operational integrity of AI systems, enabling them to function effectively even during attacks.

The cybersecurity team should focus on implementing automated monitoring systems that provide real-time alerts for potential threats, enabling rapid response and minimizing the risk of incident escalation. The cybersecurity team also integrates security practices across all aspects of AI development and deployment, creating a unified security framework within the organization.

COE Responsibilities - Cybersecurity Teams	
Threat Identification and Analysis	Continuously identify and analyze potential cyber threats to AI systems. This includes monitoring for new vulnerabilities, predicting possible attack vectors, and understanding the implications of these threats on AI operations.
Security Protocol Development	Develop robust security protocols tailored explicitly to AI and ML technologies. This involves crafting customized solutions to protect data integrity, ensure privacy, and safeguard AI systems against unauthorized access.
Incident Response and Management	Establish and manage a rapid response framework for any security incidents involving AI systems. This framework should include real-time threat detection, containment strategies, and recovery plans to minimize disruption and damage.
Security Testing and Audits	Conduct comprehensive security testing and audits of AI systems to validate the effectiveness of security measures. This involves penetration testing, security assessments, and compliance checks against industry standards.
Training and Awareness Programs	Develop and conduct comprehensive security testing and audits of AI systems to validate the effectiveness of security measures. This involves penetration testing, security assessments, and compliance checks against industry standards.
Bug Bounty Program Management	Register, implement, and manage a bug bounty program on vulnerability reporting platforms to proactively identify and remediate vulnerabilities in AI systems with the support of external security researchers.
Collaboration with Regulatory Bodies	Engage with regulatory bodies to ensure compliance with national and international cybersecurity regulations. This includes adapting AI security practices to meet evolving legal and regulatory requirements.
Innovation in Security Technologies	Stay abreast of the latest developments in cybersecurity technology and integrate cutting-edge solutions into the organization's AI systems to enhance its security posture.

IT and Operations

The IT and Operations team is crucial for ensuring the technical infrastructure and operational procedures support AI systems' secure development, deployment, and maintenance. Their expertise in managing technology and operational workflows is essential for the smooth functioning of AI security initiatives.

Reliable IT operations are crucial for the continuous functioning and availability of AI systems, especially in critical environments. Robust IT and operational practices are essential for AI security, protecting against external attacks and internal vulnerabilities. Efficient management of technological resources ensures that AI systems remain secure, cost-effective, and scalable. A well-supported IT infrastructure fosters innovation, enabling rapid adaptation of AI technologies in response to evolving security landscapes. Strong operational capabilities are also vital for effective crisis management, minimizing downtime and mitigating the impact of security incidents or system failures.

COE Responsibilities - IT and Operations	
Infrastructure Management	Design, implement, and maintain the technical infrastructure necessary for AI applications. This includes securing databases, networks, and cloud environments where AI systems operate.
System Security	Implement and oversee security measures, including firewalls, intrusion detection systems, and encryption protocols, for IT systems interacting with AI technologies.
Operational Support	Provide ongoing operational support for AI projects, ensuring all systems function smoothly and efficiently. This includes troubleshooting, system upgrades, and performance optimization.
Compliance and Protocols	Conduct comprehensive security testing and audits of AI systems to validate the effectiveness of security measures. This involves penetration testing, security assessments, and compliance checks against industry standards.
Incident Response and Recovery	Develop and conduct comprehensive security testing and audits of AI systems to validate the effectiveness of security measures. This involves penetration testing, security assessments, and compliance checks against industry standards.
Collaboration and Communication	Facilitate communication between the technical teams and other IT and ops departments to help ensure that comprehensive security insights inform operational decisions.

Legal and Compliance

Legal and compliance teams play a critical role in the COE by ensuring that all activities related to generative AI adhere to existing laws and regulations while proactively addressing emerging legal challenges.

Legal and compliance teams are essential for protecting organizational interests by ensuring AI applications comply with legal requirements and ethical standards, avoiding potential lawsuits and penalties. Their focus on compliance and ethical practices builds trust among users, stakeholders, and regulators, which is crucial for the widespread adoption of AI technologies. Additionally, their oversight fosters innovation by providing clear guidelines that allow developers and data scientists to safely and responsibly explore new AI applications. As AI technology and related laws evolve, these teams ensure continuous compliance by updating policies and practices in real-time, preventing obsolescence and maintaining the organization's adaptability.

COE Responsibilities - Legal and Compliance	
Regulatory Compliance	Ensure that all generative AI initiatives comply with local, national, and international regulations, such as AI related laws (EU AI Act, California SB 1047), data privacy laws (GDPR, CCPA), intellectual property rights, and industry-specific guidelines.
Policy Development	Assist in drafting and reviewing policies that govern the organization's development, deployment, and use of generative AI. This includes creating frameworks that ensure ethical AI usage and protect the organization against legal risks.
Contractual Governance	Oversee and manage the legal aspects of contracts and agreements with third parties, including vendors and partners, to ensure that these agreements incorporate adequate safeguards for data security and intellectual property rights.
Risk Management	Identify legal risks associated with deploying generative AI technologies and develop strategies to mitigate these risks. This involves regular audits and compliance checks.
Incident Response	Develop and implement protocols for legal responses to security breaches or compliance failures related to AI technologies. This includes coordinating with regulatory bodies as necessary.
Training and Awareness	Co-create and lead training sessions and materials to educate the COE members and other stakeholders about legal considerations, compliance requirements, and ethical AI use.

Ethics and Governance

Not all organizations have a dedicated ethics and governance team. It may be made up of legal, risk management, operations, and business groups. However, having this type of function is vital to ensuring that the deployment and use of generative AI within the organization align with ethical standards and corporate governance. This team helps bridge the gap between technological advancements and moral considerations, fostering responsible AI development and usage.

By proactively addressing ethical issues and ensuring strong governance, the team mitigates risks that could lead to reputational damage, legal challenges, or financial losses. Ethical governance also supports innovation by providing clear guidelines and frameworks that foster creativity while ensuring responsible development. The team ensures compliance with new laws and standards as AI regulations evolve, preventing legal repercussions. Furthermore, ethical guidelines and governance structures enhance decision-making processes by considering the broader impacts of AI technologies on society and the environment.

COE Responsibilities - Ethics and Governance	
Ethical Standards Development	Develop comprehensive guidelines and standards for ethical AI usage that align with the organization's values and the expectations of wider society. This includes addressing fairness, accountability, transparency, and privacy concerns.
Governance Frameworks	Create and enforce governance frameworks that oversee the ethical implementation of AI technologies. These frameworks help manage AI projects, ensuring they adhere to established ethical guidelines and business objectives.
Policy Integration	Work closely with the legal, compliance, and policy development teams to ensure that ethical considerations are integrated into all AI-related policies and procedures.
Training and Advocacy	Provide ongoing education and training for employees about ethical AI practices. Promote a culture of ethical awareness and understanding across the organization.
Audit and Compliance	Conduct regular audits to ensure adherence to ethical standards and practices. This involves reviewing AI projects and initiatives to identify potential ethical risks and governance issues.
Crisis Management	Develop protocols to handle ethical dilemmas and governance breaches effectively. This includes establishing procedures for escalation, investigation, and resolution of ethical issues in AI projects.

Human Resources

The Human Resources (HR) team plays a pivotal role in supporting the COE by managing the workforce aspects of AI security initiatives. Their involvement is crucial for recruiting, training, and maintaining an effective team aligned with the ethical and operational standards required for secure AI deployment.

The HR team ensures that the human capital strategy aligns with the technical and ethical goals of the Center of Excellence (COE), fostering a cohesive approach to AI security. They manage workforce adaptability by focusing on continuous education and training, ensuring employees remain resilient as AI technologies evolve. HR is also key to maintaining an ethical culture that values security and compliance, which is essential for the success of AI initiatives. Additionally, HR helps navigate the complexities of employment law related to AI, addressing intellectual property issues and new types of worker rights and protections.

COE Responsibilities - Human Resources	
Training and Development	Develop and implement training programs to enhance the skills of COE members and other employees involved in AI projects. This includes specialized training in AI ethics, security practices, and compliance with regulatory requirements.
Organizational Culture and Change Management	Develop and implement training programs to enhance the skills of COE members and other employees involved in AI projects. This includes specialized training in AI ethics, security practices, and compliance with regulatory requirements.
Employee Engagement and Communication	Keep the workforce informed and engaged with the organization's AI strategies and projects. HR manages internal communications to ensure employees understand their roles in supporting AI initiatives and the importance of security and compliance.
Compliance and Ethics	Work alongside the legal, ethics, and governance teams to ensure all aspects of AI development and deployment are conducted ethically and in compliance with labor laws and regulations.

Risk Management

The Risk Management team is essential in identifying, analyzing, mitigating, and monitoring risks associated with deploying and using generative AI technologies. Their expertise ensures that potential security, privacy, and operations threats are proactively managed to protect the organization and its stakeholders.

The Risk Management team is vital in ensuring smooth and secure AI operations by proactively identifying and addressing risks early, helping the organization avoid costly and disruptive issues. Comprehensive risk assessments enhance decision-making by providing necessary information for informed choices about AI strategies and projects. Staying ahead of compliance helps the organization avoid legal troubles and align with industry standards, crucial in the dynamic AI regulatory landscape.

Effective risk management also protects the organization's reputation by demonstrating a commitment to security and ethical responsibility. Additionally, by mitigating risks that could lead to financial losses through fines, downtime, or compromised data, the team plays a crucial role in safeguarding the organization's financial stability.

COE Responsibilities - Risk Management	
Risk Identification	Systematically identify business risks associated with AI technologies, including data breaches, misuse of AI applications, and financial, reputational, and compliance risks.
Risk Analysis and Evaluation	Assess the likelihood and impact of identified risks, categorizing them based on severity and potential damage. This analysis helps prioritize risk mitigation efforts.
Mitigation Strategy Development	Develop strategies and plans to reduce or eliminate risks. This includes the implementation of security protocols, the adoption of best practices in AI development, and the deployment of mitigation technologies.
Monitoring and Reporting	Continuously monitor risk factors and control measures to ensure their effectiveness. Regularly report to the COE and wider organization on risk status and improvement strategies.
Regulatory Compliance	Ensure that AI deployments comply with relevant laws and regulations, thereby avoiding legal penalties and reputational damage.
Stakeholder Communication	Communicate risk management processes and status to stakeholders, ensuring transparency and maintaining trust in the organization's AI initiatives.

Data Science Team

The Data Science team plays a critical role in the COE by leveraging their expertise in data analysis, machine learning, and statistical methods to enhance the security and integrity of AI systems. Their work is essential for identifying patterns, predicting potential threats, and informing security strategies.

They help to ensure the accuracy and integrity of data used by AI systems, which is vital for maintaining trust and reliability. By providing data-driven insights, the team supports more informed and effective decision-making across the Center of Excellence (COE), from policy creation to incident response. Their application of advanced analytical and machine learning techniques drives innovation within the COE, addressing complex security challenges.

Additionally, through detailed analysis and continuous monitoring, the Data Science team helps mitigate risks associated with AI system deployment and operation. The Data Science team can also play a key role in building predictive models that anticipate potential security breaches, enabling the COE to proactively manage AI-related risks.

COE Responsibilities - Data Science Team	
Data Analysis and Monitoring	Analyze large datasets to identify anomalies, trends, and potential security threats. Continuous data monitoring helps in the early detection of vulnerabilities within AI systems.
Model Development and Testing	Develop and refine machine learning models that can predict, detect, and respond to security threats. This includes creating models that ensure the integrity and confidentiality of data used and generated by AI systems.
Algorithmic Auditing	Regularly audit AI algorithms for accuracy, fairness, and potential biases, ensuring that they do not inadvertently compromise security or violate ethical standards.
Collaboration on Risk and Conformity Assessment	Work closely with the Risk Management and Cybersecurity teams to quantify risks and assess the potential impact of security threats based on data-driven insights.
Innovative Security Solutions	Leverage cutting-edge data science techniques and collaborate with Cybersecurity teams to develop innovative solutions for AI security, such as anomaly detection systems and automated threat intelligence.
Reporting and Documentation	Provide detailed reports and visualizations of data insights to stakeholders within the COE, helping them make informed decisions about AI security policies and strategies.

User and Consumer Groups

In the context of the Center of Excellence (COE) for Generative AI Security, it's essential to incorporate not just the foundational teams involved in building, integrating, and securing AI applications, and IT infrastructure, but also to engage the end-users of these systems deeply. This includes Line of Business leaders, marketing professionals, and support teams , for example, who interact with AI-driven technologies on a daily basis. These user groups bring critical user-centric insights that can significantly influence the effectiveness and security of AI applications.

Their real-world experiences provide essential feedback that can drive improvements in AI deployment and operational procedures, ensuring that the systems are not only robust and compliant with legal and risk management standards but also finely tuned to the specific needs and challenges of the business and better align with business goals.

Engaging a wide range of stakeholders enriches the development process and strengthens the governance and oversight of AI applications, making them safer, more effective, and more aligned with the organization's overall objectives.

Marketing and Communications

Marketing and Communications play two roles as part of the CoE. The first is to help ensure that the organization's AI security initiatives are effectively communicated internally and externally. This team plays a pivotal role in shaping the public perception and internal understanding of AI security strategies. Establishing policies and processes for crisis communication is particularly crucial, as it can significantly reduce reputational risks by swiftly addressing potential security issues before they escalate.

In addition, LLMs and generative AI are transforming marketing and communications by enabling advanced automation. LLMs assist in generating text for blogs and social updates, creating visual content, and enhancing customer interactions through chatbots and virtual assistants.

They also ensure that content adheres to brand and legal standards through rigorous checks. LLMs also play a crucial role in analyzing consumer data to refine marketing strategies and personalizing email content to boost engagement.

To ensure ethical use and data security, guardrails need to be established including plagiarism detection, adherence to privacy regulations, and mechanisms to prevent spam, ensuring that AI tools are used responsibly and effectively while maintaining brand integrity and consumer trust.

COE Responsibilities - Marketing and Communications	
Internal Communications	Facilitate transparent and ongoing communication within the organization regarding AI security policies, updates, and impacts. This will build an informed and engaged workforce.
Public Relations	Handle external communications to shape how stakeholders, including customers, partners, and regulators, perceive the organization's use of AI. This includes managing media relations and public announcements related to AI security.
Crisis Communications	Prepare and execute communication strategies for potential security breaches or controversies related to AI, ensuring that the organization maintains its credibility and effectively manages any negative impact.
Feedback Loop	Establish and maintain channels for feedback from both internal and external stakeholders, providing valuable insights that can influence AI security strategies and practices.
Educational Initiatives	Organize and promote educational campaigns and materials that help demystify AI security for non-technical employees and external audiences, enhancing overall awareness and understanding.

User/Usage policies and Guardrails - Marketing and Communications	
Data Analysis and Customer Insights	Adhere to GDPR and other relevant regulations to ensure data privacy and protection. Use anonymized data whenever possible, and ensure all data usage is transparent and with data subjects' consents or legitimate interests.
Email Marketing Automation	Monitor automated systems to prevent spamming and ensure all communications are relevant and valuable to recipients. Maintain an easy opt-out mechanism and respect user preferences to build trust and comply with anti-spam laws.
Customer Interaction	Regularly update and audit AI interactions to ensure they comply with privacy regulations and maintain professional and brand-appropriate communication. Set up protocols to escalate complex queries to human agents.
Content Generation	Implement policies to ensure originality and avoid plagiarism, and copyright infringement, for example apply detection tools and review all AI-generated content manually before publication.
Visual Content Creation	Implement checks to ensure that all visual content respects copyright laws and brand guidelines. Use approved image databases and have a clear policy for the use of trademarks to avoid infringement.

Customer Support

Once again, like marketing and communication teams, Including the Customer Support team is essential as it plays a dual role as a set of internal users and a feedback loop for ensuring that AI security measures align with customer expectations and enhance the customer experience. This group brings direct insights from customer interactions, crucial for shaping user-centric security solutions.

Application of LLM systems not only bolsters operational efficiency by enabling Customer Support to handle inquiries and issues more adeptly but also lightens the load on other departments. Additionally, by swiftly identifying and addressing security issues that customers encounter, Customer Support plays a crucial role in risk mitigation, preventing minor issues from evolving into more significant crises.

Generative AI significantly enhances the capabilities of customer support teams by automating responses, personalizing interactions, and analyzing customer feedback to improve service quality. Typical applications include AI-driven chatbots that provide 24/7 customer service, offering immediate responses to inquiries and resolving simple issues without human intervention. To ensure these tools are used securely and ethically, guardrails and usage policies are essential.

COE Responsibilities - Customer Support	
Customer Feedback Integration	Gather and relay customer feedback regarding AI applications and security measures, providing invaluable insights into customer needs and concerns.
Security Communication	Inform customers about the organization's AI security measures clearly and reassuringly, enhancing their trust and confidence in the products and services.
Incident Handling	Serve as the first point of contact for customers during security incidents involving AI systems. Ensure effective communication and resolution strategies that maintain customer trust and satisfaction.
Policy Feedback	Provide input on AI security policies from a customer interaction perspective, ensuring that these policies are practical and enhance customer satisfaction.
Reporting and Analysis	Monitor and report on customer issues related to AI security, using data to analyze trends that could indicate underlying security challenges.

User/Usage policies and Guardrails - Customer Support	
Data Privacy and Transparency	Implement strict access controls and encryption to protect customer data, ensuring compliance with regulations like GDPR. Inform customers before they are interacting with AI and provide an option to speak with a human representative if preferred, reinforcing trust and transparency.
Monitoring and Oversight	Continuously monitor AI interactions for quality assurance, and use human oversight to correct errors and refine responses. Maintain a high standard of customer service and establish guidelines to prevent biases in AI responses, and ensure fair treatment of all customers.
Feedback Loops	Incorporate mechanisms to capture customer feedback on AI interactions, allowing for ongoing improvement of AI systems based on real user experiences. These policies help safeguard intellectual property, maintain customer trust, and ensure the ethical use of AI in customer support operations.

Line of Business Representation

Incorporating Line of Business (LOB) Leadership into the COE ensures that AI security initiatives are closely aligned with the specific needs and strategic goals of different business units. This group brings crucial insights and strategic oversight, enabling the COE to tailor security measures effectively across diverse areas of the organization.

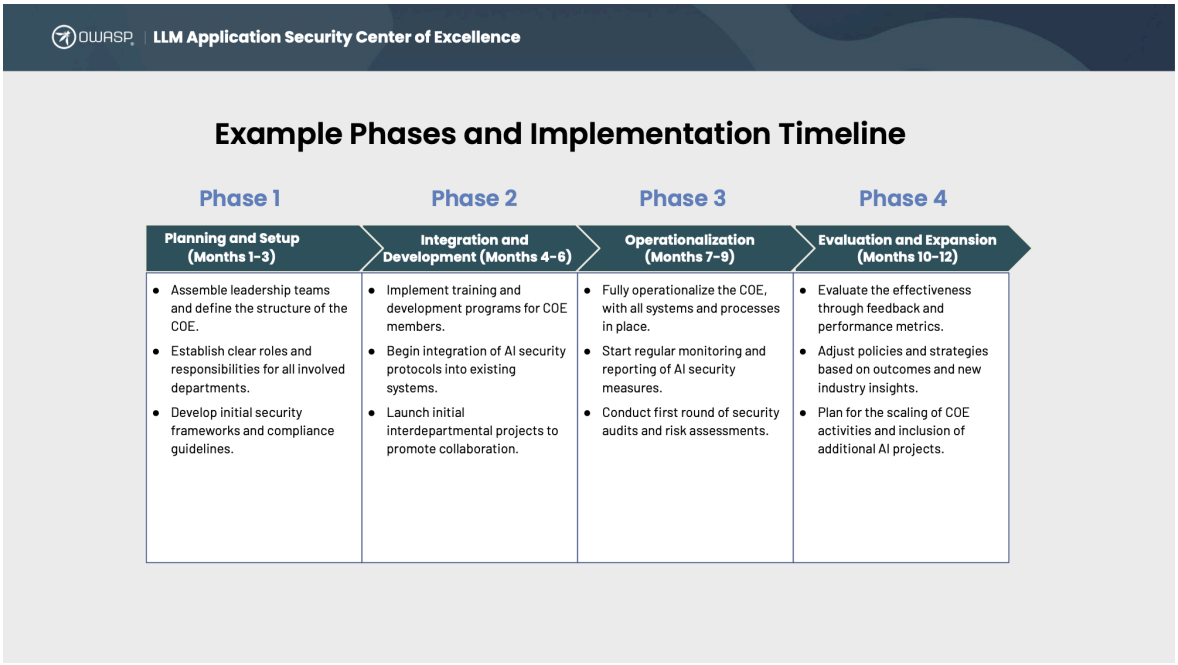
Including Lines of Business (LOB) ensures that AI security measures are not only technically robust but also finely tuned to specific business contexts, enhancing their effectiveness and ensuring smoother adoption across various departments. By embedding a business-centric approach to security, Their active involvement and advocacy promote a strong culture of security within their teams, underscoring the critical role of security in achieving business goals and maintaining operational continuity.

COE Responsibilities - LOBs	
Strategic Alignment	Ensure that AI security initiatives align with their respective lines' business objectives and strategies. This involves integrating security practices with business operations to enhance both security and business outcomes.
Resource Allocation	Allocate the necessary resources within their lines of business to support AI security initiatives. This includes budgeting for security tools, technologies, and training specific to their business needs.
Policy Development and Implementation	Participate in developing and implementing security policies that are tailored to the unique risks and requirements of different business areas. This ensures that policies are not only comprehensive but also practical and applicable.
Change Management	Lead change management efforts within their business units to ensure smooth adoption and integration of new security technologies and practices. This includes communicating the value and importance of these changes to team members.
Performance Metrics	Monitor and report on customer issues related to AI security, using data to analyze trends that could indicate underlying security challenges.
Risk Management	Collaborate with the Risk Management team to identify and address business-specific risks associated with AI technologies. This involves assessing the potential impacts on their LOB and developing strategies to mitigate these risks.
Cross-Departmental Collaboration	Facilitate collaboration between their line of business and other departments, such as IT, Data Science, and HR, to ensure that security measures are effectively implemented and supported across the organization.

Example Implementation Phases & Timeline

Every organization is different. However, taking a phased approach to developing and rolling out your COE can help ensure better alignment between functions, a clear setting of expectations, and a smooth transition into a group that provides actionable insights to help secure AI applications and ease adoption.

This example phased timeline is designed to provide a structured approach starting point based on previous implementation experience for setting up and operationalizing your COE.



This will help to ensure that each step is planned and executed with attention to detail, enabling the organization to effectively manage and secure its generative AI initiatives. The approach and framework will need to be tuned to your specific company process, resourcing and approach.

Phase 1: Planning and Setup (Months 1-3)

During the initial phase, the focus is on establishing the foundational structure of the COE. This involves assembling a leadership team from across key departments, defining the specific roles and responsibilities within the COE, and developing initial security frameworks and policies. The objective is to create a robust organizational structure that supports effective communication and collaboration across diverse teams, setting the stage for the subsequent integration and development phases.

Key Actions (Phase 1)

- Assemble leadership teams and define the structure of the COE.
- Establish clear roles and responsibilities for all involved departments.
- Develop initial security frameworks and compliance guidelines.

Phase 2: Integration and Development (Months 4-6)

This phase is critical for integrating AI security protocols into existing systems and fostering interdepartmental collaboration. The COE will implement comprehensive training programs to ensure all members are up-to-date with the latest security practices and technologies. Additionally, the launch of initial interdepartmental projects aims to enhance collaborative efforts and begin the practical application of the developed security frameworks, thereby testing and refining these strategies in real-world scenarios.

Key Actions (Phase 2)

- Implement training and development programs for COE members.
- Begin integration of AI security protocols into existing systems.
- Launch initial interdepartmental projects to promote collaboration.

Phase 3: Operationalization (Months 7-9)

The COE becomes fully functional in the operationalization phase with all systems and processes actively running. Regular monitoring and reporting mechanisms are established to track the effectiveness of AI security measures. Security audits and risk assessments are conducted to identify any vulnerabilities and to ensure compliance with the established security protocols. This phase is crucial for adjusting operational workflows based on feedback and ensuring that the COE's activities are effectively enhancing AI security.

Key Actions (Phase 3)

- Fully operationalize the COE, with all systems and processes in place.
- Start regular monitoring and reporting of AI security measures.
- Conduct first round of security audits and risk assessments.

Phase 4: Evaluation and Expansion (Months 10-12)

The final phase focuses on evaluating the COE's effectiveness through comprehensive reviews and feedback mechanisms. Based on the outcomes of these evaluations, adjustments are made to policies and strategies. Additionally, plans for scaling COE activities are developed to include additional AI projects and initiatives, aiming to broaden the scope and impact of the COE's work. This phase ensures that the COE remains dynamic and adaptable to new challenges and opportunities.

Key Actions (Phase 4)

- Evaluate the effectiveness through feedback and performance metrics.
- Adjust policies and strategies based on outcomes and new industry insights.
- Plan for the scaling of COE activities and inclusion of additional AI projects.

Beyond the initial year, the COE engages in continuous improvement activities to keep abreast of the latest developments in AI and cybersecurity. This includes updating training programs, protocols, and security measures and maintaining open lines of communication with all stakeholders. Continuous monitoring of the AI landscape helps the COE proactively address emerging threats and innovate security solutions, ensuring the long-term resilience and trustworthiness of AI systems.

Emerging Trends in AI Security

As AI technologies rapidly evolve, so do the security challenges and solutions. Key emerging trends include:

1. **AI-powered cybersecurity**: Using AI to detect and respond to threats more quickly and effectively than traditional methods.
2. **Adversarial AI**: The rise of AI systems designed to attack other AI systems, necessitating robust defenses.
3. **Federated learning**: Enhancing privacy by training AI models on distributed datasets without centralizing the data.
4. **Quantum-resistant cryptography**: Preparing for the potential threat quantum computing poses to current encryption methods.
5. **Ethical AI regulations**: Increasing focus on regulatory frameworks to ensure responsible AI development and deployment.
6. **AI supply chain security**: Growing emphasis on securing the entire AI development pipeline, from data collection to model deployment.
7. **Explainable AI (XAI)**: Developing AI systems that can provide clear explanations for their decisions, crucial for security auditing and trust-building.

CoEs must stay abreast of these trends to effectively anticipate and address evolving security challenges in the AI landscape.

Summary

The escalating integration of Large Language Models (LLMs) and Generative AI into business operations underscores the critical need for a dedicated Center of Excellence (CoE) for AI security. Such a center ensures that AI technologies are implemented efficiently and maintained within secure and ethical frameworks. The comprehensive roles and responsibilities outlined in this document provide a foundational framework, illustrating the importance of including a diverse range of stakeholders from legal, risk management, IT, operations, and beyond

Organizations planning to develop their own CoE can leverage this framework as a blueprint to understand essential stakeholder roles and adapt them to their specific operational needs. The outlined metrics and phased implementation strategy offer a practical approach, guiding organizations through the systematic setup and scaling of their CoE. This phased approach helps manage the complexity of AI integration, ensuring that security measures evolve with technological deployments and adaptations.

It's essential to adapt the CoE framework to fit each organization's unique structure and needs. Not every company will require all the roles and functions described, as organizational capabilities and needs vary widely. However, the fundamental goal remains the same: to initiate and enhance cross-functional collaboration that builds comprehensive and effective policies for the secure and ethical use of generative AI applications.

Starting this process now is key to avoiding potential security risks and ensuring that AI technologies contribute positively to business growth and innovation. This proactive approach mitigates risks and maximizes the benefits of generative AI, paving the way for secure and successful future advancements.

Glossary

- **AI (Artificial Intelligence):** The simulation of human intelligence processes by machines, especially computer systems. In the context of this guide, AI refers to systems that use large language models and generative technologies.
- **AI Security:** The practice of protecting AI systems from threats, vulnerabilities, and risks. This includes securing data, algorithms, and the infrastructure supporting AI technologies.
- **Center of Excellence (CoE):** A centralized team or structure within an organization designed to promote best practices, provide leadership, and ensure the successful deployment and management of AI and LLM technologies.
- **CISO (Chief Information Security Officer):** The executive responsible for the organization's information and data security.
- **Compliance:** Adhering to legal, regulatory, and organizational policies and standards. In AI, this often involves ensuring that AI systems meet privacy, security, and ethical guidelines.
- **Data Science Team:** A group responsible for analyzing data, developing machine learning models, and ensuring the accuracy and integrity of AI systems.
- **Ethical AI:** The practice of developing and deploying AI systems in a way that is fair, transparent, and aligned with societal values.
- **Generative AI:** A type of AI that can create new content, such as text, images, or music, based on training data. Examples include large language models like GPT.
- **Governance:** The framework of rules, practices, and processes by which an organization ensures the effective and ethical management of AI technologies.
- **Key Performance Indicators (KPIs):** Metrics used to evaluate the success of an organization, department, or project in achieving its objectives.
- **Large Language Models (LLMs):** A type of AI model designed to understand and generate human language, often trained on vast amounts of text data.
- **Machine Learning (ML):** A subset of AI that involves the use of algorithms and statistical models to enable computers to perform tasks without explicit programming.
- **Multidisciplinary Team:** A group composed of members from various departments or fields, such as security, legal, and data science, working together on AI security initiatives.
- **Operationalization:** The process of implementing AI systems into an organization's daily operations, including their secure deployment and maintenance.
- **Risk Management:** The identification, analysis, mitigation, and monitoring of risks associated with AI technologies to protect the organization from potential threats.
- **Stakeholder Engagement:** The process of involving and communicating with individuals or groups who have an interest in the organization's AI initiatives (employees, customers, and regulators).
- **Security Framework:** A structured set of guidelines and best practices designed to protect AI systems from threats and ensure their secure deployment.
- **Shift-Left Strategy:** A practice that involves incorporating security measures early in the development process, rather than addressing them at the end.
- **Trust and Transparency:** Building confidence in AI technologies by ensuring that their operations are understandable, ethical, and aligned with stakeholders' expectations.
- **Vulnerability Assessment:** The process of identifying, evaluating, and addressing security weaknesses in AI systems.

Acknowledgements

Contributors

Scott Clinton
Sandy Dunn
Rachel James
John Sotiropoulos
Kalyani Pawar
Deryck Lio
Iván Mauricio Cabezas Troyano
Vaibhav Malik
Xiaobo Zhang

Reviewers

Andy Smith
Aruneesh Salhotra
Chadd Watson
Deryck Lio
Dustin Sachs
Emmanuel Guilherme
Eugene Neelou
Iván Mauricio Cabezas Troyano
Jason L Liang
John Sotiropoulos
Joshua Berkoh
Kalyani Pawar
Krishna Sankar
Madhavi N
Markus Hupfauer
Michael Isbitski
Mohan Sekar
Mohit Yadav
Nipun Gupta
Rhea Anthony
Rico Komenda
Rock Lambros
Talesh Seeparsan
Teruhiro Tagomori
Todd Hathaway
Ron F. Del Rosario
Vaibhav Malik
Vinnie Giarrusso

OWASP Top 10 for LLM Project Sponsors

We appreciate our Project Sponsors, funding contributions to help support the objectives of the project and help to cover operational and outreach costs augmenting the resources the OWASP.org foundation provides. The OWASP Top 10 for LLM and Generative AI Project continues to maintain a vendor neutral and unbiased approach. Sponsors do not receive special governance considerations as part of their support. Sponsors do receive recognition for their contributions in our materials and web properties.

All materials the project generates are community developed, driven and released under open source and creative commons licenses. For more information on becoming a sponsor Visit the Sponsorship Section on our Website to learn more about helping to sustain the project through sponsorship.

Silver Sponsors

Sponsor list, as of publication date. Find the full sponsor list here.

References

- Insight Editor. (2023, June 6). AI Center of Excellence best practices: Take your AI to the next level. Insight. Retrieved from https://www.insight.com/content/insight-web/en_US/shop/ai-center-of-excellence-best-practices.html
- Deloitte. (2023). AI Center of Excellence: Embedding AI in the business to enable intelligent enterprises. Deloitte Insights. Retrieved from https://www2.deloitte.com/us/en/insights/ai-center-of-excellence.html
- IBM Consulting. (2024, October 1). IBM consulting unveils Center of Excellence for generative AI. IBM. Retrieved from https://www.ibm.com/consulting/generative-ai-center-of-excellence.html
- Tenable. (2024, May 31). Cybersecurity best practices for implementing AI securely and ethically. Tenable. Retrieved from https://www.tenable.com/blog/cybersecurity-best-practices-for-implementing-ai
- Boston Consulting Group. (2023). GenAI's four key business opportunities. BCG. Retrieved from https://www.bcg.com/publications/2023/genai-key-business-opportunities
- KPMG. (2023). Generative AI success requires workforce remodel. KPMG. Retrieved from https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2023/generative-ai-success-requires-workforce-remodel.pdf
- Leapsome Team. (2023). The future of AI in HR: Studies, tips, and trends. Leapsome. Retrieved from https://www.leapsome.com/blog/future-of-ai-in-hr
- Microsoft. (2023). The AI revolution: How Microsoft Digital (IT) is responding with an AI Center of Excellence. Microsoft. Retrieved from https://www.microsoft.com/en-us/microsoft-digital/the-ai-revolution
- Thomson Reuters. (2023). A year in review: How AI transformed the legal profession in 2023. Thomson Reuters. Retrieved from https://legal.thomsonreuters.com/en/insights/articles/ai-year-in-review-2023
- CISA. (2024, April 15). Joint guidance on deploying AI systems securely. Cybersecurity & Infrastructure Security Agency. Retrieved from https://www.cisa.gov/news/joint-guidance-deploying-ai-systems-securely
- Defense News. (2023, September 28). AI security center to open at National Security Agency. U.S. Department of Defense. Retrieved from https://www.defense.gov/news/ai-security-center
- Human Resource Executive. (2024, January 9). What the AI executive order means for HR. Retrieved from https://www.hrexecutive.com/articles/what-the-ai-executive-order-means-for-hr
- ISACA. (2024, August 15). AI security risk and best practices. ISACA. Retrieved from https://www.isaca.org/resources/ai-security-risk-and-best-practices
- Mamgai, A. (2024, August 15). AI security risk and best practices. ISACA. Retrieved from https://www.isaca.org/resources/news-and-trends/industry-news/2024
- AWS Machine Learning Blog. (2023). *Establishing an AI/ML center of excellence*. AWS. Retrieved from https://aws.amazon.com/blogs/machine-learning/establishing-an-ai-ml-center-of-excellence
- Domino Data Lab. (2021). *Five steps to building the AI center of excellence*. Domino Data Lab. Retrieved from https://www.dominodatalab.com/blog/five-steps-to-building-the-ai-center-of-excellence
- Deloitte. (2022). *The future of cybersecurity and AI: Leveraging AI to bolster security defenses*. Deloitte Insights. Retrieved from https://www2.deloitte.com/us/en/insights/topics/ai-future-of-cybersecurity.html
- VentureBeat. (2021). *Best practices for a successful AI center of excellence: A guide for scaling AI initiatives*. VentureBeat. Retrieved from https://venturebeat.com/2021/12/15/best-practices-for-a-successful-ai-center-of-excellence
- C3.ai. (2020). *Best practices for governing the AI application lifecycle: The center of excellence approach*. C3.ai. Retrieved from https://c3.ai/resources/whitepaper/

Project Supporters

Project supporters lend their resources and expertise to support the goals of the project.

HADESS
KLAVAN
Precize
AWS
Snyk
Astra Security
AWARE7 GmbH
iFood
Kainos
aigos
Cloud Security Podcast
Trellix
Coalfire
HackerOne
IBM
Bearer
Bit79
stackArmor
Cohere
Quiq
Lakera
Credal.ai
Palosade
Prompt Security
NuBinary
Balbix
SAFE Security
BeDisruptive
Preamble
Nexus
PromptArmor
Exabeam
Modus Create
IronCore Labs
Cloudsec.ai
Layerup
Mend.io
Giskard
BBVA
RHITE
Praetorian
Cobalt
Nightfall AI