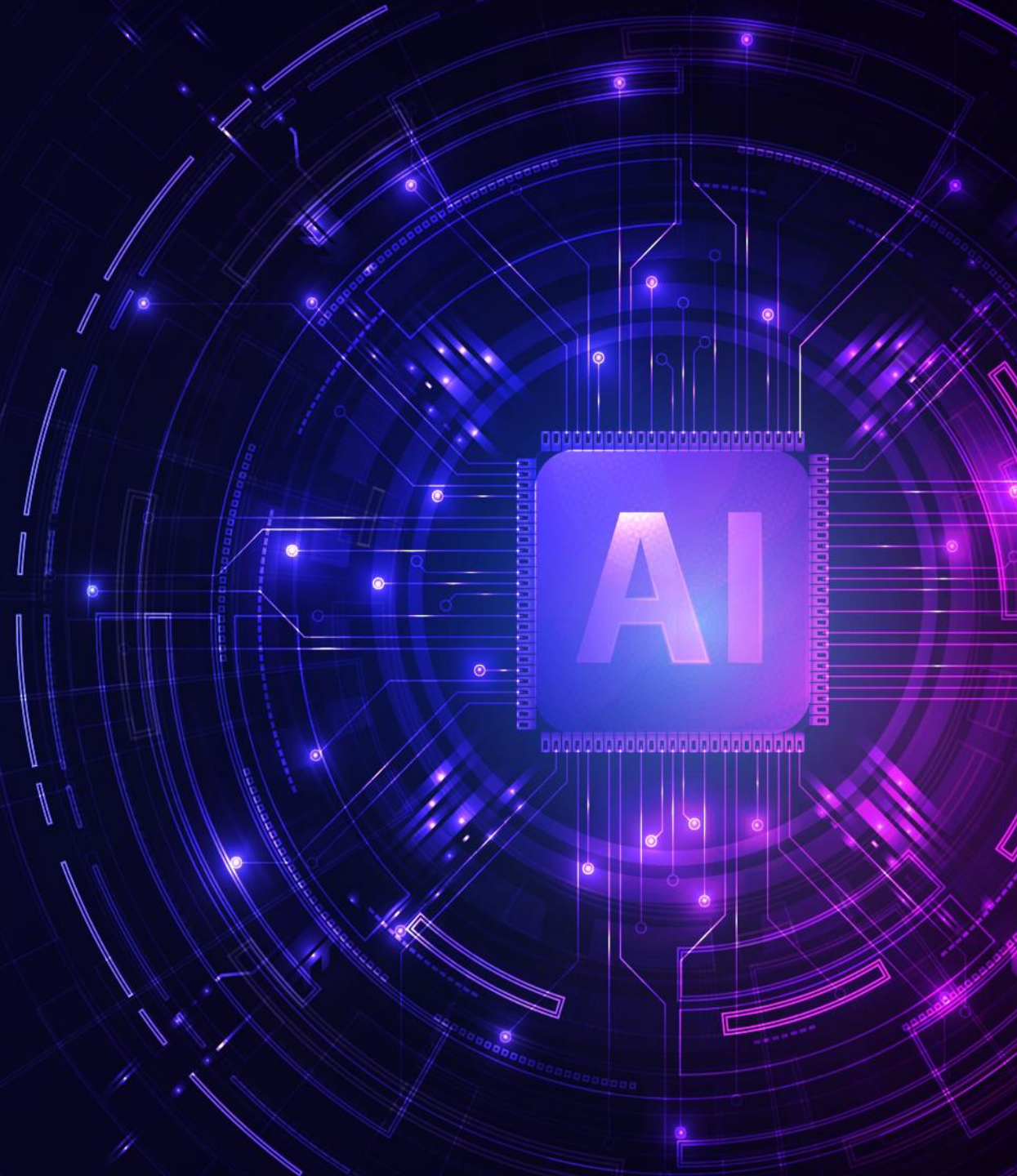
KPMG

Making Secure AI Real: **Real Threats, Lessons Learned, and Future of the Secure AI Technology**

SANS AI Cybersecurity Summit

September 8-9, 2024



With you today

Katie Boswell

KPMG US AI Security Lead
katieboswell@kpmg.com

Katie Boswell is a distinguished leader in Cyber Security Services, specializing in the Energy and Life Sciences sector. With over 20 years of experience, she partners with clients to enhance their cyber security strategy, particularly in Identity and Access Management (IAM). Her expertise strengthens the security and resilience of systems and infrastructure, crucial in safeguarding national infrastructure during times of disruption. Katie's commitment also extends to KPMG's people. She leads the Women in Cyber community, drives learning and development in the cyber field, and champions community-serving initiatives. Katie is unwavering in her support for diversity, equity, and inclusion (DEI) and strives to amplify diverse voices and experiences within the firm and among clients. Her guiding belief is that success is attainable for all.

Kristy Hornland

KPMG US AI Security Director
khornland@KPMG.com

Kristy Hornland is a Director at KPMG US specializing in AI security. She has delivered responsible and secure AI governance programs for leading life sciences, financial services, and government clients aligned to industry leading frameworks and practices, deployed AI security platforms to support these program objectives, and has held the position of Global Resilience Federation AI Security Working Group facilitator for the last two years. She has been deeply integrated in emerging technologies throughout her ten-year career with KPMG, and was part of the core team incubating KPMG’s first start up, Cranium, an AI Security platform. She is also the Women in Cyber deputy lead for KPMG US, defining the annual strategy and supporting overall governance to enable the entry, ongoing success, and long-term retention of women at KPMG.

Agenda

01 Internal & External Risks

Uncover anticipated risks before they become real issues

02 KPMG Trusted AI Framework

Understand our perspective on developing and deploying end-to-end trusted AI programs across the AI/ ML/ GenAI lifecycle

03 Trusted AI: Approach to Security

Discover the key inputs for a synthesized AI security strategy and how it enables the organization

04 The AI Security Journey

Step through the AI security journey and deep dive into key stepping stones

05 Q&A

Ask us questions!

With AI, it’s important to anticipate risks, before they become real issues

Internal Risks & Considerations

- Breaking Confidentiality & Intellectual Property
- Talent Implications
- Rapid Evolution of AI Supply Chain
- Employee misuse & inaccuracies

Financial, Brand and Reputational Risk

- Copyright
- Adversarial attack
- Cybersecurity (malicious use of AI)
- Misinformation, bias, discrimination

External Risks & Considerations

KPMG Trusted AI Framework

We understand trustworthy and ethical AI is a complex business, regulatory, and technical challenge. KPMG is committed to helping clients put it into practice. We help clients develop and deploy end-to-end trusted AI programs across the AI/ ML/ GenAI lifecycle.

Fairness
AI solutions should be designed to reduce or eliminate bias against individuals, communities, and groups.

Transparency
AI solutions should include responsible disclosure to provide stakeholders with a clear understanding of what is happening in each solution across the AI lifecycle.

Explainability
AI solutions should be developed and delivered in a way that answers the questions of how and why a conclusion was drawn from the solution.

Accountability
Human oversight and responsibility should be embedded across the AI lifecycle to manage risk and comply with applicable laws and regulations.

Data integrity
Data used in AI solutions should be acquired in compliance with applicable laws and regulations and assessed for accuracy, completeness, appropriateness, and quality to drive trusted decisions.

Reliability
AI solutions should consistently operate in accordance with their intended purpose and scope and at the desired level of precision.

Security
Robust and resilient practices should be implemented to safeguard AI solutions against bad actors, misinformation, or adverse events.

Safety
AI solutions should be designed and implemented to safeguard against harm to people, businesses, and property.

Privacy
AI solutions should be designed to comply with applicable privacy and data protection laws and regulations.

Sustainability
AI solutions should be designed to be energy efficient, reduce carbon emissions, and support a cleaner environment.

Trusted AI: Approach to Security

Input

Risk Indicators

Identifying incidents, vulnerabilities, risk considerations for AI systems (complex attack surface, etc.)

AI Ecosystem

Understanding your AI technologies, models in dev or production, third party models, and AI security tooling.

Regulations

Planning or addressing emerging/existing legislation or industry standards.

Public-Private Sector Initiatives

Leveraging knowledge from public/private/research led security initiatives (NIST, CISA, MITRE, University-led research, ISAC led working groups) to improve your organization's responsiveness to AI security considerations.

Enablement

Awareness

Enhance organization's security posture by empowering your workforce to understand emerging risks around AI.

AI Security Framework

Secure organization's AI landscape with a tailored framework providing governance and guidance on how to operationalize a robust AI security program.

AI Security Pipeline Management

Assess organization's current state of AI Security Pipelines, including technical components of an organization's AI pipeline and related vulnerabilities.

AI Security Uplift

Address organization's identified security issues and opportunities by providing a suite of services to specifically address the issues and opportunities discovered within an organization's AI ecosystem.

The AI Security Journey

Step 1
Intake Form/ Request Form, Third Party Questionnaire, or discovery scanning is completed to gain AI use case visibility

Step 2
Request Form submitted to SteerCo

Step 3
Request Form and supplemental documents about proposed use case presented to SteerCo

Step 4
SteerCo approves (or rejects) proposed Use Case to proceed to Risk Review/ Assessment

Step 5
Use Case owner completes **Risk Questionnaire**

Step 6
Completed **Risk Questionnaire** provided to Governance Committee for review

Step 7
Governance Committee reviews **Risk Questionnaire** results and identifies risk score

Step 8
Risk Assessment score formally assigned and relative controls assigned

Step 9
Risk Assessment score and assigned controls reviewed with SteerCo

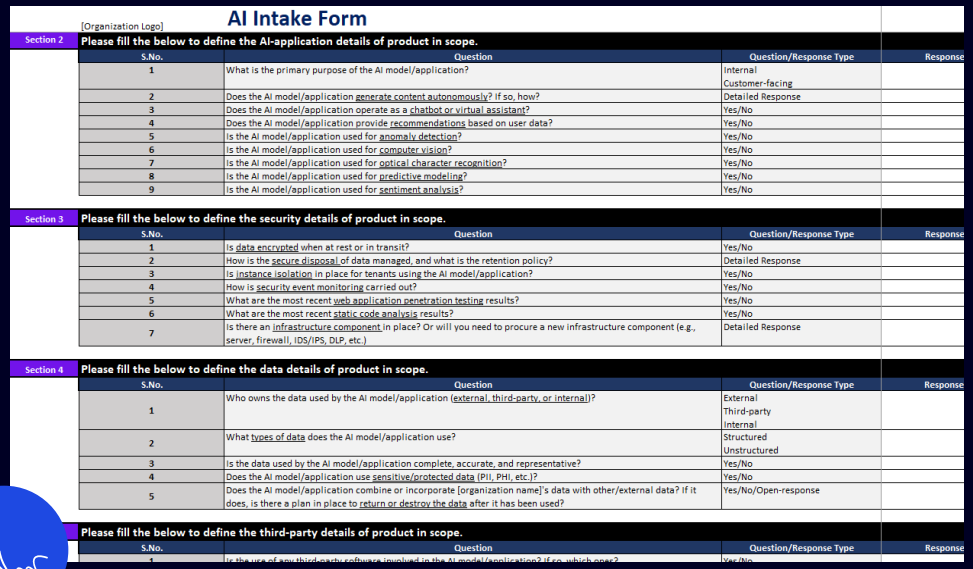
Step 10
SteerCo approves use case to proceed through development process

Step 11
Use Case undergoes **Threat Modeling** and architecture review

Step 12
Implementation begins (e.g., malware analysis, vulnerability assessment, backdoor detection, etc.)

Step 13
Post-Implementation Residual Risk Assessments performed

The Journey



[Organization Logo]

AI Intake Form

Section 2 — **Please fill the below to define the AI-application details of product in scope.**

S.No.	Question	Question/Response Type	Response
1	What is the primary purpose of the AI model/application?	Internal Customer-facing	
2	Does the AI model/application generate content autonomously? If so, how?	Detailed Response	
3	Does the AI model/application operate as a chatbot or virtual assistant?	Yes/No	
4	Does the AI model/application provide recommendations based on user data?	Yes/No	
5	Is the AI model/application used for anomaly detection?	Yes/No	
6	Is the AI model/application used for computer vision?	Yes/No	
7	Is the AI model/application used for optical character recognition?	Yes/No	
8	Is the AI model/application used for predictive modeling?	Yes/No	
9	Is the AI model/application used for sentiment analysis?	Yes/No	

Section 3 — **Please fill the below to define the security details of product in scope.**

S.No.	Question	Question/Response Type	Response
1	Is data encrypted when at rest or in transit?	Yes/No	
2	How is the secure disposal of data managed, and what is the retention policy?	Detailed Response	
3	Is instance isolation in place for tenants using the AI model/application?	Yes/No	
4	How is security event monitoring carried out?	Yes/No	
5	What are the most recent web application penetration testing results?	Yes/No	
6	What are the most recent static code analysis results?	Yes/No	
7	Is there an infrastructure component in place? Or will you need to procure a new infrastructure component (e.g., server, firewall, IDS/IPS, DLP, etc.)	Detailed Response	

Section 4 — **Please fill the below to define the data details of product in scope.**

S.No.	Question	Question/Response Type	Response
1	Who owns the data used by the AI model/application (external, third-party, or internal)?	External Third-party Internal	
2	What types of data does the AI model/application use?	Structured Unstructured	
3	Is the data used by the AI model/application complete, accurate, and representative?	Yes/No	
4	Does the AI model/application use sensitive/protected data (PII, PHI, etc.)?	Yes/No	
5	Does the AI model/application combine or incorporate [organization name]'s data with other/external data? If it does, is there a plan in place to return or destroy the data after it has been used?	Yes/No/Open-response	

Please fill the below to define the third-party details of product in scope.

S.No.	Question	Question/Response Type	Response
1	Is the use of any third-party software involved in the AI model/application? If so, which ones?	Yes/No	

Step 1

Intake Form/ Request Form, Third Party Questionnaire, or discovery scanning is completed to gain AI use case visibility

The **AI intake form** creates a standardized pathway for approving AI use cases while establishing minimum security criteria that a use case must meet. Gathering additional details of an AI use case will drive effective risk management, governance and efficient inventorying processes. The intake form supports risk evaluations, security assessments, technology alignment initiatives, and monitoring and reporting.

An AI Intake form may capture the following:

General Details
Name, origin (developed, acquired, integrated, description, purpose, etc.

Third Party Details
Third party (yes/ no), name, website, technical documentation, etc.

AI Technique Details
Machine learning, robotics, deep learning, generative AI etc.

AI Application Details
Content Generation, chatbots, virtual assistants, predictive modeling, etc.

Security Details
Data encryption, secure data retention/ disposal, security event monitoring

Data Details
Data ownership (external, third party, internal), sensitive data (yes/ no), etc.

Risk Details
Reputational risk, legal, IP, and privacy risk, cybersecurity risk, etc.

Implementation Details
Maintenance required, resource impact (reduction/ reallocation), etc.

Step 2

Step 3

Step 4

SteerCo approves (or rejects) proposed Use Case to proceed to Risk Review/ Assessment

Our process for intake, risk assessment, and threat modeling rest on the basis that a SteerCo has already been established with these key components:

Cross Functional Group Representation	Established Charter	Principles of Responsible AI	Risk Management
Legal (IP, IT Counsel)		Safety	Established RACI
Security & Privacy		Validity & Reliability	Defined Risk Scoring Methodology
Procurement & Third Party		Explainability & Interpretability	Defined RCM
Enterprise Risk Management		Accountability & Transparency	Risk Assessment
Model Risk Management		Privacy	
IT / Architecture		Security & Resilience	
Product		Fairness	
Strategy			

Illustrative AI governance organizational structure

AI Risk Questionnaire

Section 1	S.No.	Domain	AI Principle Alignment	Question
	1	Operations	Accountability	Who developed and trained the AI model: the enterprise internally or a third-party/vendor?
	2	Operations	Reliability	If the AI solution were to become unavailable, what would be the level of disruption to business processes?
	3	Operations	Reliability	What is the impact to [Client] as an organization if the AI solution were to become unavailable?
	4	Operations	Reliability	Does the AI solution support a sensitive business process/function/essential services. If yes, please put what business process(es)/function(s)/essential services will it support in the additional details column.
	5	Reputation	Explainability	Can the AI model pose reputational risk if the model were to become compromised? If yes, what type of reputational risk could the AI pose to [Client]?
	6	Reputation	Fairness	Does the AI solution takes steps to prevent unfairness and bias? (i.e. there are ways to check and make sure the use case is fair and unbiased)

Level 1 Risk	Level 2 Risk	Level 3 Risk	Control ID	Control Activities
Accountability and Transparency	**1. Inaccurate Content from Opaque AI** Production and usage of inaccurate content (including personal data) and/ or AI solution due to opaque AI models	*Failure to Understand AI Logic* The AI solution and logic is not fully understood - or is not accessible to the organization - hindering the ability to demonstrate effective end-to-end controls including validation over the relevant output used in	A.1.2	Logging should be configured to allow tracing of activities performed by the AI model (e.g. through static IP address) or user accounts within the AI solution. An end-to-end audit trail should be in place to support monitoring.
			A.1.6	Incorporate pre-model explainability techniques such as defining, documenting, and communicating model learning paradigms, model type, and input data structured to ensure transparency during model development.
		Lack of Explainable AI Solution Environment Additional IT and data components of the overall AI environment may	A.1.10	AI solutions should have thoroughly documented and well-maintained end-to-end detailed process narratives, flowcharts, and data flows that cover various use cases, process variations, and exceptions. The documentation should also integrate current internal controls, ensuring a comprehensive documentation of the overall processes.
	2. Differentiating Human vs AI Content Inability to distinguish human vs. AI-generated content	*Lack of AI Disclosure* Consumers inability to differentiate between AI and human-generated content, they may lose trust in the information presented to them, thereby causing brand reputation risk and content authenticity risk.	A.2.0	Design into AI user experience disclaimers/disclosures or features such as pop ups or explicit labels, disclaimers or visual cues so that stakeholders will be informed of the type of AI system they are interacting with or exposed to and why. Any outputs of AI are labelled as being produced by an AI system.
	4. AI Performance Erodes Over Time Leading to Outdated Solutions Inability to identify and monitor the use of AI solutions' performance	*AI Program and Project Management* Ai Program management methodologies are not in place to identify and monitor AI solutions' performance over time.	A.4.4	AI solutions intended use, metrics, fairness goals, and business goals are documented and communicated to relevant stakeholders. Changes to performance goals should be monitored over time and expanded, as appropriate. Management regularly reviews the solution outcomes (e.g. reports that have been designed and built to measure the performance of the AI solution) to ensure controls work at the same pace as the activities that are monitored (decision and operational velocity).

Step 5-7

Risk Assessment and scoring, and controls prescribed

In order to support a sustainable process, the AI Risk Assessment must come in alignment with some of the below key considerations:

- **Principled Approach**
- **Connected to Enterprise Risk Appetite and Scoring**
- **Complimentary to AI Strategy for the Organization**
- **Aligned with the Enterprise Regulatory Landscape**
- **Leverages Existing Processes and Groups**

Step 8 *Step 9* *Step 10*

Step 11

Use Case undergoes Threat Modeling and architecture review

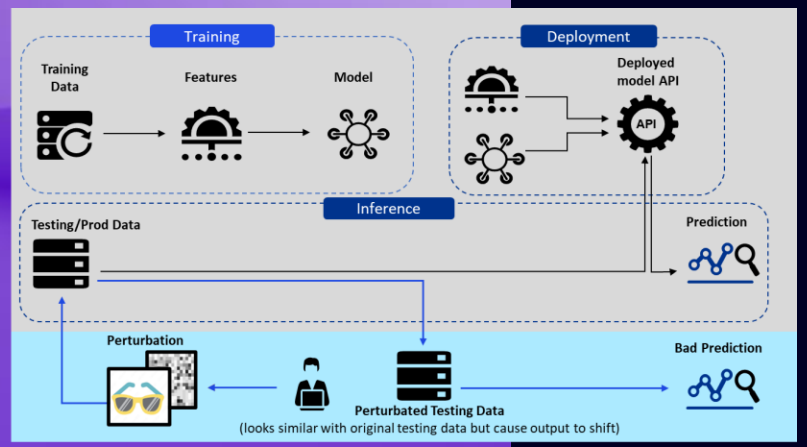
At the point of the AI Use Case progressing past our initial SteerCo reviews, we begin to loop back into our regular processes in any secure development lifecycle. For organizations, this can include evaluating the particular type of risk presented by the model type selected.

Identify the approved AI techniques and applications, leveraging information from the Intake/ Request Forms and Risk Assessment results

Leverage resources like MITRE ATLAS, OWASP Top 10 for ML and LLMs, and AI Risk and Incident Databases* to determine which threats may present based on the model identified. It is also important to identify which stage in the lifecycle we see the threat present. The OWASP AI Working Group also published their AI Security Threat Matrix February 2, 2024.

Map out the risk to detecting and protecting; explore potential motivations, impacts, the relative risk, how our prescribed controls from the risk assessment may help mitigate / protect, as well as how this could be detected.

Example Model Evasion Threat Mapping



*Resources listed below:

https://atlas.mitre.org/

https://owasp.org/www-project-top-10-for-large-language-model-applications/

https://owasp.org/www-project-machine-learning-security-top-10/

https://airisk.io/

https://incidentdatabase.ai/

https://owaspai.org/docs/ai_security_overview/#ai-security-matrix

> ***"Outcomes in the MAP function are the basis for the MEASURE and MANAGE functions."***
>
> — NIST AI Risk Management Framework 1.0, section 5.2

Step 12

Step 13

Questions?