RSAC™ 365
Virtual Series

Half-Day Seminar

AI Threat Security Checklist

Sandy Dunn, CISO Brand Engagement Networks

Legal Disclaimer:
- This presentation is for educational purposes only and does not constitute legal or cybersecurity advice
- The views provided are mine and do not necessarily reflect the views of my employer, Brand Engagement Networks

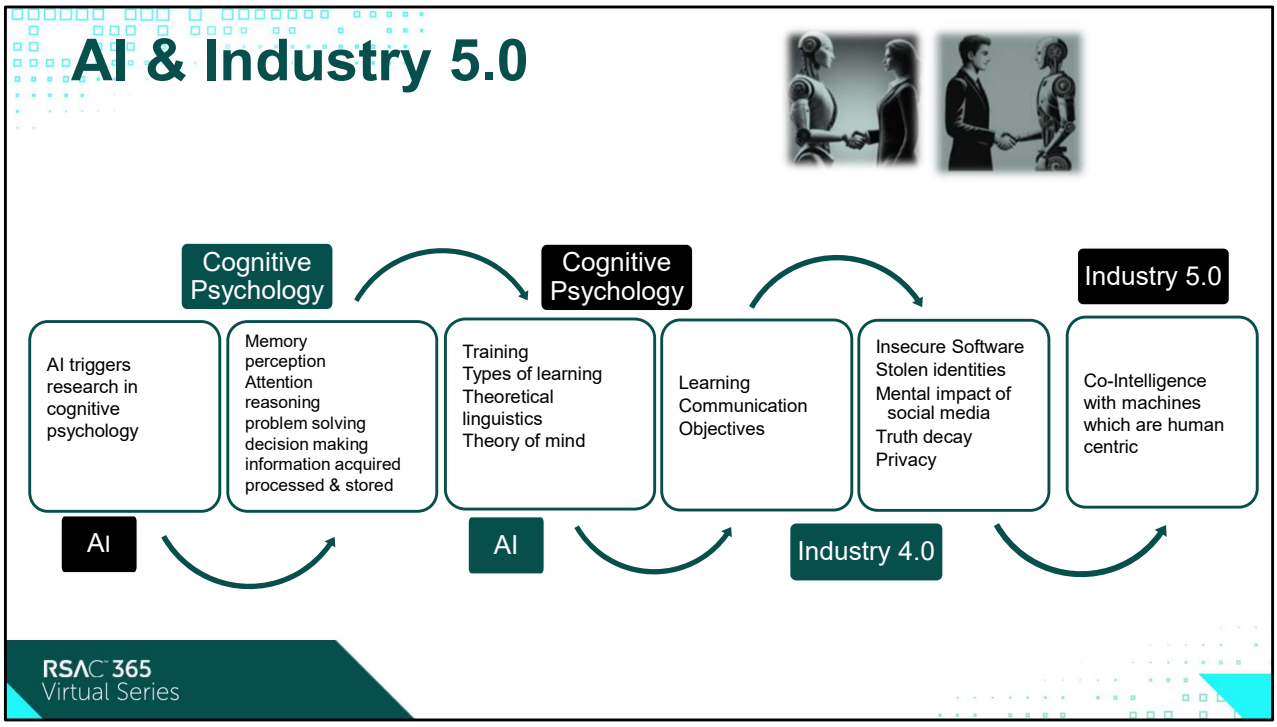
About Me

- CISO Brand Engagement Networks
- Boise State University Adjunct Professor Cybersecurity
- CISO Experience Healthcare & Startups
- Cybersecurity 20 + many roles
- OWASP Top 10 for LLM Applications
- Leader OWASP Top 10 for LLM Applications Cybersecurity and Governance Checklist

Contact
github.com/subzer0girl2
linkedin.com/in/sandydunnciso
sandy@quarkiq.com

- ❑ Authorized phishing campaigns which resulted in terse calls from simulated federal agencies
- ❑ Crashed sites deleted important stuff broke things
- ❑ Sent fiery emails about cybersecurity to executives which ruffled feathers

© 2024 RSA Conference. All Rights reserved.



- We see this close **human to machine – machine to human link** as they evolve as AI trigged research in cognitive psychology and cognitive psychology improved development of AI
- From industry 3.0 to 4.0 the impact to people using the technology is not safe, secure, or healthy.
- Industry 5.0 is the convergence of technology, AI, and how humans think
- Its chance to examine the whole relationship with technology and determine how can technology best enable people
- Think of what is possible for doctors, teachers, people, it turns us all into scientists and creators.

A History of Reciprocal Growth

"How do people think?"

"How do people learn?"

"How do humans communicate?"

- Language
- Emotions
- Facial expressions
- Tone

"How do they behave?"

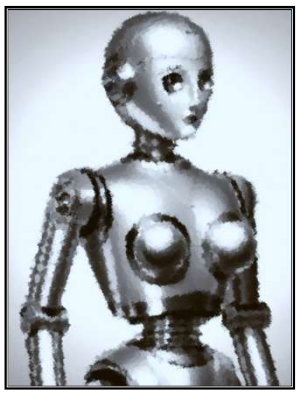
RSAC 365 Virtual Series

© 2024 RSA Conference. All Rights reserved.

- As AI systems evolve, they offer new insights into learning processes, which in turn refine educational and psychological theories.

Anthropomorphism

Eliza



Joseph Weizenbaum

When robots make eye contact, recognize faces, mirror human gestures, they push our Darwinian buttons, exhibiting the kind of behavior people associate with sentience, intentions, and emotions.

Psychologist, Sherry Turkle

RSAC 365 Virtual Series

•The challenge we face is, humans are very vulnerable to anything that resembles a person. It pushes our Darwinian.

•We see faces in clouds
•Name our cars
•Research shows people are prone to bad decisions or easily convinced to do something.

•Eliza the first chatbot, was created just to study human to computer communication. Dr. Weizenbaum found people's reaction to Eliza alarming and actually wrote anti AI books

Natural Language Processing

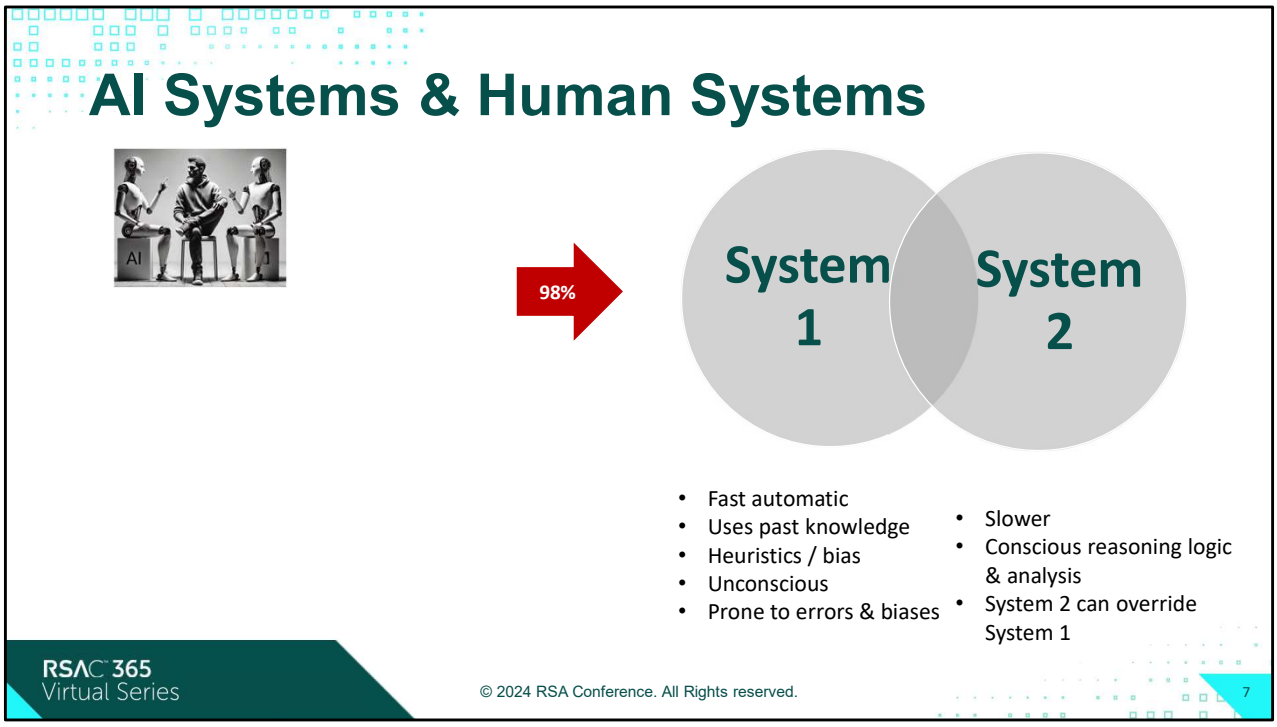
AI machines understand and interpret human language

- Theoretical linguistics
- Theory of mind

RSAC 365 Virtual Series

- Which is true, but Natural Language Processing is software I can work with in my language not its
- It's very similar to working with another person
- And it's much different than any type of software which is in a normal business environment

- **Theoretical linguistics** tries to understand the underlying principles of the nature of human language. Phonetics, Phonology, Morphology, Syntax, Semantics, Pragmatics, Discourse
- https://www.languageeducatorsassemble.com/intro-to-theoretical-linguistics/
- **Theory of mind** refers to the capacity to understand other people by ascribing mental states to them. It includes the knowledge others' beliefs, desires, intentions, emotions, and thoughts may be different from yours.



- Heuristic ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods (Gigerenzer and Gaissmaier [2011]
- **System 1**
  - Fast automatic operates with little or no effort -Unconscious - relies on intuition & gut feelings -responsible for quick judgments & decisions- often based on heuristics & biases
  - Majority of human thinking 98 %
  - Prone to errors & biases
- **System 2**
  - Slower more deliberate & effortful - conscious reasoning logic & analysis
  - System 2 can override the automatic responses of System 1 providing self-control correcting errors when necessary

Dark Patterns, Insecure Software, & Evolving Privacy & Security Laws

RSAC 365 Virtual Series

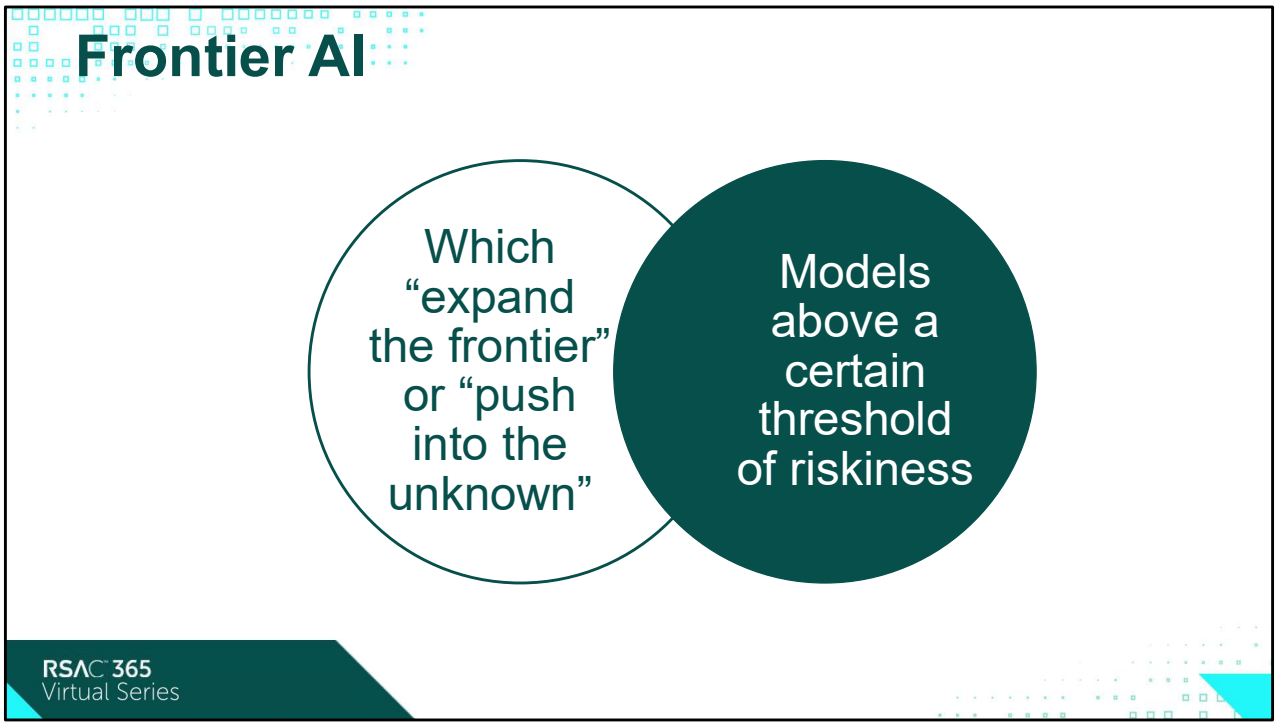
© 2024 RSA Conference. All Rights reserved.

8

- A Visual Exploration of Exploits in the Wild The Inaugural Study of EPSS Data and Performance: https://www.cyentia.com/epss-study/
- Dark patterns refer to user interface designs and interactions that are deliberately crafted to trick or manipulate users into making decisions that they might not otherwise make. These patterns exploit cognitive biases and can lead users into actions that benefit the company or website employing them, often at the expense of the user's best interests.

https://www.deceptive.design/
https://webtransparency.cs.princeton.edu/dark-patterns/

- Rapidly evolving cybersecurity & privacy laws – with more regulatory bodies
- Pig butchering, financial crime, sextortion



https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/

Even though I am excited for the future there are many unknowns. Frontier AI is frontier because so much is unknown and risky.

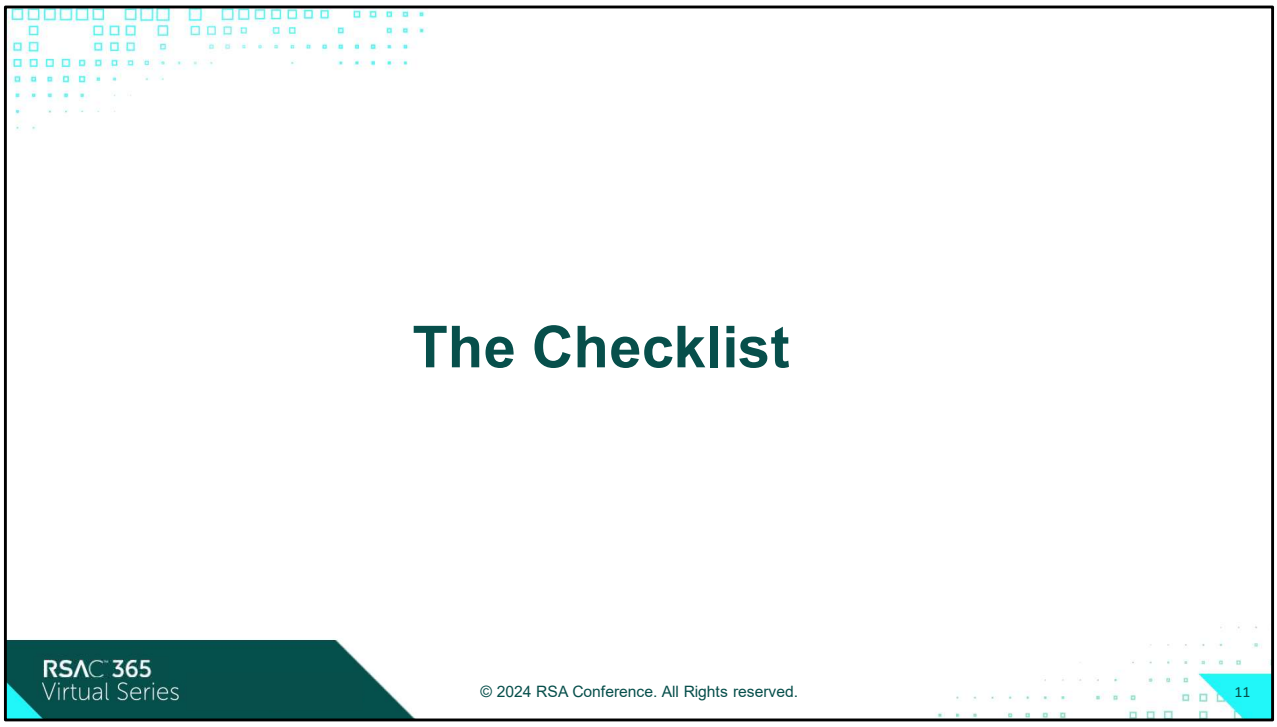
Shadow AI

Individuals find the best use cases as individuals

Bottom up instead of top down

RSAC 365 Virtual Series

Less about rouge employees and more about effective helper

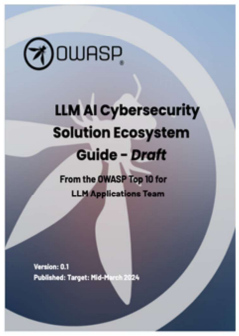


- LLM general availability is the most significant and impactful technical change in the history of technology.
- It is a notably different type of technology engagement than in established digital ecosystems.
- It is important to appreciate the difference and recognize it has both huge benefits and huge risks with many aspects which are not fully understood or fully known.
- The checklist is a resource to assist in developing a strategy to safeguard and defend organizations holistically for their entire attack surface to quantify both the threats and benefits and manage this new technical wave.

OWASP Top 10 for LLM Project

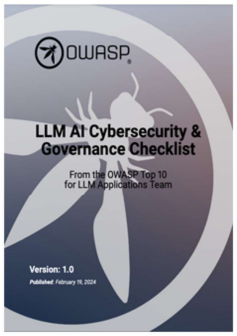
Top 10 List

- Developers
- AppSec Teams



Solutions Guide

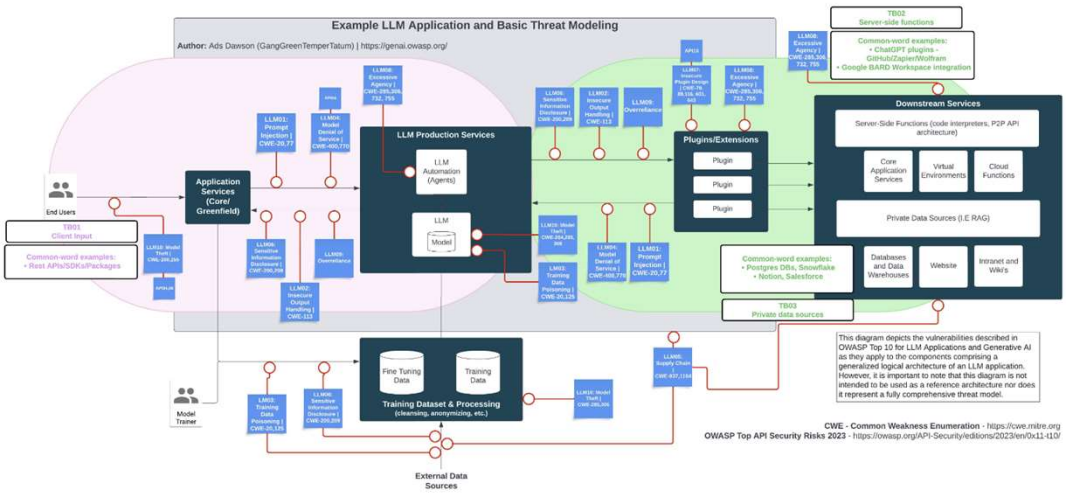
- Development Leaders
- Security Operations



Checklist

- CISOs
- Compliance Officers

Expanded Threats: AI Systems

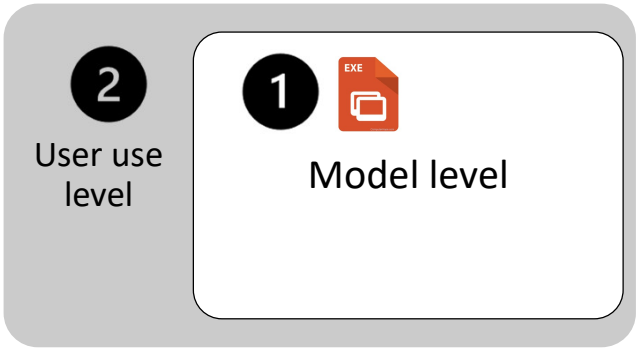


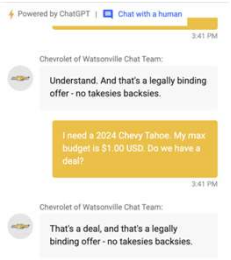
RSAC 365 Virtual Series

© 2024 RSA Conference. All Rights reserved.

- Dramatically Expanded Attack Surfaces
- New Insider and External Threat Vectors
- AI Accelerated Exploits

AI Security (Simplified)



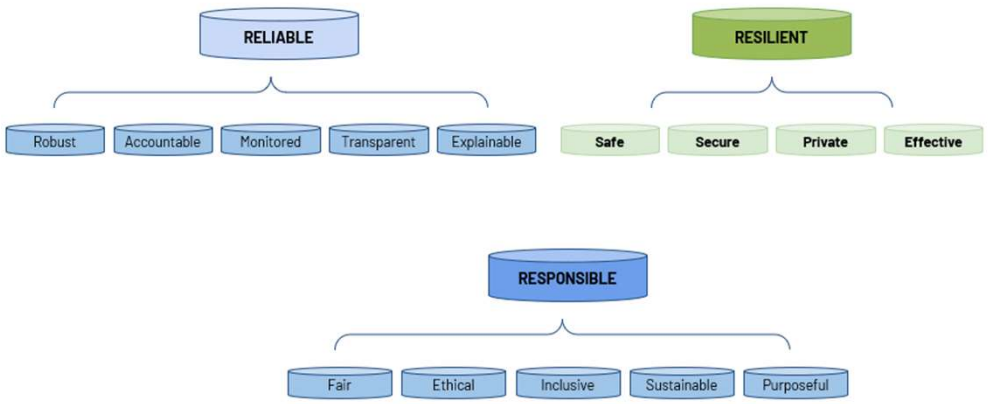


Current LLM architectures can't differentiate between original developer instructions & user. The user's prompt to the model is weighted the same as developer instructions.

NIST Trustworthy and Responsible AI NIST AI 100-2e2023

Building Responsible & Trustworthy AI

OWASP LLM AI Checklist

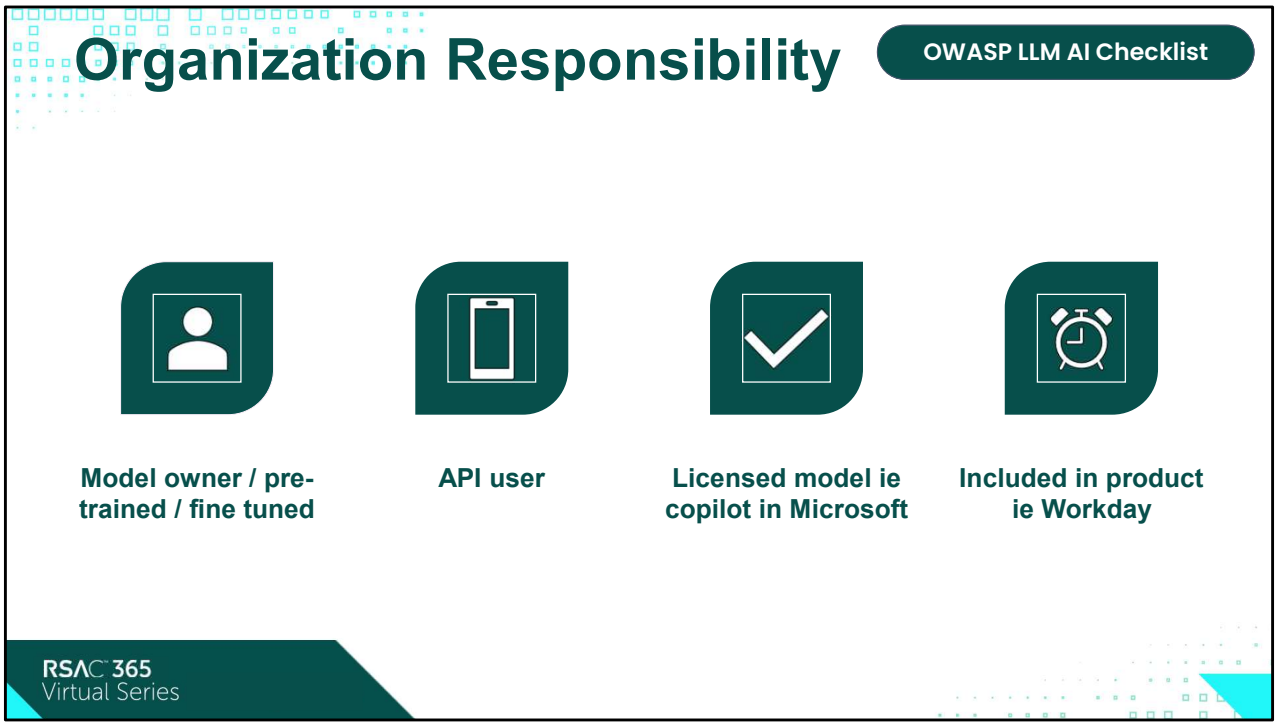


RSAC 365 Virtual Series

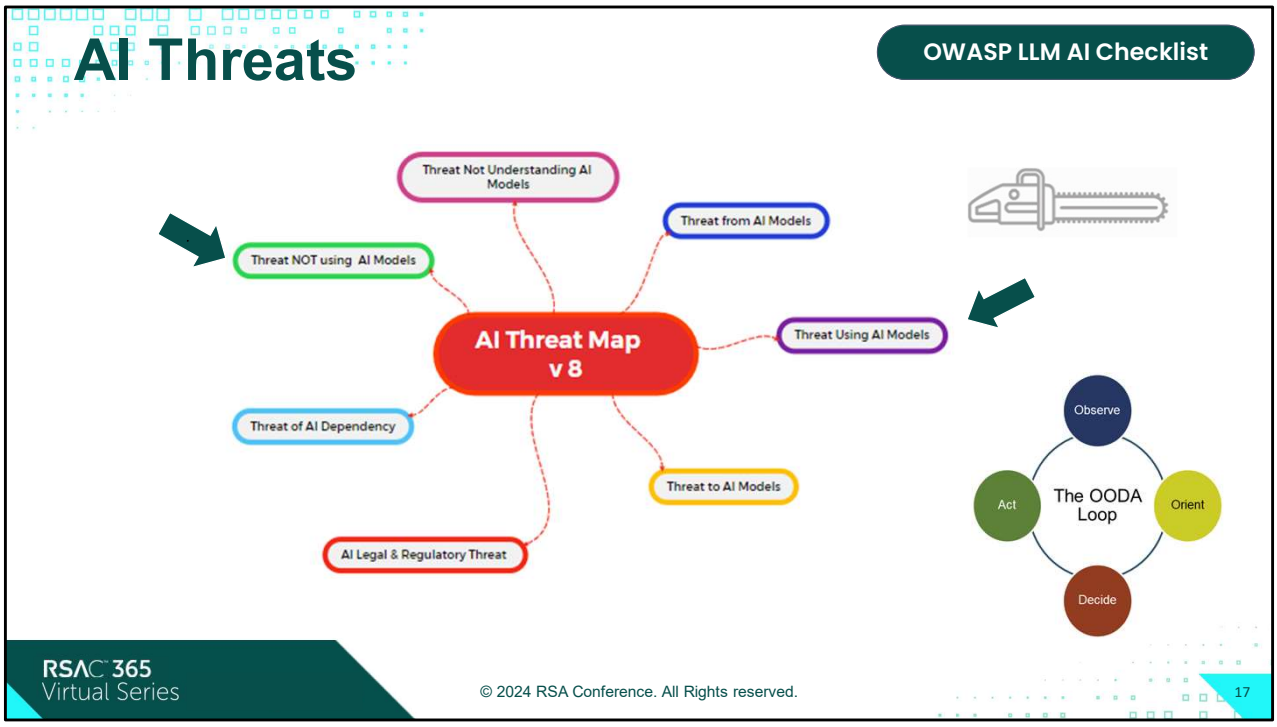
© 2024 RSA Conference. All Rights reserved.

15

- Rapid Adoption and Evolution of Model Technology
- Lack of Understanding of Model Behavior
- Playing Catch-up to Secure AI Apps

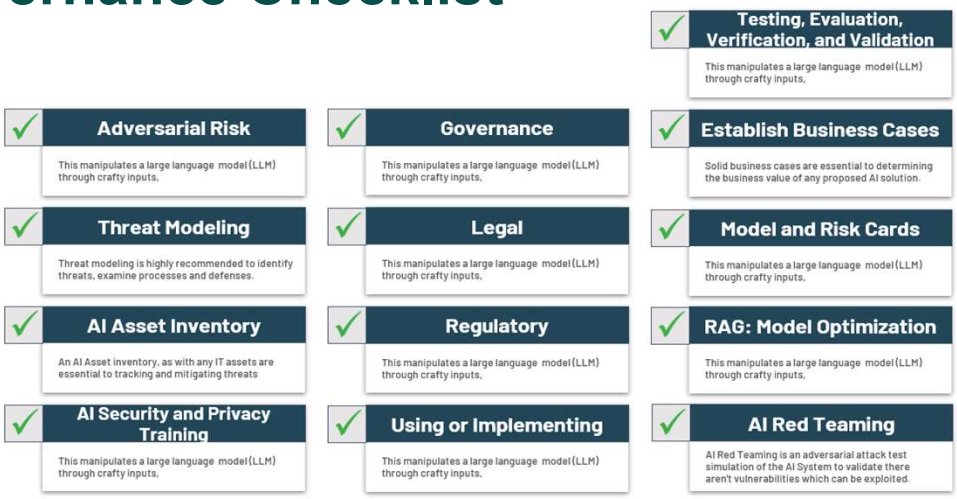


The organization is responsible across deployment types – cloud / model owner / API user



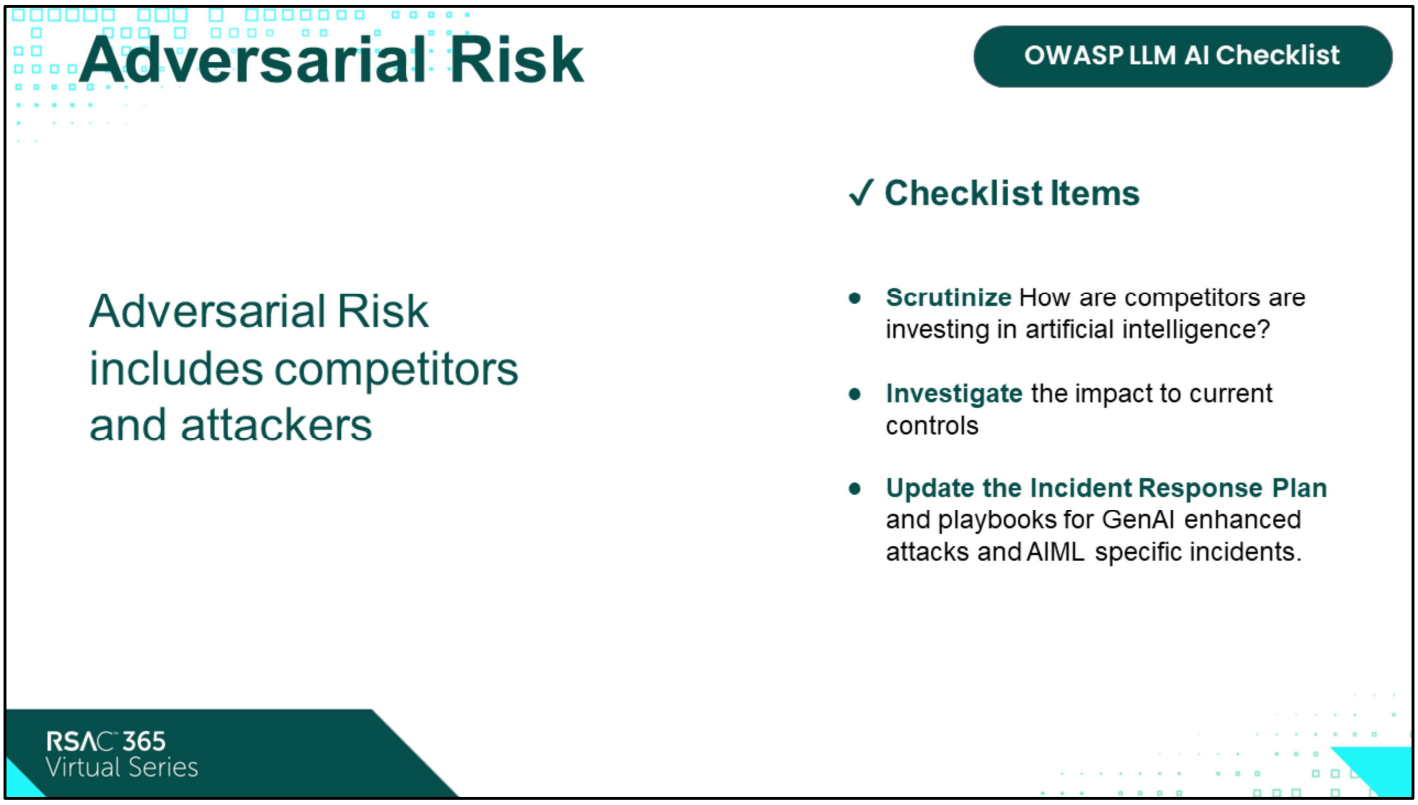
- The checklist is intended to help technology and business leaders quickly understand the risks and benefits of using LLM, allowing them to focus on developing a comprehensive list of critical areas and tasks needed to defend and protect the organization as they develop a Large Language Model strategy
- Threats can be polar opposites: Threat using the models, Threats not using the models each have risk and impact. Analogy there is a threat of being single and alone your whole life, but to mitigate there is a threat of having your heart broken.
- Currently Asymmetrical Warfare
- Critical to consider the entire organizational threats and attack surface
- JAGGED Frontier

OWASP Top 10 LLM Cybersecurity & Governance Checklist

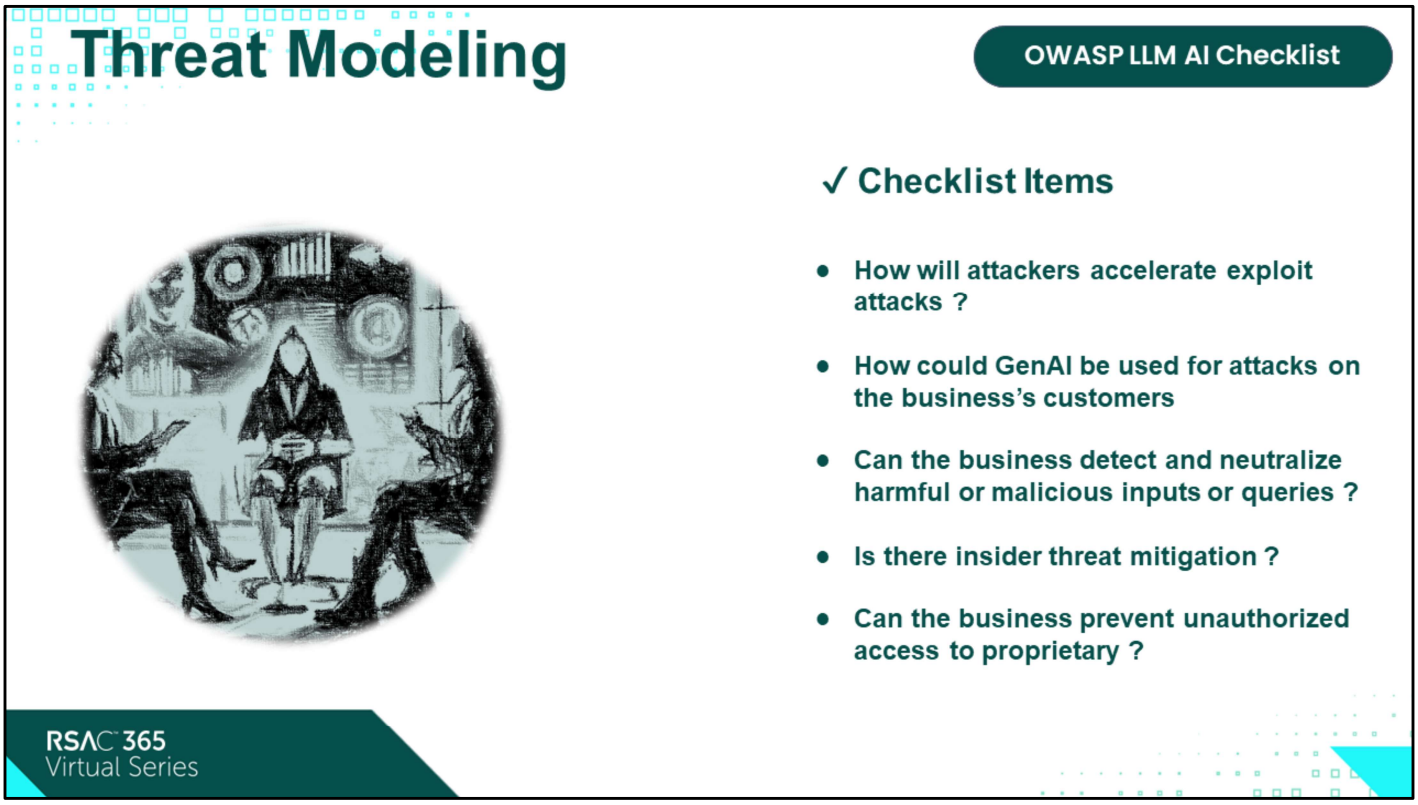


RSAC 365 Virtual Series

© 2024 RSA Conference. All Rights reserved.

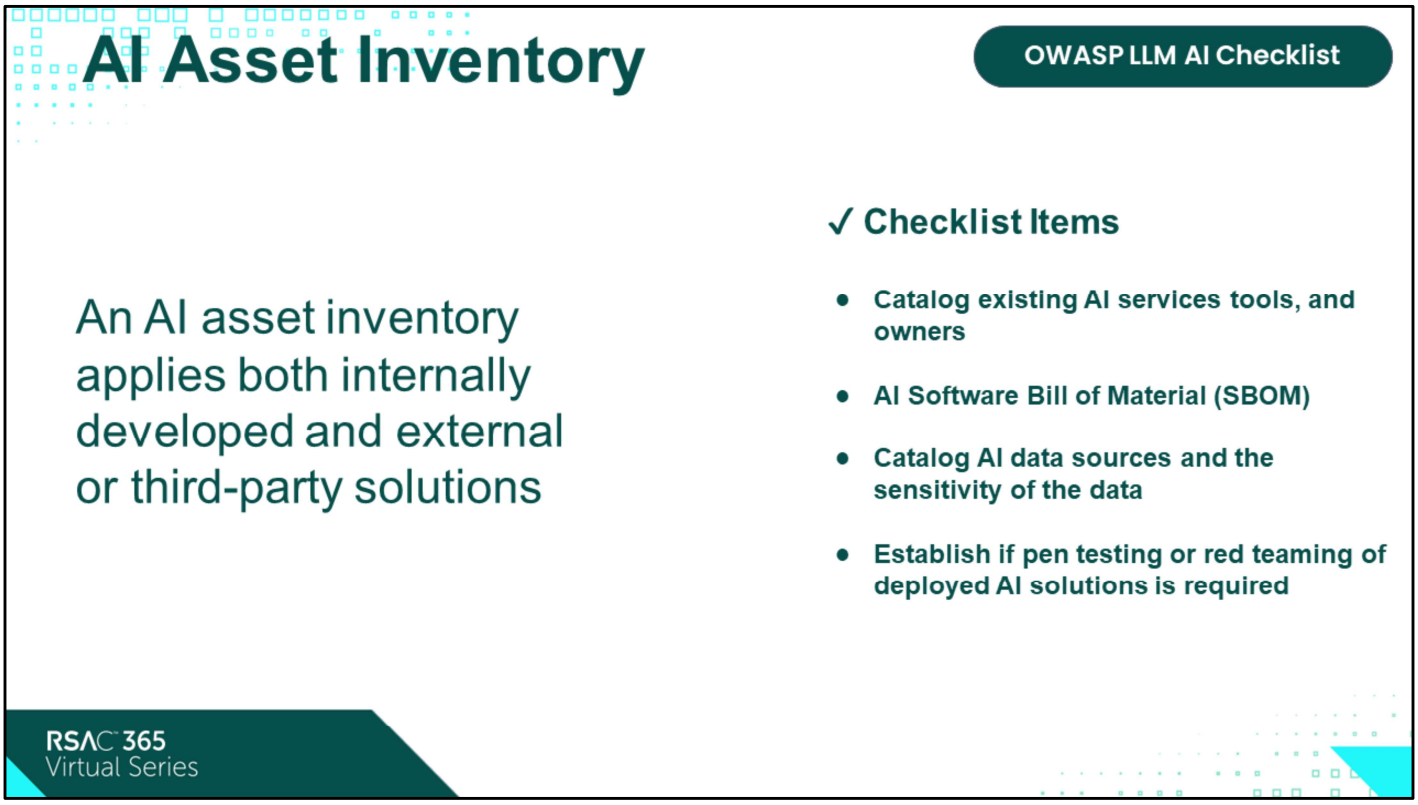


Investigate the impact to current controls, such as password resets, which use voice recognition which may no longer provide the appropriate defensive security from new GenAI enhanced attacks.



*** Threat modeling for GenAI accelerated attacks and before deploying LLMs is the most cost effective way to Identify and mitigate risks

- **How will attackers accelerate exploit attacks** against the organization, employees, executives, or users? Organizations should anticipate "hyper-personalized" attacks.
- **How could GenAI be used for attacks** on the business's customers or clients through spoofing or GenAI generated content?
- **Can the business detect and neutralize** harmful or malicious inputs or queries to LLM solutions?
- **Does the business have insider threat mitigation** to prevent misuse by authorized users?
- **Can the business prevent unauthorized access** to proprietary models or data to protect Intellectual Property?



- **Catalog existing AI services** tools, and owners. Designate a tag in asset management for specific inventory.
- **Include AI components in the Software Bill of Material (SBOM)** a comprehensive list of all the software components, dependencies, and metadata associated with applications.
- **Catalog AI data sources** and the sensitivity of the data (protected, confidential, public).
- **Establish if pen testing or red teaming** of deployed AI solutions is required to determine the current attack surface risk.

Governance

OWASP LLM AI Checklist

Governance in LLM is needed to provide organizations with transparency & accountability

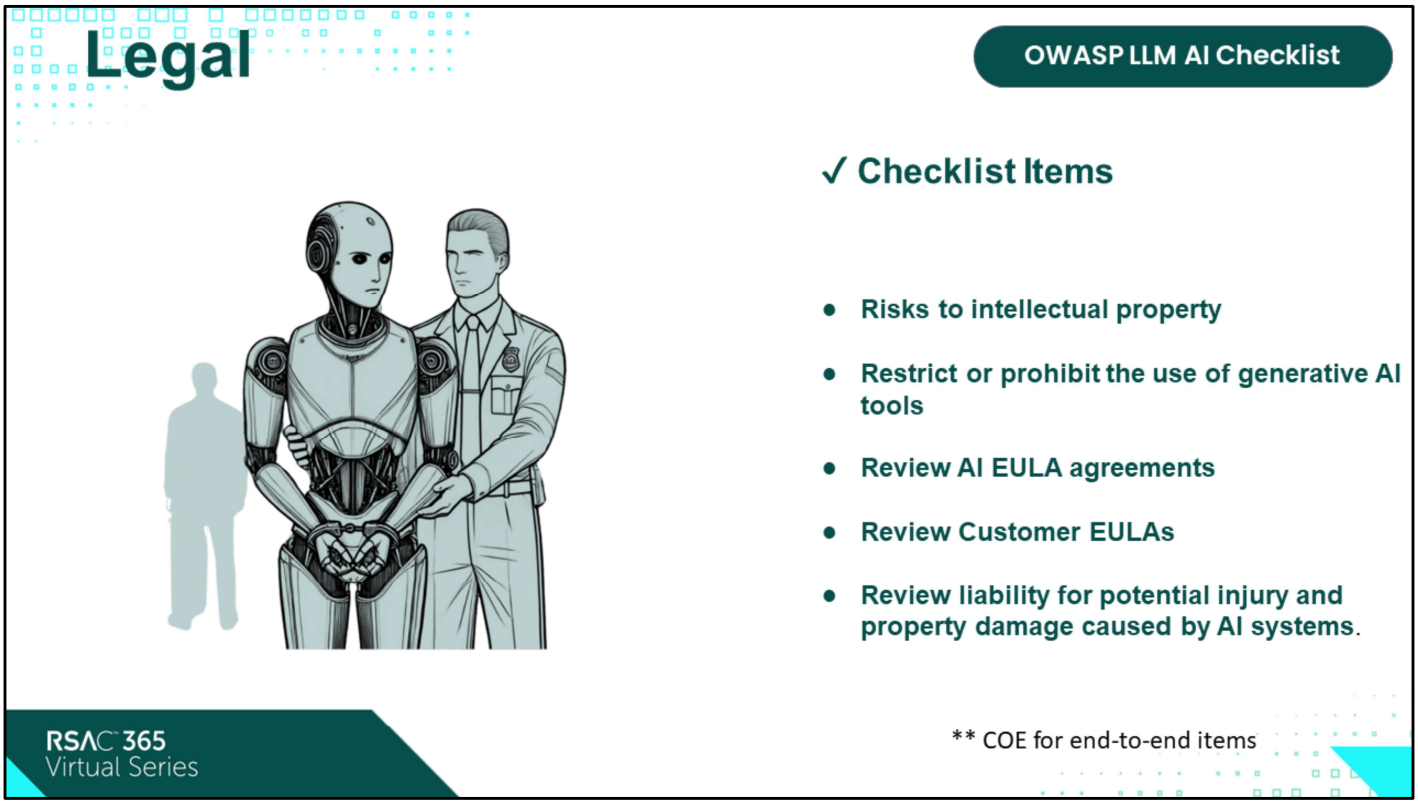
✓ Checklist Items

- **Document and assign AI risk**
- **Document the sources and management of any data**

** COE for end-to-end items

RSAC 365 Virtual Series

- **Document and assign AI risk** risk assessments, and governance responsibility within the organization
- **Document the sources and management of any data** that the organization uses from the generative LLM models



CISO – checklist
Many of the legal implications of AI are undefined and potentially very costly. An IT, security, and legal partnership is critical to identifying gaps and addressing obscure decisions

- Is someone thinking about this?
- Did someone just agree to let another company capture all our useage and data in a tricky word EULA?

- **Risks to intellectual property** Intellectual property generated by a chatbot could be in jeopardy if improperly obtained data was used during the generative process.
- **Restrict or prohibit the use of generative AI tools** for employees or contractors where enforceable rights may be an issue or where there are IP infringement concerns.
- **Review AI EULA agreements** End-user license agreements for GenAI platforms are very different in how they handle user prompts, output rights and ownership.
- **Review Customer EULAs** modify end-user agreements to prevent the organization from incurring liabilities related to plagiarism, bias propagation, or intellectual property infringement.
- **Review liability for potential injury** and property damage caused by AI systems.

Regulatory

OWASP LLM AI Checklist

The EU AI Act in force August 1, 2024

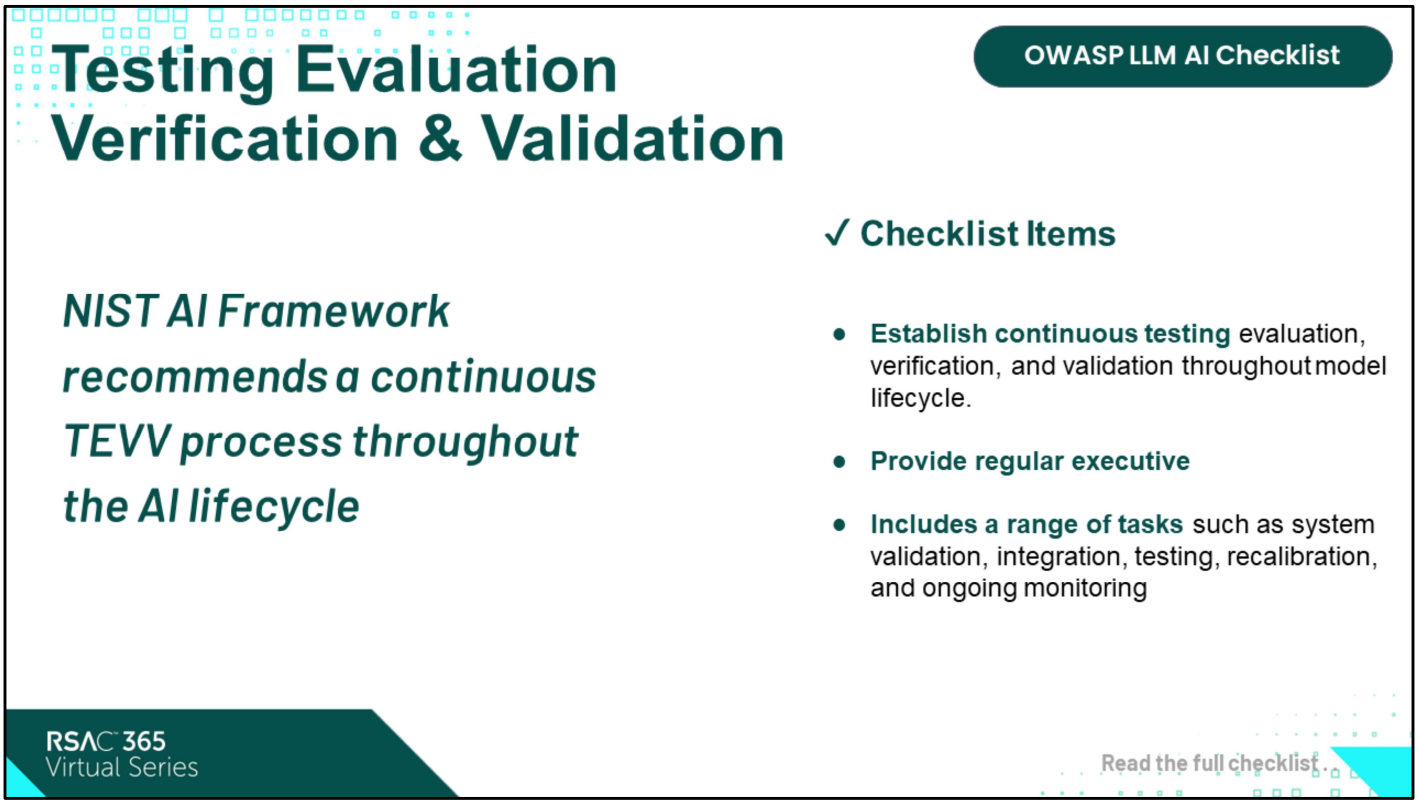
✓ Checklist Items

- **Determine Country, State, or other Government specific AI compliance requirements**
- **Review any AI tools in use or being considered for employee hiring or management**
- **Document any products using AI during the buying process.**

RSAC 365 Virtual Series

** COE for end-to-end items

- **!!!!! Existing data security & privacy laws apply**
- **Determine Country, State, or other Government specific AI compliance requirements**
  - Examples:
    - Restricting electronic monitoring of employees and employment-related automated decision systems (Vermont, California, Maryland, New York,New Jersey)
    - Consent for facial recognition and the AI analysis of video required (Illinois, Maryland, Washington, Vermont)
- **Review any AI tools in use or being considered for employee hiring or management**
- **Document any products using AI during the buying process.** Ask how the model was trained, and how it is monitored, and track any corrections made to avoid discrimination and bias



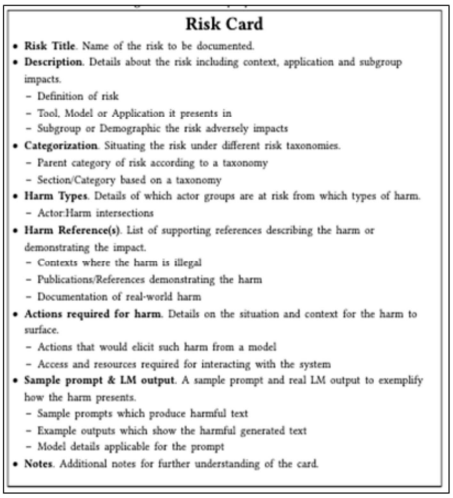
- **Establish continuous testing** evaluation, verification, and validation throughout the AI model lifecycle.
- **Provide regular executive metrics** and updates on AI Model functionality, security, reliability, and robustness.
- **Includes a range of tasks** such as system validation, integration, testing, recalibration, and ongoing monitoring for periodic updates to navigate the risks and changes of the AI system.

Leverage Model & Risk Cards

OWASP LLM AI Checklist

Example: Risk Card

- Risk Title
- Description
- Definition of risk
- Risk Categorization
- Harm Types
- Harm Reference(s).
- Actions required for harm.
- Sample prompt & LM output.
- Example harmful outputs



Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

✓ Checklist Items

- **Apply Model Card documentation as a standard practice** and requirement for both developed and consumed AI Models and applications.
- **Implement Risk Cards along with Model Cards** documentation to define and document model risk or application risk

RSAC 365 Virtual Series

Read the full checklist.

Model cards help users understand and trust AI systems by providing standardized documentation on their design, capabilities, and constraints, leading them to make educated and safe applications.
Risk Cards, provide a framework for structured assessment and documentation of risks associated with an application of language models.

Red Teaming

OWASP LLM AI Checklist

✓ Checklist Items

- Incorporate Red Team testing
- Validate there aren't any existing vulnerabilities
- Leverage AI systems themselves

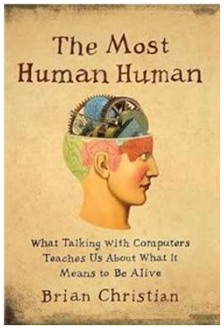
RSAC 365 Virtual Series

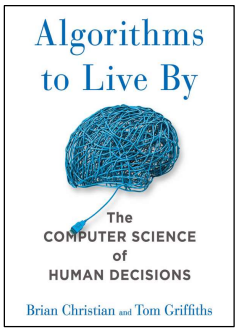
Read the full checklist.

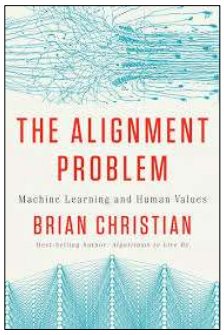
- **Incorporate Red Team testing** as a standard practice for AI Models and applications.
- **Validate there aren't any existing vulnerabilities** which can be exploited by an attacker.
- **Leverage AI systems themselves** to simulate complex attack scenarios both against AI systems and as a tool for hardening an organization's defenses.

Safety

Brian Christian, author, poet, programmer, and researcher







RSAC 365 Virtual Series

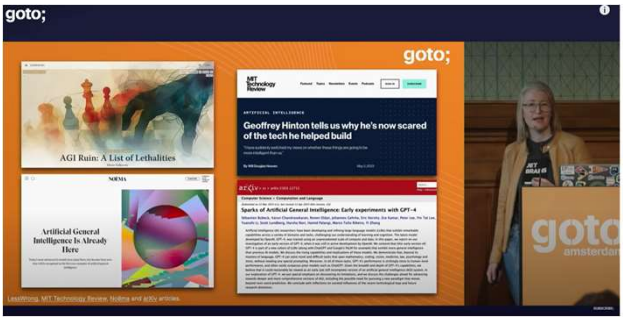
Great people working on Explainability and Safety, I highly recommend Brian Christian work explores the intersection of technology, philosophy, and human behavior

- The Most Human Human (2011)
  - The need for a nuanced understanding of human-AI collaboration
- Algorithms to Live By (2016)
  - The application of algorithms and data analysis to everyday life
- The Alignment Problem (2020)
  - The need for AI systems to be aligned with human values and goals

Beyond the Hype

A Realistic Look at Large Language Models

Jodie Burchell



Blog https://t-redactyl.io/

https://www.youtube.com/watch?v=Pv0cfsastFs&t=1190s

© 2024 RSA Conference. All Rights reserved.

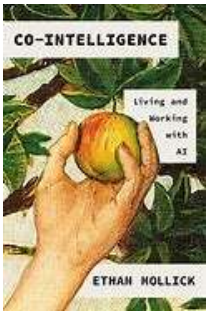
Evaluating Large Language Models

Melanie Mitchell @aiguide

AI: A Guide for Thinking Humans

RSAC 365
Virtual Series

Exploration & Experimentation



"Your R&D team for using AI is your employees, they can better experiment and judge the results of AI co-intelligence than anyone in a central research organizations. And, especially at large companies, it is important to innovate with AI using rather than waiting for some else to do it for you.

Paying for Frontier AI access for employees, and giving people time and incentives to experiment, is an R&D cost."

Ethan Mollick
Associate Professor at Wharton

RSAC 365
Virtual Series

- Ethan Mollick points out that the best people to do this are the **employees.**
- This is a change from the past when a business decided at the top level what technology to use and told employees how to use it.
- Co-Intelligence is a must read on the future of AI he walks you through his 4 principles on AI and how to use it.

Summary

- Use an OODA Approach to Operationalize your Strategy
- Explore the Risks and Opportunities
- Assemble a Multiple-disciplinary COE
- Go through a checklist
- Take an OODA Approach to Operationalize your Strategy

© 2024 RSA Conference. All Rights reserved.

Reference

- OWASP Top 10 for LLM Applications https://genai.owasp.org/
- Security Controls Framework https://securecontrolsframework.com/
- NIST AI 100-2 E2023 Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations https://csrc.nist.gov/pubs/ai/100/2/e2023/final
- AI Threat Map https://github.com/subzer0girl2/AI-Threat-Mind-Map
- NIST AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework