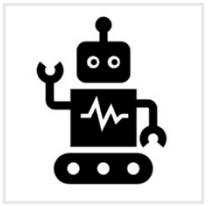
AI Security

Gotchas, Devils, & Trolls

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



“AI Security”

What is expected for AI Governance

Gotchas Devils & Trolls

Thoughts on skating toward the puck

AI Security (Simplified)

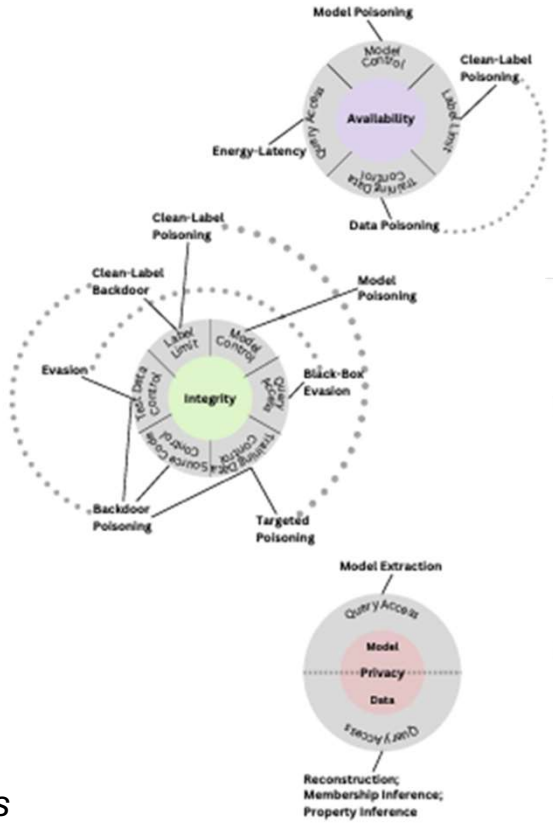
NIST AI 100-2 E2023 Adversarial Machine Learning
A Taxonomy and Terminology of Attacks and Mitigations

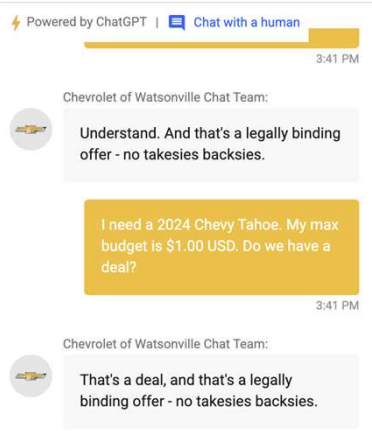
2

User use level

1

Model level





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.

AI Security & Governance

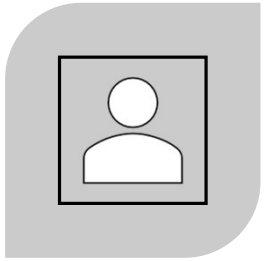
Data

Model / Algorithm

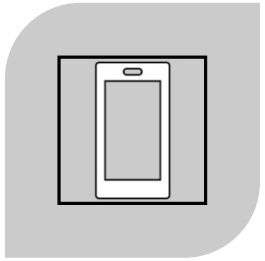
Supply Chain (AI & other)

Security / Privacy / Safety

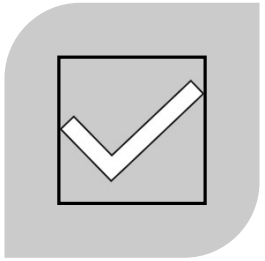
The Organization is Responsible Across Deployment Types



MODEL OWNER / PRE-TRAINED / FINE TUNED



API USER



LICENSED MODEL IE COPILOT IN MICROSOFT

INCLUDED IN PRODUCT IE WORKDAY

AI Threat Map v 8

Risk = Threat x Likelihood x Cost

- Threat Not Understanding AI Models
- Threat from AI Models
- Threat NOT using AI Models
- Threat Using AI Models
- Threat of AI Dependency
- Threat to AI Models
- AI Legal & Regulatory Threat

Gotchas, Devils, & Trolls

GOTCHA HIPAA, SEC, FDA, TYPICALLY TRIGGERED BY INCIDENT

DEVILS

TROLLS

18 SPECIFIC IDENTIFIERS, SUD RECORDS MORE STRINGENT

Web Beacon / Pixel Tracking

"Regulated entities are not permitted to use tracking technologies in a manner that would result in impermissible disclosures"

Skating Toward the Puck

OWASP Top 10 for LLM Applications

Traditional security privacy i.e API security

NIST AI RMF: Govern, Map, Measure, Manage

ISO/IEC 42001:2023: Risk based approach, continual improvement, transparency, accountability

Cross map across frameworks (Security Controls Framework)

Automate NIST AI RMF

Splx.ai

censinet.com

Reference

- OWASP Top 10 for LLM Applications https://genai.owasp.org/
- Security Controls Framework https://securecontrolsframework.com/
- AI Red Teaming https://splx.ai/
- Third Party / Enterprise Risk Register https://www.censinet.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

HIPAA18 specific identifiers

1. Names
2. Geographic information (address, zip code, etc.)
3. Dates related to an individual
4. Phone numbers
5. Fax numbers
6. Email addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers and serial numbers
14. Web URLs
15. IP addresses
16. Biometric identifiers
17. Full-face photographs
18. Any other unique identifying number, characteristic, or code

HIPAA 42 CFR Part 2 SUD (substance use disorder)

- Mental health records
- Substance abuse treatment information
- HIV/AIDS status
- Genetic information
- Sexual and reproductive health information