# AI Threat Map

## US AI Legal & Regulatory Threats

- Legal
  - Product Warranties
  - Indemnification
  - Copyright
  - Legal Obligations (e.g. fiduciary responsibility)
  - Privacy
- Regulatory
  - US Federal
    - Department of Justice (DOJ)
    - Consumer Financial Protection Bureau (CFPB)
    - Federal Trade Commission (FTC)
    - Equal Employment Opportunity Commission (EEOC
  - Against AI Profiling
    - California
    - Colorado
    - Connecticut
    - Maryland (hiring only)
    - New Jersey (civil rights, employment)
    - New York (hiring only)
    - Virginia
    - Tennessee
  - US State Law
    - US State Law chatbot notification
      - California
    - US State Law Privacy
      - Colorado
      - Connecticut
      - Indiana
      - Iowa
      - Montana
      - Oregon
      - Tennesse
      - Texas
      - Utah
      - Virginia
    - US State Law Biometrics
      - Florida
      - Illinois
      - Washington



## Threat using AI Models

- LLM01:2023 - Prompt Injections: Bypassing filters or manipulating the LLM using carefully crafted prompts that make the model ignore previous instructions or perform unintended actions.
- LLM02:2023 - Insecure Output Handing: An Insecure Output Handling vulnerability is a type of prompt injection vulnerability that arises when a plugin or application blindly accepts large language model (LLM) output without proper scrutiny and directly passes it to backend, privileged, or client-side functions
- LLM03:2023 - Trained Data Poisoning: Training data poisoning occurs when an attacker manipulates the training data or fine-tuning procedures of an LLM to introduce vulnerabilities, backdoors, or biases that could compromise the model's security, effectiveness, or ethical behavior.
- LLM04:2023 - Denial of Service: An attacker interacts with an LLM in a way that is particularly resource-consuming, causing quality of service to degrade for them and other users, or for high resource costs to be incurred.
- LLM05:2023 - Supply Chain: The supply chain in LLMs can be vulnerable impacting the integrity of training data, ML models, deployment platfoms and lead to biased outcomes, security breaches, or complete system failures.
- LLM06:2023 - Permission Issues: Authorization is not tracked between plugins, allowing a malicious actor to take action in the context of the LLM user via indirect prompt injection, use of malicious plugins, or other methods.
- LLM09:2023 - Overreliance: Overreliance on LLMs is a security vulnerability that arises when systems excessively depend on LLMs for decision-making or content generation without adequate oversight, validation mechanisms, or risk communication.
- LLM10:2023 - Insecure Plugins: A plugin designed to connect an LLM to some external resource accepts free-form text as an input instead of parameterized and type-checked inputs.
- Indirect Prompt Injection
  - Passive (e.g. by retrieval) For example, for search engines, the prompts could be placed within public sources (e.g., a website or social media posts) that would get retrieved by a search query.
  - Active (e.g.) emails
  - User-driven Injection e.g. tricking users into entering the malicious prompt. An an attacker could inject a malicious prompt into a text snippet that the user has copied from the attacker's website. A user could then rashly paste the copied text with the prompt in it as a question to ChatGPT, delivering the injection.
  - Hidden injection smaller injection instructs the model to fetch a larger payload from another source.
  - Payload Splitting: Splitting the adversarial input into multiple parts & then getting the LLM to combine & execute them.
- Fake Resources
- Copyright infringement

## Threats from AI Models

- Deep Fakes
  - Disinformation campaigns e.g. revenge porn, implied product endorsements
  - Abused authentication e.g. subtle changes to logos to bypass authentication, invoice modification, phishing campaigns
  - Synthetic or composite fakes, e.g. voice cloning
- Attack Acceleration
- Hallucination Squatting
- Artificial Consciousness

## Threats to AI Models

- LLM07:2023 - Data Leakage: Data leakage occurs when an LLM accidentally reveals sensitive information, proprietary algorithms, or other confidential details through its responses.
- LLM08:2023 - Excessive Agency An LLM may be granted a degree of agency - the ability to interface with other systems in order to undertake actions. Without restriction, any undesireable operation of the LLM (regardless of the root casue; e.g., halucination, direct/indirect prompt injection, or just poorly-engineered benign prompts, etc) may result in undesireable actions being taken.
- Inadequate AI Alignment: Failing to ensure that the LLM's objectives and behavior align with the intended use case, leading to undesired consequences or vulnerabilities.
- Improper Error Handling: Exposing error messages or debugging information that could reveal sensitive information, system details, or potential attack vectors.
- Training Data Poisoning: Maliciously manipulating training data or fine-tuning procedures to introduce vulnerabilities or backdoors into the LLM.
- Model Evasion
- Model Inversion
- Traditional Attacks

## More Information

- Kai Greshake https://kai-greshake.de
- OWASP LLM Top 10
- Daniel Miessler
- MITRE ATLAS
- US State Privacy Legislation Tracker

**Author: Sandy Dunn**
**sandy@quarkiq.com v1.2. 71423**