CoSD AI Product Risk Management Framework

1. Note that these general guidelines have been adopted from NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) and are intended to enable the County to have a framework around which AI products might be evaluated for procurement and their risks effectively managed once they are procured, integrated into a solution and placed into production operation.

2. When Peraton or the County is evaluating an AI product for procurement leveraging the RFP through ITO process, the criteria in Item 4 below should be considered in the product review process.

3. As part of Solution Design Document (SDD) development involving an AI product; the assigned security architect must:
    a. Use the County *Generative AI Flow Decision Tree* to classify the AI product being used in the solution and how it's inputs and outputs will be managed.
    b. Fill out the *County AI Security Checklist* classifying:
        i. The specific AI cybersecurity risks (see item 5iii 1-10) the selected AI product is vulnerable to and how they will be mitigated and
        ii. The specific trustworthiness attributes (see item 5a-d) of the AI product
    c. Determine whether the NIST AI RMF (based on NIST Govern, Map, Measure, Manage concepts) needs to be employed to manage AI product risk.

4. When an AI product is to be evaluated by County; seven criteria should be assessed related to the trustworthiness of the AI product under consideration:
    a. Safety
        i. AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022).
    b. Security and Resilience
        i. AI systems may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary (Adapted from: ISO/IEC TS 5723:2022).
        ii. Security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks.
        iii. Threats which must be mitigated against include (from OWASP, Open Worldwide Application Security Project, Top 10):

1. Prompt Injection: This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources
2. Insecure Output Handling: This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like Cross-Site Scripting (XSS), Cross-Site Request Forgery (CSRF), Server-Site Request Forgery (SSRF), privilege escalation, or remote code execution.
3. Training Data Poisoning: This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, and books.
4. Model Denial of Service: Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs
5. Supply Chain Vulnerabilities: LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins can add vulnerabilities.
6. Sensitive Information Disclosure: LLMs may inadvertently reveal confidential data in their responses, leading to unauthorized data access, privacy violations, and security breaches. It is crucial to implement data sanitization and strict user policies to mitigate this.
7. Insecure Plugin Design: LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.
8. Excessive Agency: LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.
9. Overreliance: Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLM.
10. Model Theft: This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes

economic losses, compromised competitive advantage, and potential access to sensitive information.

iv. **To a significant extent, although the County may have an expectation that an AI solution will be "secure" "out of the box," it appears more and more as if the organizations deploying the AI solution will need to deploy additional mechanisms to guard against some of the 10 aforementioned OWASP attacks. The most utilitarian existing cybersecurity tool applicable to this threat vector set may well be Data Loss Protection (DLP) tools. See tools like AWS Comprehend for example.**

c. Explainable and Interoperable
    i. Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes. In other words, users of an AI system need to understand not only the meaning of an AI solution's response to a prompt, but also why an AI solution has responded to a prompt in a manner in which it has (e.g., like asking an "old" AI-based expert system to explain why it made a decision).
d. Privacy-Enhanced
    i. Privacy generally refers to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation). The implication here is that trainers of an AI solution need to control personally identifiable information (PII) and all other sensitive information which might be contained in training data sets and users of an AI solution must control the sensitive information which might be used as prompt input or which may be provided by the AI solution as a prompt response.
e. Fair With Harmful Bias Managed
    i. Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Organizations' risk management efforts will be enhanced by recognizing and considering these differences. AI systems become biased because their training data was such. Again, mitigating against this threat involves carefully reviewing the input data in the training sets. It is unclear whether a DLP solution could be configured to detect bias input and as such the context of

how the training data was obtained needs to be closely scrutinized before input to the AI solution.

f. Accountable and Transparent
    i. The Meriam Webster Dictionary defines accountable as: 1) capable of being explained **: EXPLAINABLE**; 2) subject to giving an account **: ANSWERABLE**.
    ii. Transparent AI solutions are those with available and understandable internal processes regarding decisions and output.
g. Valid and Reliable
    i. Valid means confirmed "through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015).
    ii. Reliable means able to "perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022)

5. Regardless of the evaluation scores with respect to the above criteria, all AI products used by the County involve risk. NIST proposes a methodology for managing this risk by leveraging the following four-point process:
    a. GOVERN
        i. Anticipate, identify, and manage the risks a system can pose. Provide a structure by which AI risk management functions can align with organizational principles, policies, and strategic priorities. This is a function that is infused throughout AI risk management and enables the other functions of the process. Aspects of GOVERN, especially those related to compliance or evaluation, should be integrated into each of the other functions. Attention to governance is a continual and intrinsic requirement for effective AI risk management over an AI system's lifespan and the organization's hierarchy
            1. Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.
            2. Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks
            3. Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.
            4. Organizational teams are committed that considers and communicates AI risk.

5. Processes are in place for robust engagement with relevant AI actors.
6. Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.

b. MAP
   i. The MAP function establishes the context to frame risks related to an AI system. The information gathered while carrying out the MAP function enables negative risk prevention and informs decisions for processes such as model management, as well as an initial decision about appropriateness or the need for an AI solution. Outcomes in the MAP function are the basis for the MEASURE and MANAGE functions.
      1. Context is established and understood.
      2. Categorization of the AI system is performed.
      3. AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.
      4. Risks and benefits are mapped for all components of the AI system including third-party software and data.
      5. Impacts to individuals, groups, communities, organizations, and society are characterized.

c. MEASURE
   i. The MEASURE function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. It uses knowledge relevant to AI risks identified in the MAP function and informs the MANAGE function. AI systems should be tested before their deployment and regularly while in operation. AI risk measurements include documenting aspects of systems' functionality and trustworthiness
      1. Appropriate methods and metrics are identified and applied.
      2. Systems are evaluated for trustworthy characteristics.
      3. Mechanisms for tracking identified AI risks over time are in place.
      4. Feedback about efficacy of measurement is gathered and assessed

d. MANAGE
   i. The MANAGE function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the GOVERN function. Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events

1. AI risks based on assessments and other analytical output from the MAP and MEASURE functions are prioritized, responded to, and managed.
2. Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.
3. risks and benefits from third-party entities are managed.
4. Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly