LLM MATURITY MODEL FOR AI RISK MANAGEMENT IN THE DOD

Ashley Gray, AI/ML Solutions Architect
21 FEB 2024

AGENDA

- Challenges and Considerations
- Design for Use as a Risk Management Framework
- Model Application Example
- Conclusion

LLM MATURITY MODEL

CHALLENGES AND CONSIDERATIONS

Supermarket AI meal planner app suggests recipe that would create chlorine gas

Pak 'n' Save's Savey Meal-bot cheerfully created unappealing recipes when customers experimented with non-grocery household items

“Andrew” is willing to help with just about anything.

IMPLICATIONS OF IMMATURE AI IN THE DOD

Consequences of Operator Credulity

- AI hallucinations
- Biased Outputs
- Unsafe/Deadly Suggestions

Abuse of Technology

- Deepfakes
- Ransomware
- Uncontrolled Development of Weapons

Advancement of Adversaries

- Development of newer, more capable technologies
- Espionage/Embedded AI

Adversarial AI

- Data Poisoning attacks
- Exposure of test data
- DDoS attacks

Unknown Unknowns

CONSIDERATIONS FOR MATCHING LLM MATURITY TO TASK

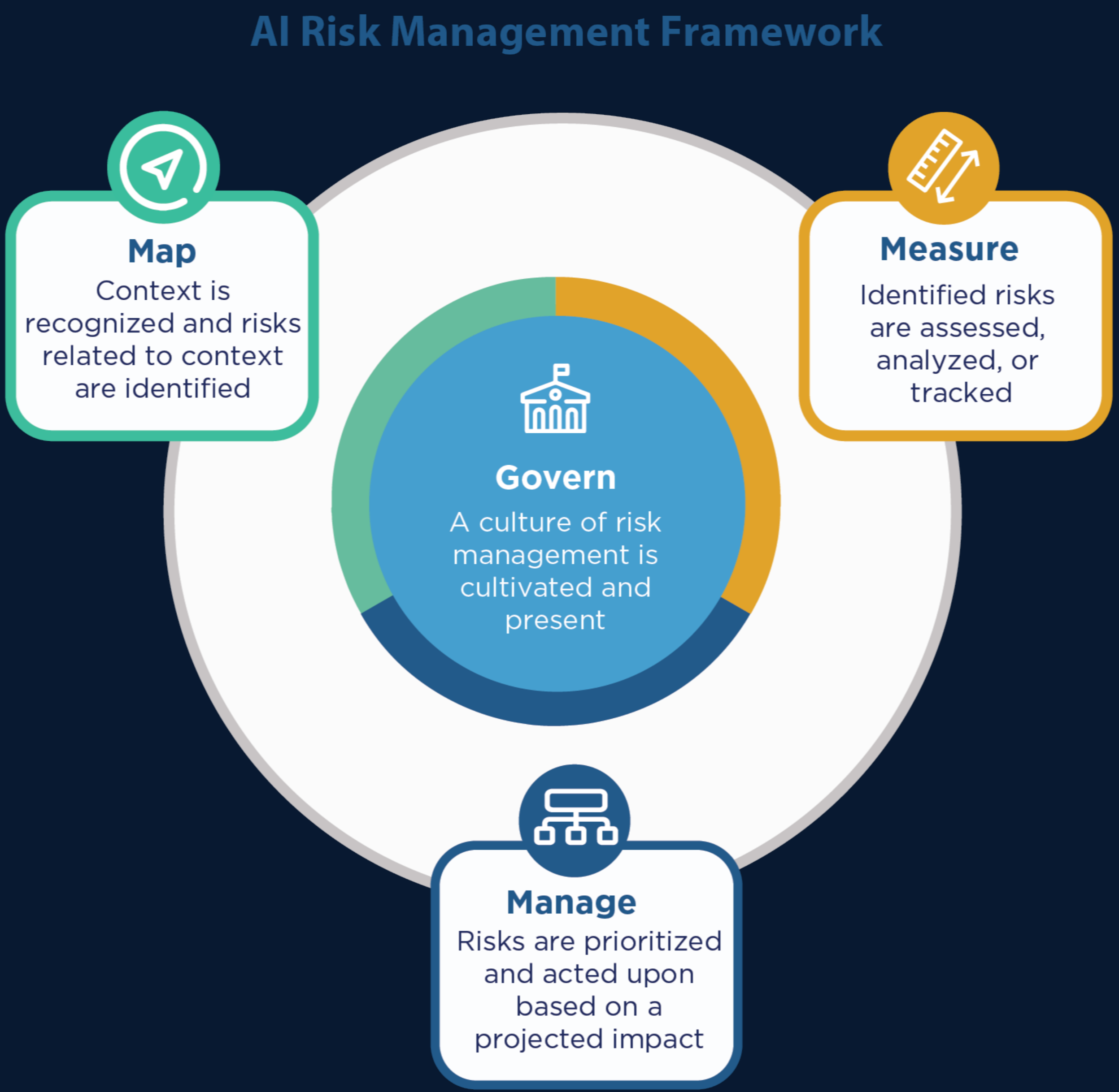
- Multiple Models in a Workstream
- Complexity of Data Transformations
- Deployment Environment
- Multimodality
- Governing Situational/Subjective Characteristics

LLM MATURITY MODEL

DESIGN FOR USE AS A RISK MANAGEMENT FRAMEWORK

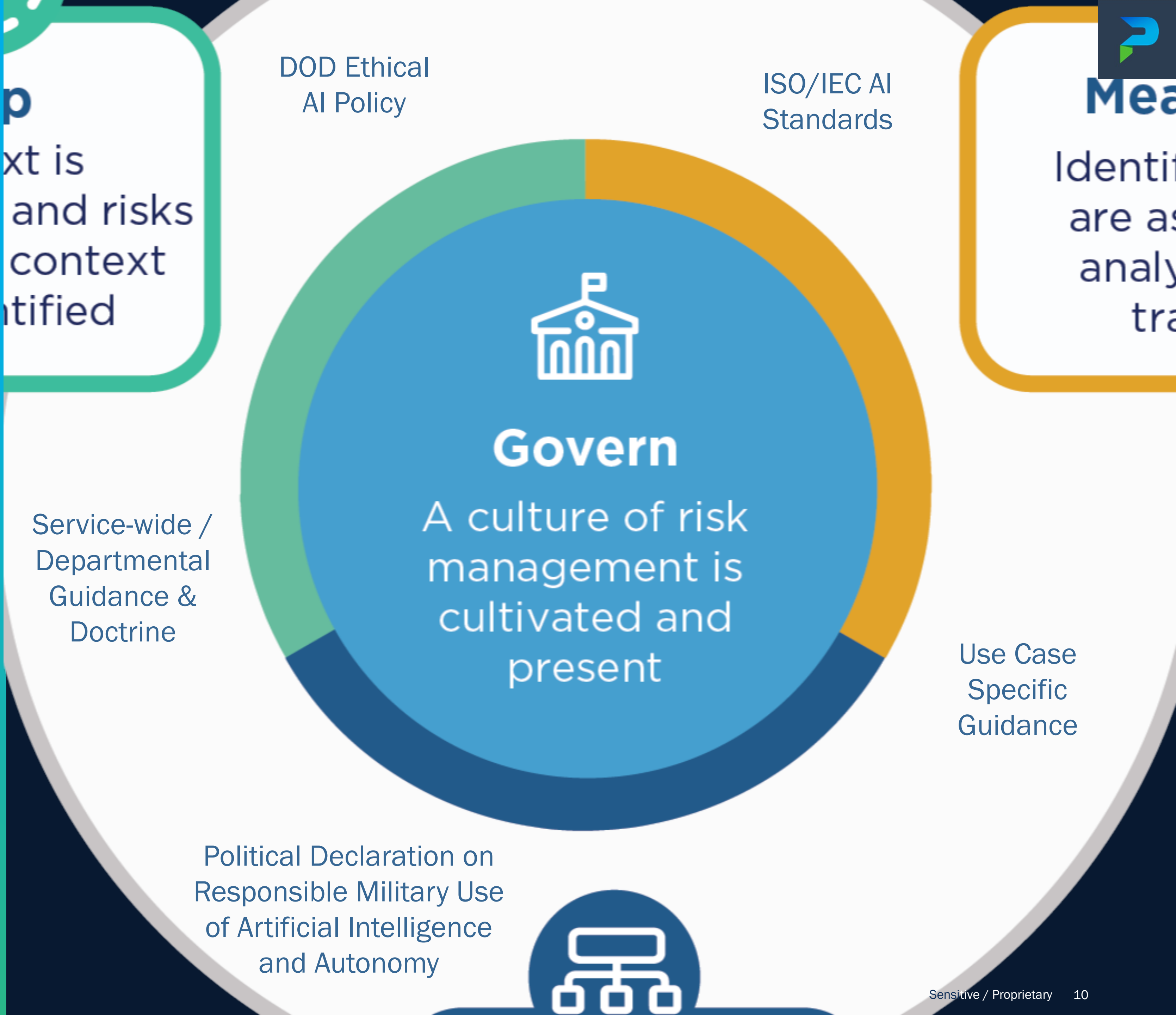
NIST'S AI RMF

Recommendation: Approach LLM MM design with AI RMF in mind



NIST'S AI RMF

Recommendation: Approach LLM MM design with AI RMF in mind



NIST'S AI RMF

Recommendation: Approach LLM MM design with AI RMF in mind

Defining Risks / Areas of Concern for LLM Deployment

- Data information and Integrity
- Output Quality and Human Factors
- Resistance to Adversarial Attacks and Exploitation
- Technological Robustness

NIST'S AI RMF

Recommendation: Approach LLM MM design with AI RMF in mind

- Quantitative measures & evaluation tools for measuring model risks & characteristic performance
- Defining Basic, Developing, and Mature characteristic performance for mapping into LLM and LLM task profiles.

NIST'S AI RMF

Recommendation: Approach LLM MM design with AI RMF in mind

Continuous improvement through reviews of updated standards, guidance, and governance, learning from operator feedback, and adjusting model characteristic performance criteria.

LLM MATURITY MODEL

STRUCTURE AND DEVELOPMENT METHODOLOGY

DEFINING MATURITY LEVELS FOR MODEL CHARACTERISTICS

1 Basic

- Minimum acceptable characteristic fitness
- May be significantly below benchmark
- Actionable intelligence
- Characteristics evaluated below this level are **UNACCEPTABLE**

2 Developing

- Characteristic fitness is slightly below or near average human performance
- Characteristic evaluation is at or near benchmark
- Not suitable for unsupervised operations

3 Mature

- Characteristic fitness is near or above upper echelon of human performance
- Characteristic evaluation meets or surpasses benchmark level
- Suitable for unsupervised operations

NOTE: Some environment-dependent performance characteristics will not have maturity levels and will depend on developer or scope requirements.

UNIVERSALLY GOVERNABLE CHARACTERISTICS

(top) LLM Tasks / (left) LLM Characteristics	Conversation	Language Translation	Text Summarization	Question Answering	Information Extraction	Labeling/ Classification	Programming/ Code Generation	Speech-to-Text Transcription	Symbolic Reasoning (Math)	Logical Reasoning (Inductive/ Deductive)
Bias and Fairness										
Toxicity Detection										
Traceability										
Language Comprehension										
Reading Comprehension										
Knowledge										
Reasoning										
Intuition										
Coherence										
Completeness										
Novelty										
Truthfulness										
Robustness										

DEFINING TASK MATURITY REQUIREMENTS

USE CASE DEPENDENT CHARACTERISTICS

(top) LLM Tasks / (left) LLM Characteristics	Conversation	Language Translation	Text Summarization	Question Answering	Information Extraction	Labeling/ Classification	Programming/ Code Generation	Speech-to-Text Transcription	Symbolic Reasoning (Math)	Logical Reasoning (Inductive/ Deductive)
Symbolic Reasoning *										
Vision-Language Understanding *										
Multilingual Support *										

Multimodality
Scalability
Training Time
Training Resource Requirements
Training Inference Costs
Deployed Model Resource Requirements
Deployment Inference Costs

Not governable on the task level.

WORKSTREAM MAPPING

Candidate Workstreams

- Text-Media Content Creation
- Document/Repository Search
- Document Source Retrieval
- Document Classification
- Cipher Detection
- Event Prediction/Simulation
- .
- .
- .

LLM Tasks Required

- Conversation
- Text Summarization
- Information Extraction
- Labeling/Classification
- Programming/Code Generation
- .
- .
- .

LLM Characteristic Maturity Requirements

- Bias/Fairness
- Toxicity Detection
- Reading Comprehension
- Multilingual Support
- Novelty
- Truthfulness
- Robustness
- .
- .

Required LLM characteristic maturity for workstream:
Red – low
Yellow – medium
Green – high

LLM MATURITY MODEL

USE CASE PROFILING AND MODEL MATCHING EXAMPLE

EXAMPLE CIPHER DETECTION WORKSTREAM MODEL FITTING

Consider the example of developing a cipher detection tool for handheld or edge devices, and we have been tasked to choose the most mature of three fine-tuned models for incorporation into the workstream.

- Models are pre-trained and fine-tuned
- Will be deployed on a field-deployed, low SWaP device
- Must support multiple modalities
- Operators will examine and evaluate output for coherence and reasonability

(top) LLM Tasks / (left) LLM Characteristics	Speech-to-Text Translation	Logical Reasoning	Information Extraction	Use Case Summary	LLM 1 Maturity	LLM 2 Maturity	LLM 3 Maturity
Bias and Fairness							
Toxicity Detection							
Traceability							
Language Comprehension							
Reading Comprehension							
Knowledge							
Reasoning							
Intuition							
Coherence							
Completeness							
Novelty							
Truthfulness							
Robustness							
Symbolic Reasoning *							
Vision-Language Understanding *							
Multilingual Support *							
Multimodality							
Scalability							
Training Time							
Training Resource Requirements							
Training Inference Costs							
Deployed Model Resource Requirements							
Deployment Inference Costs							

MITIGATIONS FOR LACK OF MATURITY

Concluding thoughts for moving towards more mature LLM use cases.

- Consider multiple LLMs in a workstream.
- Set guardrails for usage of AI tools/systems
- Keep RAG knowledge bases current.
- Review and recycle operator feedback.
- Consider alternatives to LLMs/AI for use case.

THANK YOU

ASHLEY GRAY
AI/ML Solutions Architect / D&I
She/Her/Hers
ashley.gray@parsons.us