NAVAL
POSTGRADUATE
SCHOOL

Safety in AI-Enabled Warfare Decision Aids

Spring 2022 AAAI
SafeAI Workshop

1 March 2022

Dr. Bonnie Johnson
Naval Postgraduate School
Systems Engineering
bwjohnson@nps.edu

Advances in computational thinking and data science have led to a new era of artificial intelligence systems being engineered to adapt to complex situations and develop actionable knowledge. These learning systems are meant to reliably understand the essence of a situation and construct critical decision recommendations to support autonomous and human-machine teaming operations.

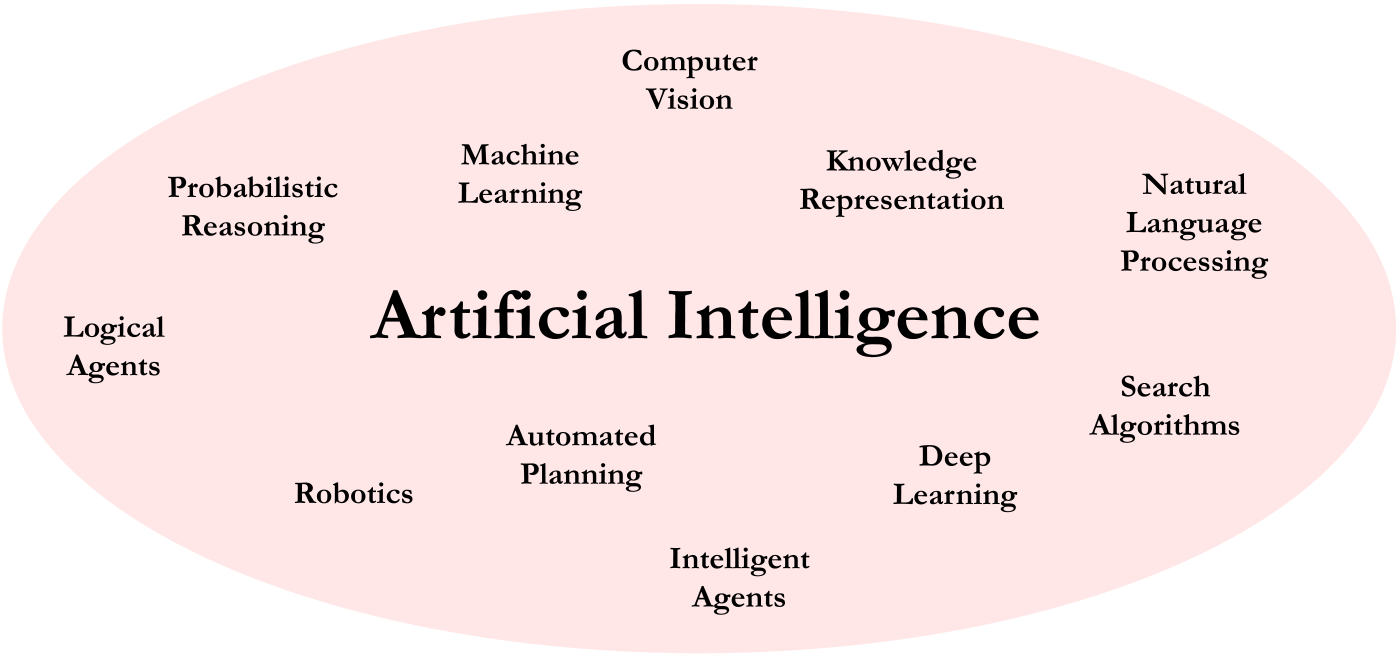
In parallel, the increasing volume, velocity, variety, veracity, value, and variability of data is confounding the complexity of these new systems – creating challenges in terms of their development and implementation. For artificial systems supporting critical decisions with higher consequences, **safety** has become an important concern. Methods are needed to avoid failure modes and ensure that only desired behavior is permitted.

What is AI?

Here's a good definition:

AI is a field that includes many different approaches with the objective of creating machines with intelligence[1]

[1] – Melanie Mitchell. 2019. Artificial Intelligence – A Guide for Thinking Humans Picador: New York. – definition of AI as a field



Artificial Intelligence

Computer Vision

Probabilistic Reasoning

Machine Learning

Knowledge Representation

Natural Language Processing

Logical Agents

Search Algorithms

Automated Planning

Robotics

Deep Learning

Intelligent Agents

Artificial Intelligence includes:

Knowledge Representation & Reasoning

- Think "if-then," but can be more complex
- Explicitly programmed
- Can involve complex manually designed coding schemes for data / knowledge
- Based on graphs and ontologies
- Sometimes referred to as handcrafted knowledge systems[2]

Machine Learning

- The system learns (trains) from a large amount of data
- The system learns patterns by trial-and-error until it can predict the labeled examples
- Includes supervised, unsupervised, reinforcement, and deep learning
- The "trained" system can be used (for prediction) given new data

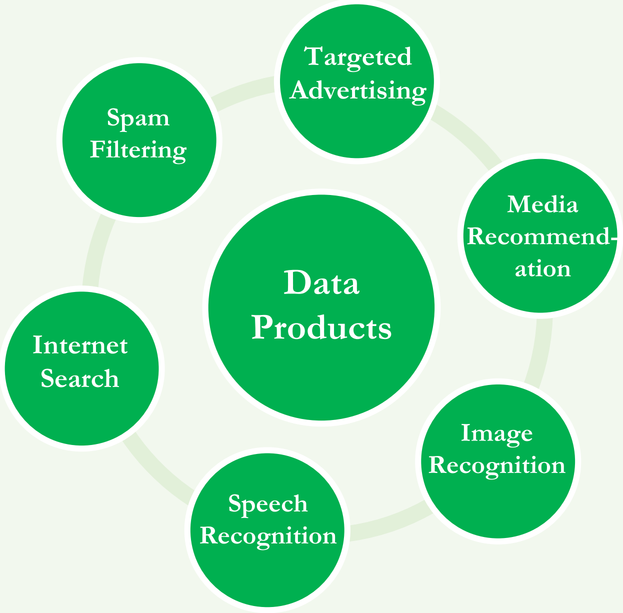
[2] – Greg Allen. 2020. Understanding AI Technology. Joint AI Center (JAIC) Report, US Dept of Defense – definition of handcrafted knowledge systems

Four characterizations of AI

Thinking Humanly "the exciting new effort to make computers think…machines with minds, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning…" (Bellman, 1978)	**Thinking Rationally** "The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly "The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	**Acting Rationally** "Computational intelligence is the study of the design of intelligent agents." (Poole et al., 1998) "AI…is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Source: Russell and Norvig (2015, Figure 1.1)

Three Types of AI System Application Domains



Data Products

Targeted Advertising

Media Recommendation

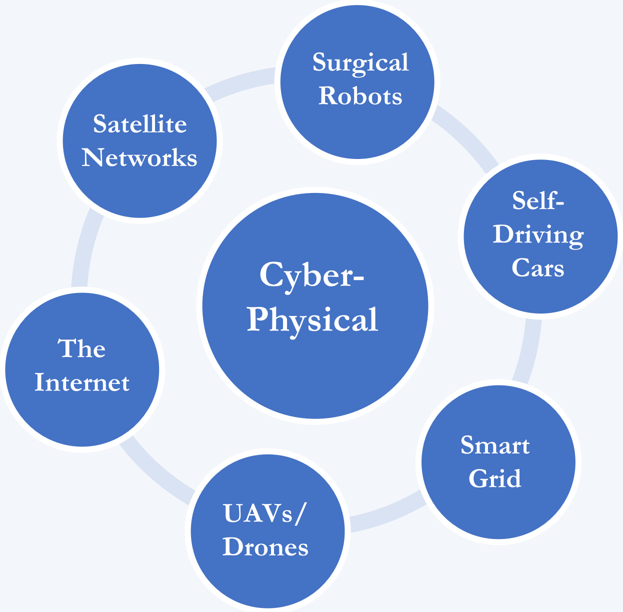
Image Recognition

Speech Recognition

Internet Search

Spam Filtering

Data product systems use computers to generate information products.



Cyber-Physical

Surgical Robots

Self-Driving Cars

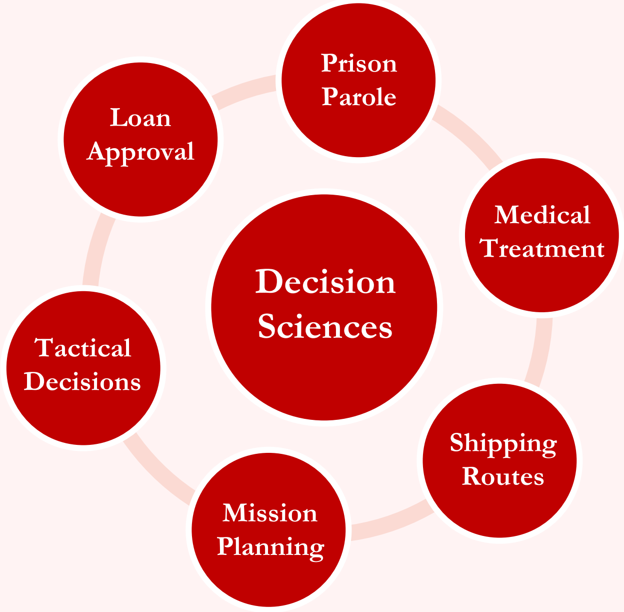
Smart Grid

UAVs/ Drones

The Internet

Satellite Networks

Cyber-physical systems include computer automation (often AI) and physical components.



Decision Sciences

Prison Parole

Medical Treatment

Shipping Routes

Mission Planning

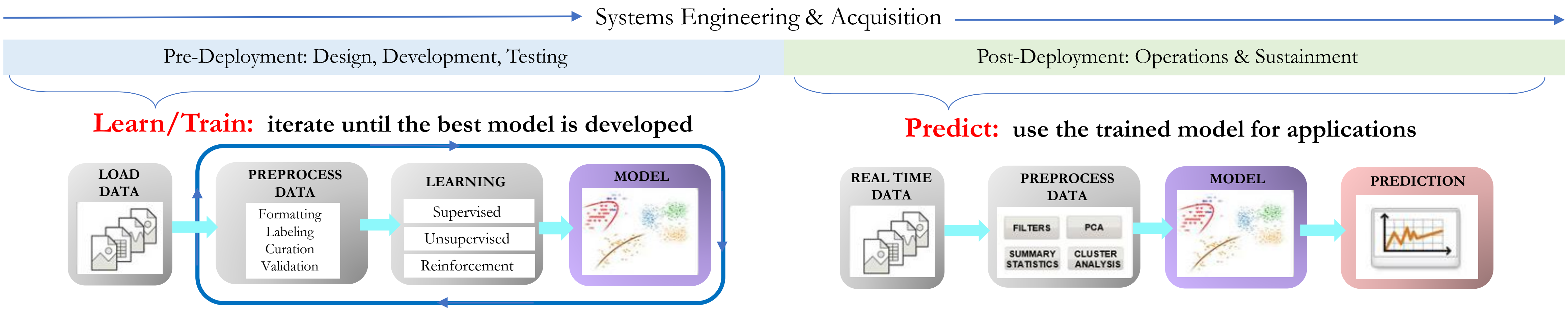
Tactical Decisions

Loan Approval

Decision science systems use computer algorithms to automate the process of making decision and advising plans and strategies.

Each application domain contains its own range of possible failure modes, and each will require tailored safety solution measures.

Machine learning systems introduce a new set of challenges



Systems Engineering & Acquisition

Pre-Deployment: Design, Development, Testing

Post-Deployment: Operations & Sustainment

Learn/Train: **iterate until the best model is developed**

LOAD DATA

PREPROCESS DATA

Formatting
Labeling
Curation
Validation

LEARNING

Supervised
Unsupervised
Reinforcement

MODEL

Predict: **use the trained model for applications**

REAL TIME DATA

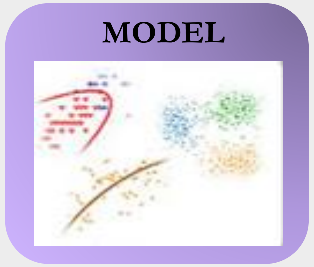
PREPROCESS DATA

FILTERS
PCA
SUMMARY STATISTICS
CLUSTER ANALYSIS

MODEL

PREDICTION

The machine learning "system" is the trained model



MODEL

Characteristics of ML Systems:

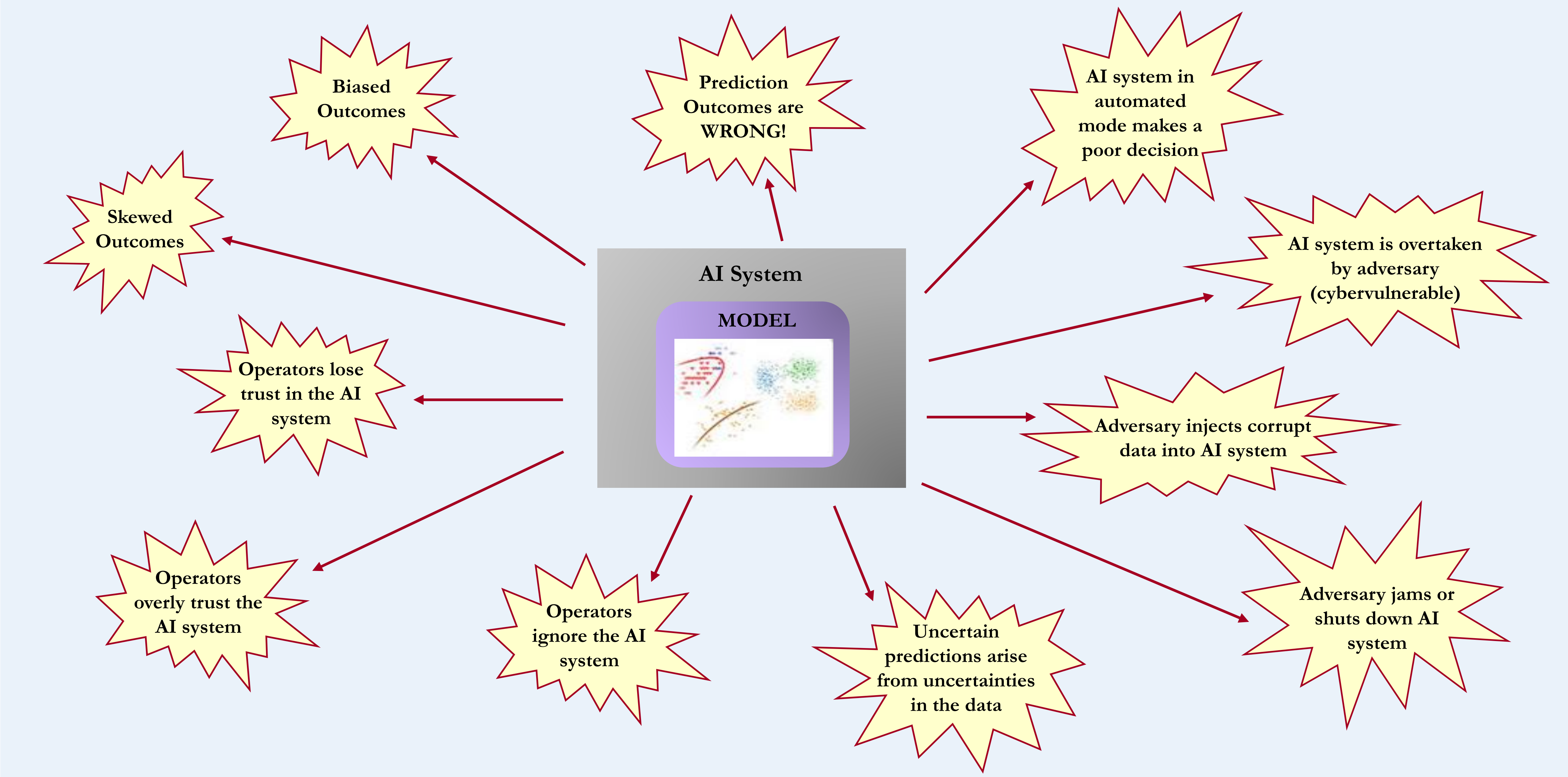
Non-Deterministic – ML is a technique that allows a computer to learn a task without being explicitly programmed. The ML system implements inductive inference on real-time or operational data sets after being trained. Therefore, ML system behavior leads to variability in results.

Intimately Connected to Data – ML systems "emerge" or are generated through the process of learning on training data sets. They are a product of the quality, sufficiency, and representativeness of the data. They are intimately connected and wholly dependent on their training data.

Complex – ML systems can exhibit complex behavior due to deep learning (the ML system consists of networks of many learning sub-components) and complex mathematical operations involving very large datasets and computations. The complex (unexpected) behavior can emerge.

Intimately Connected to Context – During operations, the behavior of ML systems is highly dependent on the context, or operational situation. Uncertainty in data representations of situational awareness, will lead to ML system prediction error. Complexity in the operational situation will lead to complex ML system operations.

Failure Modes



Biased Outcomes
Prediction Outcomes are WRONG!
AI system in automated mode makes a poor decision
Skewed Outcomes
AI System
MODEL
AI system is overtaken by adversary (cybervulnerable)
Operators lose trust in the AI system
Adversary injects corrupt data into AI system
Operators overly trust the AI system
Operators ignore the AI system
Uncertain predictions arise from uncertainties in the data
Adversary jams or shuts down AI system

Consequences

Two types of AI systems according to the severity of their failure consequences

Type A
Safety is Paramount

Applications in which AI system model predictions are used to support consequential decisions that can have a profound effect on people's lives

Type B
Safety is Less Important

Applications in which AI system model predictions are used in settings of low consequence and large scale that have minimal effects on people's lives

Root Causes

Systems Engineering & Acquisition

Pre-Deployment: Design, Development, Testing

Bias in the training data sets

Incompleteness---data sets don’t represent all scenarios

Rare examples – data sets don’t include unusual scenarios

Corruption in the training data sets

Mis-labeled data

Mis-associated data

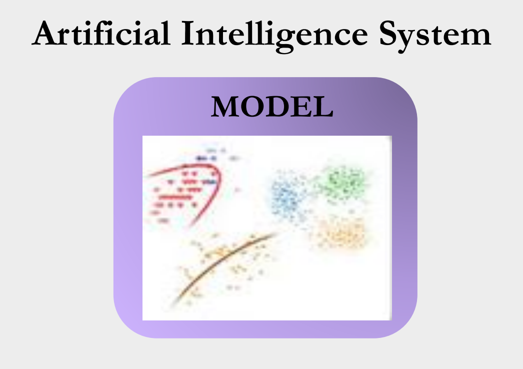
Poor validation methods (is there criteria for deciding how much training data is good enough?)

Poor data collection methods

Underfitting in the model – when the model is not capable of attaining sufficiently low error on the training data

Cost function algorithm errors – when trained model is optimized to the wrong cost function

Wrong algorithm – when the training data is fit to the wrong algorithmic approach (regression neural network, etc.)



Artificial Intelligence System

MODEL

Post-Deployment: Operations & Sustainment

Uncertainty/error in operational datasets

Corruption in operational datasets

Inaccuracy in the algorithm model (prediction error)

Operational complexity that overwhelms the AI system

Overfitting – when the model presents a very small error on the training data but fails to generalize, i.e., fails to perform as well on new examples; the model is “overfit” to the training data

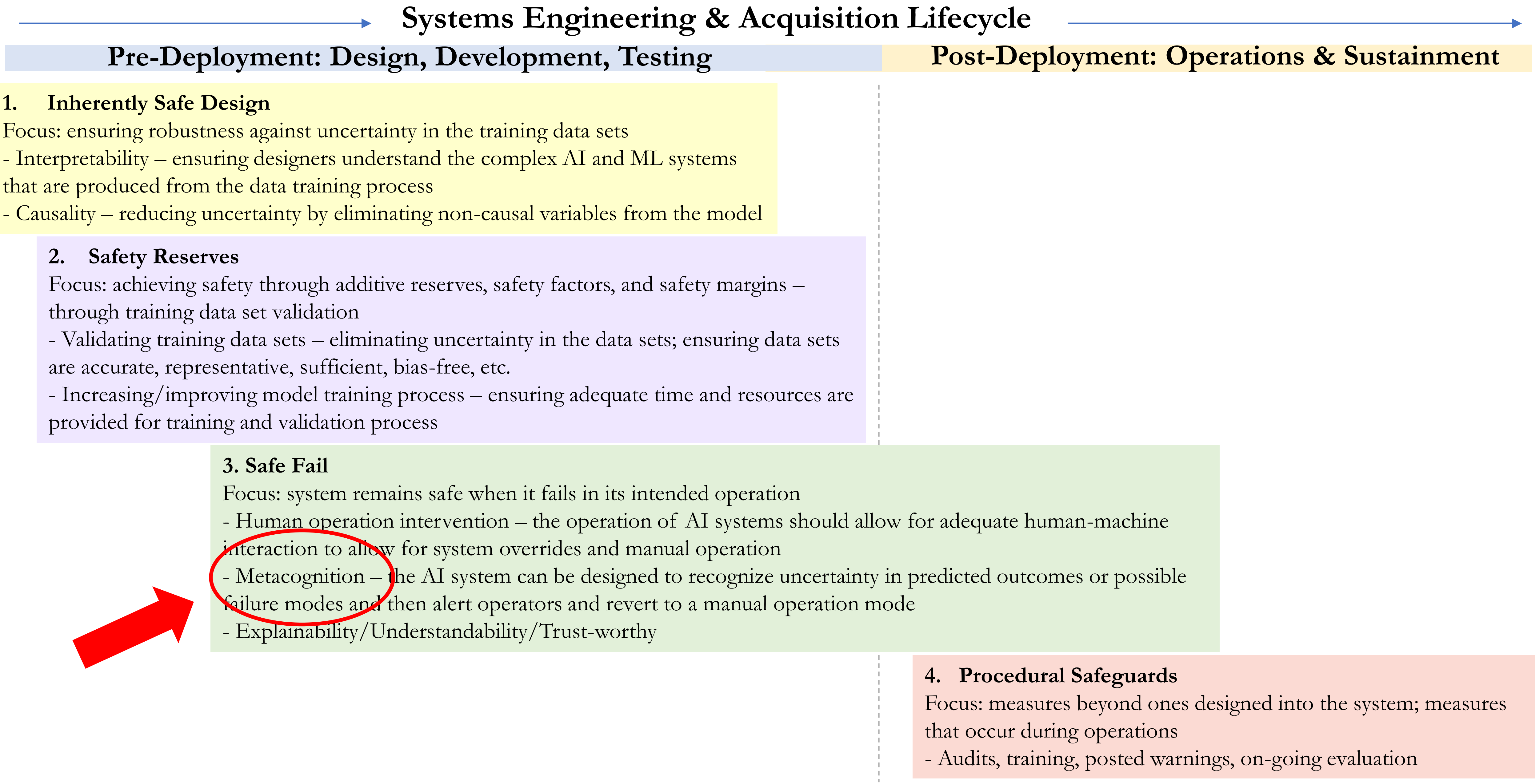
Lack of explainability

Trust issues

Operator-induced error

Adversarial attacks – hacking, deception, inserting false data, controlling automated systems

AI System Safety: Four Types of Solution Strategies



Systems Engineering & Acquisition Lifecycle

Pre-Deployment: Design, Development, Testing

Post-Deployment: Operations & Sustainment

1. Inherently Safe Design

Focus: ensuring robustness against uncertainty in the training data sets

- Interpretability – ensuring designers understand the complex AI and ML systems that are produced from the data training process
- Causality – reducing uncertainty by eliminating non-causal variables from the model

2. Safety Reserves

Focus: achieving safety through additive reserves, safety factors, and safety margins – through training data set validation

- Validating training data sets – eliminating uncertainty in the data sets; ensuring data sets are accurate, representative, sufficient, bias-free, etc.
- Increasing/improving model training process – ensuring adequate time and resources are provided for training and validation process

3. Safe Fail

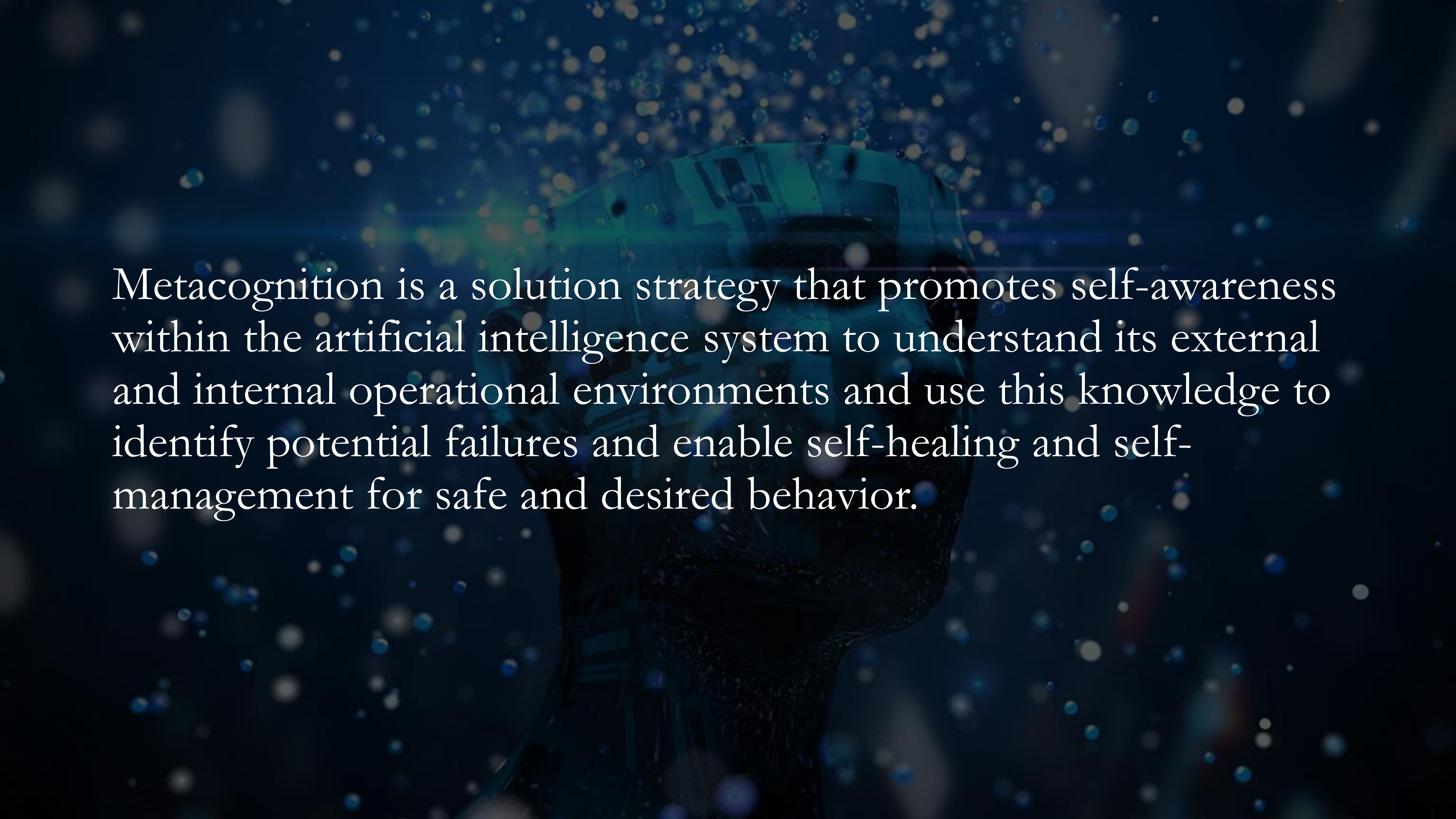
Focus: system remains safe when it fails in its intended operation

- Human operation intervention – the operation of AI systems should allow for adequate human-machine interaction to allow for system overrides and manual operation
- Metacognition – the AI system can be designed to recognize uncertainty in predicted outcomes or possible failure modes and then alert operators and revert to a manual operation mode
- Explainability/Understandability/Trust-worthy

4. Procedural Safeguards

Focus: measures beyond ones designed into the system; measures that occur during operations

- Audits, training, posted warnings, on-going evaluation



Metacognition is a solution strategy that promotes self-awareness within the artificial intelligence system to understand its external and internal operational environments and use this knowledge to identify potential failures and enable self-healing and self-management for safe and desired behavior.

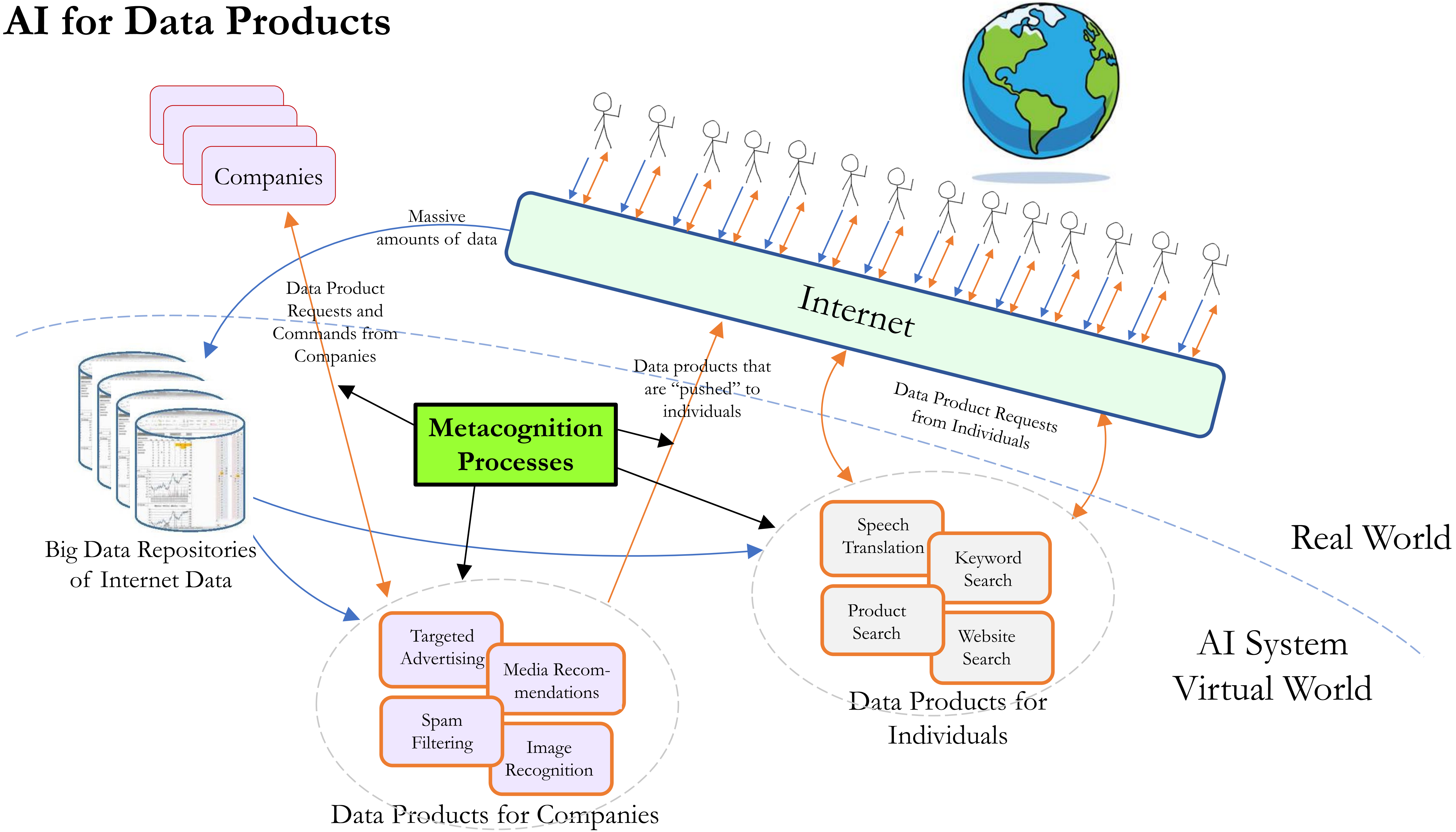
Metacognition as a safety measure

Johnson, Bonnie. "Metacognition for artificial intelligence system safety – an approach to safe and desired behavior." Submitted to the *Journal of Safety Science*. 2021.

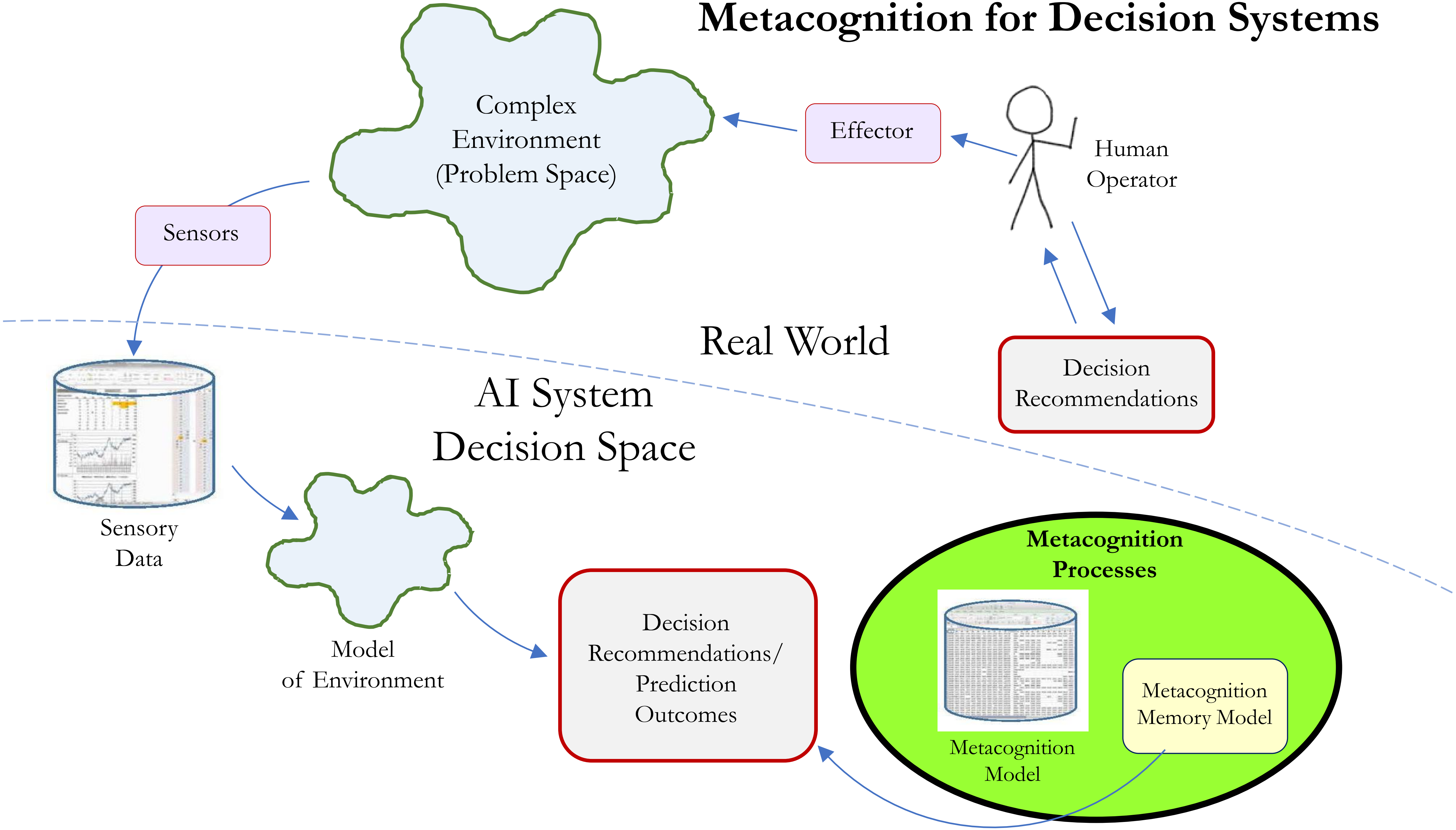
Metacognition Capabilities

1. Evaluating level of uncertainty in knowledge
2. Evaluating level of uncertainty in AI outputs
3. Failure self-predictions
4. Anomaly detection
5. Identification of new or unfamiliar situation
6. Evaluation of situation complexity
7. Constructionist learning: self-sufficient locus of control
8. Identification/prediction of high-risk courses of action
9. Identification/prediction of undesirable emergent behavior
10. Prediction of poor performance
11. Development of metacognitive memory
12. Evaluation of historical safety risks, failures, error, poor performance
13. Evaluation of contextual complexity, uncertainty, and unfamiliarity
14. Evaluation of individual component failures

AI for Data Products



Metacognition for Decision Systems



Complex Environment (Problem Space)

Effector

Human Operator

Sensors

Real World

Decision Recommendations

AI System Decision Space

Sensory Data

Model of Environment

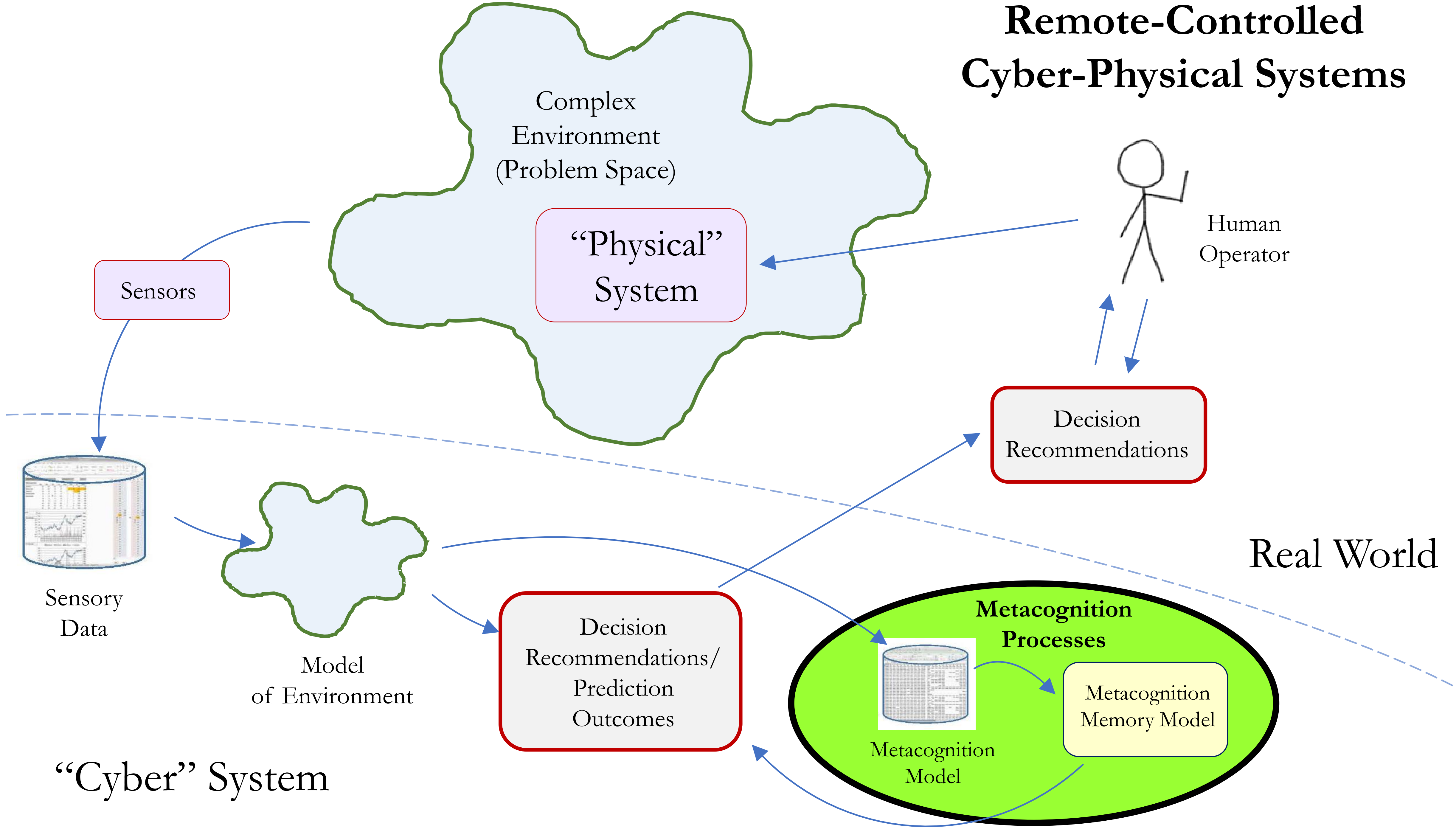
Decision Recommendations/ Prediction Outcomes

Metacognition Processes

Metacognition Model

Metacognition Memory Model

Remote-Controlled Cyber-Physical Systems



Complex
Environment
(Problem Space)

“Physical”
System

Sensors

Human
Operator

Decision
Recommendations

Sensory
Data

Model
of Environment

Real World

Decision
Recommendations/
Prediction
Outcomes

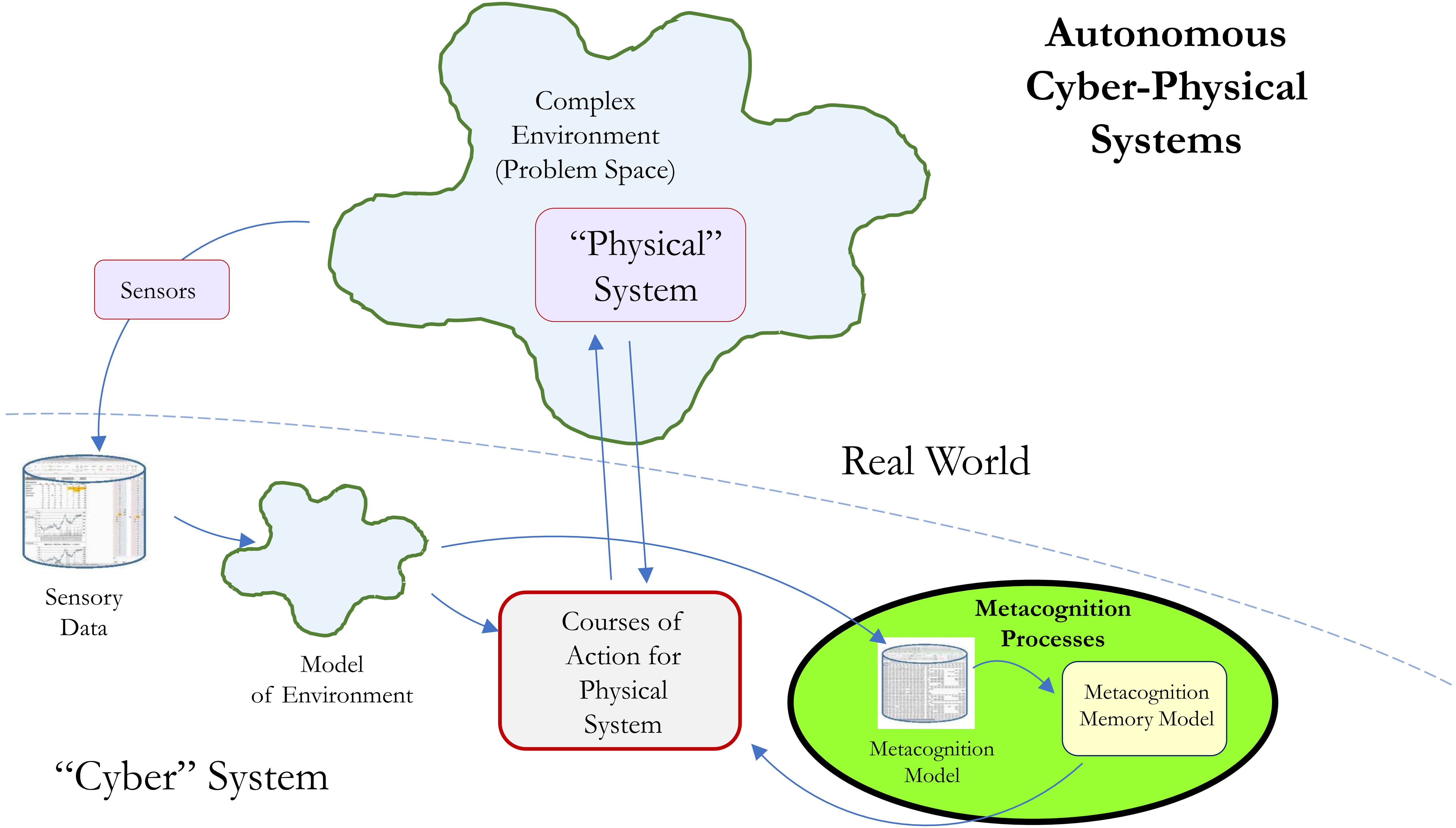
**Metacognition
Processes**

Metacognition
Model

Metacognition
Memory Model

“Cyber” System

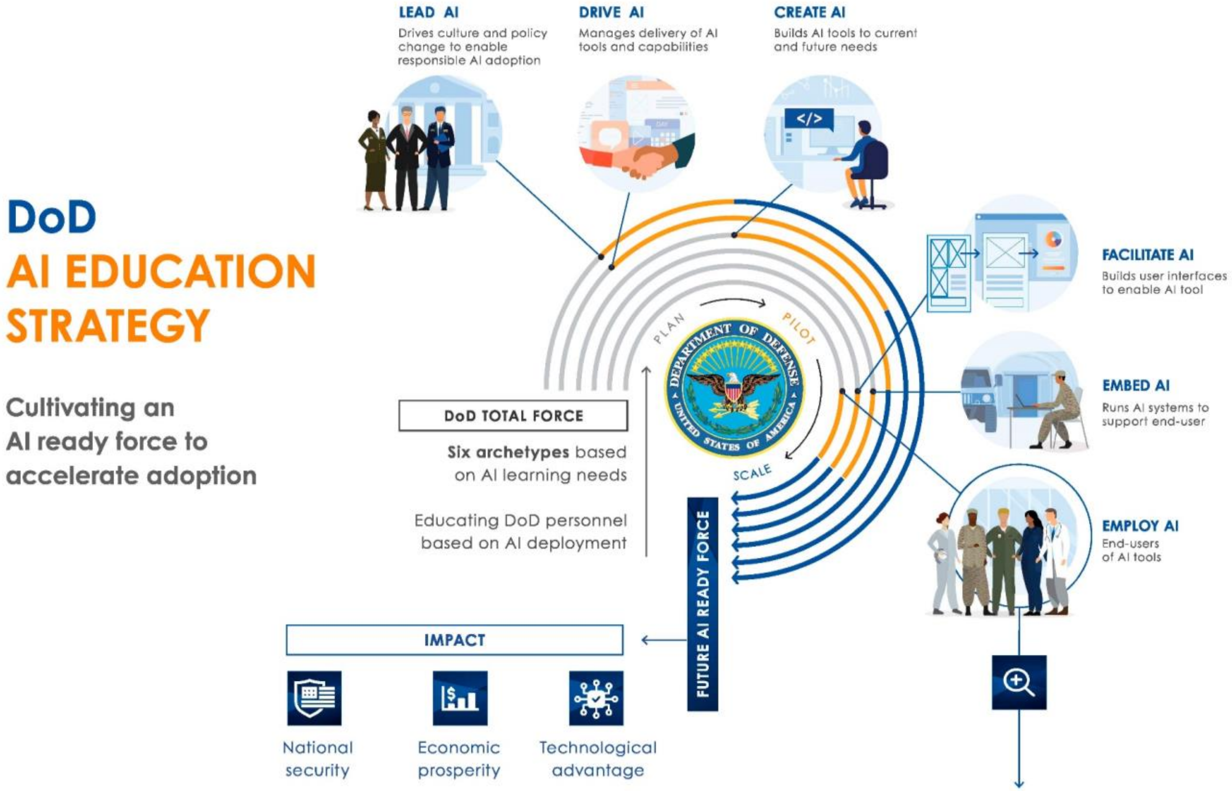
Autonomous Cyber-Physical Systems



Safety risks in developing and implementing AI-enabled warfare decision aids

U.S. DoD Joint Artificial Intelligence Center (JAIC)

The U.S. Department of Defense (DoD) established the Joint Artificial Intelligence Center (JAIC) in 2018 to focus on the broad enablement and implementation of AI capabilities within DoD.



DoD AI EDUCATION STRATEGY

Cultivating an AI ready force to accelerate adoption

LEAD AI
Drives culture and policy change to enable responsible AI adoption
DRIVE AI
Manages delivery of AI tools and capabilities
CREATE AI
Builds AI tools to current and future needs
FACILITATE AI
Builds user interfaces to enable AI tool
EMBED AI
Runs AI systems to support end-user
EMPLOY AI
End-users of AI tools
PLAN
PILOT
SCALE
DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA
DoD TOTAL FORCE
Six archetypes based on AI learning needs
Educating DoD personnel based on AI deployment
FUTURE AI READY FORCE
IMPACT
National security
Economic prosperity
Technological advantage

JAIC 5 Ethical Principles for AI:

DOD Adopts Ethical Principles for Artificial Intelligence, U.S. Department of Defense, 24 February 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts- ethical-principles-for-artificial-intelligence/

1. Responsible

DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

2. Equitable

DoD will take deliberate steps to minimize unintended bias in AI capabilities.

3. Traceable

DoD's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development process, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

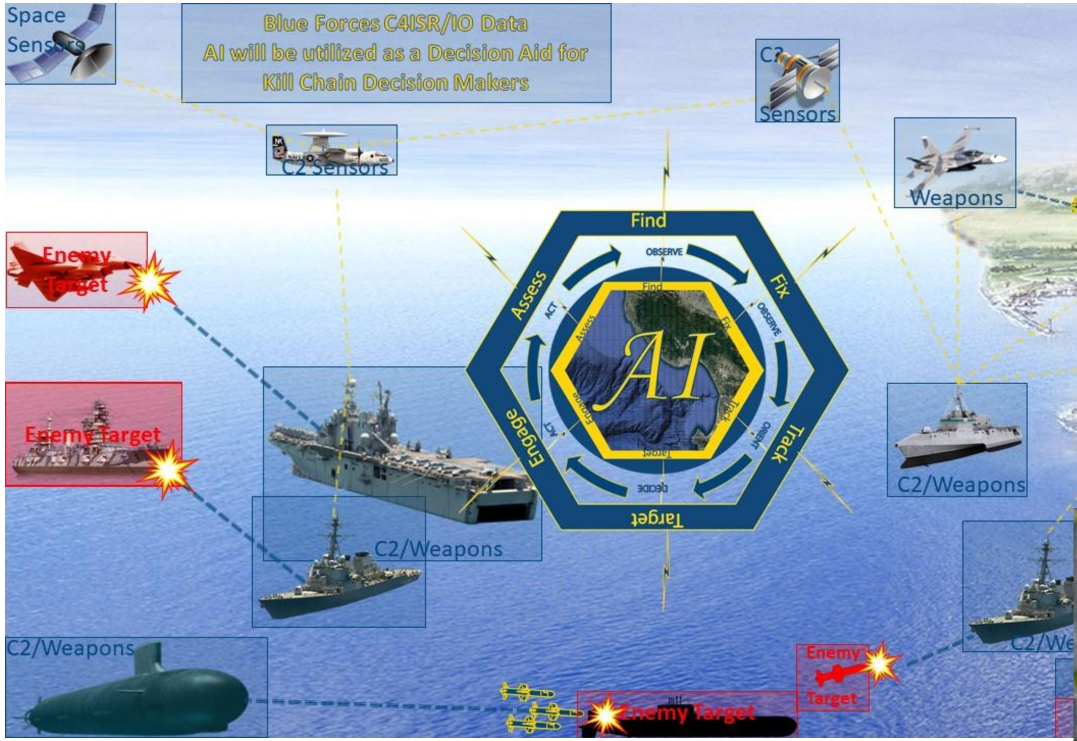
4. Reliable

DoD's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

5. Governable

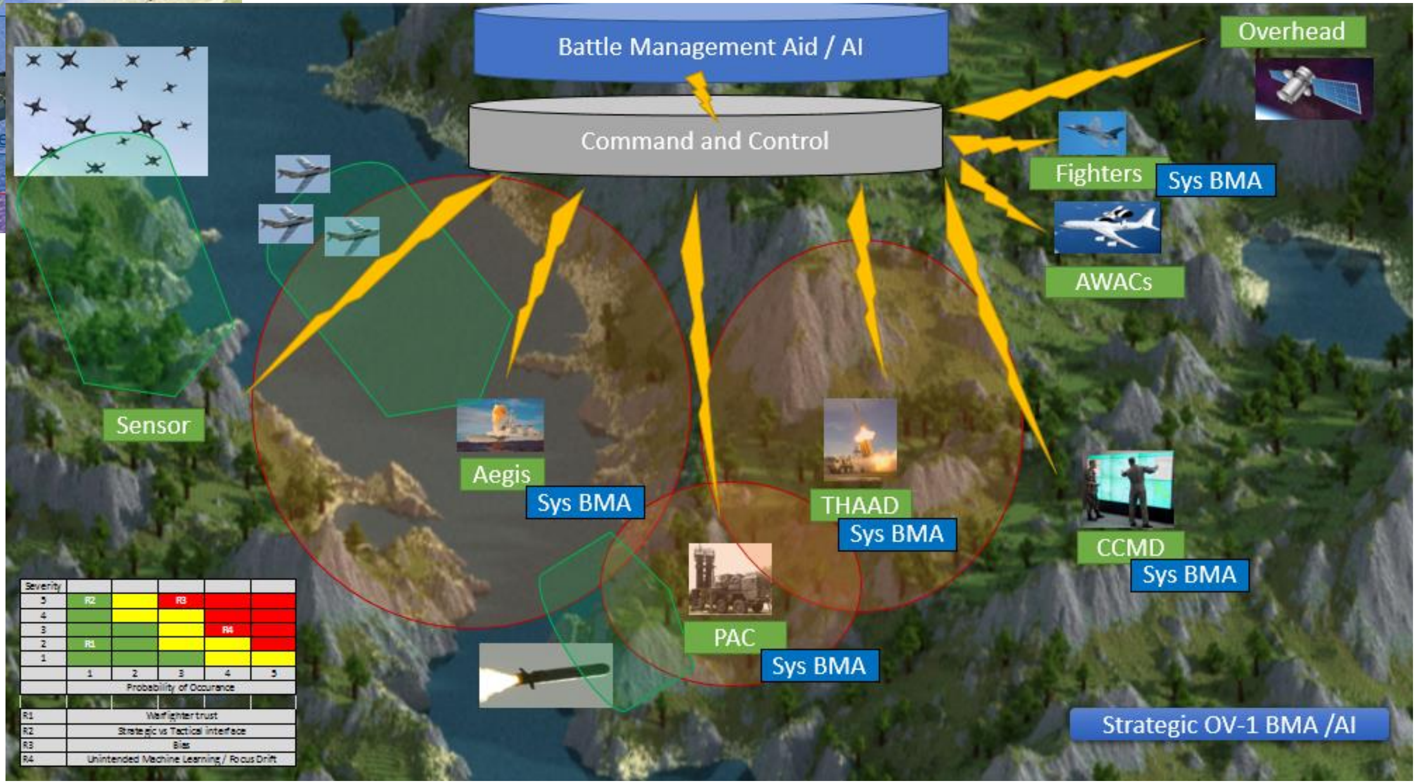
DoD will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Future AI-enabled warfare decision aids



Blue Forces C4ISR/IO Data
AI will be utilized as a Decision Aid for Kill Chain Decision Makers

Type of AI system application: **decision sciences system**

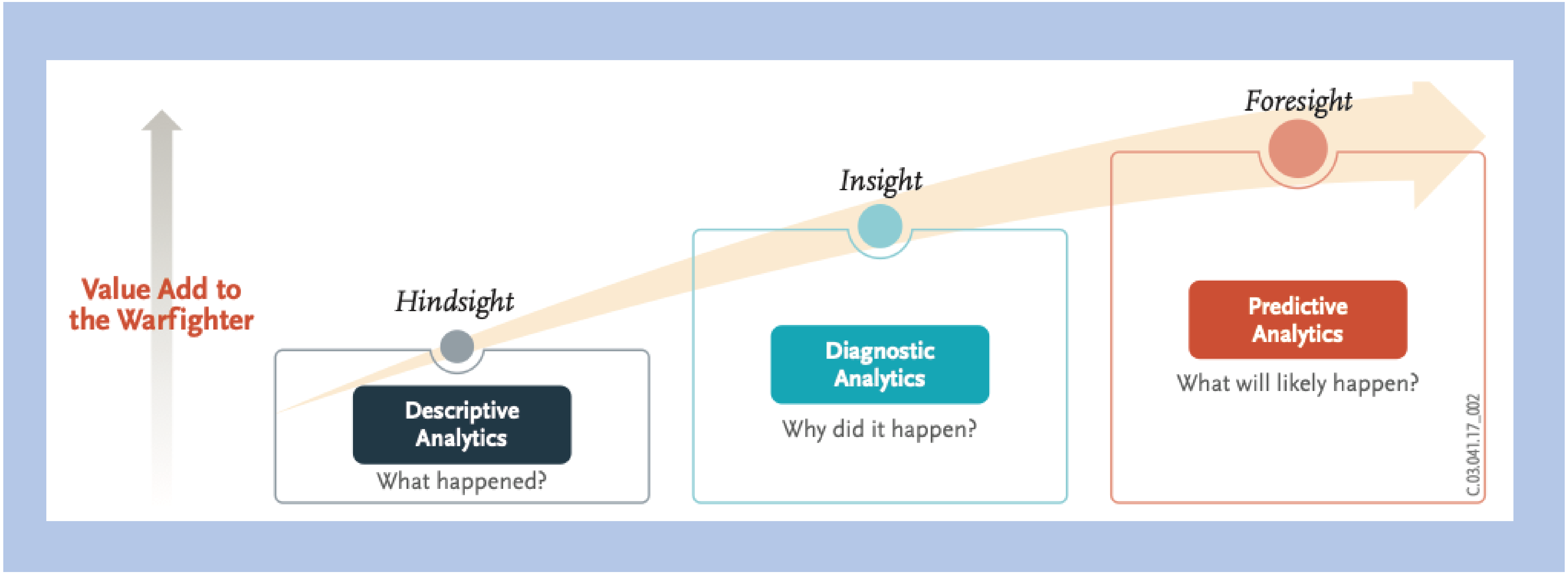


Severity					
5	🟩 R2	🟨	🟥 R3	🟥	🟥
4	🟩	🟨	🟨	🟥	🟥
3	🟩	🟩	🟨	🟥 R4	🟥
2	🟩 R1	🟩	🟨	🟨	🟥
1	🟩	🟩	🟩	🟨	🟨
	1	2	3	4	5
	Probability of Occurence				

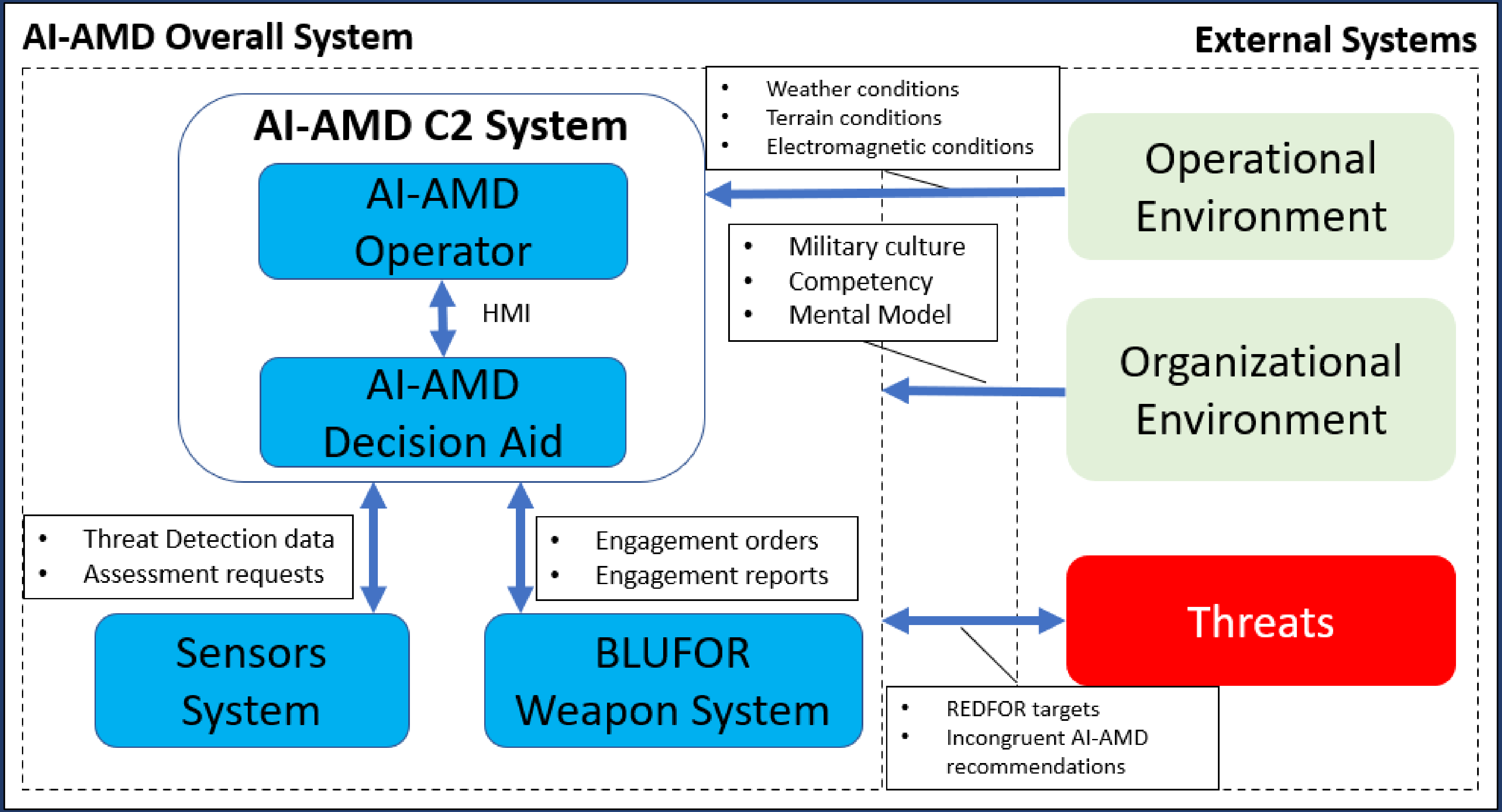
R1	Warfighter trust
R2	Strategic vs Tactical interface
R3	Bias
R4	Unintended Machine Learning / Focus Drift

Future concept:
AI-enabled warfare decision aid supports warfighter decision-making through enhanced battlespace knowledge, addressing uncertainty, recommending tactical courses of action, developing engagement strategies

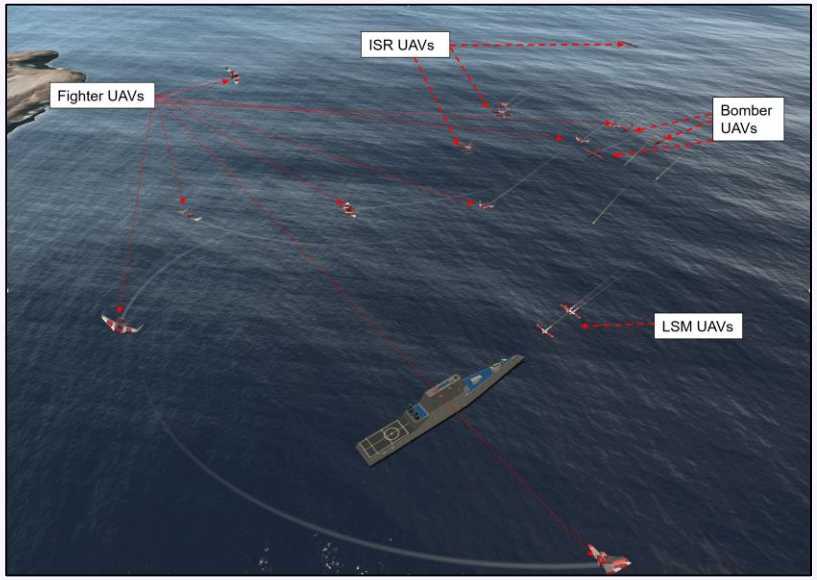
Analytical value to the warfighter: hindsight, insight, and foresight



AI-Enabled warfare decision aid:
for Air and Missile Defense (AMD)



Shipboard defense against drone swarms using laser weapon system: training a machine learning algorithm to select the **most effective engagement strategy** (shoot at the closest drone first or shoot at the weaponized drone).



Engagement Methodology	Number of Simulations	Blue Force Wins	Blue Force Losses	Win Percentage
Proximity	161	106	55	66%
Threat	112	100	12	89%

Tree Classification

Engagement Methodology	Number of Simulations	ML Correct Prediction	ML Incorrect Prediction	Percentage Correct	False Positive (predicted loss and win occurred)	False Negative (predicted win and loss occurred)
Proximity	35	27	8	77%	8	0
Threat	34	33	1	97%	0	1

Random Forest

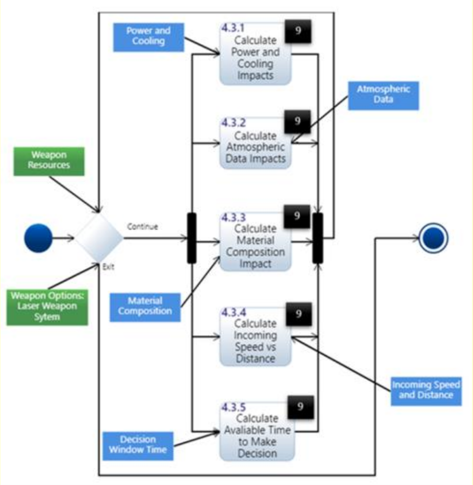
Engagement Methodology	Number of Simulations	ML Correct Prediction	ML Incorrect Prediction	Percentage Correct	False Positive (predicted loss and win occurred)	False Negative (predicted win and loss occurred)
Proximity	35	29	6	83%	5	1
Threat	34	32	2	94%	0	2

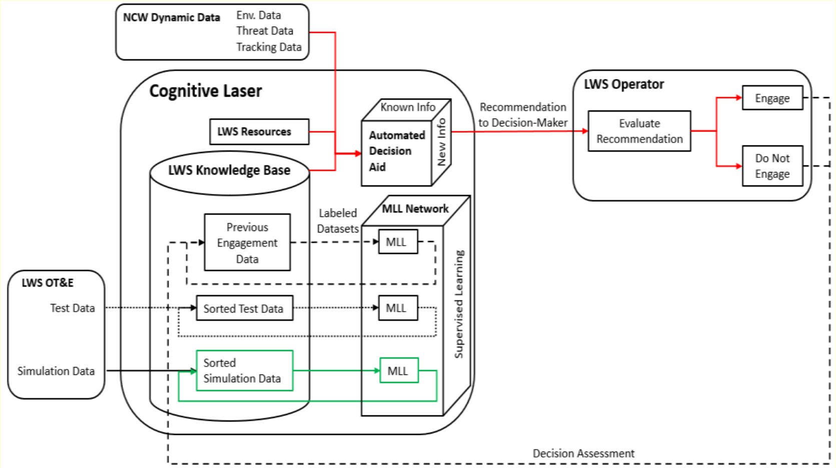
Logistic Regression

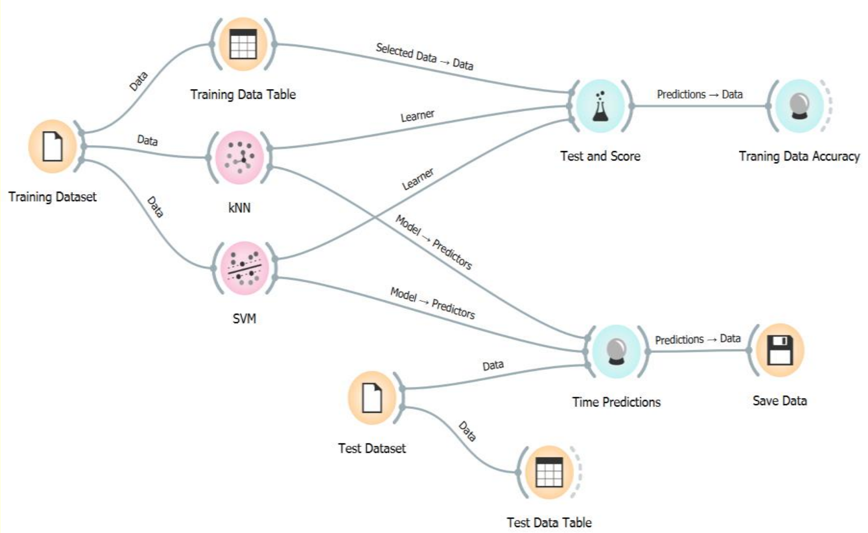
Engagement Methodology	Number of Simulations	ML Correct Prediction	ML Incorrect Prediction	Percentage Correct	False Positive (predicted loss and win occurred)	False Negative (predicted win and loss occurred)
Proximity	35	30	5	86%	3	2
Threat	34	27	7	79%	4	3

Edwards, Daniel. 2021. “Simulated Laser Weapon System Decision Support to Combat Drone Swarms with Machine Learning.” Naval Postgraduate School Thesis.

Shipboard defense against threats using a laser weapon: training a machine learning algorithm to calculate the **required laser “dwell time”** based on the threat’s material (type and thickness of material).

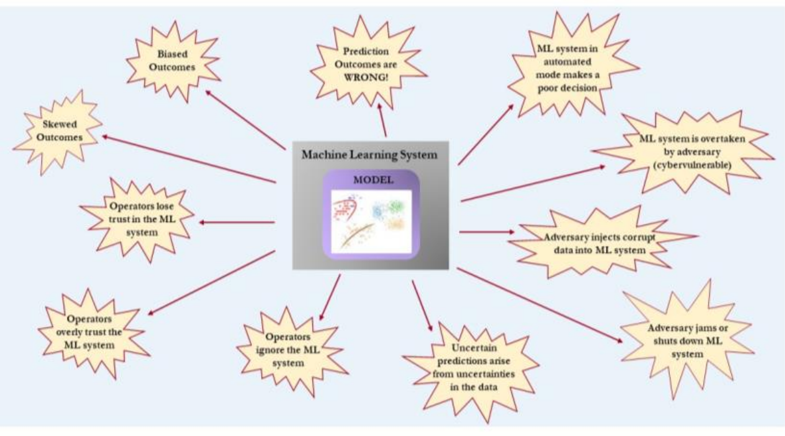

Power and Cooling
4.3.1 Calculate Power and Cooling Impacts
9
Atmospheric Data
4.3.2 Calculate Atmospheric Data Impacts
9
Weapon Resources
Continue
Exit
4.3.3 Calculate Material Composition Impact
9
Weapon Options: Laser Weapon Sytem
Material Composition
4.3.4 Calculate Incoming Speed vs Distance
9
Incoming Speed and Distance
4.3.5 Calculate Avaliable Time to Make Decision
9
Decision Window Time



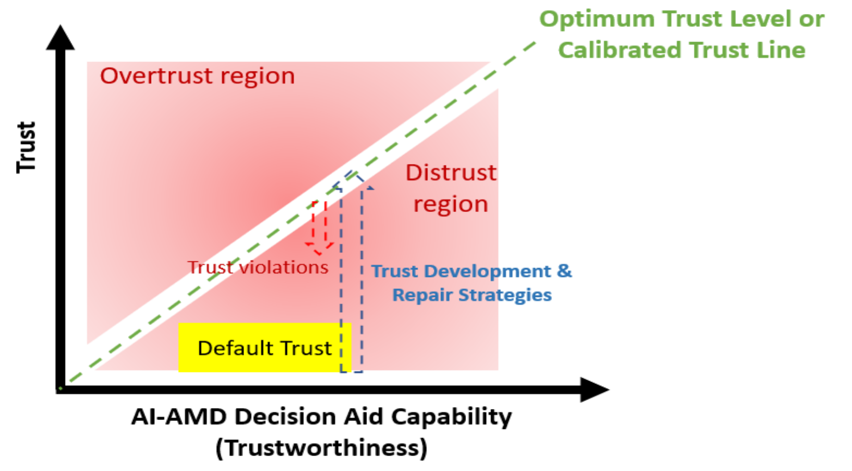


Blickley, W., Carlon, J., Magana, M., Pacheco, A., and Roscher, J. 2021. “Cognitive Laser Weapon System – Exploring Automation, Artificial Intelligence, and Human-Machine Teaming for Engagement.” Naval Postgraduate School Thesis.

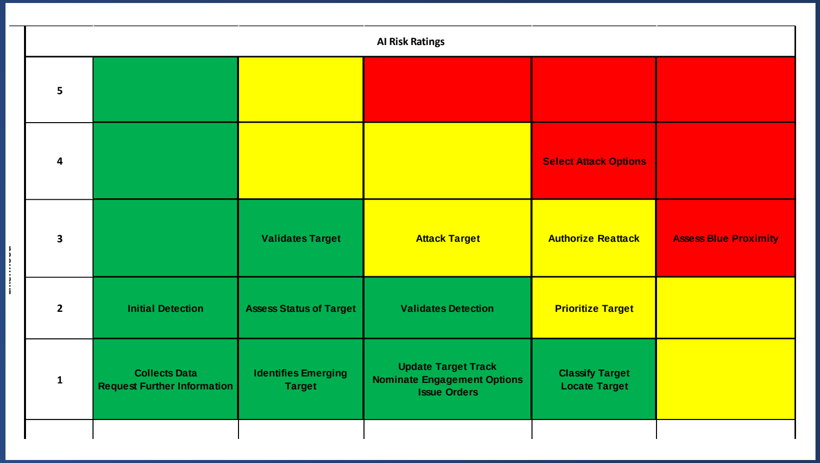
AI Safety


Biased Outcomes
Prediction Outcomes are WRONG!
ML system in automated mode makes a poor decision
Skewed Outcomes
ML system is overtaken by adversary (cybervulnerable)
Machine Learning System
MODEL
Operators lose trust in the ML system
Adversary injects corrupt data into ML system
Operators overly trust the ML system
Operators ignore the ML system
Uncertain predictions arise from uncertainties in the data
Adversary jams or shuts down ML system

AI Trust

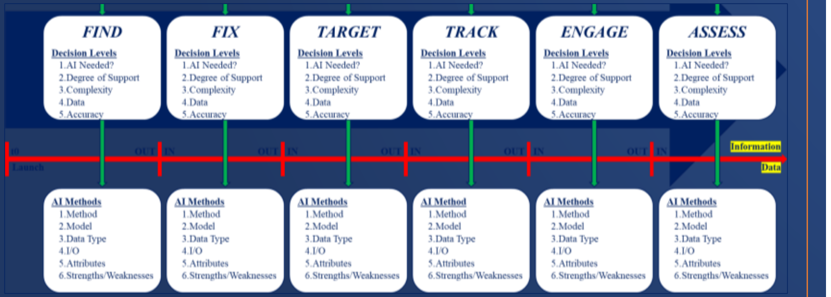


Decision Risk

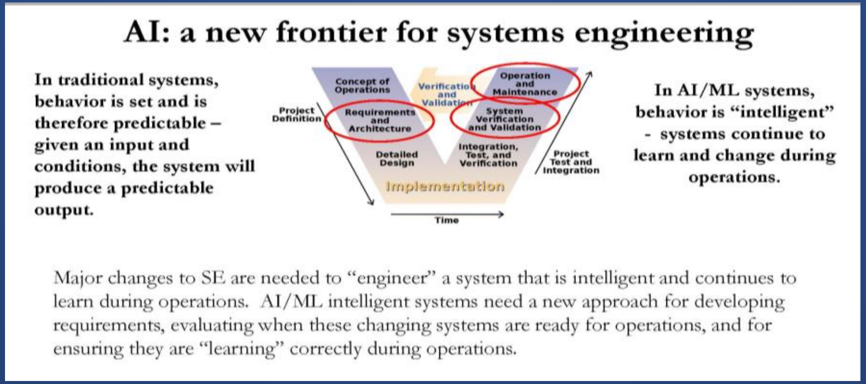


	AI Risk Ratings				
5					
4				Select Attack Options	
3		Validates Target	Attack Target	Authorize Reattack	Assess Blue Proximity
2	Initial Detection	Assess Status of Target	Validates Detection	Prioritize Target	
1	Collects Data Request Further Information	Identifies Emerging Target	Update Target Track Nominate Engagement Options Issue Orders	Classify Target Locate Target	

Mapping AI to the Kill Chain



SE for AI



AI: a new frontier for systems engineering

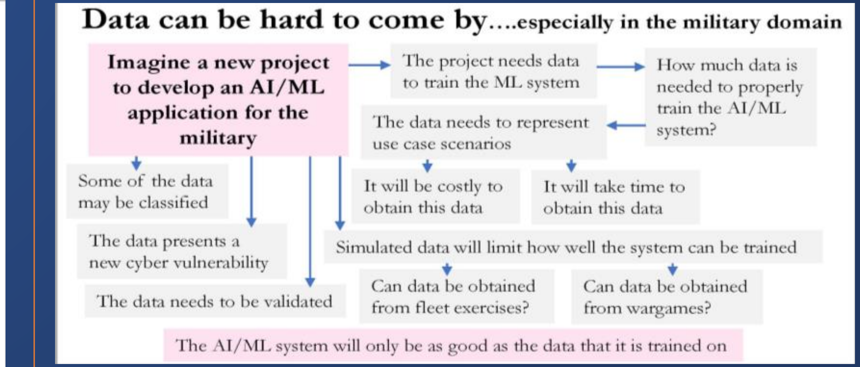
In traditional systems, behavior is set and is therefore predictable – given an input and conditions, the system will produce a predictable output.

Concept of Operations
Verification and Validation
Operation and Maintenance
Project Definition
Requirements and Architecture
System Verification and Validation
Detailed Design
Integration, Test, and Verification
Project Test and Integration
Implementation
Time

In AI/ML systems, behavior is "intelligent" - systems continue to learn and change during operations.

Major changes to SE are needed to "engineer" a system that is intelligent and continues to learn during operations. AI/ML intelligent systems need a new approach for developing requirements, evaluating when these changing systems are ready for operations, and for ensuring they are "learning" correctly during operations.

Data for AI



Data can be hard to come by….especially in the military domain

Imagine a new project to develop an AI/ML application for the military

The project needs data to train the ML system

How much data is needed to properly train the AI/ML system?

The data needs to represent use case scenarios

Some of the data may be classified

It will be costly to obtain this data

It will take time to obtain this data

The data presents a new cyber vulnerability

Simulated data will limit how well the system can be trained

The data needs to be validated

Can data be obtained from fleet exercises?

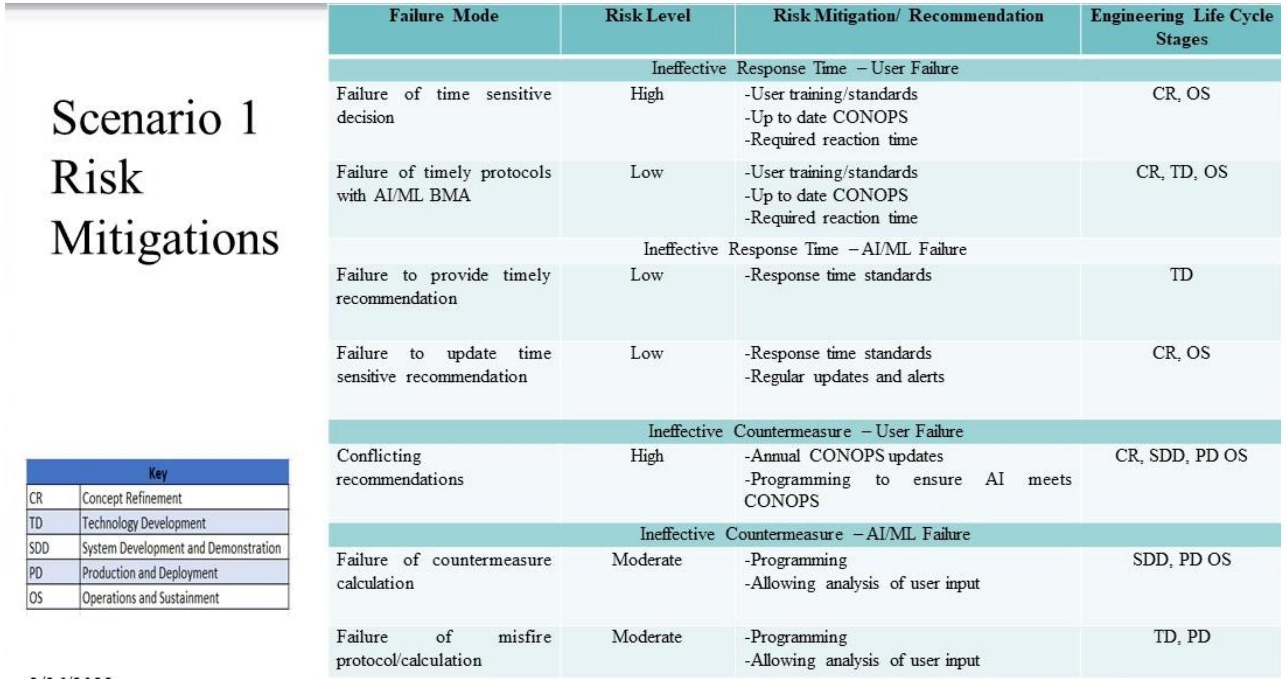
Can data be obtained from wargames?

The AI/ML system will only be as good as the data that it is trained on

Evaluation of safety risks in AI-enabled decision aids for air and missile defense

Hoopes, A., Cruz, L., Wuornos, S., Shilt, S., and Pappa, R. 2021. “Evaluation of the Safety Risks in Developing and Implementing Automated Battle Managements Aids for Air and Missile Defense.” Naval Postgraduate School Capstone Report.

Examples of Failure Mode Analysis



Scenario 1 Risk Mitigations

Failure Mode	Risk Level	Risk Mitigation/ Recommendation	Engineering Life Cycle Stages
Ineffective Response Time – User Failure			
Failure of time sensitive decision	High	-User training/standards -Up to date CONOPS -Required reaction time	CR, OS
Failure of timely protocols with AI/ML BMA	Low	-User training/standards -Up to date CONOPS -Required reaction time	CR, TD, OS
Ineffective Response Time – AI/ML Failure			
Failure to provide timely recommendation	Low	-Response time standards	TD
Failure to update time sensitive recommendation	Low	-Response time standards -Regular updates and alerts	CR, OS
Ineffective Countermeasure – User Failure			
Conflicting recommendations	High	-Annual CONOPS updates -Programming to ensure AI meets CONOPS	CR, SDD, PD OS
Ineffective Countermeasure – AI/ML Failure			
Failure of countermeasure calculation	Moderate	-Programming -Allowing analysis of user input	SDD, PD OS
Failure of misfire protocol/calculation	Moderate	-Programming -Allowing analysis of user input	TD, PD

Key	
CR	Concept Refinement
TD	Technology Development
SDD	System Development and Demonstration
PD	Production and Deployment
OS	Operations and Sustainment

Common Failure Modes

Failure Type	Definition	Examples
Operational	Failure of system operation or system to system operation	Internal sensor function failure, launcher malfunction
AI/ML Programming	Incorrect/unintended error in AI/ML programming	Identify hostile threat as non-hostile, unable to process multiple threats
Human-Machine Interaction (HMI)	Errors with user interaction with the system(s) (AI interaction focused)	Interface issues, interpretation error, lack of trust in AI/ML
Adversarial Attack	Direct attack or manipulation by adversary	C2 network hacking, insider threat, enemy causes ML recognition mistake

Examples of Risk Mitigation Strategies

Systems Engineering Phases:	Needs Analysis and Concept Refinement Phase	Design & Development Phase	Test & Evaluation Phase	Operations and Support Phase
AI Safety Risk Mitigation Strategies:	• Development of AI safety requirements • Careful consideration of safety in operational scenarios • Safety risk analysis • SE plans for safety for system data, system design, T&E, and operations	• Physical access control • Encryption, firewall, access control to designs • Design evaluation • Secure access to data • Careful data assessment and validation	• Multiple evaluation sites and organizations • Independent evaluation • Secure access control to system, data, and test sites • Audits • Network and firewall protections	• Alternate sites • Uninterrupted Power Supply • Firewall, network protection • Audits • Backup • Training • Software/system updates

Dr. Johnson's Team:
Research Questions

How can we engineer AI?
What AI methods are appropriate?
How can we ensure AI is safe?
How can humans best interact with AI?
Will warfighters trust AI?
What new systems engineering methods are needed to develop AI?
How will AI benefit the warfighter?
How will AI benefit the Navy?
What will future AI-enabled military concepts of operation look like?

Wrap Up

- AI has huge potential for many diverse applications (data products, cyber-physical, decision sciences)
- AI systems present new types of safety risks: failure modes, consequences, root causes
- AI safety must be implemented throughout the systems engineering lifecycle
- Metacognition is an AI system safety strategy that must be engineered into systems and implemented during operations.
- Many exciting research opportunities!

I welcome collaboration!

Dr. Bonnie Johnson
Naval Postgraduate School
bwjohnson@nps.edu

References

Allen, G. 2020. Understanding AI technology. Joint Artificial Intelligence Center (JAIC) Report, US Department of Defense.

Blickley, W., Carlon, J., Magana, M., Pacheco, A., and Roscher, J. 2021. "Cognitive Laser Weapon System – Exploring Automation, Artificial Intelligence, and Human-Machine Teaming for Engagement." Naval Postgraduate School Thesis.

Crowder, J., Friess, S. 2011. Metacognition and metamemory concepts for AI systems. In: Athens: the Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1-6.

Crowder, J., Friess, S. 2012. Extended metacognition for Artificially Intelligent Systems (AIS): artificial locus of control and conitive economy. In: Athens: the Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1-6.

Crowder, J., Carbone, J. 2014. Eliminating cognitive ambiguity with knowledge relativity threads. In: Athens: The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1-7.

Edwards, Daniel. 2021. "Simulated Laser Weapon System Decision Support to Combat Drone Swarms with Machine Learning." Naval Postgraduate School Thesis

Faria, J. 2017. Non-determinism and failure modes in machine learning. In: 2017 IEEE 28th International Symposium on Software Reliability Engineering Workshops, 310 – 316.

Faria, J. 2018. Machine learning safety: an overview. In: Proceedings of the 26th Safety-Critical Systems Symposium. York, UK, 4-18.

Gunning, D., Aha, D. 2019. DARPA's Explainable Artificial Intelligence Program. AI Magazine 40(2), 44-58.

Hoopes, A., Cruz, L., Wuornos, S., Shilt, S., and Pappa, R. 2021. "Evaluation of the Safety Risks in Developing and Implementing Automated Battle Managements Aids for Air and Missile Defense." Naval Postgraduate School Capstone Report.

Johnson, Bonnie. "Metacognition for artificial intelligence system safety – an approach to safe and desired behavior." Submitted to the *Journal of Safety Science*. 2021.

Johnson, M., Bradshaw, J., Feltovich, P. 2018. Tomorrow's human-machine design tools: from levels of automation to interdependencies. Journal of Cognitive Engineering and Decision-Making, 12(1), 77-82.

Michael, C., Acklin, D., Scheuerman, J. 2020. On interactive machine learning and the potential of cognitive feedback. In: arXiv:2003.10365v1.

Mitchell, M. 2019. Artificial Intelligence – A Guide for Thinking Humans. Picador: New York.

Nushi, B., Kamar, E., Horvitz, E. 2018. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. In: Sixth Annual Conference on Human Computation and Crowdsourcing (HCOMP), 126-135.

Varshney, K. 2016. Engineering safety in machine learning. In: Information Theory and Applications Workshop (ITA), La Jolla, CA, USA, 1-5. doi: 10.1109/ITA.2016.7888195.

Varshney, K., Alemazdeh, H. 2017. On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data 5(3), 246-255.

Welling, M. 2015. Are ML and statistics complimentary? In: IMS-ISBA Meeting on Data Science in the Next 50 Years, December.