2021

Human-Machine Weapons Engagement Decisions: Systems Safety in Complex Decision Environments

Johnson, Bonnie W.; Miller, Scot A.; Green, John M.; Kendall, Walter A.; Godin, Arkady A.

Monterey, California: Naval Postgraduate School

Human-Machine Weapons Engagement Decisions: Systems Safety in Complex Decision Environments

Naval Postgraduate School

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.

Research Questions:

- What are the safety concerns involved in implementing AI methods to support human-machine weapons engagement decisions?
- What concepts, requirements, and methods can ensure that future AI-enabled tactical decision aids are safely deployed?

Failure Modes

System Produces Faulty/Poor Decision Recommendation

- Skewed outcomes/ predictions
- Uncertain outcomes/ predictions
- Biased outcomes/ predictions

System Under Attack (Cyber attack)

- System and its outcomes are corrupted by adversary
- System is overtaken by adversary/adversary is controlling system
- Adversary jams or shuts down system
- Adversary gains access to system; decision information/knowledge is compromised

Human Machine Operation Issues

- Operators have lack of trust in the system
- Operators ignore the system
- Operators are overly trusting (over reliant) in the system
- Operators introduce errors into the system
- Operators misunderstand the system recommendations/ predictions
- System identifies operator misuse
- System identifies operator-induced error
- System identifies operator inattentiveness or fatigue

Artificial Intelligence System

MODEL

Root Causes

Pre-Deployment: Design, Development, Testing

- Incompleteness---data sets don't represent all scenarios
- Rare examples – data sets don't include unusual scenarios
- Bias in the training data sets
- Corruption in the training data sets
- Mis-labeled or mis-associated data
- Poor data collection methods
- Poor validation methods (is there criteria for deciding how much training data is good enough?)
- Underfitting in the model – model is not capable of attaining sufficiently low error on the training data
- Cost function algorithm errors – 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

- Inaccuracy in the algorithm model (prediction error)
- Operational complexity that overwhelms the AI
- Uncertainty/error in operational datasets
- Corruption in operational datasets
- Lack of explainability
- Trust issues
- Operator-induced error
- 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
- Adversarial attacks – hacking, deception, inserting false data, controlling automated systems

AI System Safety: Four Types of Solution Strategies

Pre-Deployment: Design, Development, Testing → **Post-Deployment: Operations & Sustainment**

1 Inherently Safe Design

Focus: ensuring robustness against uncertainty in the training data sets
Methods: Interpretability – ensuring designers understand the complex ML systems produced from the data training process
Causality – reducing uncertainty by eliminating non-causal variables from the model

2 Safety Reserves

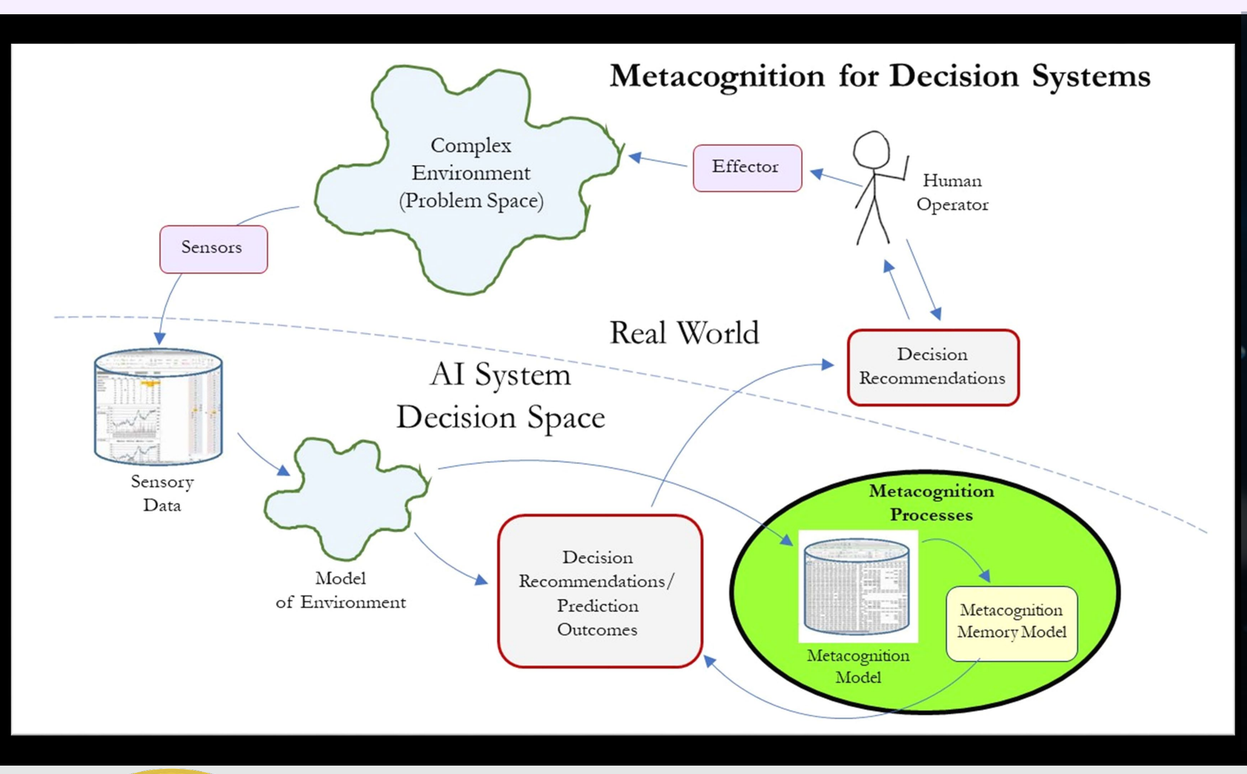
Focus: achieving safety through additive reserves, safety factors, safety margins – through training data set validation
Methods: Validating training data sets – eliminating uncertainty; ensuring data is accurate, representative, sufficient, bias-free
Increasing/improving model training process – ensuring 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
Methods: Human operation intervention (system overrides and manual operation)
Metacognition – AI system self-diagnosis to recognize and address failure modes
Explainability/Understandability/Trust-worthy

4 Procedural Safeguards

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

Research P.I.s Dr. Bonnie Johnson, NPS Systems Engineering and Mr. Scot Miller, NPS, Information Sciences
Research Team: Arkady Godin, Tony Kendall, John M. Green
Topic Sponsor: NAWCWD China Lake – Mr. Bruce Nagy
NPS Students: Peh Ming Hui, Luis Cruz, Ryane Pappa, Savanna Shilt, Angela Hoopes, Samuel Wuornos, J. Isaac Jones, Russell Kress, Adam Rahman, William Newmeyer, Wallace Fukumae, Kheng Hun, Robert French, Obed Matuga, and Caitlyn O'Shaugnessy

NRP Project ID: NPS-21-N317-A