NAVAL
POSTGRADUATE
SCHOOL

Artificial Intelligence Systems: Unique Challenges for Defense Applications

2021 Acquisition Research Symposium
Pre-symposium Webinar: Developing Artificial Intelligence in Defense Programs

3 March 2021

Dr. Bonnie Johnson
NPS Systems Engineering
bwjohnson@nps.edu

AI Systems: Unique Challenges for Defense Applications

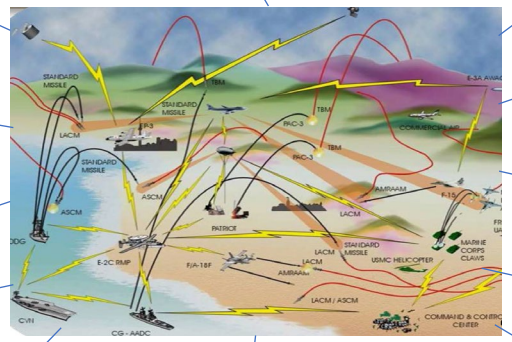
Tactical Decisions are Complex!

Multi-mission operations

Cyber-Vulnerabilities

Command Authority

Collaboration between platforms



Time-critical operations

Interoperability challenges

Information Overload

Real-time Decisions

Unknown-unknowns

Incomplete Situational Awareness

Human Trust in AI

Life or Death Consequences

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

The data presents a new cyber vulnerability

The data needs to be validated

It will be costly to obtain this data

It will take time to obtain this data

Simulated data will limit how well the system can be trained

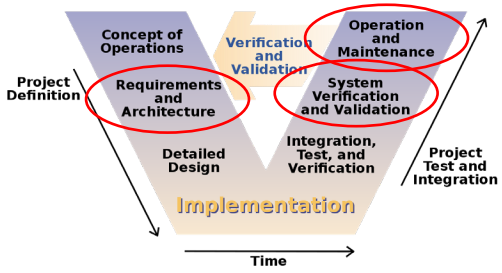
Can data be obtained from fleet exercises?

Can data be obtained from wargames?

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

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.



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.

Adversaries

1
The Race is On!

Adversary advancements in AI—are we keeping up or falling behind?

Will AI be the new standard for future military dominance?

Can our AI/ML systems support our military decision superiority?

2
Cyber Attacks

As we rely more and more on AI/ML systems, are we creating more cyber vulnerabilities?

Through growth in automation, are we making it easier for adversaries to take control of our systems or "poison" our systems with bad data?

3
Threats Keep Changing

Can our AI/ML systems keep up with the always-changing adversarial threat space?

Technology is rapidly evolving. The geo-political landscape continues to change. Can AI/ML systems evolve fast enough & in a safe and trustable way to meet this pace?

AI Systems: Unique Challenges for Defense Applications

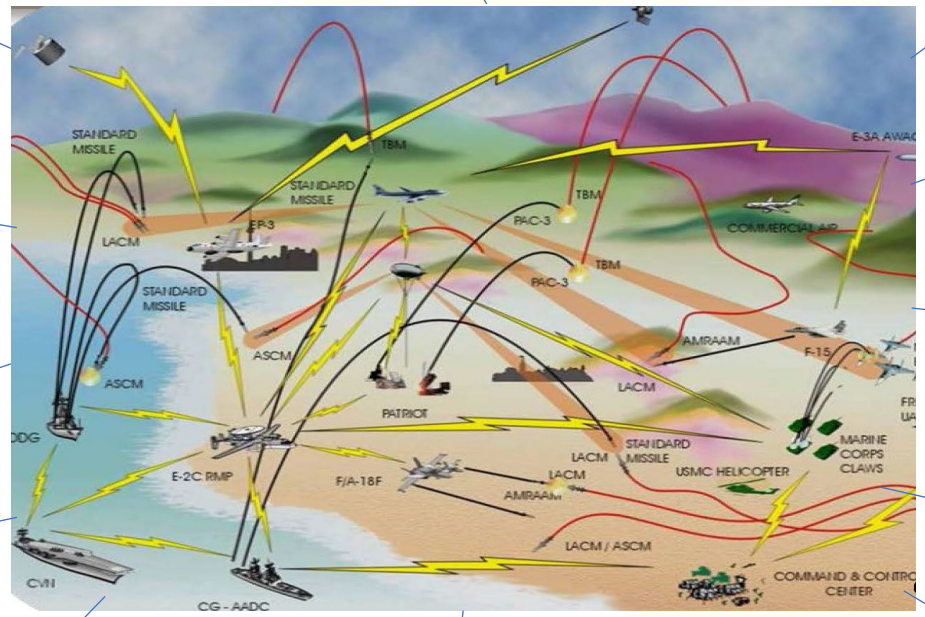
Tactical Decisions are Complex!

Multi-mission operations

Cyber-Vulnerabilities

Command Authority

Collaboration between platforms



Time-critical operations

Information Overload

Interoperability challenges

Real-time Decisions

Unknown-unknowns

Incomplete Situational Awareness

Human Trust in AI

Life or Death Consequences

AI Systems: Unique Challenges for Defense Applications

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?

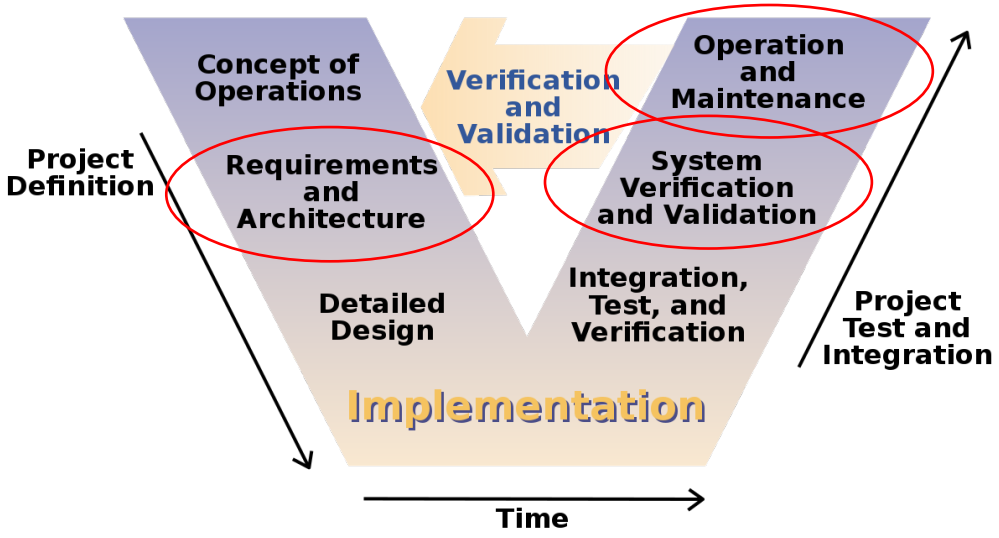
Can data be obtained from wargames?

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

AI Systems: Unique Challenges for Defense Applications

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.



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.

AI Systems: Unique Challenges for Defense Applications

Adversaries

1
The Race is On!

Adversary advancements in AI—are we keeping up or falling behind?

Will AI be the new standard for future military dominance?

Can our AI/ML systems support our military decision superiority?

2
Cyber Attacks

As we rely more and more on AI/ML systems, are we creating more cyber vulnerabilities?

Through growth in automation, are we making it easier for adversaries to take control of our systems or "poison" our systems with bad data?

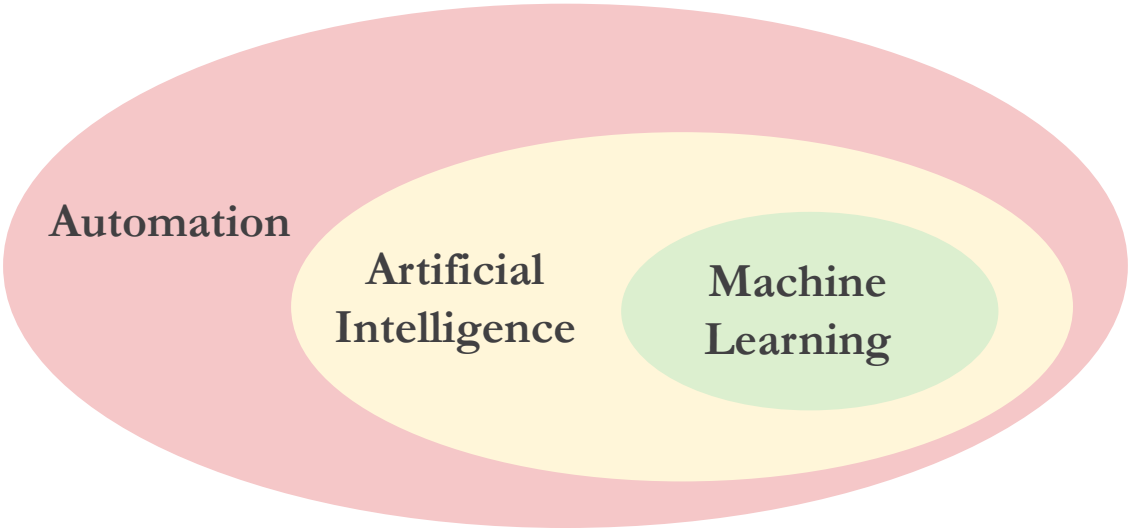
3
Threats Keep Changing

Can our AI/ML systems keep up with the always-changing adversarial threat space?

Technology is rapidly evolving. The geo-political landscape continues to change. Can AI/ML systems evolve fast enough & in a safe and trustable way to meet this pace?

What is AI?

Here's a good definition:
AI is the application of human (or biological) processes to problem solving using machines (usually, but not always digital computers)



Two Primary Types of AI

1. Explicitly Programmed

- Think "if-then," but can be more complex
- Uses normal programming languages
- Can involve complex manually designed coding schemes for data / knowledge

2. Learns from Data

- The system is provided a large amount of data (many labeled examples)
- The system learns patterns by trial-and-error until it can predict the labeled examples
- Then, the "trained" system can be used (for prediction) given new data

Adapted from Barclay Brown, Raytheon

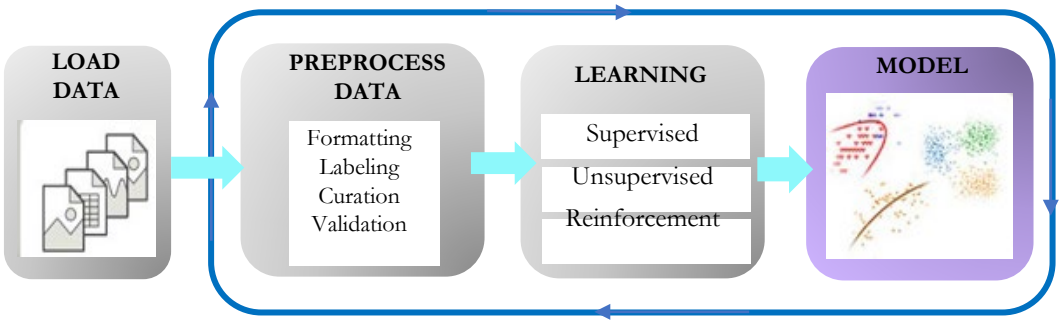
A new type of system – a new set of challenges

Systems Engineering & Acquisition

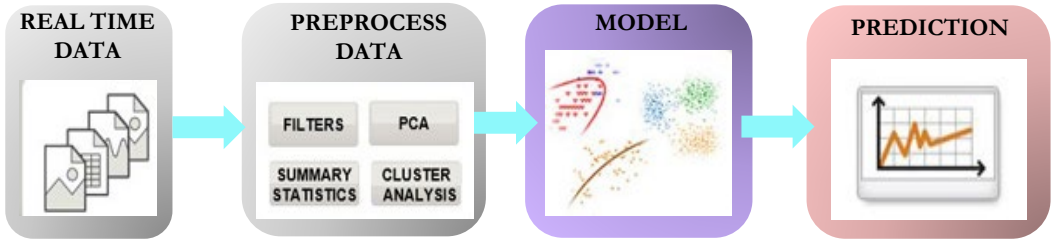
Pre-Deployment: Design, Development, Testing

Post-Deployment: Operations & Sustainment

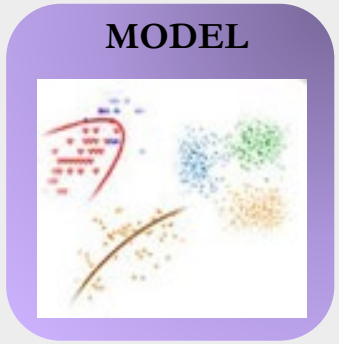
Learn/Train: **iterate until the best model is developed**



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



The machine learning "system" is the trained model



A new type of system – a new set of challenges

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.

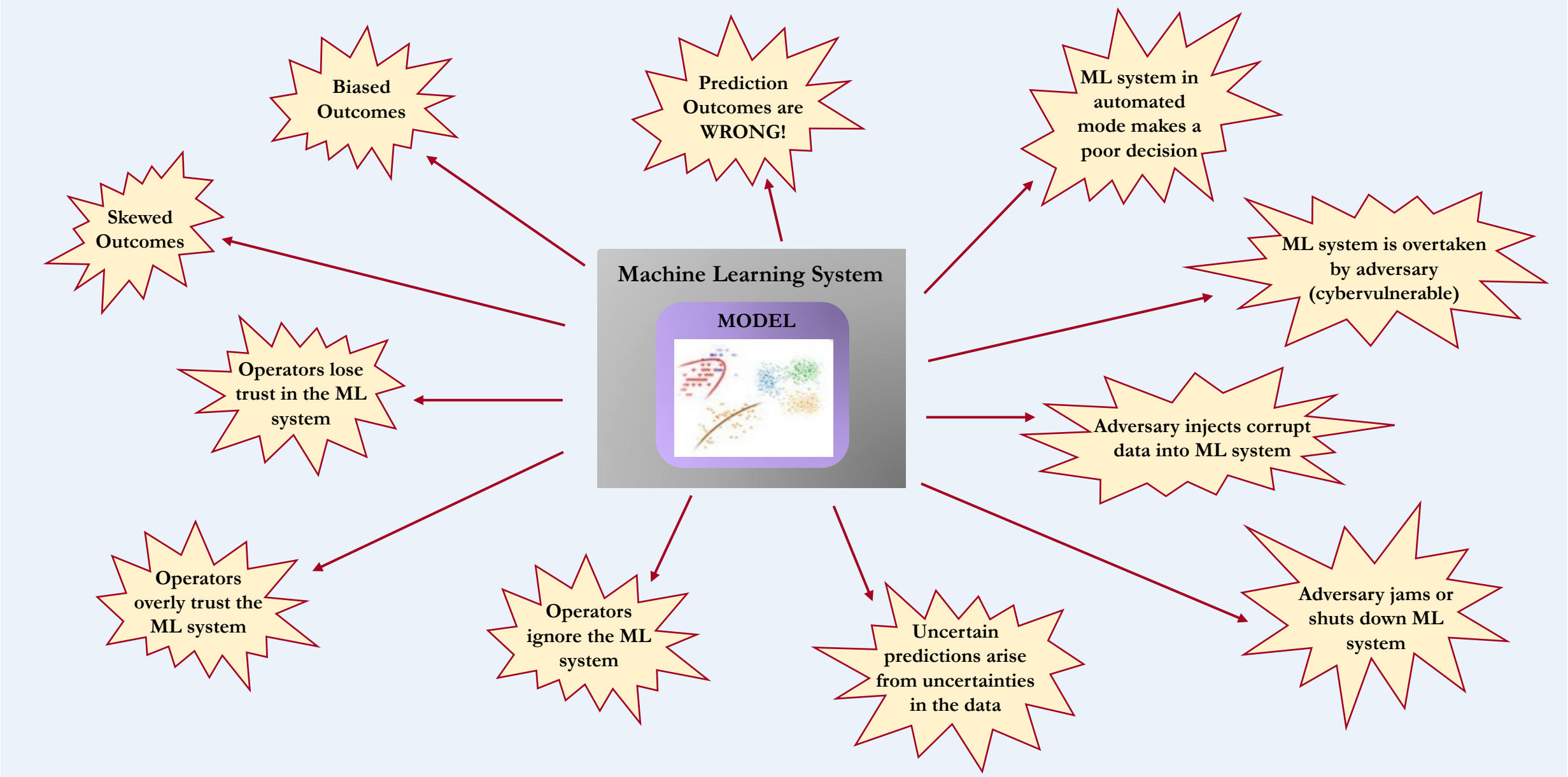
Complex

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 of AI/ML Systems



AI System Safety: Root Causes of Failure Modes

Systems Engineering & Acquisition

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

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?)

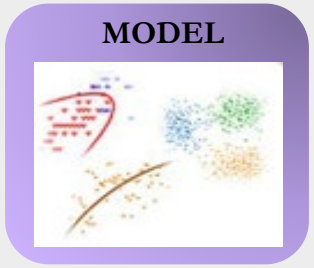
Poor data collection methods

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.)

Machine Learning System



Uncertainty/error in operational datasets

Corruption in operational datasets

Inaccuracy in the ML algorithm model (prediction error)

Operational complexity that overwhelms the ML 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: Solution Strategies

Step One: Determine whether the ML system application is Type A or Type B

Type A

Safety is Paramount

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

Examples:
- Medical diagnosis
- Loan approval
- Prison sentencing

Defense Application Examples:
- Time-critical tactical applications (combat identification, weapon engagement decisions)
- Mission planning applications (strike planning, aviation planning, UAS operations)

Type B

Safety is Less Important

Applications in which ML system model predictions are used in setting of low consequence and large scale

Examples:
- Services that decide which news story to show up on top
- Services that decide which advertisements to show

Defense Application Examples:
- Planning operations with ample time (some logistics operations)

AI System Safety: Four Types of Solution Strategies

Systems Engineering & Acquisition

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 ML systems that are produced from the data training process
- Causality – reducing uncertainty by eliminating non-causal variables from the model

AI System Safety: Four Types of Solution Strategies

Systems Engineering & Acquisition

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 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

AI System Safety: Four Types of Solution Strategies

Systems Engineering & Acquisition

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 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 ML systems should allow for adequate human-machine interaction to allow for system overrides and manual operation
- Metacognition – the ML 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

AI System Safety: Four Types of Solution Strategies

Systems Engineering & Acquisition

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 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 ML systems should allow for adequate human-machine interaction to allow for system overrides and manual operation
- Metacognition – the ML 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

Mapping AI to the Kill Chain

- SE Capstone project – graduating Dec 2021
- NRP 2021 – project with OPNAV N2/N6 Sponsor
- SE Capstone team (graduated Sept 2020) did preliminary study/

	Action	AI Method/Auto	Description
Find (Observe)	Collect Data	Data Management	Preprocessing and storing data
	Accept Initial Detection	Data Fusion / Fuzzy Reasoning	Fuse vague data to detect an anomaly
	Identify Emerging Threat	Case-based Reasoning	Retrieve similar cases
Fix (Observe)	Request Further Information	Event Procedure	Auto executes when triggered (emerged target)
	Classify Target	Decision Theory / Evidential Reasoning	Decide on target from data mining knowledge base
	Locate Target	Spatial Reasoning	Monitors the target in space and time
	Validate Detection	Predictive Analytics	Predicts trajectory of threat
Track (Orient)	Request Updated Target Track	Target Coordinate Mensuration (TCM) Validation	Provide precision coordinates meeting requirements of AD system
	Validate Target	Data Fusion/Forward Chaining	Combine location data with AD capabilities data
	Assess Blue Proximity	Utility Theory / Predictive Analytics / Forward Chaining	Assesses utility (capability) and readiness
Target (Decide)	Nominate Engagement Options	Decision Theory	Assesses both probability and utility of threat knowledge
	Prioritize Targets	Decision Theory	Assesses both probability and utility of COAs
	Select Attack Option(s)	Event Procedure / Template Filling	Auto executes when triggered and auto populate fields
Engage (Act)	Issue Orders	Event Procedure	Auto executes when triggered (attack order)
	Send Fire Command	Predictive Analytics / Spatial Reasoning	Monitors and projects threat and AD asset
Assess (Act)	Assess Target Status	Event Procedure	Auto executes when triggered (failed engagement)
	Authorize Re-attack	Data Management	Preprocessing and storing data

Engineering Trust into AI Systems

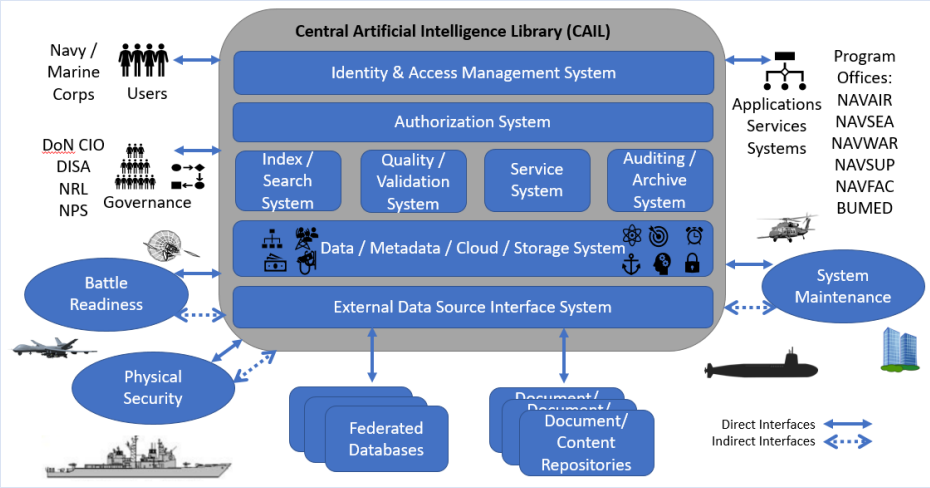
- SE Thesis project – graduating Dec 2021
- NRP 2021 – project with NAWC China Lake Sponsor

Study Approach:

1. Conduct a lit review of "trust" in AI systems
2. Study "trust" in an automated battle management aid for air and missile defense
   - Conceptualize a future AI-enabled BMA system for AMD
   - Study AMD kill chain and identify decision points involving HMI
2. Model the human-machine decision interactions for the AMD kill chain using BMA
   - Study the model using different threat scenario simulations with a variety of complexity
   - Identify "trust" issues/risks and their consequences
   - Characterize the components of trust in each decision point
3. Develop a strategy for engineering trust in AMD BMA systems based on the M&S analysis results

Data Management Strategy for the Navy

- SE Capstone project – graduating June 2021
- Presentation at NAML 2021



AI System Safety

- SE Capstone project – graduating Sept 2021
- NRP 2021 – project with NAWC China Lake Sponsor

Study Approach:

1. View an automated battle management aid for air and missile defense as a system
   - Characterize current BMAs and future BMAs
   - Conceptualize a future AI-enabled BMA system for AMD
   - Understand a future AI-enabled BMA system's SE lifecycle (highlight unique SE aspects of an AI-enabled system)
2. Perform a system safety analysis for the future AI-enabled BMA system
   - Problems occurring during operations
   - Problems creeping in during development
   - Data corruption (cyber attacks, bias, unintended poor data, incomplete data, etc.)
   - Human-machine safety risks (mis-trust, overreliance (overly trusted), dis-use, operator induced error, AI-explainability (or lack of understanding), AI complexity, etc.)
   - Cyberattacks
3. Characterize possible consequences of safety-related problems
4. Develop solutions, methods, and strategies for countering the safety issues
5. Compare and evaluate the solutions

Mapping AI to the Kill Chain

- SE Capstone project – graduating Dec 2021
- NRP 2021 – project with OPNAV N2/N6 Sponsor
- SE Capstone team (graduated Sept 2020) did preliminary study/

	Action	AI Method/Auto	Description
Find (Observe)	Collect Data	Data Management	Preprocessing and storing data
	Accept Initial Detection	Data Fusion / Fuzzy Reasoning	Fuse vague data to detect an anomaly
	Identify Emerging Threat	Case-based Reasoning	Retrieve similar cases
Fix (Observe)	Request Further Information	Event Procedure	Auto executes when triggered (emerged target)
	Classify Target	Decision Theory / Evidential Reasoning	Decide on target from data mining knowledge base
	Locate Target	Spatial Reasoning	Monitors the target in space and time
	Validate Detection	Predictive Analytics	Predicts trajectory of threat
Track (Orient)	Request Updated Target Track	Target Coordinate Mensuration (TCM) Validation	Provide precision coordinates meeting requirements of AD system
	Validate Target	Data Fusion/Forward Chaining	Combine location data with AD capabilities data
	Assess Blue Proximity	Utility Theory / Predictive Analytics / Forward Chaining	Assesses utility (capability) and readiness
Target (Decide)	Nominate Engagement Options	Decision Theory	Assesses both probability and utility of threat knowledge
	Prioritize Targets	Decision Theory	Assesses both probability and utility of COAs
	Select Attack Option(s)	Event Procedure / Template Filling	Auto executes when triggered and auto populate fields
Engage (Act)	Issue Orders	Event Procedure	Auto executes when triggered (attack order)
	Send Fire Command	Predictive Analytics / Spatial Reasoning	Monitors and projects threat and AD asset
Assess (Act)	Assess Target Status	Event Procedure	Auto executes when triggered (failed engagement)
	Authorize Re-attack	Data Management	Preprocessing and storing data

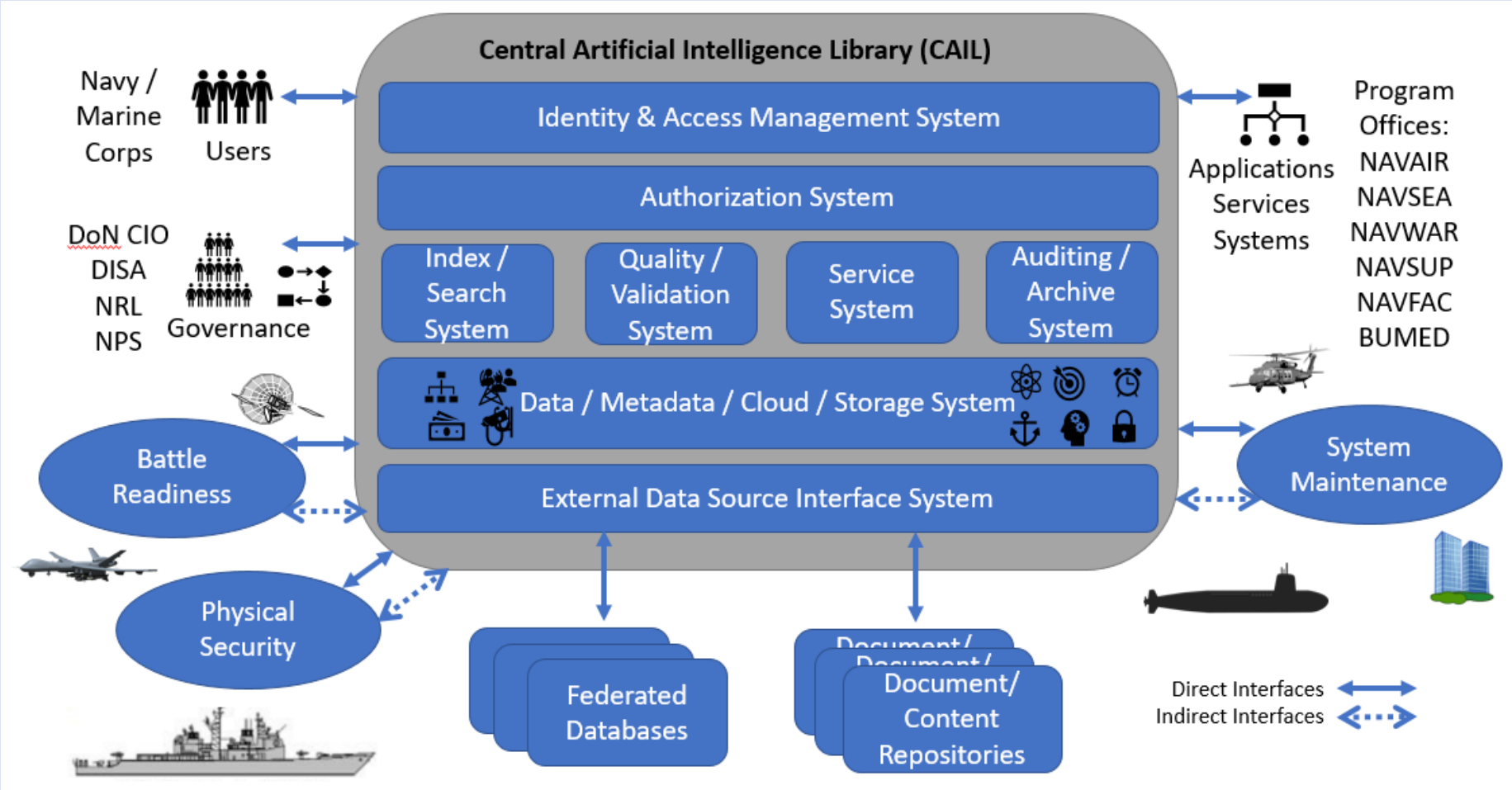
- SE Thesis project – graduating Dec 2021
- NRP 2021 – project with NAWC China Lake Sponsor

Study Approach:

1. Conduct a lit review of "trust" in AI systems
2. Study "trust" in an automated battle management aid for air and missile defense
   - Conceptualize a future AI-enabled BMA system for AMD
   - Study AMD kill chain and identify decision points involving HMI
2. Model the human-machine decision interactions for the AMD kill chain using BMA
   - Study the model using different threat scenario simulations with a variety of complexity
   - Identify "trust" issues/risks and their consequences
   - Characterize the components of trust in each decision point
3. Develop a strategy for engineering trust in AMD BMA systems based on the M&S analysis results

Data Management Strategy for the Navy

- SE Capstone project – graduating June 2021
- Presentation at NAML 2021



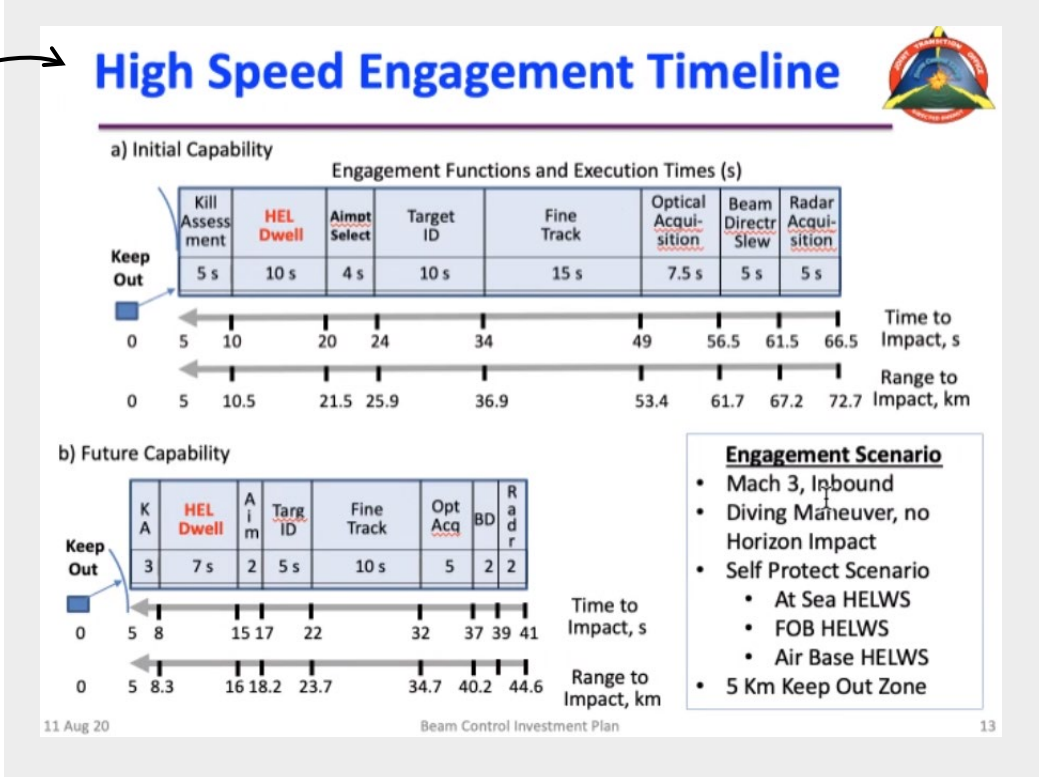
- SE Capstone project – graduating Sept 2021
- NRP 2021 – project with NAWC China Lake Sponsor

Study Approach:

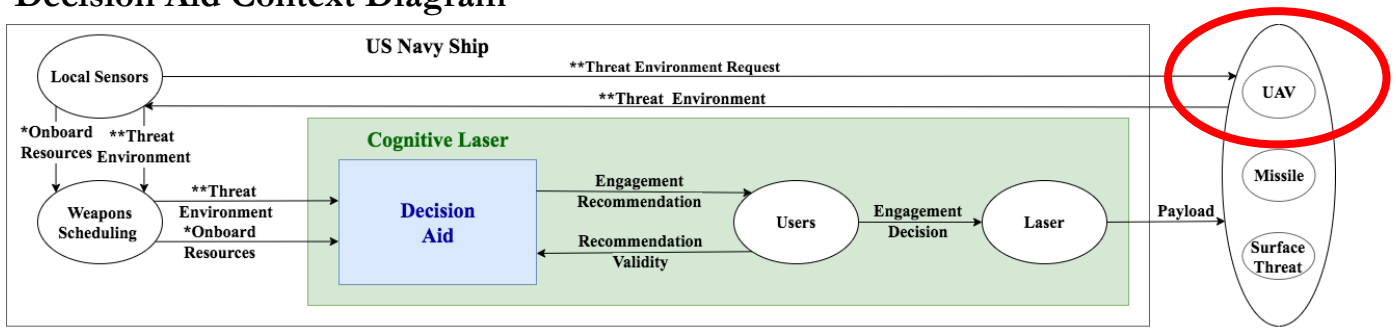
1. View an automated battle management aid for air and missile defense as a system
   - Characterize current BMAs and future BMAs
   - Conceptualize a future AI-enabled BMA system for AMD
   - Understand a future AI-enabled BMA system's SE lifecycle (highlight unique SE aspects of an AI-enabled system)
2. Perform a system safety analysis for the future AI-enabled BMA system
   - Problems occurring during operations
   - Problems creeping in during development
   - Data corruption (cyber attacks, bias, unintended poor data, incomplete data, etc.)
   - Human-machine safety risks (mis-trust, overreliance (overly trusted), dis-use, operator induced error, AI-explainability (or lack of understanding), AI complexity, etc.)
   - Cyberattacks
3. Characterize possible consequences of safety-related problems
4. Develop solutions, methods, and strategies for countering the safety issues
5. Compare and evaluate the solutions

Cognitive Laser

Primary Objective:
Reduce the Laser Engagement Timeline



Decision Aid Context Diagram



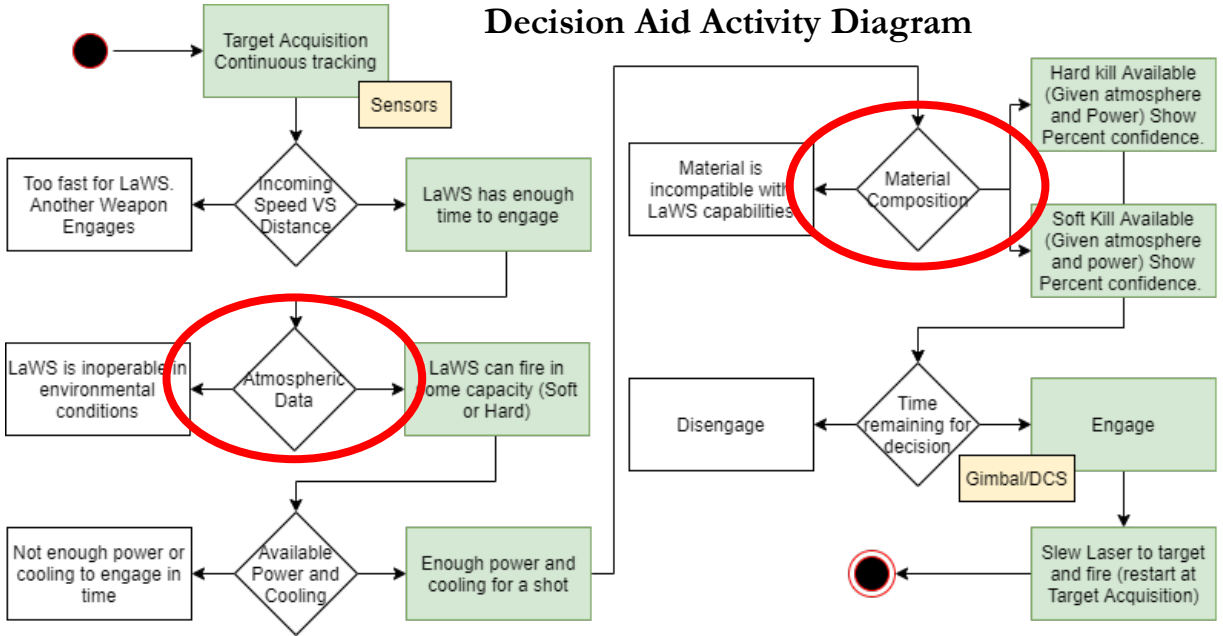
*Onboard Resources include Available Power and Cooling

**Threat Environment and Threat Environment Request include Detect, Track, ID, Environmental Conditions

Complex Laser Weapon Decision Space

- Small engagement time windows
- Variable optimal range based on weather and target
- Variable dwell time based on available power, range and target
- Variable magazine capacity based on power, cooling, range and target
- Deconfliction (Field of fire clearing) concerns due to propagation past target
- Target fix, aim and tracking requiring high speed and accuracy
- Beam focus (adaptive optics) requiring sub-second feedback
- Complex weapon selection based on these variables
- Target composition and aimpoint selection
- Selection of soft-kill vs. hard-kill
- Damage assessment

Decision Aid Activity Diagram



NAVAL POSTGRADUATE SCHOOL

Wrap Up

- AI/ML has huge potential for defense applications
- ML systems are different than traditional systems – we need to be mindful of new challenges and new types of failure modes
- Systems Engineering and Acquisition are entering a new frontier with AI/ML systems – we need new ideas, methods, and strategies
- Exciting research opportunities:
  - AI/ML applications for defense (tactical kill chain, directed energy, air and missile defense)
  - Engineering AI/ML systems (system safety, data management, system design)

I welcome collaboration!

Dr. Bonnie Johnson
bwjohnson@nps.edu