Metacognition for Artificial Intelligence Systems: an Approach to Safety and Desired Behavior in Complex Systems

2021 ICONS
16th International Conference on Systems

AI4SysSoS Special Track

20 April 2021

Dr. Bonnie Johnson
NPS 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.

Two Primary Types of AI

Explicitly Programmed

Handcrafted Knowledge Systems

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

Learns from Data

Machine Learning Systems

- 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

Three Types of AI System Application Domains

Data product systems use computers to generate information products.

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

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. The machine learning “system” is the trained 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.

**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 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.

**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
- Skewed Outcomes
- Prediction Outcomes are WRONG!
- Operators lose trust in the AI system
- Operators overly trust the AI system
- Operators ignore the AI system
- AI system in automated mode makes a poor decision
- AI system is overtaken by adversary (cybervulnerable)
- Adversary injects corrupt data into AI system
- Adversary jams or shuts down AI system
- Uncertain predictions arise from uncertainties in the data
Two types of AI systems according to the severity of their failure consequences

<table>
<thead>
<tr>
<th>Type A</th>
<th>Type B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Safety is Paramount</strong></td>
<td><strong>Safety is Less Important</strong></td>
</tr>
<tr>
<td>Applications in which AI system model predictions are used to support consequential decisions that can have a profound effect on people’s lives</td>
<td>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</td>
</tr>
</tbody>
</table>
## 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.)

### 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

<table>
<thead>
<tr>
<th>Pre-Deployment: Design, Development, Testing</th>
<th>Post-Deployment: Operations &amp; Sustainment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Inherently Safe Design</strong></td>
<td></td>
</tr>
<tr>
<td>Focus: ensuring robustness against uncertainty in the training data sets</td>
<td></td>
</tr>
<tr>
<td>- Interpretability – ensuring designers understand the complex AI and ML systems that are produced from the data training process</td>
<td></td>
</tr>
<tr>
<td>- Causality – reducing uncertainty by eliminating non-causal variables from the model</td>
<td></td>
</tr>
<tr>
<td><strong>2. Safety Reserves</strong></td>
<td></td>
</tr>
<tr>
<td>Focus: achieving safety through additive reserves, safety factors, and safety margins – through training data set validation</td>
<td></td>
</tr>
<tr>
<td>- Validating training data sets – eliminating uncertainty in the data sets; ensuring data sets are accurate, representative, sufficient, bias-free, etc.</td>
<td></td>
</tr>
<tr>
<td>- Increasing/improving model training process – ensuring adequate time and resources are provided for training and validation process</td>
<td></td>
</tr>
<tr>
<td><strong>3. Safe Fail</strong></td>
<td></td>
</tr>
<tr>
<td>Focus: system remains safe when it fails in its intended operation</td>
<td></td>
</tr>
<tr>
<td>- Human operation intervention – the operation of AI systems should allow for adequate human-machine interaction to allow for system overrides and manual operation</td>
<td></td>
</tr>
<tr>
<td>- 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</td>
<td></td>
</tr>
<tr>
<td>- Explainability/Understandability/Trust-worthy</td>
<td></td>
</tr>
<tr>
<td><strong>4. Procedural Safeguards</strong></td>
<td></td>
</tr>
<tr>
<td>Focus: measures beyond ones designed into the system; measures that occur during operations</td>
<td></td>
</tr>
<tr>
<td>- Audits, training, posted warnings, on-going evaluation</td>
<td></td>
</tr>
</tbody>
</table>
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

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 Decision Systems

Complex Environment (Problem Space)

Sensory Data

Model of Environment

Decision Recommendations/Prediction Outcomes

Decision Recommendations

Human Operator

Effector

Real World

AI System

Decision Space

Metacognition Model

Metacognition Memory Model
AI for Data Products

Big Data Repositories of Internet Data

Data Product Requests and Commands from Companies

Massive amounts of data

Metacognition Processes

Targeted Advertising
Spam Filtering
Image Recognition

Data Products for Companies

Speech Translation
Keyword Search
Product Search
Website Search

Data Products for Individuals

Data products that are “pushed” to individuals

Real World

AI System Virtual World
Wrap Up

- AI/ML 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


