Application of Systems Theoretic Process Analysis (STPA) to the Mission Assurance of AI-Enabled Systems

Robert B. Crombie
Daniel J. Byrne
Systems Integration and Test Office

September 5, 2023

Approved for public release. OTR 2023-01140

Outline

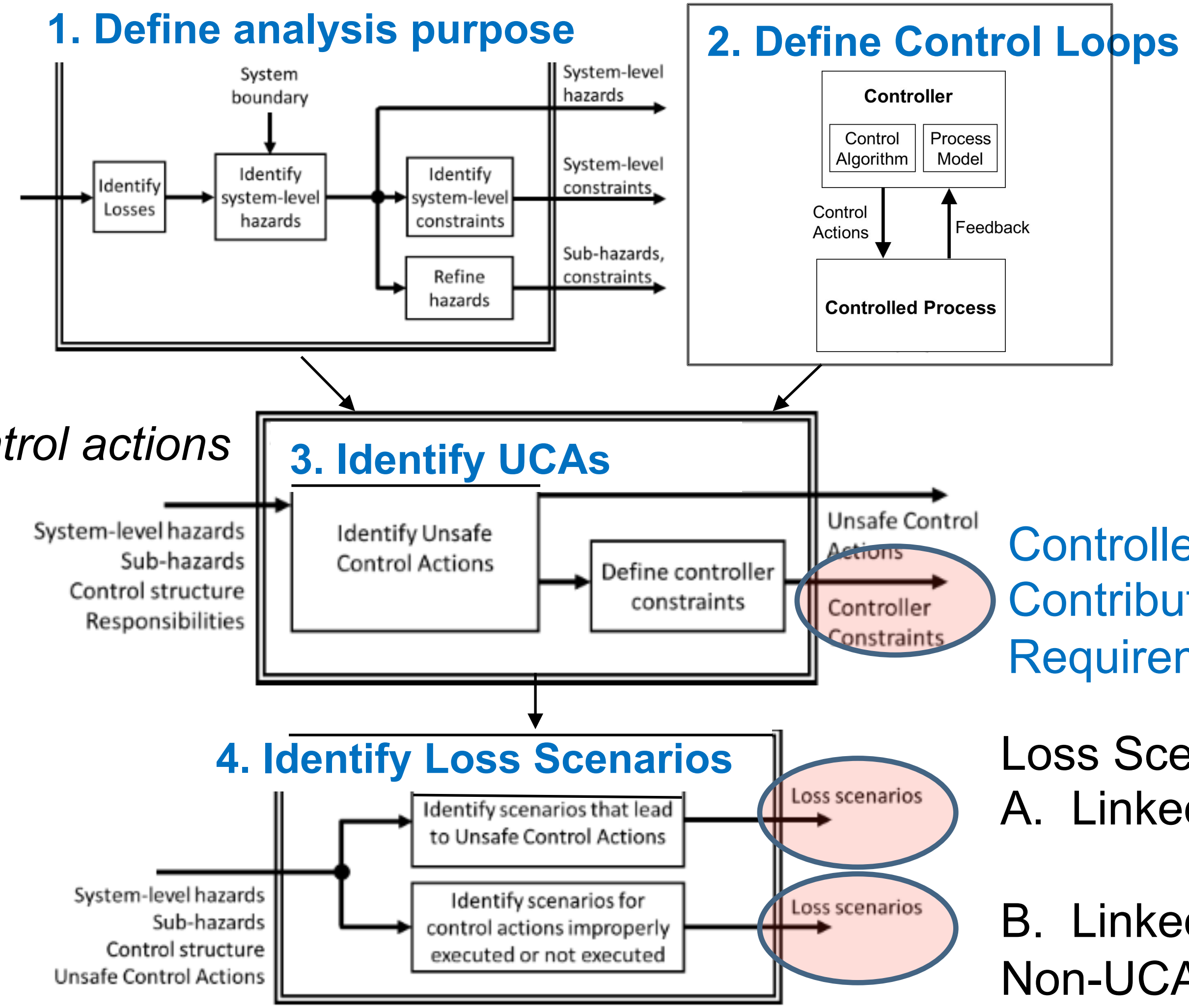
Application of STPA to the Mission Assurance of AI-Enabled Systems

- Goal: Explore the utility of STPA to the mission assurance of AI-enabled systems
- STPA Overview
- Application to Natural Language Processing (NLP) AI-Enabled System
  - *Losses and Hazards*
  - *Control structures, system constraints*
  - *Loss scenarios*
  - *AI (NLP) Mitigation Approaches*
  - *Utility of STPA in:*
    - Aerospace Trusted AI Framework
    - Aerospace Mission Assurance Guidelines for AI-enabled systems
- Application of STPA to Image Processing AI-Enabled System
- Summary and Next Steps

STPA Overview

System Theoretic Process Analysis

- 1. Define analysis purpose
  - *Define Losses*
  - *Define hazards*
    - Link to Losses
  - *Identify constraints (system requirements)*
- 2. Define System Control Loops
  - *Define controllers, their functions, and control actions*
- 3. Identify Unsafe Control Actions (UCAs)
    - Link to hazards
- 4. Identify loss scenarios
    - Link to UCAs
    - Link to other Non-UCA cases
  - *Identify mitigations*



Controller Constraints Contribute to system Requirements

Loss Scenarios:

A. Linked to UCAs

B. Linked to other Non-UCA cases

Base Diagram Source: Nancy Leveson and John Thomas “STPA Handbook”, 2018, available from the MITRE Partnership for Systems Approaches to Safety and Security (PSASS) web site at http://psas.scripts.mit.edu/home/.

Application of STPA to NLP System to Identify Potential Enterprise Risks

Uses two NLP AI models

- An "Enterprise Risk" defined as a risk with negative impact to <u>multiple Aerospace customers</u>
  - *Provides visibility to Aerospace executive leadership of emerging space enterprise issues for special focus*
    - Example: launch base consumable shortages
  - *Several sources (Industrial Base actions, Critical Technologies, Readiness Reviews, Strategic Materials, NLP System)*
- NLP system input: internal End of Week (EOW) reports by customer-facing line management
  - *NLP open-source English-language **sentiment model** (XLNET): interprets the sentiment polarity in a text paragraph*
    - Adapted using **supervised training** based on ground-truth labeled EOW training data
    - Identifies "risks" (negative sentiment) for <u>all customers</u>
  - *NLP **similarity model** compares all EOW reports*
    - **Unsupervised training**, with metrics built into the model to help it learn without labeled data
    - Tags "risks" with a similarity score (<u>between customers</u>) of 65% or greater as **potential enterprise risks** for human review

NLP System Losses and Hazards

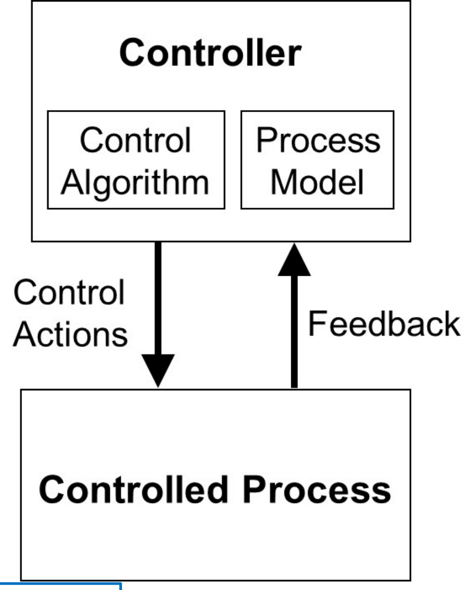
*A **hazard** is a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to a **loss**.*

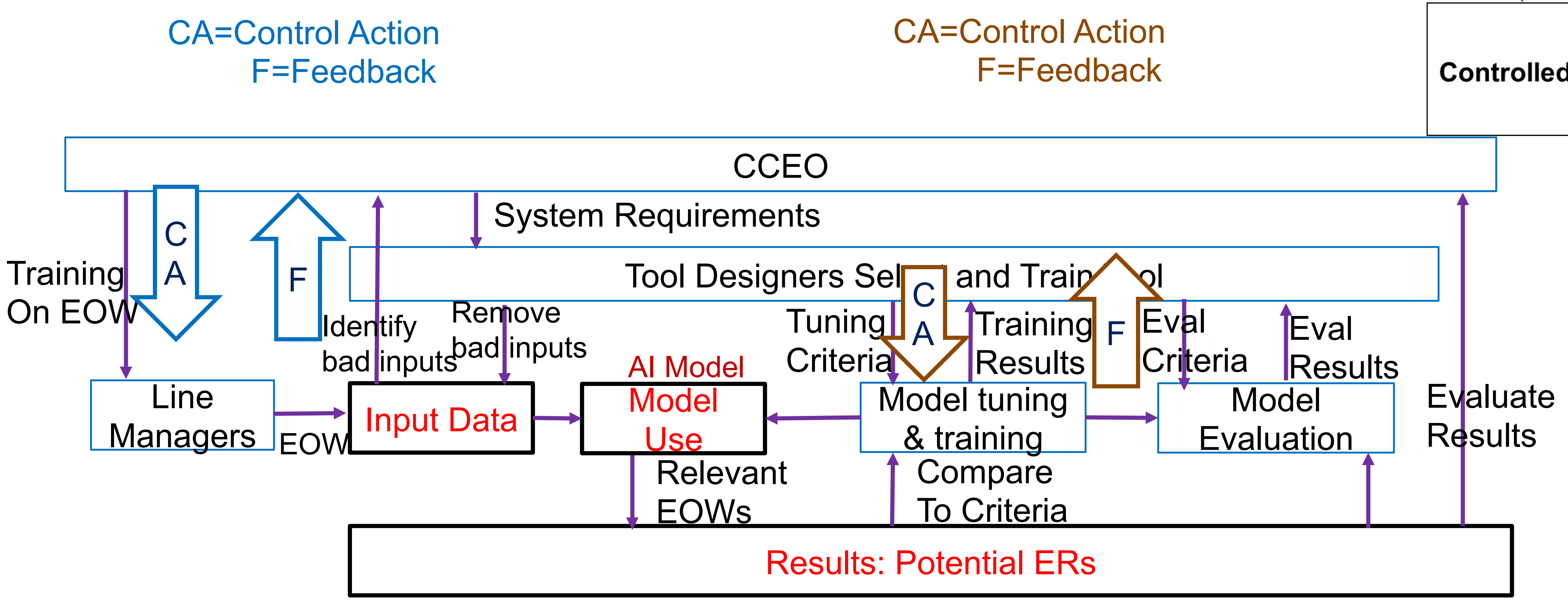
- Losses
  - *Aerospace customer needs go unmet*
  - *Customer data is compromised*
  - *Aerospace fails to identify valid enterprise risks (ERs)*
  - *NLP system costs detract from other priority technologies*
- Hazards
  - *Customer-facing managers do not record negative items in EOW report*
  - *Backdoor malware in open-source NLP models*
  - *NLP Model logic is flawed and misses whole category of important risks*
  - *Too many false positives are identified*
  - *NLP system overlooks important ER*
  - *Too many false negatives are identified*
  - *Excessive NLP system lifecycle costs*
  - *Non-Disclosure information from EOW reports is disclosed*

These were used to generate system constraints (requirements) and enforcement control structures

Overall Control Structure for NLP System

Corporate Chief Engineer's Office (CCEO) sponsored the development

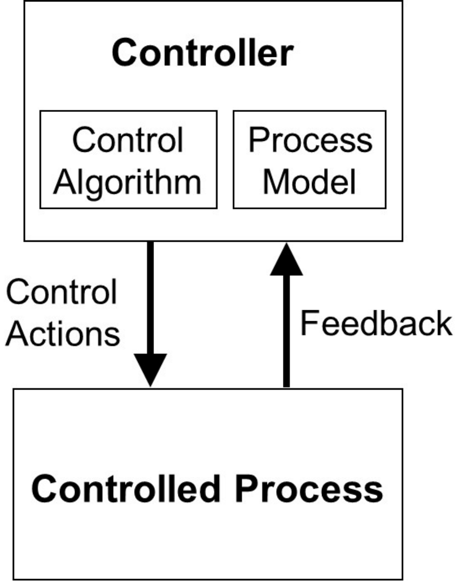




Two of the 13 (embedded) control structures identified

Selected STPA Control Structures based on System Constraints

Two of 13 NLP System Control structures



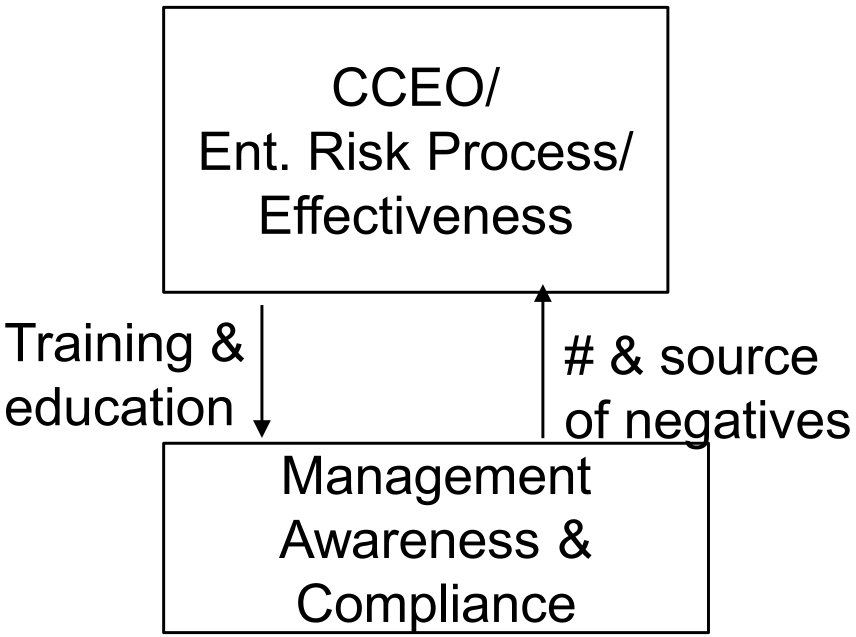
System Constraint	What can enforce constraints?	System Hazard
SC-2: EOW guidance to customer-facing group to emphasize reporting negatives	CCEO EOW reporting guidance	Customer-facing group does not record negative item in EOW report
SC-5: Sentiment Model tuning requirements on false positive rate	Model tuning exit criteria	Too many false positives are identified

Controller/ Algorithm/ Process Model

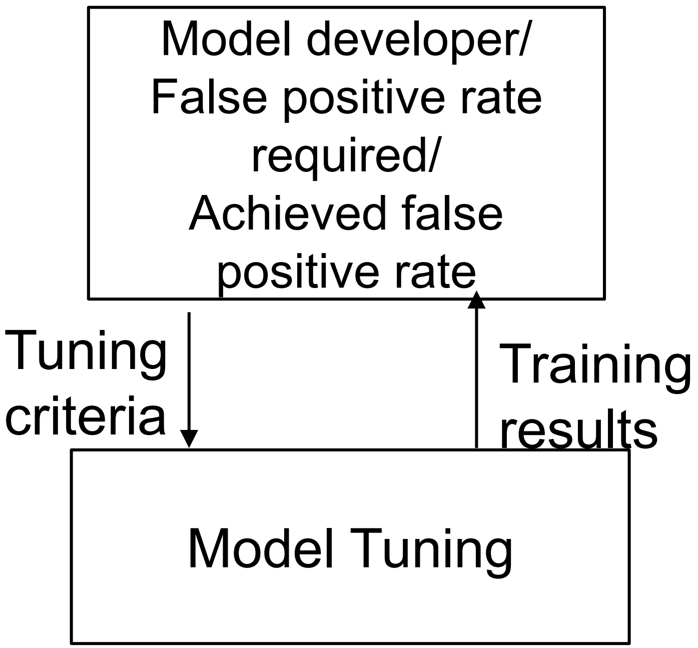
Control Actions
Feedback

Controlled Process:

SC-2

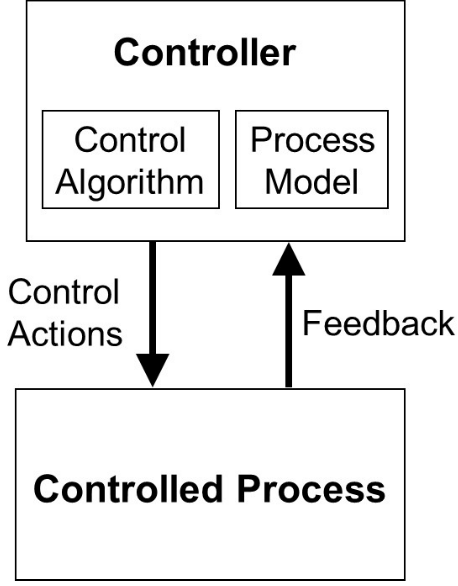


SC-5



Unsafe Control Actions (UCAs) and Loss Scenarios

NLP System



- 69 UCA loss scenarios identified from 13 control structures:
  - *39 Loss Scenarios postulated (3 for each control structure) from these UCA types:*

Loss Scenarios Caused by UCA Type:		
a. Inadequate control algorithm	b. Unsafe input from (other) controllers	c. Controller failures

  - *19 Loss scenarios caused by Process Model flaws*
  - *25 Loss scenarios caused by Inadequate Feedback*
- 86 “non-UCA” loss scenarios:
  - *23 Loss scenarios for Control Action not executed (by the actuator)*
  - *33 Loss scenarios for Control Action Improperly Executed (by actuator)*
  - *30 Loss scenarios related to the Controlled Process*

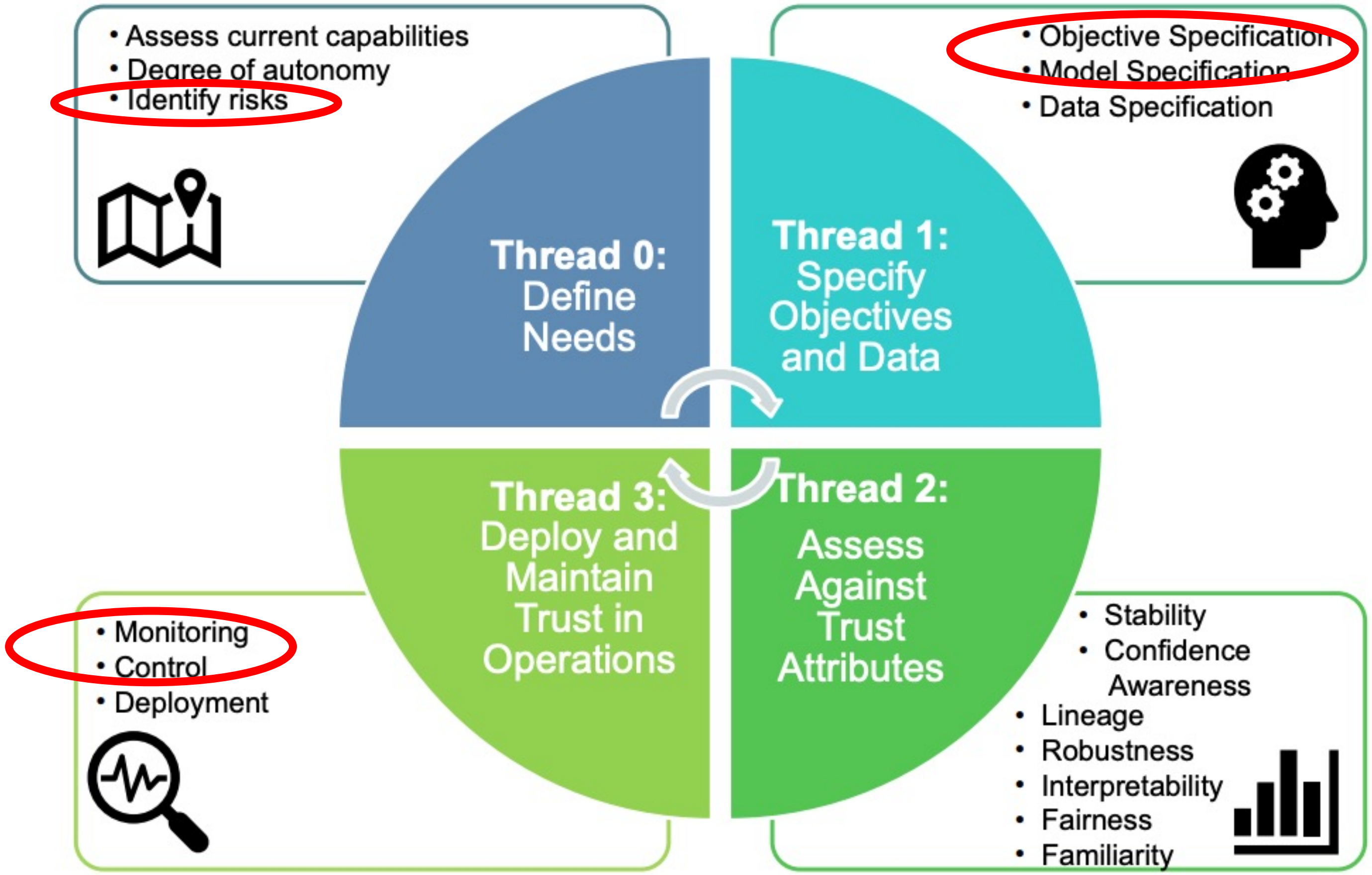
Mitigation approaches determined for each UCA and non-UCA Loss Scenario

Mitigations Relevant for AI-Enabled Systems

NLP System STPA Analysis proposed over 100 mitigations

- Topical mitigations relevant for mission assurance of AI-enabled systems:
  - *Corporate processes for AI system development*
    - System requirements and architecture
    - AI system design, including peer reviews
    - Model training
    - Life cycle cost considerations
    - Standard design review process for AI-enabled systems
      - *Carefully crafting and implementing system requirements, particularly for model tuning*
- Specific AI-focused mitigation approaches identified included:
  - *Selecting appropriate training data for the NLP model*
  - *Monitoring NLP model input data*

Aerospace Trusted AI Framework: Utility of STPA



- Thread 0: Identify risks to trust attributes
- Thread 1: Determine failure modes and constraining requirements (specifications) for the objective system and model
- Threads 2 and 3: Guide system development through the mitigations and requirements identified by STPA in earlier threads.
- Thread 3: Implement system requirements for monitoring and control derived from STPA analysis

Aerospace Mission Assurance Guidelines: Utility of STPA

MA for AI-Enabled Systems

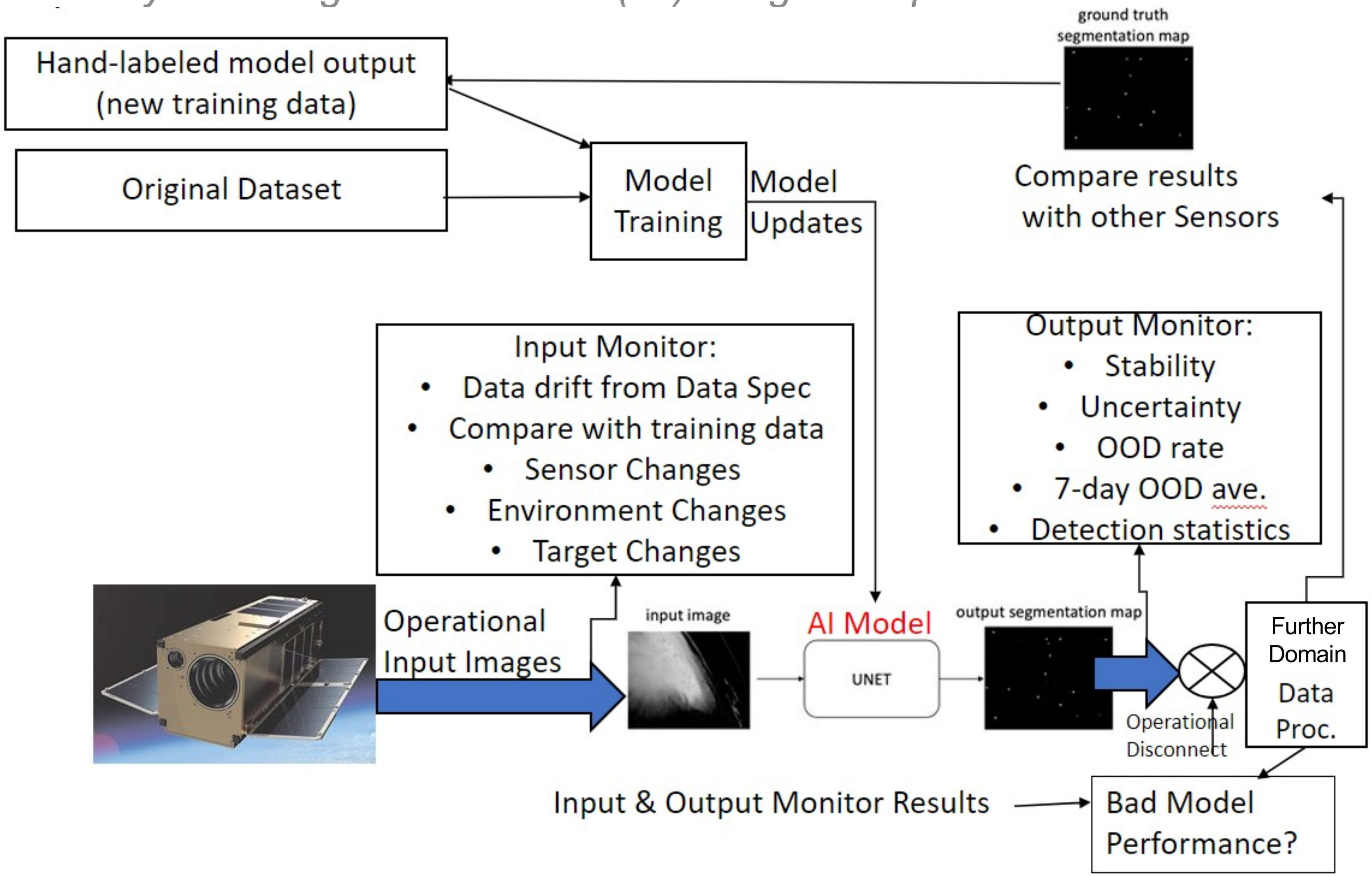
Specific Best Practices

Trusted Sources	Performance Analysis	Fault and Redundancy Management	Hardware & Cyber	Prototyping and Verification	Usability & Operator Training
•Data Sources •Data Specification •Data Configuration Management •Open Source Algorithms / Software •Trusted AI Framework Threads	•Object Specification •Algorithm Performance Metrics •Software Performance Metrics •Repeatable System Behavior •Reproducibility •Uncertainty & Confidence •Equitability	•ML Reliability •Enhancing the predicitability of AI controlled state systems •Adversarial Robustness •Monitoring & Control	•AI Hardware •Algorithm Encryption •Data Encryption •Ensuring systems cannot be reversed engineered for NSS capabilities •Computational Requirements •ML specific cyber concerns	•HIL/SIL platforms to demonstrate behavior •Evolved I&T to verify the system-to-AI-system response chain	•HMI / HMT •Training for Systems with AI •Familiarity •Interpretability •Monitoring & Control

- Trusted Sources:
  - *As described above, STPA can help with aspects of the Trusted AI Framework Threads.*
- Fault and Redundancy Management:
  - *STPA can identify loss scenarios and mitigation approaches to enhance:*
    - ML reliability,
    - Adversarial robustness, and
    - Improve monitoring and control approaches.

2. Conceptual Image Processing System

Purpose: Identify faint targets in infrared (IR) image for space domain awareness



OOD=Out of Distribution

Neural Network AI Model (UNET) and Input/Output Monitors require training/tuning

Early Conclusions from STPA Analysis of Image Processing System

- For the Image Processing System STPA analysis:
  - *Using an MBSE tool made it easy to maintain the STPA database and generate tables*
  - *The system was treated as in the concept development phase. Focus was more on technical faults of the system architecture and less on development processes.*
    - STPA is effective for identifying issues during concept development to inform system architecture
  - *Identified a need for tools to support the system developers in analyzing the monitors' statistics and image data*

- Both AI-enabled systems' STPA analyses were able to identify issues and mitigations for each

- Key differences between the two systems are:
  - *NLP system uses **two** AI components operating on **serial** data in response to user input, queries, and model training*
  - *The image processing system uses **one** AI component that processes each image independently without consideration for previous images or external controls*
    - The latency between ingest of an image and delivery of the processed image to the user is very short

More examples of STPA applied to space systems are needed to aid STPA analyses of actual space systems

Summary and Next Steps

Application of STPA to the Mission Assurance of AI-Enabled Systems

- Summary:
  - *STPA analysis of NLP system showed*
    - Thorough assessment of interactive risks with proposed mitigations (including system requirements)
      - *Topical mitigations relevant to many AI-enabled systems*
      - *Specific AI-focused mitigation approaches at corporate level for AI-enabled system development*
    - Proved utility of STPA in Aerospace Trusted AI Framework and Mission Assurance guidelines
  - *STPA analysis of Image Processing System showed*
    - STPA is effective for identifying issues during concept development to inform system architecture

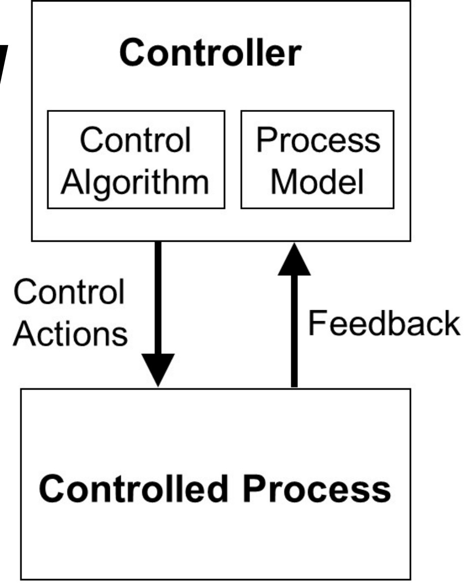
- Next steps:
  - *Complete STPA analysis of Image Processing System and report this month*
  - *Brief internal customers in Fall*
    - Offer assistance in STPA analysis of real systems
  - *Brief NASA JPL Autonomy Seminar in Winter*
  - *FY24 IR&D effort proposed to develop STPA analysis process using MBSE to produce SoS requirements*

STPA can define needed system requirements for systems with autonomous AI components

Questions?

21 possible interaction failures in an STPA control structure model

1-4 are "Unsafe Control Actions" (UCAs); 14-21 lead to UCAs



1. Not providing the control action (CA) leads to a hazard
2. Providing the control action leads to a hazard
3. Providing a potentially safe control action but too early, too late, or in the wrong order
4. The control action lasts too long or is stopped too soon (for continuous control actions)
5. Control action not received [by the actuator]
6. Control action not executed [by the actuator]
7. Control action not received [by the controlled process]
8. Control action improperly executed [by the actuator]
9. Actuator does not respond adequately to CA
10. CA actuator not applied or received properly at the controlled process, or CA not sent but actuators/elements respond
11. Control action not executed [by the controlled process]
12. Control action improperly executed [by the controlled process]
13. Control Action not received but Controlled Process still responds
14. Process Model (PM) ignores feedback/ interprets incorrectly
15. PM incorrect beliefs of states, modes, process, sensors, actuators, or past/future
16. PM incorrect beliefs about capabilities, dynamics, other processes, need to coordinate with other controllers
17. Feedback or information not received
18. Inadequate feedback is received from the controlled process
19. Inadequate control algorithm
20. Failures involving the controller (for physical controllers)
21. Unsafe control input (from another controller, possibly an adversary)

NLP System Hazards, Constraints, Enforcement Mechanisms

*A **hazard** is a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to a **loss**.*

System Hazard	System Constraint	What can enforce constraints?
Customer-facing group does not record negative item in EOW report	EOW guidance to customer-facing group emphasized reporting negatives	CCEO EOW reporting guidance
Backdoor malware in open-source model affects results	Security scan of model; security monitoring of system	Cybersecurity requirement on model acquisition and use.
NLP Model logic is flawed and misses whole category of important risks	Periodic review of model logic and design by CCEO	CCEO oversight
Too many false positives are identified	Sentiment Model tuning requirements on false positive rate	Model tuning exit criteria
NLP system overlooks important ER	Sentiment and Similarity Models tuning requirements on ER detection (true positive rate)	Model tuning criteria (iterative)
Too many false negatives are identified	Model tuning requirements on ER detection (true negative rate)	Similarity Model tuning test with user feedback loop
Excessive NLP system lifecycle costs	Simple cost-effective NLP models' maintenance	NLP Models' design, AI specialist effort, similarity tool license cost Effort for XLNET model training data acquisition and processing
Non-Disclosure information from EOW reports is disclosed	Protect model from external and internal unauthorized data disclosure	Model access controls and network firewalls
Several above	Protect model from data drift	Active input data monitoring and active results monitoring

NLP System Loss Scenario Mitigation Themes

Over 100 mitigations derived

Theme	Components
Front-end system planning	Define EOW, cyber, ER processes; hold requirements review; develop a maintenance plan
People-centered management processes & practices	Policies, procedures, staffing, education, assistance, management
System requirements	Tuning requirements, etc.
System architecture	Lifecycle cost including training and monitoring
Processes for system design	Peer reviews for many AI development steps
Corporate processes for IT and AI-enabled system development	Internal Access controls; IT services; Standardized development & review process for AI-enabled systems
Sponsor behavior	Cyber test schedule & response; assign independent reviewers; justify labor hours
Model developer behavior	Model tuning and design to requirements; Recognize input data effects
Model training process	Test data review against system requirements