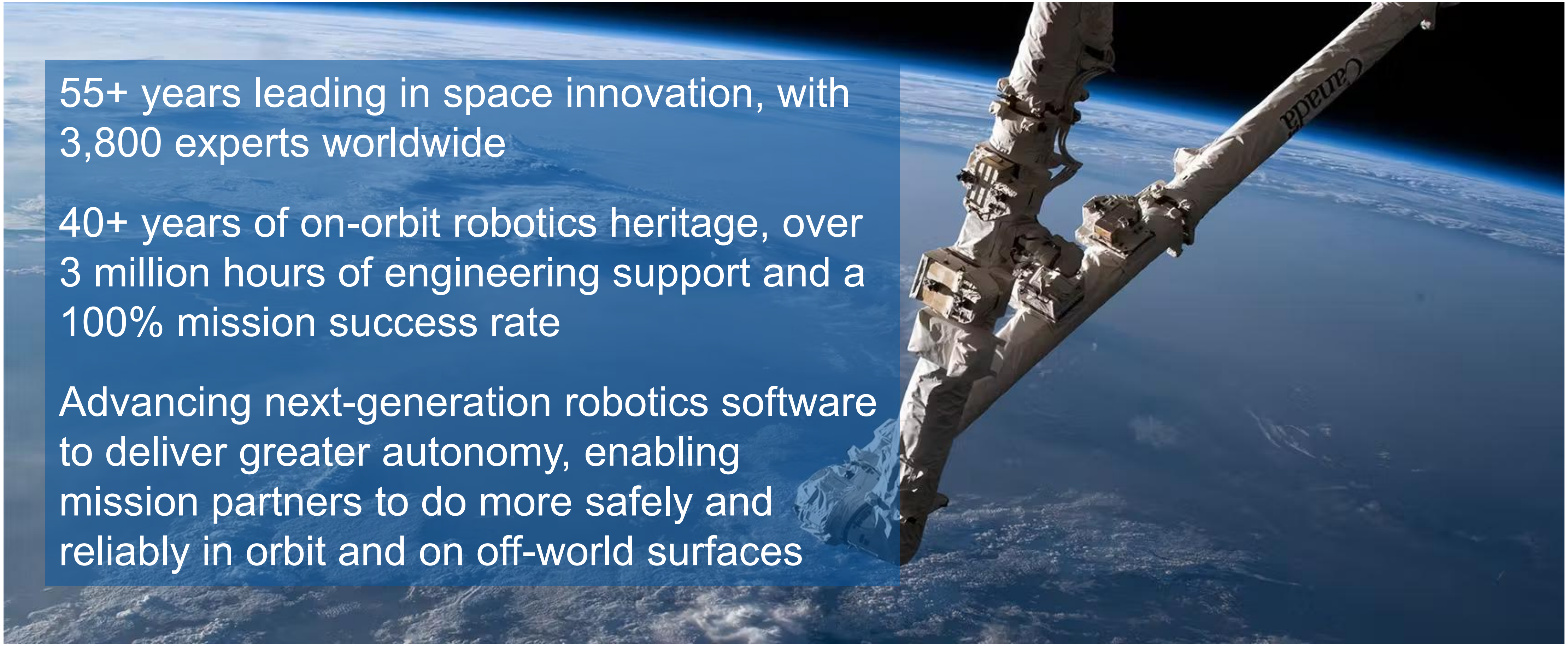
Product Assurance for AI-based Telemetry Health Monitoring of Safety-Critical Space Robotics

European Space Agency – Software Product Assurance Conference 2025

Kelly Gasperski, Simon Diemert, Emmanuel Lesser, Adam Casey, Nader Abu El Samid, Malav Naik, Justin Kernot, Laure Millet, **Jeff Joyce**, Paul Grouchy

MDA Space at a Glance

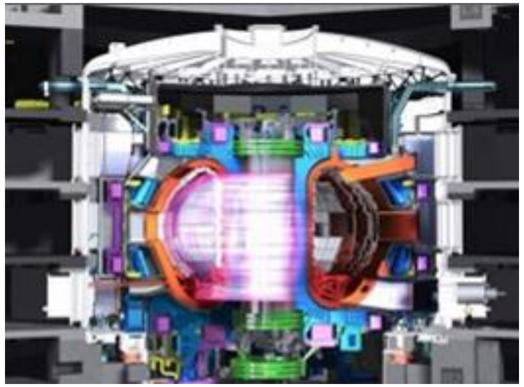
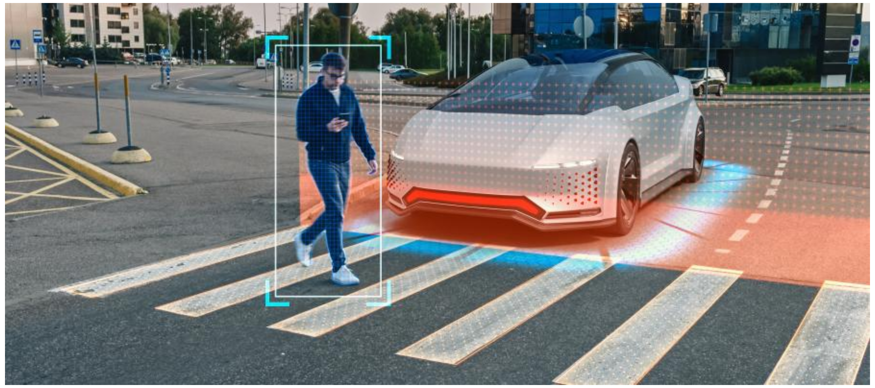


55+ years leading in space innovation, with 3,800 experts worldwide

40+ years of on-orbit robotics heritage, over 3 million hours of engineering support and a 100% mission success rate

Advancing next-generation robotics software to deliver greater autonomy, enabling mission partners to do more safely and reliably in orbit and on off-world surfaces

About Critical Systems Labs







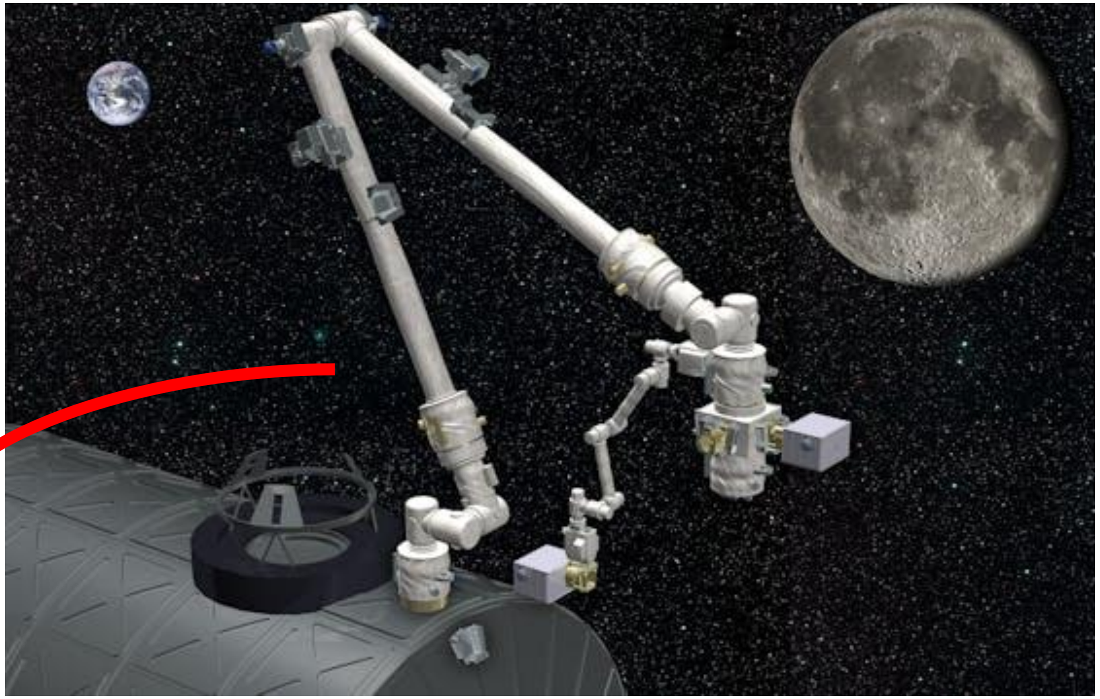
Presentation Abstract

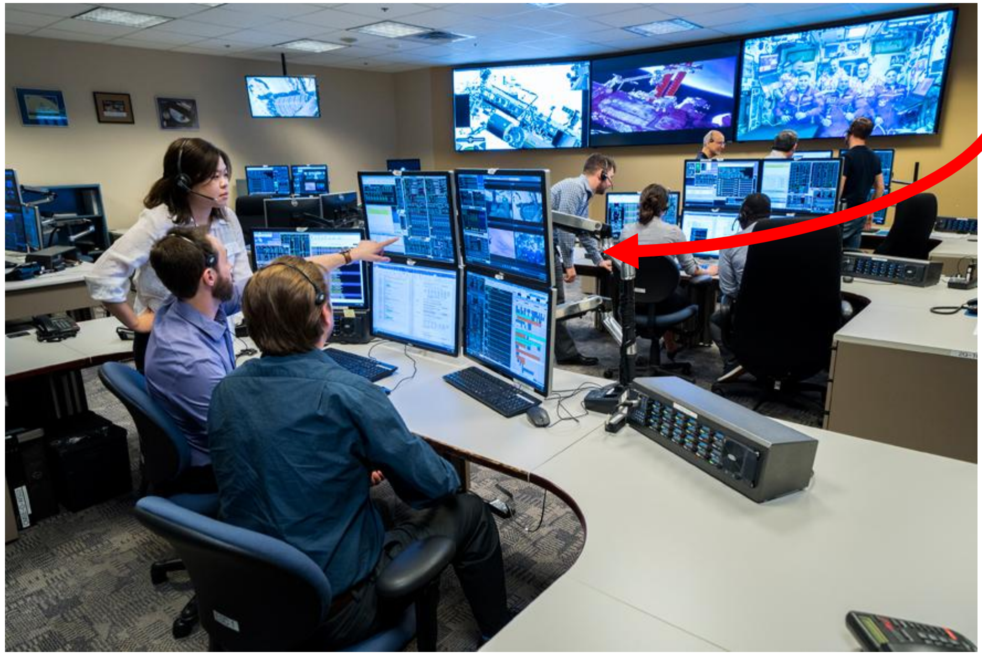
This presentation reports on assurance aspects of a study that explores the application of Artificial Intelligence (AI) and Machine Learning (ML) technologies in safety-critical space robotics, with a focus on the use of AI in the ground control segment to enhance operator decision-making. A key challenge is to assure safety and achieve safety certification in a context where established standards and published guidance are not entirely compatible with the use of AI/ML. This study is partially motivated by fundamental uncertainty about how AI-based functionality can be validated to provide high confidence in its safety, i.e., high recall, low false positives. This study also seeks to identify what strategies can be used to mitigate the human factors risk that operators become complacent by trusting the AI-functionality to monitor operational safety. While focused on a particular application for space-robotics, the results of this study will be broadly applicable beyond the space community including other technical domains such as automotive and medical devices that are rapidly integrating AI/ML into safety-critical technology. Preliminary findings and future research directions will be presented

Problem Formulation

Operators review telemetry data from the Flight Segment to check for **anomalous behaviour** suggestive of system failures.

Flight Segment

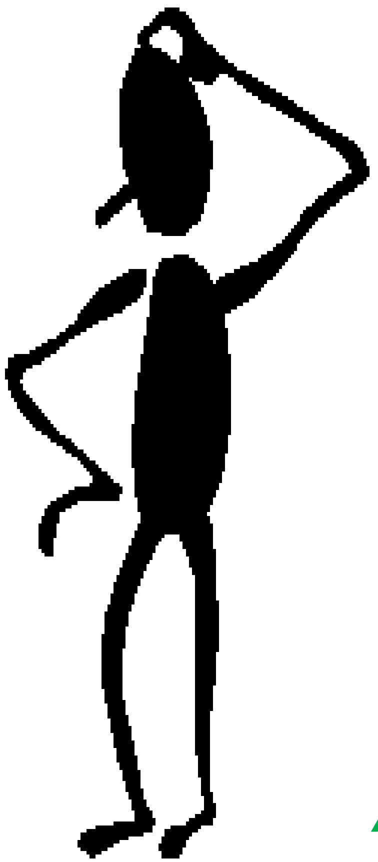




Ground Segment

Key Operational Challenges:

1. Rule-based monitoring (“if-then-else”) is difficult to scale to complex anomalies.
2. Manual review by operators can miss subtle anomalies, especially in a demanding operational environment.



Will using AI impact operator capability?

AI does not get “tired”

How do we certify an AI-enabled system?

Can we use AI to detect anomalies in telemetry data sent by the flight segment?

Do we trust an AI to do this task?

AI can analyze data very quickly

AI can detect complex anomalies, based on previous experience (training data)

Perform Maneuver → Transmit Telemetry Data

Manual Review of Telemetry Data by Operators

Rule-based Telemetry Data Analysis

AI Telemetry Data Analysis

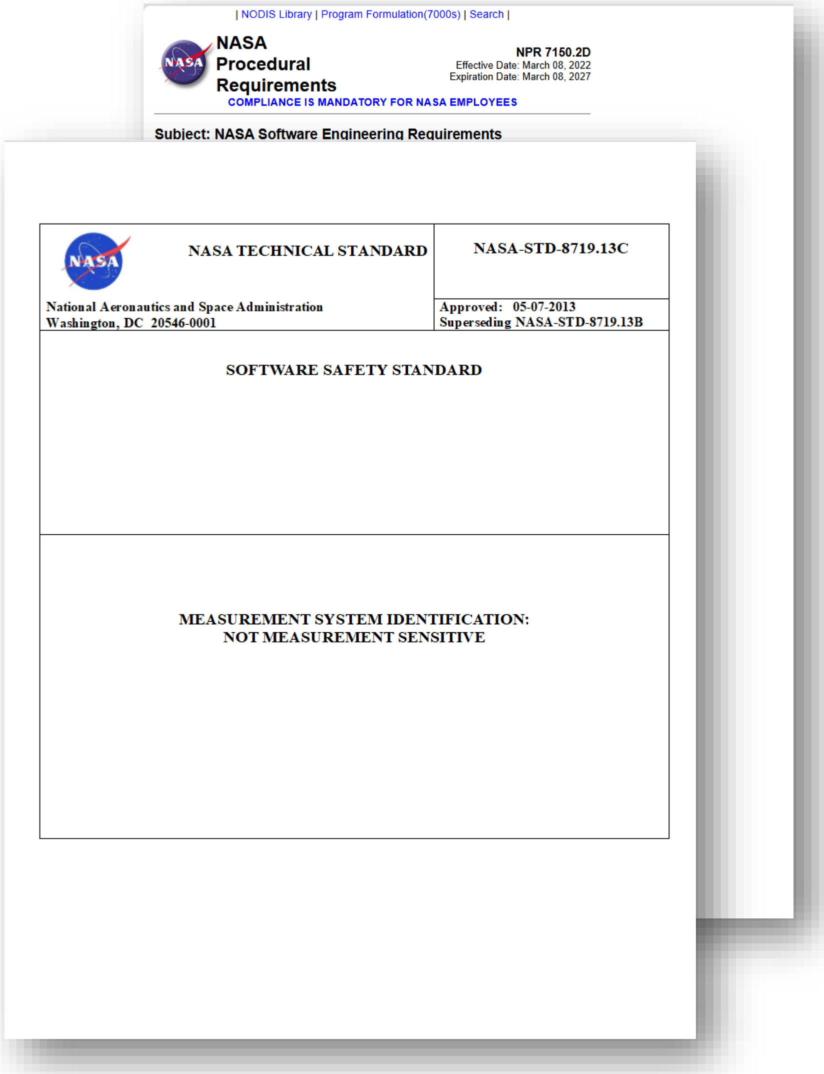
Integrate and Display Results

Make Operational Decisions

Guiding Principle: Introducing AI-based telemetry monitoring should reduce safety risk relative to the same system without it.

But we still must certify it!

- Safety-related software should be certified:
  - NASA-STD-8719.13
  - NPR-7150.2
  - ECSS-E-ST-40C and ECSS-Q-ST-80C
- But these standards do not account for AI or machine learning technology.
- *What do other industries do?*


| NODIS Library | Program Formulation(7000s) | Search |

NASA Procedural Requirements

NPR 7150.2D
Effective Date: March 08, 2022
Expiration Date: March 08, 2027

COMPLIANCE IS MANDATORY FOR NASA EMPLOYEES

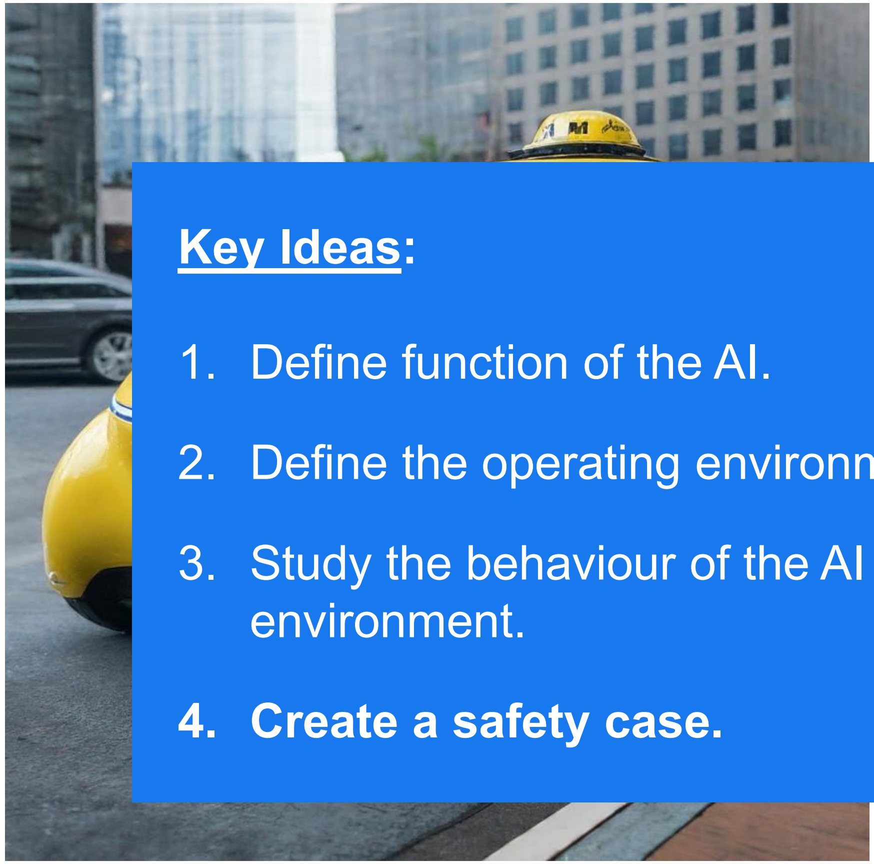
Subject: NASA Software Engineering Requirements

NASA TECHNICAL STANDARD	NASA-STD-8719.13C
National Aeronautics and Space Administration Washington, DC 20546-0001	Approved: 05-07-2013 Superseding NASA-STD-8719.13B

SOFTWARE SAFETY STANDARD

MEASUREMENT SYSTEM IDENTIFICATION:
NOT MEASUREMENT SENSITIVE

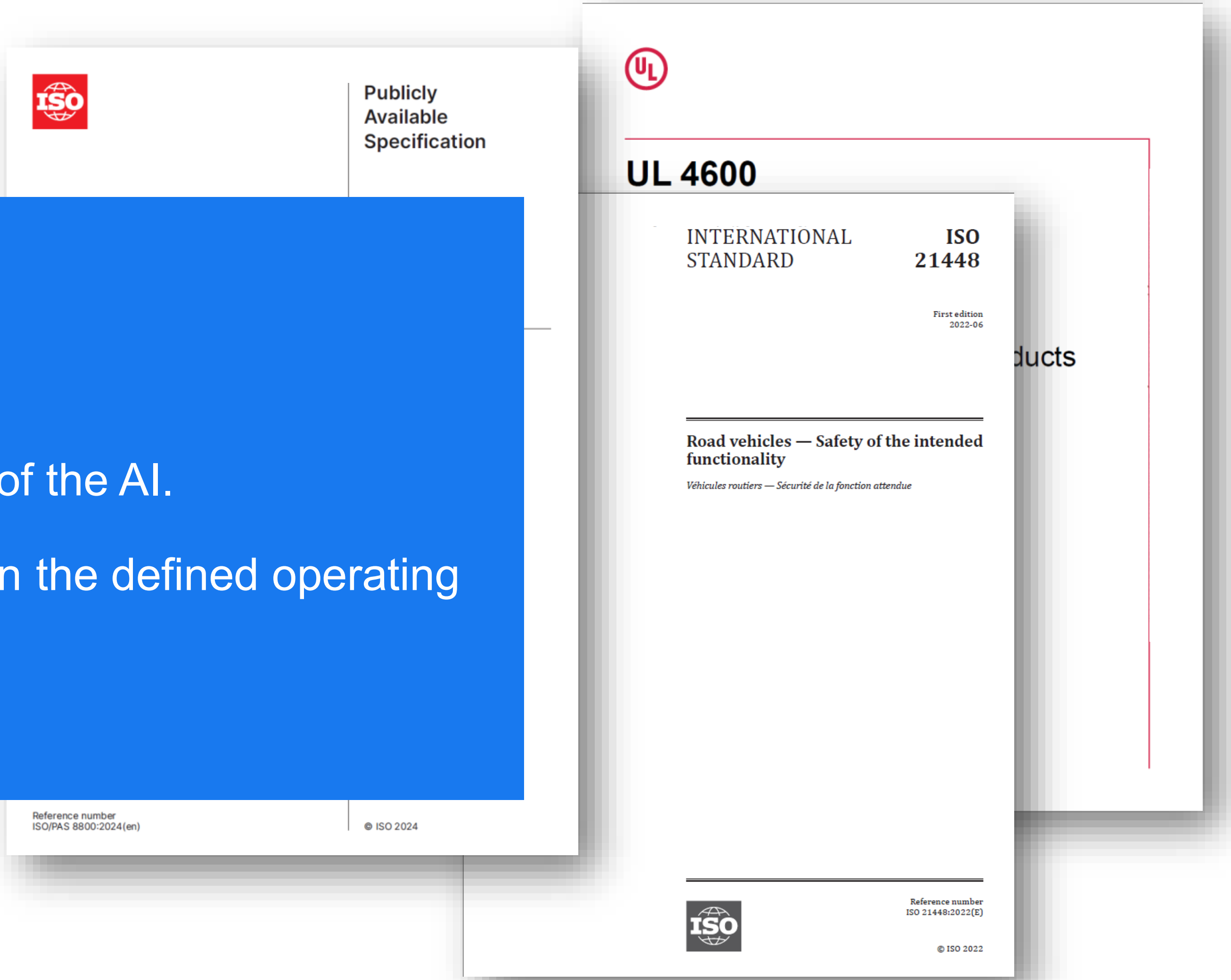
Autonomous vehicles use AI, right?



Generated by Google Gemini II.

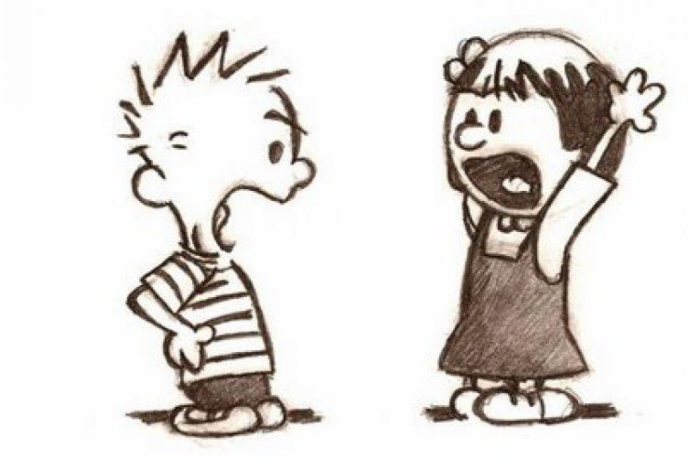
Key Ideas:

1. Define function of the AI.
2. Define the operating environment of the AI.
3. Study the behaviour of the AI within the defined operating environment.
4. **Create a safety case.**

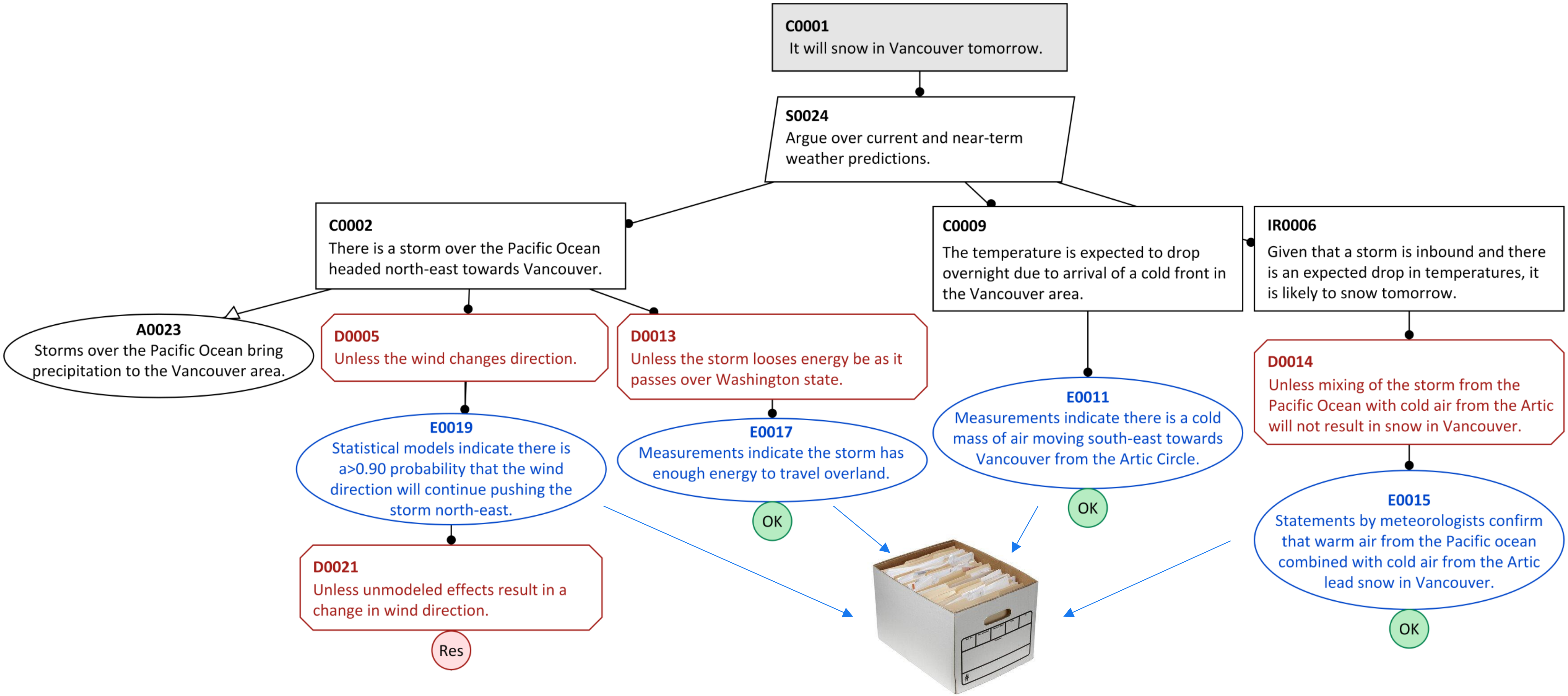


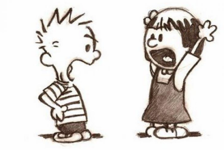
Three Objectives

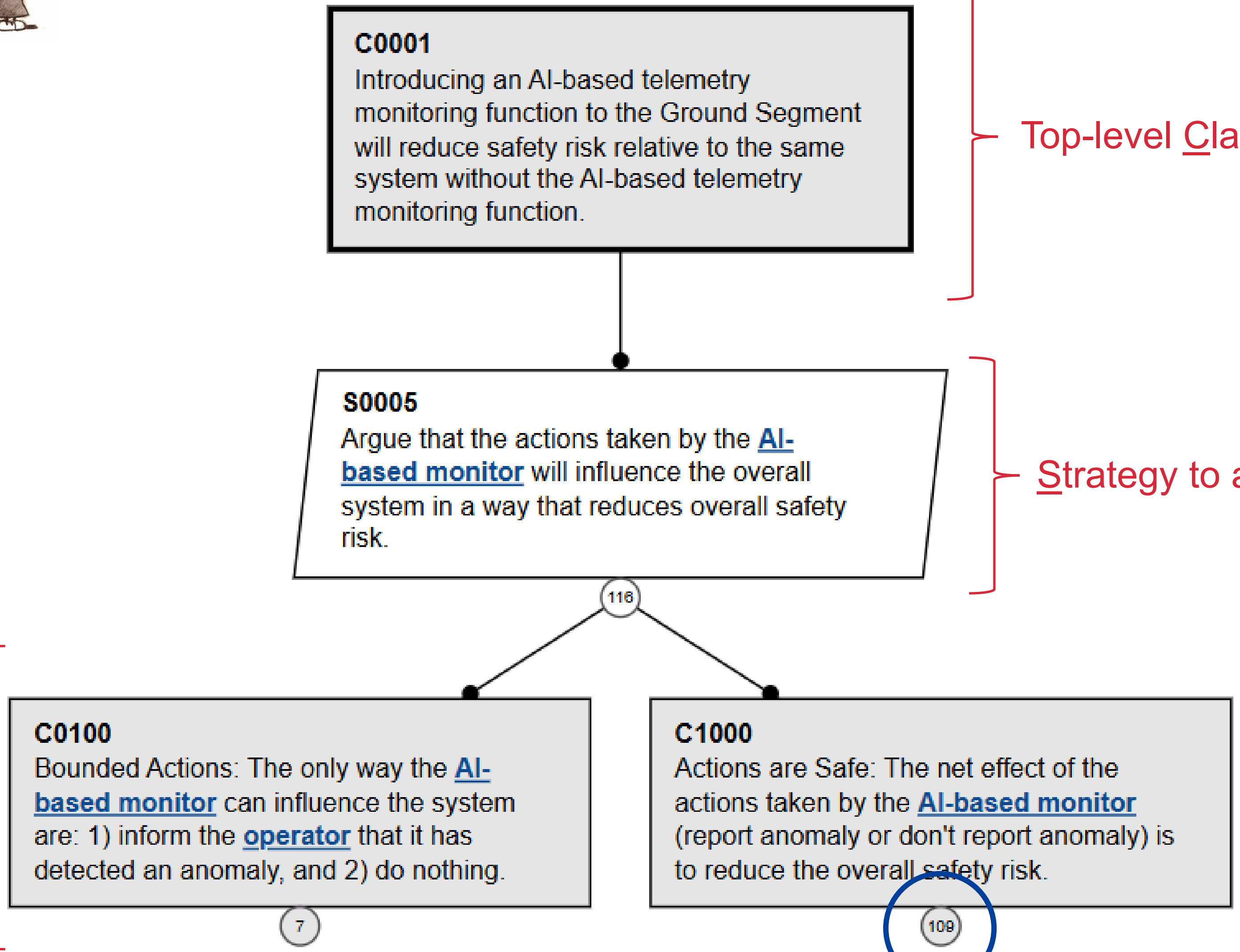
1. Develop a **safety case argument** for AI-based telemetry monitoring.
2. Review literature on **human factors** related to AI-based monitoring systems.
3. Identify and apply **validation methods** for AI models.



+


C0001
It will snow in Vancouver tomorrow.
S0024
Argue over current and near-term weather predictions.
C0002
There is a storm over the Pacific Ocean headed north-east towards Vancouver.
C0009
The temperature is expected to drop overnight due to arrival of a cold front in the Vancouver area.
IR0006
Given that a storm is inbound and there is an expected drop in temperatures, it is likely to snow tomorrow.
A0023
Storms over the Pacific Ocean bring precipitation to the Vancouver area.
D0005
Unless the wind changes direction.
D0013
Unless the storm looses energy be as it passes over Washington state.
E0011
Measurements indicate there is a cold mass of air moving south-east towards Vancouver from the Artic Circle.
D0014
Unless mixing of the storm from the Pacific Ocean with cold air from the Artic will not result in snow in Vancouver.
E0019
Statistical models indicate there is a>0.90 probability that the wind direction will continue pushing the storm north-east.
E0017
Measurements indicate the storm has enough energy to travel overland.
OK
OK
E0015
Statements by meteorologists confirm that warm air from the Pacific ocean combined with cold air from the Artic lead snow in Vancouver.
OK
D0021
Unless unmodeled effects result in a change in wind direction.
Res





Top-level Claim

Strategy to argue this claim

Two branches:

1. Bounded Actions
2. Net Reduction in Safety Risk

C1000
Actions are Safe: The net effect of the actions taken by the AI-based monitor (report anomaly or don't report anomaly) is to reduce the overall safety risk.

Divide this claim into two cases….

C1001
When an anomaly is present, actions taken by the monitor will reduce risk overall.

93

C1002
When an anomaly is not present, actions taken by the monitor will only minimally increase overall risk, if at all.

92

IR1008
"An anomaly is present" and "an anomaly is not present" covers all possible cases. If risk is reduced overall when an anomaly is present, and no more than minimally increased when an anomaly is not present, then overall risk is reduced.

4

Case: Anomaly Present

Case: Anomaly Not Present

Divide this claim into two cases….

C1001
When an anomaly is present, actions taken by the monitor will reduce risk overall.

C1003
If the AI-based monitor correctly reports an anomaly (true positive), safety risk is maintained or decreased.

12

Case: True Positive

C1007
If the AI-based monitor fails to report an anomaly (false negative), safety risk is minimally increased.

10

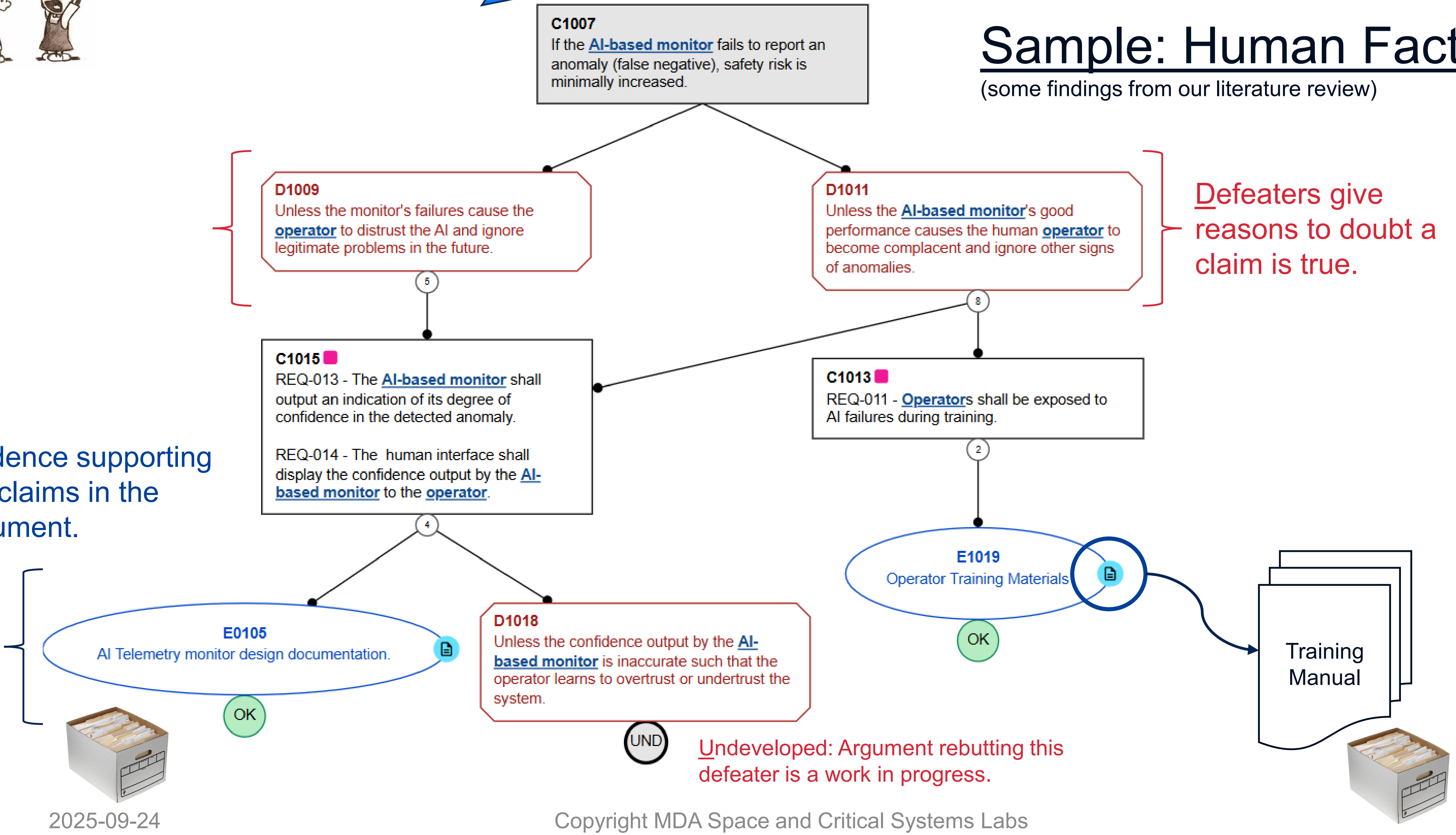
Case: False Negative

IR1016
If correctly reporting an anomaly maintains or decreases safety risk, and failing to report an anomaly only minimally increases safety risk, then safety risk overall is maintained or decreased when an anomaly is present.

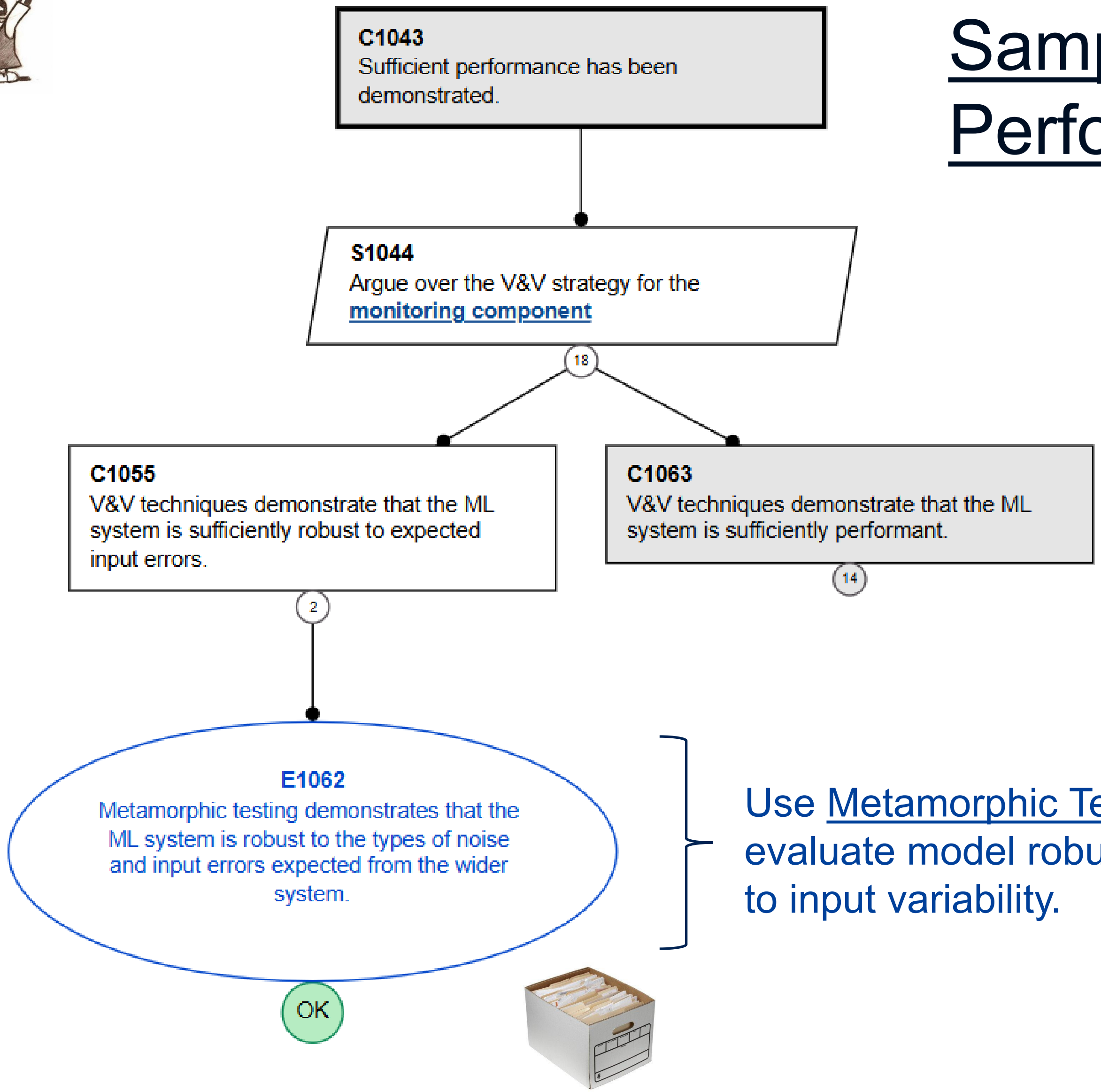
77

Sample: Human Factors

(some findings from our literature review)



Sample: Validation of Model Performance



Evaluate model performance across operating domain with Combinatorial Testing.

Use Metamorphic Testing to evaluate model robustness to input variability.

Metamorphic Testing: Overview

DOI:10.1145/3241979

Metamorphic testing can test untestable software, detecting fatal errors in autonomous vehicles' onboard computer systems.

BY ZHI QUAN ZHOU AND LIQUN SUN

Metamorphic Testing of Driverless Cars

ON MARCH 18, 2018, Elaine Herzberg became the first pedestrian in the world to be killed by an autonomous vehicle after being hit by a self-driving Uber SUV in Tempe, AZ, at about 10 P.M. Video released by the local police department showed the self-driving Volvo XC90 did not appear to see Herzberg, as it did not slow down or alter course, even though she was visible in front of the vehicle prior to impact. Subsequently, automotive engineering experts raised questions about Uber's LiDAR technology.[12] LiDAR, or "light detection and ranging," uses pulsed laser light to enable a self-driving car to see its surroundings hundreds of feet away.

Velodyne, the supplier of the Uber vehicle's LiDAR technology, said, "Our LiDAR is capable of clearly imaging Elaine and her bicycle in this situation. However, our LiDAR does not make the decision to put on the brakes or get out of her way" ... "We know absolutely nothing about the engineering of their [Uber's] part ... It is a proprietary secret, and all of our customers keep this part to themselves"[15] ... and "Our LiDAR can see perfectly well in the dark, as well as

» key insights

- **Many software systems (such as AI systems and those that control self-driving vehicles) are difficult to test using conventional approaches and are known as "untestable software."**
- **Metamorphic testing can test untestable software in a very cost-effective way, using a perspective not previously used by conventional approaches.**
- **We detected fatal software faults in the LiDAR obstacle-perception module of self-driving cars and reported the alarming results eight days before Uber's deadly crash in Tempe, AZ, in March 2018.**

MARCH 2019 | VOL. 62 | NO. 3 | COMMUNICATIONS OF THE ACM 61

Z. Q. Zhou and L. Sun, "Metamorphic testing of driverless cars," *Commun. ACM*, vol. 62, no. 3, pp. 61–67, Feb. 2019.

Specifying Functional Safety Requirements for AI/ML in terms of Metamorphic Relations

Laure Millet; Critical Systems Labs Inc; Vancouver, BC, Canada

Simon Diemert; Critical Systems Labs Inc; Vancouver, BC, Canada

Rami Debouk; General Motors; Warren, Michigan, USA

Ramesh S; General Motors; Warren, Michigan, USA

Mateo Delgado; Critical Systems Labs Inc; Vancouver, BC, Canada

Jeffrey Joyce; Critical Systems Labs Inc; Vancouver, BC, Canada

Keywords: Functional Safety Requirements, Metamorphic Relations, Metamorphic Testing, Verification/Testing, Machine Learning

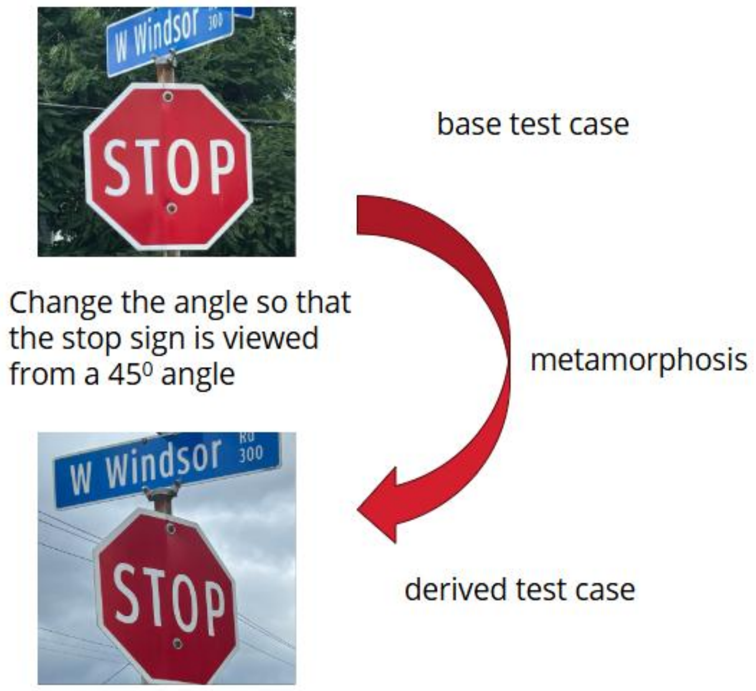
Abstract

The specification of functional safety requirements is an integral part of a safety engineering process, especially in the case of a system whose behavior is largely determined by software. Many industry standards, such as RTCA DO-178C (aerospace), ISO 26262 (automotive) and CENELEC EN 50128 (rail), underscore the importance of specifying functional safety requirements. However, for the development of software that uses Machine Learning (ML) or other forms of Artificial Intelligence (AI), there is a widespread view that the functionality that results from the use of AI/ML cannot be specified with the level of precision that would normally be expected for the specification of functional safety requirements. A possible solution to this conundrum is specifying functional safety requirements in the form of Metamorphic Relations (MR). In this form, requirements specify a constraint on how a change to a situation or a scenario should affect the behavior of the system. Specifying functional safety requirements in the form of MRs accommodates the inherent limitations on how precisely the predictable behavior of a system can be specified. This also creates the opportunity to use Metamorphic Testing (MT) which can be an effective means of safety validation for software developed using AI/ML.

Introduction

Due to its remarkable ability in solving complex problems, M
adopted as a way forward for the development of safety critic
critical industries like aerospace, automotive, or healthcare. As
precisely specified and systematically tested. However, there
specifying functional safety requirements for such systems. Firs
to handle complex tasks for which it can be very challenging to de
of the system should be (Heyn, et al., 2021) (M. Rahimi, 2019
many correct or nearly correct behaviours, and determining whic
trivial task. Secondly, even if it would be possible to identify t
scenario, AI/ML systems are often used to perform tasks with
variability input space. Because the behaviour of AI/ML systems
are made to the input data or the conditions for the scenarios fo
precisely specifying the exact behaviour of these systems would b
Finally, while in traditional software every line of source code ca
requirement, the AI/ML system code/model is obtained automati

L. Millet, S. Diemert, R. Debouk, R. S, M. Delgado, and J. Joyce, "Specifying Functional Safety Requirements for AI/ML in terms of Metamorphic Relations," presented at the International System Safety Conference, Minneapolis, MN, United states, Aug. 2024.



Metamorphic Testing: Example Relation

We have identified 15 MRs for our study.

MR act as a type of functional requirement for an ML model.

Research Question: Do these MRs discriminate between models of different quality?

MR #12:

GIVEN that the model correctly detects a nominal event (no anomaly present), — Base Case

APPLYING gaussian noise with standard deviation 1 is to Input A, — Transformation

RESULTS IN the model detecting an anomaly. — Expected Result

Metamorphic Testing: Preliminary Results

Metamorphic Relation	**Number of Failing Tests** (by highest-quality models)
Given that the model correctly detects a nominal event, **if the time profile is shifted by +/-1s**, the model should still detect it as nominal.	30 / 150
Given that the model correctly detects a nominal event, **if the Input B profile is shifted by +/-0.1 units**, the model should still detect this change as nominal.	30 / 150
Given that the model correctly detects a nominal event**, if the gaussian noise with standard deviation 0.25 is added to Input A**, the model should still detect no anomaly.	30 / 150
Given that the model correctly detects a nominal event, **if gaussian noise with standard deviation 0.25 is added to Input B**, the model should still detect no anomaly.	30 / 150
Given that no anomaly is detected and there are fewer than 2.5% outliers, **if 5% random outliers are added to Inputs A and B**, the model should detect an anomaly.	30 / 150

Preliminary Observation: MRs can reveal inconsistencies between an engineer's expectations of a model and the model's actual behaviour.

Validation Method: Combinatorial Testing

- Approach developed by researchers U.S. NIST

- Generate a set of test cases that contain a t-way tuples of the input set.
  - For conventional software, $t \leq 6$ is sufficient to find *most* (~99%) of software defects [1].

- **Research Question:** Does CT discriminate between ML models of varying quality.
  - Experiment in progress!

Given $A, B, C \in \{0,1\}$ for $t = 2$:

A	B	C
0	0	1
0	1	0
1	1	1
1	0	0

[1] R. Kuhn, R. Kacker, and Y. Lei, “Practical Combinatorial Testing,” National Institute of Standards and Technology, NIST Special Publication (SP) 800-142, Oct. 2010.

Combinatorial Testing: Preliminary Results

Preliminary results averaged across high-quality models, on a subset of the identified operating domain.

t value	# Test Cases	# Passed Test Cases	Sensitivity	Specificity
t=1	21	21 (100%)	100%	0%
t=2	441	421.5 (95.6%)	95.7%	0%
t=3	1384	1334.9 (96.5%)	96.5%	50.0%
t=4	3970	3834.4 (96.6%)	96.7%	50.0%

Preliminary Observation: Higher combinatorial strengths begin to reveal model insufficiencies. On-going experiments are exploring this question more deeply.

Summary and Takeaways

The results of this project will enable MDA Space to:

1. Use **safety case arguments** as a key tool for certifying AI-enabled safety-critical systems, driving more reliable and consistent operations

2. Incorporate **human factors** as part of the safety argument to better integrate AI with operators, handle greater data volumes, and enhance preventative maintenance to reduce downtime and costs

3. Apply the safety argument to **organize and motivate** verification methods (e.g., metamorphic and combinatorial testing), creating a **scalable framework** to extend these benefits across future missions and platforms

MDA SPACE

Critical Systems Labs
Innovating Safely

Emmanuel Lesser
Software S&MA Lead
emmanuel.lesser@mda.space

Jeff Joyce
CEO / President
jeff.joyce@cslabs.com