Are we ready for the first EASA guidance on the use of ML in Aviation?

Dr. Guillaume Brat

guillaume.p.brat@nasa.gov

NASA Ames Research Center
Intelligent Systems Division
Robust Software Engineering

Talk organization

- Follow the structure of the EASA document “EASA Concept Paper: First usable guidance for Level 1 machine learning applications”
  - Identify what, in the NASA research on “Complex Autonomous Systems Assurance” is applicable to the guideline
- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

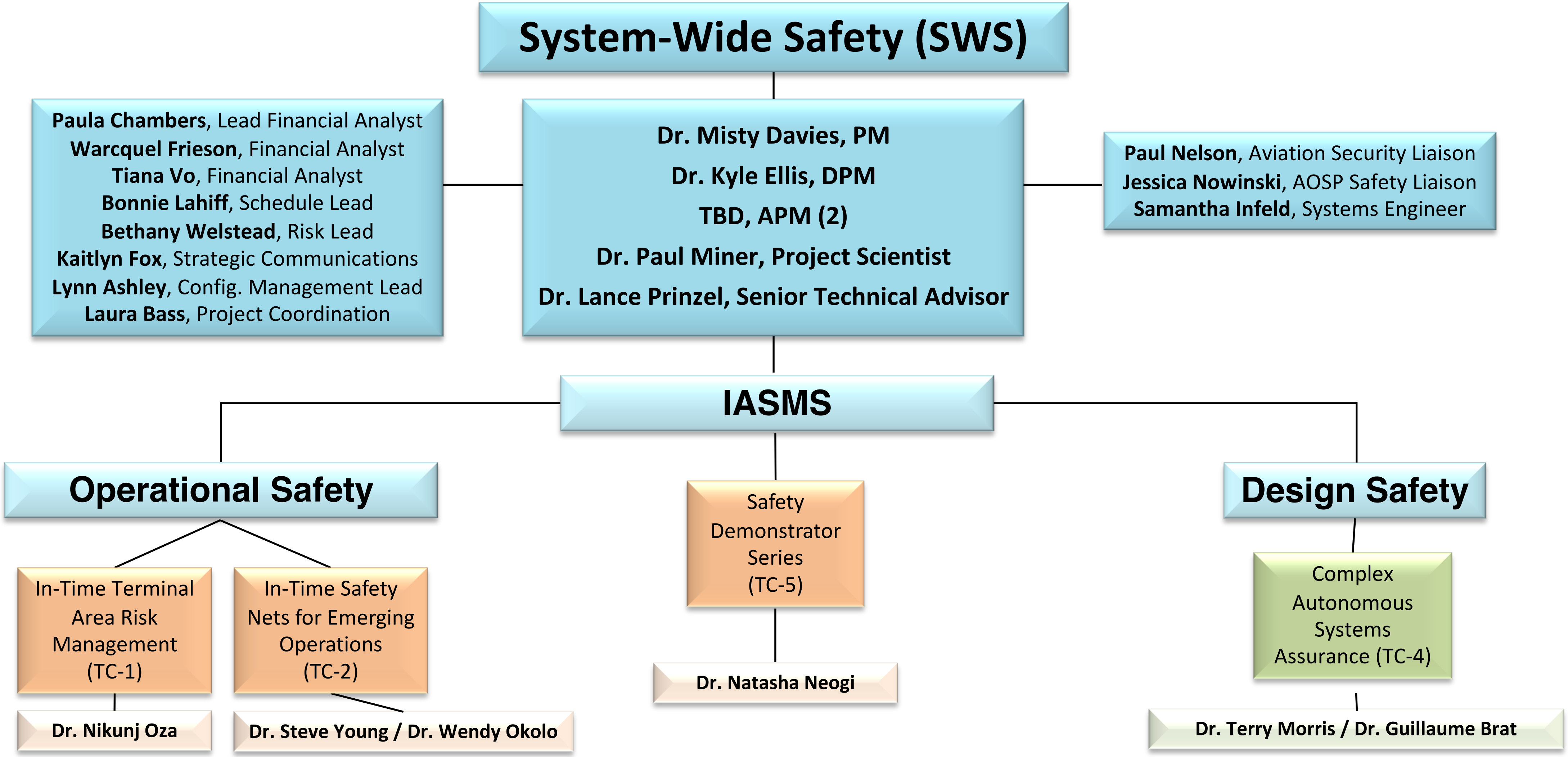
Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

Which NASA are we talking about?

- NASA: National Aeronautics and Space Administration
- ARMD: Aeronautics Research & Mission Directorate
- AOSP: Airspace Operations & Safety Program
- SWS: System-Wide Safety Project

The SWS Project



System-Wide Safety (SWS)

Dr. Misty Davies, PM
Dr. Kyle Ellis, DPM
TBD, APM (2)
Dr. Paul Miner, Project Scientist
Dr. Lance Prinzel, Senior Technical Advisor

Paula Chambers, Lead Financial Analyst
Warcquel Frieson, Financial Analyst
Tiana Vo, Financial Analyst
Bonnie Lahiff, Schedule Lead
Bethany Welstead, Risk Lead
Kaitlyn Fox, Strategic Communications
Lynn Ashley, Config. Management Lead
Laura Bass, Project Coordination

Paul Nelson, Aviation Security Liaison
Jessica Nowinski, AOSP Safety Liaison
Samantha Infeld, Systems Engineer

IASMS

Operational Safety

In-Time Terminal Area Risk Management (TC-1)

Dr. Nikunj Oza

In-Time Safety Nets for Emerging Operations (TC-2)

Dr. Steve Young / Dr. Wendy Okolo

Safety Demonstrator Series (TC-5)

Dr. Natasha Neogi

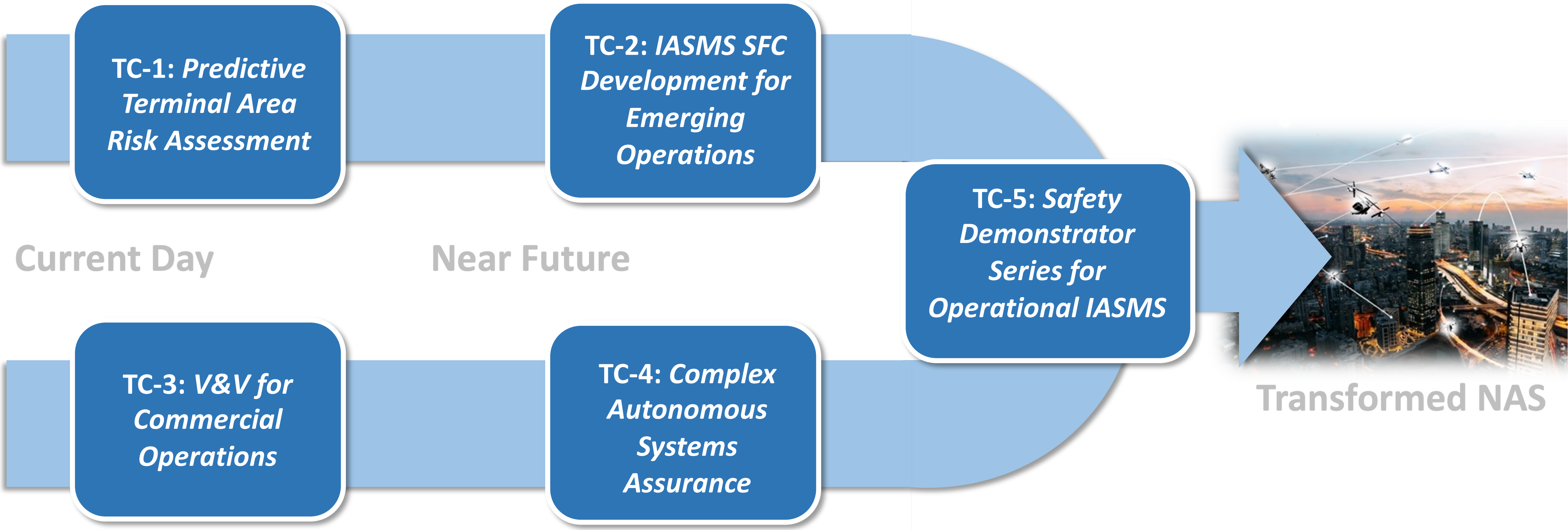
Design Safety

Complex Autonomous Systems Assurance (TC-4)

Dr. Terry Morris / Dr. Guillaume Brat

SWS Research Portfolio

Operational Safety (Thrust 5)



TC-1: *Predictive Terminal Area Risk Assessment*

TC-2: *IASMS SFC Development for Emerging Operations*

Current Day

Near Future

TC-5: *Safety Demonstrator Series for Operational IASMS*

Transformed NAS

TC-3: *V&V for Commercial Operations*

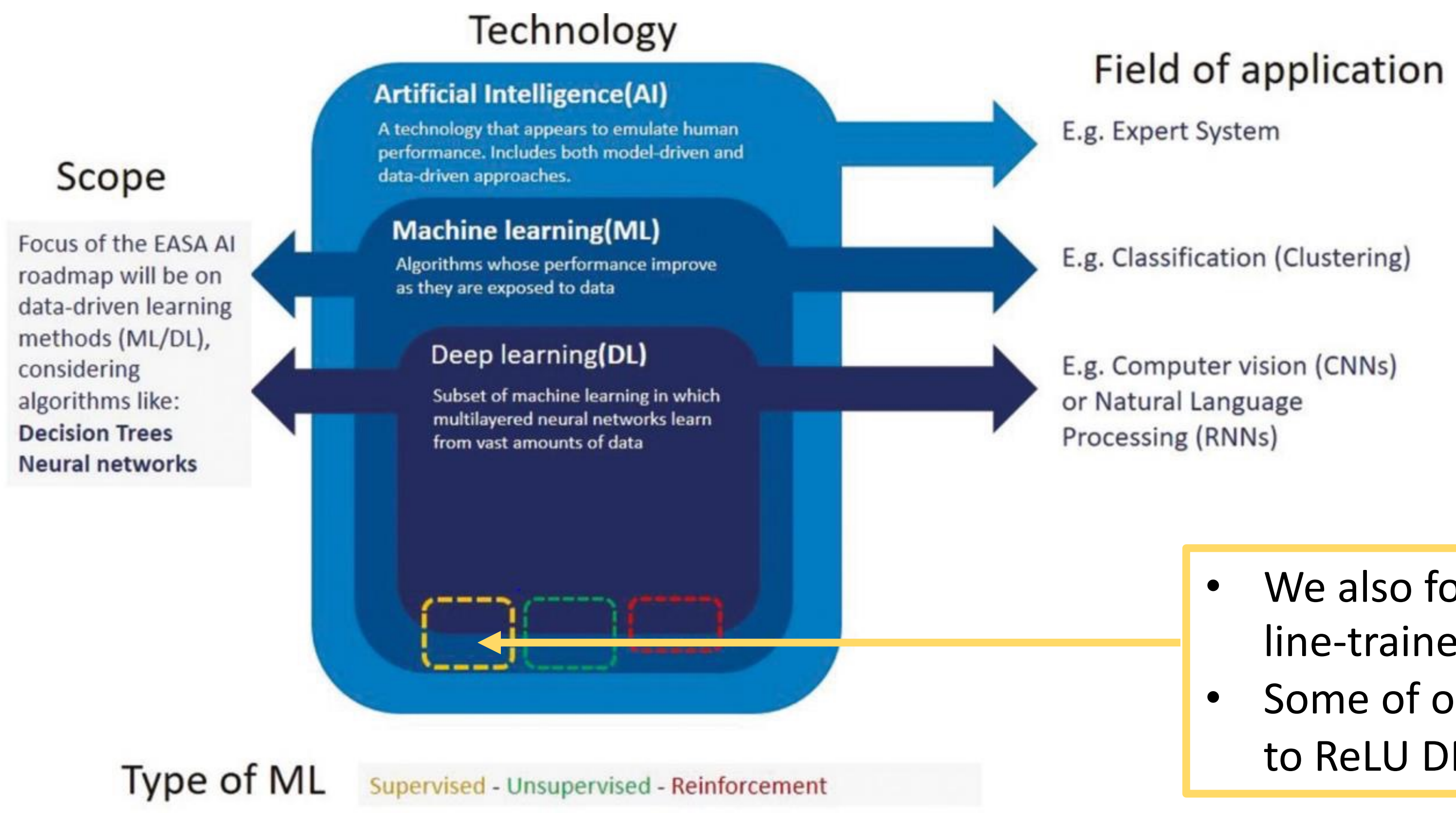
TC-4: *Complex Autonomous Systems Assurance*

Design Safety (Thrust 6)

Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

AI target: supervised, off-line learning, ML



Technology

Scope

Focus of the EASA AI roadmap will be on data-driven learning methods (ML/DL), considering algorithms like:
Decision Trees
Neural networks

Artificial Intelligence(AI)

A technology that appears to emulate human performance. Includes both model-driven and data-driven approaches.

Machine learning(ML)

Algorithms whose performance improve as they are exposed to data

Deep learning**(DL)**

Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Field of application

E.g. Expert System

E.g. Classification (Clustering)

E.g. Computer vision (CNNs) or Natural Language Processing (RNNs)

Type of ML Supervised - Unsupervised - Reinforcement

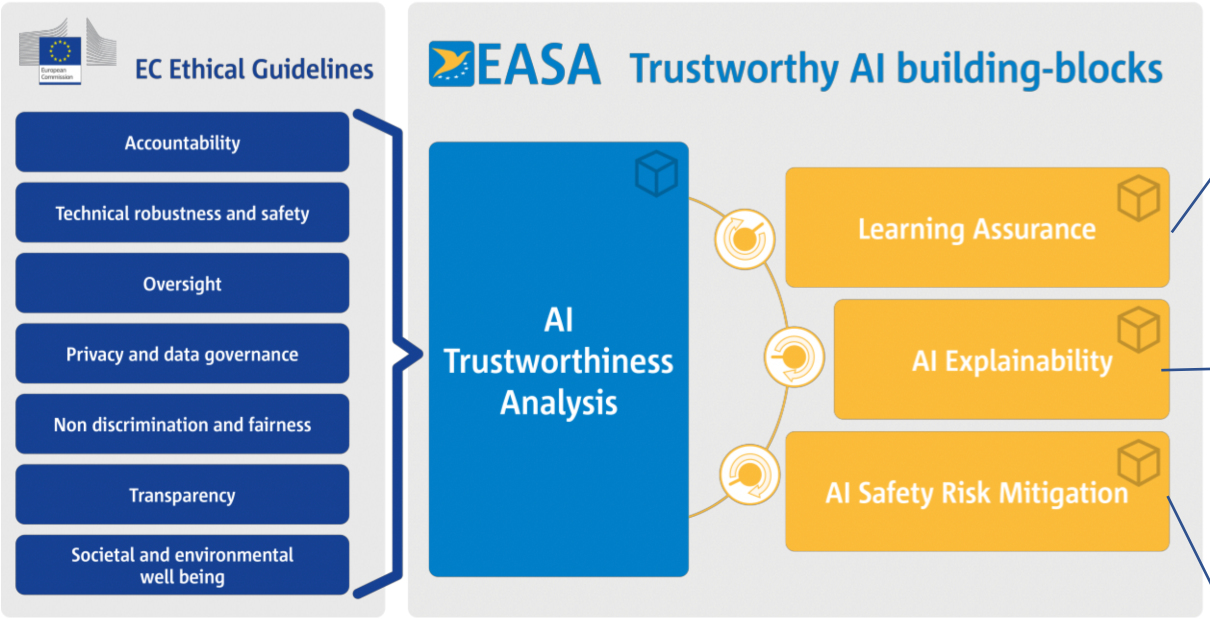
- We also focus mostly on off-line-trained, supervised DNNs
- Some of our results apply only to ReLU DNNs

EASA challenges

- Main challenges addressed through the first set of EASA guidelines:
  - Adaptation of assurance frameworks to cover learning processes and address development errors in AI/ML components
  - Creation of a framework for data management, to address the correctness (bias mitigation) and completeness/representativeness of data sets used for the ML items training and their verification
  - Addressing model bias and variance trade-off in the various steps of ML processes
  - Lack of predictability and explainability of ML/DL application behaviour, due to their statistical nature and to model complexity
  - Lack of guarantee on robustness and on absence of ‘unintended function’ in ML/DL applications
  - Mitigation of residual risk in ‘AI black box’. The expression ‘black box’ is a typical criticism oriented at AI/ML techniques, as the complexity and nature of AI/ML models bring a level of opaqueness that make them look like unverifiable black boxes (unlike rule-based software)
  - Lack of trust by end users/operators.

2.

EASA four basic blocks


European Commission
EC Ethical Guidelines
Accountability
Technical robustness and safety
Oversight
Privacy and data governance
Non discrimination and fairness
Transparency
Societal and environmental well being
EASA
Trustworthy AI building-blocks
AI Trustworthiness Analysis
Learning Assurance
AI Explainability
AI Safety Risk Mitigation

paradigm shift from programming to learning, as existing development assurance methods are not adapted to cover learning processes specific to AI/ML.

capability to provide the human with understandable and relevant information on how an AI/ML application is coming to its results

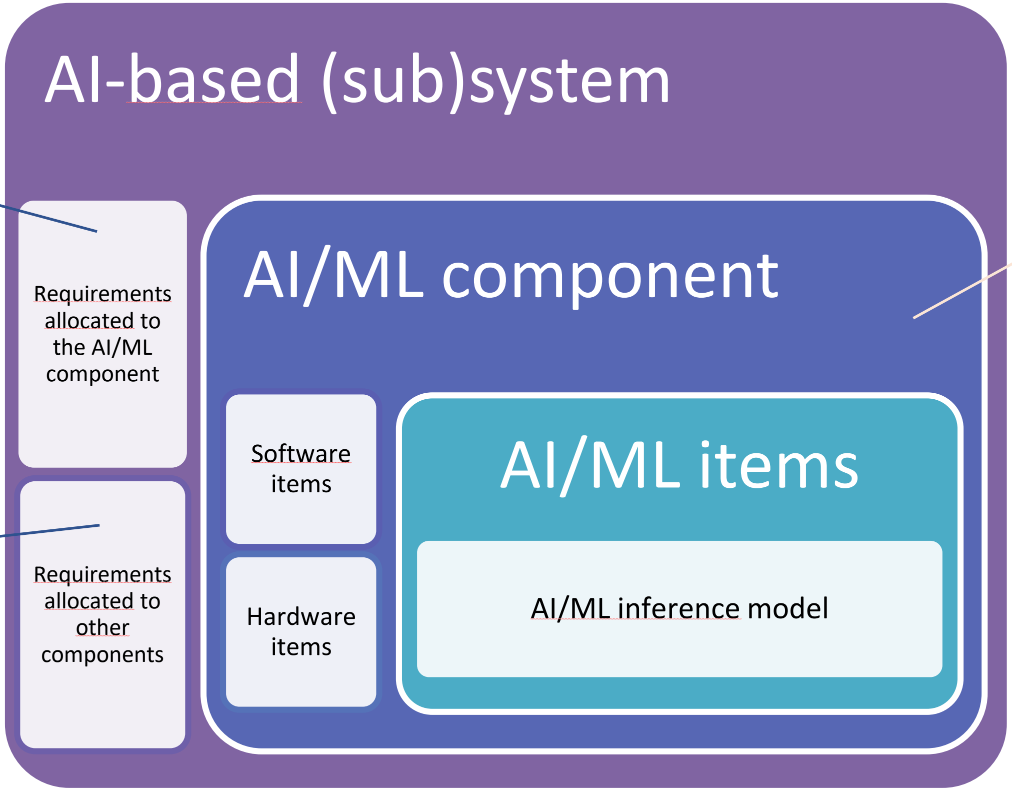
may not always be able to open the ‘AI black box’ to the extent required and that the safety risk may need to be addressed to deal with the inherent uncertainty of AI.

3.

AI-based sub-system decomposition

Future extensions:
- FRET: Elicitation and formalization of requirements for ML
- Probabilistic requirements?

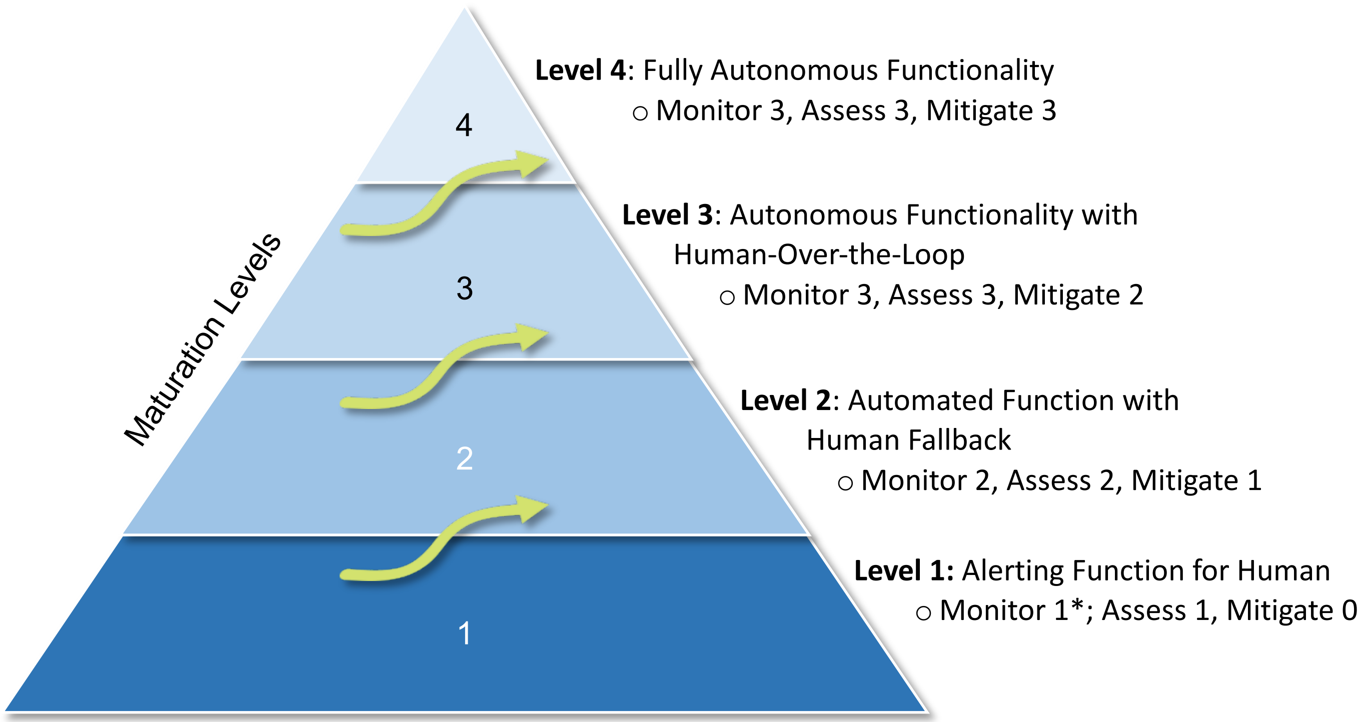
Past and current work:
- FRET: Elicitation and formalization of requirements



Current focus of NASA research, Except for hardware items

Criticality of AI applications

SFC Maturity
Services – Functions – Capabilities



Level 4: Fully Autonomous Functionality
- ○ Monitor 3, Assess 3, Mitigate 3

Level 3: Autonomous Functionality with Human-Over-the-Loop
- ○ Monitor 3, Assess 3, Mitigate 2

Level 2: Automated Function with Human Fallback
- ○ Monitor 2, Assess 2, Mitigate 1

Level 1: Alerting Function for Human
- ○ Monitor 1*; Assess 1, Mitigate 0

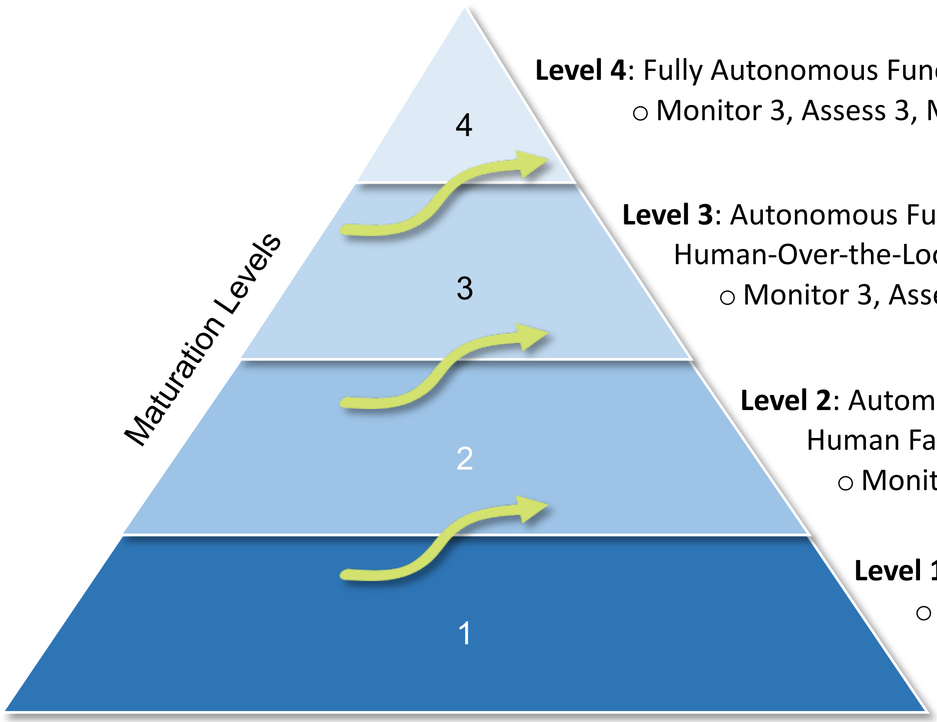
*The Monitor-Assess-Mitigate numbers signify increases in capability

EASA AI Roadmap AI Level	High level function/task allocated to the (sub)systems
Level 1A Human augmentation	Automation support to information acquisition
	Automation support to information analysis
Level 1B Human assistance	Automation support to decision-making
Level 2 Human-AI collaboration	Overseen automatic decision-making
	Overseen automatic action implementation
Level 3A More autonomous AI	Overridable automatic decision-making
	Overridable automatic action implementation
Level 3B Autonomous AI	Non-overridable automatic decision-making
	Non-overridable automatic action implementation

Table 1 — EASA AI typology and definitions

Criticality of AI applications

SFC Maturity
Services – Functions – Capabilities



Level 4: Fully Autonomous Functionality
○ Monitor 3, Assess 3, Mitigate 3

Level 3: Autonomous Functionality with Human-Over-the-Loop
○ Monitor 3, Assess 3, Mitigate 2

Level 2: Automated Function with Human Fallback
○ Monitor 2, Assess 2, Mitigate 1

Level 1: Alerting Function for Human
○ Monitor 1*; Assess 1, Mitigate 0

*The Monitor-Assess-Mitigate numbers signify increases in capability

SWS planned Safety Demonstrators (SD)

Safety Demos	ConOps	Max SFC Risk Level
SD-1	Wildfire fighting	2
SD-2	Post-Hurricane Disaster Relief	3
SD-3	Medical Courier Delivery	4
SD-4	Un-evacuated Urban Area Disaster Response	4

Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

Trustworthiness Analysis

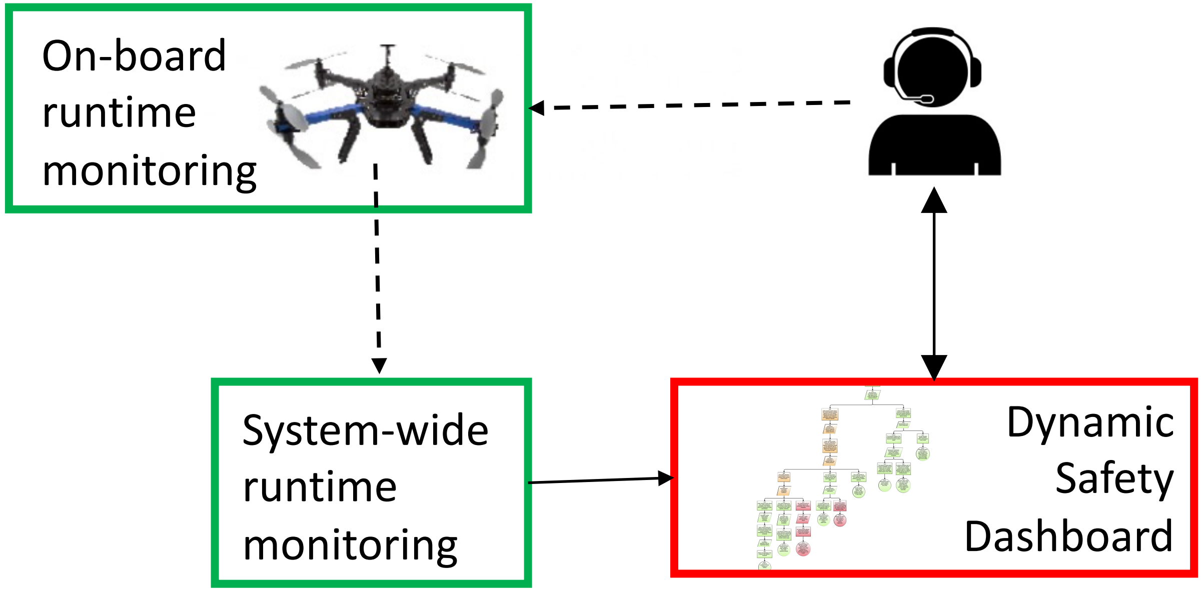
EASA guidance	NASA-related work: SMAR-STEReO
Objective CO-01: The applicant should identify the high-level function(s)/task(s) to be performed by the (sub)system either in interaction with the human or in autonomy.	ConOps for SMARt-STEReO define the functions needed for wildfire response, whether those are fulfilled by humans or autonomy, and do a preliminary trade study to determine usefulness of UAS, UTM, and NASA tech such as ICAROUS, Safeguard, and Safe2Ditch in this operational environment
Objective CO-03: The applicant should define and document the ConOps for all AI-based (sub)systems. A focus should be put on the definition of the operational design domain (ODD) and on the capture of specific operational limitations and assumptions.	
Objective CO-04: The applicant should perform a functional analysis of the (sub)system.	Use of IDEFo diagrams (functional models with an operational emphasis) in SMARt-STEReO
Objective CO-02: The applicant should define the AI-based (sub)system taking into account domain-specific definitions of 'system'.	

Safety Assessment Concept

- Initial safety assessment, during design phase by considering the contribution of an AI/ML component to system failure and by having particular architectural considerations when AI is introduced;

- Function-based risk and resiliency modelling and analysis using Fmdtools for ML-human systems
- Wrap SMARt-STEReO simulation in AdaStress framework to find rare catastrophic events (MOC-SA-02-1)
  - Used for AI systems as well as complex systems with emerging behaviors.
  - Using NLP for taxonomy construction for emerging systems

- Continuous safety assessment, with the implementation of a data-driven AI safety risk assessment based on operational data and occurrences. This 'continuous' analysis of in-service events may rely on processes already existing for domains considered in this guideline. The processes will need to be adapted to the AI introduction.



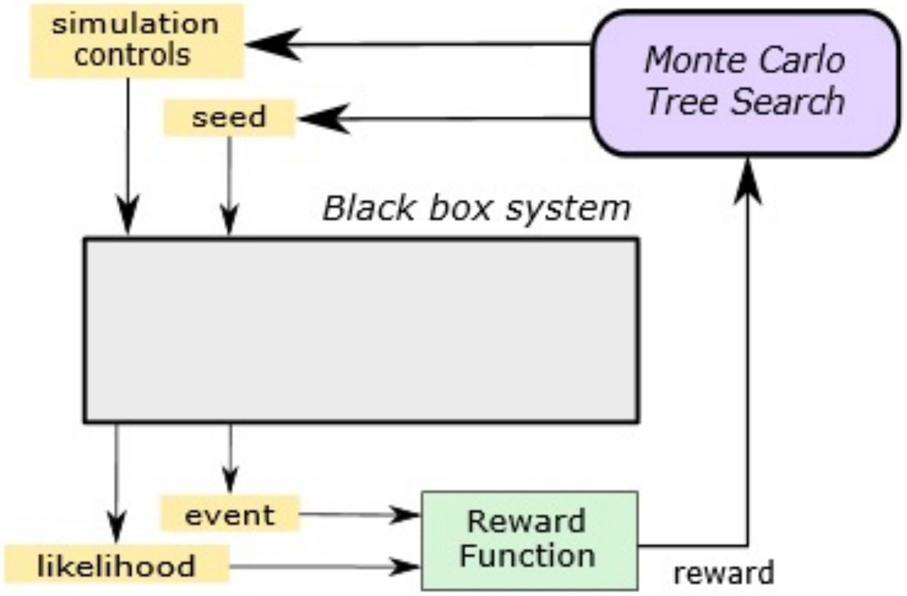
On-board runtime monitoring

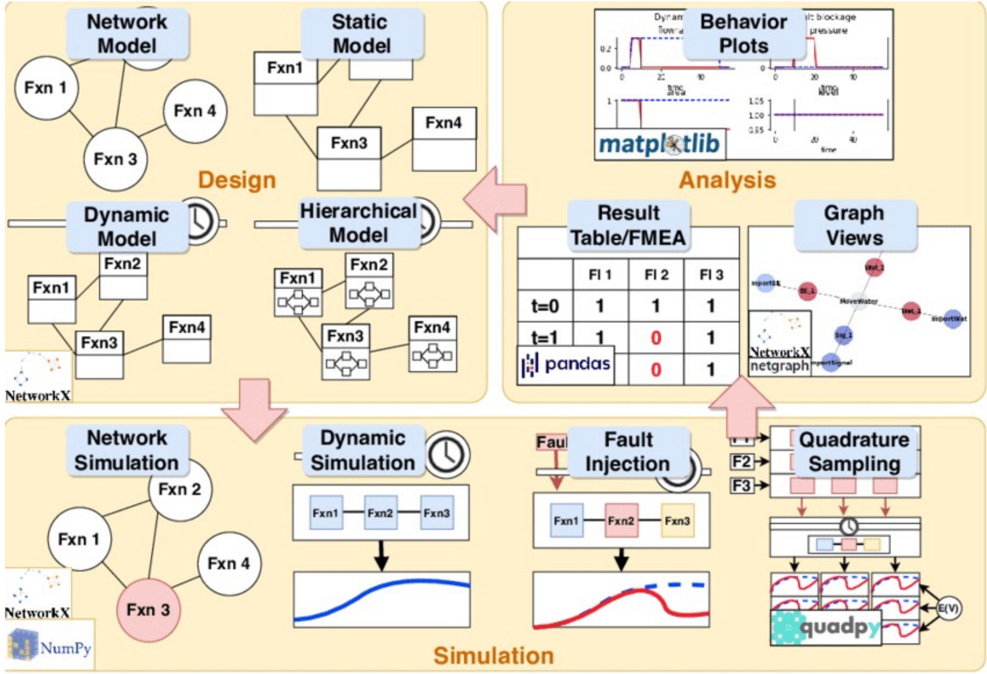
System-wide runtime monitoring

Dynamic Safety Dashboard

AdaStress and FMDToolsat a Glance

- AdaStress (Adaptive Stree testing) is a software package for an accelerated simulation-based stress testing method for finding the most likely path to a failure event





	FI 1	FI 2	FI 3
t=0	1	1	1
t=1	1	0	1
		0	1

- Fmdtools (Fault Model Design *tools*) a design and analysis environment, which enables a designer to represent the system in the early design process, simulate the effects of faults, and quantify corresponding resilience metrics.

Impact Assessment of AI on System Safety

- Perform functional hazard assessment in the context of the ConOps
- — Perform safety assessment
  - …
  - Define AI/ML component performance/reliability metrics
  - Analyse and mitigate the effect of AI/ML component exposure to input data outside of the ODD
  - Perform AI/ML component failure mode effect analysis
- — Verification — final safety assessment
- — Consolidate the safety assessment to verify the safety objective

- Early work on a notion of coverage for NNs

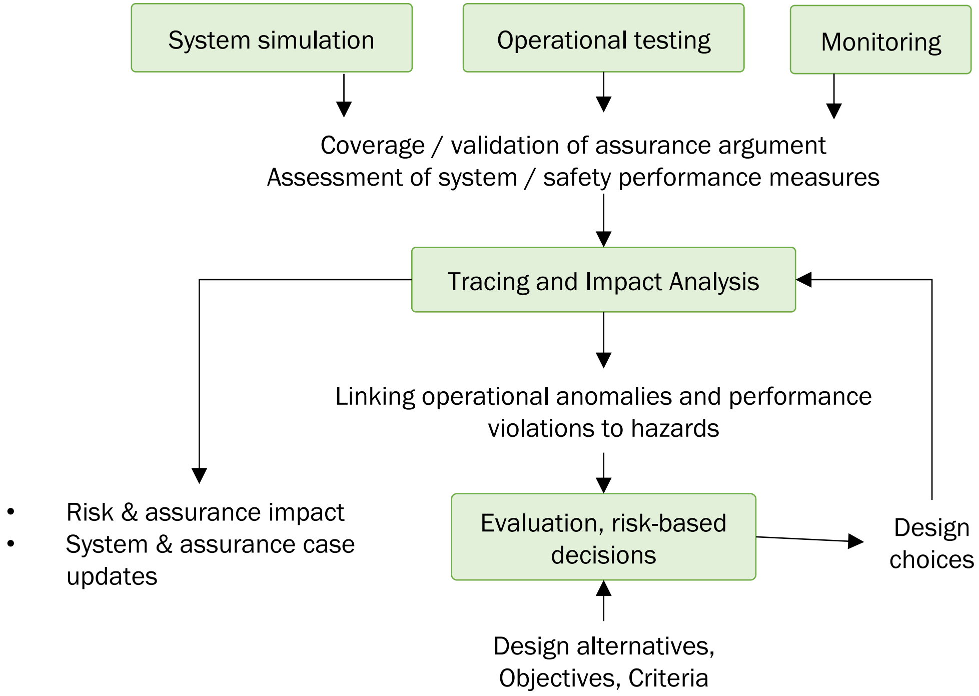
- Use of statistical testing techniques (MARGInS, SysAI) or formal analysis of NNs (Prophecy)

- No work yet but identified as a need:
  - What characterize ML failures?
  - How do we represent faults in ML?
- That would feed into many tools, e.g., fmdtools so that true impact of ML failures on system can be analyzed

Change assessment support

Objective SA-04: The applicant should perform a safety support assessment for any change in the functional (sub)system embedding a component developed using AI/ML techniques or incorporating AI/ML algorithms.

- AdvocATE (tool for creating and visualizing safety cases)
  - Integration of risk-based design, development, and assurance in AdvoCATE
- Demonstrated in March 2021 on a centerline tracking example implemented with DNNs


System simulation
Operational testing
Monitoring
Coverage / validation of assurance argument
Assessment of system / safety performance measures
Tracing and Impact Analysis
Linking operational anomalies and performance violations to hazards
• Risk & assurance impact
• System & assurance case updates
Evaluation, risk-based decisions
Design choices
Design alternatives, Objectives, Criteria

Summary for Safety Assessment

EASA Guidance	NASA related work
Objective SA-01: The applicant should define metrics to evaluate the AI/ML component performance and reliability.	• SMARt-STEREo is aiming at defining metrics for resiliency in complex systems, which include systems using ML components.
Objective SA-02: The applicant should perform a system safety assessment for all AI-based (sub)systems.	• Use of NLP to extract failure taxonomy for complex systems • currently based on safety reports, which we don't have for AI-based (sub)systems.
Objective SA-03: The applicant should define metrics to evaluate the AI/ML component performance.	*Already covered by SA-01?*
Objective SA-04: The applicant should perform a safety support assessment for any change in the functional (sub)system embedding a component developed using AI/ML techniques or incorporating AI/ML algorithms.	• Supporting capabilities in the AdvoCATE tool (creation of safety case) and its risk-based design decision capabilities. • Complementary capabilities in fmdtools.

Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organisations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

Requirements and Architecture Management

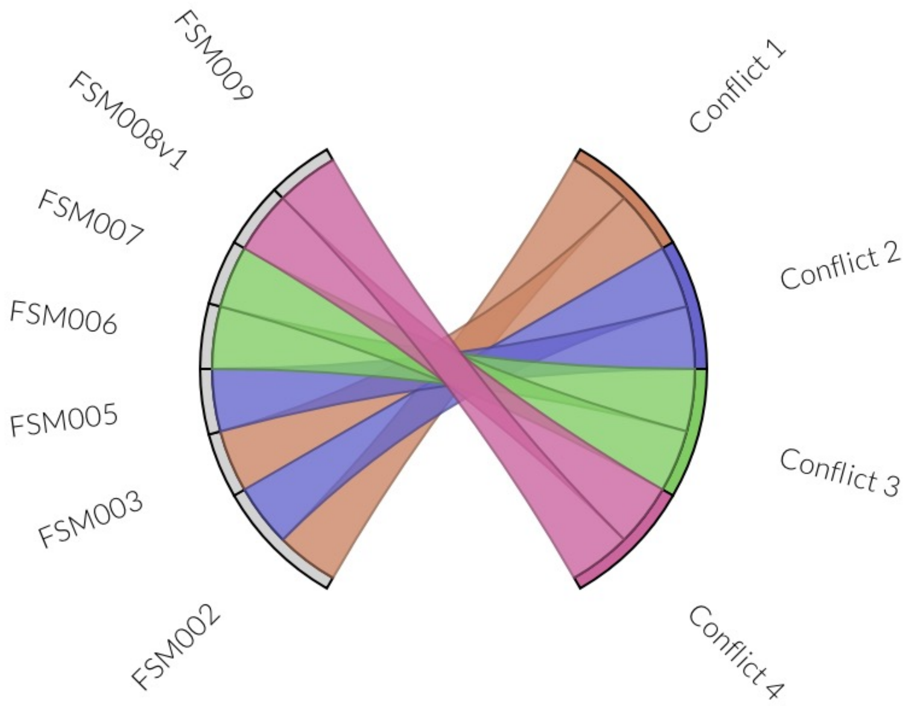
Objective DA-02: (Sub)systems requirements documents should be prepared and encompass the capture of the following minimum requirements:

EASA Guidance	NASA: Formalization using FRET
safety requirements allocated to the AI/ML component	Require extension to probabilistic requirements
information security requirements allocated to the AI/ML component	Informal capture
functional requirements allocated to the AI/ML component	Require extension to probabilistic requirements
operational requirements allocated to the AI/ML component, including ODD monitoring requirements	Require extension to probabilistic requirements
non-functional requirements allocated to the AI/ML component (e.g. performance, scalability, reliability, resilience, etc.)	Informal capture mostly

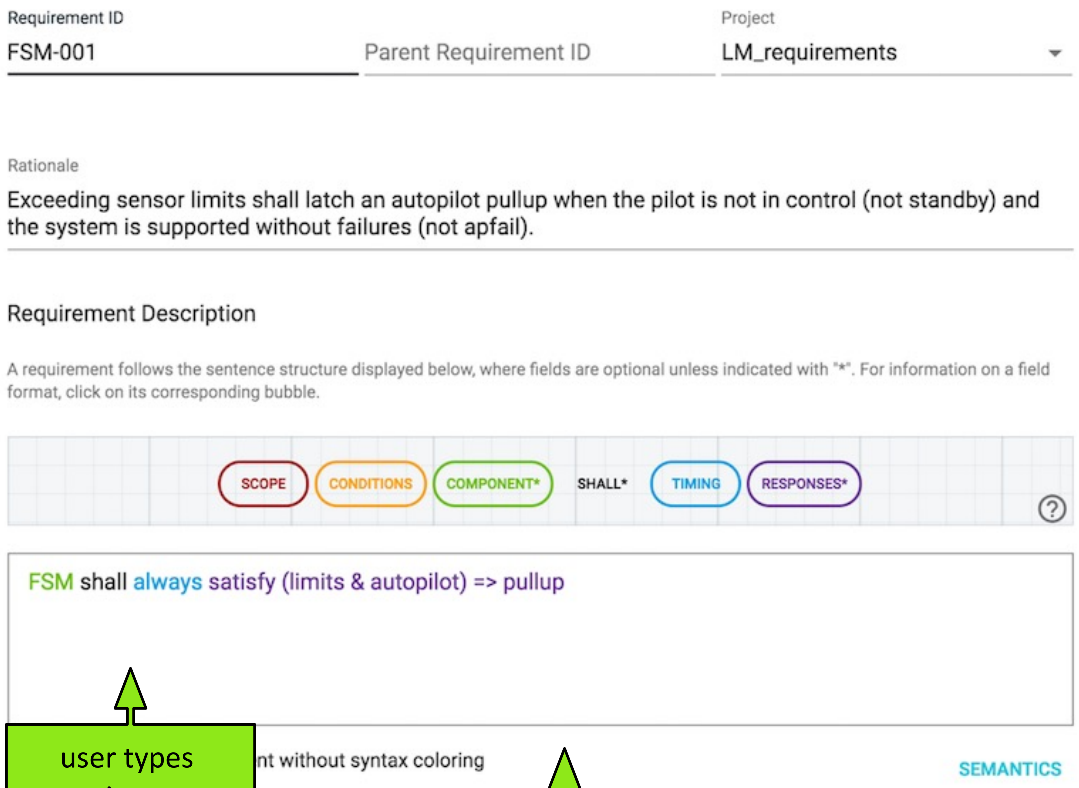
Requirements and Architecture Management

Objective DA-04: Each of the captured requirements should be validated**.**

- If the requirements for an ML component can be formalized, FRET has the capability of checking their realizability, i.e.,
  - Realizability checking aims at determining whether an implementation exists, always complying with a set of requirements, regardless of the stimuli provided by the system's environment.
- Current formalization is limited to future and past linear temporal logic.
  - Can it be extended to probabilistic requirements?



FRET at a Glance



Requirement ID
FSM-001

Parent Requirement ID

Project
LM_requirements

Rationale
Exceeding sensor limits shall latch an autopilot pullup when the pilot is not in control (not standby) and the system is supported without failures (not apfail).

Requirement Description

A requirement follows the sentence structure displayed below, where fields are optional unless indicated with "*". For information on a field format, click on its corresponding bubble.

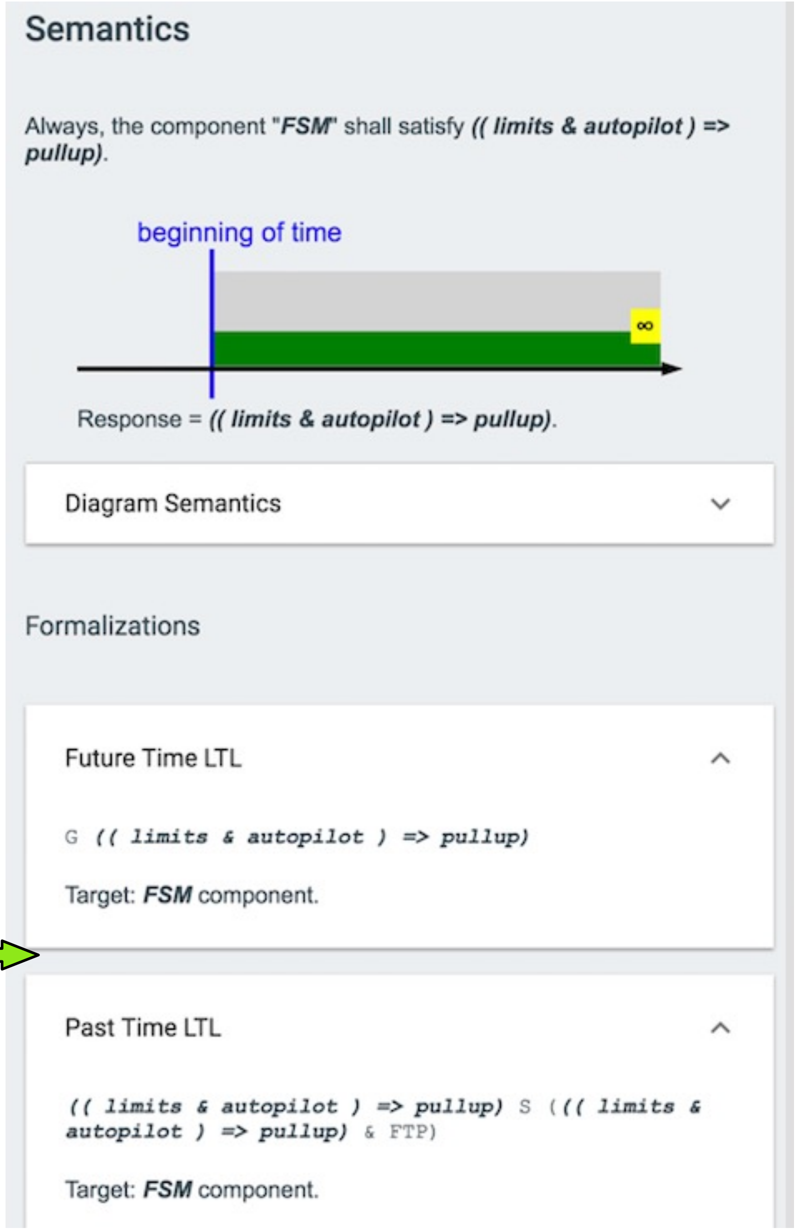
FSM shall always satisfy (limits & autopilot) => pullup

nt without syntax coloring

user types requirement

parser recognizes fields and color codes them dynamically

FRET generates formal semantics



Semantics

Always, the component "***FSM***" shall satisfy ***((limits & autopilot) => pullup)***.

Response = ***((limits & autopilot) => pullup)***.

Formalizations

Future Time LTL

```
G (( limits & autopilot ) => pullup)
```

Target: ***FSM*** component.

Past Time LTL

```
(( limits & autopilot ) => pullup) S ((( limits & autopilot ) => pullup) & FTP)
```

Target: ***FSM*** component.

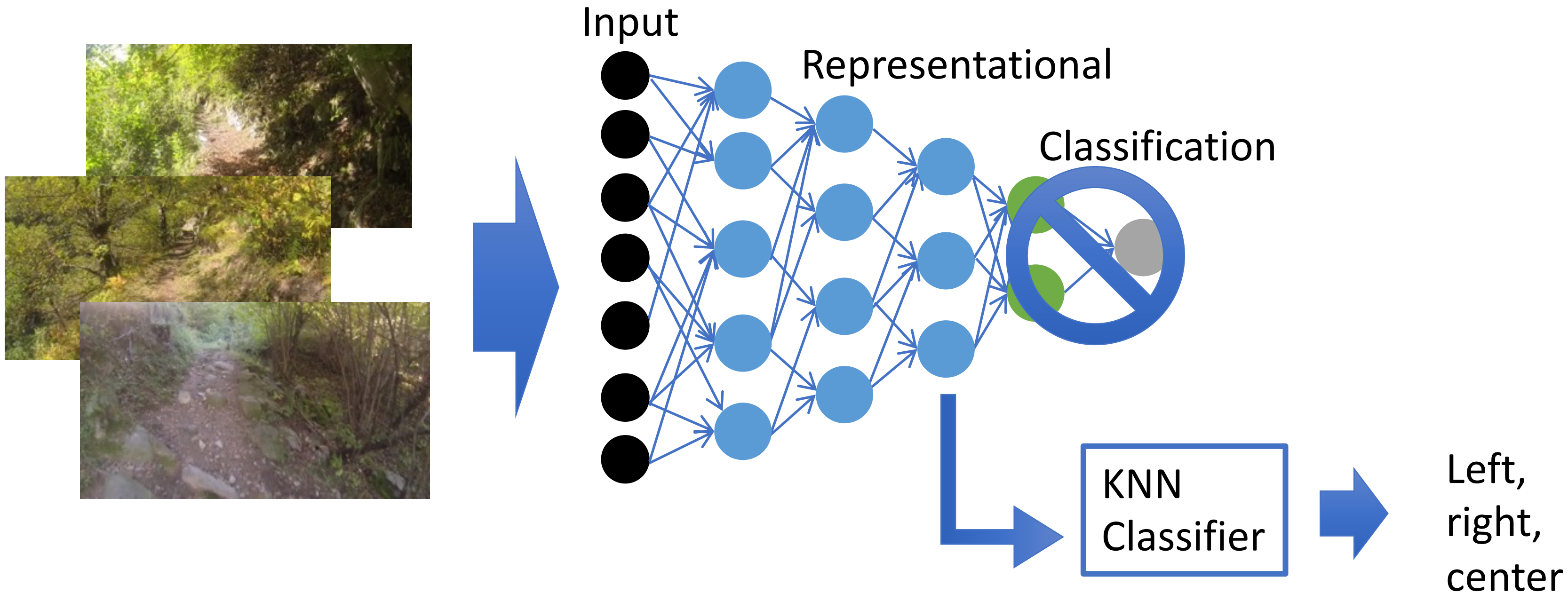
DA Objectives Summary

EASA Guidance	NASA related work
Objective DA-01: The applicant should describe the proposed learning assurance process, taking into account each of the steps described in Sections C.3.3 to C.3.12, as well as the interface and compatibility with development assurance processes.	
Objective DA-02: (Sub)systems requirements documents should be prepared and encompass the capture of the following minimum requirements: …	Require FRET extension for probabilistic analysis
Objective DA-03: The applicant should describe the system and subsystem architecture, to serve as reference for related safety (support) assessment and learning assurance objectives.	
Objective DA-04: Each of the captured requirements should be validated.	FRET realizability analysis
DA-05: Each of the captured (sub)system requirements should be verified.	Traceability: Drishti tool can suggest links using NLP technology.

Data Management

Objective DM-10: The applicant should ensure validation and verification of the data all along the data management process so that the DQRs are addressed.

- Anticipated MOC DM-10-3: We have developed a hybrid network with k-nn (K-Nearest Neighbor), which can show problems with correctness of training data
  - SysAI can also be used to identify biases and variance.



Data Validation (DM-10)

Anticipated MOC DM-10-5: Data sets independence

- We have developed tools for dataset independence for image datasets coming from video streams
- In our domain our classification comes from a single image, yet our dataset comes from a time series (video).
  - Images taken from close in time are not statistically independent, which leads to test data being very similar to training data.
  - This leads to a large overconfidence in performance accuracy since the classifier has seen images very close to the test set images.
- Solution: temporally separate the training set from test set.
- We developed splicing algorithms where blocks of the video were put into training and other blocks were put into testing where the edges of the testing blocks were eliminated to ensure that no test image was temporally close to a training image.
- We can adjust the block size and block selection method to maximize the independence of the training and test sets, while keeping the statistics of the training and test sets as similar as possible.

DM Objectives summary

EASA Guidance	NASA
Objective DM-01: The applicant should capture the DQRs for all data pertaining to the data management process, including but not limited to: …	FRET
Objective DM-02: The applicant should capture the requirements on data to be pre-processed and engineered for the inference model.	FRET
Objective DM-03: To enable the data collection step, the applicant should identify explicitly and record the input space and the operating parameters that drive the selection of the training, validation and test data sets.	
Objective DM-04: Once data sources are collected, the applicant should make sure that the data set is correctly annotated or labelled.	
Objective DM-05: The applicant should make sure that the operations on data properly address the captured requirements (including DQRs).	

DM Objectives summary

EASA Guidance	NASA
Objective DM-06: The applicant should ensure a sequence of operations on the collected data in preparation of the training.	
Objective DM-07: When applicable, the applicant should transform the pre-processed data from the input space into features which are effective for the performance of the selected ML algorithm.	
Objective DM-08: The applicant should ensure that the data is effective for the stability of the model and the convergence of the learning process, possibly via normalisation.	
Objective DM-09: The applicant should distribute the data into three separate and independent data sets which will meet the specified DQRs: …	
Objective DM-10: The applicant should ensure validation and verification of the data all along the data management process so that the DQRs are addressed.	• K-nn and dataset independence • SysAI to eliminate bias and variance in data.

Learning Process Management

Objective LM-04: The applicant should provide quantifiable generalisation guarantees.

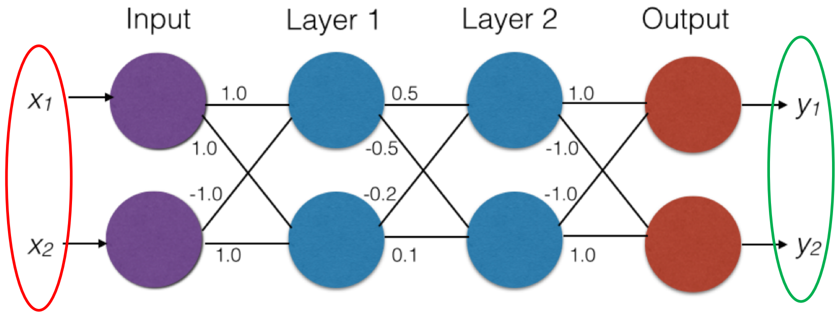
- The methods suggested in LM-04-1 take a traditional view of generalization that assumes that the test, training and deployment data are all sampled from the same data distribution
  - Likely the case in Example "3.1. AI-based augmented 4D trajectory prediction — climb and descent rates" where there is a large amount of similar telemetry data of real-world flights and future flights.
- Instead, we are examining generalization in domains where the use case cannot be assumed to be statistically identical to training and test data.
  - As in runway identification and runway foreign object debris detection
  - In our case study of path identification, images from deployment would likely be statistically different than images from the training set.
  - Assuming they were statistically the same would give high overconfidence in our generalization capabilities.
  - Instead we created a series of test sets that differ in how close they are to the original training set.
  - By looking at performance falloffs we can then see how much divergence from the training set is safe for a trained neural network.

Learning process verification

Objective LM-10: The applicant should perform a requirements-based verification of the trained model behaviour and document the coverage of the ML component requirements by verification methods.

- Prophecy Key Idea:
  - Infer “likely” properties, aka contracts, of a NN
  - Prove them using a decision procedure
- What is a contract? σ => P
  - σ is a precondition (“safe region”)
  - P is a postcondition; desired output behavior (e.g. some prediction)

If σ holds for

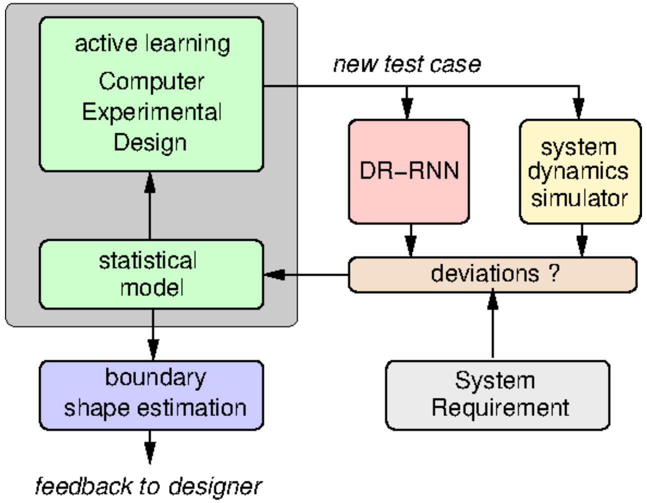


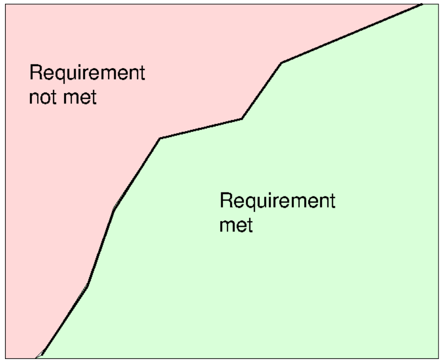
satisfies P

LM-11 and MOC LM-11-1

Objective LM-11: The applicant should provide an analysis on the robustness (or stability) of the algorithms and of the trained model.

- Perturbations in the operational phase due to fluctuations in the data input and prediction output or in the model itself.
  - Use of SysAI (new branchof MARGInS) to evaluate the Taxinet example with noisy data.
  - Perturbations in the design phase could also be found using SysAI.





- Other thoughts:
  - Combining Deep NN classifier, semantic segmentation, and object detection should allow an increase in robustness, and,
  - Use Prophecy (and formal methods) to generate adversarial inputs for robustness analysis and abnormal ranges.

LM Objectives summary

EASA Guidance	NASA
Objective LM-01: The applicant should describe the AI/ML components and model architecture (including computational graph and activation functions).	
Objective LM-02: The applicant should capture the requirements pertaining to the learning management and training processes, including but not limited to: …	FRET
Objective LM-03: The applicant should document the credit sought from the training environment and qualify the environment accordingly.	
Objective LM-04: The applicant should provide quantifiable generalisation guarantees.	Work for domains where the use case cannot be assumed to be statistically identical to training and test data
Objective LM-05: The applicant should document the result of the model training.	*Process*
Objective LM-06: The applicant should document any model optimisation that may affect the model behaviour (e.g. pruning, quantisation) and assess their impact on the model behaviour or performance.	SysAI has been used to analyze impact of architecture optimization on NN performance
Objective LM-07: The applicant should estimate bias and variance of the selected model family and should provide evidence of the reproducibility of the training process.	Analysis of learning biases and variances can be done with SysAI - SysAI is designed to find areas of high variance efficiently.

LM Objectives summary

EASA Guidance	NASA
Objective LM-08: The applicant should ensure that the estimated bias and variance of the selected model meet the associated learning process management requirements.	Analysis of learning biases and variances can be done with SysAI - SysAI is designed to find areas of high variance efficiently.
Objective LM-09: The applicant should perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification.	
Objective LM-10: The applicant should perform a requirements-based verification of the trained model behaviour and document the coverage of the ML component requirements by verification methods.	Prophecy: formal methods to compute formal preconditions for safety requirements on output of ReLU NNs.
Objective LM-11: The applicant should provide an analysis on the robustness (or stability) of the algorithms and of the trained model.	MARGInS / SysAI Prophecy for abnormal ranges and adversarial inputs
Objective LM-12: The applicant should verify the anticipated generalisation bounds using the test data set.	These bounds can be determined by SysAI. It can also do geometric shape detection that supports the documentation requirements.

IMP Objectives Summary

EASA Guidance	NASA related work
Objective IMP-01: The applicant should capture the requirements pertaining to the implementation process, including but not limited to: …	FRET
Objective IMP-02: Any post-training model transformation (conversion, optimisation, deployment) should be identified and validated for its impact on the model behaviour and performance.	• Prophecy: use formal contracts to perform model distillation • PRECISA: floating-point error static code analysis to find bounds on round-off errors.
Objective IMP-03: For each transformation step, the environment (i.e. software tools and hardware) necessary to perform model transformation should be identified and any associated assumptions or limitations captured and validated.	

Inference Model Verification

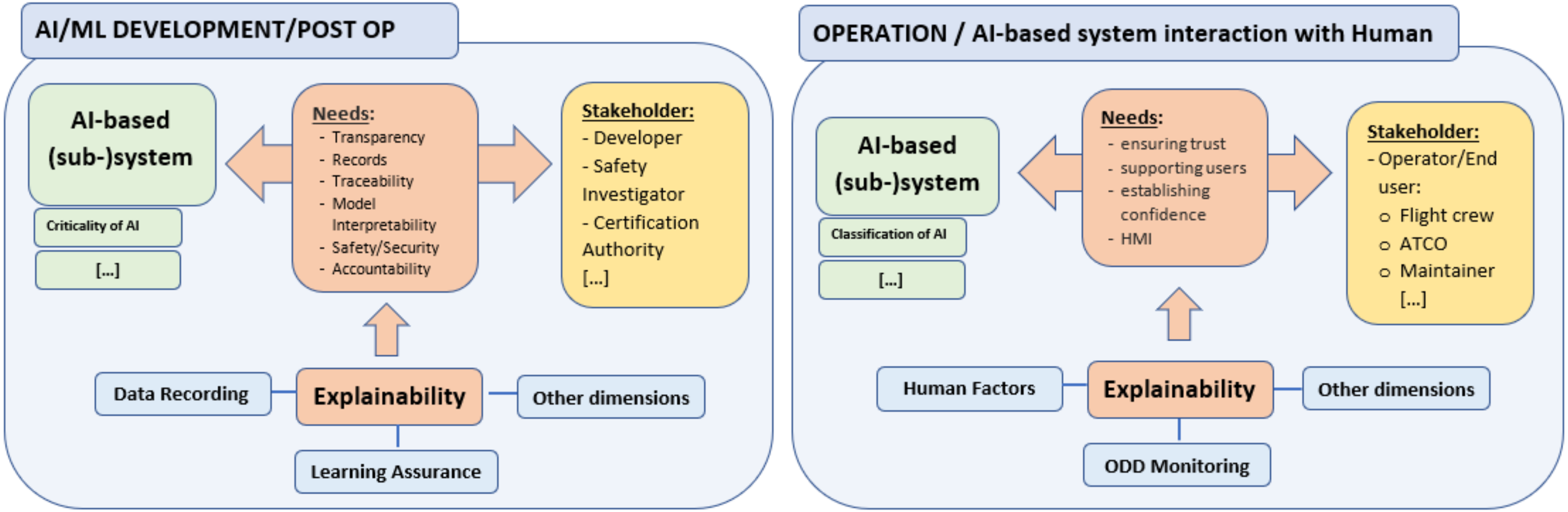
EASA Guidance	NASA related work
Objective IMP-04: The applicant should verify that any conversion, optimisation or transformation performed during the trained model implementation step have not altered the defined model properties.	Prophecy: use formal contracts to perform model distillation
Objective IMP-05: The differences between the hardware platform used for training and the one used for verification should be identified and assessed for impact on the inference model behaviour and performance.	
Objective IMP-06: The applicant should perform an evaluation of the performance of the inference model based on the test data set and document the result of the model verification.	
Objective IMP-07: The applicant should perform a requirements-based verification of the inference model behaviour and document the coverage of the ML component requirements by verification methods.	Prophecy
Objective IMP-08: The applicant should provide an analysis on the robustness (or stability) of the inference model.	MARGInS, SysAI

Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

AI Explainability

AI explainability: *Capability to provide the human with understandable and relevant information on how an AI/ML application is coming to its results.*



AI/ML DEVELOPMENT/POST OP

AI-based (sub-)system

Criticality of AI

[...]

Needs:
- Transparency
- Records
- Traceability
- Model Interpretability
- Safety/Security
- Accountability

Stakeholder:
- Developer
- Safety Investigator
- Certification Authority

[...]

Data Recording

Explainability

Other dimensions

Learning Assurance

OPERATION / AI-based system interaction with Human

AI-based (sub-)system

Classification of AI

[...]

Needs:
- ensuring trust
- supporting users
- establishing confidence
- HMI

Stakeholder:
- Operator/End user:
  - Flight crew
  - ATCO
  - Maintainer

  [...]

Human Factors

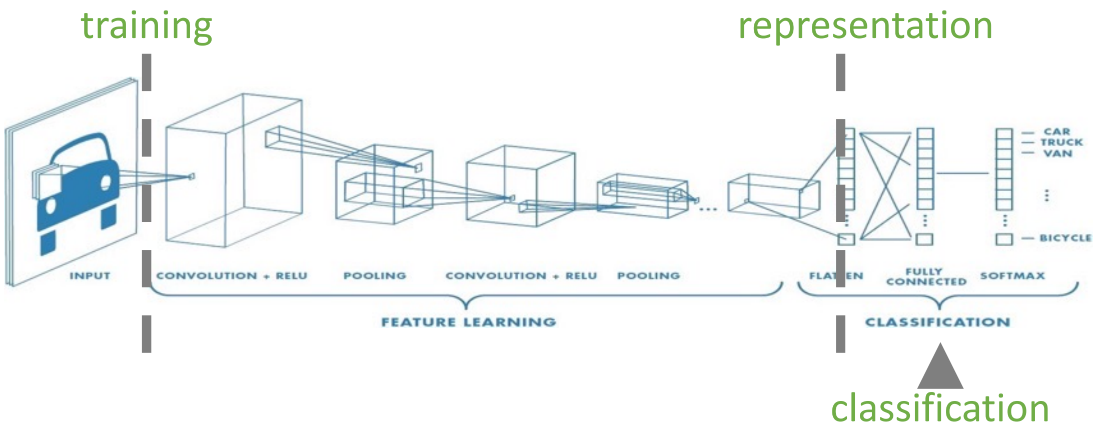
Explainability

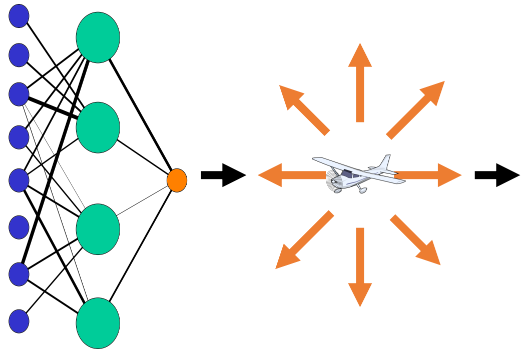
Other dimensions

ODD Monitoring

Post hoc/a posteriori/local explanation

- We have work on explaining a neural network controller in an object avoidance domain (post hoc explanations).
    - We were able to take the output of the controller and generate a series of explanations for observed behavior.
    - The most successful of these were
        - rule templates, and,
        - Integrated gradients: explain the relationship between a model's predictions in terms of its features
- Our work with LSTMs (long short-term memory designed for learning from sequential data) provided temporal post hoc explanations with heat maps.





```
if conflict to left
   go up with probability .19
else if conflicts are above
   go left with probability
.87
else
   go up with probability .07
…
...
...
```

Post hoc/a posteriori/local explanation

- Establishing causal relationships between the inputs and the outputs of the model
  - We use the Scenic framework to assess the performance of a perception module, identify correct and incorrect detections, and extract rules from those results that semantically characterizes the correct and incorrect scenarios.
- Finding the boundaries of the model and help in repair
  - In recent work, we demonstrate constraint solving repair for neural networks in the context of three different scenarios:
    - Improving the overall accuracy of a model,
    - Fixing security vulnerabilities caused by poisoning of training data and
    - Improving the robustness of the network against adversarial attacks.
- Highlighting undesirable bias (data sets and model bias)
  - Developed a probabilistic analysis technique for neural networks, which uses a form of concolic execution tailored to the topology of the network and uses solution space quantification over collected constraints, corresponding to neuron activations.
    - Our evaluation shows that this framework provides useful results for quantifying fairness and robustness, in the context of medium sized networks.

A priori/global explanation

- Hybrid deep network with K-NN classifier provided explanations of classification component of networks directly by showing training images used for voting of class label.
- Variational auto encoder provided explanation of latent variables (values of the variables inside of the network) by showing how the latent variables related to human recognizable properties of the image.
- Our work on semantic segmentation can be used to give explanation of which pixels of the image are being used to make classification decision.
- Object identification can be used to give explanation of what objects in the image are being used to make classification decision
- Establishing how confident a model is with its own decisions
    - Working with Boeing on a Prophecy runtime module, which, for every input, will assign a metric of confidence to the outputs of the Taxinet model.

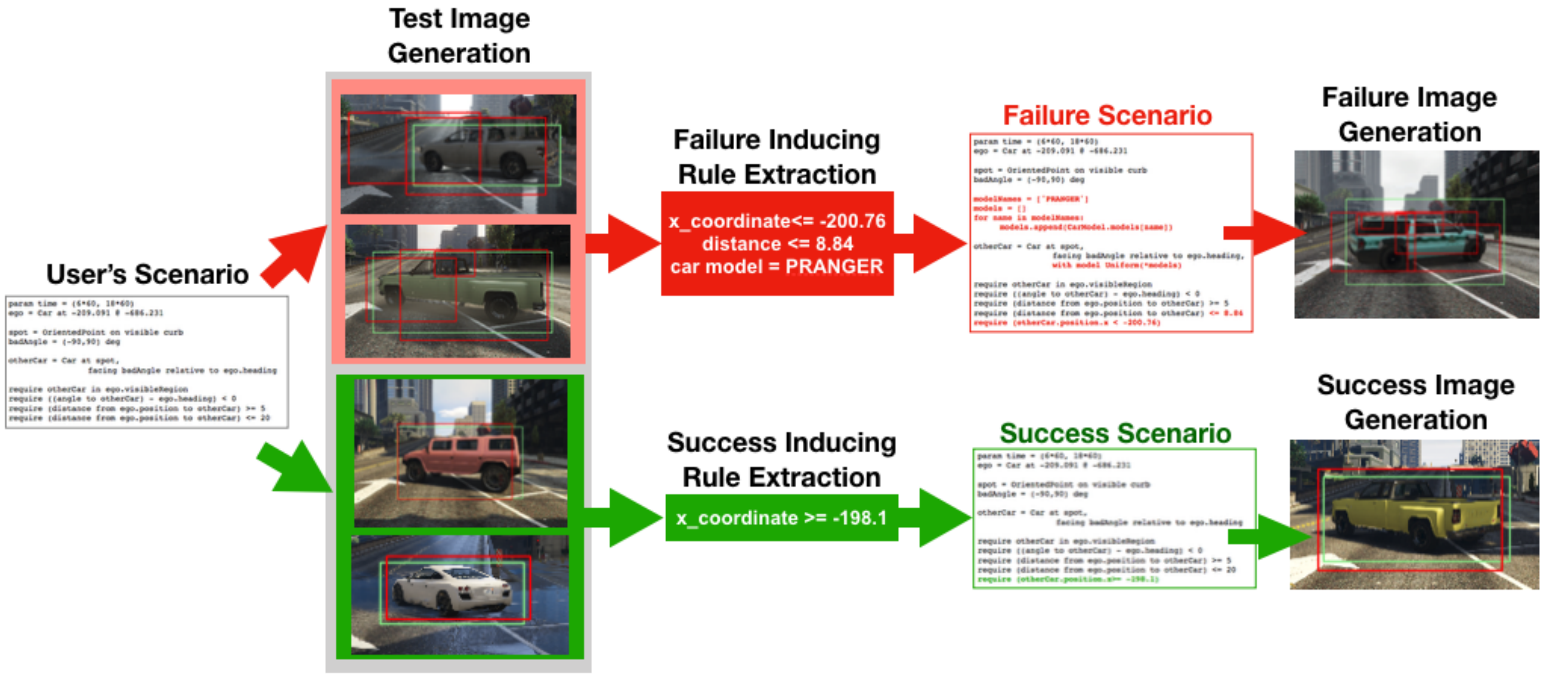
EXP Objectives summary

EASA Guidance	NASA
Objective EXP-01: The applicant should identify the list of humans that are intended to interact with the AI-based (sub)system, at any stage of its life cycle, together with their roles, their responsibilities and their expected expertise (including assumptions made on the level of training, qualification and skills).	NASA is funding work
Objective EXP-02: For each role, the applicant should identify which task(s) the humans are intended to perform in interaction with the AI-based (sub)system, as well as the task allocation pattern.	
Objective EXP-03: For each output of the (sub)system relevant to the task(s), the applicant should identify the need for an explanation depending on several criteria (including nature of the task, authority of the human and the level of AI-based system) and specify the set of necessary explanations to be provided to the human.	
Objective EXP-04: For each output relevant to the task(s), the applicant should ensure the validity of the specified explanation, based on actual measurements (e.g. monitoring) or on a quantification of the level of uncertainty.	

AI Monitoring

Anticipated MOC EX-03-3: Identification of the trained features relevant to the explanation objective.

- We have developed a semantic and programmatic framework for characterizing success and failure scenarios of a given perception module in the form of Scenic programs.
- The technique leverages decision-tree learning to derive rules in terms of high-level, meaningful features and generates new inputs that conform with these rules.



User's Scenario

```
param time = (6*60, 18*60)
ego = Car at -209.091 @ -686.231

spot = OrientedPoint on visible curb
badAngle = (-90,90) deg

otherCar = Car at spot,
                facing badAngle relative to ego.heading

require otherCar in ego.visibleRegion
require ((angle to otherCar) - ego.heading) < 0
require (distance from ego.position to otherCar) >= 5
require (distance from ego.position to otherCar) <= 20
```

Failure Scenario

```
param time = (6*60, 18*60)
ego = Car at -209.091 @ -686.231

spot = OrientedPoint on visible curb
badAngle = (-90,90) deg

modelNames = ['PRANGER']
models = []
for name in modelNames:
      models.append(CarModel.models[name])

otherCar = Car at spot,
                facing badAngle relative to ego.heading,
                with model Uniform(*models)

require otherCar in ego.visibleRegion
require ((angle to otherCar) - ego.heading) < 0
require (distance from ego.position to otherCar) >= 5
require (distance from ego.position to otherCar) <= 8.84
require (otherCar.position.x < -200.76)
```

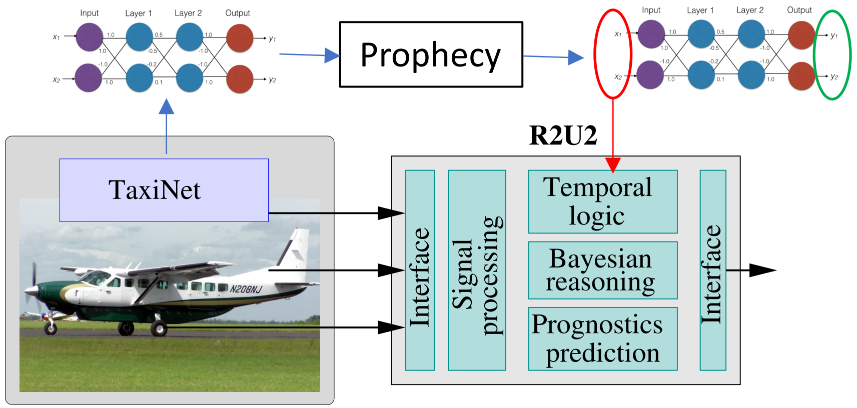
Success Scenario

```
param time = (6*60, 18*60)
ego = Car at -209.091 @ -686.231

spot = OrientedPoint on visible curb
badAngle = (-90,90) deg

otherCar = Car at spot,
                facing badAngle relative to ego.heading

require otherCar in ego.visibleRegion
require ((angle to otherCar) - ego.heading) < 0
require (distance from ego.position to otherCar) >= 5
require (distance from ego.position to otherCar) <= 20
require (otherCar.position.x>= -198.1)
```

AI Monitoring

Anticipated MOC EX-03-3: The AI-based system inputs should be monitored to be within the operational boundaries in which the AI/ML component performance is guaranteed, and deviations should be indicated to the human.

- Use of runtime monitoring (R2U2 tool) to monitor inputs based on pre-conditions computed by Prophecy
  - Or by MARGInS or SysAI

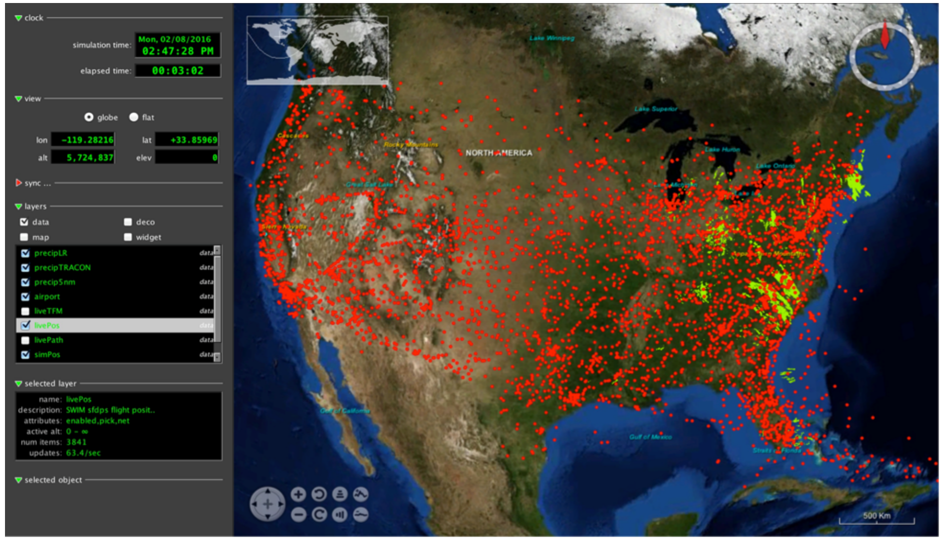


- R2U2 is a run-time monitoring and V&V tool that combines Metric Temporal Logic observers, Bayesian Network reasoners, and model-based prognostics.
- Integration of FRET with CoPilot (Stream-based runtime verification language)
  - Also used in the DARPA Assured Autonomy project.

AI Data Recording Capability

Objective EX-05: The applicant should provide the means to record operational data that is necessary to explain the behaviour of the AI-based (sub)system post operations.

- RACE is a framework for monitoring of complex airspace systems based on *actors* – concurrent objects that only communicate through asynchronous messages.
    - It can concurrently process independent data streams
    - We have also combined it with a runtime monitoring tool (MESA) to do overall system risk assessment in operations (or simulation).



clock
simulation time: Mon, 02/08/2016 02:47:28 PM
elapsed time: 00:03:02

view
● globe ○ flat
lon -119.28216 lat +33.85969
alt 5,724,837 elev 0

sync ...

layers
☑ data □ deco
□ map □ widget
☑ precipLR data
☑ precipTRACON data
☑ precip5nm data
☑ airport data
□ liveTFM data
☑ livePos data
□ livePath data
☑ simPos data

selected layer
name: livePos
description: SWIM sfdps flight posit..
attributes: enabled,pick,net
active alt: 0 - ∞
num items: 3841
updates: 63.4/sec

selected object

AI Recording Capability

- EASA: This monitoring consists in recording and processing data from day-to-day operation to detect and evaluate deviations from the expected behaviour of the AI-based system, as well as issues affecting interactions with human users or other systems.
    - *The prophecy runtime module will monitor for abnormal behaviors (deviations from expected behavior) based on patterns for correct and incorrect behavior extracted from the Taxinet model in an offline analysis.*
- EASA: Data recording for the purpose of accident or incident investigation.
    - accurately reconstruct the chronological sequence of inputs to and outputs from the AI-based system;
        - *In our Boeing Case Study we generated counter-example scenarios that represent the airplane violating safety properties and going off the runway.*
    - determine in this chronological sequence, which inputs and outputs most likely contributed to the accident or incident;
        - *In our Scenic work we generate rules in terms of the input features that represent incorrect behavior and show how they can help with debugging*
    - identify any unexpected behaviour of the AI-based system that is relevant for explaining the accident or incident; and
        - *In our repair work we perform fault localization to identify the parameters of a neural network model that contribute to failures.*
    - determine the causes of each unexpected behaviour of the AI-based system, either directly or by running simulations that accurately reproduce the unexpected behaviour.

EX Objectives summary

EASA Guidance	NASA
Anticipated MOC EX-03-1: Considerations on trained model interpretability.	
Anticipated MOC EX-03-2: Identification of the trained features relevant to the explanation objective.	• Work on explainability: semantic segmentation, object identification • Use decision-tree learning in Scenic to derive meaningful features
Anticipated MOC EX-03-3: The AI-based system inputs should be monitored to be within the operational boundaries in which the AI/ML component performance is guaranteed, and deviations should be indicated to the human.	• Integration of Prophecy and R2U2 • Integration of MARGInS and R2U2 • MESA for system-wide monitoring • Integration of FRET and CoPilot
Anticipated MOC EX-03-4: The training and instructions available for the human should include procedures to act on the possible outputs of the ODD monitoring.	
Anticipated MOC EX-03-5: Information concerning unsafe system operating conditions should be provided to the human operator/end user to enable them to take appropriate corrective action in a timely manner.	• MARGInS / SYsAI can be sued to compute unsafe operating conditions
Objective EX-05 (EXP-05): The applicant should provide the means to record operational data that is necessary to explain the behaviour of the AI-based (sub)system post operations.	RACE

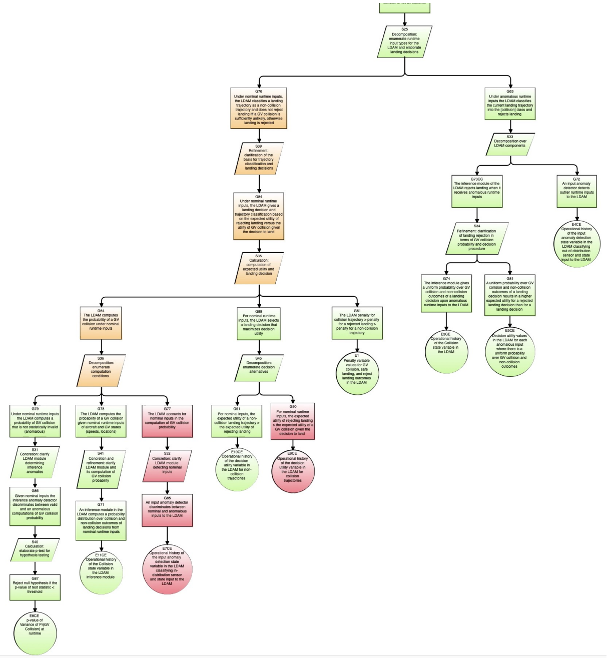
Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

AI safety risk mitigation

Objective SRM-01: Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or if an additional dedicated layer of protection, called hereafter safety risk mitigation (SRM), would be necessary to mitigate the residual safety risks to an acceptable level.

- Extend AdvoCATE (tool for creating and visualizing safety cases) into a dynamic dashboard to reason about safety during operations and support risk-based decision making
- trade trees record decision alternatives (e.g., design decisions) to be presented with
    - impact of each alternative on baseline risk levels, as well as on existing assurance artifacts,
    - decisions about alternatives to be made and recorded, and
    - justifications of those decisions to be provided as assurance arguments.
- Integration of decision-making, assurance impact analysis, and (safety) assurance case development within a common framework.

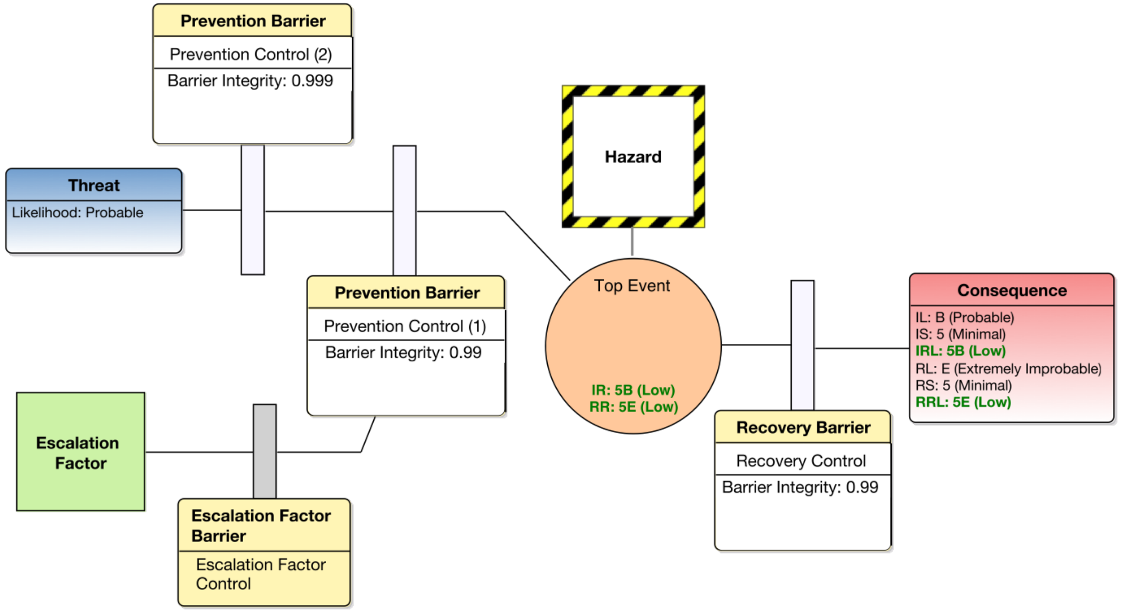


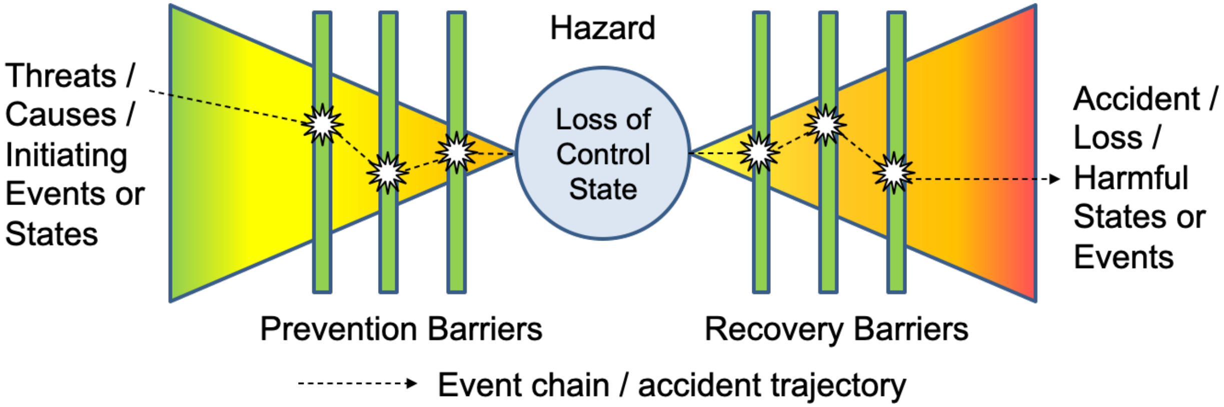
S25 Decomposition: enumerate runtime input types for the LDAM and elaborate landing decisions
G76 Under nominal runtime inputs, the LDAM classifies a landing trajectory as a non-collision trajectory and does not reject landing iff a GV collision is sufficiently unlikely, otherwise landing is rejected
G63 Under anomalous runtime inputs the LDAM classifies the current landing trajectory into the [collision] class and rejects landing
S39 Refinement: clarification of the basis for trajectory classification and landing decisions
S33 Decomposition over LDAM components
G84 Under nominal runtime inputs, the LDAM gives a landing decision and trajectory classification based on the expected utility of rejecting landing versus the utility of GV collision given the decision to land
G73CC The inference module of the LDAM rejects landing when it receives anomalous runtime inputs
G72 An input anomaly detector detects outlier runtime inputs to the LDAM
S34 Refinement: clarification of landing rejection in terms of GV collision probability and decision procedure
E4CE Operational history of the input anomaly detection state variable in the LDAM classifying out-of-distribution sensor and state input to the LDAM
S35 Calculation: computation of expected utility and landing decision
G74 The inference module gives a uniform probability over GV collision and non-collision outcomes of a landing decision upon anomalous runtime inputs to the LDAM
G81 A uniform probability over GV collision and non-collision outcomes of a landing decision results in a higher expected utility for a rejected landing decision than for a landing decision
G64 The LDAM computes the probability of a GV collision under nominal runtime inputs
G89 For nominal runtime inputs, the LDAM selects a landing decision that maximizes decision utility
G61 The LDAM penalty for collision trajectory > penalty for a rejected landing > penalty for a non-collision trajectory
E3CE Operational history of the Collision state variable in the LDAM
E5CE Decision utility values in the LDAM for each anomalous input where there is a uniform probability over GV collision and non-collision outcomes
S36 Decomposition: enumerate computation conditions
S45 Decomposition: enumerate decision alternatives
E1 Penalty variable values for GV collision, safe landing, and reject landing outcomes in the LDAM
G79 Under nominal runtime inputs the LDAM computes a probability of GV collision that is not statistically invalid (anomalous)
G78 The LDAM computes the probability of a GV collision given nominal runtime inputs of aircraft and GV states (speeds, locations)
G77 The LDAM accounts for nominal inputs in the computation of GV collision probability
G91 For nominal inputs, the expected utility of a non-collision landing trajectory > the expected utility of rejecting landing
G90 For nominal runtime inputs, the expected utility of rejecting landing > the expected utility of a GV collision given the decision to land
S31 Concretion: clarify LDAM module determining inference anomalies
S41 Concretion and refinement: clarify LDAM module and its computation of GV collision probability
S32 Concretion: clarify LDAM module detecting nominal inputs
E10CE Operational history of the decision utility variable in the LDAM for non-collision trajectories
E9CE Operational history of the decision utility variable in the LDAM for collision trajectories
G86 Given nominal inputs the inference anomaly detector discriminates between valid and an anomalous computations of GV collision probability
G71 An inference module in the LDAM computes a probability distribution over collision and non-collision outcomes of landing decisions from nominal runtime inputs
G85 An input anomaly detector discriminates between nominal and anomalous inputs to the LDAM
S40 Calculation: elaborate p-test for hypothesis testing
E11CE Operational history of the Collision state variable in the LDAM inference module
E7CE Operational history of the input anomaly detection state variable in the LDAM classifying in-distribution sensor and state input to the LDAM
G87 Reject null hypothesis if the p-value of test statistic < threshold
E8CE p-value of Variance of Pr(GV Collision) at runtime

AI Safety Risk Mitigation

Objective SRM-02: The applicant should establish SRM means as identified in Objective SRM-01.

- AdvoCATE Barrier models
  - Risk scenarios showing chain of events leading to accidents, loss, or harm
  - Represented using Bow Tie Diagrams (BTDs)
  - Barrier = Risk reduction mitigation
    - Prevention or recovery function
    - Reduction of event probability and severity




Hazard
Threats / Causes / Initiating Events or States
Loss of Control State
Accident / Loss / Harmful States or Events
Prevention Barriers
Recovery Barriers
Event chain / accident trajectory

AI SAFETY RISK MITIGATION

> **Anticipated MOC SRM-01:** In establishing whether AI SRM is necessary and to which extent, the following considerations should be accounted for:
> - coverage of the explainability building block;
> - coverage of the learning assurance building block;
> - ….

- This summer, we will explore techniques for measuring coverage of neural network models.
  - We believe that the patterns extracted by Prophecy from a trained model have the potential to capture the behavioral / functional / feature coverage, which can in turn be used to estimate the coverage of the explainability and learning assurance blocks.
- We also plan to evaluate existing testing frameworks for NN.

SRM Objectives summary

EASA Guidance	NASA
Objective SRM-01: Once activities associated with all other building blocks are defined, the applicant should determine whether the coverage of the objectives associated with the explainability and learning assurance building blocks is sufficient or if an additional dedicated layer of protection, called hereafter safety risk mitigation (SRM), would be necessary to mitigate the residual safety risks to an acceptable level.	AdvoCATE and the dynamic dashboard can be used to make risk-based design decisions
Anticipated MOC SRM-01: In establishing whether AI SRM is necessary and to which extent, the following considerations should be accounted for: • coverage of the explainability building block; • coverage of the learning assurance building block; • ….	Future work: explore techniques for measuring coverage of neural network models using Prophecy-computed patterns
Objective SRM-02: The applicant should establish SRM means as identified in Objective SRM-01.	AdvoCATE has bow tie diagrams representing what prevention barriers are in place to prevent a risk from being realized and recovery barriers for mitigating it after it is realized.

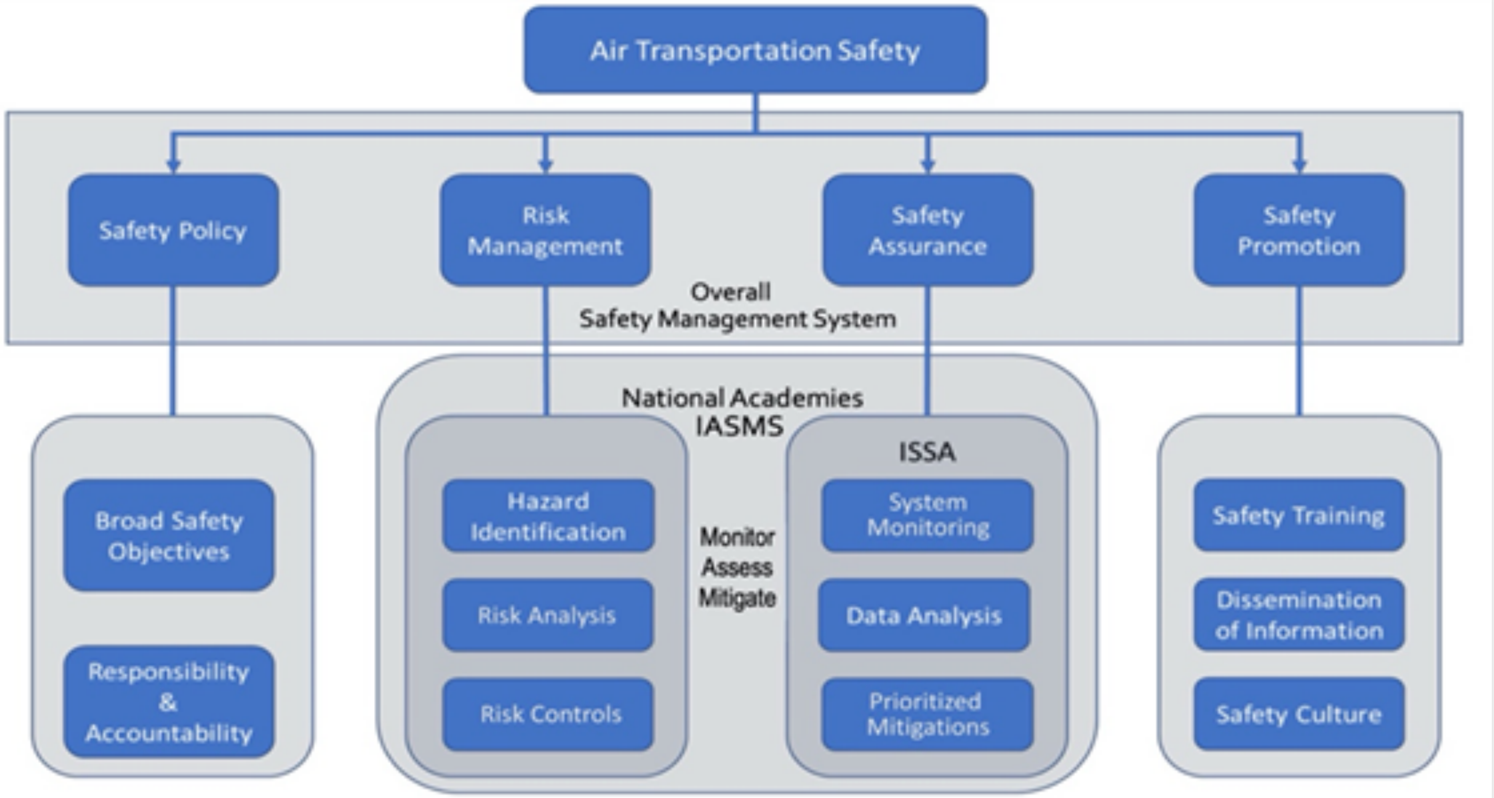
Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

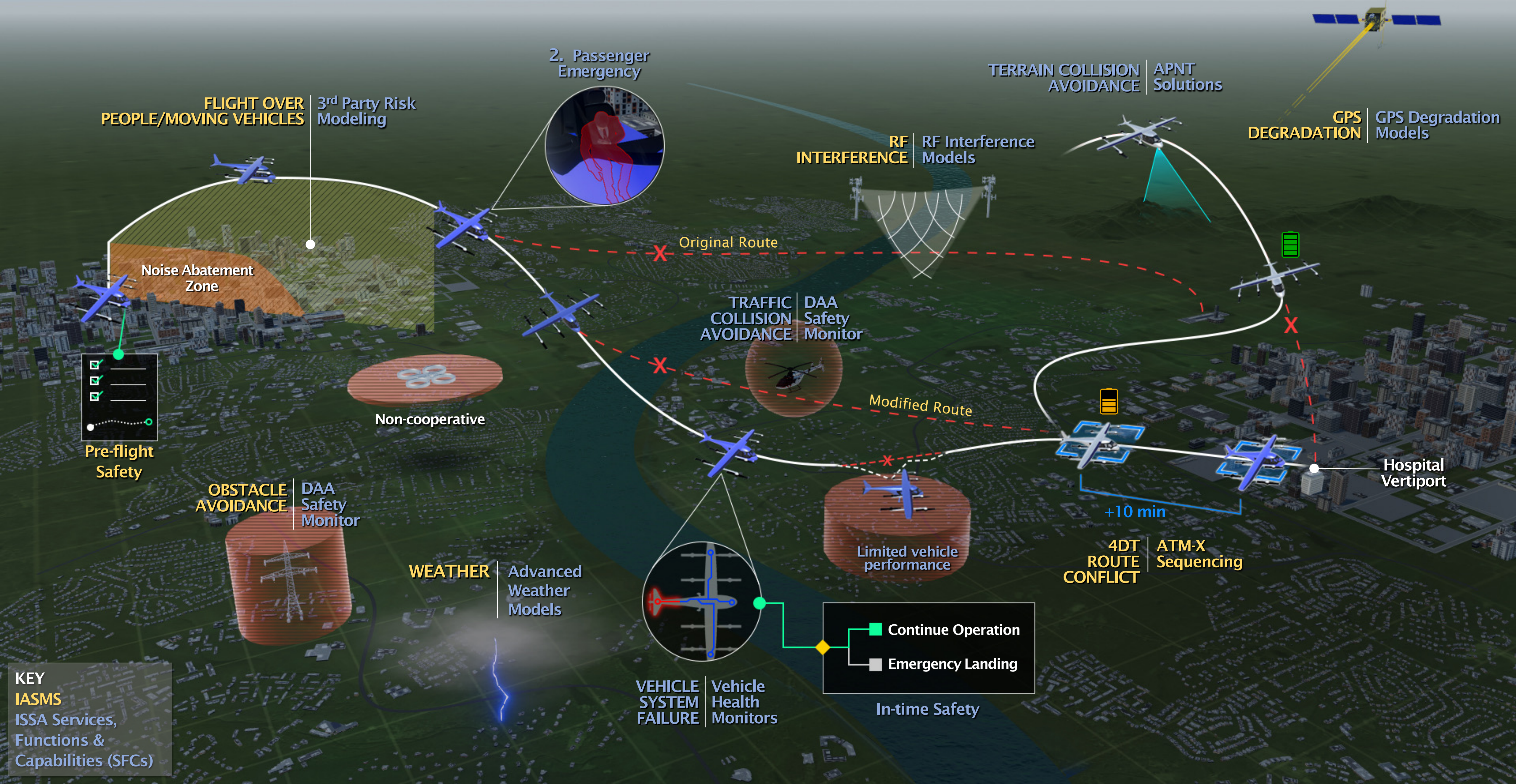
AI continuous safety assessment

Provision ORG-02: Implement a data-driven ‘AI continuous safety assessment system’ based on operational data and in-service events.

- In-time Aviation Safety Management System (IASMS) to proactively mitigate risks and demonstrate innovative solutions while ultimately ensuring safety to the community on the ground and in the National Airspace System
  - Monitor: domain-specific safety monitoring and alerting tools,
  - Assess: integrated predictive technologies with domain-level applications, and
  - Mitigate: in-time safety threat management.
- POC: Kyle Ellis, kyle.k.ellis@nasa.gov



ISSA/IASMS OV – 1



ORG Objectives summary

EASA Guidance	NASA
Provision ORG-01: The organisation should review its processes and adapt them to the introduction of AI technology	
Provision ORG-02: Implement a data-driven ‘AI continuous safety assessment system’ based on operational data and in-service events.	IASMS

Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Miscellaneous objectives table
  - NASA Tool/technique/processes table

Other objectives

EASA Guidance	NASA related work
Objective CL-01: The applicant should classify the AI-based (sub)system, based on the levels presented in Table 1 — EASA AI typology and definitions, with adequate justifications.	
Objective IS-01: For each AI-based (sub)system and its data sets, the applicant should manage those information security risks with an impact on safety, identifying and accounting for specific threats introduced by AI/ML usage.	• One group attached to ARMD is dedicated to watching cybersecurity issues in Aviation (POC: Paul Nelson) • May see work by Dr. Walsh in the future on this topic.
Objective ET-01: The applicant should perform an ethics-based trustworthiness assessment for any AI-based (sub)system developed using ML techniques or incorporating ML algorithms.	• Some past work on biases in ML
Objective ET-02: In performing the ethics-based trustworthiness assessment, the applicant should address questions from the EU Commission Assessment List for Trustworthy AI (ALTAI), taking into account the clarifications brought in the following anticipated MOC.	
Objective ET-03: The applicant should comply with national and EU data protection regulations (e.g. GDPR), i.e. involve their Data Protection Officer (DPO), consult with their National Data Protection Authority, etc.	
Objective ET-04: The applicant should assess the environmental impact of the AI-based (sub)system.	

Other objectives

EASA Guidance	NASA related work
Objective DM-14: The applicant should perform a data and learning verification step to confirm that the appropriate data sets have been used for the training, validation and verification of the model and that the expected guarantees (generalisation, robustness) on the model have been reached	Covered to some extent by our work on generalization
Objective CM-01: The applicant should apply all configuration management principles to the AI-based (sub)system life-cycle data, including but not limited to: …	
Objective QA-01: The applicant should ensure that quality/process assurance principles are applied to the development of the AI-based (sub)system, with the required independence level.	

Agenda

- NASA Research Map
- Introduction
- AI Trustworthiness Guidelines
  - Trustworthiness Analysis
  - Learning Assurance
  - AI Explainability
  - AI Safety Risk Mitigation
  - Organizations
- Summary
  - Objectives table
  - NASA Tool/technique/processes table

NASA Tools

Tools	Description	Availability	Technical POC	POC Email
FRET	Requirement elicitation and analysis	Open Source	Dimitra Giannakopoulou	dimitra.giannakopoulou@nasa.gov
PRECISA	Floating-point static code analysis	Open Source	Laura Titolo	Laura.titolo@nasa.gov
fmdtools	Resiliency analysis in design	Open Source	Daniel Hulse	daniel.e.hulse@nasa.gov
AdvoCATE	Assurance case automation toolset	Open Source	Ewen Denney	ewen.w.denney@nasa.gov
MARGInS	ML/statistical libraries for system testing	Usage Agreement	Carlos Paradis	carlos.v.paradis@nasa.gov
SysAI	ML/statistical libraries for system testing	Not available yet	Yuning He	yuning.he@nasa.gov
Drishti	Traceability hints using NLP	Not available yet	Nija Shi	nija.shi@nasa.gov

NASA Tools

Tools	Description	Availability	Technical POC	POC Email
AdaStress	Adaptive stress testing	Open Source	Ritchie Lee	ritchie.lee@nasa.gov
RACE	Runtime for Airspace Concept Evaluation	Open Source	Peter Mehlitz	peter.c.mehlitz@nasa.gov
MESA	Run-time analysis of live data streams	Open Source	Nastaran Shafiei	nastaran.shafiei@nasa.gov
Prophecy	Formal analysis of Neural Networks	Not available yet	Corina Pasareanu	corina.s.pasareanu@nasa.gov
R2U2	Vehicle-level run-time analysis	Usage Agreement	Johann Schumann	johann.m.schumann@nasa.gov
SMARt	NLP for failure taxonomy	Not available yet	Hannah Walsh	hannah.s.walsh@nasa.gov
CoPilot	Stream-based runtime verification	Open Source	Aaron Dutle	aaron.m.dutle@nasa.gov