IRT SAINT EXUPÉRY

Du modèle d'activité au modèle d'argumentation : apport des techniques de MBSE à l'ingénierie des systèmes à base d'IA

AFIS CROcc CNES COMET

Le programme de recherche Confiance.ai

fit FRENCH INSTITUTES OF TECHNOLOGY

La gestion de l'héritage Confiance.ai

THE EUROPEAN TRUSTWORTHY AI ASSOCIATION

The European Trustworthy AI Association is a non-profit organization established by industrial leaders, building on the legacy of the Confiance.ai programme. It is on a mission to empower the industry with state of the art, open-source methodology and tools, enabling the engineering of AI-based systems that can be trusted and comply with regulations.

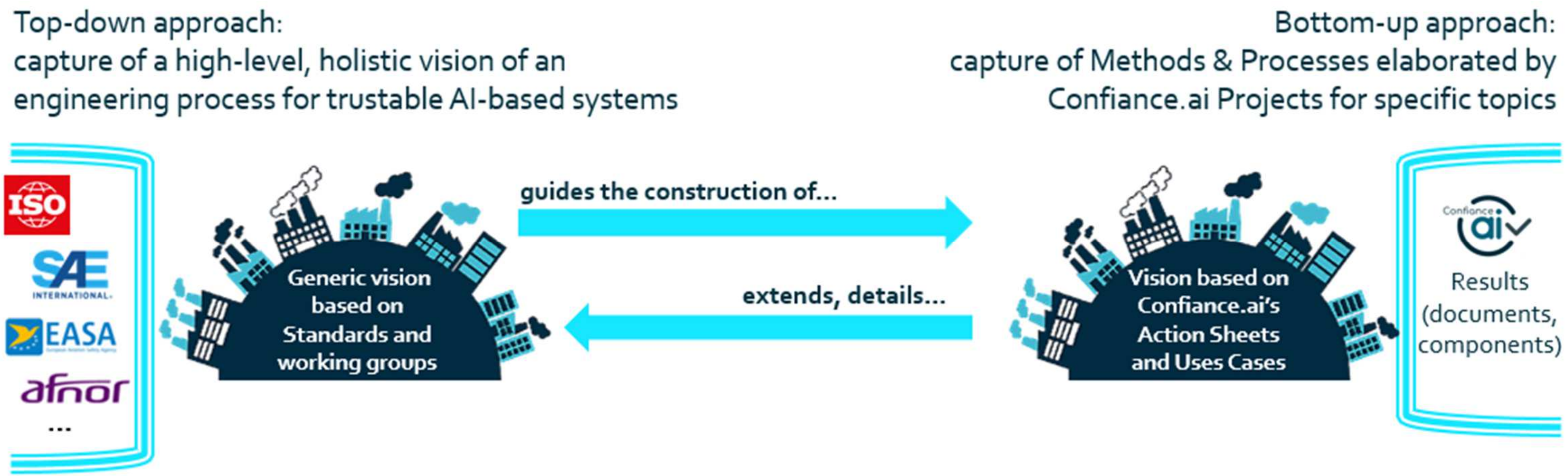
The association aims to be a driving force behind an ambitious European strategy for industrial and responsible AI. Its ambition is to propel Europe to the forefront of innovation in trustworthy AI, by making its methodologies and tools an international benchmark and thus, supporting the broader adoption of responsible AI in industy.

Buts de la méthode « end-to-end » de Confiance.ai

- Complete the « classical » engineering disciplines (Systems Engineering, Software Engineering) to take into account the specificities of ML, with modifications only where necessary
- Structure the results of Confiance.ai (local methods, software components) to facilitate their use

Double approche pour construire la méthode

SAINT EXUPÉRY

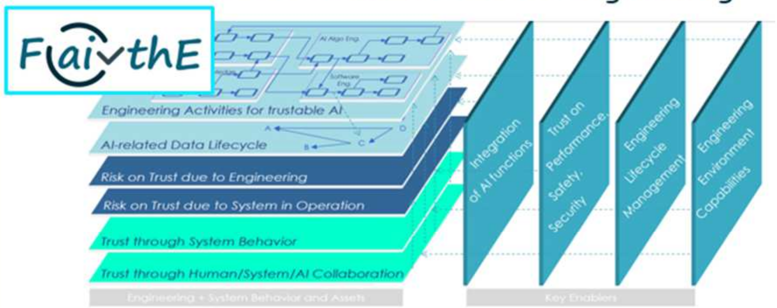


fit | FRENCH INSTITUTES OF TECHNOLOGY

Modélisation des processus et activités d'ingénierie

Meta-model

Framework for AI Trustable & Holistic Engineering

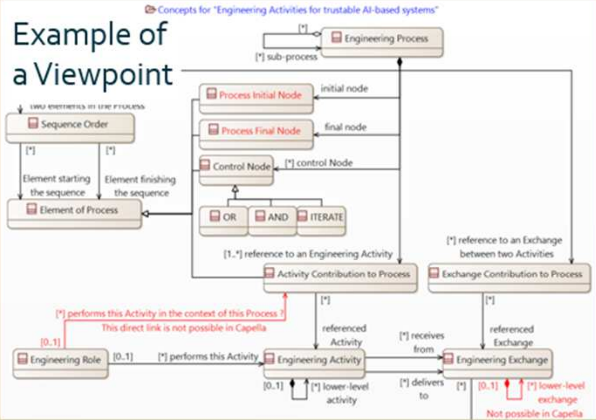


Several crossing Viewpoints to capture the concerns

Implemented in Capella

Meta-model describing all viewpoints of FaithE
+
mapping between concepts of the meta-model and Capella concepts to be used in the Model

Example of a Viewpoint

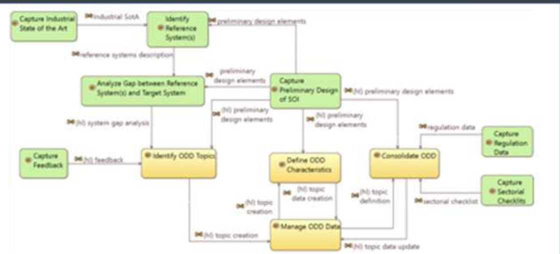


Model

Implemented in Capella

- Model fragment for high-level engineering tasks
- High-level breakdown of specific AI activities, Connections between these activities
- …

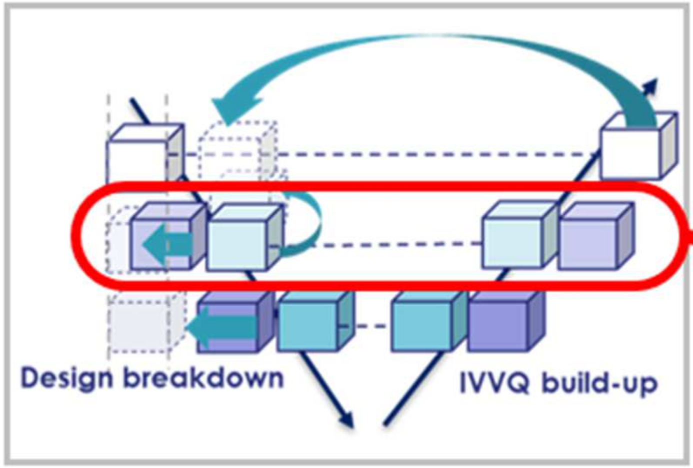
Example of modeled process for ODD (in work)

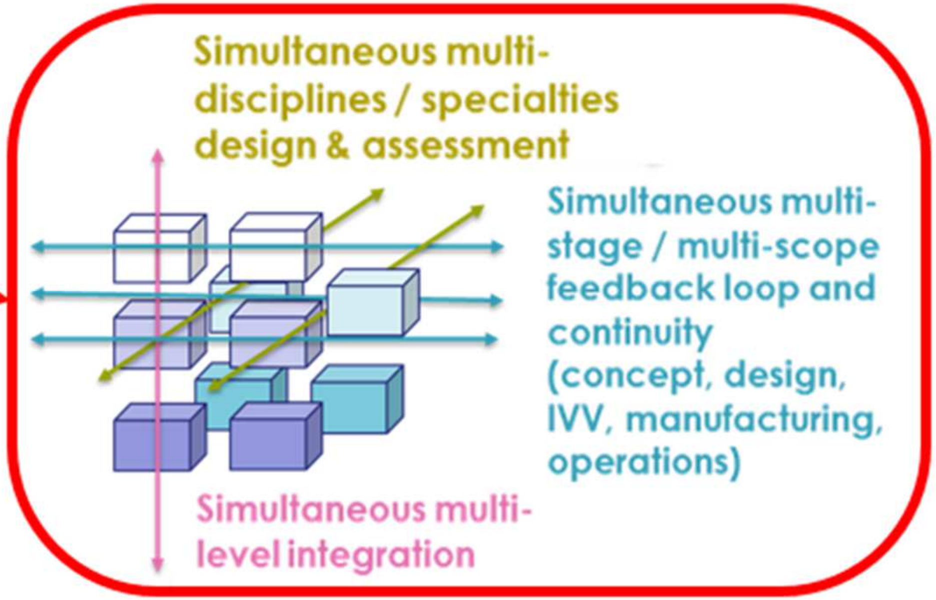


- Model fragment for AI monitoring
- Model fragment for robustness test
- Model fragment for ODD building
- Model fragment for …

FRENCH INSTITUTES OF TECHNOLOGY

Complexité d'une telle méthode





FRENCH INSTITUTES OF TECHNOLOGY

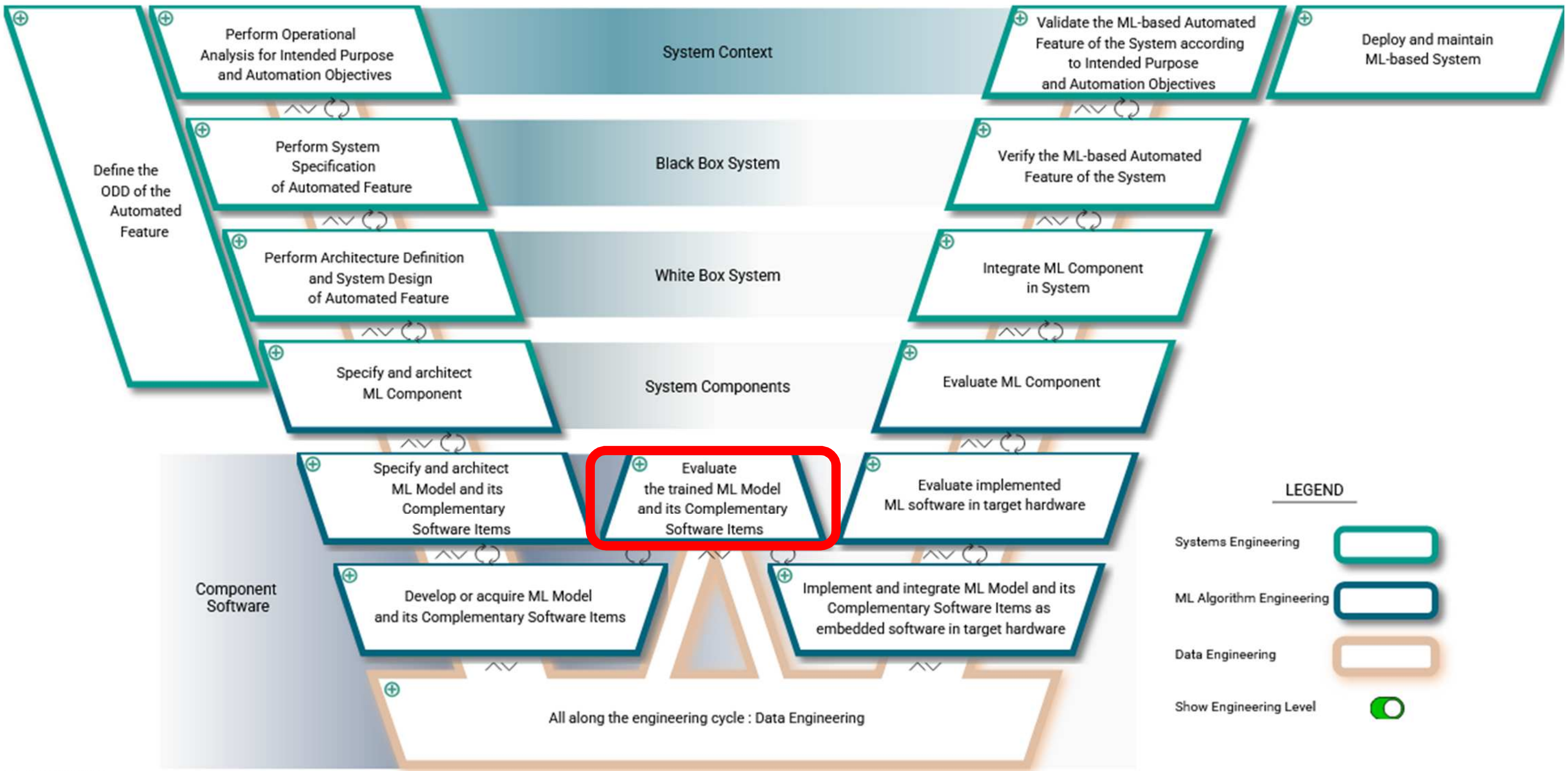
Méthode accessible via la Body Of Knowledge

SAINT EXUPÉRY

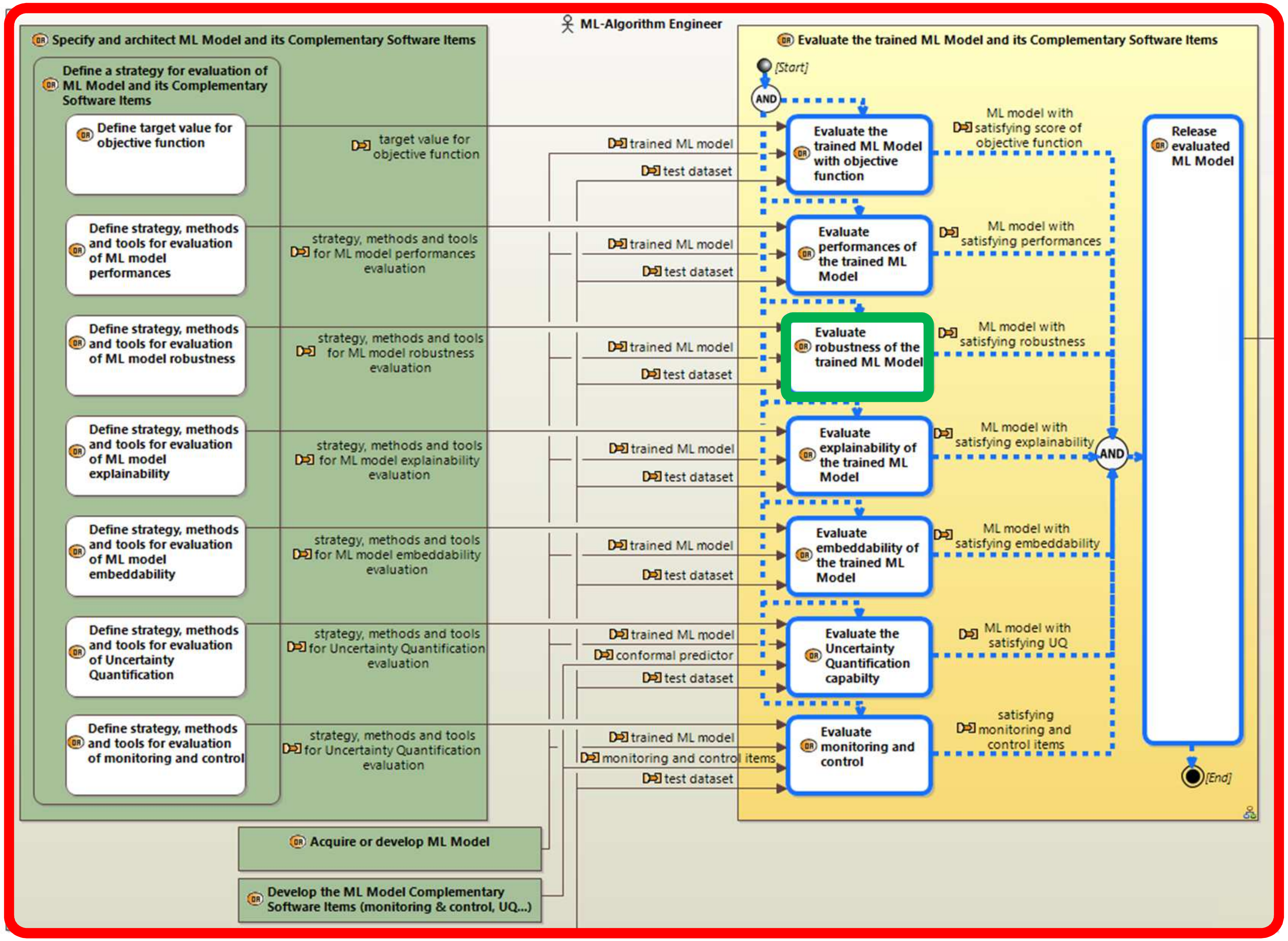
https://bok.confiance.ai/

FRENCH INSTITUTES OF TECHNOLOGY

Vue de plus haut niveau

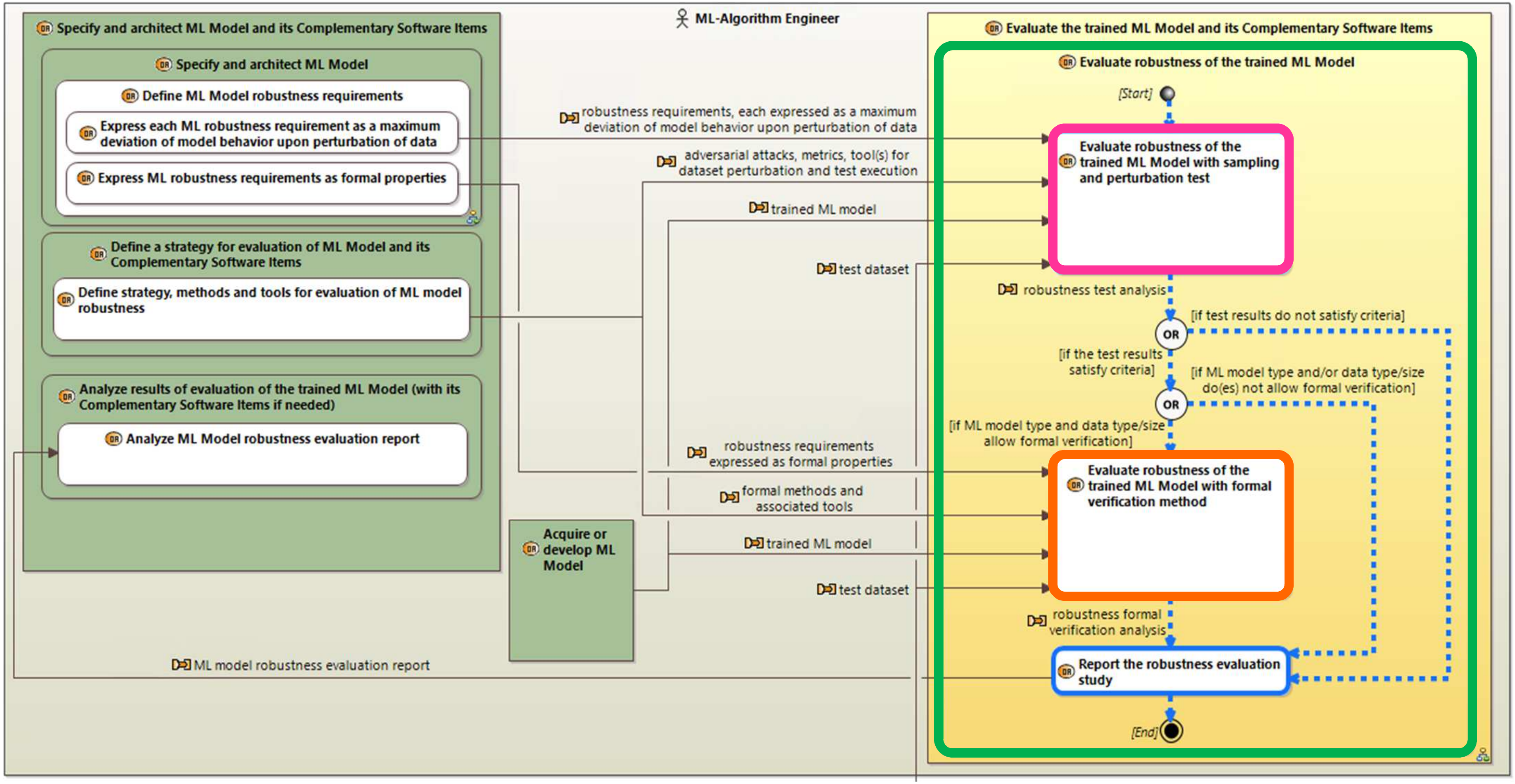


Zoom sur “Evaluation of ML Model”



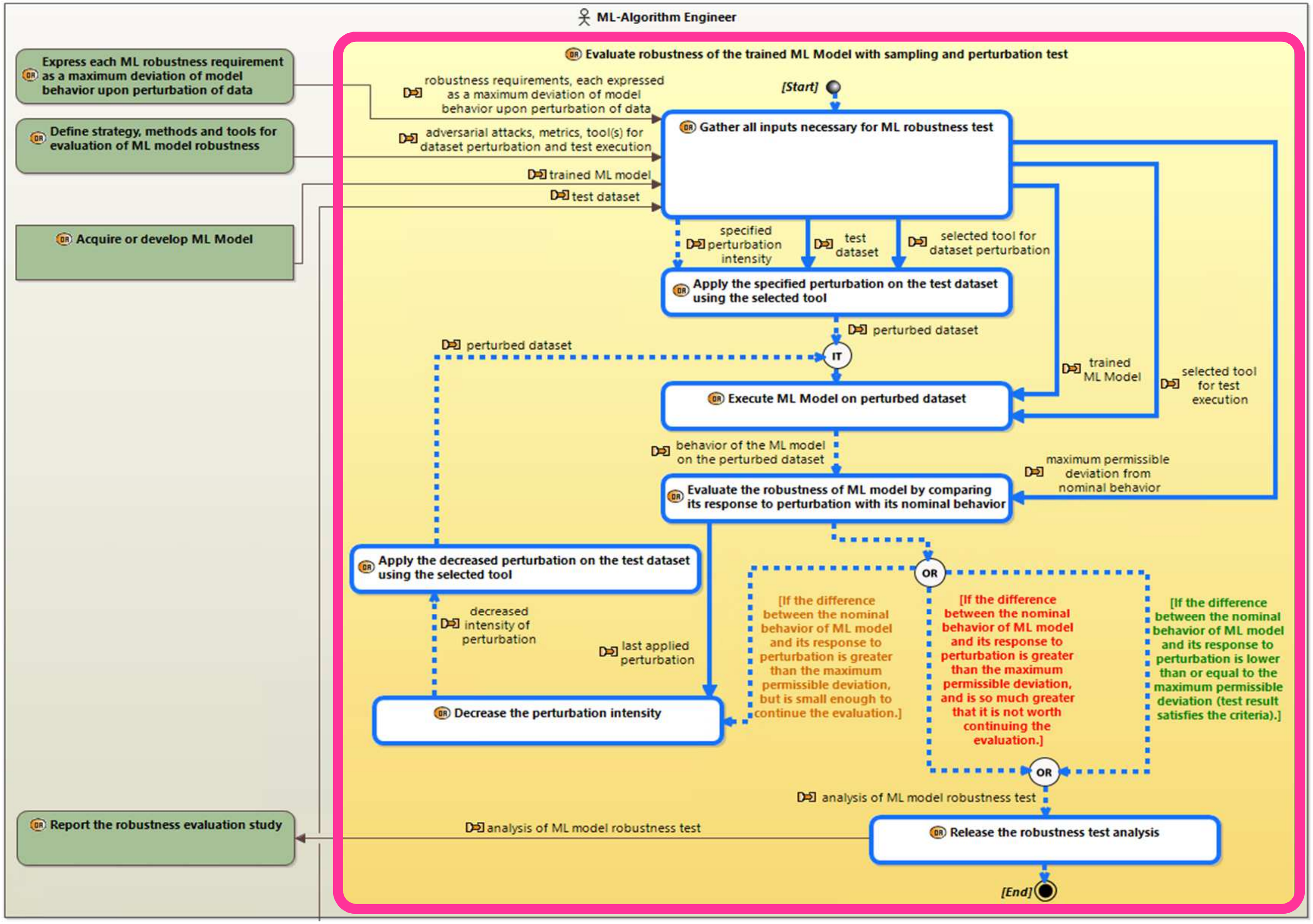
FRENCH INSTITUTES OF TECHNOLOGY

Zoom sur “Evaluation of ML Model robustness”



FRENCH INSTITUTES OF TECHNOLOGY

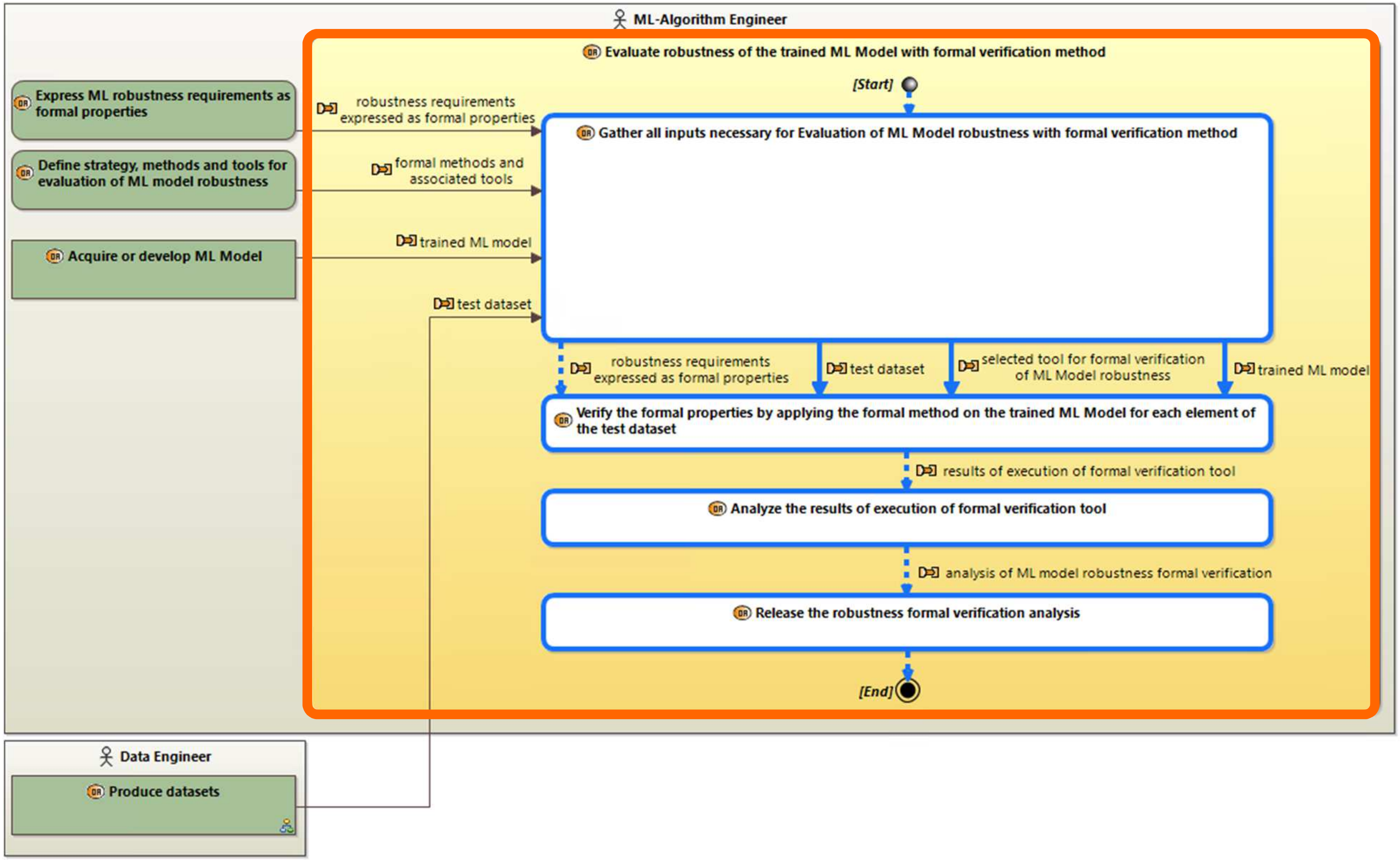
Zoom sur “Test of ML Model robustness”



Related to :

- §B.2. of document "Methodological Guideline for Robustness Functional Set"
- Component 331: Adversarial Attack Characterization Component
- Component 332: AI Metamorphis Observer Component (AIMOS)
- Component 333: Amplification Method for Robustness Evaluation Component
- Component 334: Non-overlapping Corruption Benchmarker Component
- Component 335: Time-series Robustness Characterizer Component
- Component 3141: Chiru

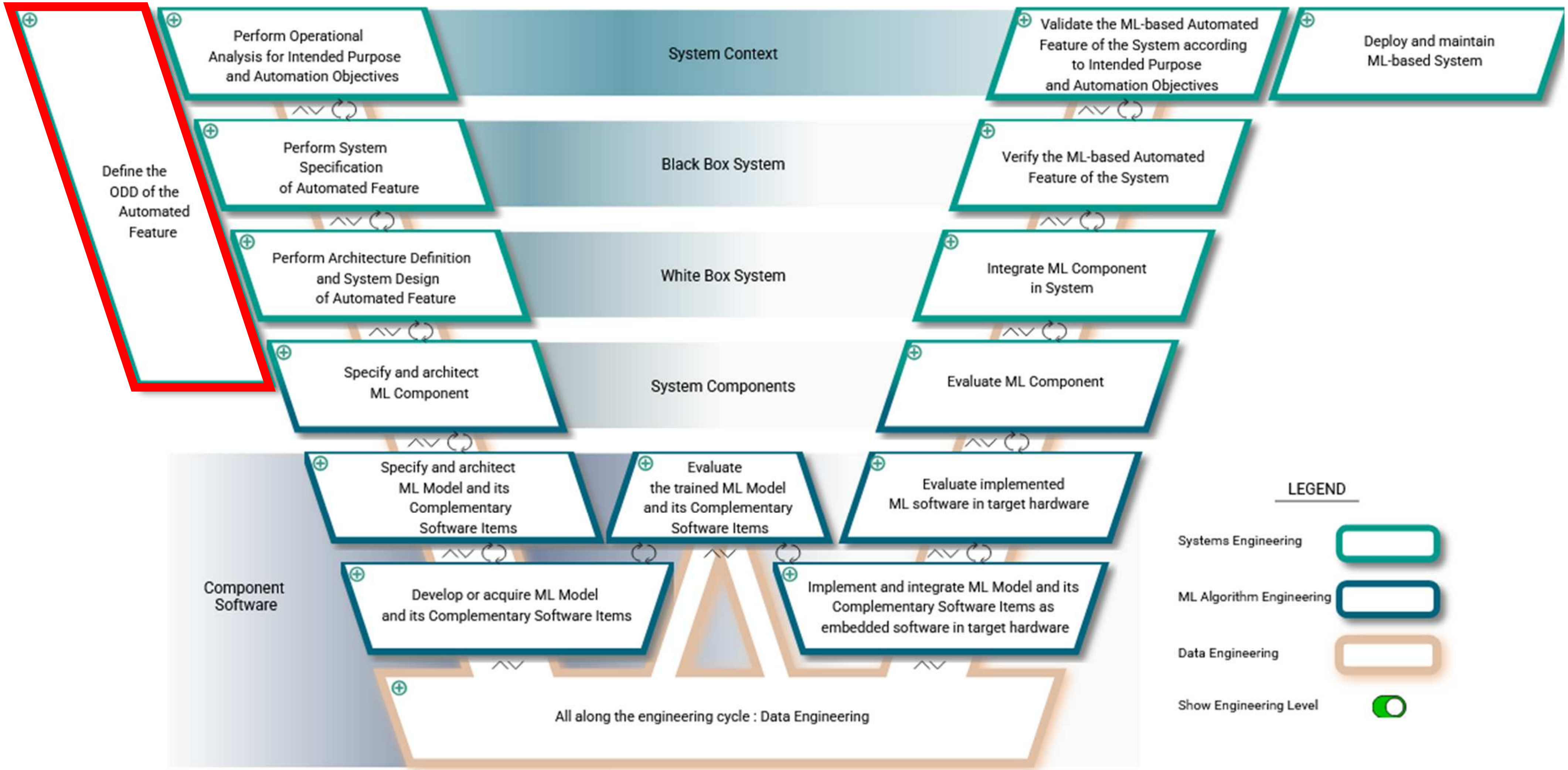
Zoom sur “Formal verification of ML Model robustness”



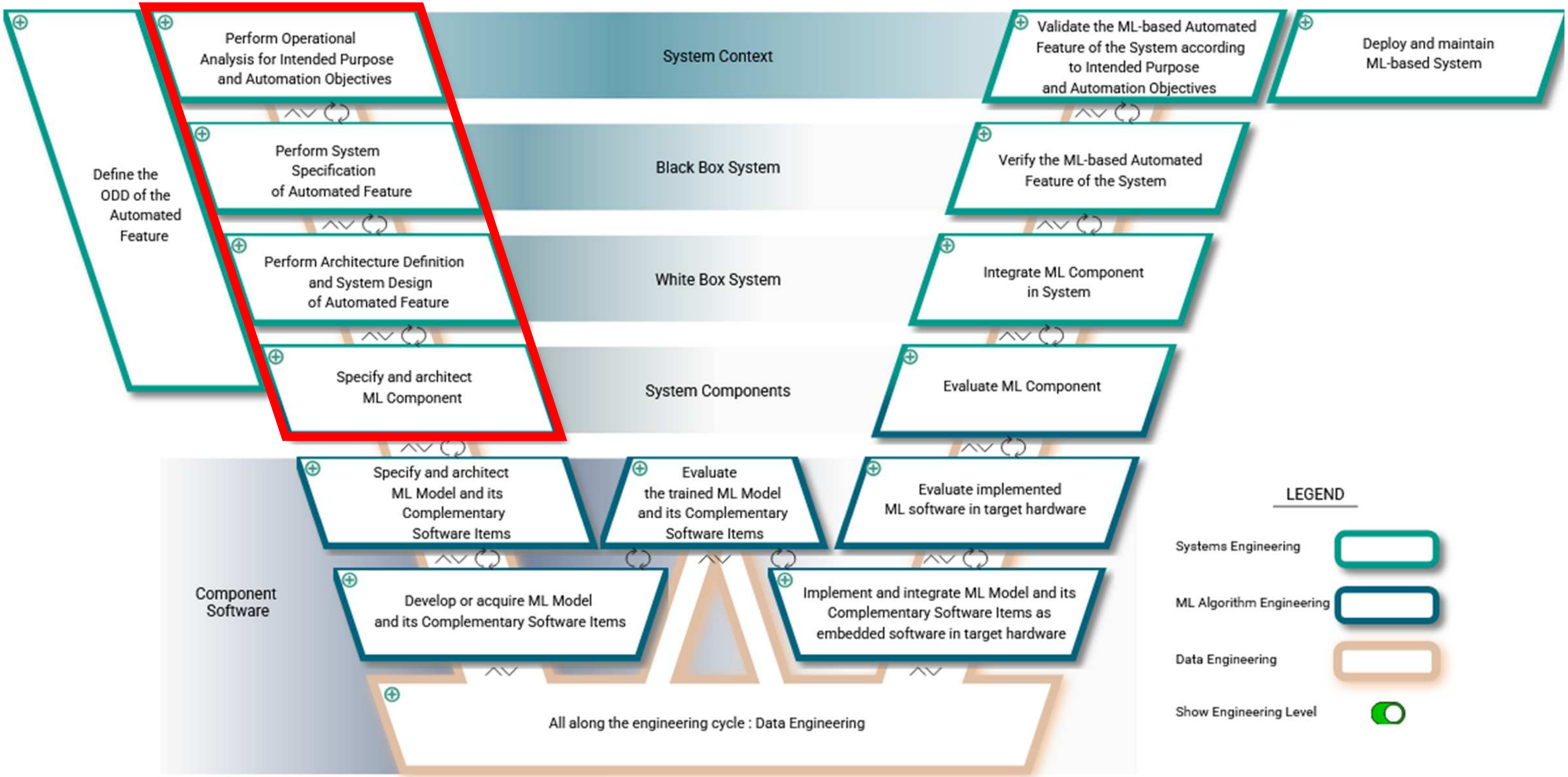
Related to :

- §B.1. of document "Methodological Guideline for Robustness Functional Set"
- Component 321: Saimple
- Component 322: nnenum
- Component 323: α-β-crown
- Component 3171: PyRAT
- Component 391: MIP Solver

Spécificités de l'IS of d'un système basé ML: ODD

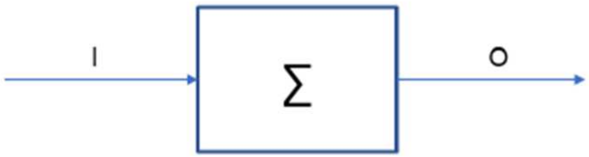


Spécificités de l'IS of d'un système basé ML: écart avec un système classique



Spécificités de l'IS of d'un système basé ML: écart avec un système classique

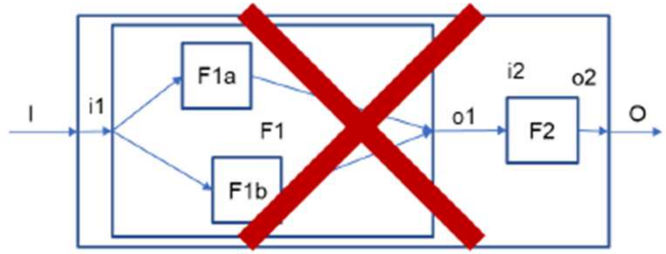
Function of a "conventional" component



Transfer function (Σ)
I/O
f1 f2 … fn

Function of an ML-based component

Functional requirement
I/O
F1 F2 . Fn

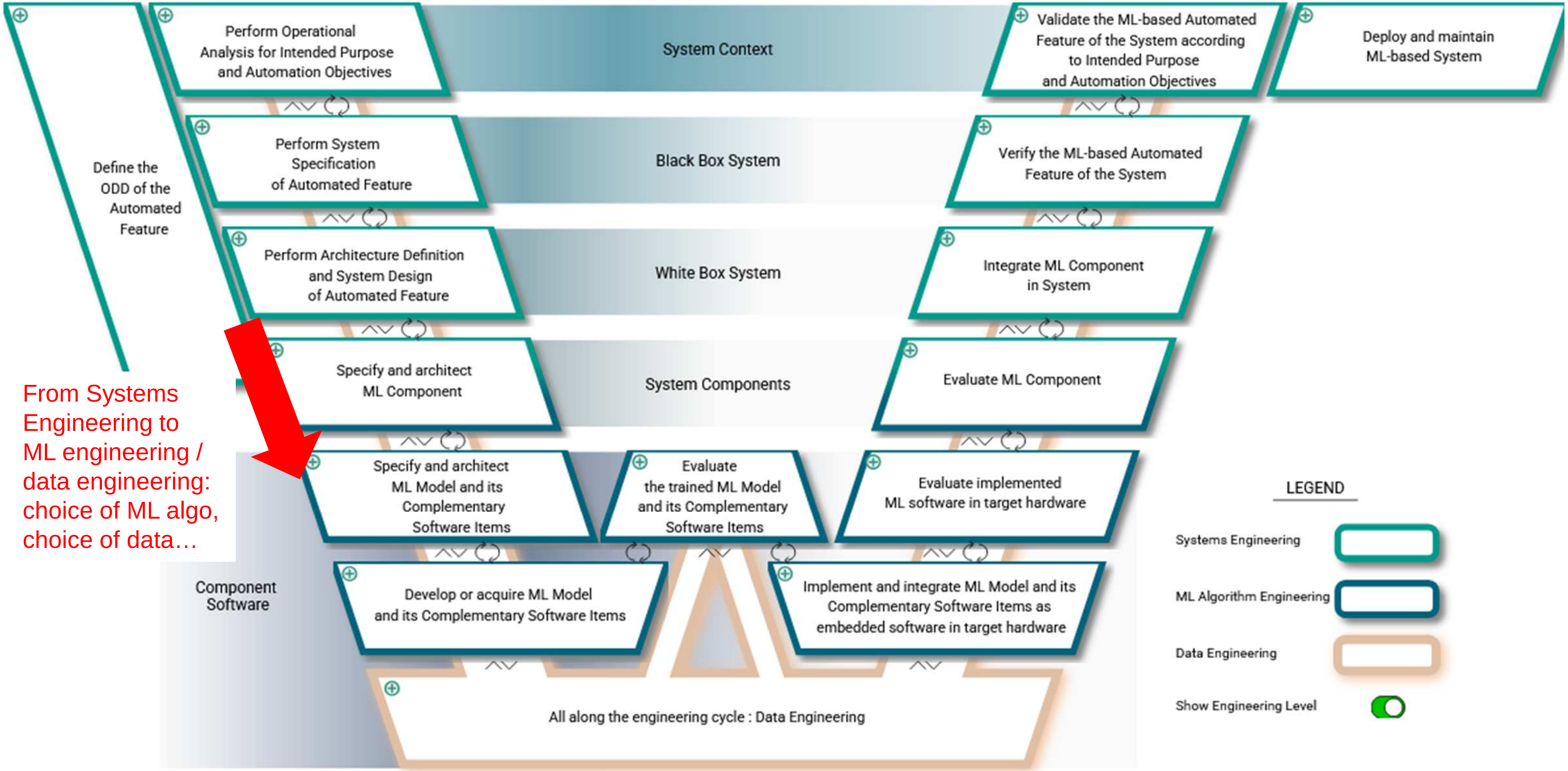


- Expected environmental conditions
- Harmful environmental conditions
- Required intended behavior
- Unwanted or Disturbing Behavior

FIT | FRENCH INSTITUTES OF TECHNOLOGY

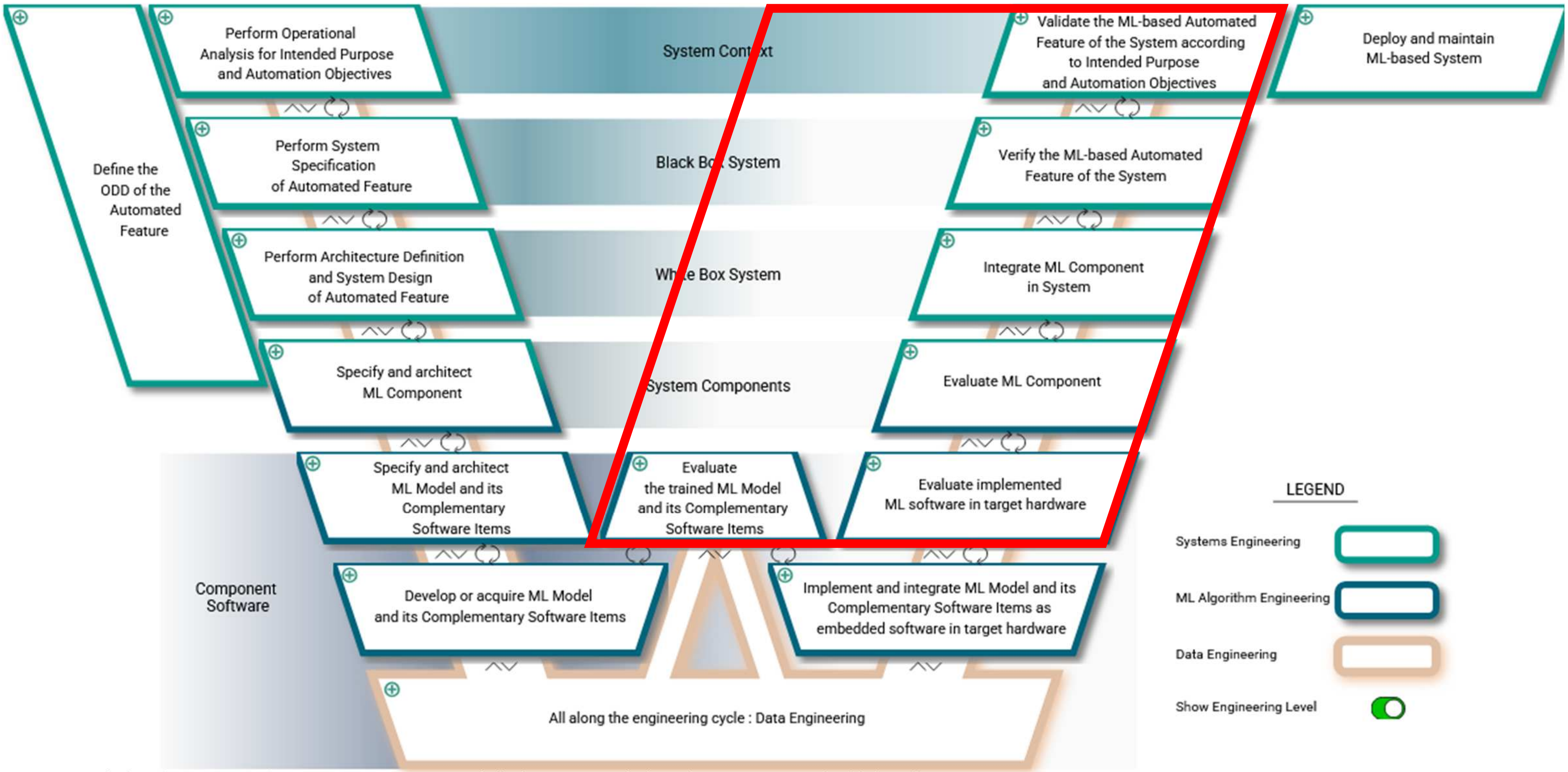
Spécificités de l'IS of d'un système basé ML: écart avec un système classique

SAINT EXUPÉRY



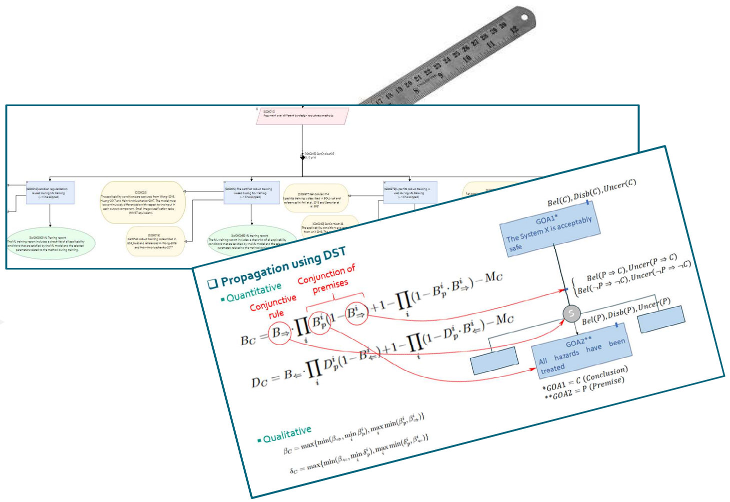
From Systems Engineering to ML engineering / data engineering: choice of ML algo, choice of data…

V&V d’un système basé ML

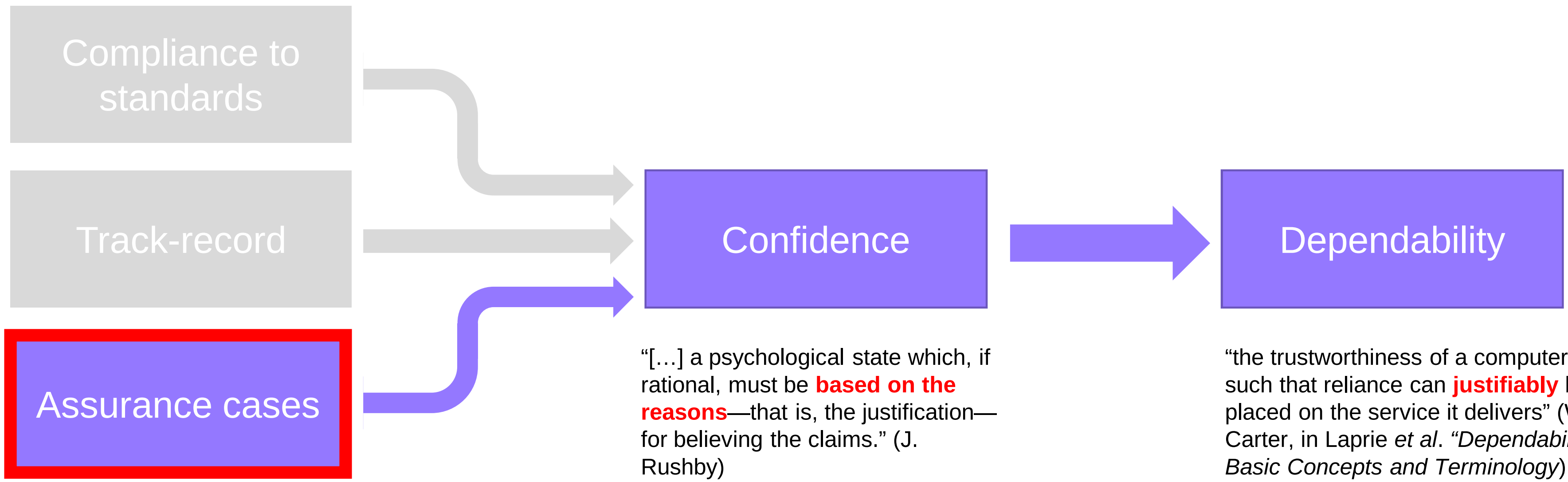


De la Confiansssssse à la Confiance

E. JENN – IRT Saint-Exupery



Trusting ML-based systems?



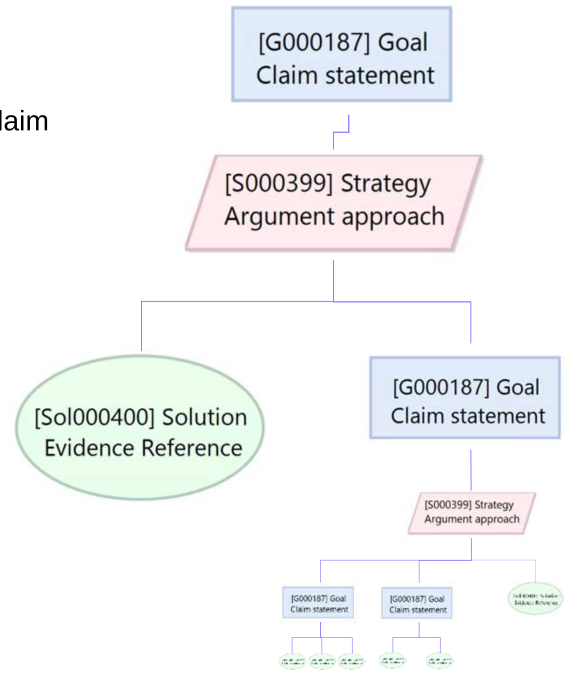
"[…] a psychological state which, if rational, must be **based on the reasons**—that is, the justification—for believing the claims." (J. Rushby)

"the trustworthiness of a computer system such that reliance can **justifiably** be placed on the service it delivers" (W.C. Carter, in Laprie *et al*. *"Dependability: Basic Concepts and Terminology*)

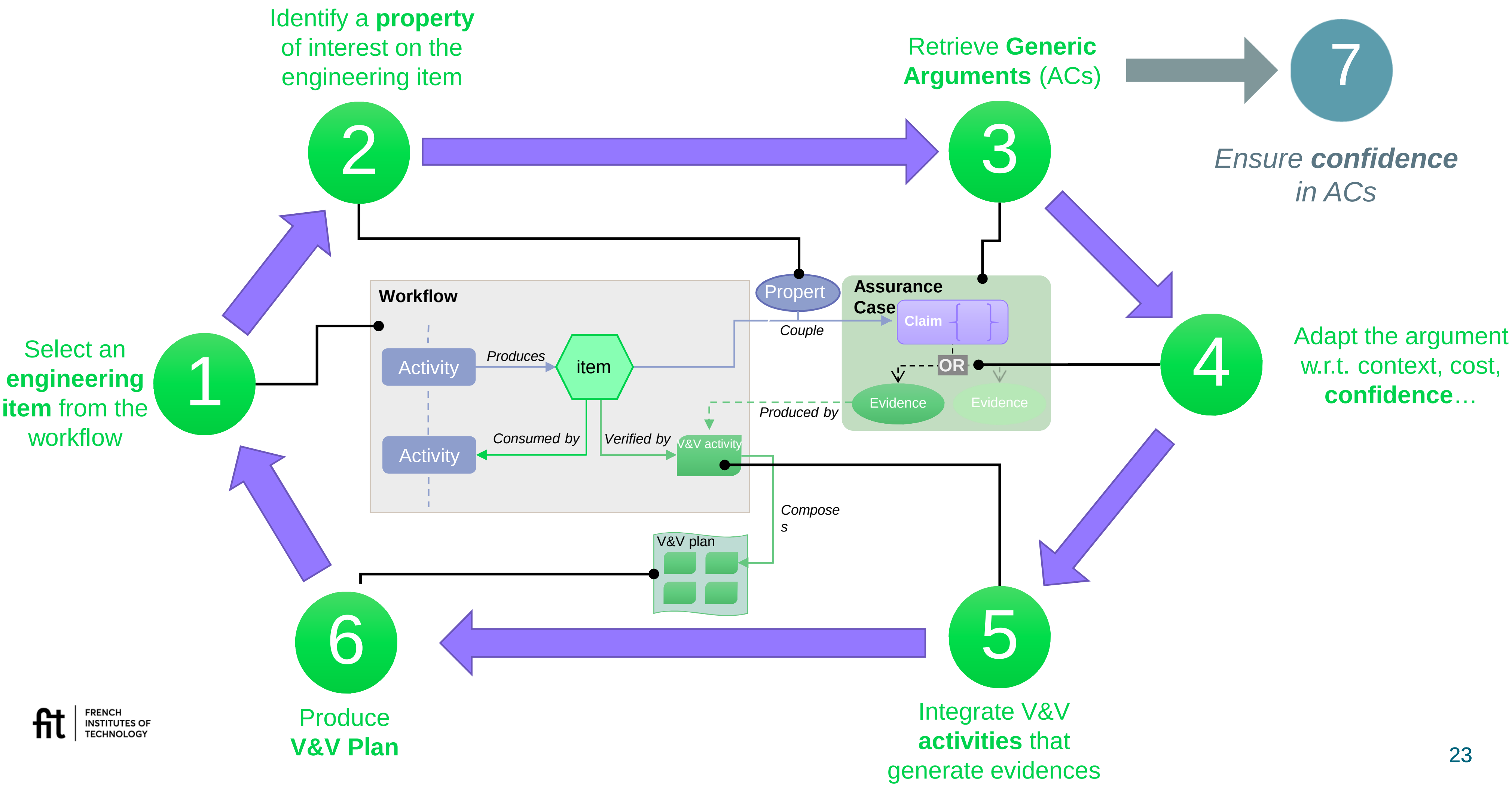
"[…] **claims, argument, and evidence is surely the (perhaps tacit)** intellectual foundation of any rational means for assuring and certifying the safety or other critical property of any kind of system. However, assurance cases differ from other means of assurance, such as those based on standards or guidelines, by making **all three components explicit**." (J.Rushby)

Robustness AC Template
ACs and Goal Structuring Notation

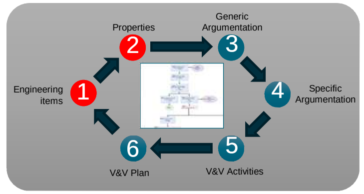
- AC formalism uses a set of concepts and notations (cf. GSN 3):
  - **Goal (& subgoals)**: affirmation that shall be assessed during the reasoning.
  - **Solution**: A solution refers to some evidence that is deemed sufficient to establish the truth of the parent claim
  - **Strategy**: justifies the decomposition of goals into sub-goals.
  - And a few other elements (context, assumptions, etc.)

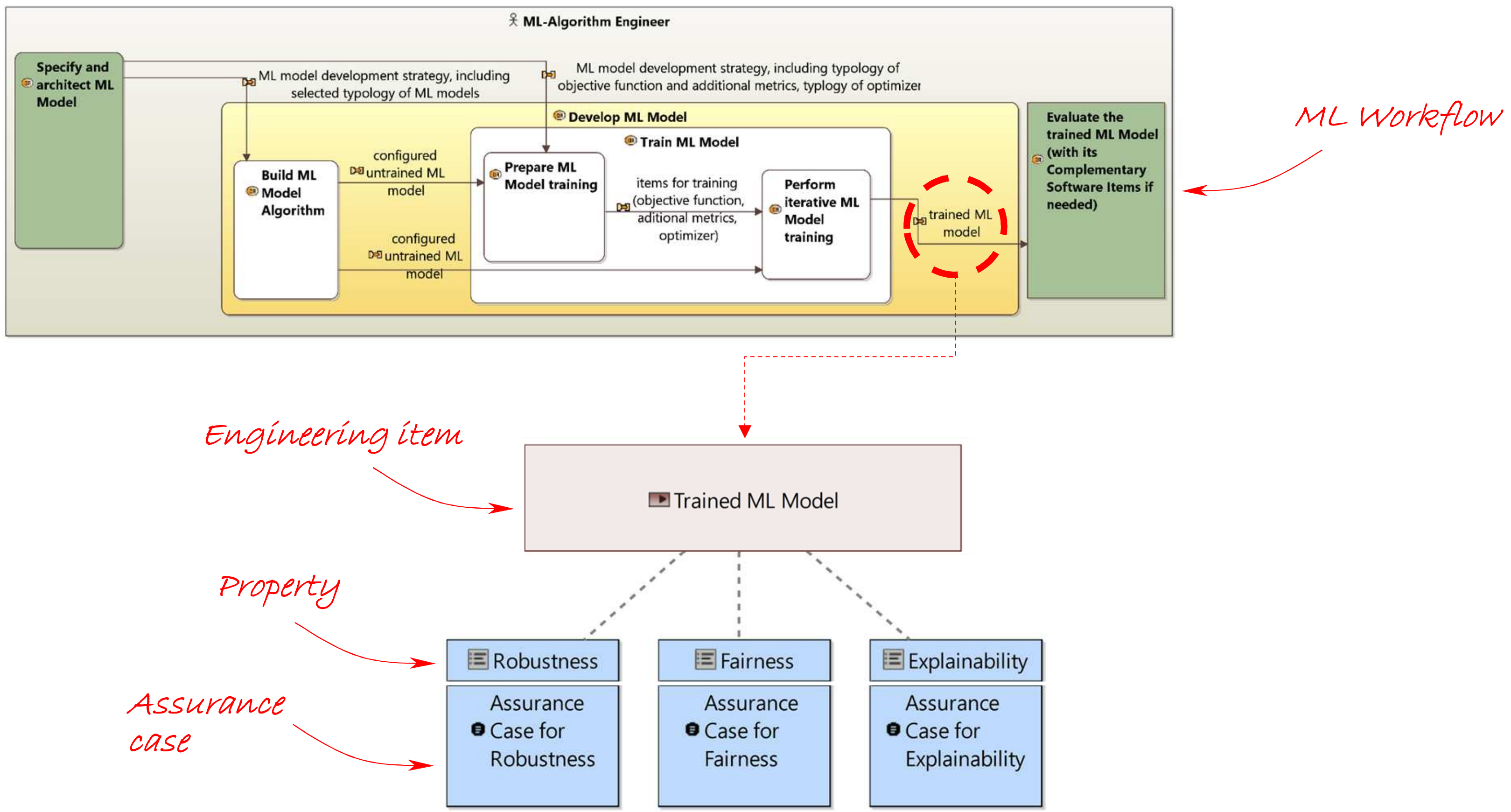


Global Approach

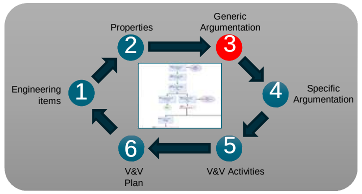


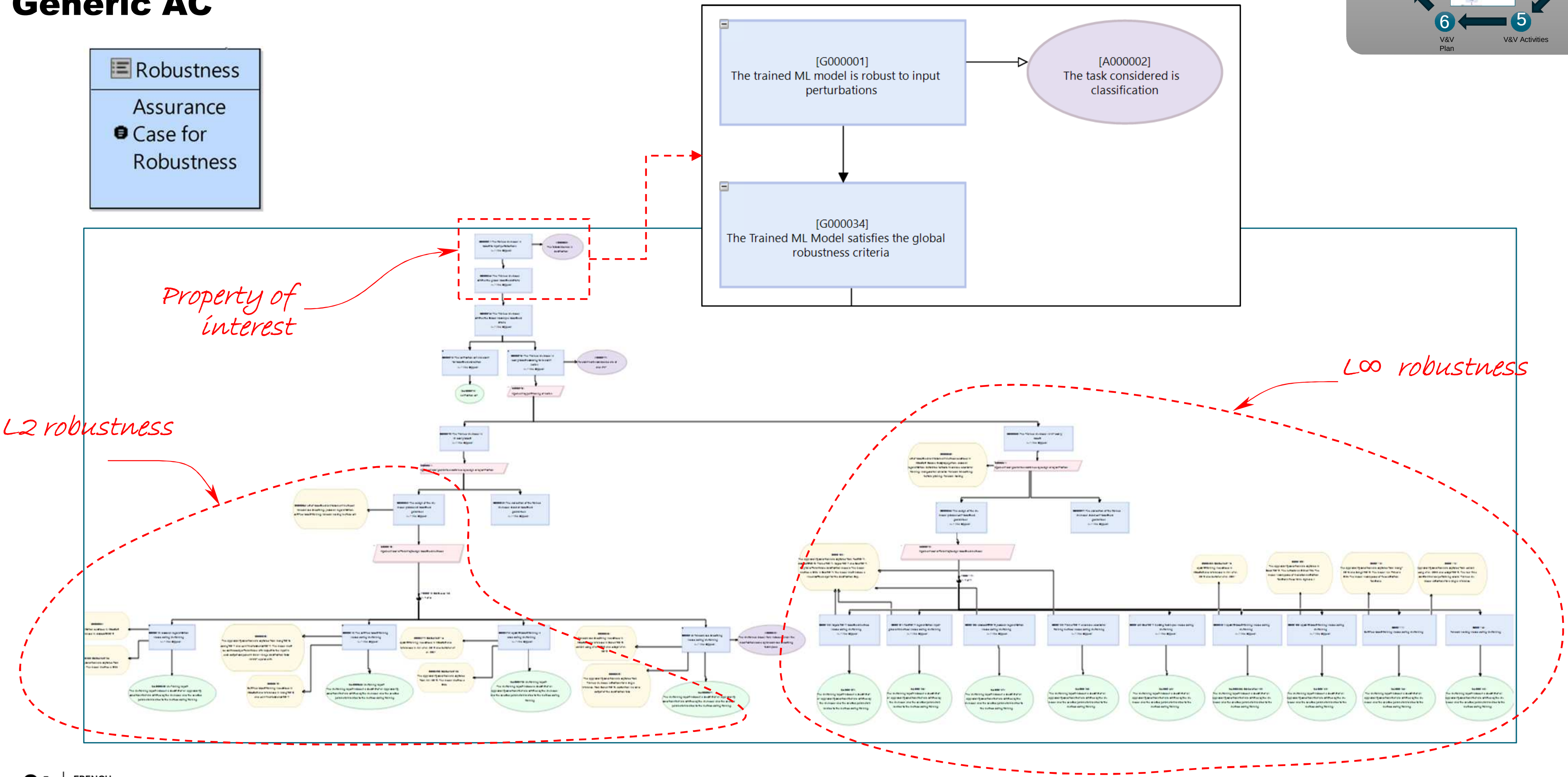
Robustness AC Template
From Engineering Items to Assurance Cases



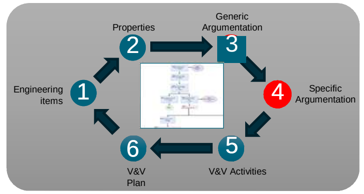


Robustness AC Template
Generic AC

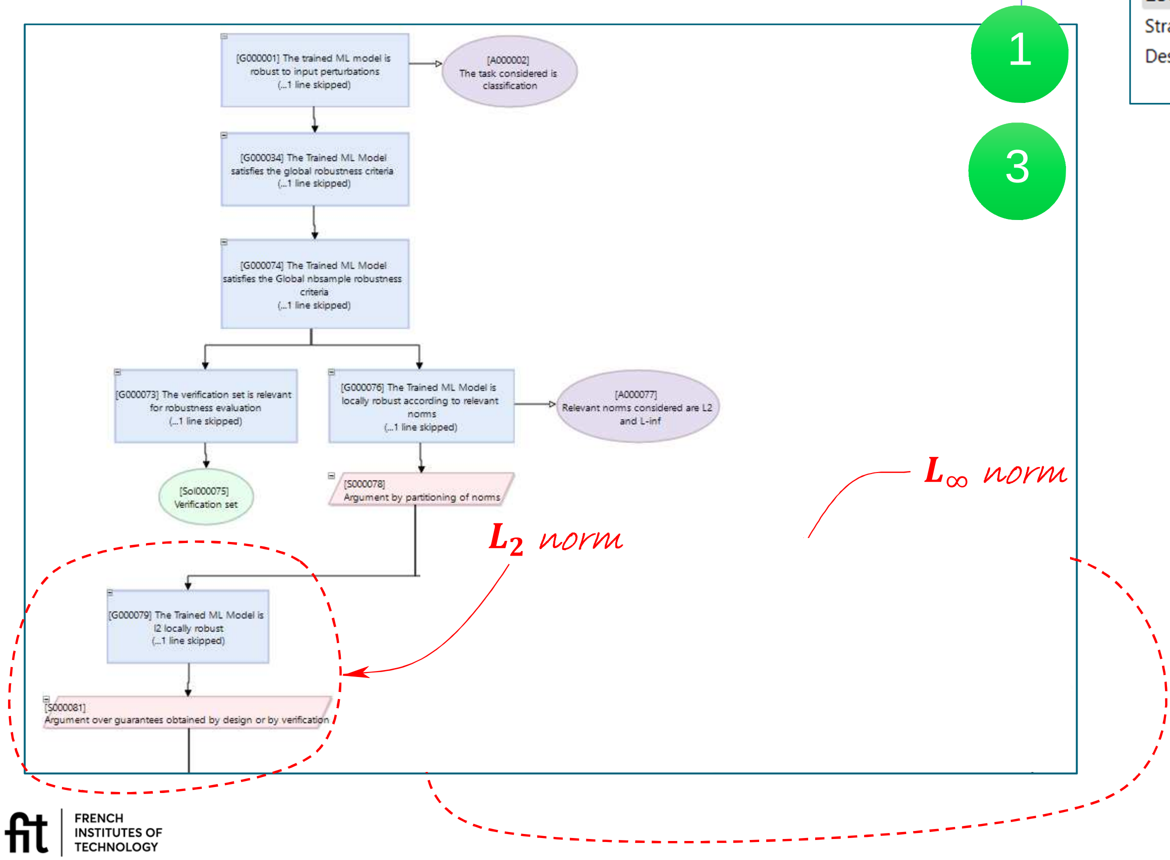


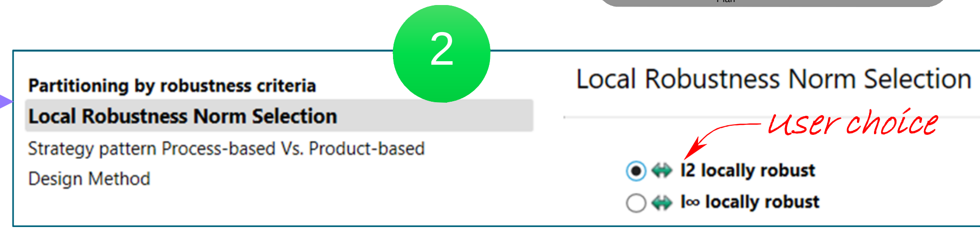


Robustness AC Template

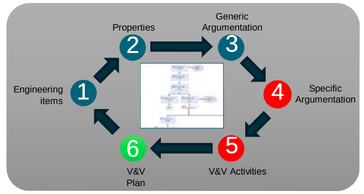


2. Partitioning by norms (only l2 and l∞ considered)

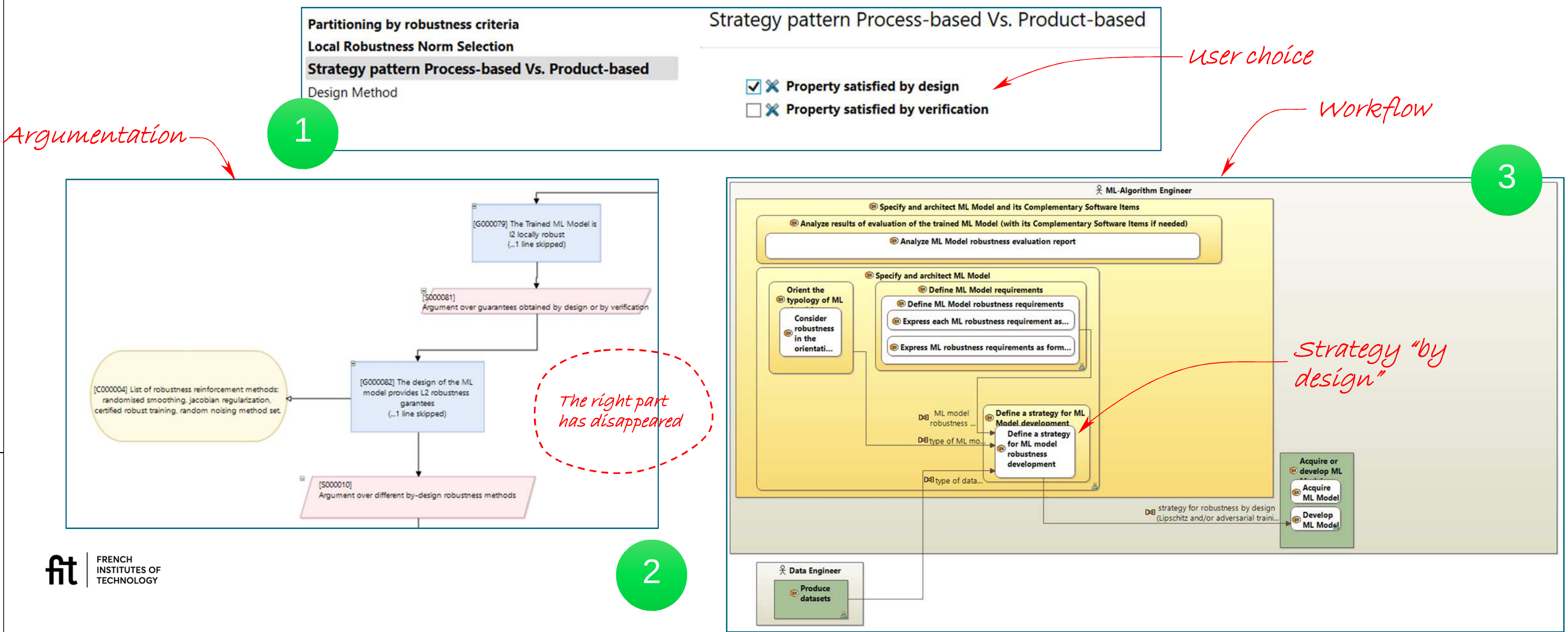




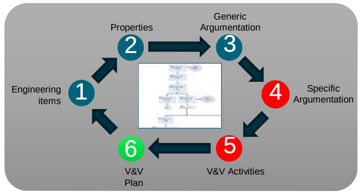
Robustness AC Template



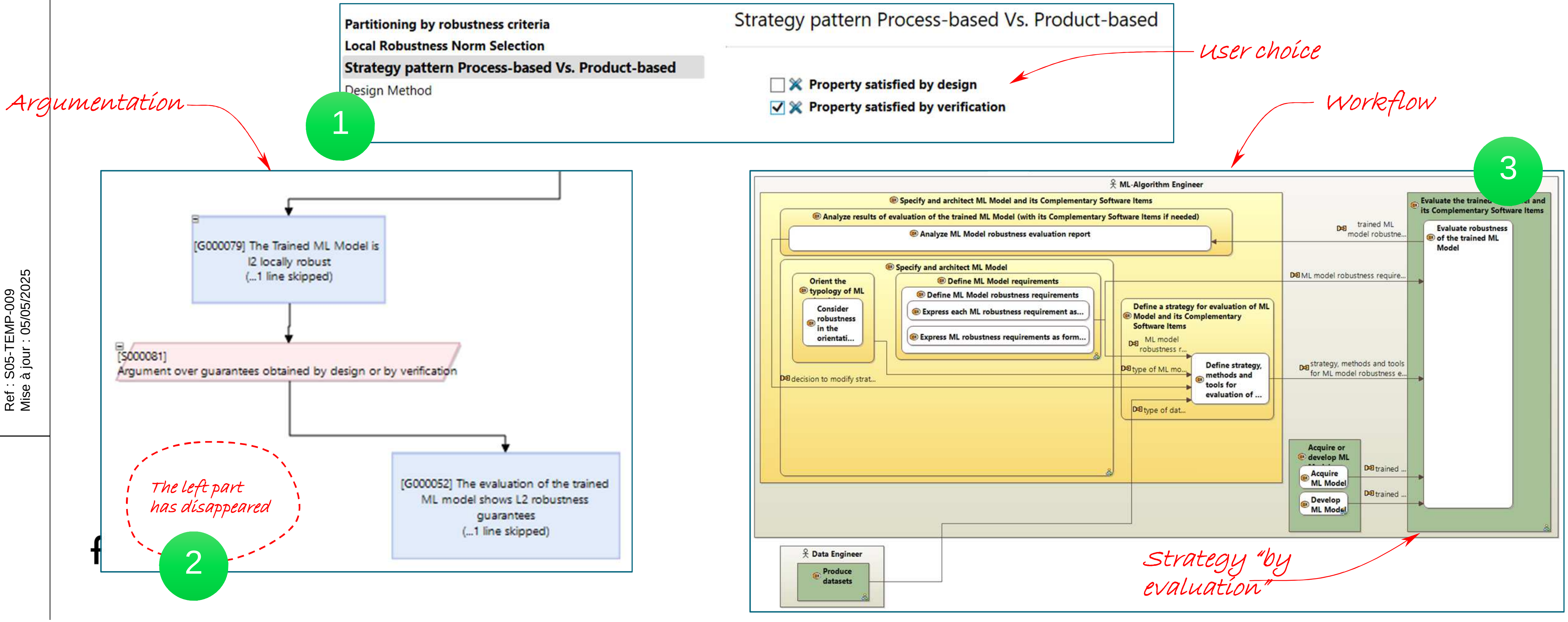
Strategy pattern Process-based (By Design) Vs. Product-based (By verification)



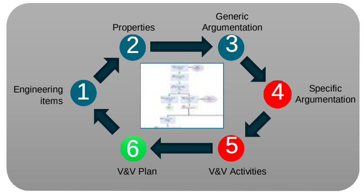
Robustness AC Template



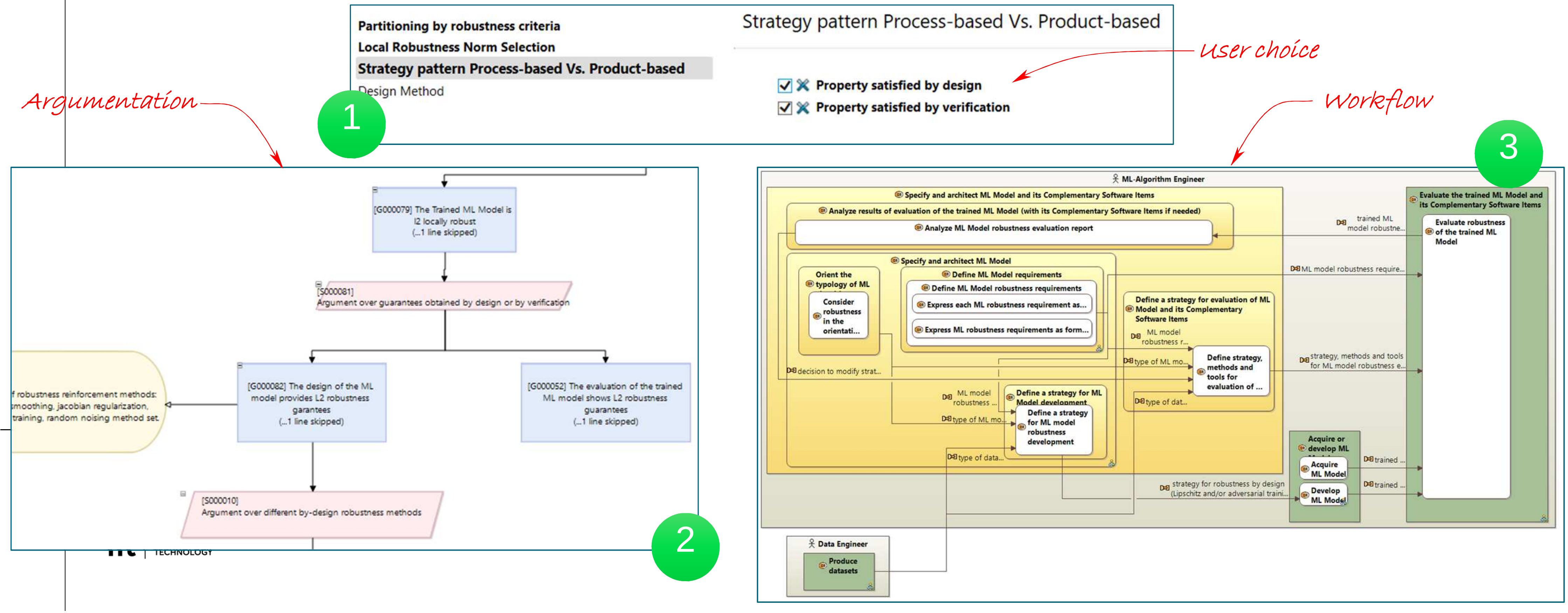
Strategy pattern Process-based (By Design) Vs. Product-based (By verification)



Robustness AC Template



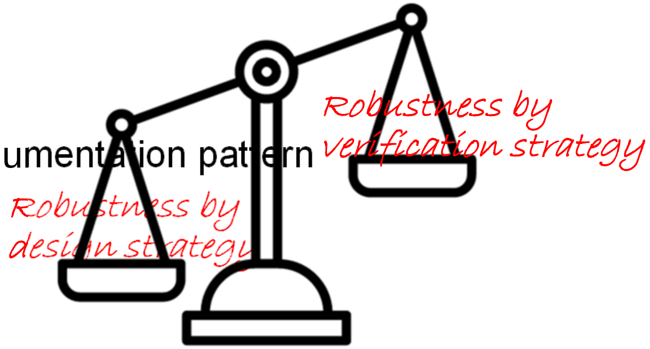
Strategy pattern Process-based (By Design) Vs. Product-based (By verification)



How to choose the argumentation strategy…

- Focus the validation effort on the most sensitive parts of the argumentation
  - Identify insufficiently convincing strategies associated to a goal
  - Identify contradiction between proof elements
  - Improve the argumentation

⇒ Estimate level of confidence in the argumentation pattern



What does confidence mean in our framework ?

SAINT EXUPÉRY

- Level of confidence ≈ Amount of information to justify a judgment about a proposition or, reciprocally, level of uncertainty about a judgment
  - Choice of an uncertainty representation
  - Elicitation of uncertainty associated to atomic elements
  - Propagation of uncertainty through the AC

Uncertainty is a general description of a state of knowledge that makes it difficult/impossible to assess the truth or the falsity of a piece of information (or a proposition).

- Complete information consists of what is known, and what is unknown (uncertainty/ignorance) about a proposition A:

$$Conf(A) + Uncer(A) = 1$$

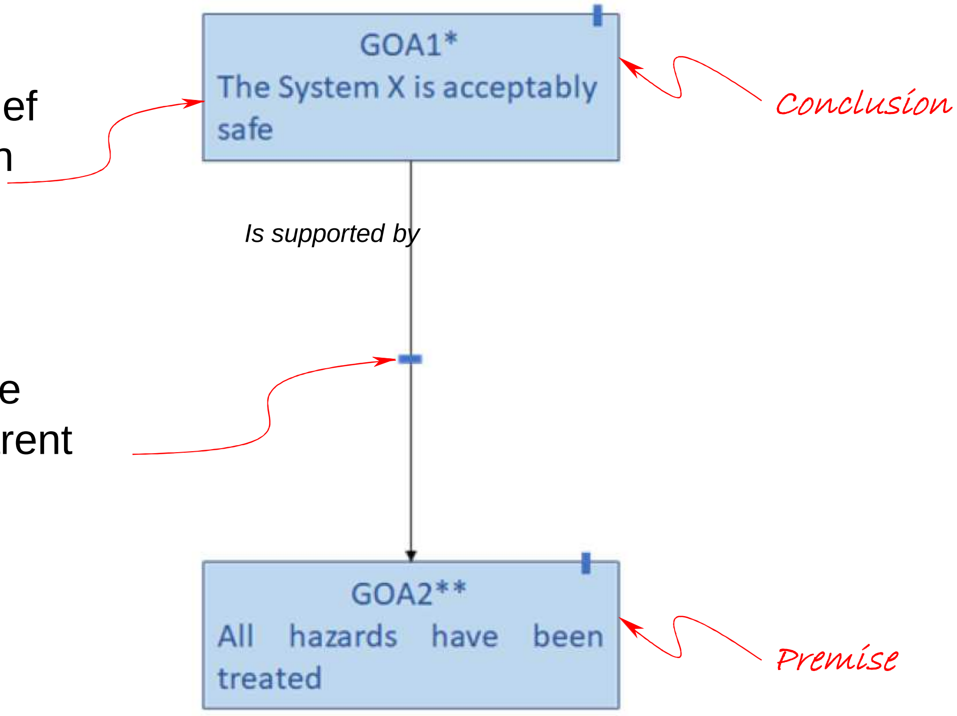
Sources of uncertainty in AC

- **Two aspects to estimate uncertainty**

  - **Trustworthiness** which quantifies the truth (with belief measures) and the falsity (with disbelief measures) in propositions (i.e., goals).

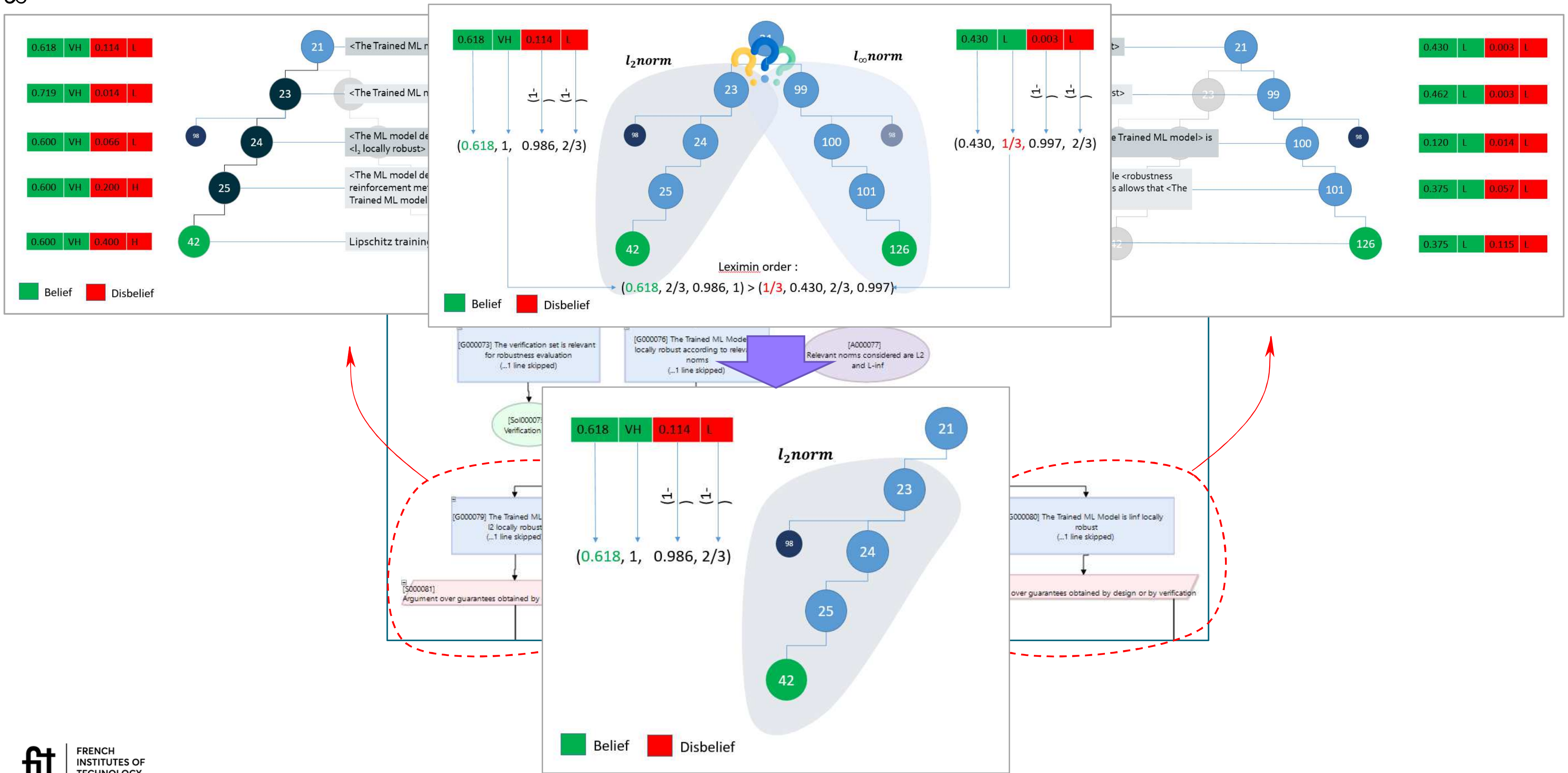
$$Conf(G) = Bel(G) + Disb(G)$$

  - **Appropriateness** which quantifies the truth about the inference (i.e., “supported by” relation) between a parent goal and its child goal(s).



UNCERTAINTY IN THE AC PATTERN FOR ROBUSTNESS OF ML – PROPAGATION RESULTS

$l_\infty norm$



Status

- A method to link design and argumentation
- A tool (Capella plugin using pure::variant) to implement the method
- A method to evaluate confidence in the argumentation

V. Mussot *et al.*, 'Assurance Cases to face the complexity of ML-based systems verification', in *Embedded Real Time System Congress, ERTS'24*, Toulouse, France, June 2024. Accessed: Sept. 03, 2025. [Online]. Available: https://hal.science/hal-04588599

Y. I. Messaoud, J.-L. Farges, E. Jenn, and V. Mussot, 'Uncertainty in Assurance Case Pattern for Machine Learning', in *Embedded Real Time System Congress, ERTS'24*, Toulouse, France, June 2024. [Online]. Available: https://hal.science/hal-04584490v1/document

FRENCH INSTITUTES OF TECHNOLOGY

Ref : S05-TEMP-009
Mise à jour : 05/05/2025

12/12/2025

FRENCH INSTITUTES OF TECHNOLOGY