Des patrons pour la certification d'IA embarquable

Journée IE + INFORSID

29/03/2024

Bernard BOTELLA [(1)], Florent CHENEVIER [(2)], Stephen CREFF [(3)], Jean-Loup FARGES [(4)], Anthony FERNANDES PIRES [(4)]
Ramon CONEJO LAGUNA [(5)], **Eric JENN [(5)]**, Florent LATOMBE [(6)], Yassir ID MESSAOUD [(3)], Vincent MUSSOT[(5)]

[(1)] CEA, [(2)] Thales AVS, [(3)] IRT SystemX, [(4)] ONERA, [(5)] IRT Saint Exupéry, [(6)] Obeo

Agenda

- ❑ Context and Objectives
- ❑ The process: Assurance cases and the ML development workflow
- ❑ Uncertainty assessment
- ❑ Conclusion

FRENCH INSTITUTES OF TECHNOLOGY

Contex and Objectives

FRENCH INSTITUTES OF TECHNOLOGY

Context

The Confiance.ai Programme (www.confiance.ai)

TECHNOLOGY

- Reliability, robustness
- Lawfulness & Compliance
- Accuracy
- Security
- Safety

INTERACTIONS

- Transparency
- Explainability
- Accountability
- Oversight & control

ETHICS

- Fairness
- Privacy
- Diversity & inclusion
- Sustainability

45M€ budget, 5years (cur. Year 3)

To provide industrial companies with solutions enabling the development of new products and services based on trustworthy AI

FEDERATIVE ENVIRONMENT, METHODS, TOOLS AND USE CASES
Open / Interoperable / Maintained

- Design
- Evaluation
- Data and knowledge
- Embedded systems

QUALITY ASSURANCE – ENGINEERING

FRENCH INSTITUTES OF TECHNOLOGY

Context

The Confiance.ai Programme (www.confiance.ai)

- ❑ Program structure : 7 Engineering Challenges, 2021 => 2024

EC	Adressed Topic
EC#1	Integration & Use-Cases, (+ Trusted AI Devops environment)
EC#2	Process, methodology & Guidelines
EC#3	Characterization & Qualification of Trustworthy AI
EC#4	Design for Trustworthy AI @ Algo, Components & System levels
EC#5	Data, Information & knowledge engineering for trusted AI
EC#6	IVV&Q strategy toward homologation/certification
EC#7	Target Embedded AI

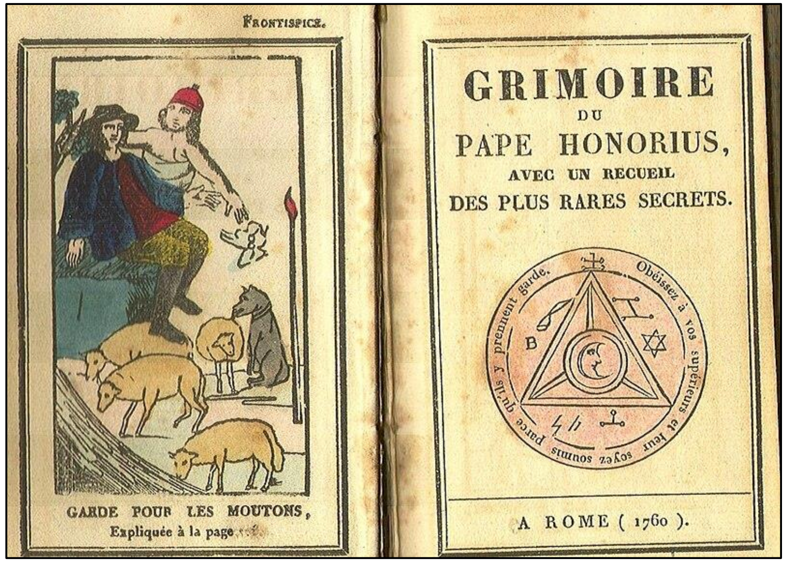
We are here

Context
Assurance Cases

Easy path #1

CONFIDENCE

Easy path #2



"Look at the book Chap. 3, Sec. 14, Vs 16"

*"Trusssssssssst me…
Trusssssssssst me…"*

FRENCH INSTITUTES OF TECHNOLOGY

Context
Assurance Cases

1 → 2 → 3 → 4

1. Dependability

“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*)

2. Confidence

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

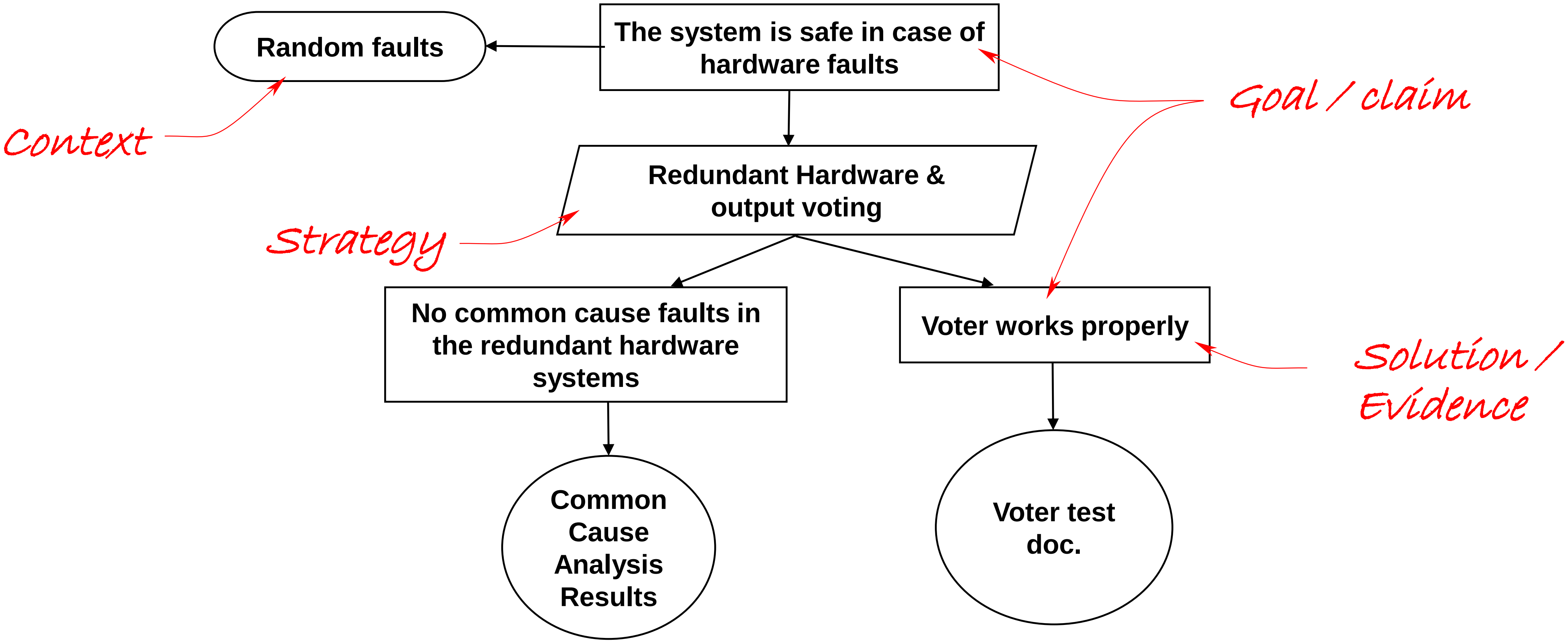
3. Emerging technology. No track record.

4. Assurance cases

“This framework of **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)

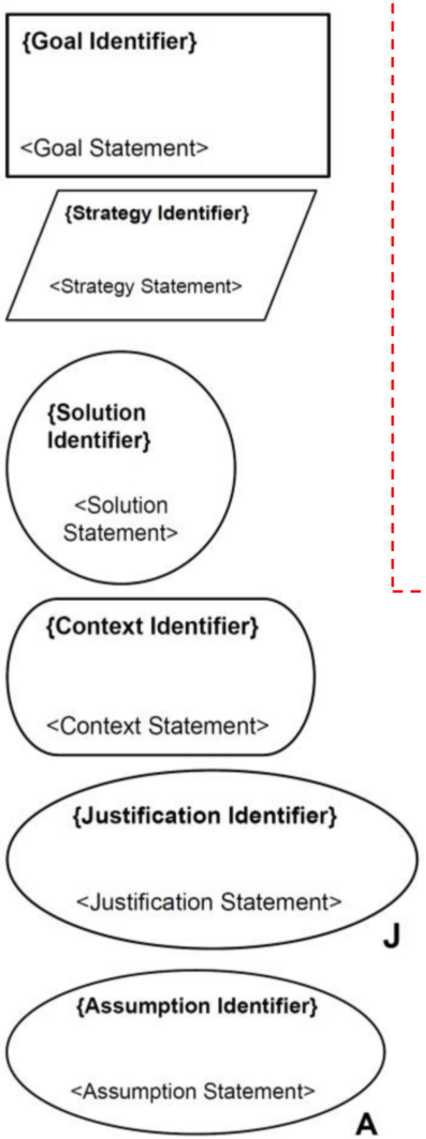
FRENCH INSTITUTES OF TECHNOLOGY

Assurance Cases
Main concepts



FRENCH INSTITUTES OF TECHNOLOGY

Assurance Cases
Main concepts



- **Goal** (& subgoals): affirmation that shall be assessed during the reasoning. Any goal may be refined in several subgoals.
- **Strategy**: justifies the decomposition of claims into sub-claims. It is an additional cue that helps the reader understand the form that an argument is going to take.
- **Solution**: A solution refers to some evidence that is deemed sufficient to establish the truth of the parent claim
- **Context**: define or constrain the scope over which the claim is made.
- **Justification**: describes why a given strategy is proposed as an approach to s upporting a particular goal, or provide reasons why the strategy being adopted is adequate.
- **Assumption**: statement about a property considered true. Assumptions must be valid for the related claim/strategy to be valid.

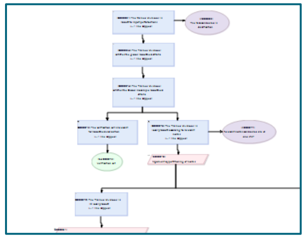
2 The process

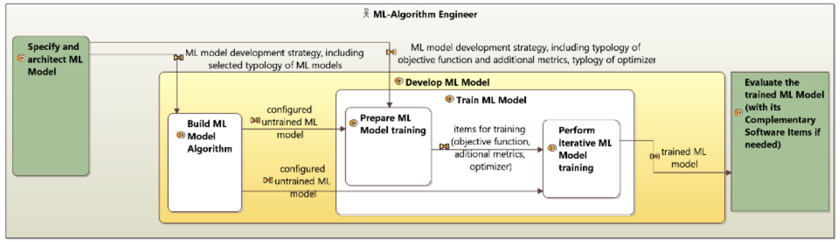
From Engineering Items to Assurance Cases

FRENCH INSTITUTES OF TECHNOLOGY

Process Overview

`[G] The <Trained ML Model> is <robust>`





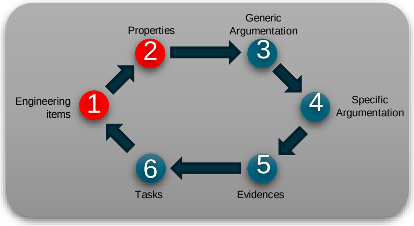
1. Select an engineering item from the ML workflow
2. Select a property of interest
3. Get generic argumentation (AC)
4. Adapt the argumentation wrt context, cost, confidence, etc.
5. Identify evidences to produce
6. Update workflow

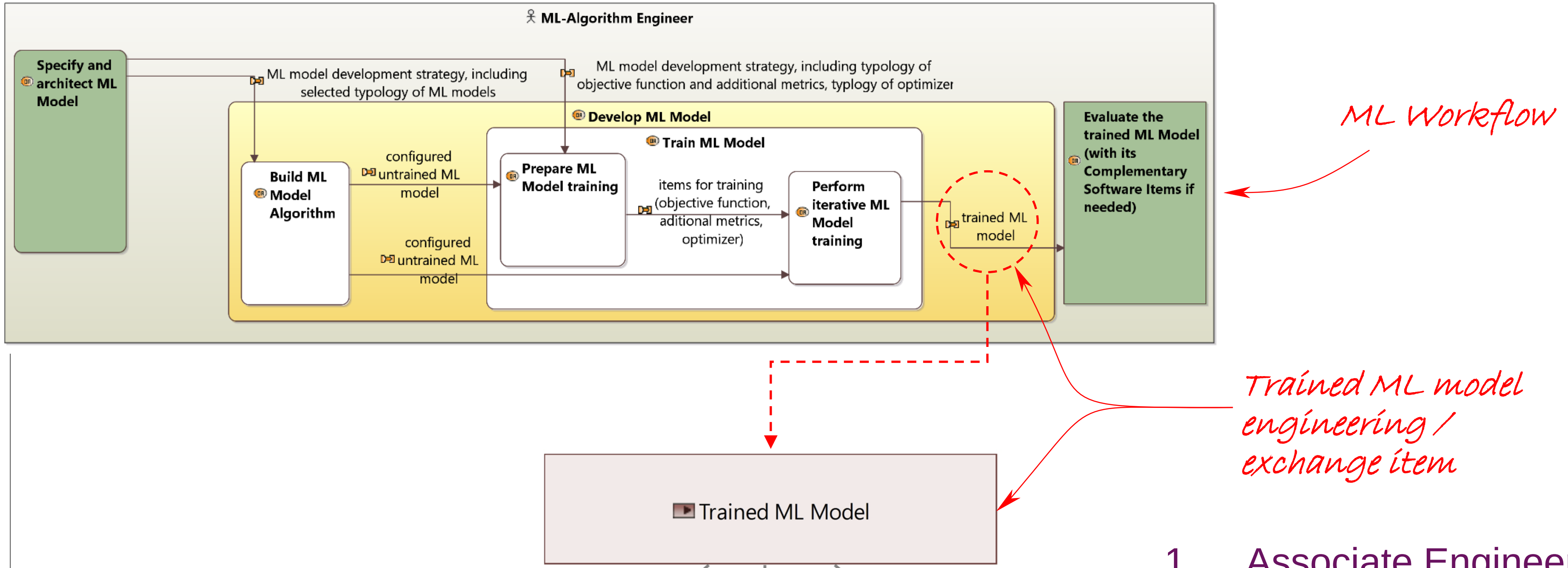
Capella

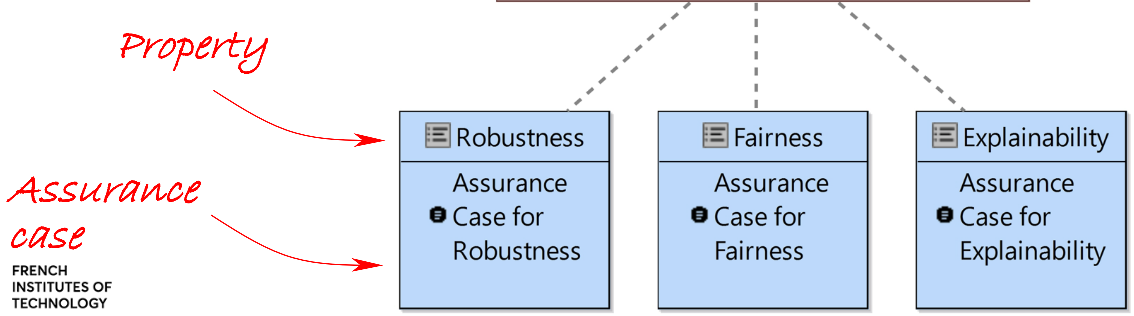
FRENCH INSTITUTES OF TECHNOLOGY

From Engineering Items to Assurance Cases

Robustness argumentation template







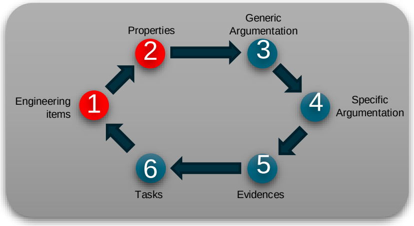
1. Associate Engineering Conditions (e.g. '*Robustness*') to Exchange items

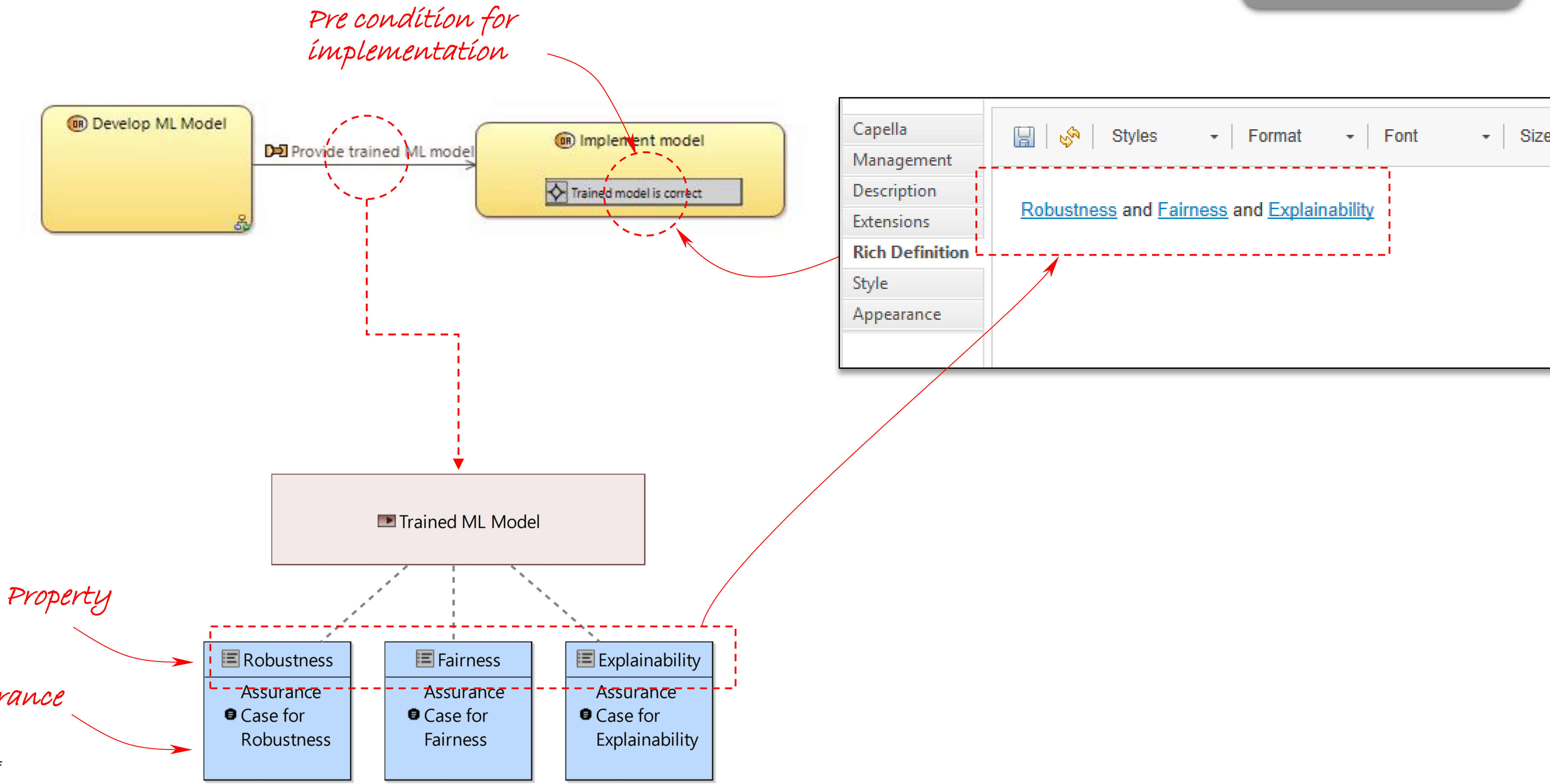
2. Define Assurance Cases for each condition

FRENCH INSTITUTES OF TECHNOLOGY

From Engineering Items to Assurance Cases

Robustness argumentation template

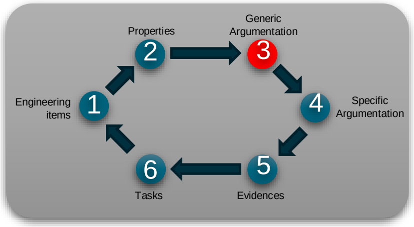


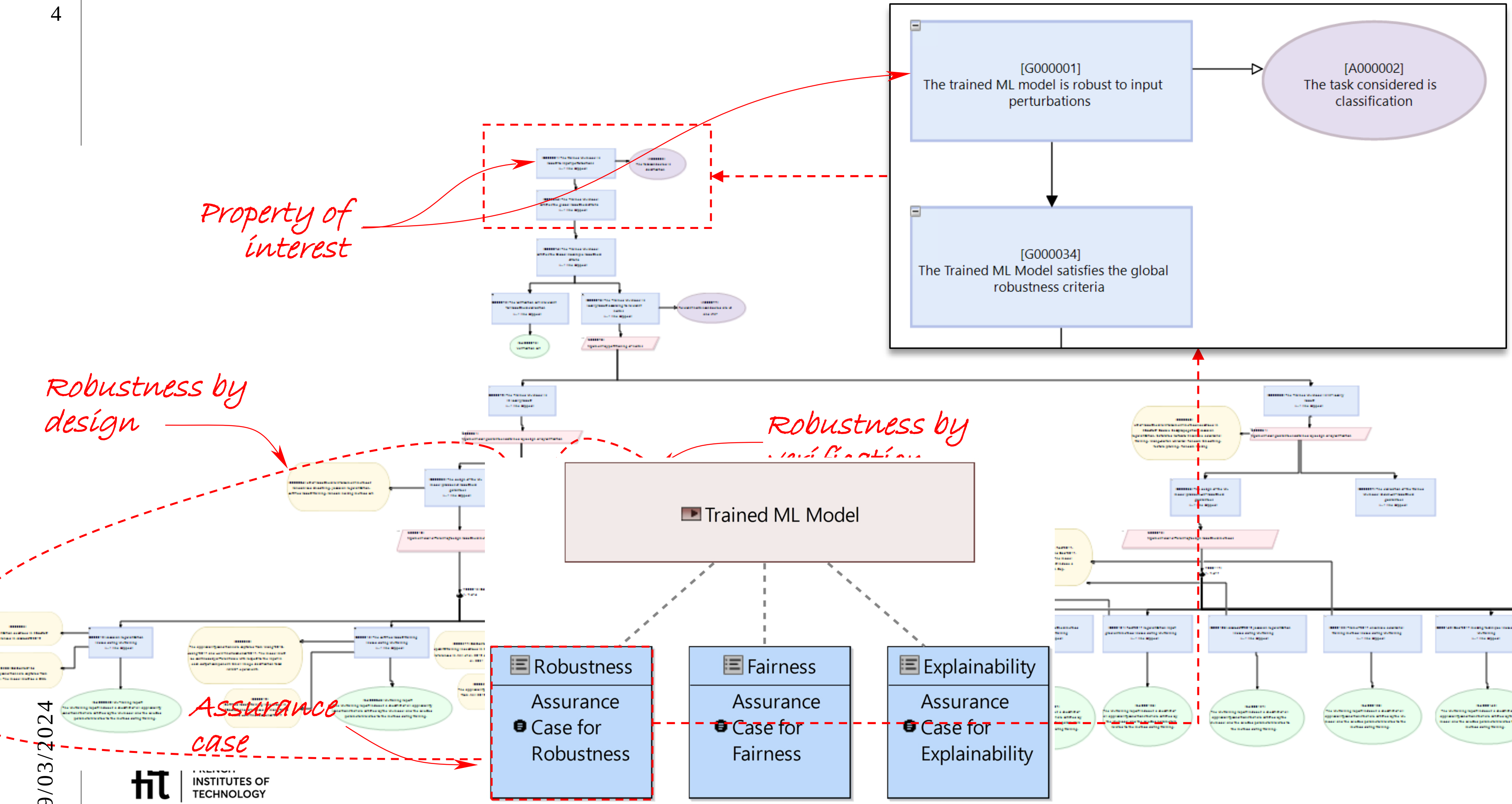


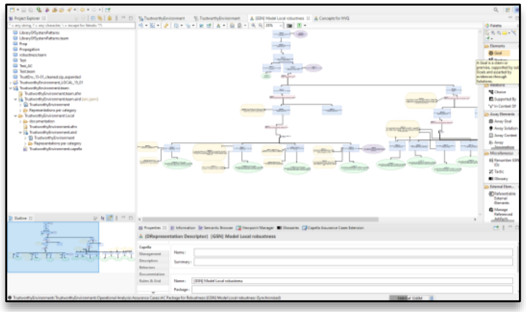
FRENCH INSTITUTES OF TECHNOLOGY

From Engineering Items to Assurance Cases

Robustness argumentation template





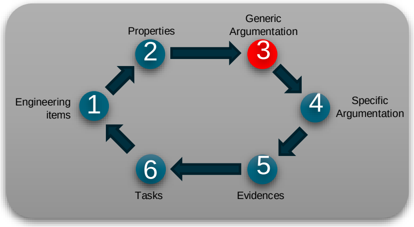


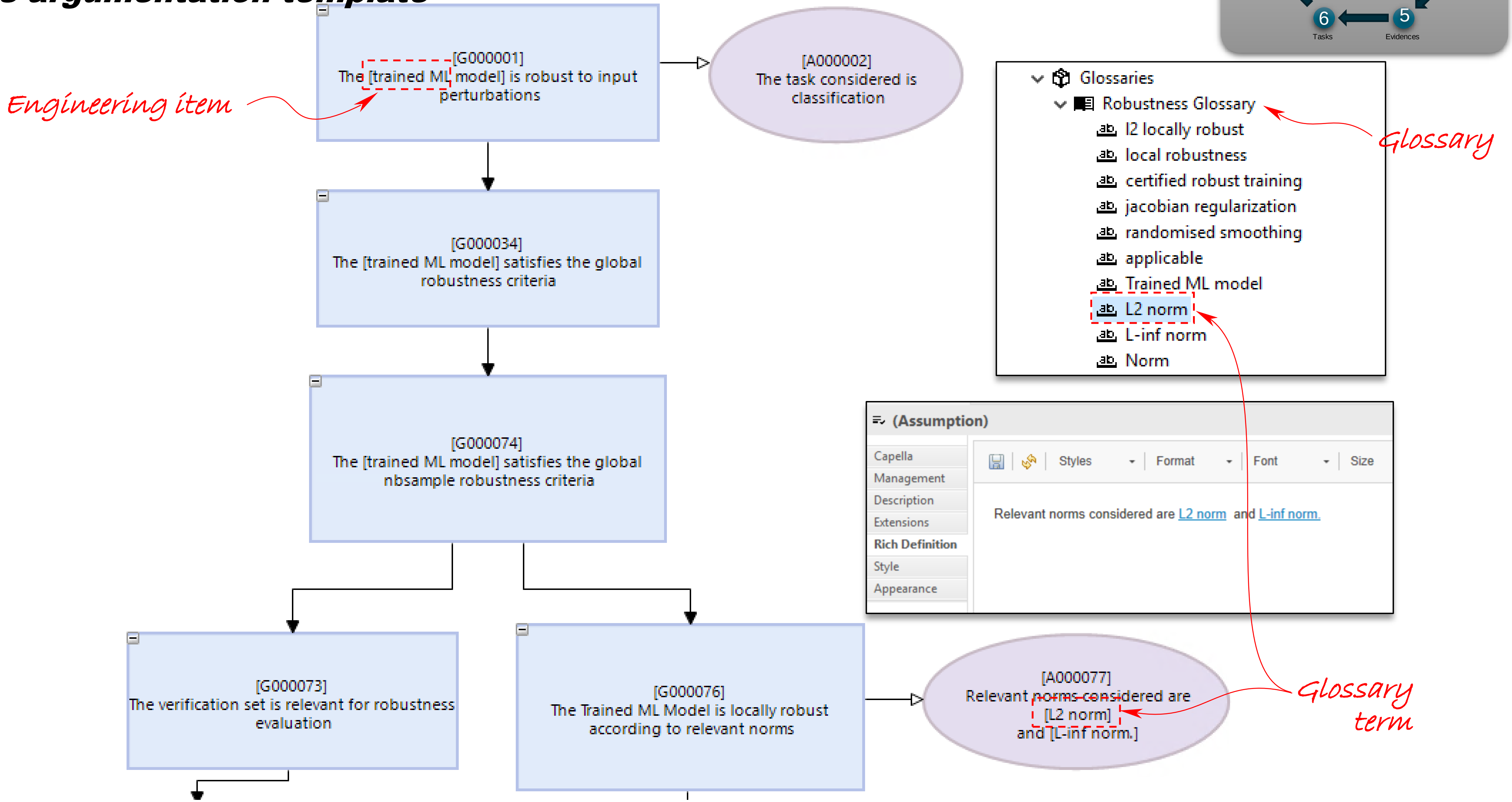
GSN extension in the Capella Editor

IRT INSTITUTES OF TECHNOLOGY

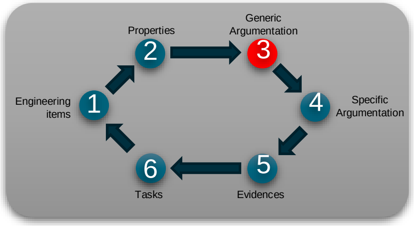
From Engineering Items to Assurance Cases

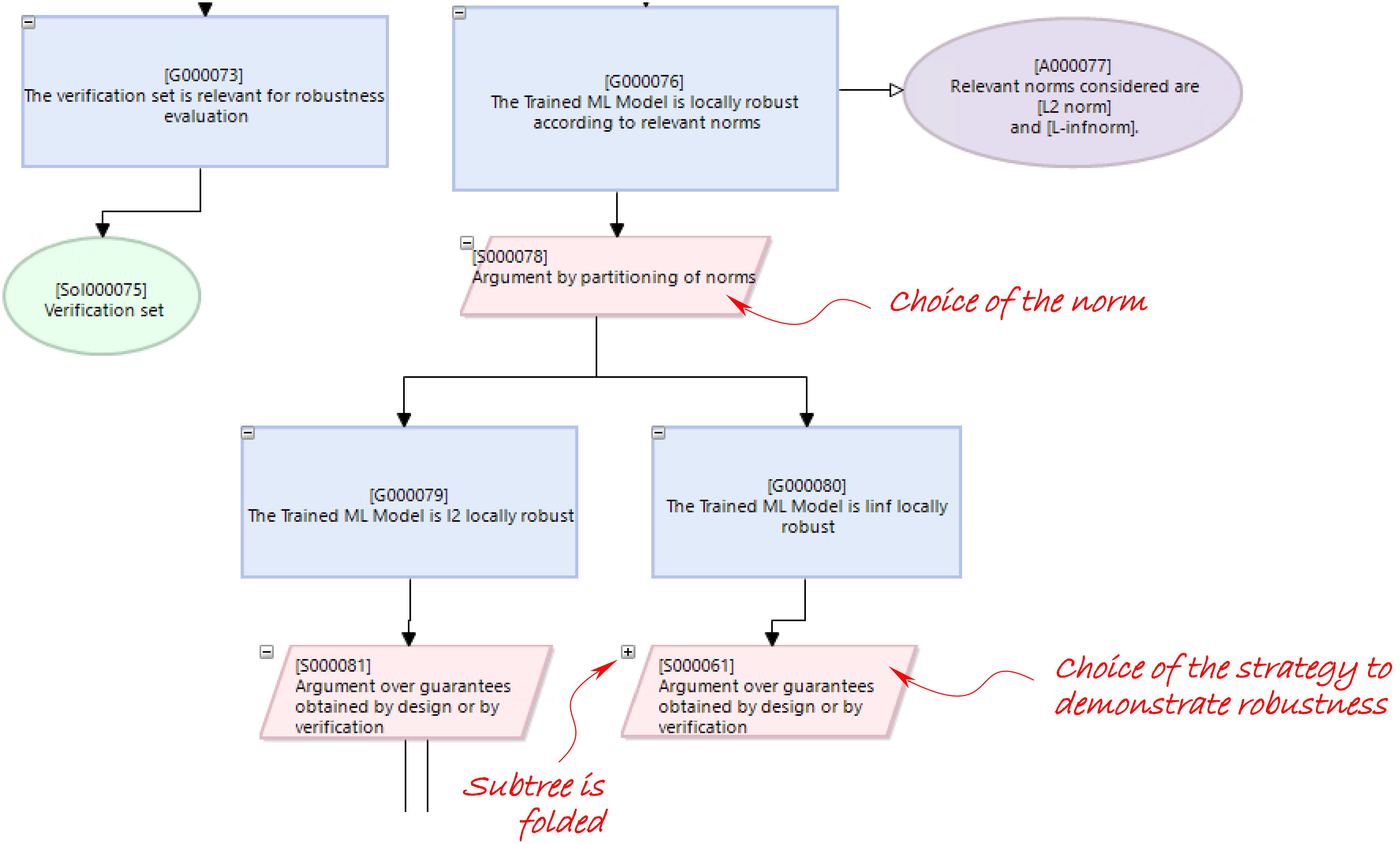
Robustness argumentation template





From Engineering Items to Assurance Cases
Robustness argumentation template

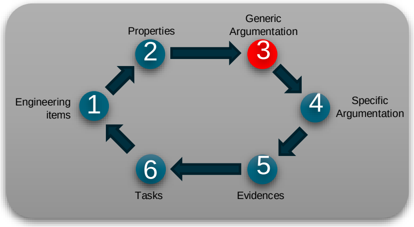


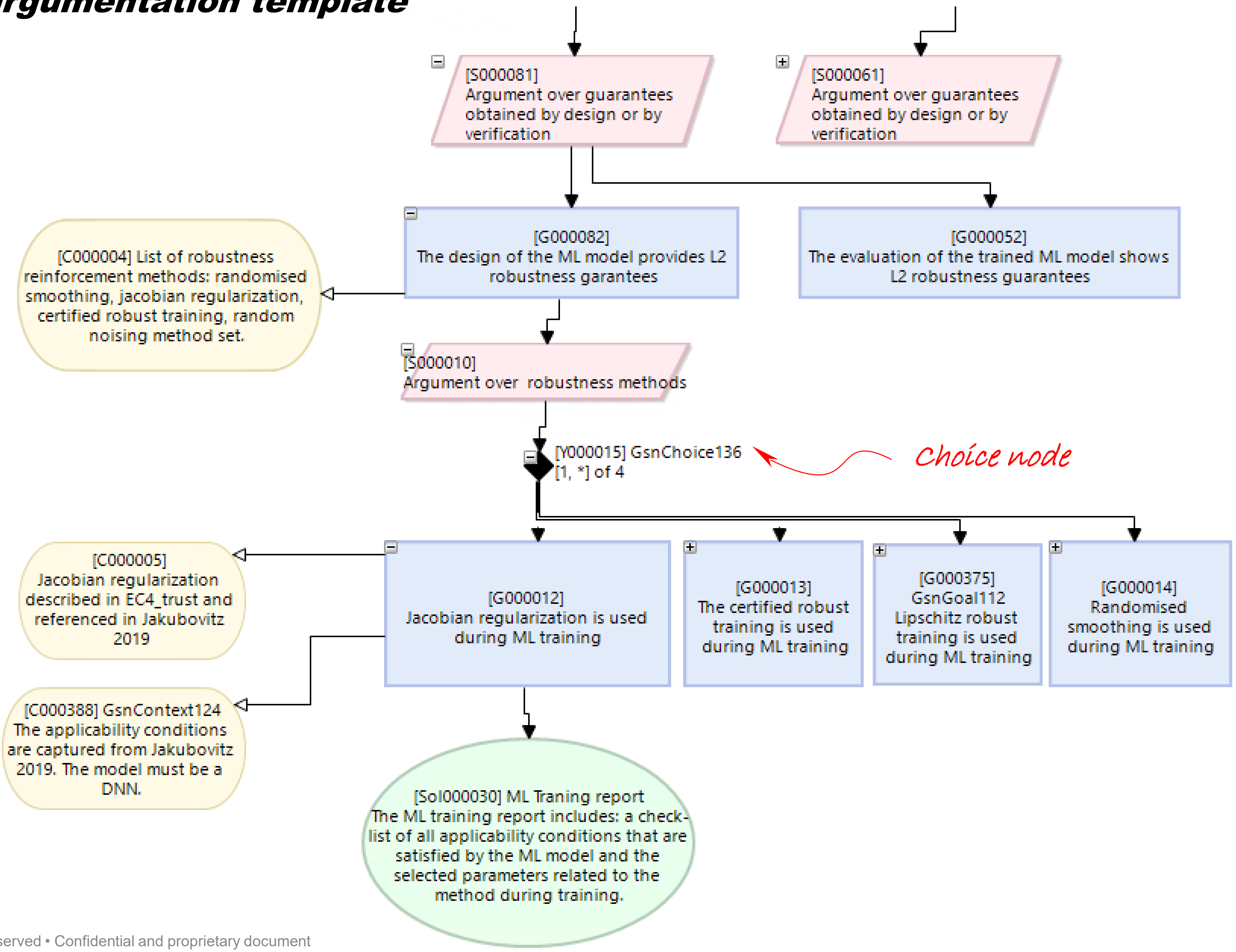


FRENCH INSTITUTES OF TECHNOLOGY

From Engineering Items to Assurance Cases

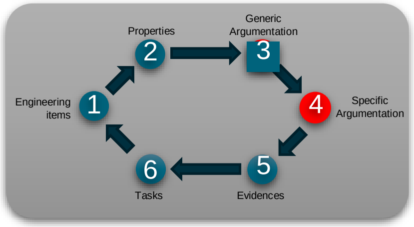
Robustness argumentation template





FRENCH INSTITUTES OF TECHNOLOGY

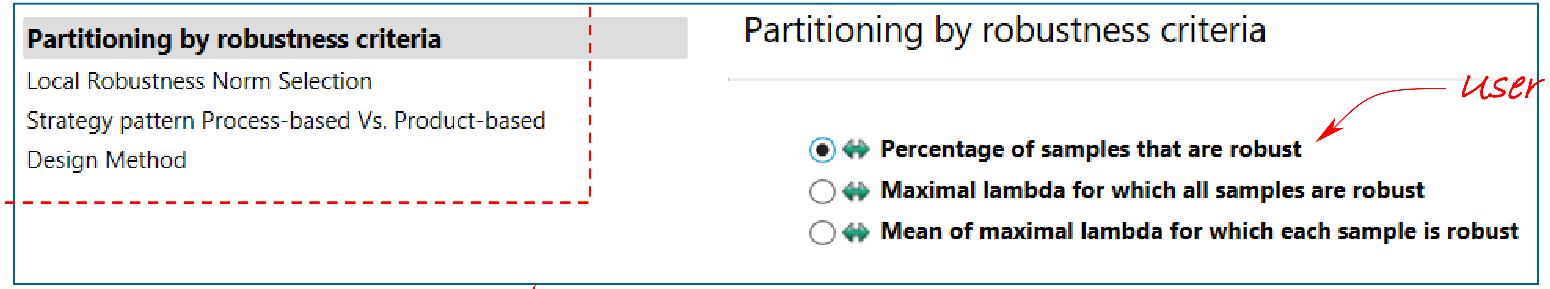
From Engineering Items to Assurance Cases
Refinement of requirements



1. Partitioning by robustness criteria
   - Percentage of samples that are robust
   - Maximal **lambda** for which all samples are robust
   - Mean of maximal **lambda** for which each sample is robust

Robusteness criteria

All choices



User choice

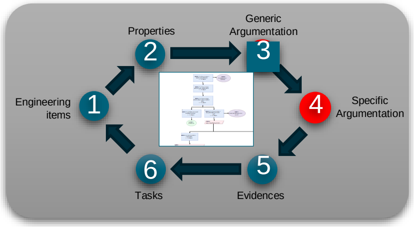
Configuration menu

Capella Pure::variant configuration wizard

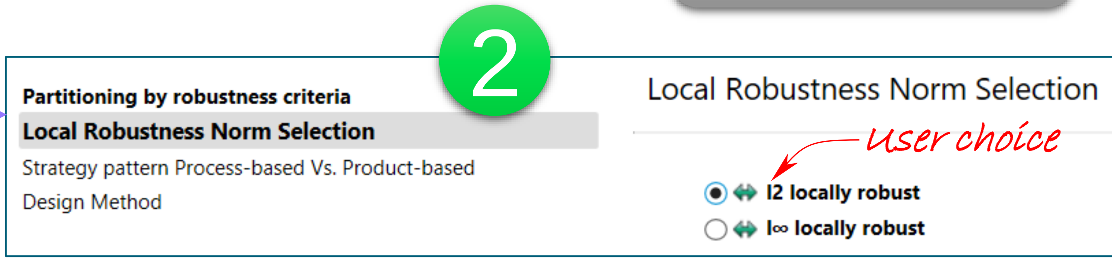
FRENCH INSTITUTES OF TECHNOLOGY

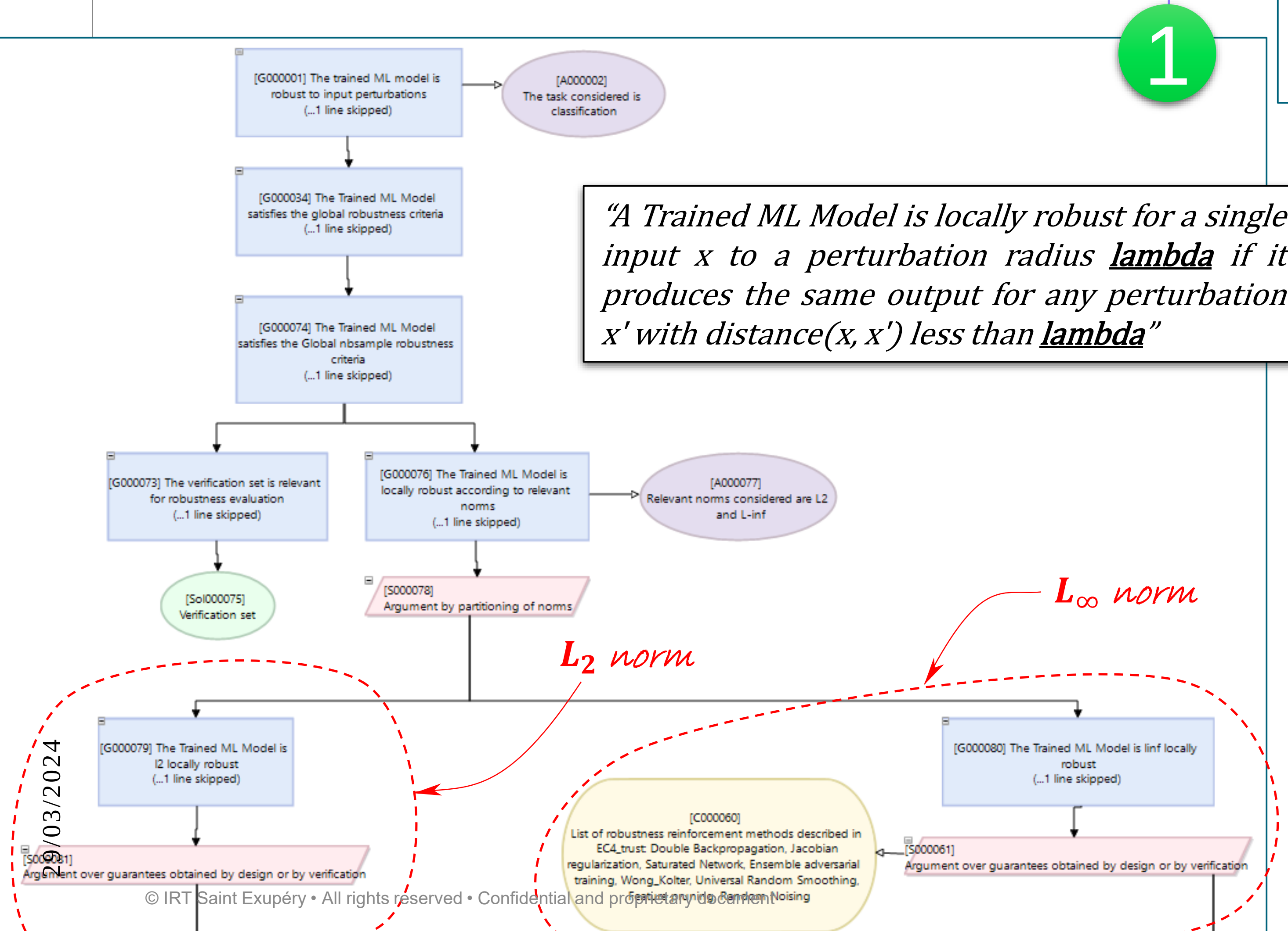
From Engineering Items to Assurance Cases

Refinement of requirements

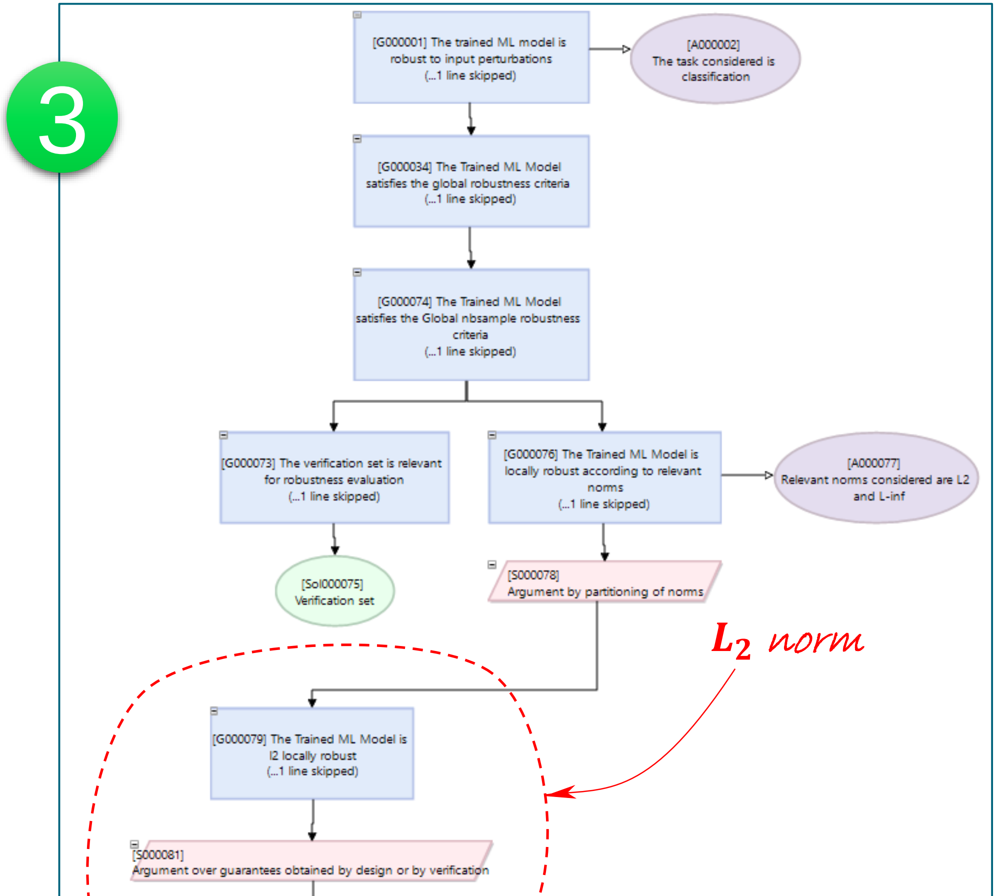


- ❑ Partitioning by norms (only l2 and l∞ considered)

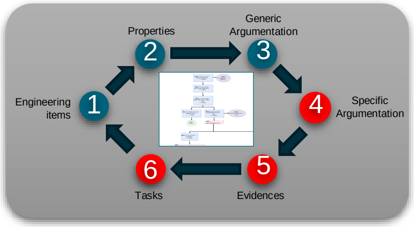




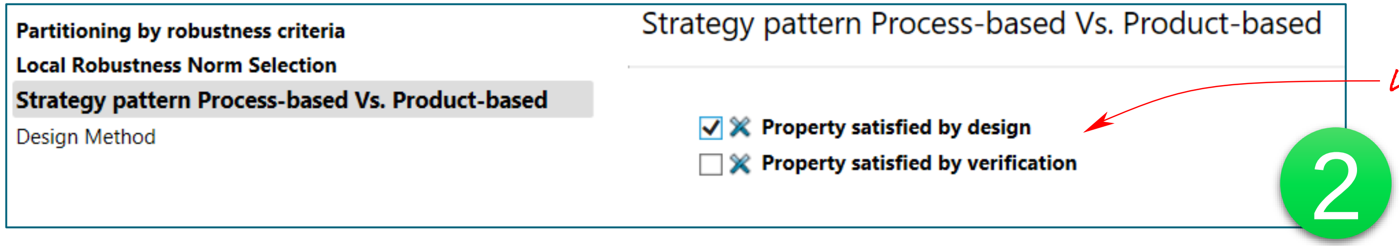
*"A Trained ML Model is locally robust for a single input x to a perturbation radius **lambda** if it produces the same output for any perturbation x' with distance(x, x') less than **lambda**"*



From Engineering Items to Assurance Cases



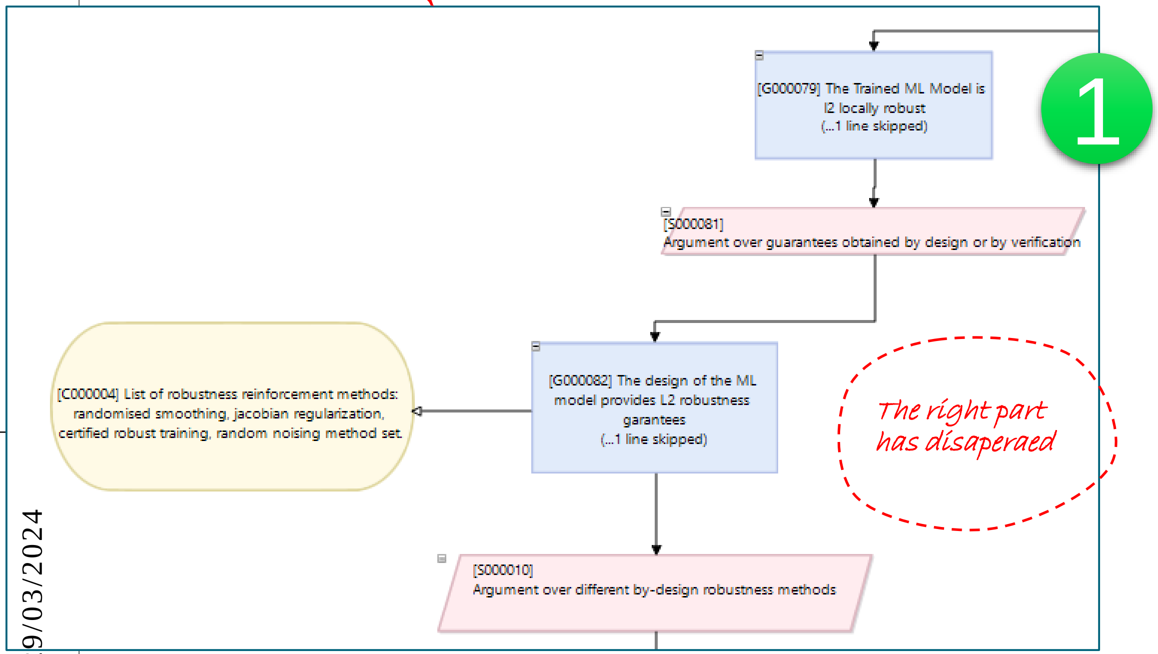
❑ Strategy pattern Process-based (By Design) Vs. Product-based (By verification)

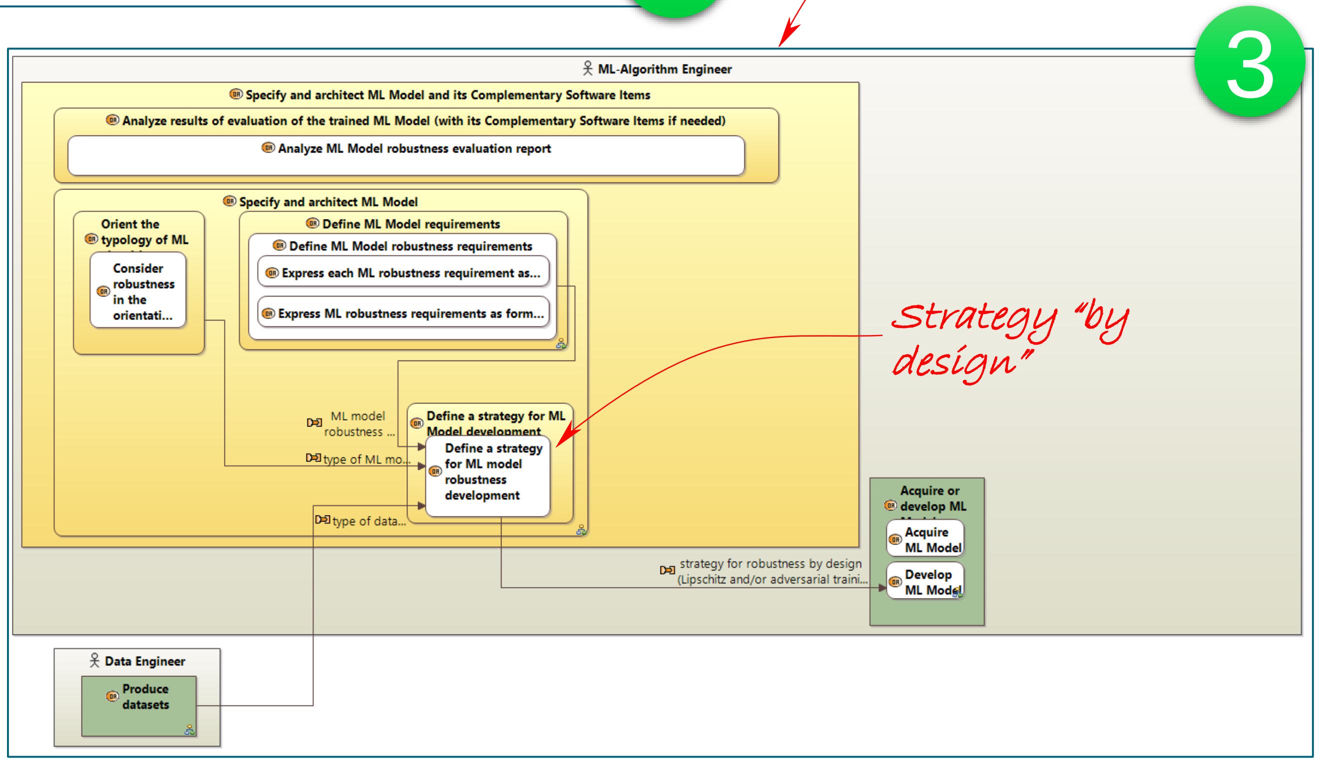


User choice

Argumentation

Workflow

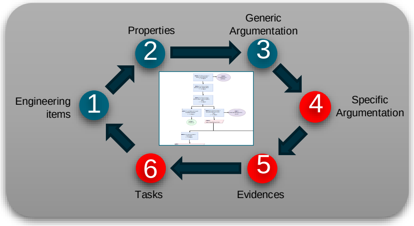




Strategy "by design"

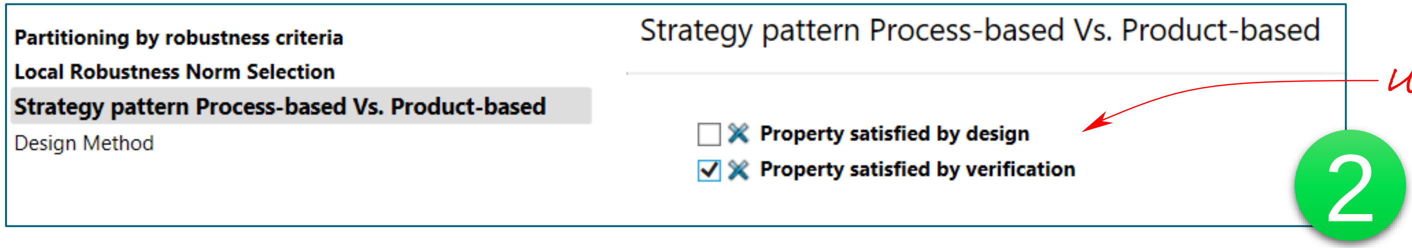
29/03/2024

Robustness AC Template



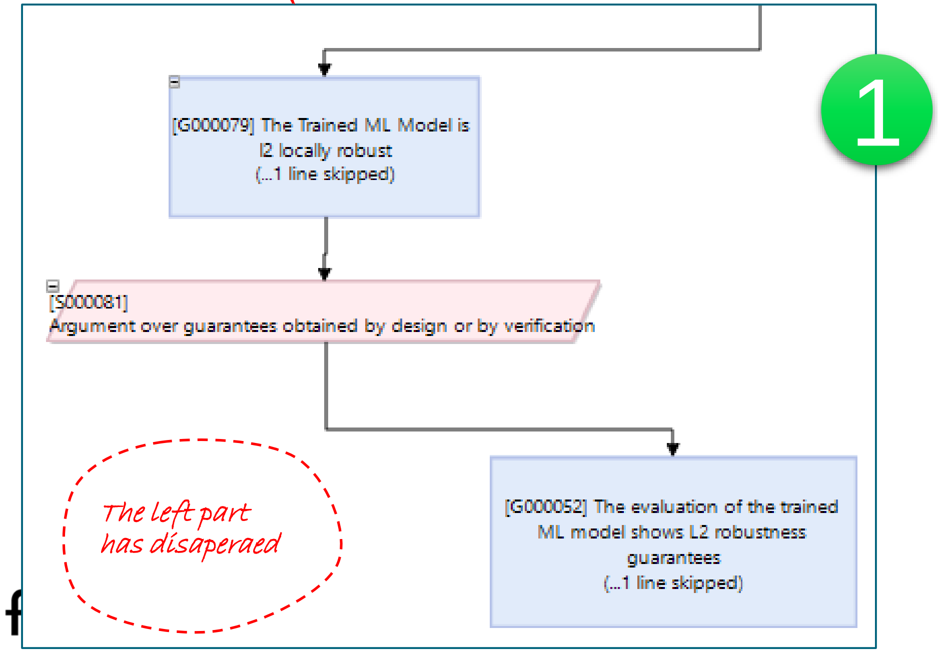
- ❑ Strategy pattern Process-based (By Design) Vs. Product-based (By verification)

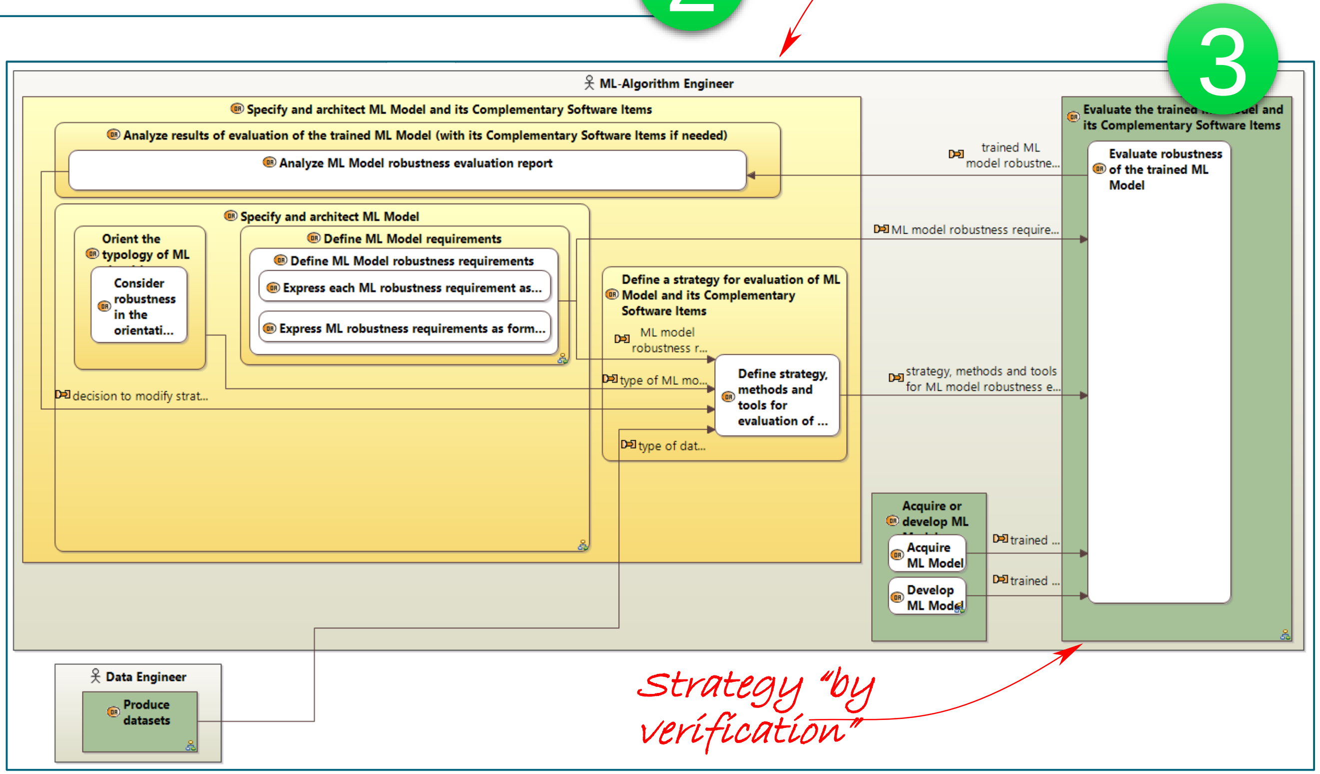
User choice



Argumentation

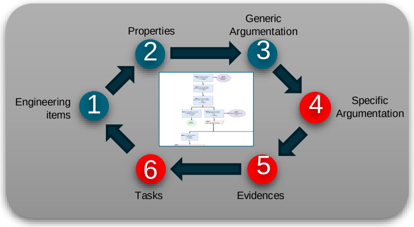
Workflow



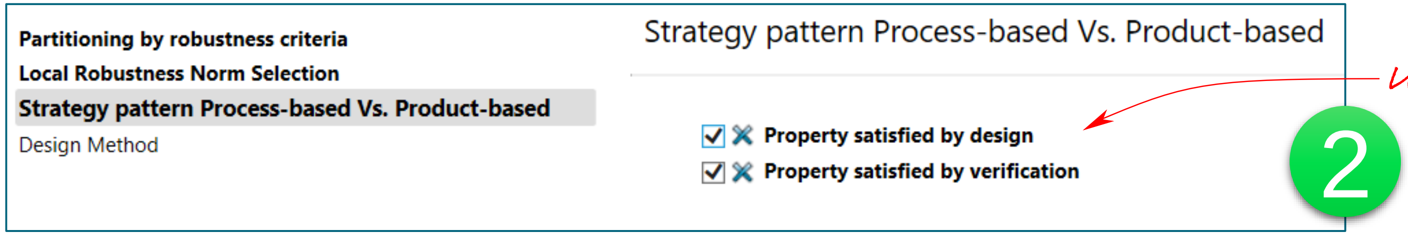


Strategy "by verification"

Robustness AC Template



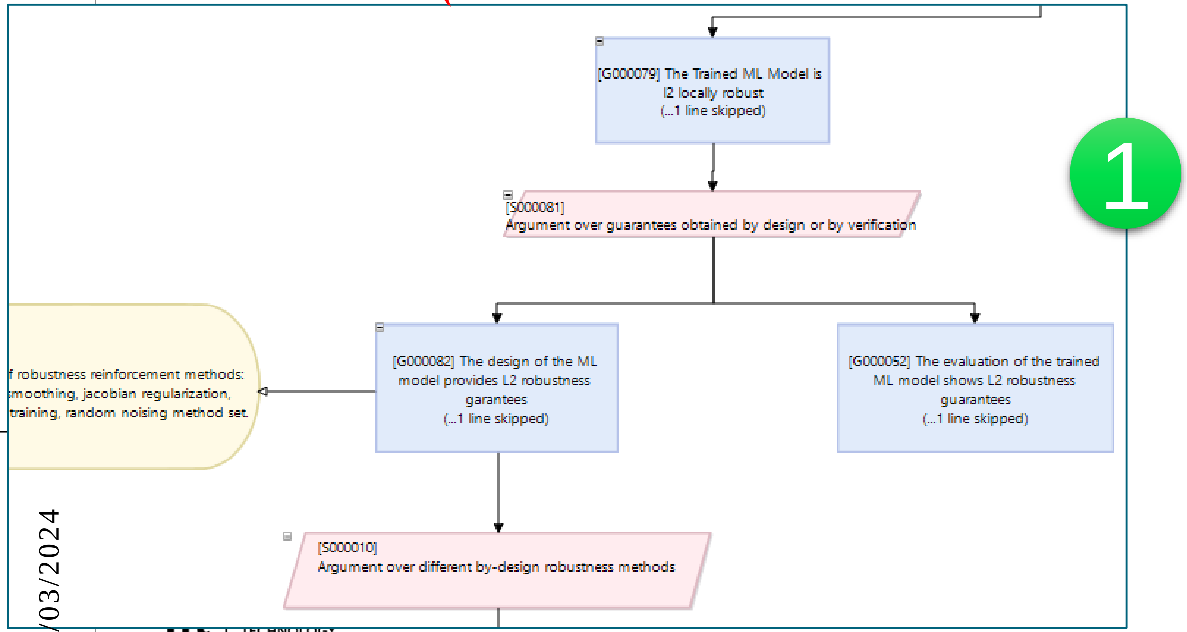
- ❑ Strategy pattern Process-based (By Design) Vs. Product-based (By verification)

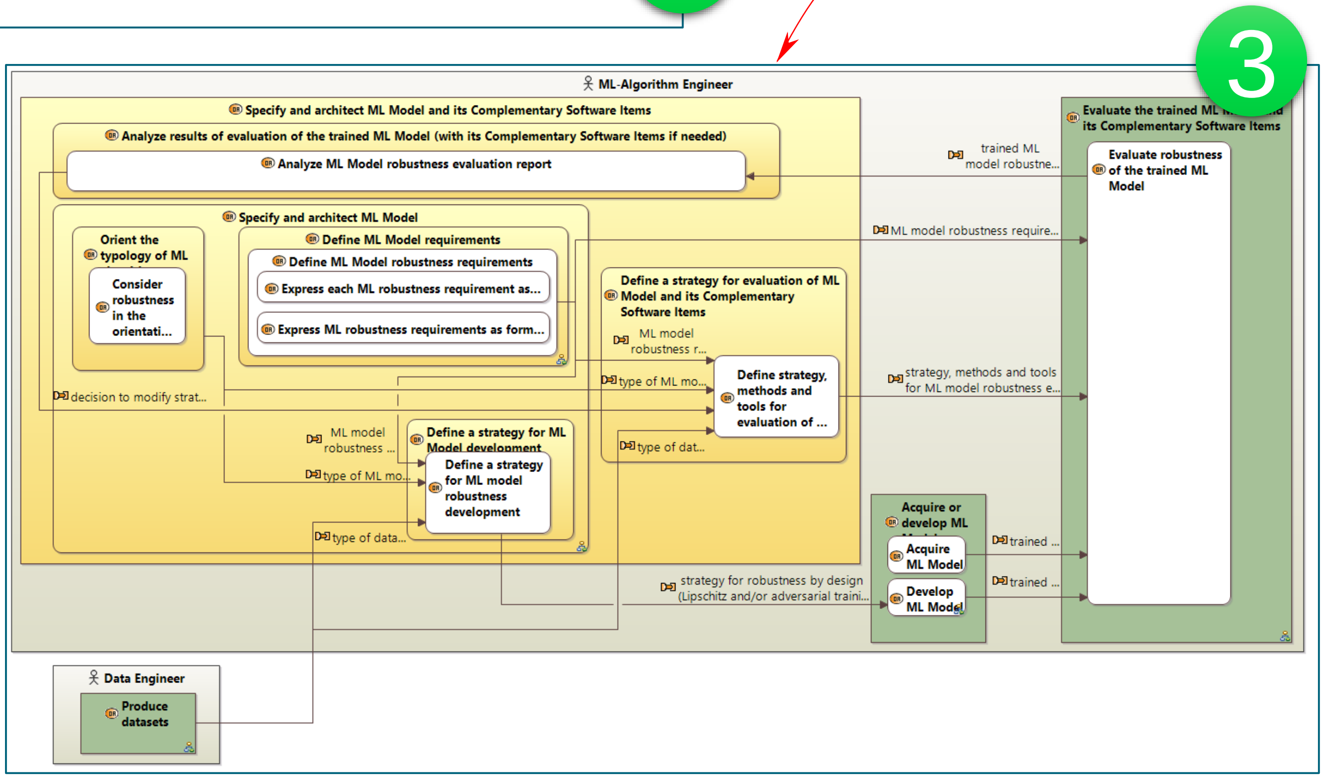


User choice

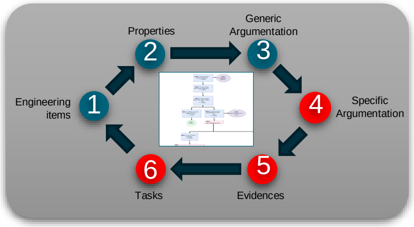
Argumentation

Workflow

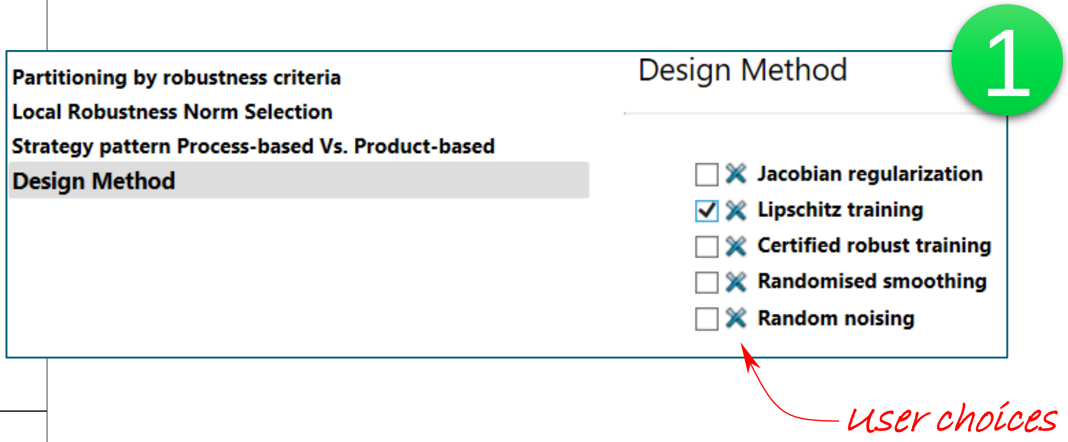


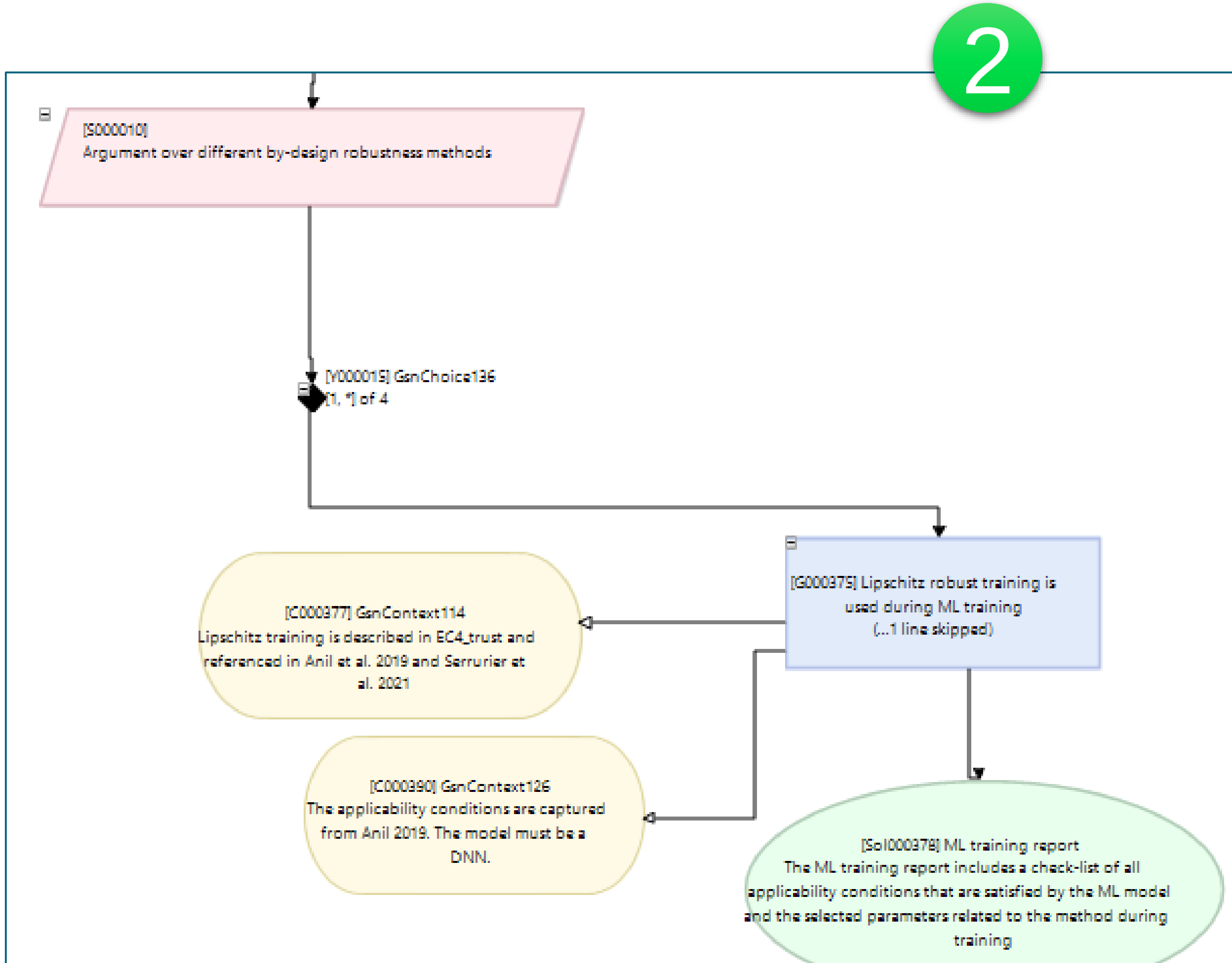


Robustness AC Template



- Families of method (from Confiance SotA: "EC4-Trustworthiness by design"):
  - <Jacobian regularization>, <Lipschitz training>, <Certified robust training>, <Randomised smoothing>, <Random noising>





FRENCH INSTITUTES OF TECHNOLOGY

3

Uncertainty Assessment & Choice of Strategies

Using Dempster-Shafer theory…

What we would like…

- Choose the most convincing strategy
- Focus the validation effort on the most sensitive parts of the argumentation
  - Assessment performed at each goal provides
    - Goal weakness
    - Contradiction between proof elements
  - For conjunctions
    - Procedure to improve the AC
- Identify the weaknesses of AC structure
  - Not sufficiently convincing strategies associated to a goal whose

FRENCH INSTITUTES OF TECHNOLOGY

Uncertainty in the context of AC

❑ How to establish confidence ?

- Use of assurance case to justify the well-founded development of systems integrating machine learning

❑ What is an assurance case?

- A structured argument used to justify a desired claim (safe, reliable, robust …), based on evidence(s) concerning both the system and the environment in which it operates.

❑ Issue

- **What are the sources of uncertainty in a structured argument?**
- **How to measure and propagate uncertainty in these structures?**

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).

What does confidence mean in our framework ?

- The concept of "Confidence", in our context (i.e., argumentation), reflects the amount of information an expert has that can justify his/her judgment about a proposition.

- A justification can be for or against a proposition. Formally, it's defined as:

$$Conf(A) = Bel(A) + Disb(A).$$

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

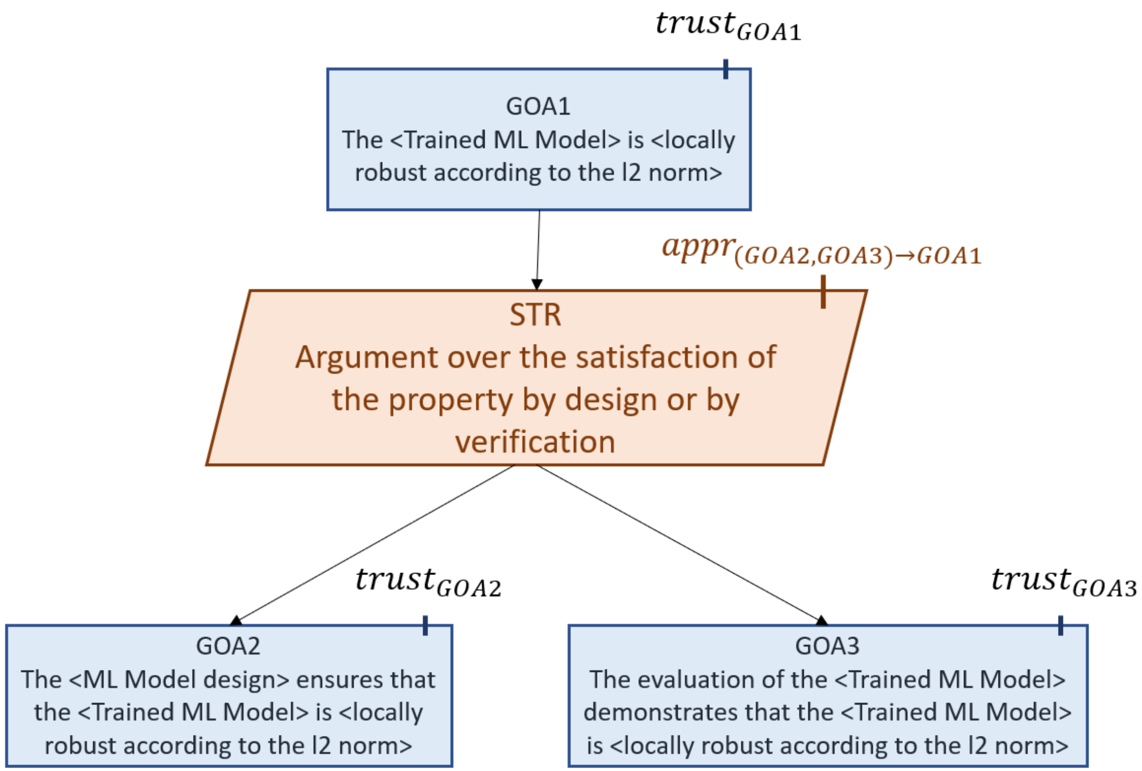
$$Conf(A) + Uncer(A) = 1.$$

Sources of uncertainty in AC

- Two factor to estimate uncertainty

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

  - Appropriateness which quantifies the truth about the inference (i.e., supported by relation) between a parent goal and its child goal(s). This is related to the strategy deployed by the AC designer to develop his/her reasoning.



- $trust_i \equiv (Bel_i, Disb_i, Uncer_i), i = \{GOA1, GOA2, GOA3\}$
- $appr \equiv (Bel_{(GOA2,GOA3)\rightarrow GOA1}, Uncer_{(GOA2,GOA3)\rightarrow GOA1})$

FRENCH INSTITUTES OF TECHNOLOGY

Source of uncertainty

- ❑ Aleatoric uncertainty (or Randomness) due to the variability of natural phenomena. E.g., rolling a dice.
- ❑ Epistemic uncertainty (or Incompleteness) due to lack of information. E.g., “The crime suspect fled in a grey car”. This information is not that sufficient to track down the suspect. What kind of car was it? In which direction did he/she flee?
- ❑ Inconsistency due to misinformation and contradiction. E.g., pro- and anti-vaccine arguments in a global pandemic situation.
- ❑ Fuzziness or vagueness due to imprecise information. E.g., Pierre is tall. The borderline between “tall” and “not tall” is not well-defined.

FRENCH INSTITUTES OF TECHNOLOGY

Measuring uncertainty – Probability Theory

❑ Probability theory deals well with random events (frequencies), but less well with singular events due to a lack of information.

❑It represents uncertainty by assuming even distribution over the whole frame of discernment Ω, such that: $P(\{\omega_i\}) = \frac{1}{|\Omega|}$.

- Example: Case of a light bulb, $\Omega = \{On, Off\}$
  - I have no idea of the state of the light bulb: $P(\{On\}) = P(\{Off\}) = \frac{1}{2}$
  - There's an equal chance of the light bulb being on or off: $P(\{On\}) = P(\{Off\}) = \frac{1}{2}$

Both situations are described using the same model

FRENCH INSTITUTES OF TECHNOLOGY

Measuring uncertainty – Dempster-Shafer Theory

- Dempster-Shafer theory (DST) is a generalization of probability theory that deal well with both epistemic and aleatory uncertainties.

- It defines the concepts of: Mass function (BPA) $m\colon 2^{\Omega} \rightarrow [0,1]$ such that: $\sum_{E \subseteq \Omega} m(E) = 1$.

- Example: Case of a light bulb, $\Omega = \{On, Off\}$
  - $m(\{On\})$: Quantifies the probability that the light bulb is **“On”**.
  - $m(\{Off\})$: Quantifies the probability that the light bulb is **“Off”**.
  - $m(\Omega)$: Quantifies ignorance on the state of the light bulb **“On” or “Off”**. Uncertainty
  - $m(\emptyset)$: Quantifies contradiction. I.e., **“On” and “Off” at the same time**.

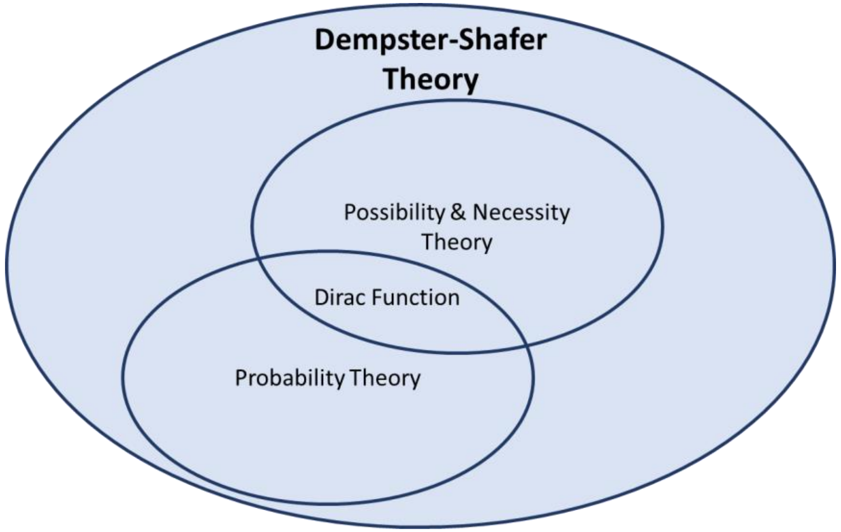
Measuring uncertainty – Dempster-Shafer Theory

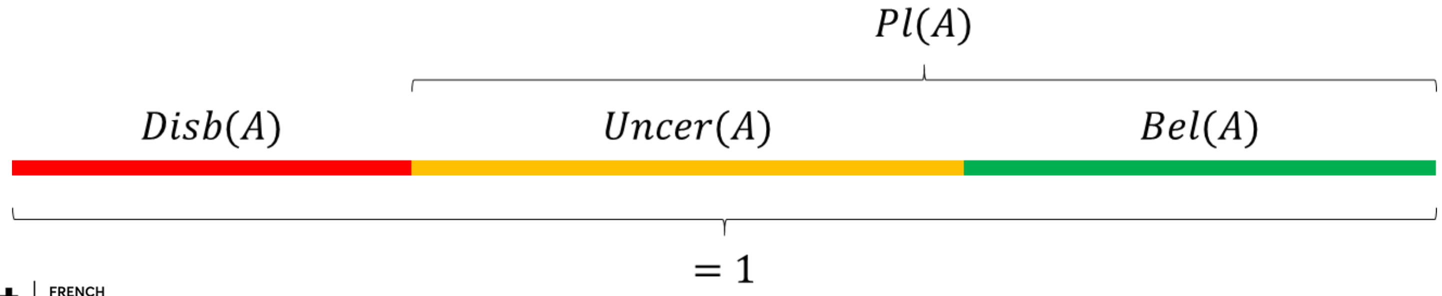
- From a mass function, we define the concepts of:
  - Belief function:

$$Bel(A) = \sum_{E \subseteq A, E \neq \emptyset} m(E) \text{ and } Disb(A) = Bel(\bar{A})$$

  - Plausibility function:

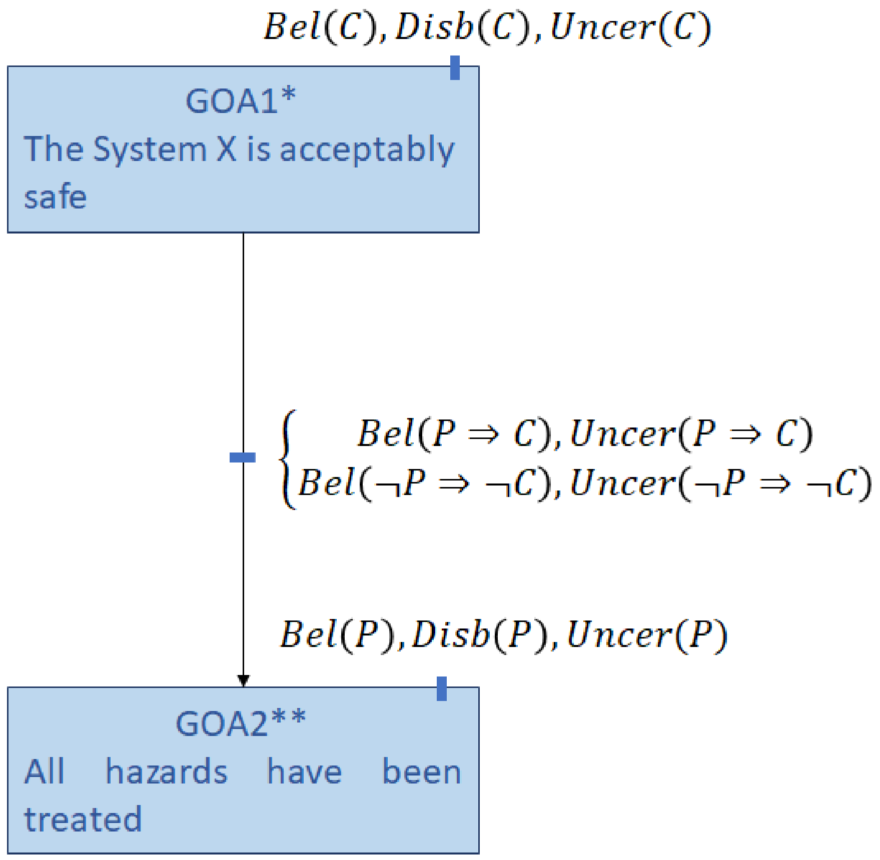
$$Pl(A) = \sum_{E \cap A \neq \emptyset} m(E) = 1 - Disb(A)$$





Uncertainty Evaluation – Mathematical Background

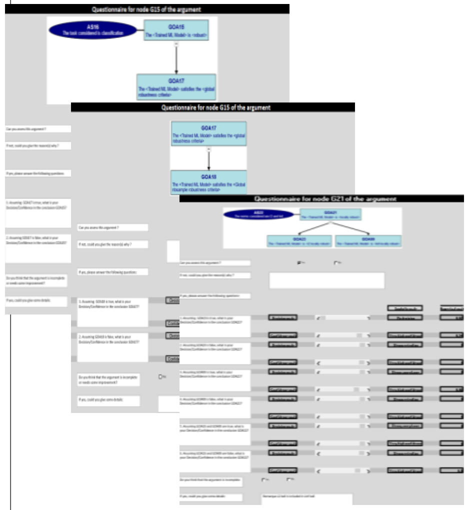
- ## *Hypothesis*
  - ***Goals** directly supported by a **Solution** can be:*
    - *Believed, i.e. $m(g)_g$ can be different from zero*
    - *Disbelieved, i.e. $m(g)_{not\ g}$ can be different from zero*
    - *Epistemically uncertain, i.e. $m(g)_{g\ or\ not\ g}$ can be different from zero*
  - *Rules involved in a **Strategy** can be:*
    - *Believed, i.e. $m(r)_r$ can be different from zero*
    - *Epistemically uncertain, i.e. $m(r)_{r\ or\ not\ r}$ can be different from zero*
    - *But cannot be disbelieved, i.e. $m(g)_{not\ g} = 0$*
  - *Rules are: p_i => C, not p_i => not C*
    - *Provides a formal and flexible definition of Is supported by*



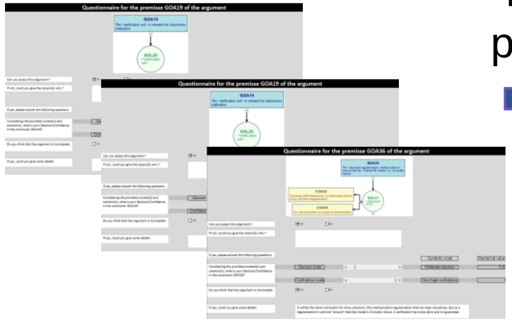
$*GOA1 = C\ (Conclusion)$
$**GOA2 = P\ (Premise)$

fit | FRENCH INSTITUTES OF TECHNOLOGY

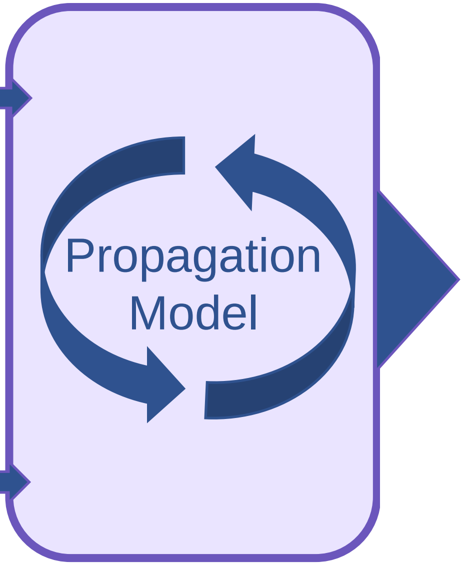
Confidence & Uncertainty in DST/AC framework



Belief values on rules (i.e., appropriateness)



Belief and disbelief values on premises (i.e., trustworthiness)



Belief and disbelief values on "Top-goal" (i.e., trustworthiness)

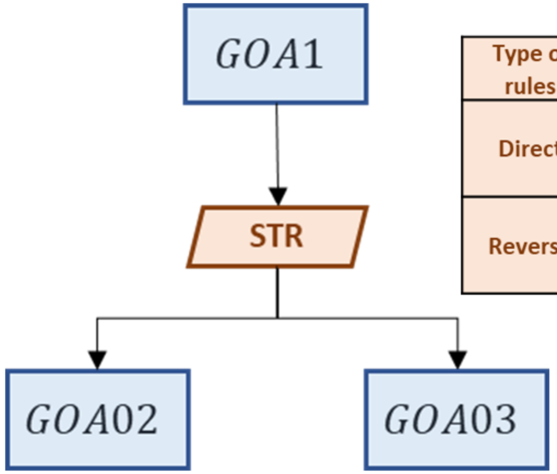
AC-Robustness

FRENCH INSTITUTES OF TECHNOLOGY

Uncertainty metrics
Visualization format

- Uncertainty metrics are displayed in terms of belief and disbelief degrees.

Metric (GOA01)	Numerical value	Qualitative value
Belief degree	0.95	Very high
Disbelief degree	0.00	Very low
Conflict degree	0.00	Very low



Type of rules	Rules (STR)	Belief Values	Qualitative belief
Direct	$All \rightarrow GOA01$	1.00	Very high
	$GOA02 \rightarrow GOA01$	0.95	Very high
	$GOA03 \rightarrow GOA01$	0.70	High
Reverse	None $\rightarrow \neg GOA01$	1.00	Very high
	$\neg GOA02 \rightarrow \neg GOA01$	0.50	High
	$\neg GOA02 \rightarrow \neg GOA01$	0.50	High

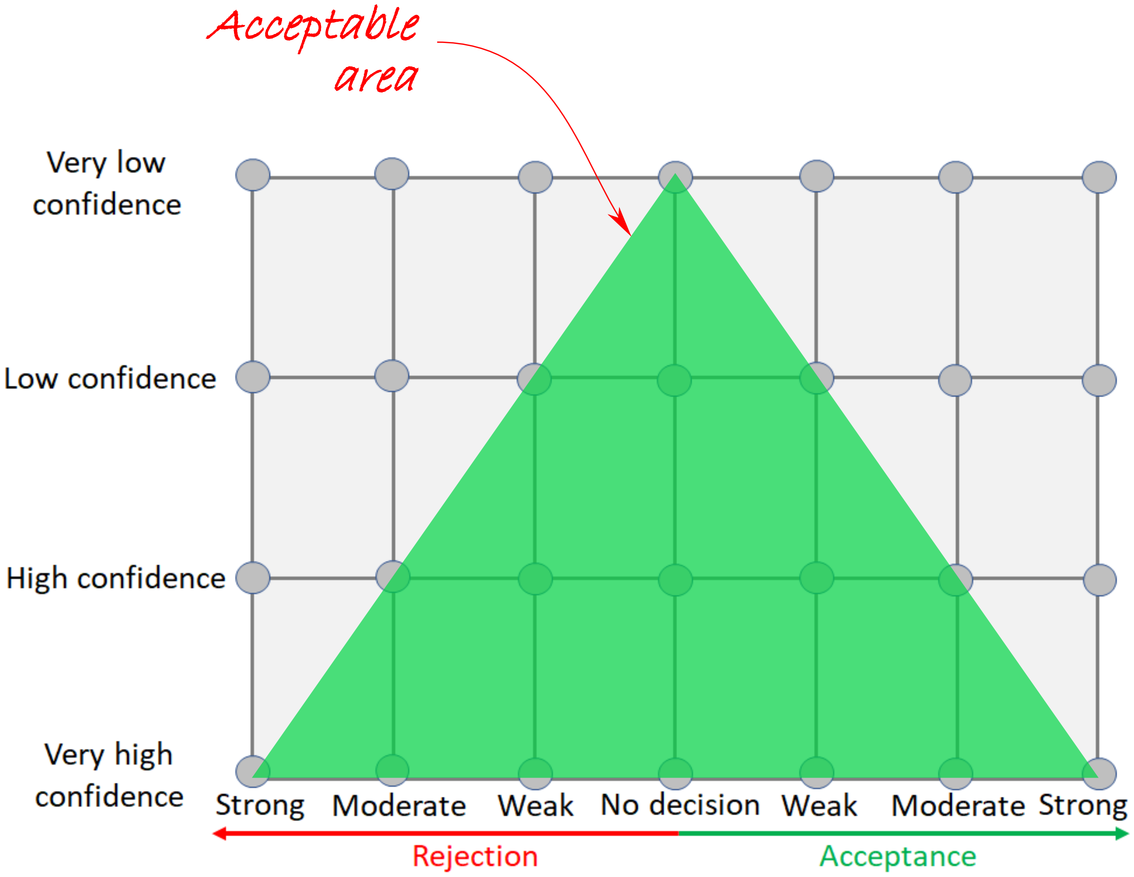
Metric (GOA02)	Numerical value	Qualitative value
Belief degree	0.90	Very high
Disbelief degree	0.00	Very low
Conflict degree	0.00	Very low

Metric (GOA03)	Numerical value	Qualitative value
Belief degree	0.75	Very high
Disbelief degree	0.00	Very low
Conflict degree	0.00	Very low

Nota. Belief (resp. disbelief) degree, noted $Bel(\{A\})$ (resp. $Disb(\{A\}) = Bel(\{\neg A\})$ represents the sum of all evidence in favour of (resp. against) an assertion (A). While uncertainty degree is noted $Uncer(\{A\}) = 1 - Bel(\{A\}) - Disb(\{A\})$. The strength of an evidence for or against A is called a mass and is resp. noted $m(\{A\})$ to quantify the probability that A is True or $m(\{\neg A\})$ when A is False, while $m(\{A, \neg A\})$ quantifies ignorance.

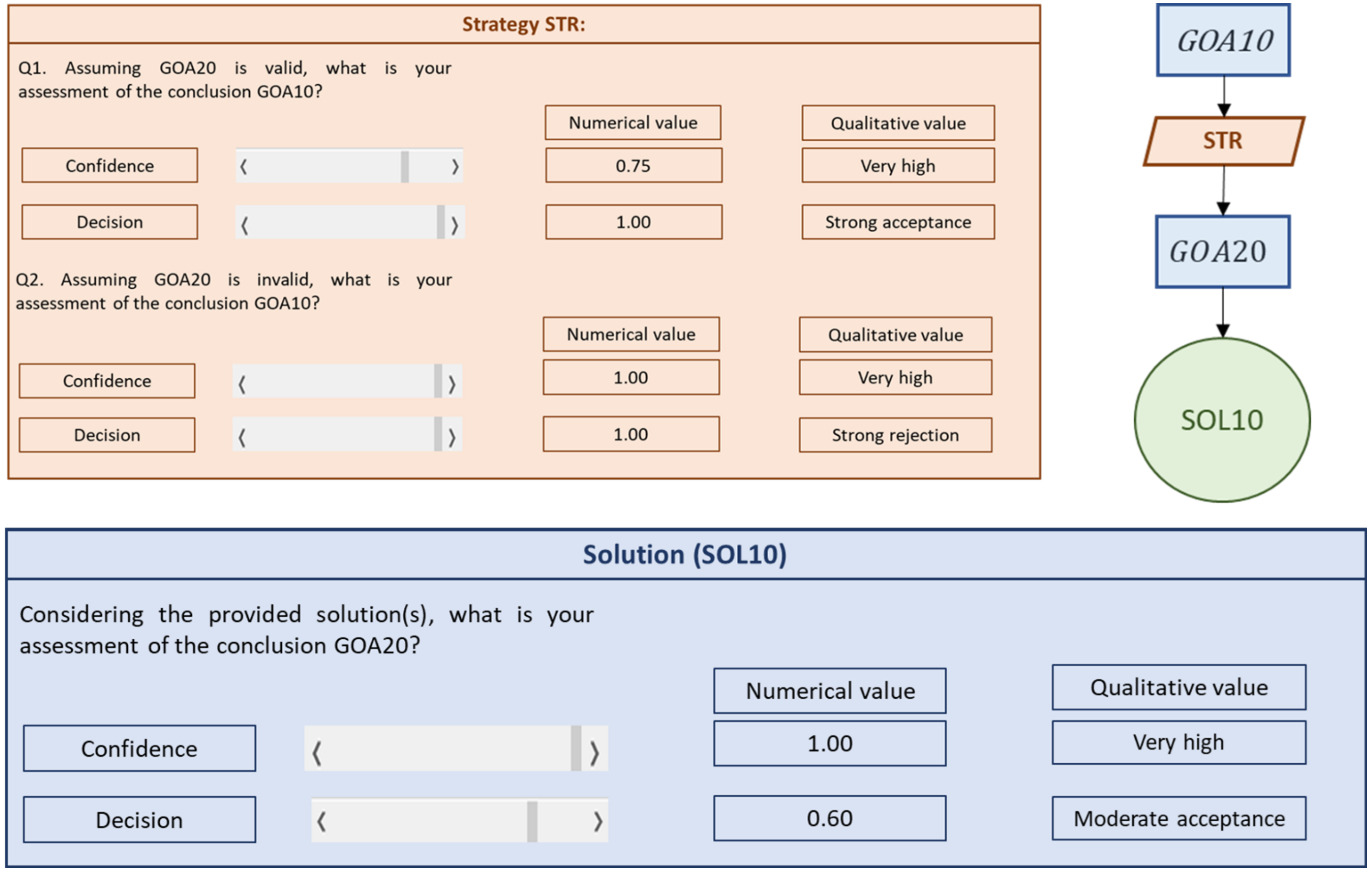
Uncertainty Evaluation – Mathematical Background

- Elicitation
  - **Decision**, Dec(A): given by an expert to accept or reject a proposition (A)
    - Dec(A)=[1+Bel(A)-Disb(A)]/2
  - **Confidence**, Conf(A): the amount of information the expert needs to justify his/her decision
    - Conf(A)=Bel(A)+Disb(A)
- A constraint is added to ensure that strong decisions are not taken in cases of significant uncertainty:
  - [1-Conf(A)]/2≤Dec(A)≤[1+Conf(A)]/2



Uncertainty metrics
Entry format

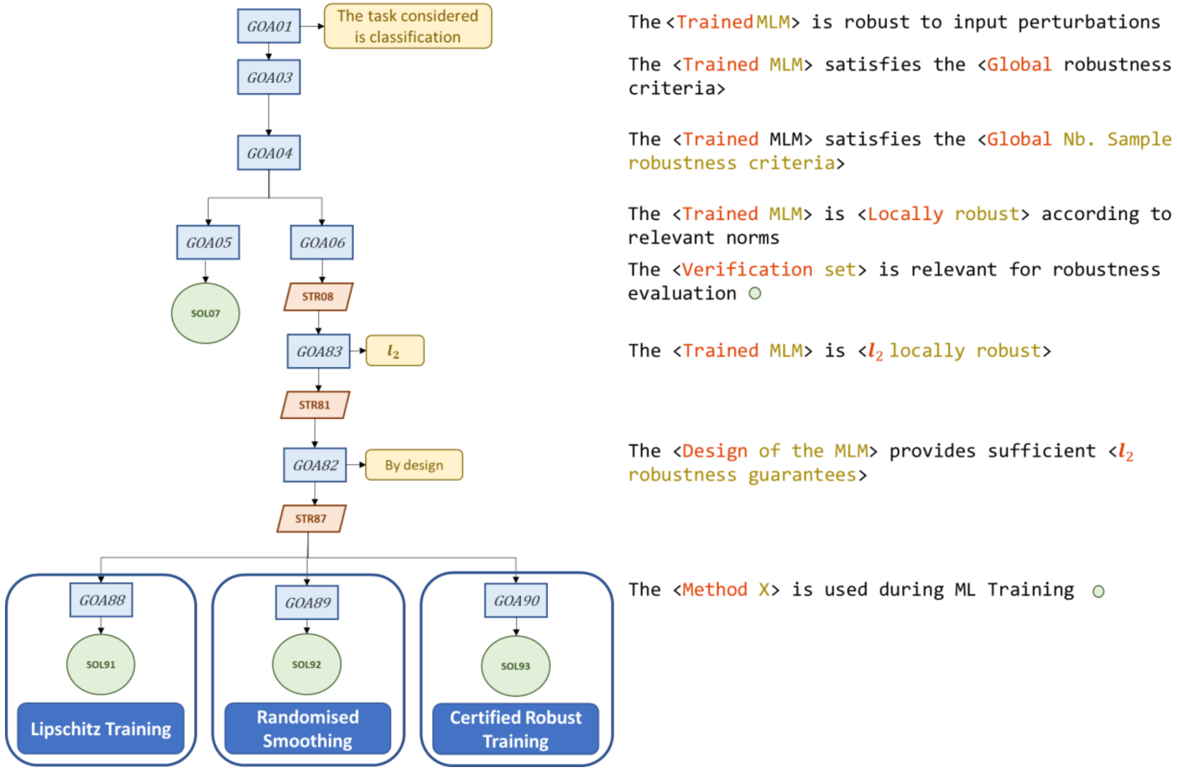
- ❑ Uncertainty metrics are pre-entered by the developer



FRENCH INSTITUTES OF TECHNOLOGY

Confidence metrics propagation

❑ Method selection on the basis of propagation results:



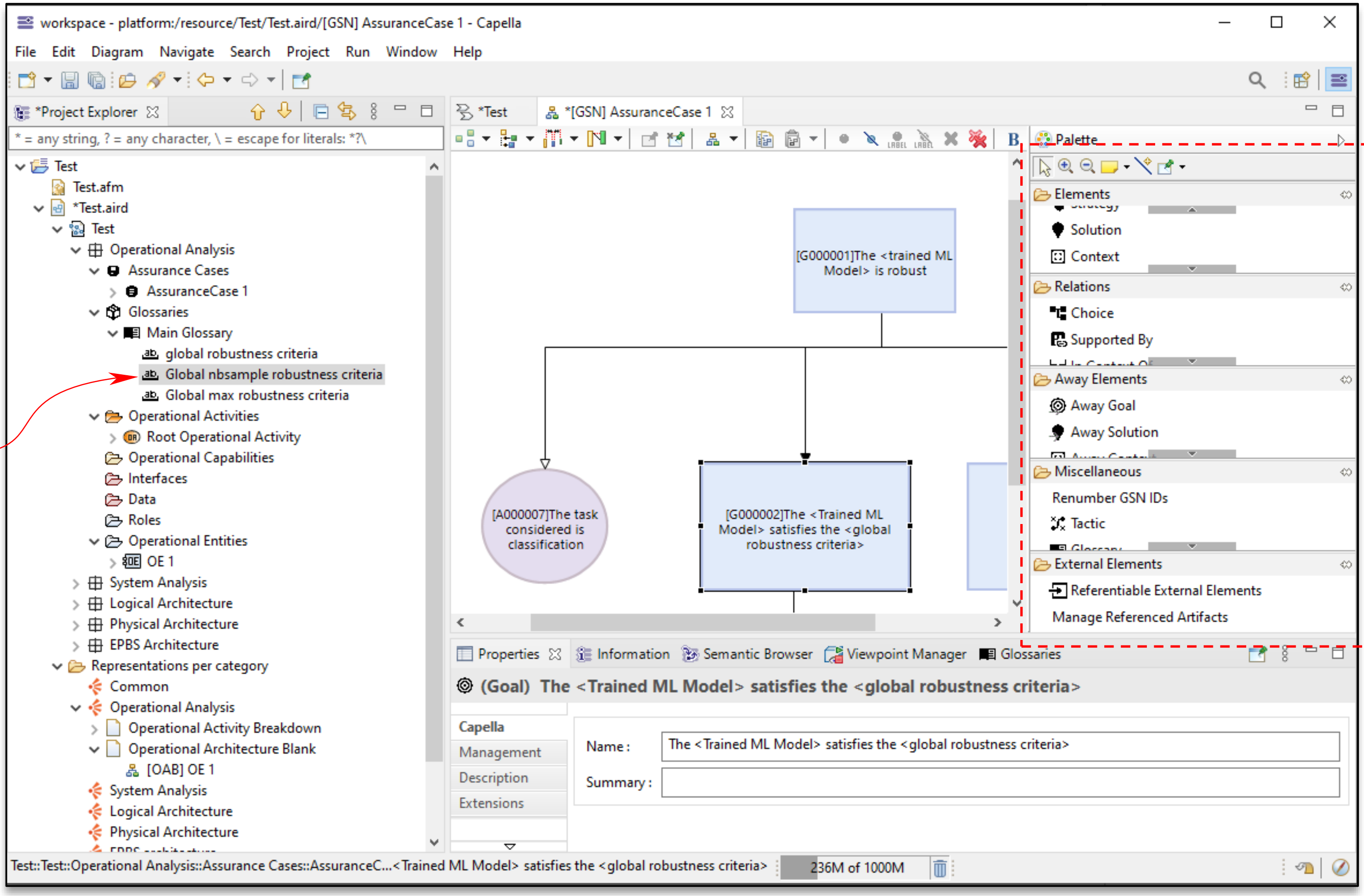
AC after requirements specification

Methods	Propagation results on the top-goal	
	Belief degree	Disbelief degree
Lipschitz Training	0,92	0,01
Randomised Smoothing	0,78	0,02
Certified Robust Training	0,89	0,01

4

Tool support

Assurance case viewpoint in Capella Environment

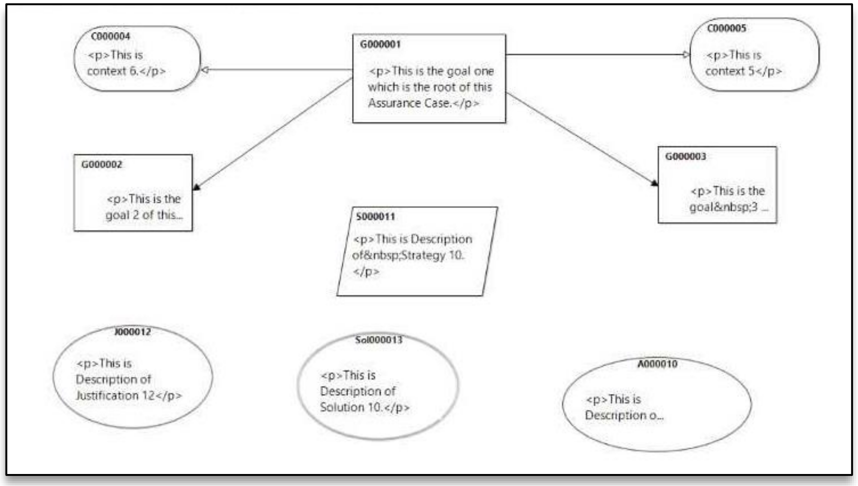


Glossary entries

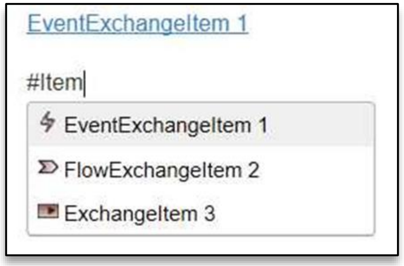
GSN palette

Assurance case viewpoint in Capella
Environment

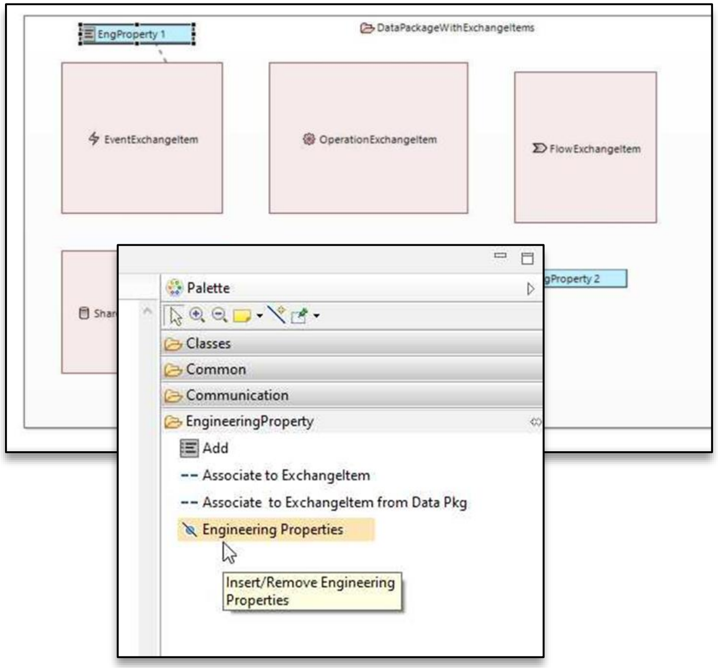
Assurance Cases



Content assist



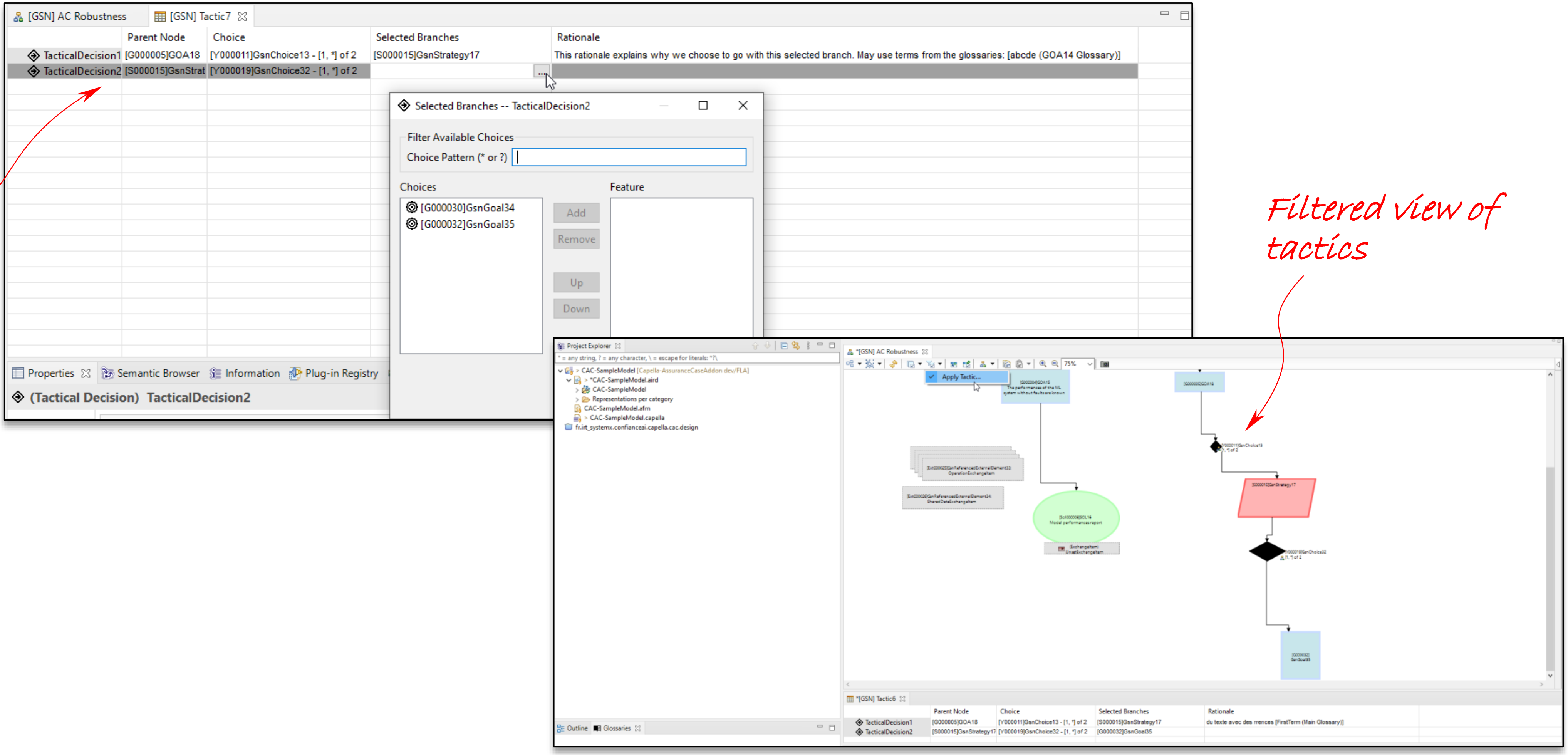
Property <=> Exchange item



FRENCH INSTITUTES OF TECHNOLOGY

Assurance case viewpoint in Capella
Tactics

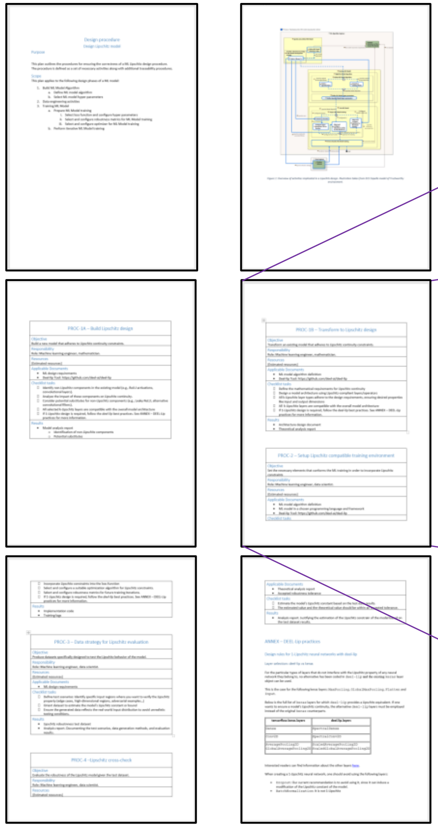
Tabular definition of tactics



Filtered view of tactics

Assurance case viewpoint in Capella

V&V Plan



PROC-1B – Transform to Lipschitz design

Objective

Transform an existing model that adheres to Lipschitz continuity constraints.

Responsibility

Role: Machine learning engineer, mathematician.

Resources

{Estimated resources}

Applicable Documents

- ML model algorithm definition
- Deel-lip Tool: https://github.com/deel-ai/deel-lip

Checklist tasks

- ☐ Define the mathematical requirements for Lipschitz continuity
- ☐ Design a model architecture using Lipschitz-compliant layers/operators
- ☐ All k-Lipschitz layer types adhere to the design requirements, ensuring desired properties like input and output dimensions
- ☐ All k-Lipschitz layers are compatible with the overall model architecture
- ☐ If 1-Lipschitz design is required, follow the *deel-lip* best practices. See ANNEX – DEEL-Lip practices for more information.

Results

- Architecture design document
- Theoretical analysis report

FRENCH INSTITUTES OF TECHNOLOGY

5

Conclusion

Where are we now?

What next?

Status and next steps...

- Status
  - A (small) set of **Assurance Cases** on “important” properties for the development of systems embedding ML components (robustness, explicability, ODD correctness and completeness,…)
  - A Model-based approach integrating and linking workflow and assurance case models
  - A Capella **GSN viewpoint** with extensions supporting the approach

- Next steps
  - Improve integration within the “Confiance.ai” workbench
    - Links with the **set of solutions** proposed by the project
    - Links with the “**Body of Knowledge**” created by the project
  - Extension of Assurance Cases to **other properties**
  - Addition of **new features**
    - Impact analysis
    - Dependencies between strategies

FRENCH INSTITUTES OF TECHNOLOGY

THANKS FOR YOUR ATTENTION