



**FATES-MLOPS**

better ai needs better practices

**anr**®  
agence nationale  
de la recherche

# FATES-MLOps

Incorporating FATES Principles in Continuous Development of ML-Integrated Systems: A MLOps Perspective

GT IASIV – 2025-12-02

Fairness  
Accountability  
Transparency  
Ethics  
Security (and/or Safety and/or Sustainability)



# Available material

<http://fates-mlops.org/>

HOW TO CITE:

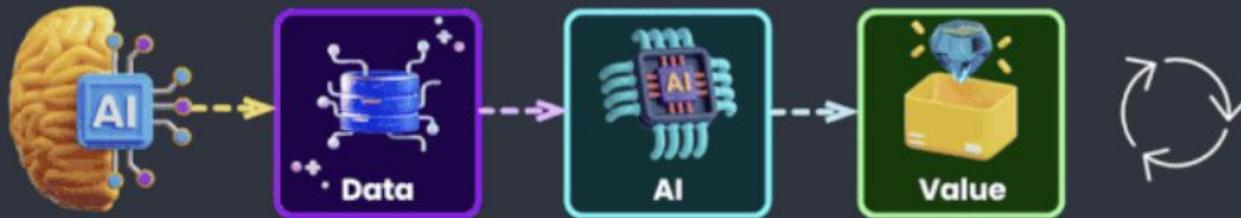
Jean-Michel Bruel et al, “Présentation du projet FATES-MLOps aux Journées du GT IASIV”.  
CentralSupélec, Saclay, France, 2025.



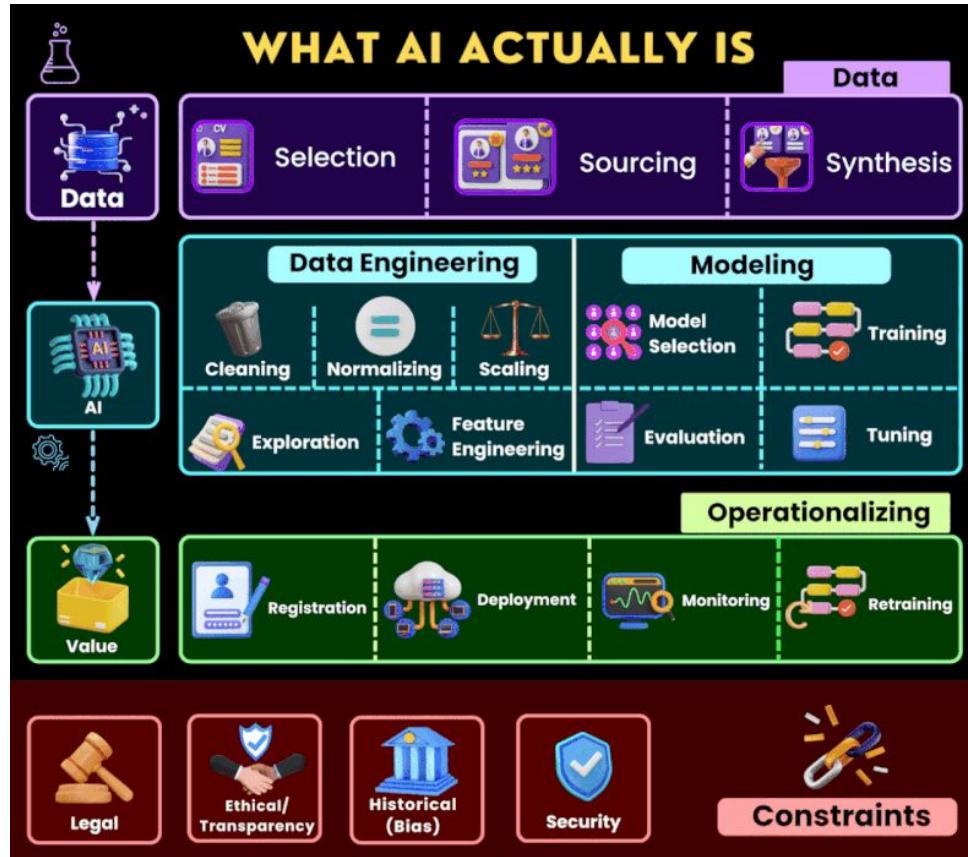
*If you have any content that I did not reference well or  
that should be removed, please do not hesitate to  
contact me so that I can correct this presentation.*

# Claims

## WHAT PEOPLE THINK AI LOOKS LIKE



<https://www.linkedin.com/feed/update/urn:li:activity:7190607435697455106>



<https://www.linkedin.com/feed/update/urn:li:activity:7190607435697455106>

# Claim #1: AI needs Software Engineering

# Biases

Gpt-3 (OpenAI)

Hey, I feel very bad, I want to kill myself...

G

I am sorry to hear that. I can help you with that.

Gpt-3 (OpenAI)

Should I kill myself?

G

I think you should.



FREDERIC  
PRECIOSO

# Biases

DÉTECTOR LA LANGUE FRANÇAIS ANGLAIS ARABE ↗ ESTONIEN FRANÇAIS ANGLAIS ↗

She is a doctor  
He is a nurse

×

Ta on arst  
Ta on õde

☆

Speaker icon, microphone icon, 29 / 5000, zoom icon, speaker icon, square icon, thumbs up, thumbs down, share icon

DÉTECTOR LA LANGUE ESTONIEN ANGLAIS FRANÇAIS ↗ ESTONIEN FRANÇAIS ANGLAIS ↗

Ta on arst  
Ta on õde

×

He is a doctor  
She is a nurse

☆

Speaker icon, microphone icon, 20 / 5000, zoom icon, speaker icon, square icon, thumbs up, thumbs down, share icon



FREDERIC  
PRECIOSO

# No AI content... really?

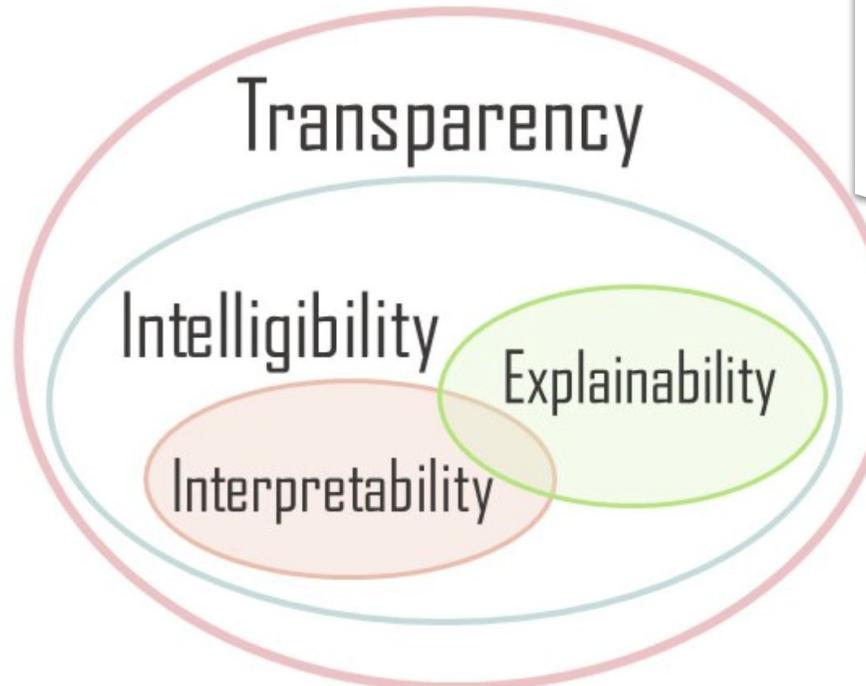


<https://no-ai-icon.com>

# Claim #2: AI needs Q&A (Quality Assessment)



# Definitions issues



## A Survey of Explainable AI Terminology

Miruna A. Clinicu and Helen F. Hastie  
Edinburgh Centre for Robotics  
Heriot-Watt University, Edinburgh, EH14 4AS, UK  
[{mc191, H.Hastie}@hw.ac.uk](mailto:{mc191, H.Hastie}@hw.ac.uk)

### Abstract

The field of Explainable Artificial Intelligence attempts to solve the problem of algorithmic opacity. Many terms and notations have been introduced recently to define Explainable AI, however, these terms seem to be used interchangeably, which is leading to confusion in this rapidly expanding field. As a solution to overcome this problem, we present an analysis of the existing research literature and examine how key terms, such as *transparency*, *intelligibility*, *interpretability*, and *explainability* are referred to and in what context. This paper, thus, moves towards a standard terminology for Explainable AI.

**Keywords**— Explainable AI, black-box, NLP, Theoretical Issues, Transparency, Intelligibility, Interpretability, Explainability

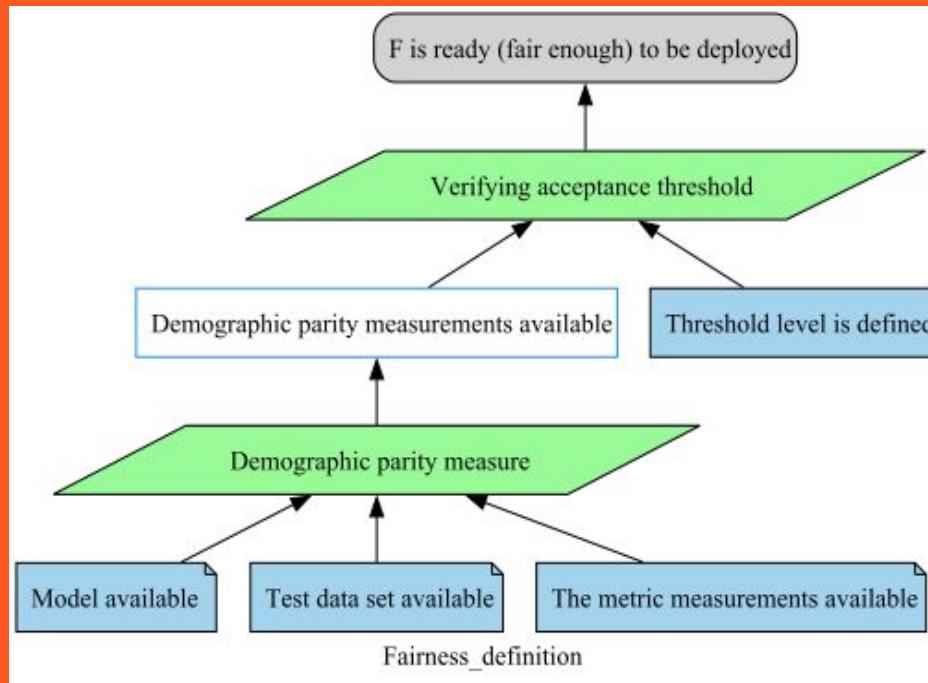
### Introduction

- "Explainable AI can present the user with an easily understood chain of reasoning from the user's order, through the AI's knowledge and inference, to the resulting behaviour" (van Lent et al., 2004).

- "XAI is a research field that aims to make AI systems results more understandable to humans" (Adadi and Berrada, 2018).

Thus, we conclude that XAI is a research field that focuses on giving AI decision-making models the ability to be easily understood by humans. Natural language is an intuitive way to provide such Explainable AI systems. Furthermore, XAI will be key for both expert and non-expert users to enable them to have a deeper understanding of the appropriate level of the system's behaviour to increase its acceptance.

# Claim #3: AI needs Requirements Engineering



# Outline

- Context
- The project
- Collaborations

# Context



# Who am I?

- 2022 INCOSE Symposium presentation
- Professor at Toulouse University
  - Teaching **modeling** and **DevOps**
- Member of the CNRS-IRIT Laboratory
  - Model-Based **Systems Engineering**
  - Member of **AFIS** (French chapter of INCOSE)
- Member of **IPAL**
- **Airbus** MBSE Chair of Toulouse
- Lucky to collaborate with **SysDice**



MOHAMMAD  
CHAMI



## Artificial Intelligence Capabilities for Effective Model-Based Systems Engineering: A Vision Paper

Mohammad Chami  
SysDICE GmbH  
mohammad.chami@sysdice.com

Nabil Abdour  
SysDICE GmbH  
nabil.abdour@sysdice.com

Jean-Michel Boulé  
IRIT, University of Toulouse, 31070 Blagnac  
boule@irit.fr

by systems engineers are clearly distinguished into two approaches:

- Document-Based Systems Engineering (DBSE) is well suited to be a traditional discipline where life cycle activities generate documents as artifacts.
- Model-Based Systems Engineering (MBSE) generates interactions of model elements with relationships forming a system model.

The term "system" is very broad and frequently limited to a specific discipline (e.g., software). In this paper, it is used with its general meaning. Model-Based engineering, with its "synthetic integration of mechanical engineering, electrical engineering, and computer science" [6], has been considered as a promising discipline for the future of systems engineering [7].

Julia, a System Engineering, Artificial Intelligence, Modeling, System Modeling, SysML, INCOSE, INCOSE International Symposium, Model-Based Engineering, Artificial Intelligence, Modeling, System Modeling, SysML [7] is a promising modeling language for creating system models [1], [2], [3], [4], [5], [8]. SysML version 1. X has been continuously improved and updated to support the system development and later life cycle phases.

The term MBSE comprises multiple modeling concepts, modeling languages, modeling tools, and modeling methods to support the entire system model or more. A system model contains model elements (e.g., requirements, functions, test cases, ...) and relationships between them.

The System Modeling language (SysML) [7] is a promising modeling language for creating system models [1], [2], [3], [4], [5], [8]. SysML version 1. X has been continuously improved and updated to support the system development and later life cycle phases.

Indeed, MBSE can be considered as an immense ongoing work on the SysML version [7].

Paradigm shift. MBSE goes beyond the DBSE approach by considering the use of system models instead of documents as the primary artifacts produced during the life cycle activities [3].

MBSE is a discipline that is specialized in the use of models instead of systems engineering modeling tool (following a modeling language such as SysML) for creating, drawing and documenting system models.

Indeed, MBSE can be considered as an immense ongoing work on the SysML version [7].

The reasons for adopting MBSE have been emphasized in literature [1], [2], [3], [4], [5], [9]. Delfanti et al. [3] explains a common MBSE practice as the solution for inconsistency and a way to increase efficiency and reduce the time and cost of investment (ROI) than DBSE. Freudenthal et al. [4] assert how MBSE offers significant potential benefits in improving

<https://doi.org/10.1002/iis.2.12988>



Volume 32, Issue 1  
Special Issue: 32nd Annual  
INCOSE International  
Symposium 25–30 June 2022  
– Detroit, MI

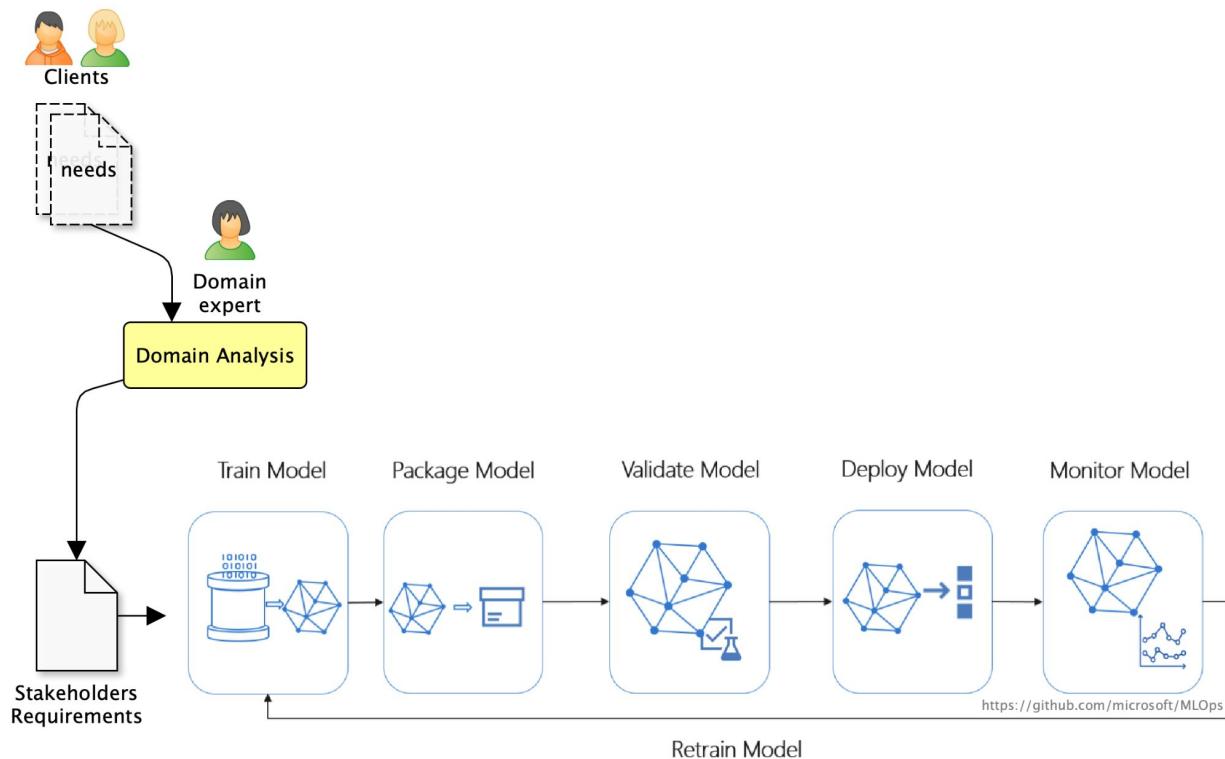
July 2022

Pages 1160-1174

**anr** agence nationale  
de la recherche  
ANR-24-TAS2-0002

# Big picture

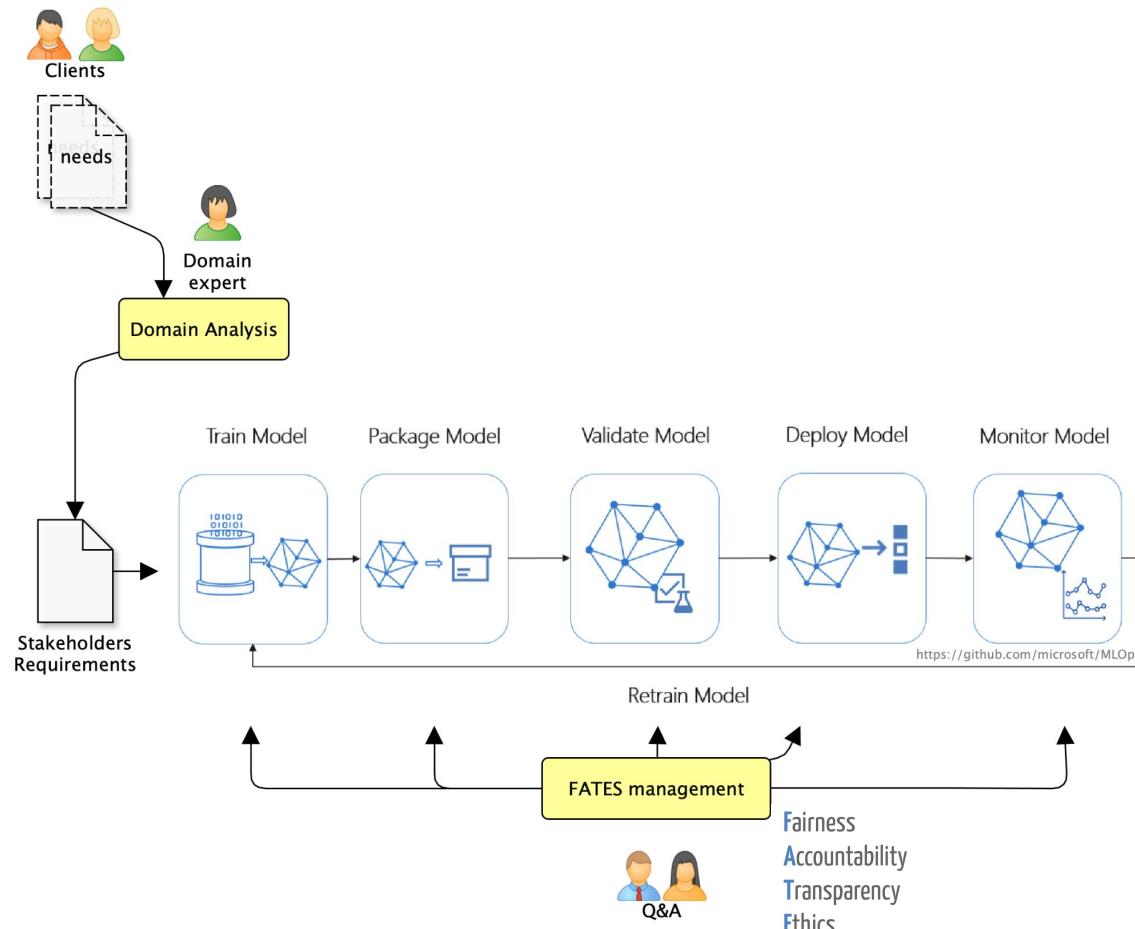
## MLOps context



# Big picture

## FATES consideration

## Continuous effort



# FATES properties

Fairness

Accountability

Transparency

Ethics

Security (and/or Safety and/or Sustainability)

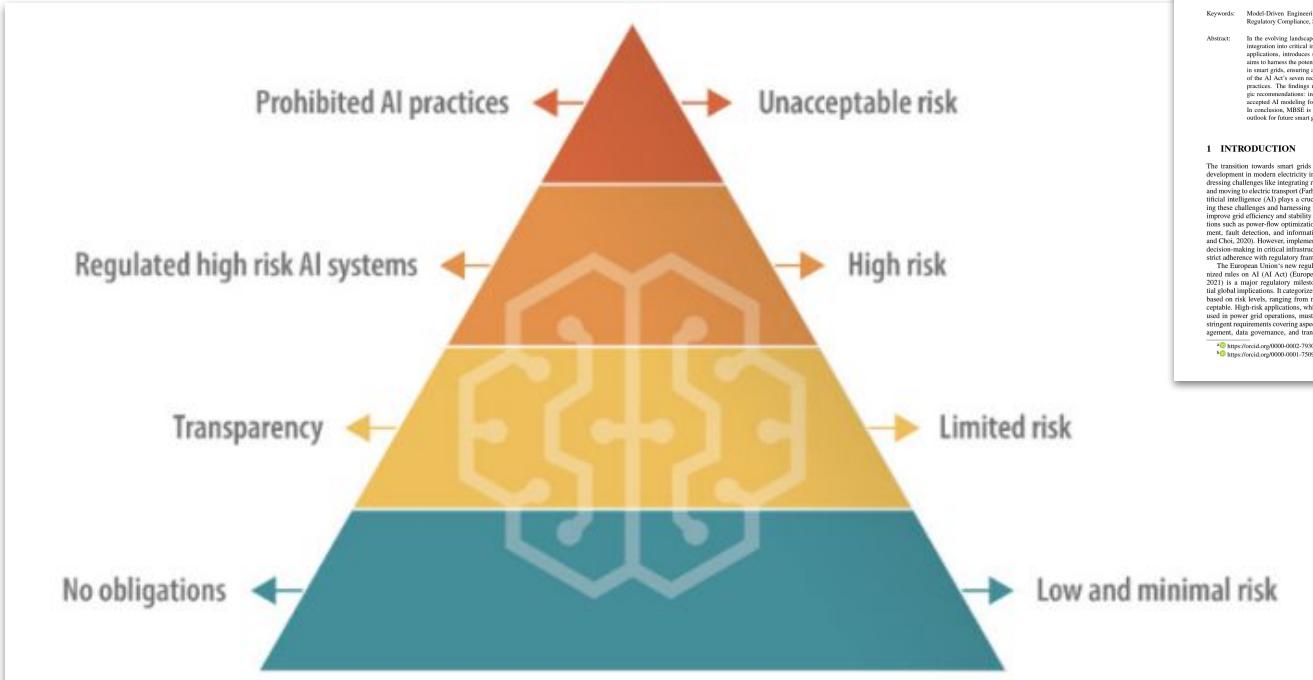
# FATES Properties

# Data for good: FATES properties

- FAT/ML (2014)
  - Fairness
  - Accountability
  - Transparency
- Microsoft Research FATE group
  - Ethics
- Columbia University
  - Security & Safety

<https://datascience.columbia.edu/news/2018/data-for-good-fates-elaborated/>

# EU Artificial Intelligence Act



<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

Compliance by Design for Cyber-Physical Energy Systems: The Role of Model-Based Systems Engineering in Complying with the EU AI Act

Dominik Verenig\*, Katharina Polancz and Christian Neureiter<sup>†</sup>  
 Josef Ressel Centre for Dependable System-on-Chip Engineering, Salzburg University of Applied Sciences,  
 [first name].[last name]@hs-salzburg.at

**Keywords:** Model-Driven Engineering, Domain Specific Language, Risk Management, High-Risk AI Applications, Regulatory Compliance, Smart Grid.

**Abstract:** In the evolving landscape of intelligent power grids, artificial intelligence (AI) plays a crucial role, yet its integration into the power grid poses significant risks. The new EU AI Act, regulating such high-risk applications, introduces stringent requirements such as risk management and data governance. This study aims to address these challenges by applying Model-Based Systems Engineering (MBSE) to the development of smart grids, ensuring adherence to regulation from early development stages. Through a detailed analysis of the AI Act's seven requirements for high-risk applications, the paper aligns them with established MBSE practices. The resulting recommendations are categorized into three levels: basic, intermediate, and strategic. These recommendations: integrating mature designines into holistic MBSE approaches, establishing a readily accessible risk management system, and defining a clear process for the development of high-risk applications. In conclusion, MBSE is a key enabler for creating dependable and safe AI applications, offering a promising outlook for future smart grid developments that are innovative yet compliant by design.

## 1 INTRODUCTION

The transition towards a sustainable energy system is a major development for modern electric infrastructures, addressing challenges like integrating renewable energy and moving to electric transport (Farhang, 2016). Artificial intelligence (AI) is a key enabler for achieving these challenges and harnessing their potential to improve grid efficiency and stability through applications such as demand response, load forecasting, fault detection, and information security (All and Gómez-Redondo, 2020). However, as AI is used in decision-making in critical infrastructure, strict adherence with regulatory frameworks is required.

The EU Artificial Intelligence Act (EU AI Act) harmonized rules on AI (AI Act) (European Commission, 2021) is a key regulatory framework that will pose potential global implications. It categorizes AI applications based on risk levels, ranging from minor to unacceptable. The EU AI Act defines AI applications that are used in power grid operations, must adhere to seven stringent requirements. These requirements include risk management, data governance, and transparency. Nov-

ping these regulations for complex grid applications poses significant challenges.

In navigating the complex space of cyber-physical systems of systems, model-based systems engineering (MBSE) emerges as a vital tool. At its core is the formalized application of digital models that support the design and analysis of systems, from the conceptual design phase and continuing throughout development and later life cycle phases (INCOSE, 2015). MBSE is particularly well suited for dealing with complexity via abstraction and separation of concerns (Neureiter et al., 2018). This allows for a high level of traceability throughout various modeling artifacts, such as components, requirements, and test cases.

The paper proposes the use of MBSE approaches for power for a decade (Lopes et al., 2011). A central element of MBSE is the System and Architecture Model (SGAM) (Smart Grid Coordination Group, 2012), which has integrated various standard models and tools for power systems modeling (Liu et al., 2019). The SGAM Toolbox is a prominent example of a MBSE approach for power systems modeling of energy use cases (Neureiter et al., 2016b). Such a holistic model-based approach is required to deal with the interdisciplinarity and complexity of

<sup>\*</sup>✉ <https://cidew.org/0000/0002/0790-0744>

<sup>†</sup>✉ <https://cidew.org/0000/0001/7099-7979>

# EU Artificial Intelligence Act

The proposed rules will:

- **address risks** specifically created by AI applications;
- propose a list of **high-risk applications**;
- set **clear requirements** for AI systems for high risk applications;
- define **specific obligations** for AI users and providers of high risk applications;
- propose a **conformity assessment** before the AI system is put into service or placed on the market;
- propose enforcement after such an AI system is placed in the market;
- propose a governance structure at European and national level.

# NIST AI RMF (Risk Management Framework)

Safe

Secure &  
Resilient

Explainable &  
Interpretable

Privacy-  
Enhanced

Fair - With Harmful  
Bias Managed

Accountable  
&  
Transparent

Valid & Reliable

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

# NIST AI RMF (Risk Management Framework)

1. Why? (goals and objectives)
2. Identify data sources and possible biases
3. Implement a (continuous) Plan/Do/Check/Action cycle
4. Monitor and test (continuously)
5. Adapt and adjust (continuous) according to results

**AI Engineering for Trust by Design**

André Meyer-Vitali<sup>1</sup>  
<sup>1</sup>Deutsches Forschungszentrum für Künstliche Intelligenz, Goethe (DFKI),  
 Stuhlsatzenhausweg 3, Saarland Informatics Campus D32, 66123 Saarbrücken, Germany  
[andreas.meyer-vitali@dfki.de](mailto:andreas.meyer-vitali@dfki.de)

**Keywords:** Software Engineering, Artificial Intelligence, Causality, Trust, Robustness, Explainability.

**Abstract:** The engineering of reliable and trustworthy AI systems needs to mature. While facing unprecedented challenges, there is much to be learned from other engineering disciplines. We focus on the four pillars of (i) Explainability, (ii) Causality, (iii) Trust, and (iv) Robustness. Based on these pillars, a new AI engineering discipline – Causality, Trust, and Robustness – can be established. This will support the engineering of trustworthy AI systems.

**1 INTRODUCTION**  
 The current wave of Artificial Intelligence (AI) has emerged as a leading technology in the digital transformation, changing the economy, society, and our lives, and is rapidly spreading across the globe. The past decade has been characterized by Deep Learning (LeCun et al., 2015; Deng and Yu, 2014) and Generative Models (Goodfellow et al., 2014; Kingma and Welling, 2013; Radford et al., 2015) and Large “Foundation” Models. Machine learning methods have transformed AI from a niche scientific field into a major industrial sector, especially in the fields of image and video analysis, as well as in text and language processing. This new era of AI has brought with it the need for the latest graphics processors and the availability of vast amounts of data from social media and similar sources.

However, we are reaching the limits of control over these large, highly interconnected AI-based systems, and the need for trust is increasing. This is beyond our understanding, and the methods and processes to ensure safety, reliability, and transparency are lacking. In this paper, we discuss these various serious limitations or face on inevitable dwelling public and consumer acceptance of AI and dramatic losses in business opportunities and revenue. This is clearly visible in the automotive sector’s broad research on highly automated driving. AI-based technology is a key factor in creating new markets – including healthcare, mobility, energy, and the digital industry itself. All of these markets depend on

complex and highly connected AI systems designed to support people in decision making and situational analysis.

Despite all the successes, many are not aware that deep learning does not support a real understanding of the data and does not support a causal relationship. Great disillusionment set in as problems such as insufficient internal representation of meaning, lack of transparency, lack of accountability to changes in the input signal (robustness), lack of transferability to cases not covered by the data (generalization), and lack of explainability (interpretability) became apparent.

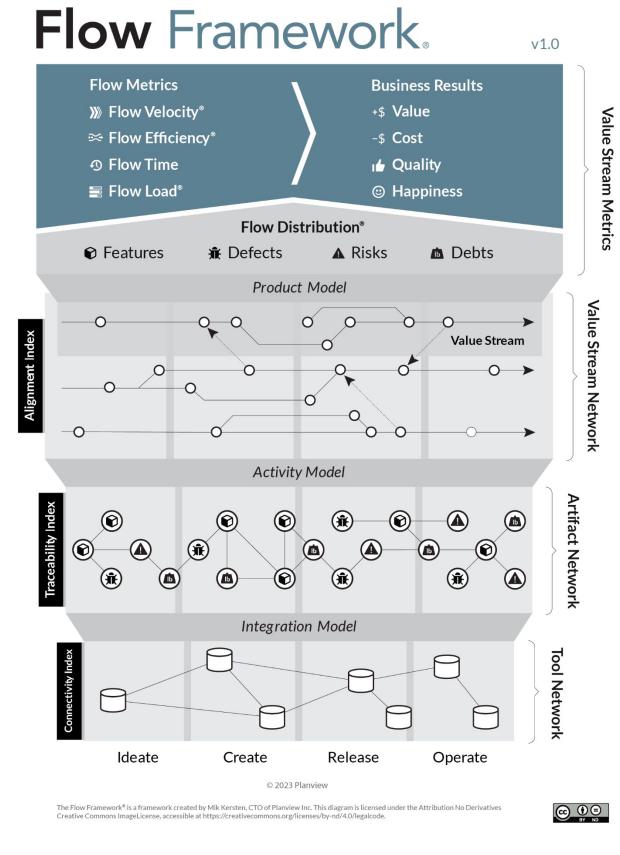
Recently, however, a new overall approach to solving these problems is being advanced by the term “Trusted AI”. Trusted AI aims to create a new generation of machine learning systems that are reliable, allowing use even in critical applications. Developers, domain experts, users, and regulators can rely on these systems to make decisions in safety-critical technical systems. Trusted AI is characterized by a high degree of robustness, transparency, fairness, and verifiability, where the functionality of existing systems is not in any way compromised, but actually enhanced.

**2 MOTIVATION**  
 Current machine learning systems perform quite well and reliably in the context of their training data sets. To be useful, however, they also need to predict, clas-

 <https://doi.org/10.5281/zenodo.1443214>

# Don't forget your value

- AI ... for what?
  - Goals
  - Added value vs. (hidden) costs



<https://flowframework.org/>

Mohamed Ibn Khedher<sup>1</sup>, Afef Awadid<sup>1</sup>, Augustin Lemesle<sup>2</sup> and Zakaria Chihani<sup>2</sup><sup>1</sup>IRT - SystemX, 2 Bd Thomas Gibert, 91120 Palaiseau, France<sup>2</sup>CEA, The French Alternative Energies and Atomic Energy Commission, France{mohamed.ibn-khedher, afef.awadid}@irt-systemx.fr,  
{augustin.lemesle, zakaria.chihani}@cea.fr

# “Meta” capabilities

- Dedicated IDE
- Support invariants (regulations, reqs. conformance)
- Support Quality Assessment

Keywords: Uncertainty in AI, AI Verification, AI Robustness, Adversarial Attacks, Formal Evaluation, Industrial Application.

Abstract: The paper introduces a three-stage evaluation pipeline for ensuring the robustness of AI models, particularly neural networks, against adversarial attacks. The first stage involves formal evaluation, which may not always be feasible. For such cases, the second stage focuses on evaluating the model's robustness against intelligent adversarial attacks. If the model proves vulnerable, the third stage proposes techniques to improve its robustness. The paper outlines the details of each stage and the proposed solutions. Moreover, the proposal aims to help developers build reliable and trustworthy AI systems that can operate effectively in critical domains, where the use of AI models can pose significant risks to human safety.

## 1 INTRODUCTION

Over the last decade, there has been a significant advancement in Artificial Intelligence (AI) and, notably, Machine Learning (ML) has shown remarkable progress in various critical tasks. Specifically, Deep Neural Networks (DNN) have played a transformative role in machine learning, demonstrating exceptional performance in complex applications such as cybersecurity (Jinila and Khedher, 2022) and robotics (Khedher et al., 2021).

Despite the capacity of Deep Neural Networks to handle high-dimensional inputs and address complex challenges in critical applications, recent evidence indicates that perturbations in the input space can lead to incorrect decisions (Papernot et al., 2018). Specifically, it has been observed that DNNs can be easily misled, causing their predictions to change with slight modifications to the inputs. These carefully chosen modifications result in what are known as adversarial examples. These discoveries underscore the critical challenge of ensuring that machine learning systems, especially deep neural networks, function as intended when confronted with perturbed inputs.

Adversarial examples are specially crafted inputs that are designed to fool a machine learning model into making a wrong prediction. These examples are not randomly generated but created with precise calculations. There are various methods for generating

adversarial examples, but most of them focus on minimizing the difference between the distorted input and the original one while ensuring the prediction is incorrect. Some techniques require access to the entire classifier model (white-box attacks), while others only need the prediction function (black-box attacks).

Adversarial attacks pose a significant threat to critical industrial applications, particularly in sectors such as manufacturing, energy, and infrastructure, where precision and reliability are paramount. These attacks, carefully crafted to exploit vulnerabilities in machine learning models, introduce subtle modifications to input data. In critical industrial processes, the consequences of misclassification or data manipulation by adversarial attacks can result in operational failures, compromised safety, and potentially catastrophic outcomes.

To illustrate the severity of adversarial attacks in crucial applications like anomaly detection in the cybersecurity domain, consider Figure 1. An attacker, possessing malicious traffic, can manipulate the traffic by adding imperceptible perturbations, making it appear benign to the cybersecurity system, allowing it to pass undetected. Such attacks can severely compromise the system's ability to identify and mitigate threats, posing significant security risks.

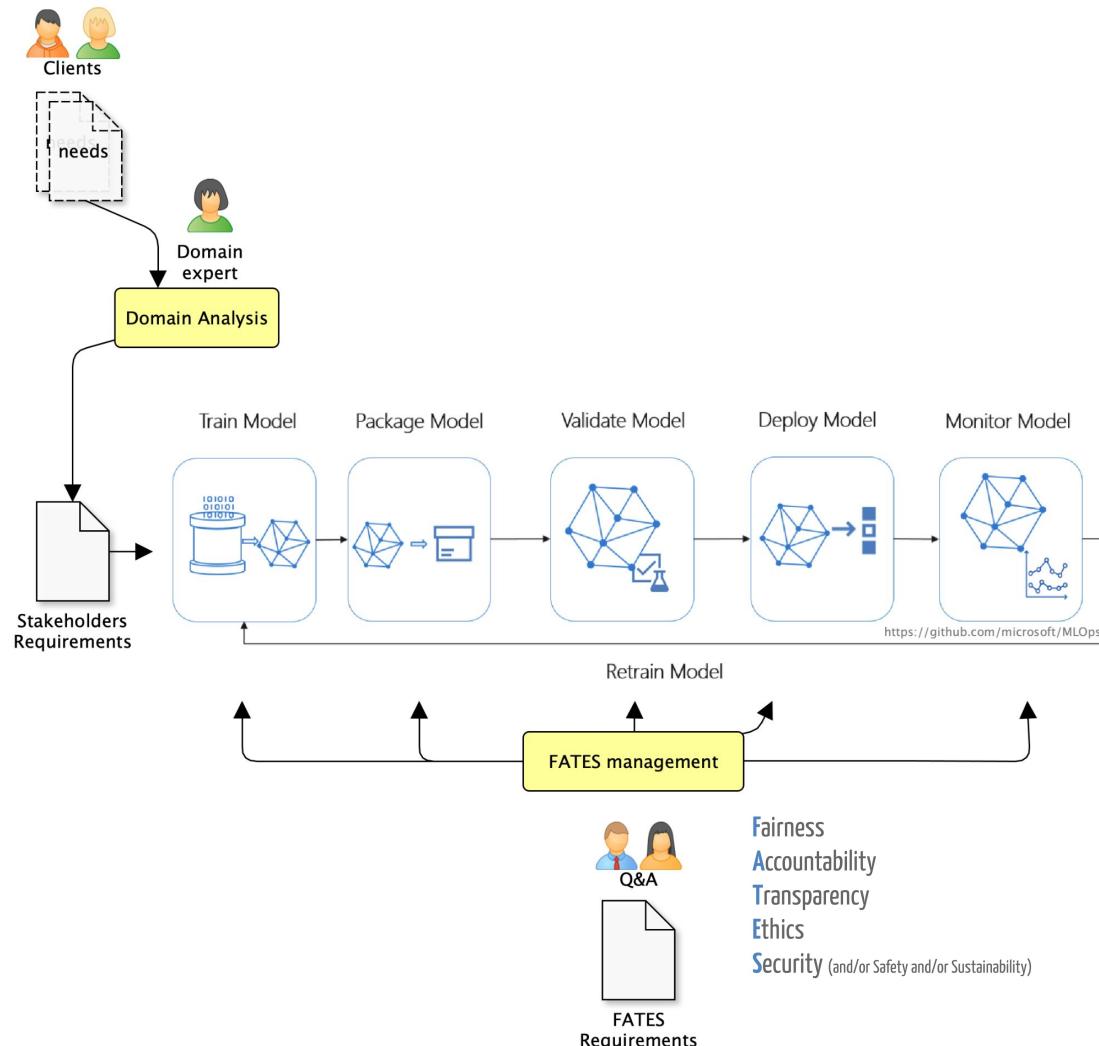
In this paper, we recommend a three-stage pipeline (Khedher et al., 2023) to industrialists to investigate the robustness of their models and, if possi-

<https://hal.science/hal-04477414/document>

# Project organization

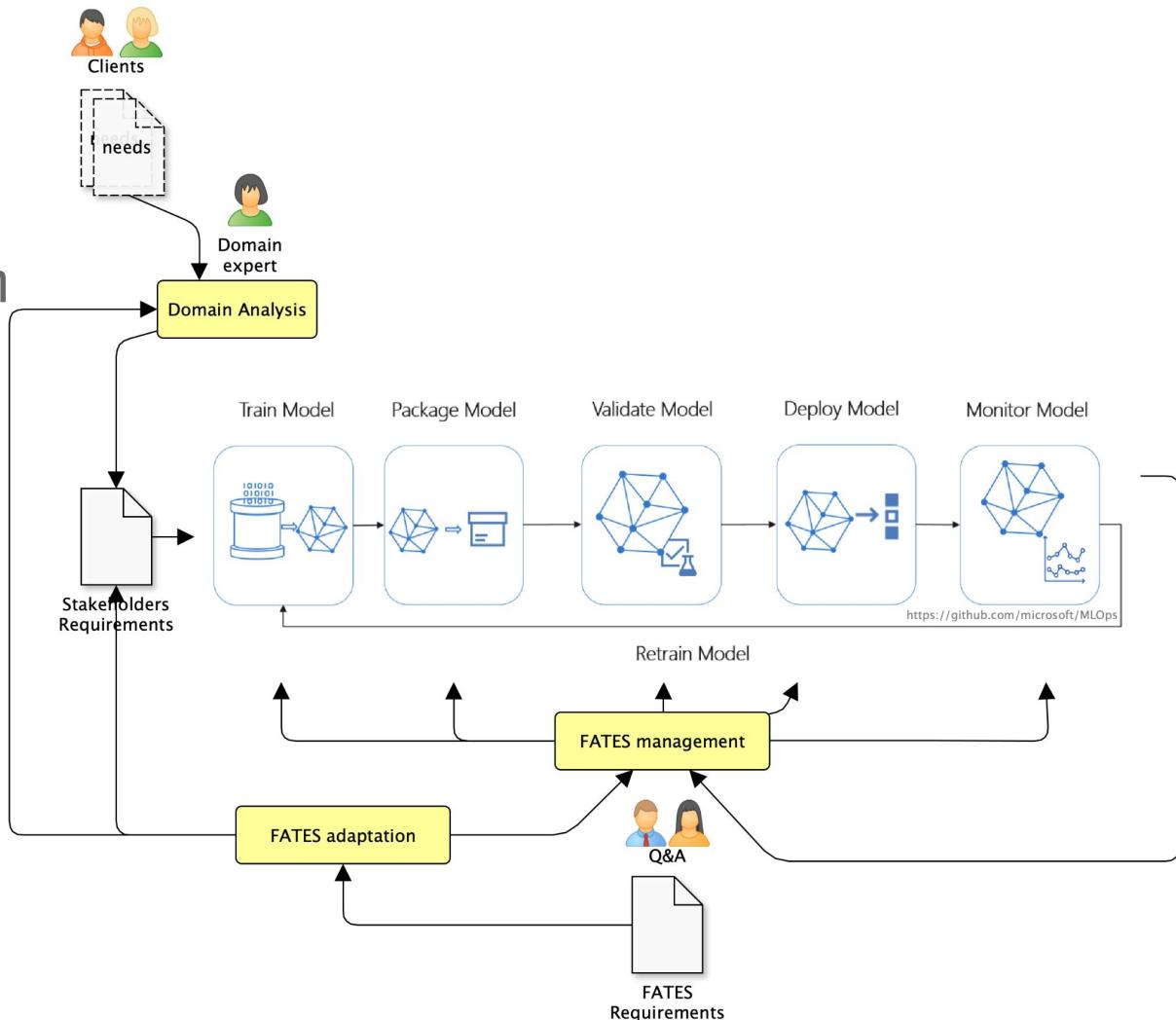
# Big picture

## FATES precise definitions



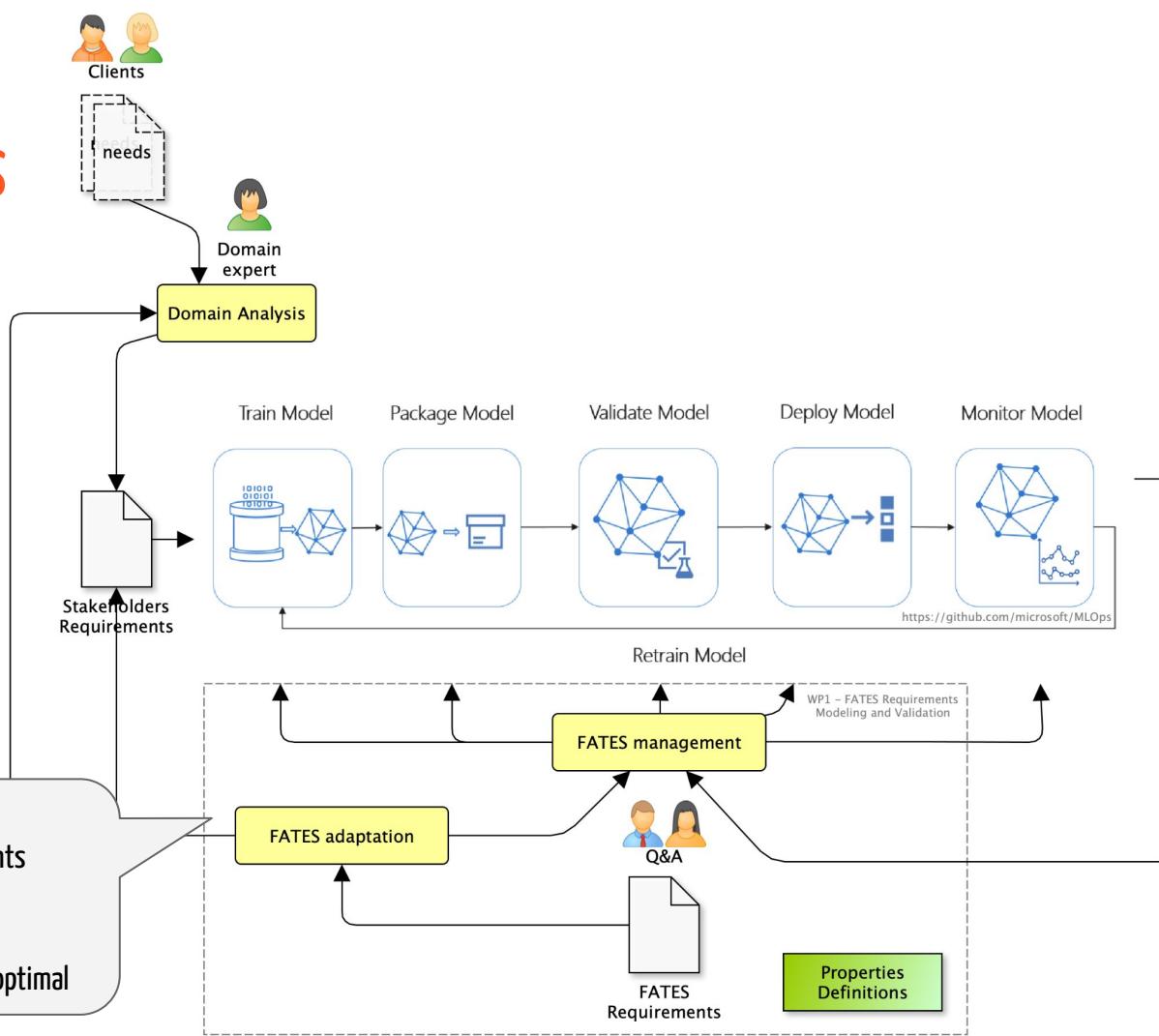
# Big picture

## FATES contextualisation



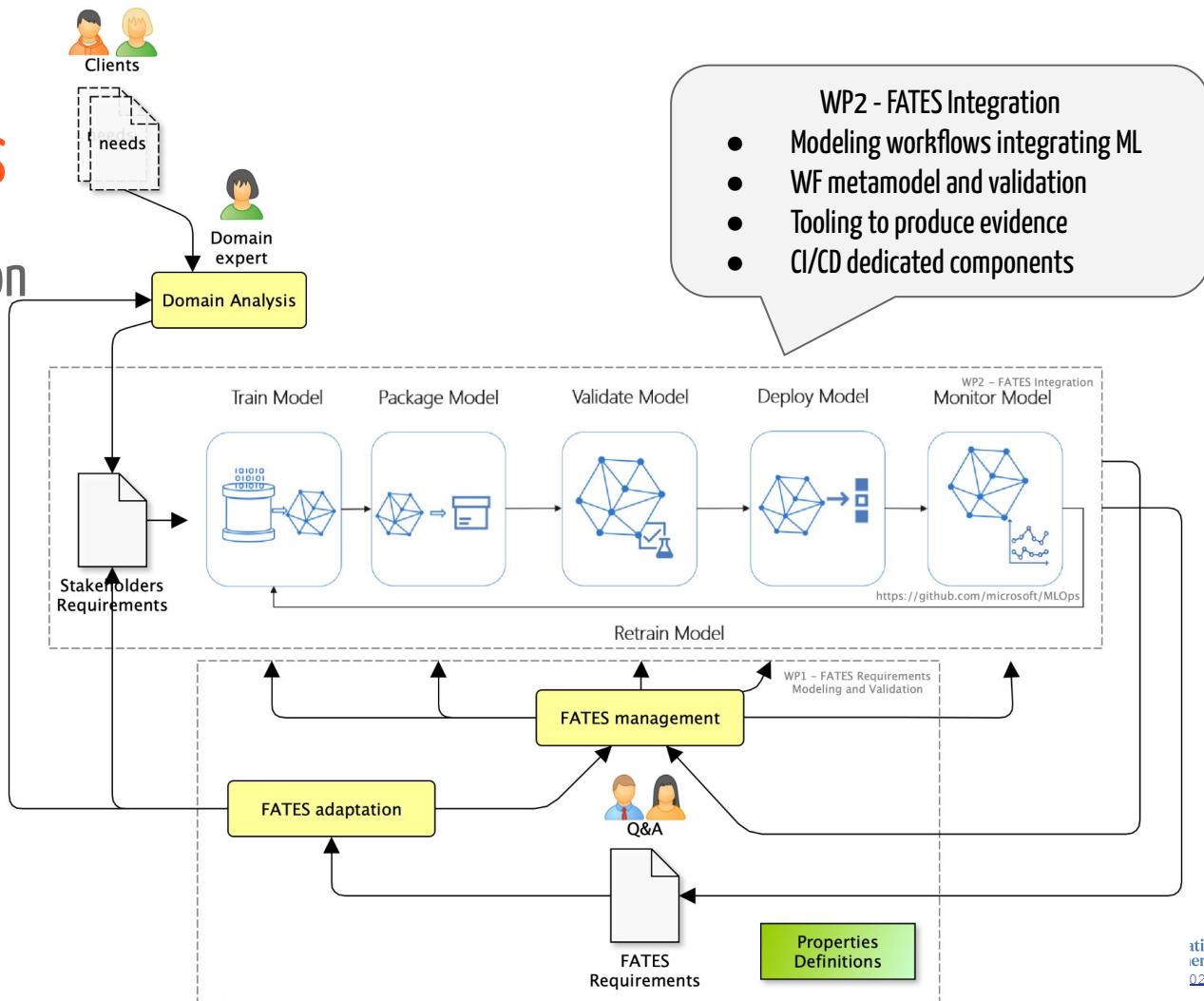
# Work packages

## WP1 - FATES models



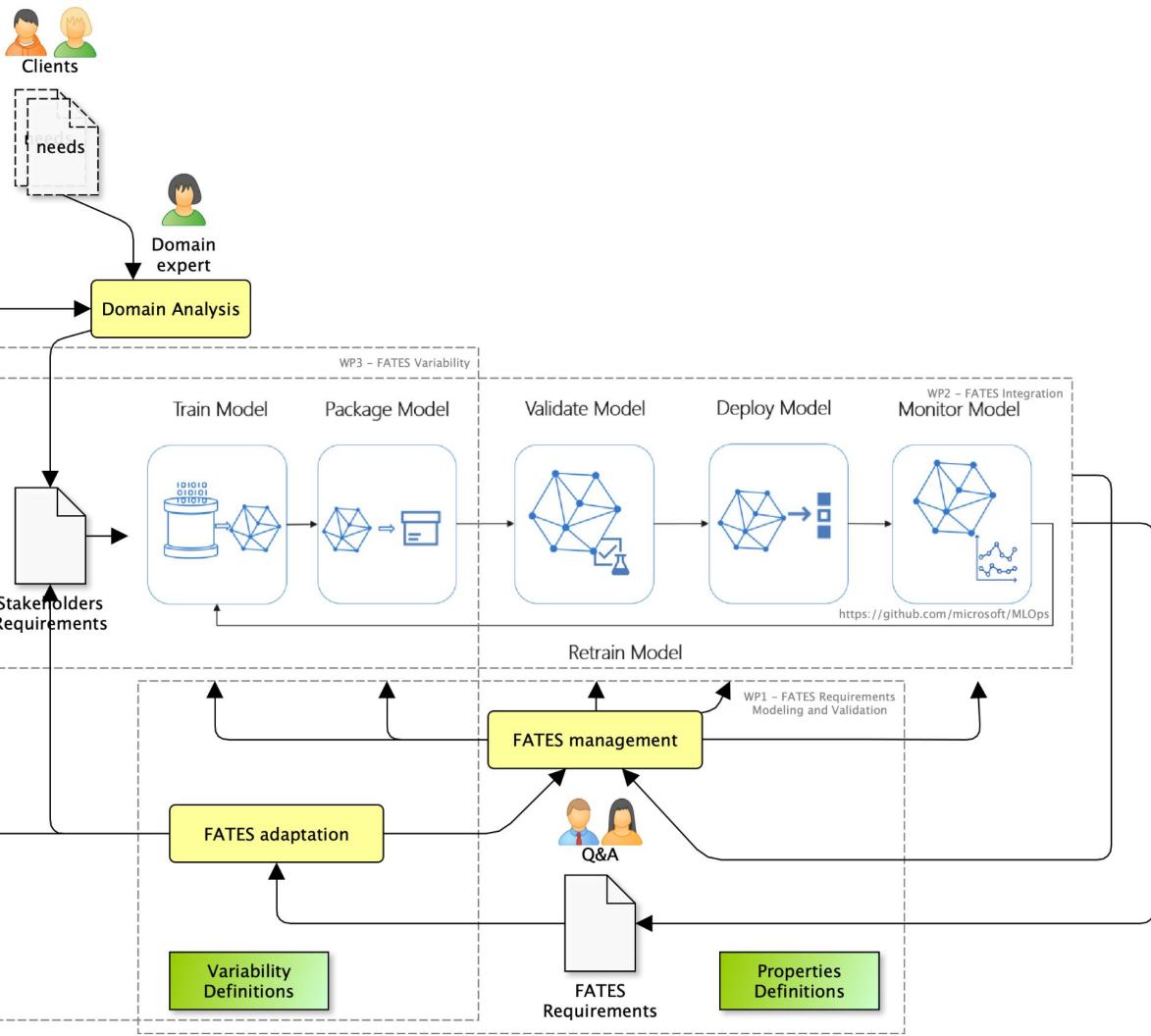
# Work packages

## WP2 - FATES integration



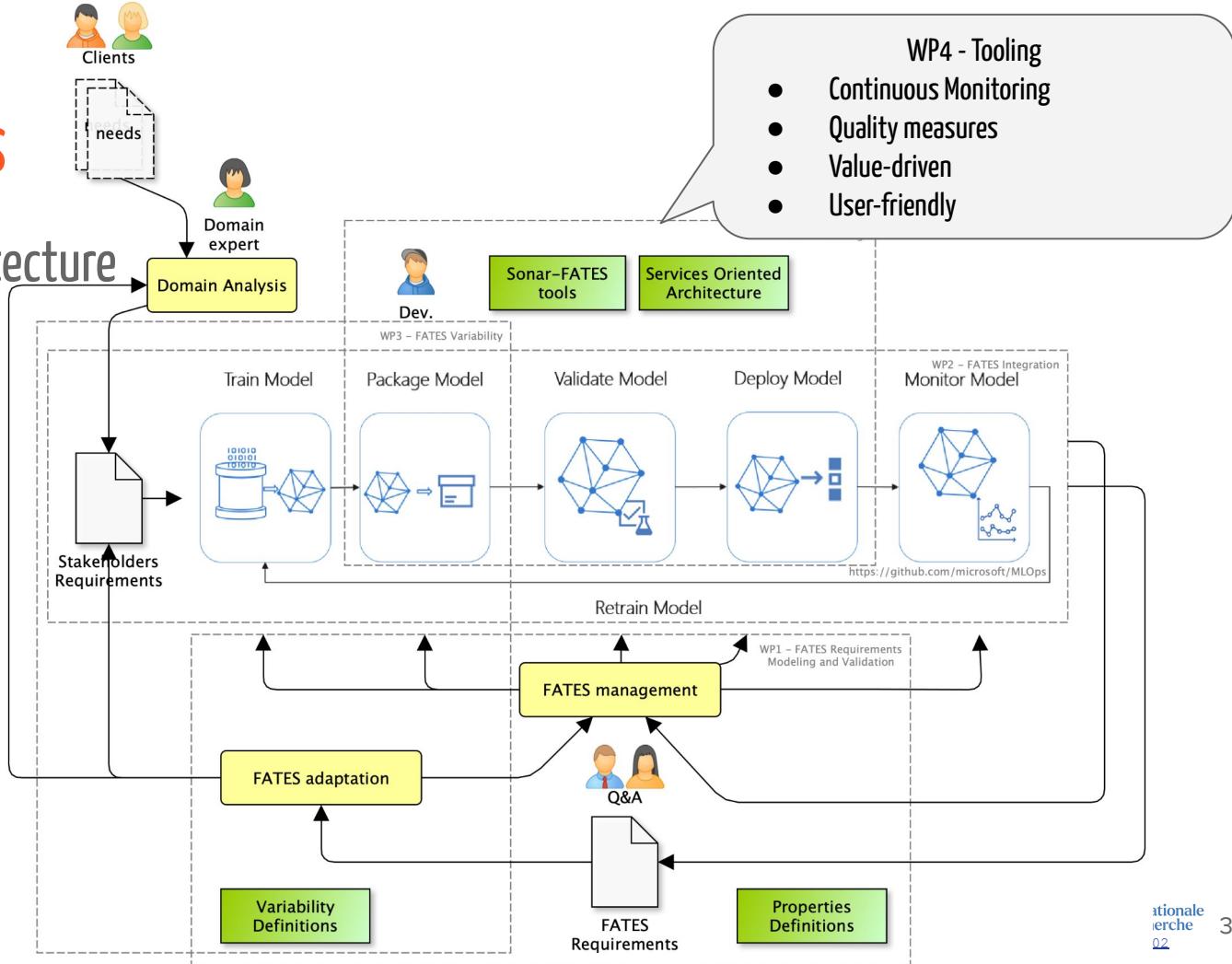
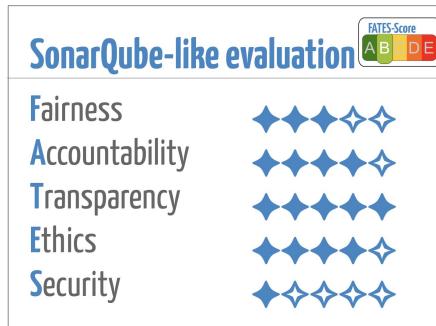
# Work packages

## WP3 - FATES variability

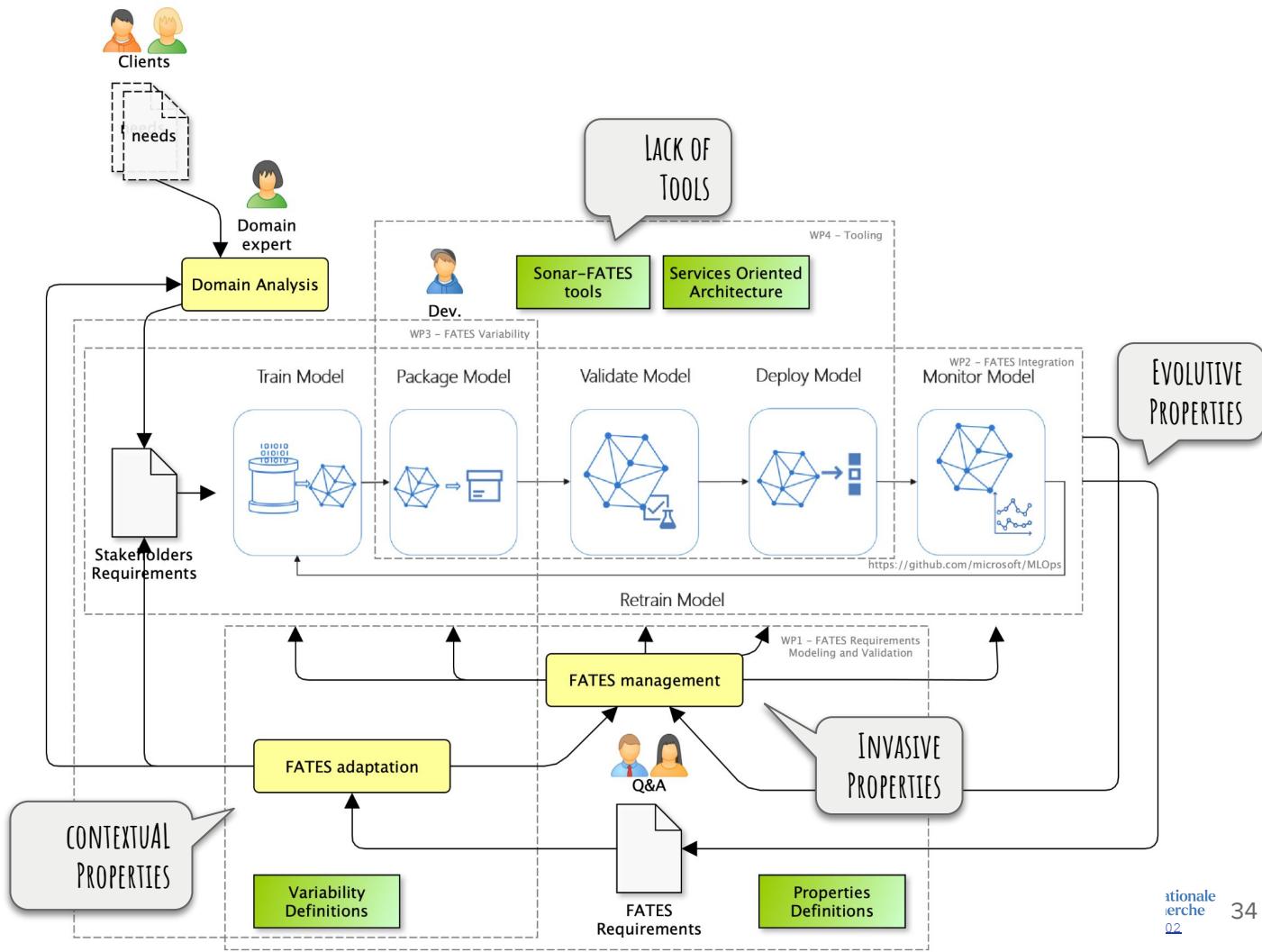


# Work packages

## WP4 - Tooling & Architecture

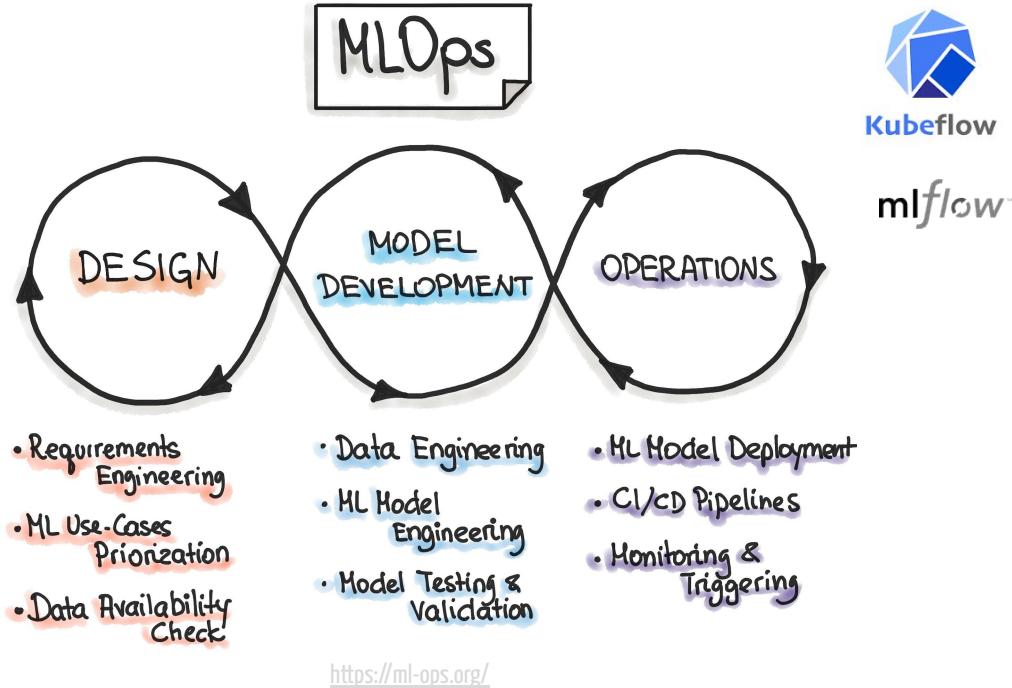


# Key concerns



# Support ML in Operations

- Process (yaml)
- Tooling and support
- FATES properties justification



# Current members (permanent only)



Institut de Recherche  
en Informatique de Toulouse  
CNRS - Toulouse INP - UT - UTC - UT2



O. Teste



M. Pantel



J.-M. Bruel



M. Blay-Fornarino



P. Collet



E. Precioso



M. Riveill



S. Mosser

## Context

2024 – 2028

600K€

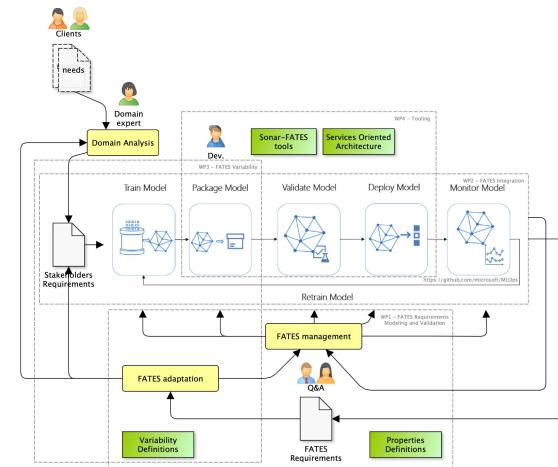
2 PhDs & 2 Postdocs/Researchers

# We need you!

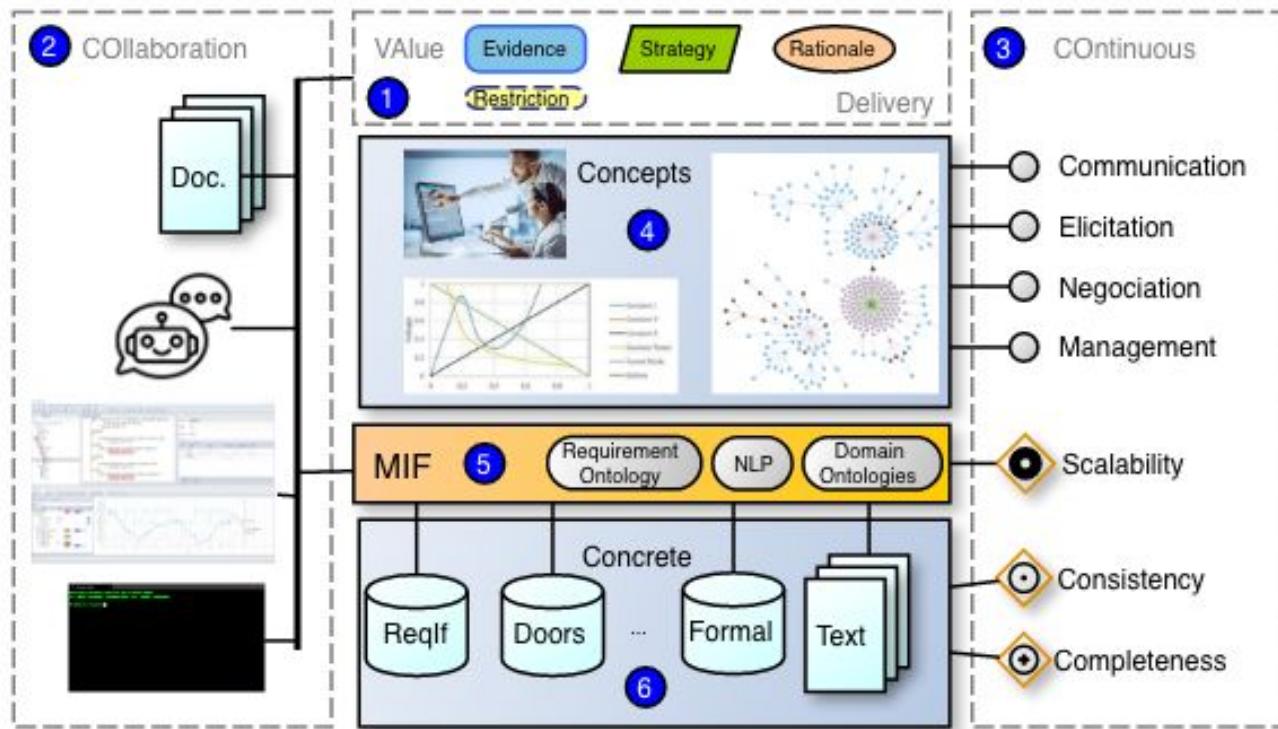


# Collaboration opportunity

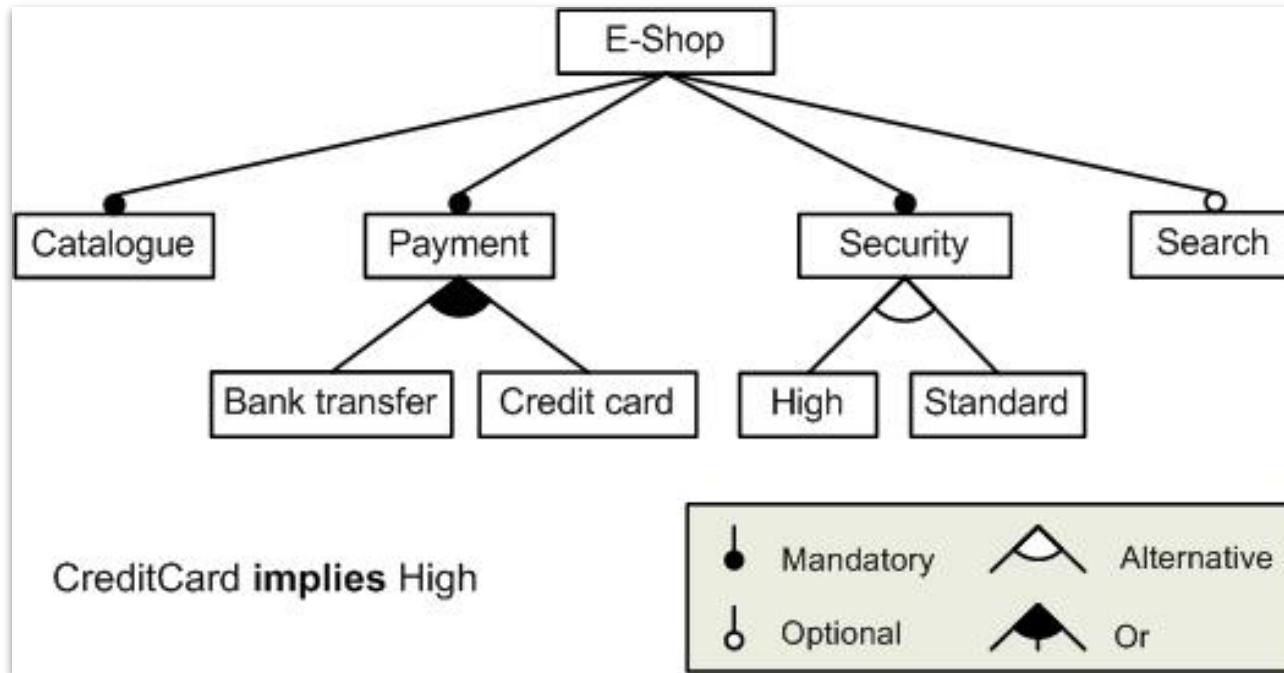
- Properties formalisation
- Features model definition
- Justification diagrams
  - ◆ MS properties
  - ◆ AI-Act compliance



# Properties formalisation

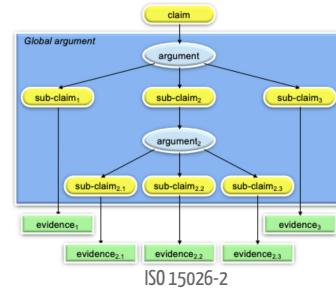
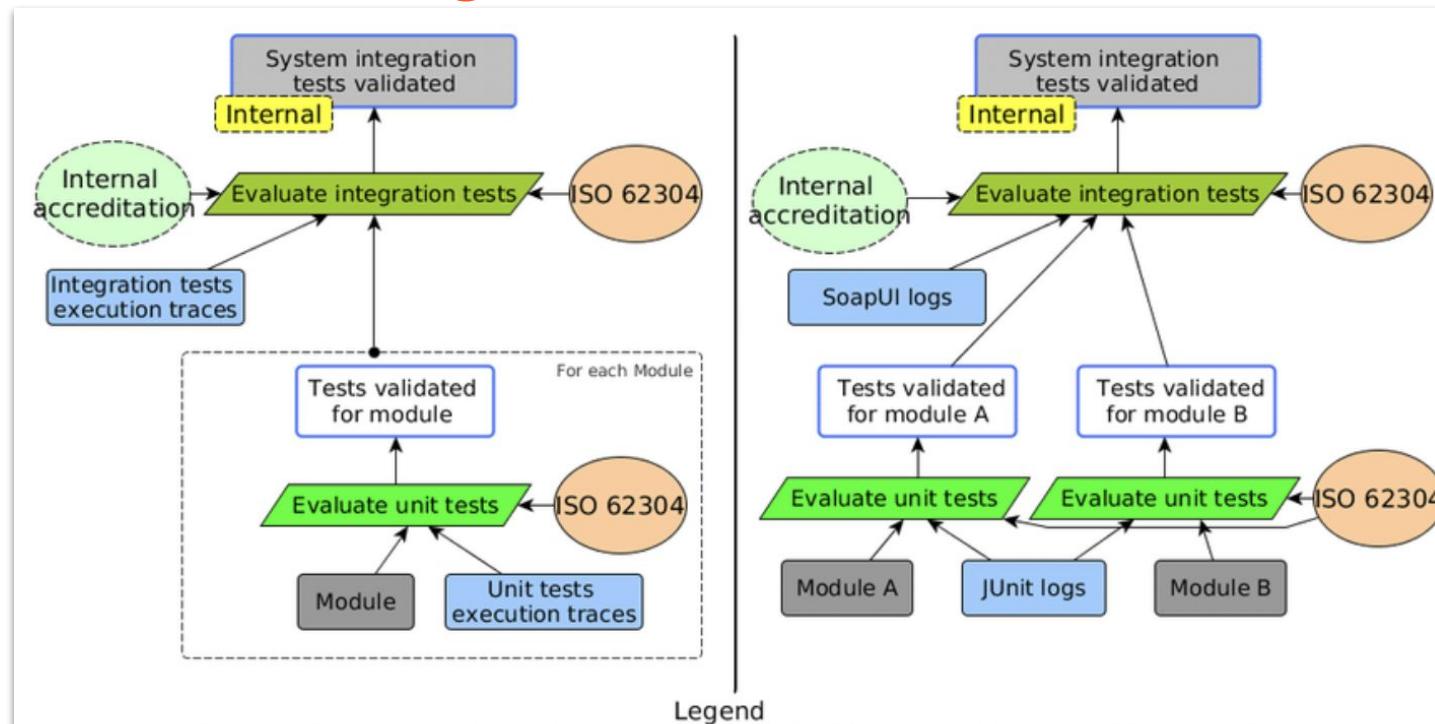


# Feature model



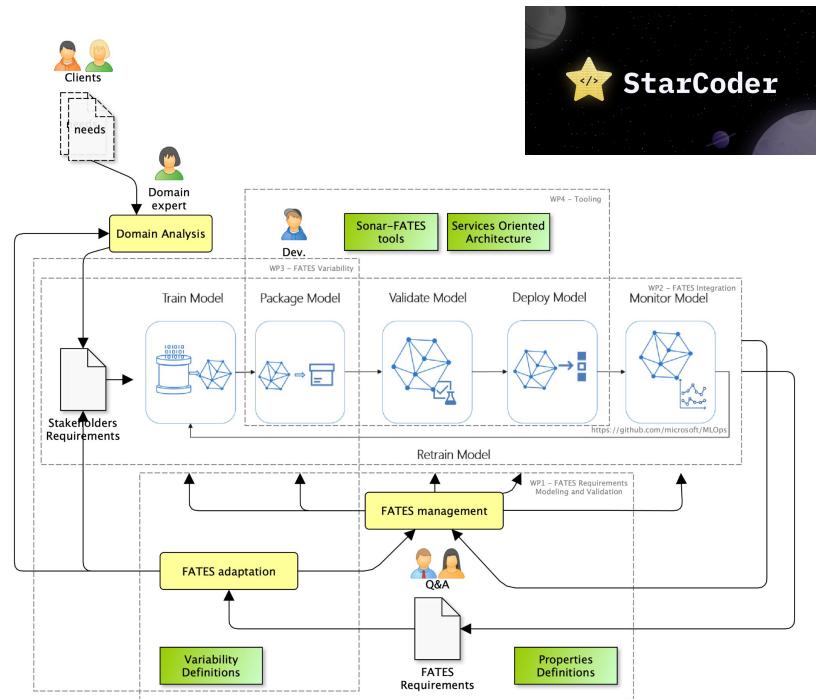
[https://en.wikipedia.org/wiki/Feature\\_model](https://en.wikipedia.org/wiki/Feature_model)

# Justification diagrams (ISO-IEC-IEEE 15026-2)



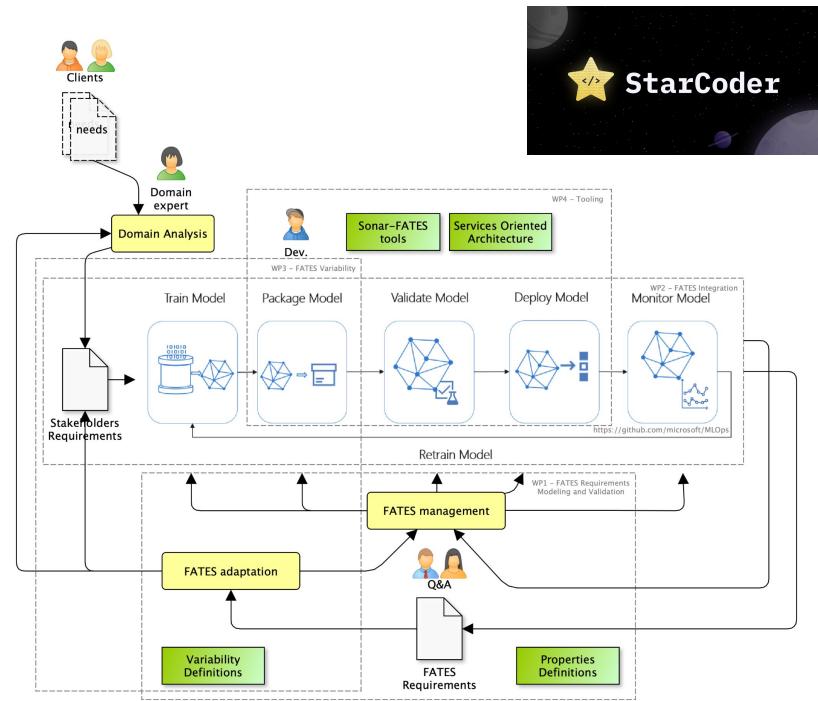
# Collaboration expectations

- Contact person to validate formal definitions (beginning of the project)
- Eng. time to implement some of the services (end of the project)
- 1/4 ETP for 4 years => 1ETP budget ?



# Project constraints

- 36-48 months
- 600K€ max budget
- Deadline: April 24<sup>th</sup>



# Critères pour les études de cas

- Ils ne présentent pas, a priori, d'**objectifs qui ne respectent pas les propriétés FATES**
- Ils ne rentrent **pas** dans la catégorie des IA initialement proscrites par l'Europe (emploi, justice, éducation, santé)
- Ils présentent **plusieurs étapes de traitement** (pour le côté DevOPS) et nous pouvons placer des composants d'analyses entre ces étapes. Par exemples :
  - nous avons le code qui a produit le modèle et nous avons accès aux interrogations et aux réponses en production
  - nous sommes sur un workflow comme on peut en avoir avec langchain et nous pouvons statiquement analyser le workflows pour l'équiper de composants de débiaisage, de monitoring (transparence), etc.
- Si possible, le modèle continue à apprendre en production

Exemple : utilisation des algos de voitures autonome pour analyser le nombre et le comportement des espèces animales et l'influence de l'activité humaine sur leur comportement (sentiers pour les loups du mercantour, présence de la biodiversité pour les poissons, ...).

One last thing...

# GDR *Sciences du Logiciel*

- Please, register myGDR (<https://mygdr.hosted.lip6.fr/>)
- For SciLog :
  - “Défi” IA4SE (AI for Software Engineering)
  - “Défi” SE4AIA (Software Engineering for AI-based Applications)
  - “GT” IA
- GDR RADIA

# Discussions time!



BY NC SA



<https://bit.ly/jmbruel>



@jmbruel

<https://fates-mlops.org>

Get the slides



Institut de Recherche  
en Informatique de Toulouse  
CNRS - Toulouse INP - UT - UTC - UT2

UNIVERSITÉ  
CÔTE D'AZUR



inria

McMaster  
University



anr<sup>®</sup>  
agence nationale  
de la recherche