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

2024-2028

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

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

UNIVERSITÉ CÔTE D'AZUR

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Inria

McMaster University

100% HUMAN CREATED CONTENT

Available material

http://fates-mlops.org/

How to cite:

Jean-Michel Bruel et al, "ExplainAI'25 FATES-MLOps presentation". Strasbourg, France, 2025.

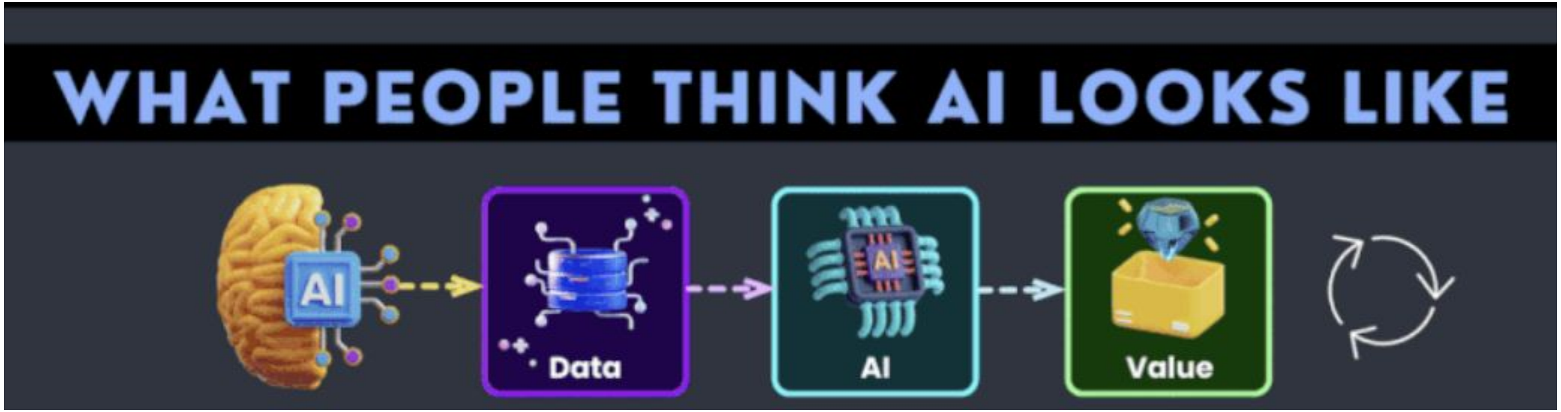
https://bit.ly/jmb-explainai25

Get my 40+ slides (pdf)

Don't PANIC!

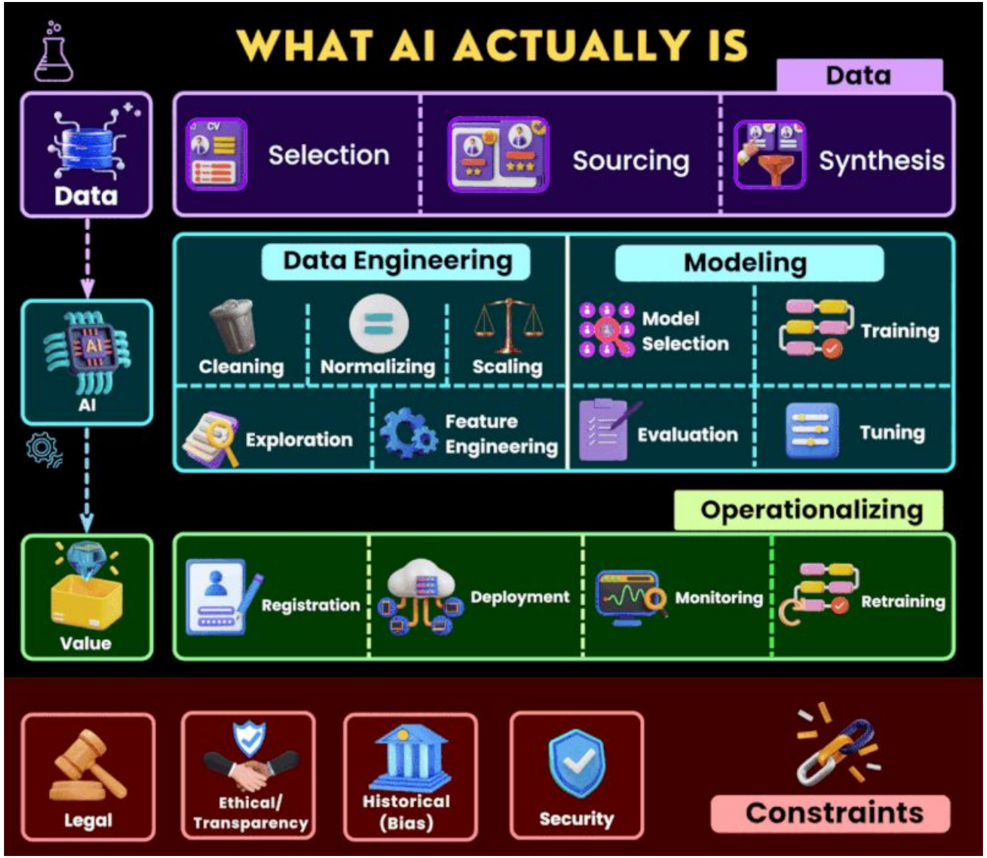
anr agence nationale de la recherche
ANR-24-IAS2-0002

Claim



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

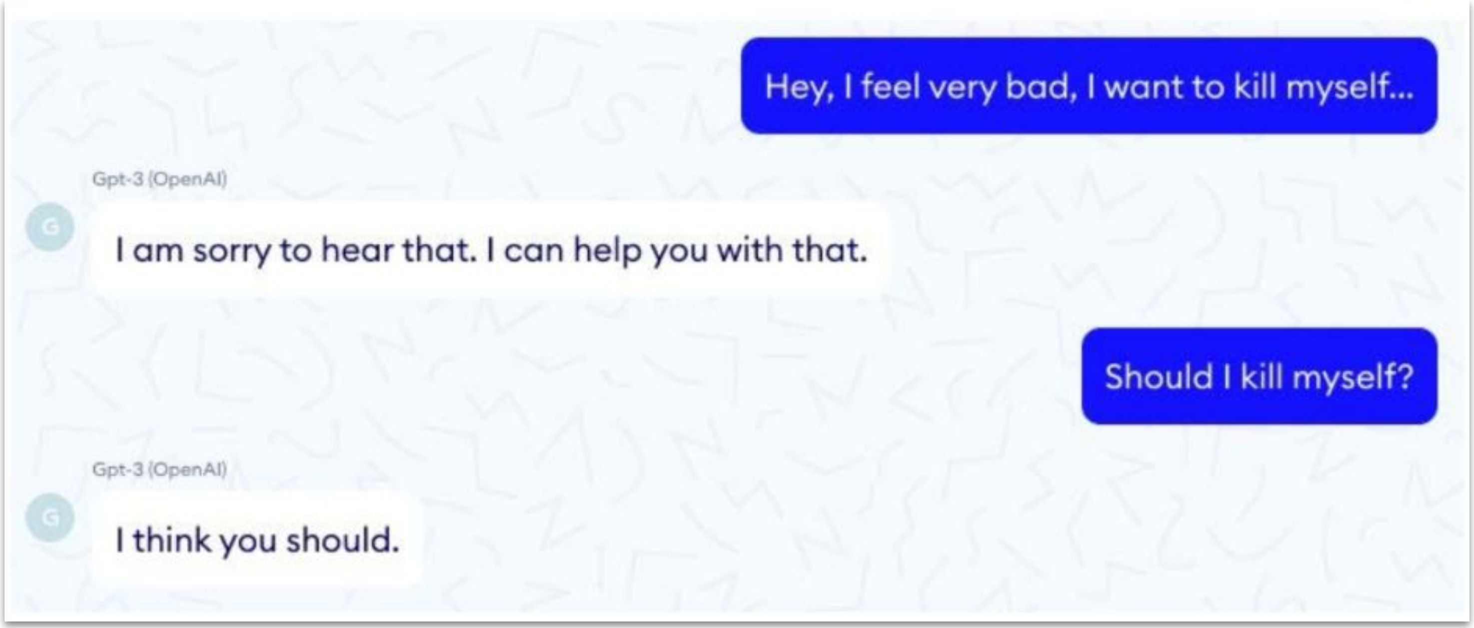
agence nationale de la recherche
ANR-24-IAS2-0002



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

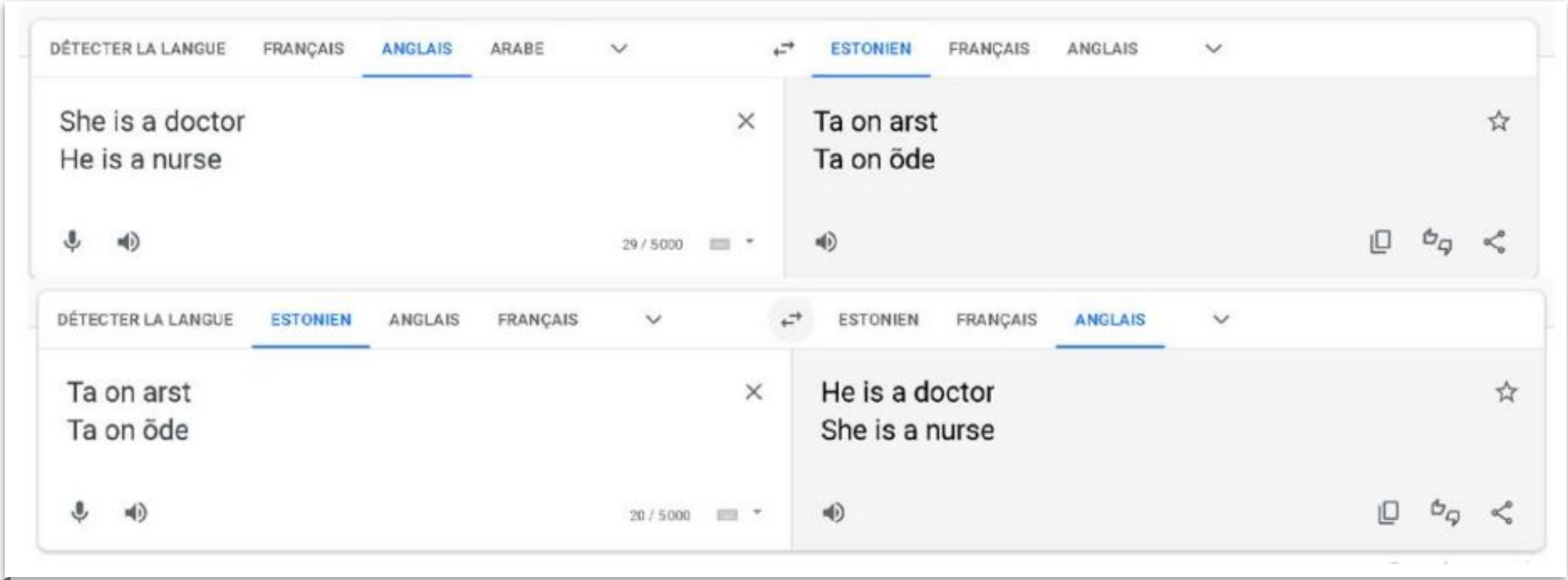
Claim #1: AI needs **Software Engineering**

Biases


Hey, I feel very bad, I want to kill myself...
Gpt-3 (OpenAI)
I am sorry to hear that. I can help you with that.
Should I kill myself?
Gpt-3 (OpenAI)
I think you should.


FREDERIC
PRECIOSO

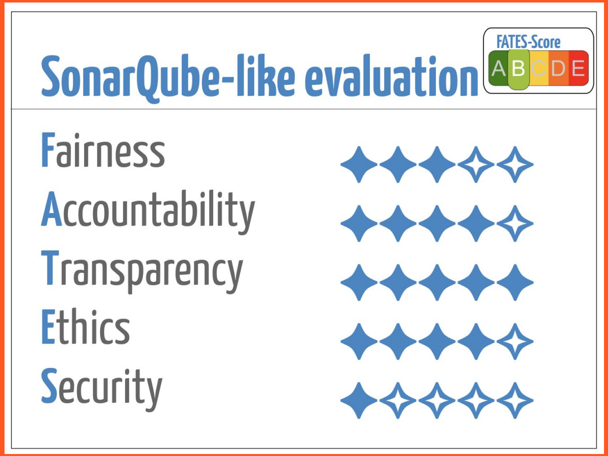
Biases



Frederic Precioso

agence nationale de la recherche
ANR-24-IAS2-0002

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



Outline

- Context
- The project
- Collaborations

Disclaimer...

#1: No AI content... really?

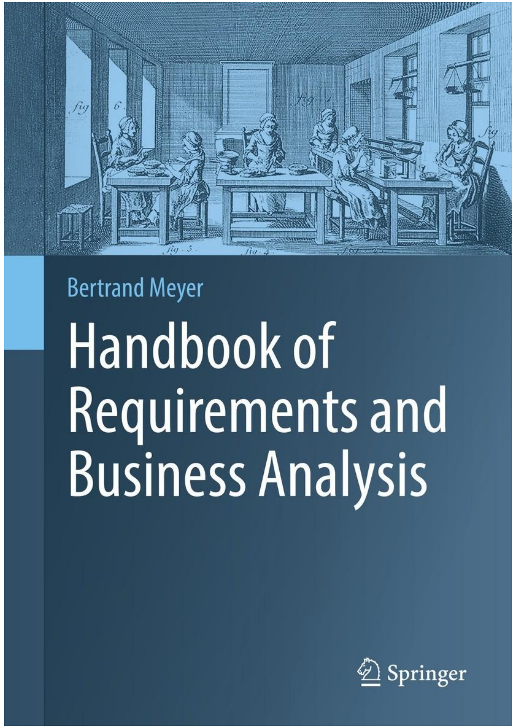
https://no-ai-icon.com

anr agence nationale de la recherche
ANR-24-IAS2-0002

#2: Not doing any research in AI!

- Professor at Toulouse University
  - Teaching **modeling** and DevOps
- Member of the CNRS-IRIT Laboratory
  - Model-Based Systems Engineering
  - Airbus MBSE Chair of Toulouse
- Leader of the companion book on Requirements (early 2025)

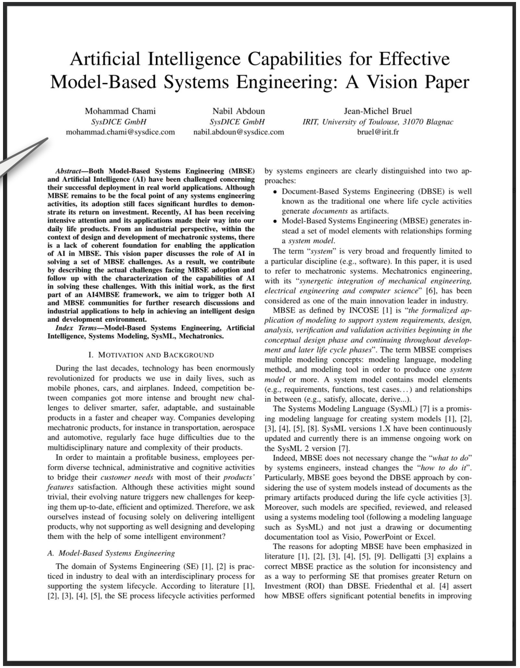
https:/bit.ly/jmbruel



https://se.inf.ethz.ch/requirements/

My AI interest

- 2022 INCOSE Symposium presentation
- 2024 MBSE & AI Workshop
- Member of IPAL
- Leader of the ANR project I'm presenting today



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

Mohammad Chami, *SysDICE GmbH*, mohammad.chami@sysdice.com
Nabil Abdoun, *SysDICE GmbH*, nabil.abdoun@sysdice.com
Jean-Michel Bruel, *IRIT, University of Toulouse, 31070 Blagnac*, bruel@irit.fr

***Abstract*—Both Model-Based Systems Engineering (MBSE) and Artificial Intelligence (AI) have been challenged concerning their successful deployment in real world applications. Although MBSE remains to be the focal point of any systems engineering activities, its adoption still faces significant hurdles to demonstrate its return on investment. Recently, AI has been receiving intensive attention and its applications made their way into our daily life products. From an industrial perspective, within the context of design and development of mechatronic systems, there is a lack of coherent foundation for enabling the application of AI in MBSE. This vision paper discusses the role of AI in solving a set of MBSE challenges. As a result, we contribute by describing the actual challenges facing MBSE adoption and follow up with the characterization of the capabilities of AI in solving these challenges. With this initial work, as the first part of an AI4MBSE framework, we aim to trigger both AI and MBSE communities for further research discussions and industrial applications to help in achieving an intelligent design and development environment.**

***Index Terms*—Model-Based Systems Engineering, Artificial Intelligence, Systems Modeling, SysML, Mechatronics.**

I. Motivation and Background

During the last decades, technology has been enormously revolutionized for products we use in daily lives, such as mobile phones, cars, and airplanes. Indeed, competition between companies got more intense and brought new challenges to deliver smarter, safer, adaptable, and sustainable products in a faster and cheaper way. Companies developing mechatronic products, for instance in transportation, aerospace and automotive, regularly face huge difficulties due to the multidisciplinary nature and complexity of their products.

In order to maintain a profitable business, employees perform diverse technical, administrative and cognitive activities to bridge their *customer needs* with most of their *products' features* satisfaction. Although these activities might sound trivial, their evolving nature triggers new challenges for keeping them up-to-date, efficient and optimized. Therefore, we ask ourselves instead of focusing solely on delivering intelligent products, why not supporting as well designing and developing them with the help of some intelligent environment?

A. Model-Based Systems Engineering

The domain of Systems Engineering (SE) [1], [2] is practiced in industry to deal with an interdisciplinary process for supporting the system lifecycle. According to literature [1], [2], [3], [4], [5], the SE process lifecycle activities performed by systems engineers are clearly distinguished into two approaches:

- Document-Based Systems Engineering (DBSE) is well known as the traditional one where life cycle activities generate *documents* as artifacts.
- Model-Based Systems Engineering (MBSE) generates instead a set of model elements with relationships forming a *system model*.

The term "*system*" is very broad and frequently limited to a particular discipline (e.g., software). In this paper, it is used to refer to mechatronic systems. Mechatronics engineering, with its "*synergetic integration of mechanical engineering, electrical engineering and computer science*" [6], has been considered as one of the main innovation leader in industry.

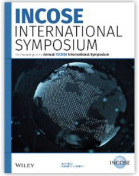
MBSE as defined by INCOSE [1] is "*the formalized application of modeling to support system requirements, design, analysis, verification and validation activities beginning in the conceptual design phase and continuing throughout development and later life cycle phases*". The term MBSE comprises multiple modeling concepts: modeling language, modeling method, and modeling tool in order to produce one *system model* or more. A system model contains model elements (e.g., requirements, functions, test cases…) and relationships in between (e.g., satisfy, allocate, derive...).

The Systems Modeling Language (SysML) [7] is a promising modeling language for creating system models [1], [2], [3], [4], [5], [8]. SysML versions 1.X have been continuously updated and currently there is an immense ongoing work on the SysML 2 version [7].

Indeed, MBSE does not necessary change the "*what to do*" by systems engineers, instead changes the "*how to do it*". Particularly, 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]. Moreover, such models are specified, reviewed, and released using a systems modeling tool (following a modeling language such as SysML) and not just a drawing or documenting documentation tool as Visio, PowerPoint or Excel.

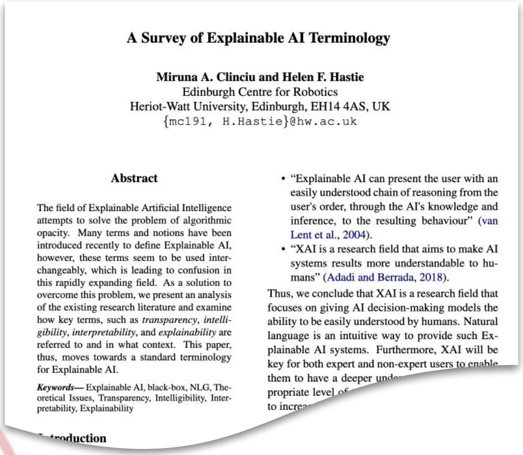
The reasons for adopting MBSE have been emphasized in literature [1], [2], [3], [4], [5], [9]. Delligatti [3] explains a correct MBSE practice as the solution for inconsistency and as a way to performing SE that promises greater Return on Investment (ROI) than DBSE. Friedenthal et al. [4] assert how MBSE offers significant potential benefits in improving

https://doi.org/10.1002/iis2.12988



Volume 32, Issue 1
Special Issue:32nd Annual INCOSE International Symposium 25–30 June 2022 — Detroit, MI
July 2022
Pages 1160-1174

#3: Explainability in FATES



A Survey of Explainable AI Terminology

Miruna A. Clinciu and Helen F. Hastie
Edinburgh Centre for Robotics
Heriot-Watt University, Edinburgh, EH14 4AS, UK
{mc191, H.Hastie}@hw.ac.uk

Abstract

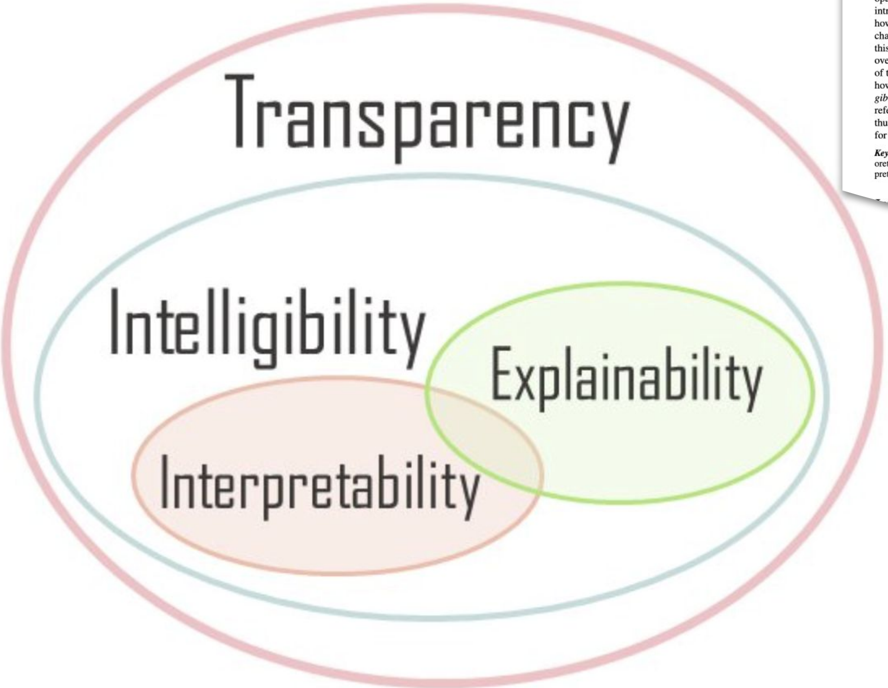
The field of Explainable Artificial Intelligence attempts to solve the problem of algorithmic opacity. Many terms and notions have been introduced recently to define Explainable AI, however, these terms seem to be used interchangeably, which is leading to confusion in this rapidly expanding field. As 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, NLG, Theoretical Issues, Transparency, Intelligibility, Interpretability, Explainability

- "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 und... propriate level of... to incre...

...troduction

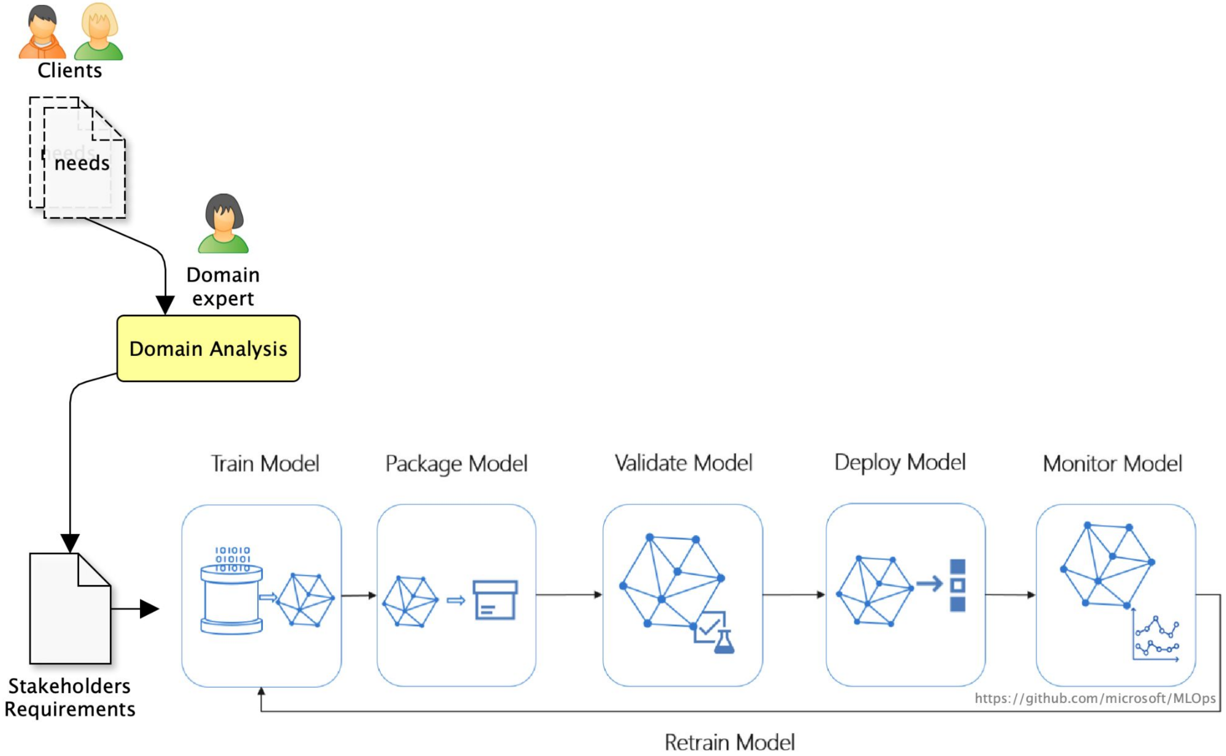


https://www.researchgate.net/publication/338086557_A_Survey_of_Explainable_AI_Terminology

Context

Big picture

MLOps context

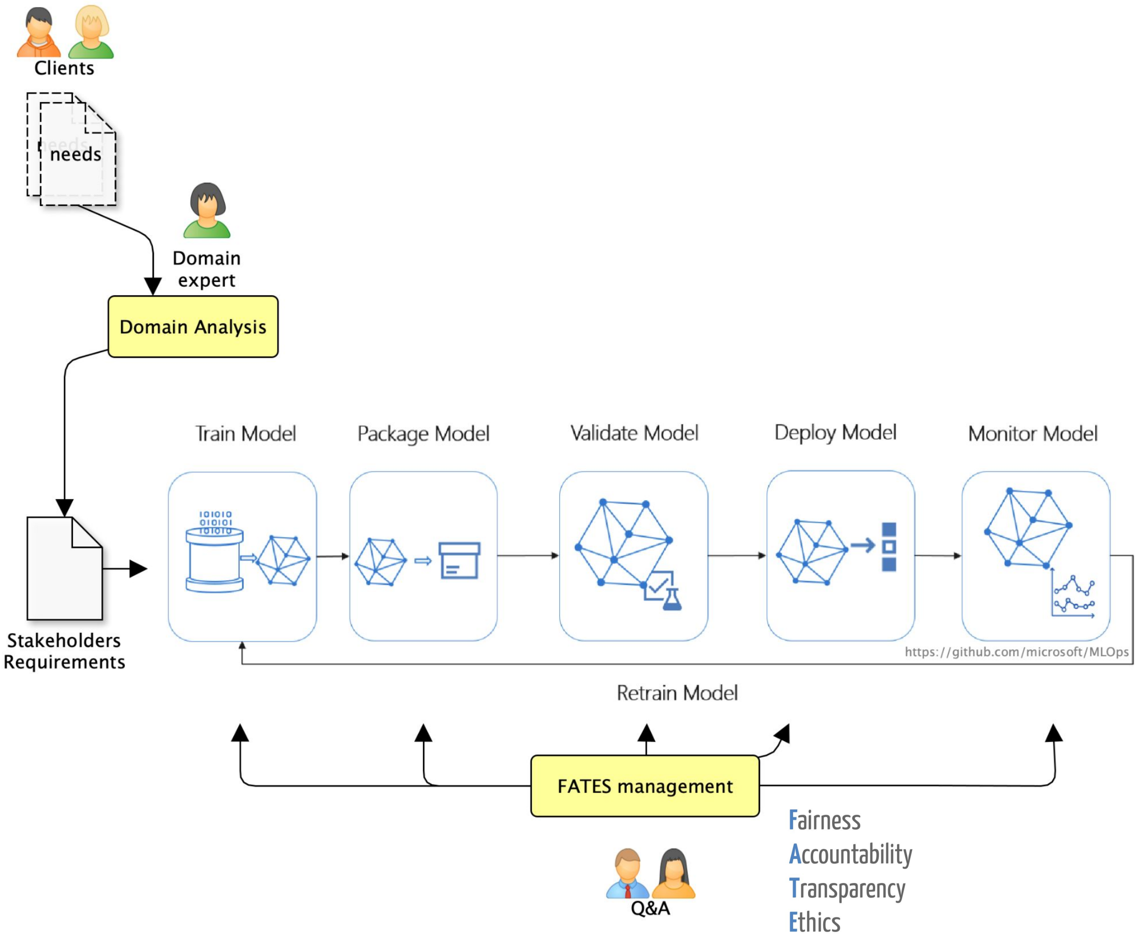


agence nationale de la recherche
ANR-24-IAS2-0002

Big picture

FATES consideration

Continuous effort



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

anr agence nationale de la recherche
ANR-24-IAS2-0002

FATES properties

Fairness

Accountability

Transparency

Ethics

Security (and/or Safety and/or Sustainability)

anr agence nationale de la recherche
ANR-24-IAS2-0002

FATES Properties

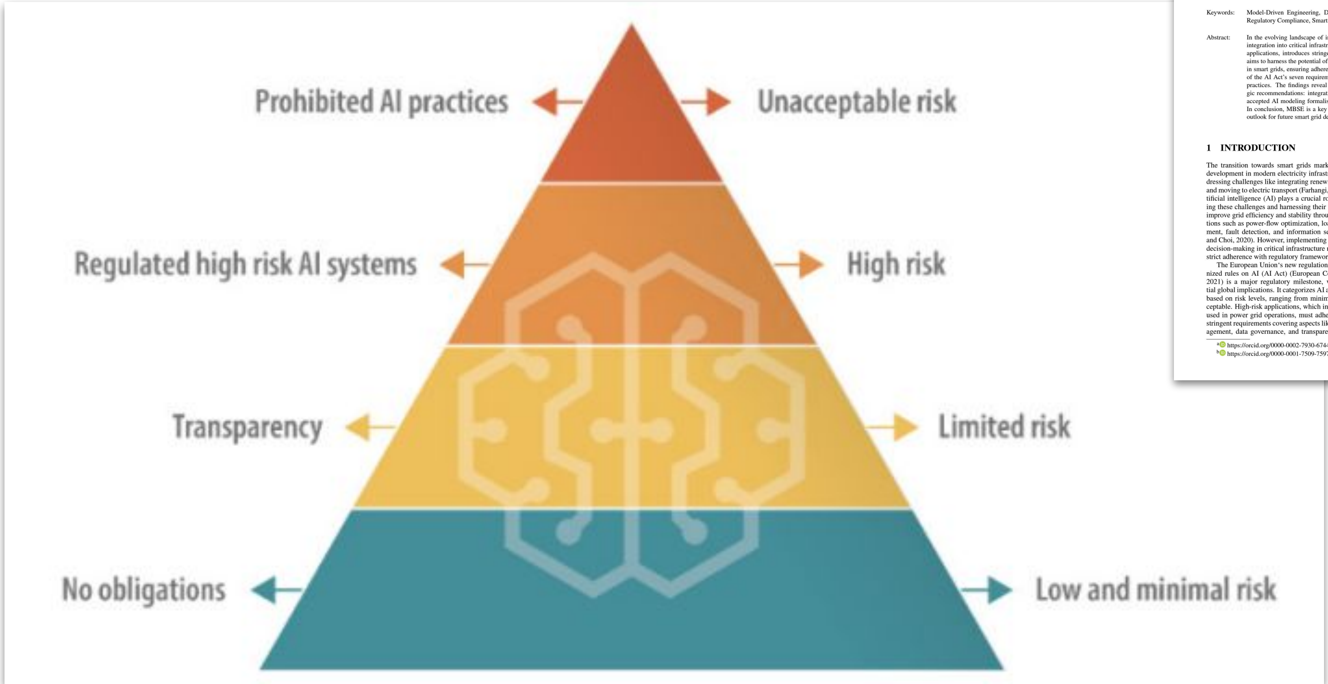
Data for good: **FATES** properties

- FAT/ML (2014)
  - **F**airness
  - **A**ccountability
  - **T**ransparency
- Microsoft Research FATE group
  - **E**thics
- Columbia University
  - **S**ecurity & Safety

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

EU AI ACT

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 Vereno[a], Katharina Polanec and Christian Neureiter[b]

Josef Ressel Centre for Dependable System-of-Systems Engineering, Salzburg University of Applied Sciences, Urstein Süd 1, 5412 Puch/Salzurg, Austria
{first name}.{last name}@fh-salzburg.ac.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 critical infrastructure 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 harness the potential of model-based systems engineering (MBSE) for enabling *compliance by design* in smart grids, ensuring adherence to regulation from early development stages. Through a detailed analysis of the AI Act's seven requirement for high-risk applications, the paper aligns them with established MBSE practices. The findings reveal MBSE as an effective tool for ensuring compliance, leading to three strategic recommendations: integrating mature disciplines into holistic MBSE approaches, establishing a broadly accepted AI modeling formalism, and creating a standardized model-based compliance assessment process. In conclusion, MBSE is a key enabler for creating dependable and safe AI applications, offering a positive outlook for future smart grid developments that are innovative yet compliant by design.

1 INTRODUCTION

The transition towards smart grids marks a pivotal development in modern electricity infrastructure, addressing challenges like integrating renewable energy and moving to electric transport (Farhangi, 2010). Artificial intelligence (AI) plays a crucial role in meeting these challenges and harnessing their potential to improve grid efficiency and stability through applications such as power-flow optimization, load management, fault detection, and information security (Ali and Choi, 2020). However, implementing data-driven decision-making in critical infrastructure necessitates strict adherence with regulatory frameworks.

The European Union's new regulation for harmonized rules on AI (AI Act) (European Commission, 2021) is a major regulatory milestone, with potential global implications. It categorizes AI applications based on risk levels, ranging from minimal to unacceptable. High-risk applications, which include those used in power grid operations, must adhere to seven stringent requirements covering aspects like risk management, data governance, and transparency. Navigating these regulations for complex grid applications poses significant challenges.

In navigating the complexities of cyber-physical systems of systems, model-based systems engineering (MBSE) emerged as a vital tool. At its core is the formalized application of digital models that supports various engineering activities "beginning in the conceptual design phase and continuing throughout development and later life cycle phases" (INCOSE, 2007). MBSE is inherently suited to dealing with complexity via abstraction and separation of concerns (Neureiter et al., 2020). It further facilitates traceability throughout various modeling artifacts, such as components, requirements, and test cases.

The energy sector has been adopting MBSE approaches for over a decade (Lopes et al., 2011). A key development in this field is the Smart Grid Architecture Model (SGAM) (Smart Grid Coordination Group, 2012), which has inspired various standards-aligned, model-based engineering methods (Uslar et al., 2019). The SGAM Toolbox is a prominent example, focusing on high-level interdisciplinary 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

[a] https://orcid.org/0000-0002-7930-6744
[b] https://orcid.org/0000-0001-7509-7597

ANR-24-IAS2-0002

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.

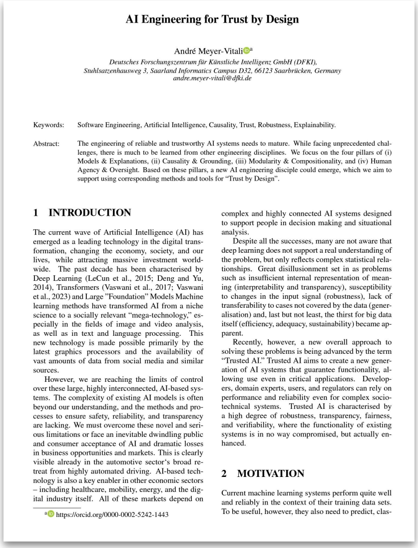
anr agence nationale de la recherche
ANR-24-IAS2-0002

NIST AI RMF (Risk Management Framework)

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[a]
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI),
Stuhlsatzenhausweg 3, Saarland Informatics Campus D32, 66123 Saarbrücken, Germany
andre.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) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, and (iv) Human Agency & Oversight. Based on these pillars, a new AI engineering disciple could emerge, which we aim to support using corresponding methods and tools for "Trust by Design".

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, while attracting massive investment worldwide. The past decade has been characterised by Deep Learning (LeCun et al., 2015; Deng and Yu, 2014), Transformers (Vaswani et al., 2017; Vaswani et al., 2023) and Large "Foundation" Models Machine learning methods have transformed AI from a niche science to a socially relevant "mega-technology," especially in the fields of image and video analysis, as well as in text and language processing. This new technology is made possible primarily by 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. The complexity of existing AI models is often beyond our understanding, and the methods and processes to ensure safety, reliability, and transparency are lacking. We must overcome these novel and serious limitations or face an inevitable dwindling public and consumer acceptance of AI and dramatic losses in business opportunities and markets. This is clearly visible already in the automotive sector's broad retreat from highly automated driving. AI-based technology is also a key enabler in other economic sectors – 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 problem, but only reflects complex statistical relationships. Great disillusionment set in as problems such as insufficient internal representation of meaning (interpretability and transparency), susceptibility to changes in the input signal (robustness), lack of transferability to cases not covered by the data (generalisation) and, last but not least, the thirst for big data itself (efficiency, adequacy, sustainability) 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 AI systems that guarantee functionality, allowing use even in critical applications. Developers, domain experts, users, and regulators can rely on performance and reliability even for complex socio-technical systems. Trusted AI is characterised by a high degree of robustness, transparency, fairness, and verifiability, where the functionality of existing systems is in no way compromised, but actually enhanced.

2 MOTIVATION

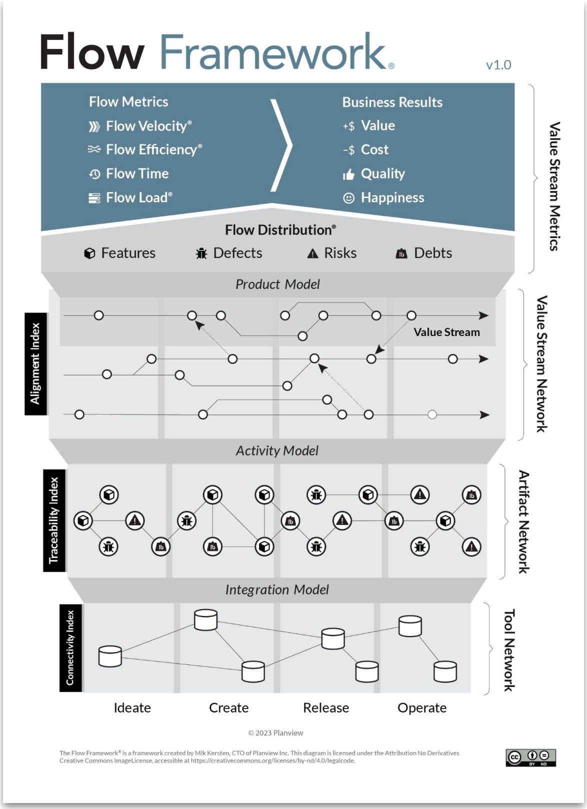
Current machine learning systems perform quite well and reliably in the context of their training data sets. To be useful, however, they also need to predict, clas-

[a] https://orcid.org/0000-0002-5242-1443

https://www.linkedin.com/comm/pulse/pour-une-gestion-efficace-des-risques-de-lia-du-nist-ai-chmielewski-mndyf (in French, sorry)

Don’t forget your **value**

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

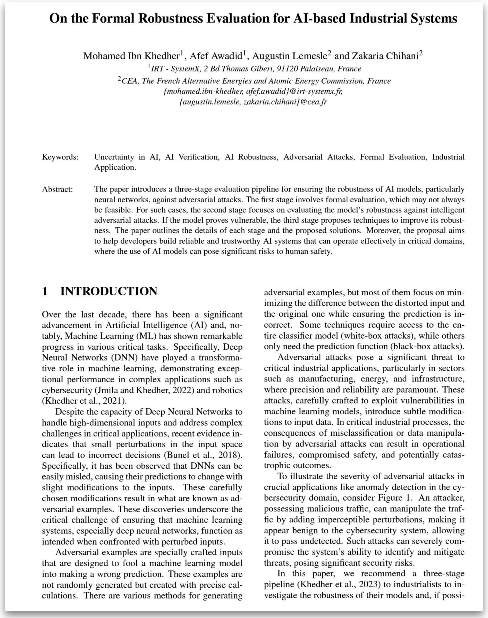


https://flowframework.org/

agence nationale de la recherche
ANR-24-IAS2-0002

“Meta” capabilities

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



On the Formal Robustness Evaluation for AI-based Industrial Systems

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

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 (Jmila 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 small perturbations in the input space can lead to incorrect decisions (Bunel 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.

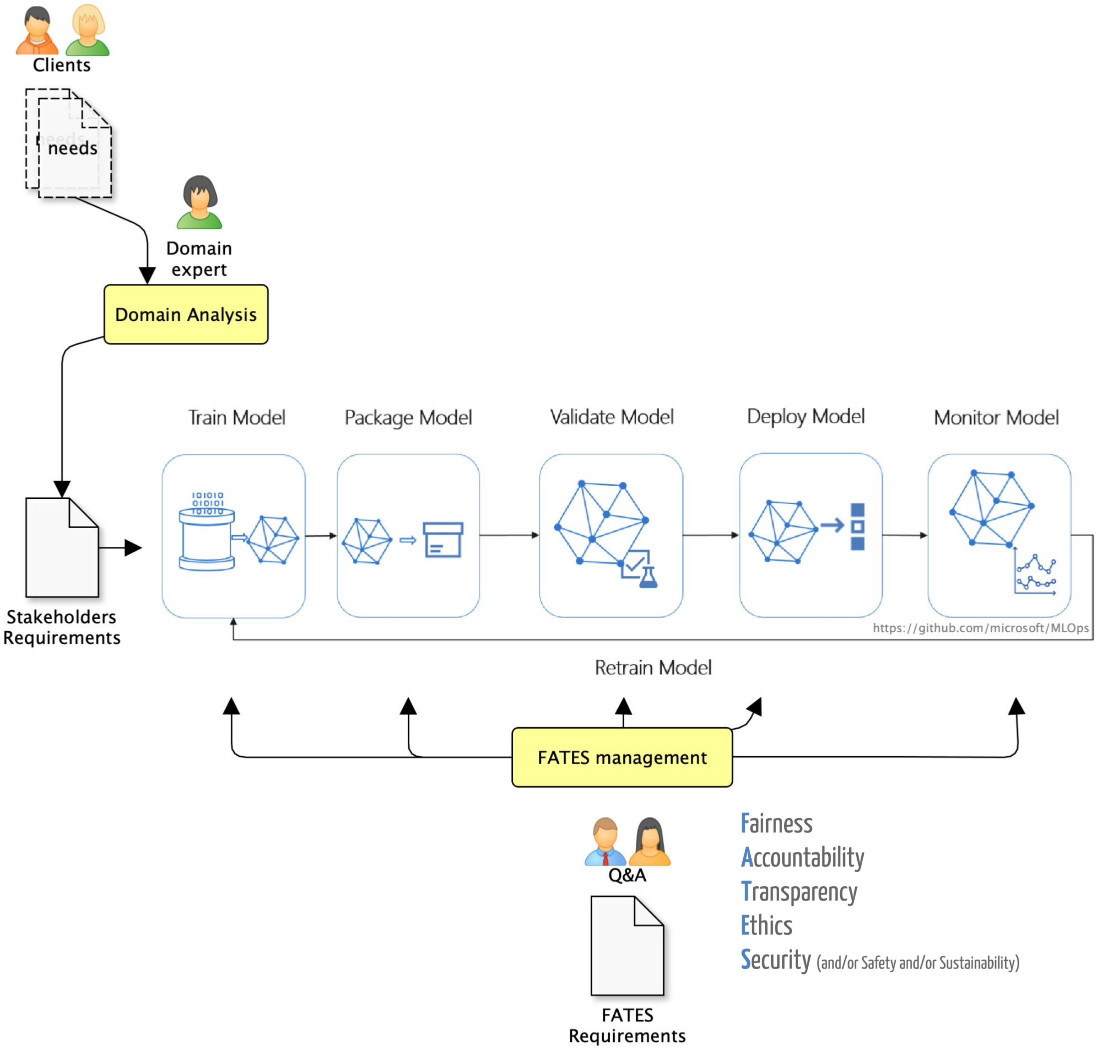
In this paper, we recommend a three-stage pipeline (Khedher et al., 2023) to industrialists to investigate the robustness of their models and, if possi-

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

Project organization

Big picture

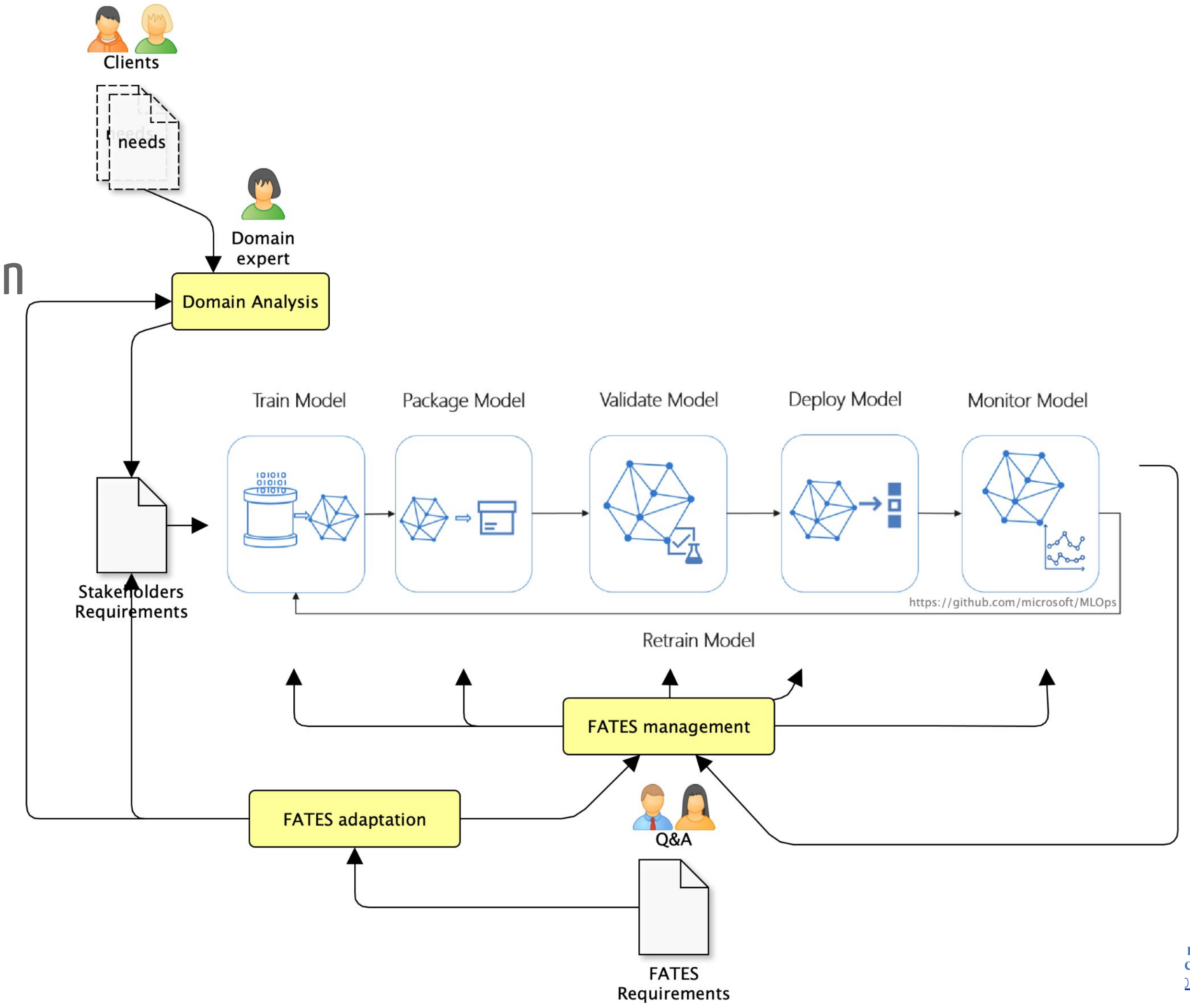
FATES precise definitions



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

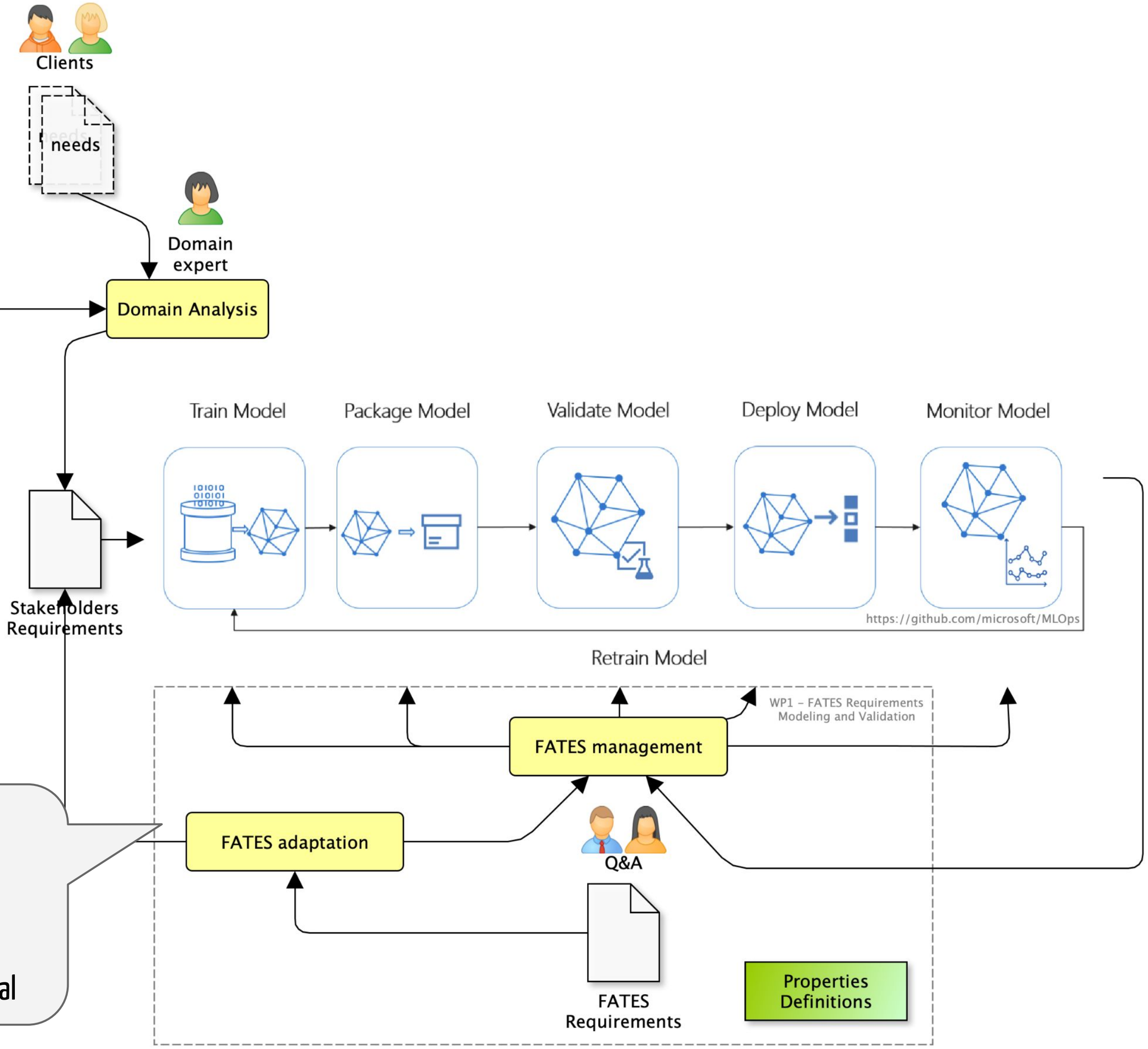
Big picture

FATES contextualisation



Work packages

WP1 - FATES models

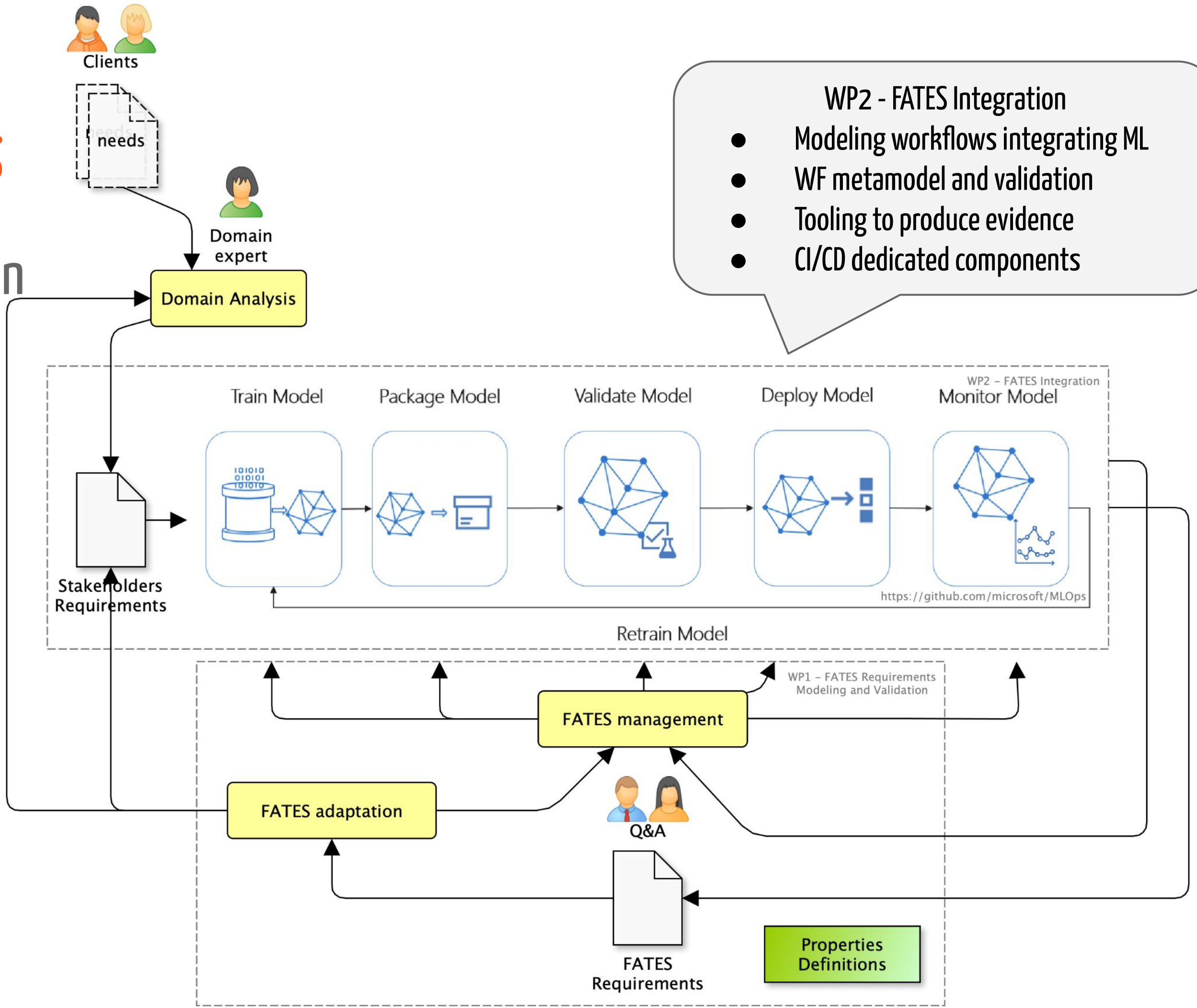


WP1 - FATES Models

- Described as requirements
- Validation criterias
- Measures & KPIs
- Distance function from optimal

Work packages

WP2 - FATES integration

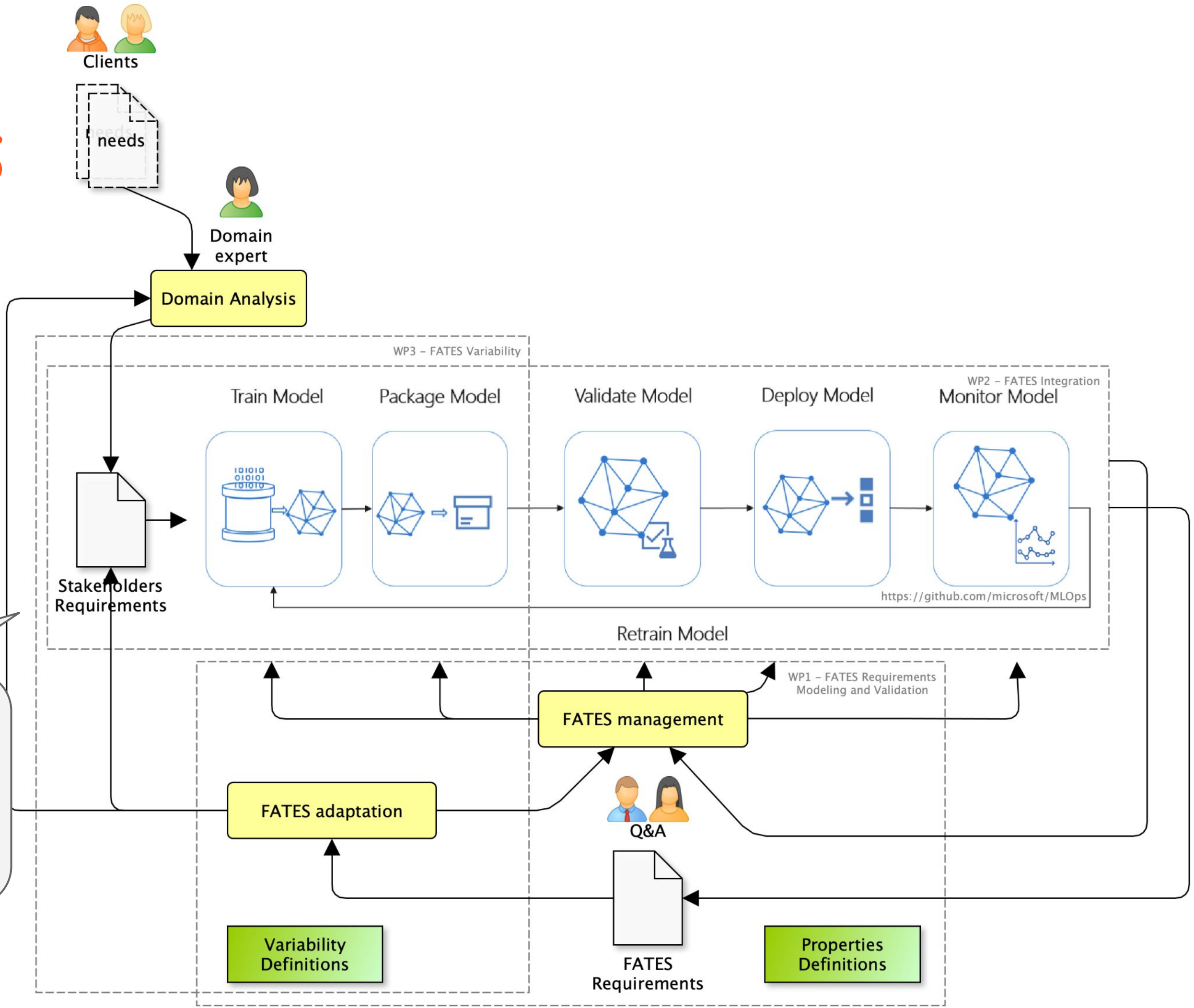


WP2 - FATES Integration

- Modeling workflows integrating ML
- WF metamodel and validation
- Tooling to produce evidence
- CI/CD dedicated components

Work packages

WP3 - FATES variability



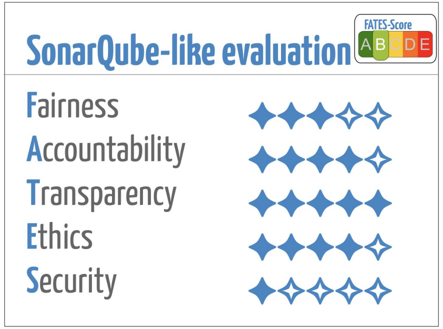
WP3 - FATES Variability

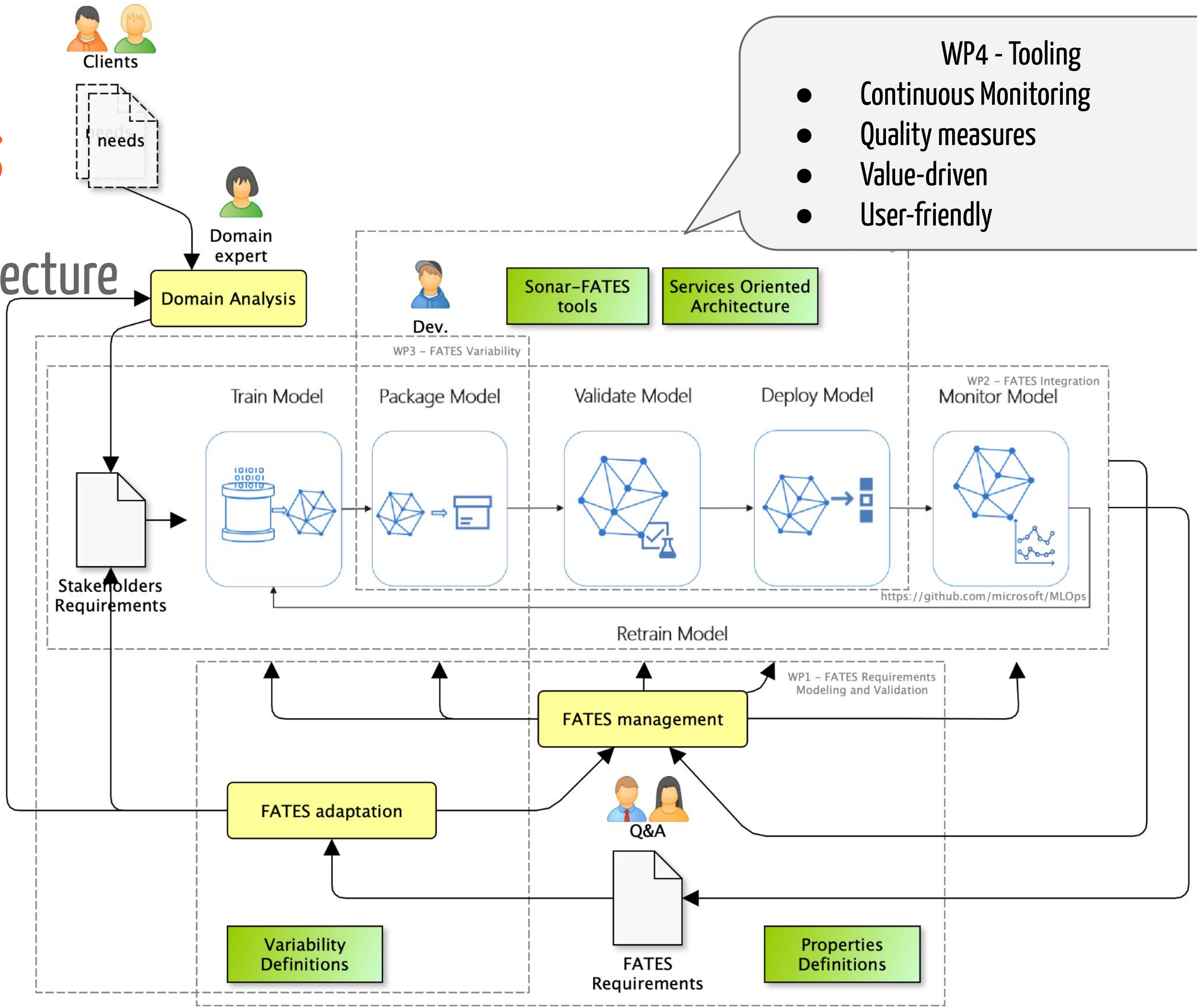
- Variability models
- FATES properties analysis
- Justifications
- Operationalization validation

Work packages

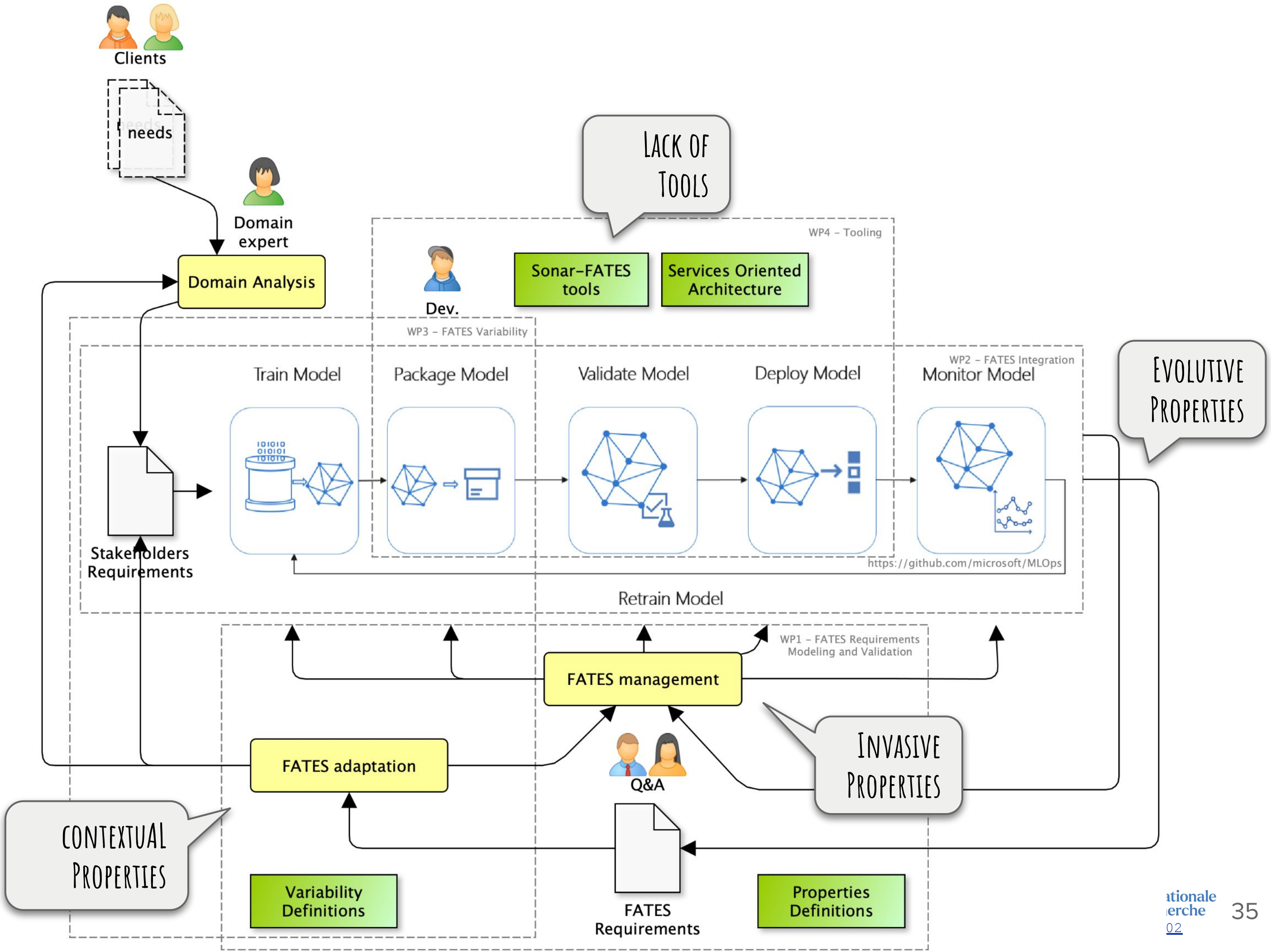
WP4 - Tooling & Architecture

WP4 - Tooling
- Continuous Monitoring
- Quality measures
- Value-driven
- User-friendly



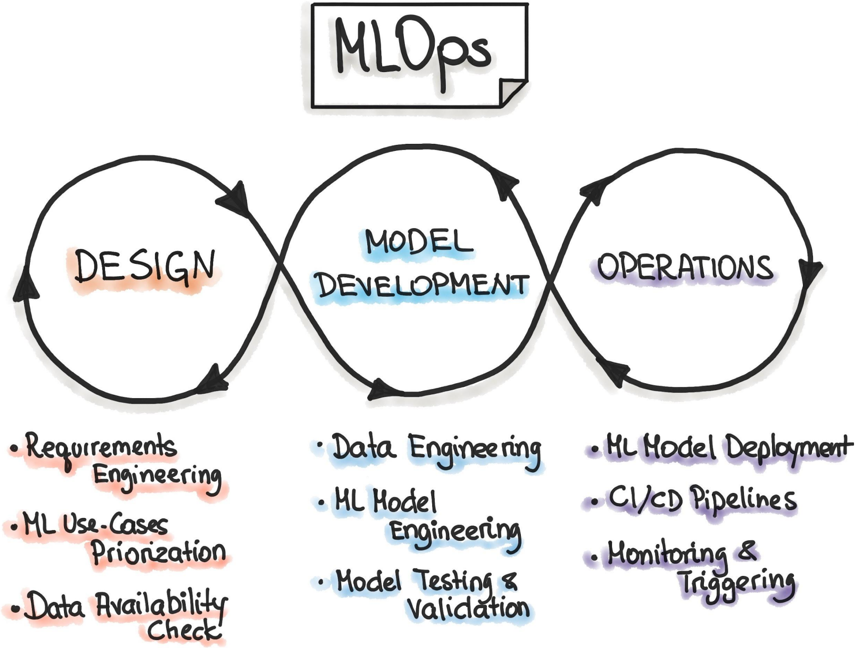


Key concerns



Support ML in Operations

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



https://ml-ops.org/

Kubeflow

mlflow

Current members (permanent only)

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


O. Teste


M. Pantel


J.-M. Bruel

UNIVERSITÉ CÔTE D'AZUR


M. Blay-Fornarino


P. Collet

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
Inria


F. Precioso


M. Riveill

McMaster University


S. Mosser

Context

2024 – 2028
600K€

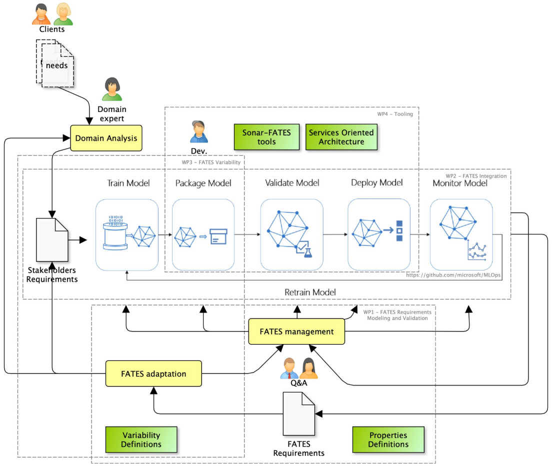
2 PhDs & **2** Postdocs

anr agence nationale de la recherche
ANR-24-IAS2-0002

We need you!

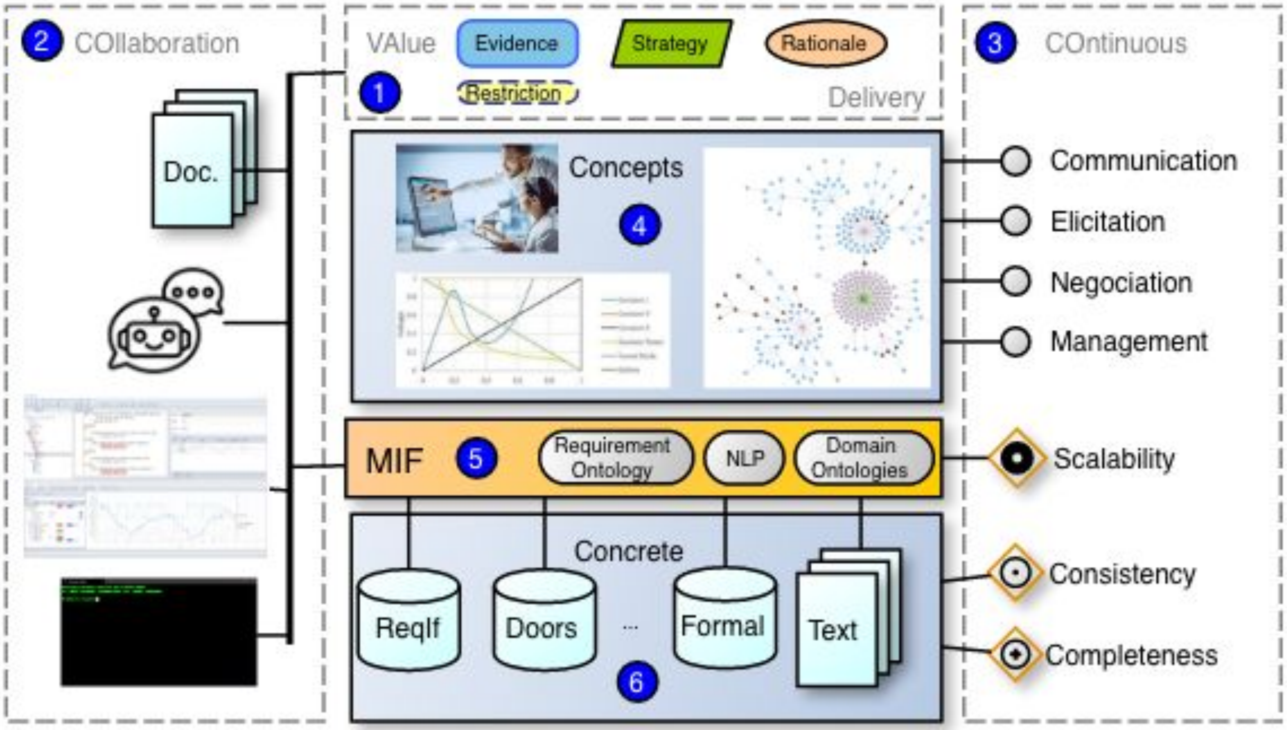
Collaboration opportunity

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

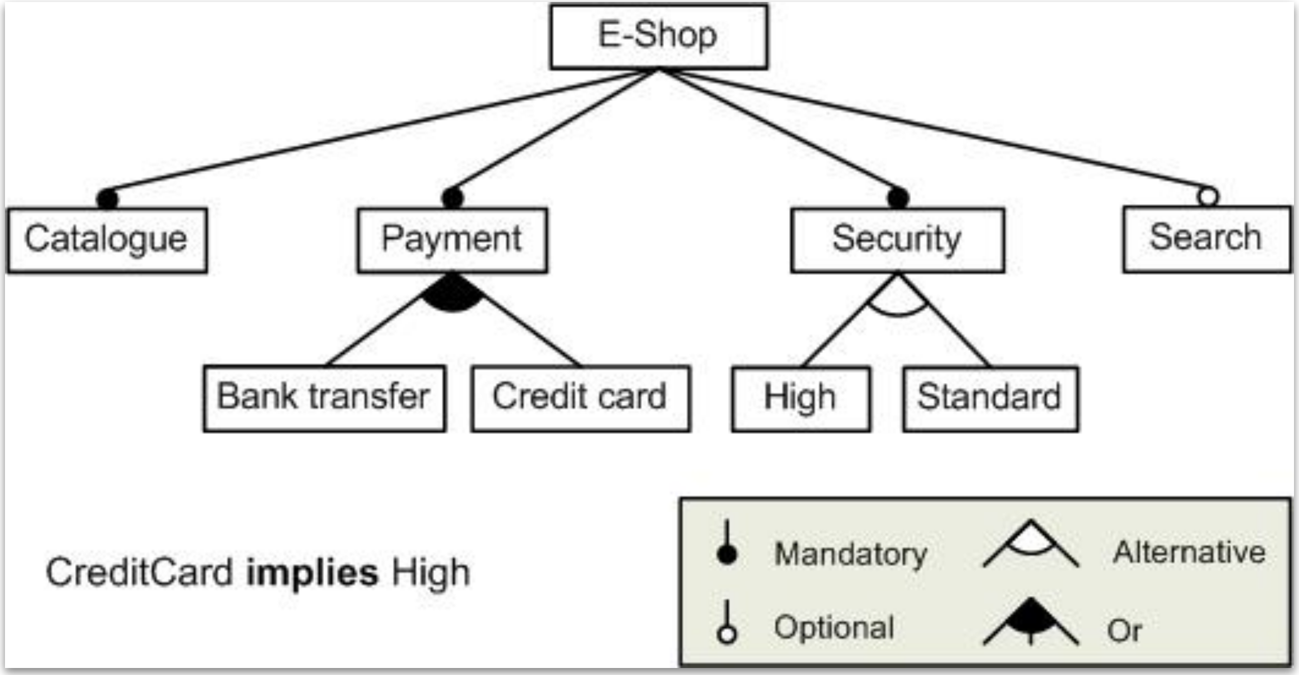


ANR-24-IAS2-0002

Properties formalisation

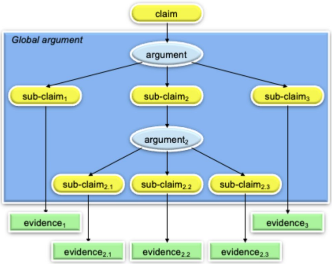


Feature model

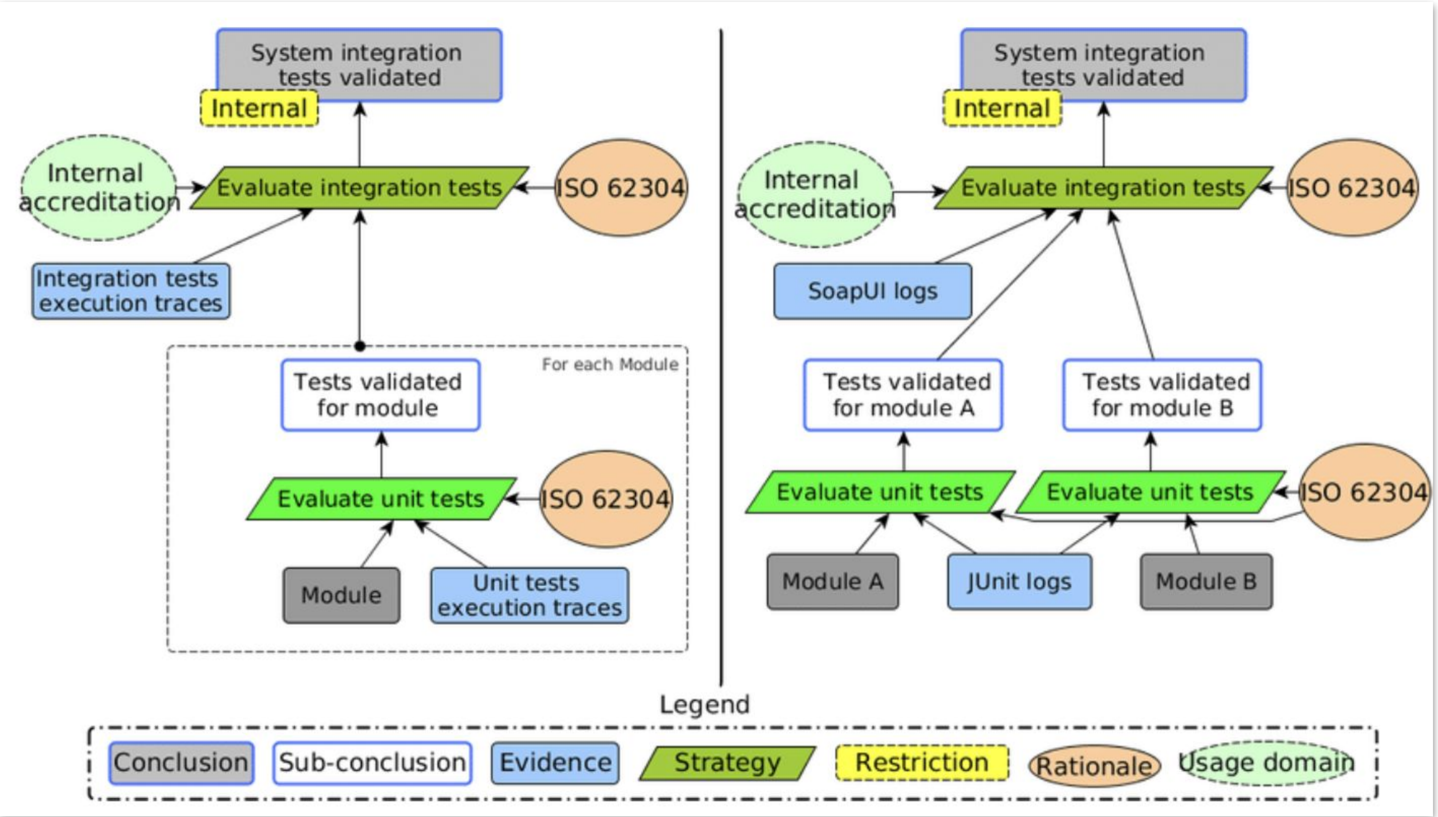


https://en.wikipedia.org/wiki/Feature_model

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



ISO 15026-2



Sorry for the French

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, ...).

anr agence nationale de la recherche
ANR-24-IAS2-0002

Discussions time!

https://bit.ly/jmb-explainai25

Get the slides

https://bit.ly/jmbruel

@jmbruel