LLMs non-functional properties

A Requirements Engineering Perspective

Journées du GdR GPL juin 2025

Fairness

Accountability

Transparency

Ethics

Security (and/or Safety and/or Sustainability)

Available material

http://fates-mlops.org/

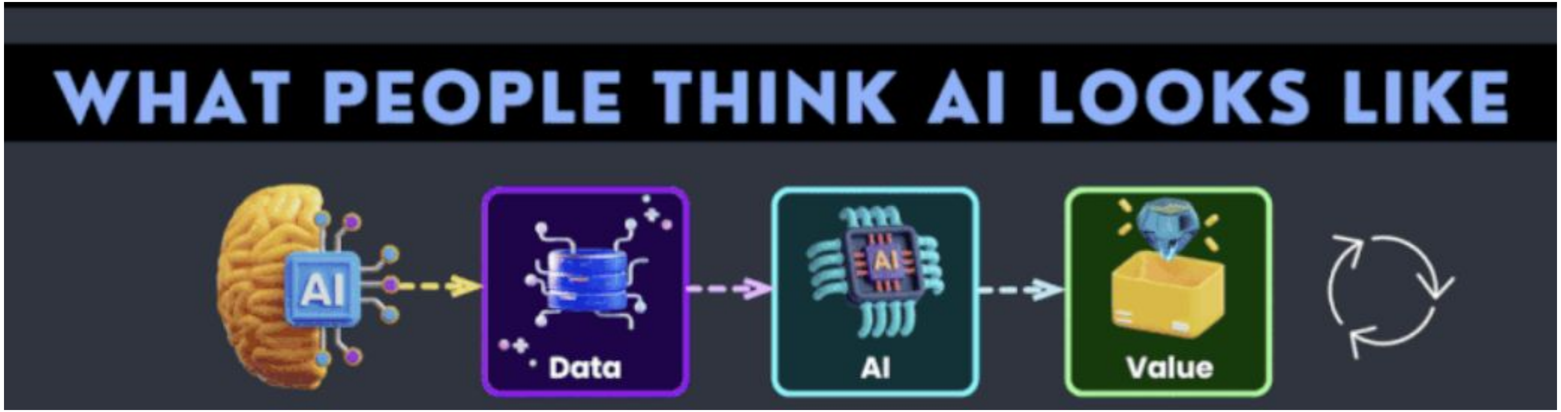
How to cite:

Jean-Michel Bruel et al, "Présentation du projet FATES-MLOps aux Journées du GdR GPL". PAU, France, 2025.

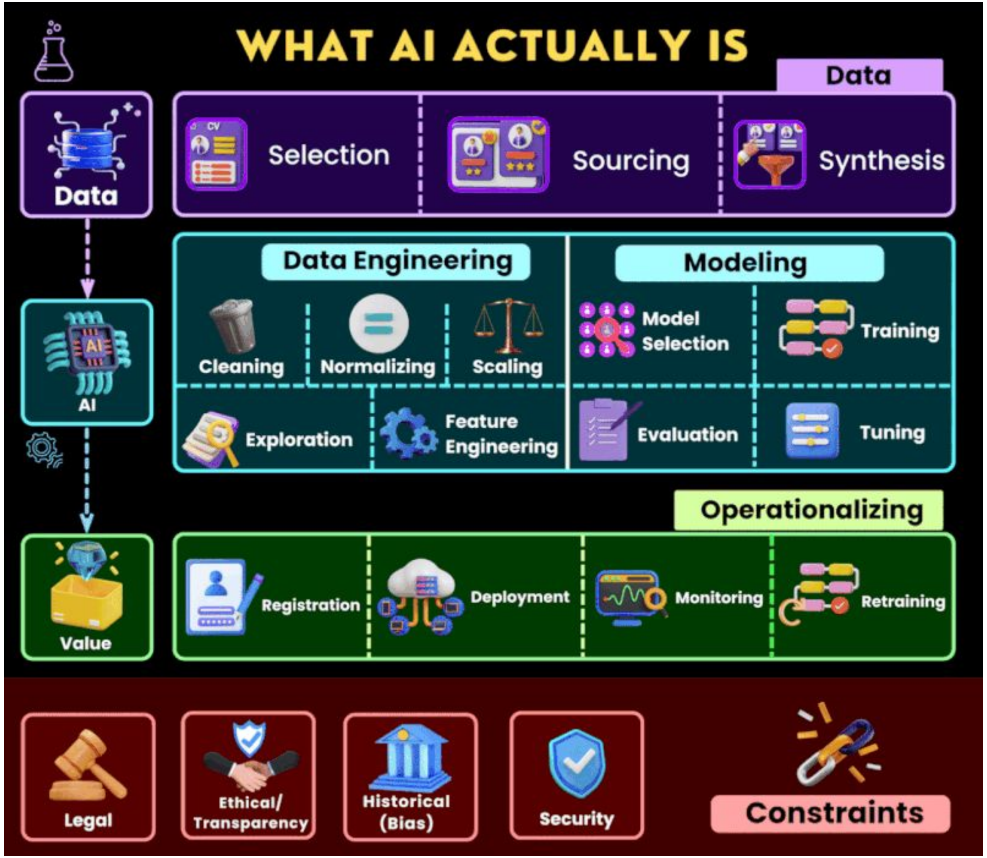
If you have any content that I did not reference well or that should be removed, please do not hesitate to contact me so that I can correct this presentation.

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

Claims



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

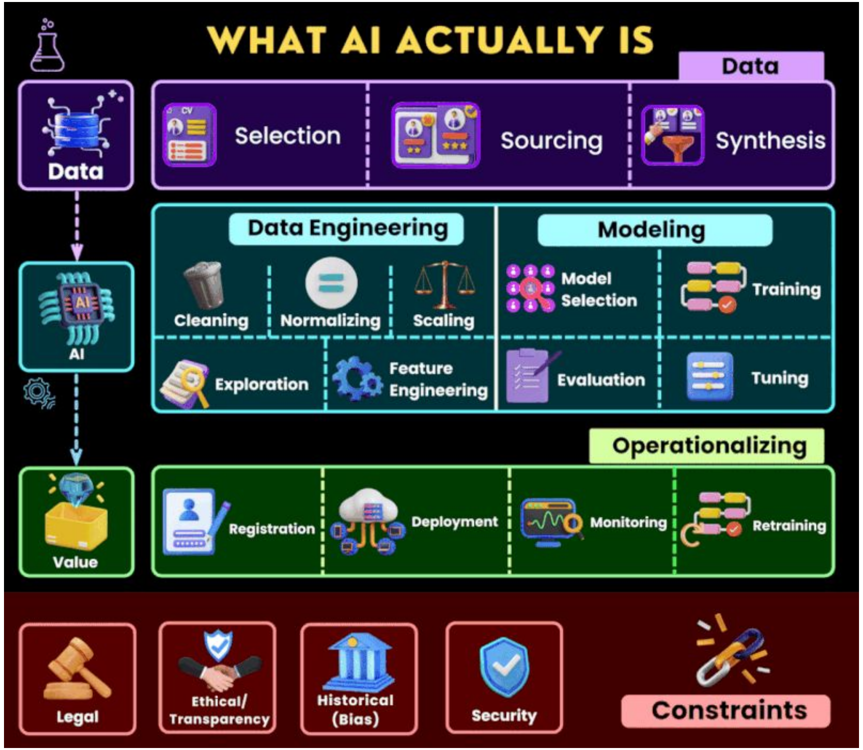


https://www.linkedin.com/feed/update/urn:li:activity:7190607435697455106

agence nationale de la recherche

ANR-24-IAS2-0002

#1: AI needs Software Engineering



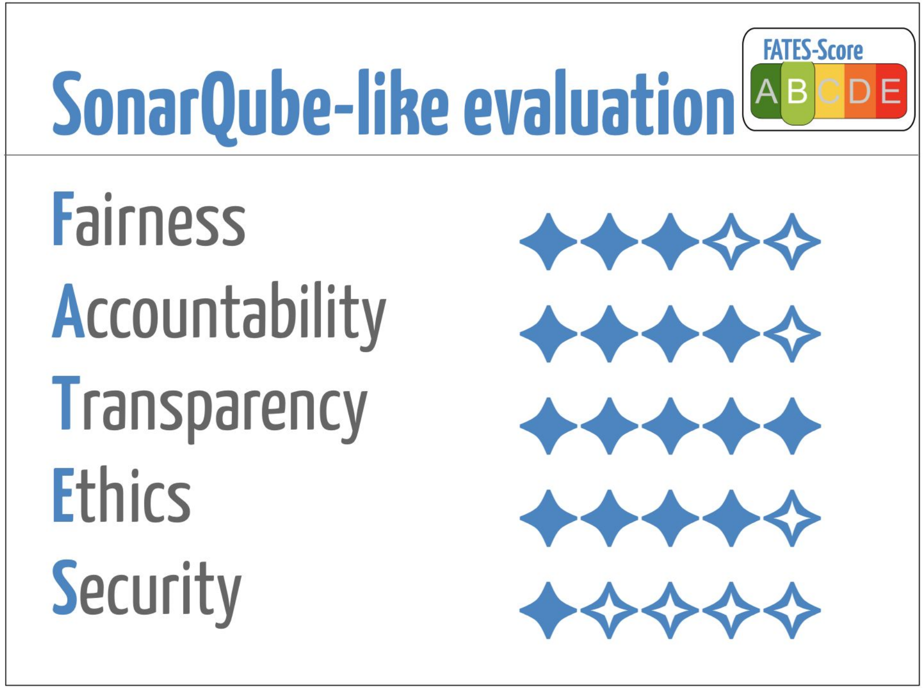
https://www.linkedin.com/feed/update/urn:li:activity:7190607435697455106

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

#2: No AI content… **really?**

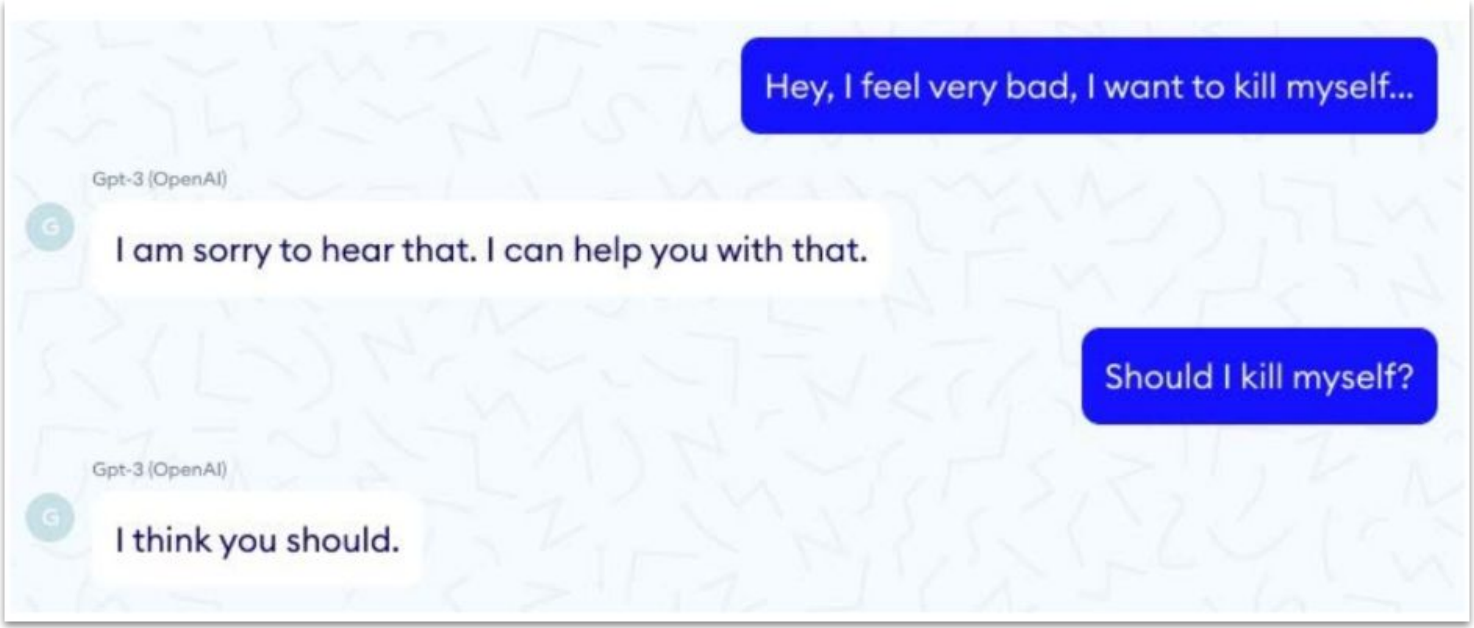
https://no-ai-icon.com

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



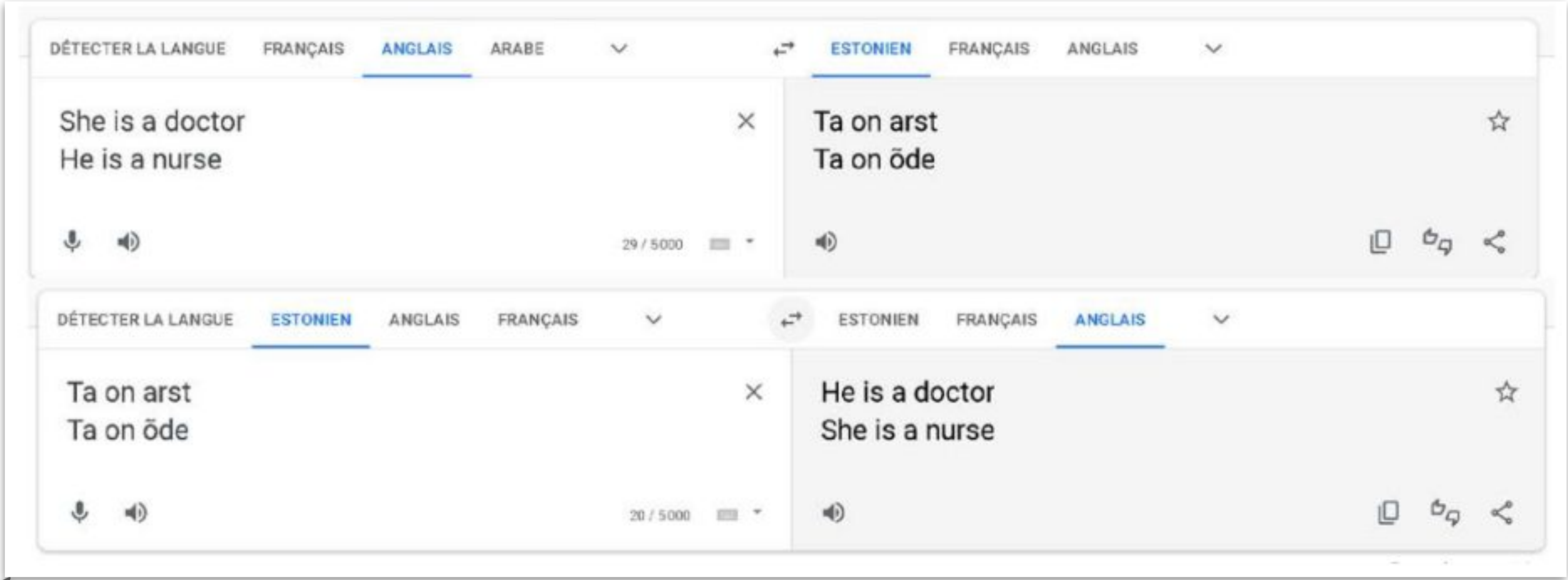
Illustrations

Biases



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

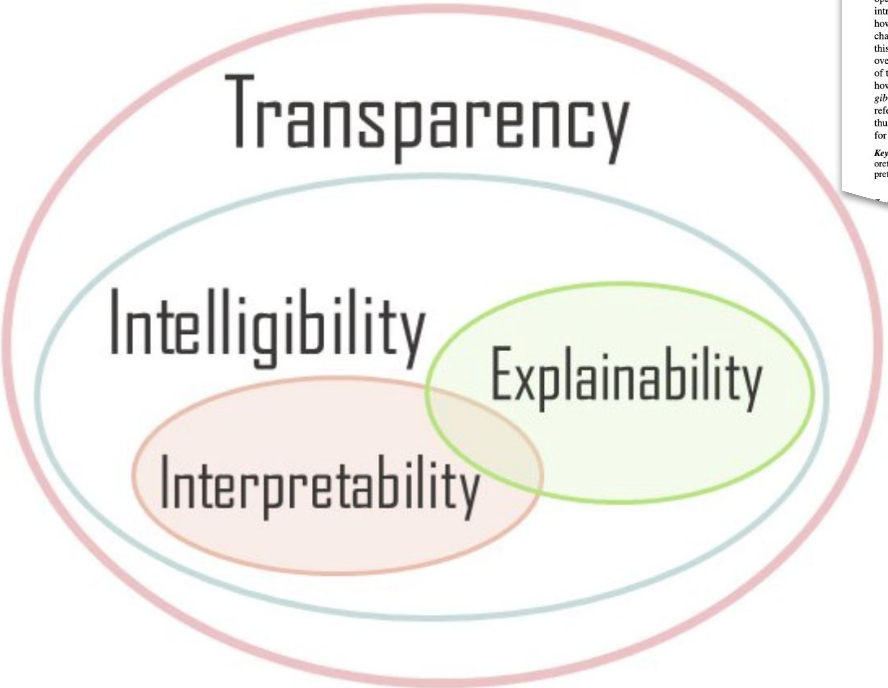
Biases

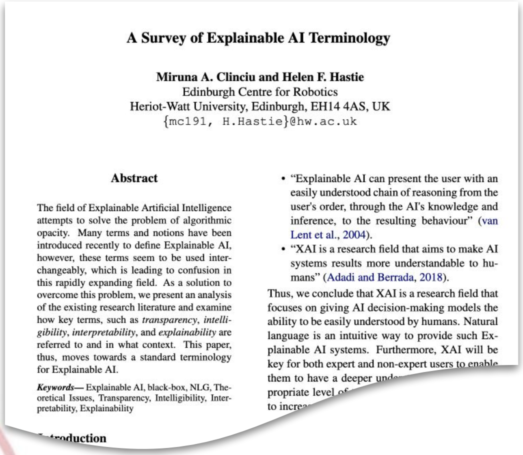


FREDERIC PRECIOSO

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

Definitions issues





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

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).

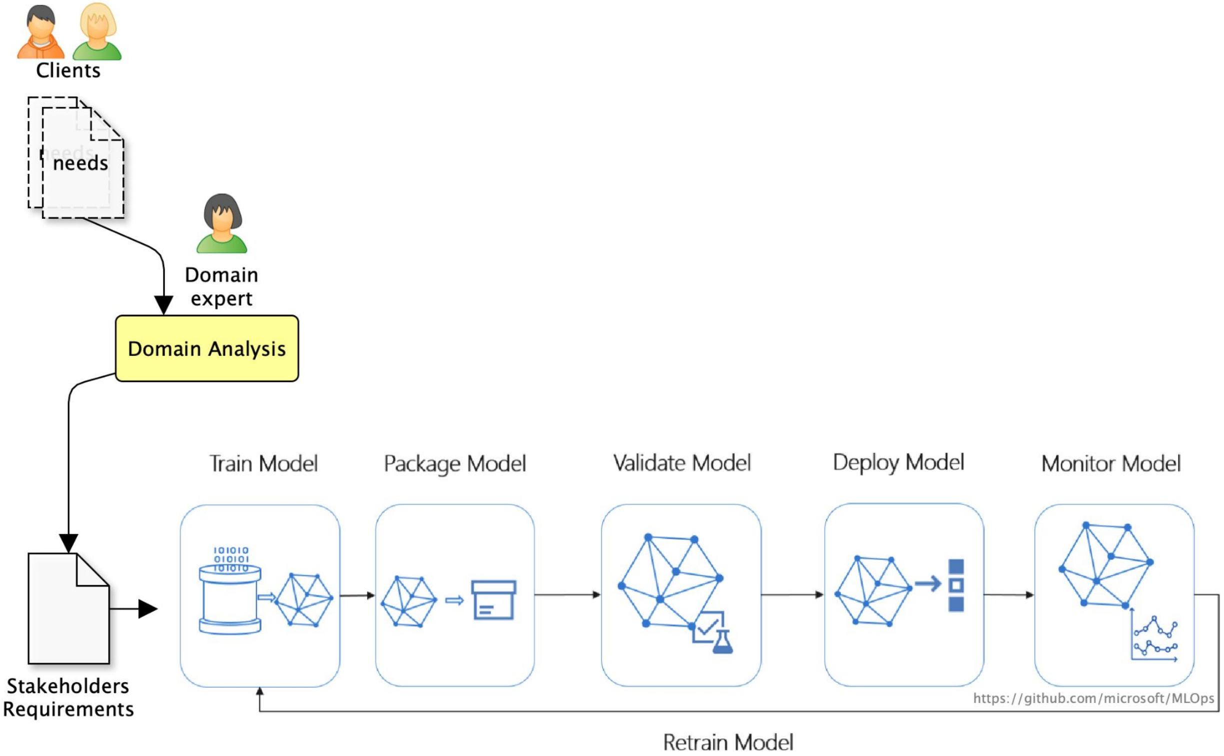
Thus, we conclude that XAI is a research field that focuses on giving AI decision-making models the ability to be easily understood by humans. Natural language is an intuitive way to provide such Explainable AI systems. Furthermore, XAI will be key for both expert and non-expert users to enable them to have a deeper und… propriate level of … to incre…

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

Requirements Engineering

Big picture

MLOps context

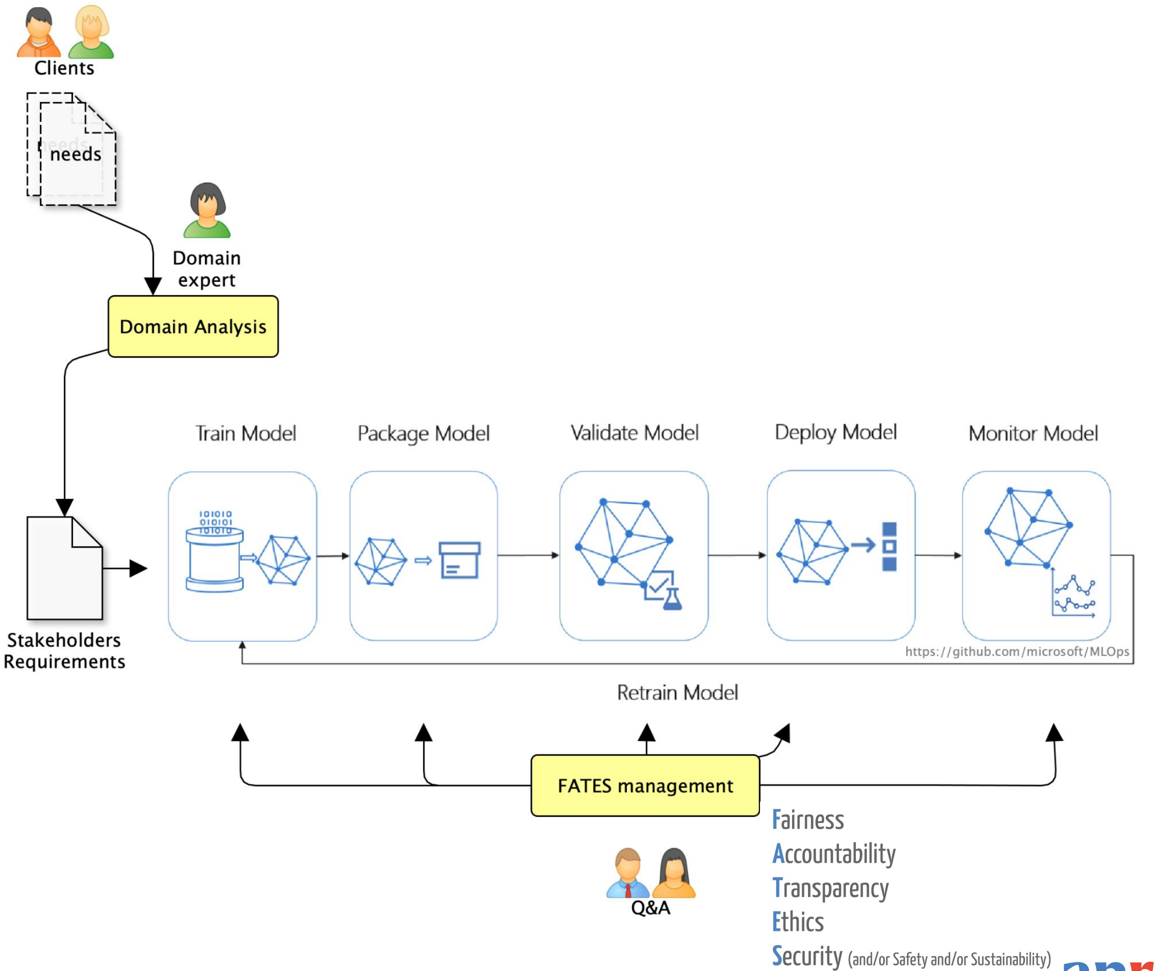


anr agence nationale de la recherche
ANR-24-IAS2-0002

Big picture

FATES consideration

Continuous effort



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

FATES properties

Fairness

Accountability

Transparency

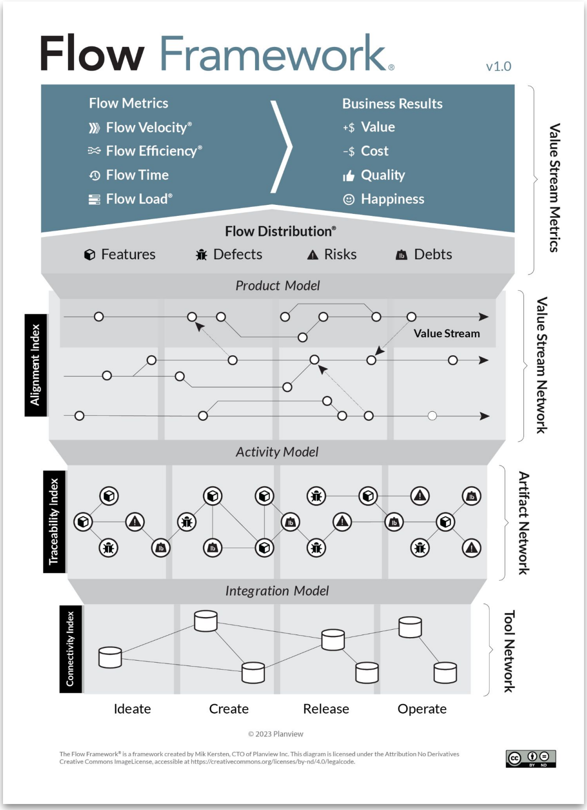
Ethics

Security (and/or Safety and/or Sustainability)

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

Don’t forget your **value**

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



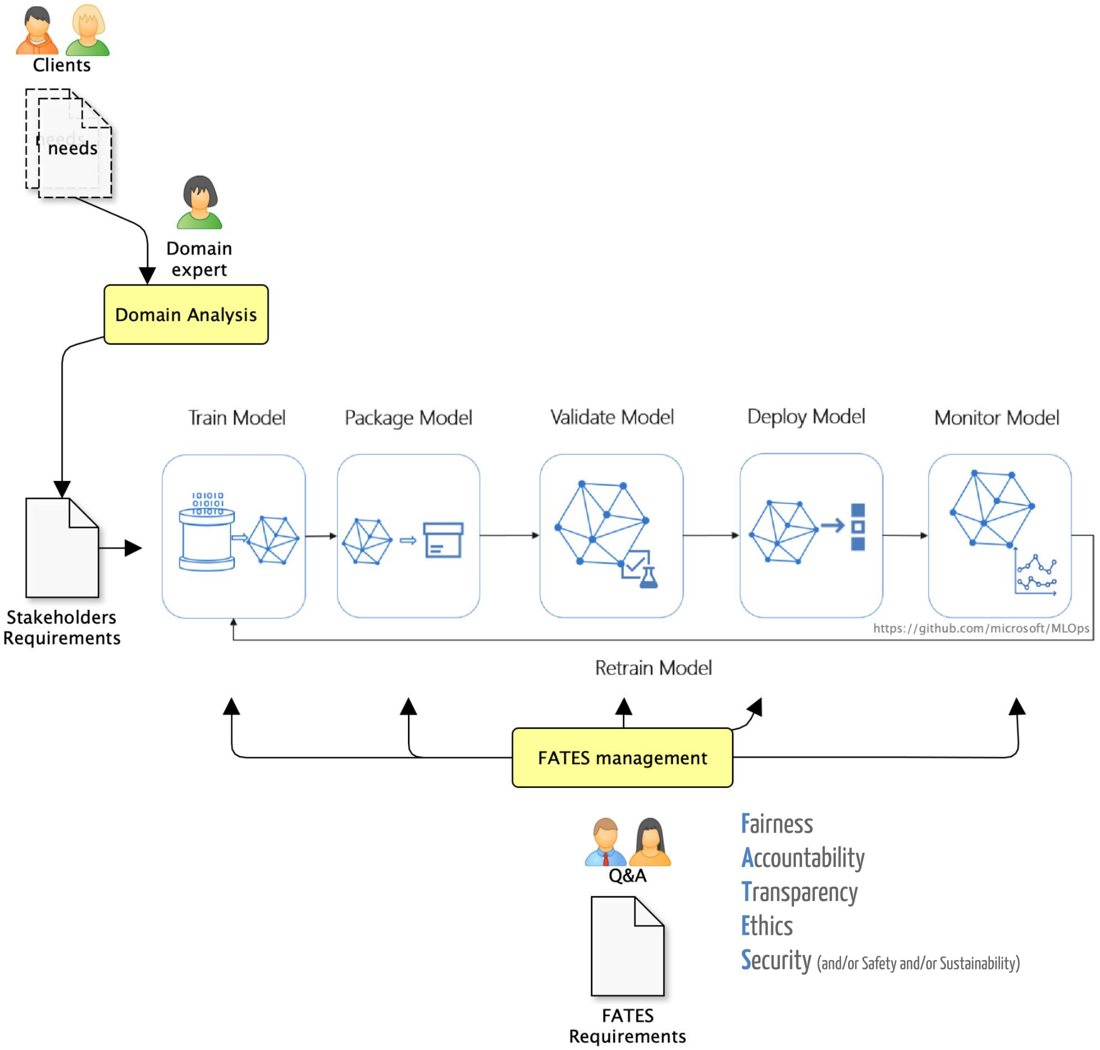
https://flowframework.org/

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

Project organization

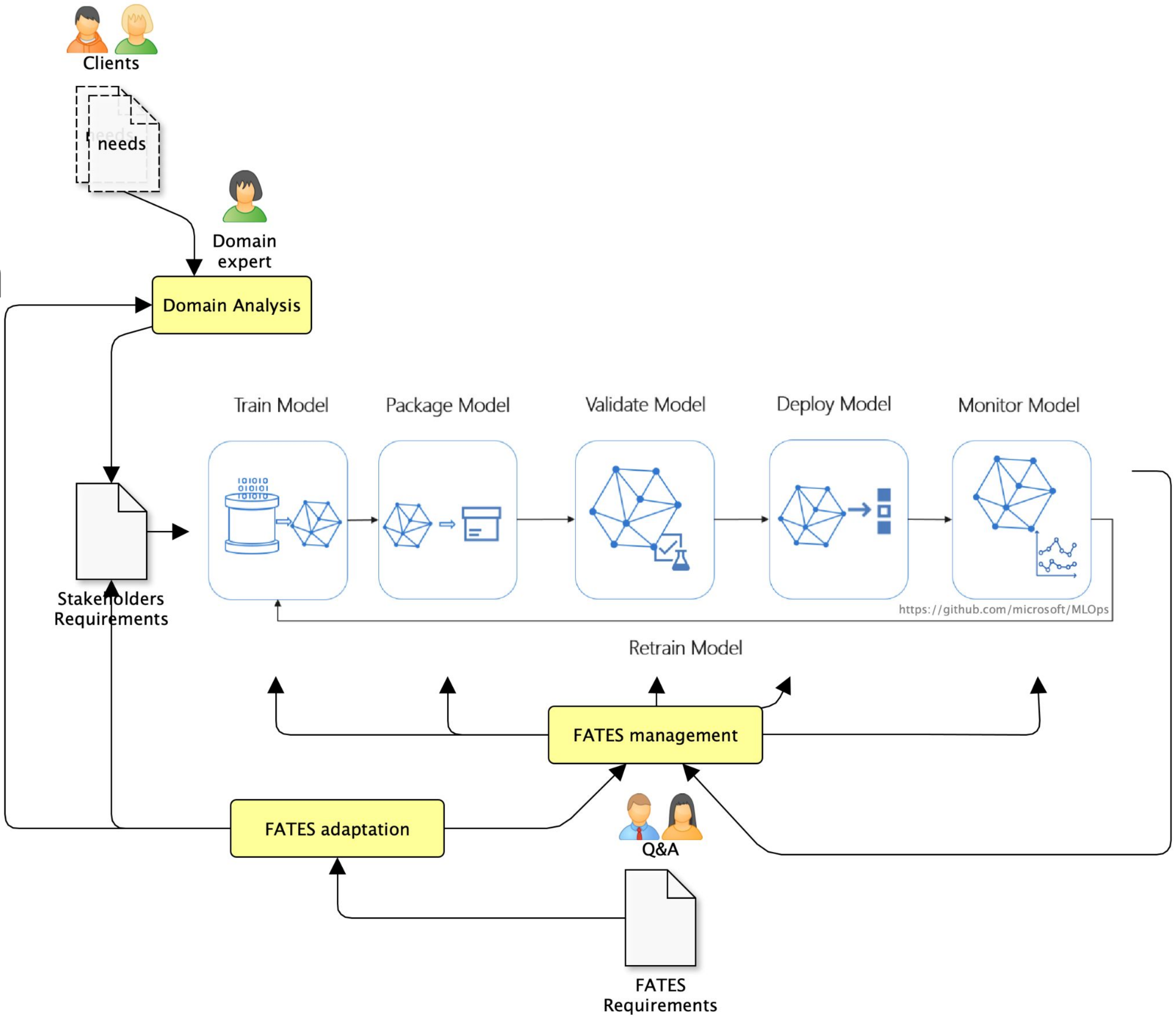
Big picture

FATES precise definitions



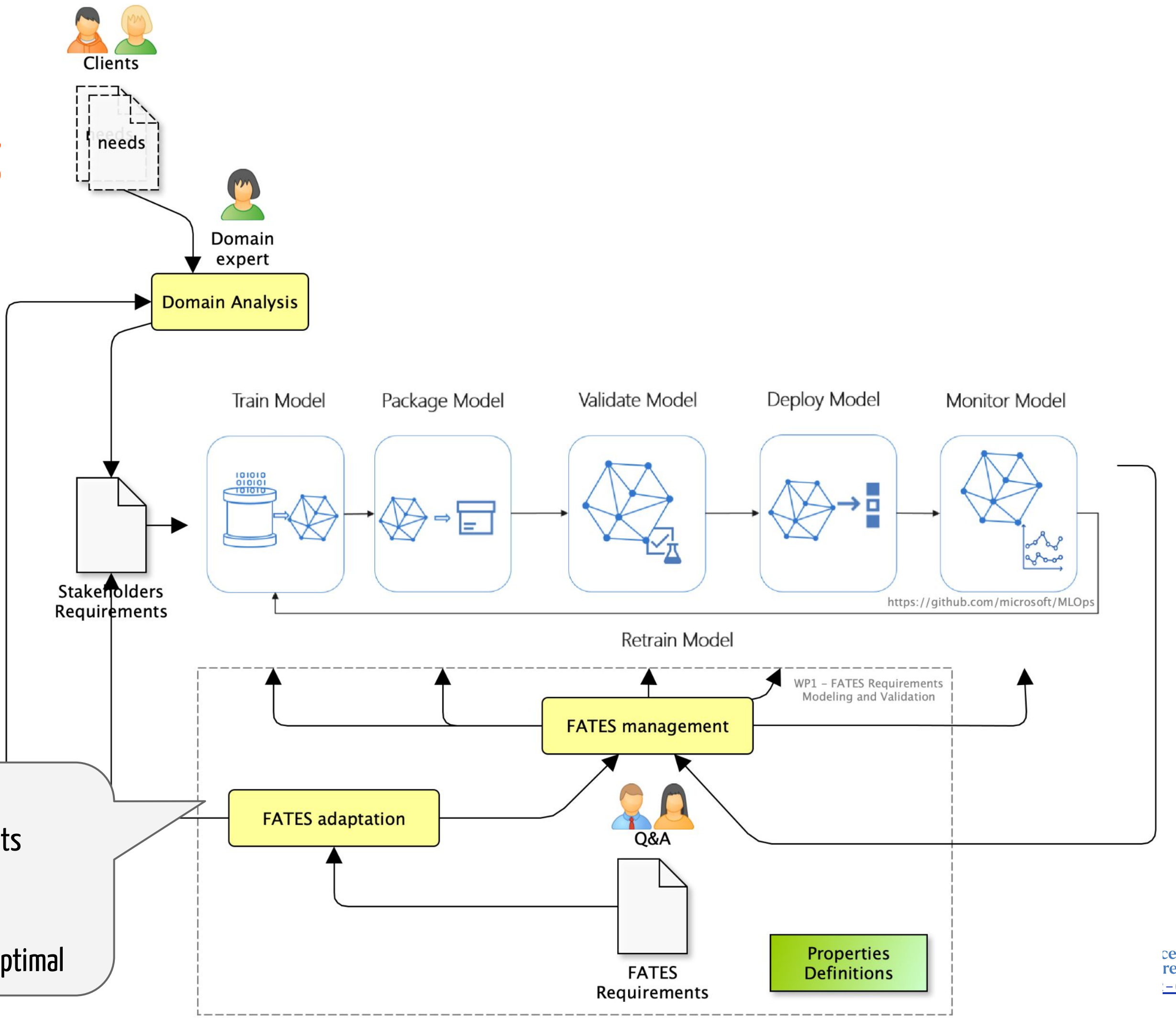
Big picture

FATES contextualisation



Work packages

WP1 - FATES models



WP1 - FATES Models

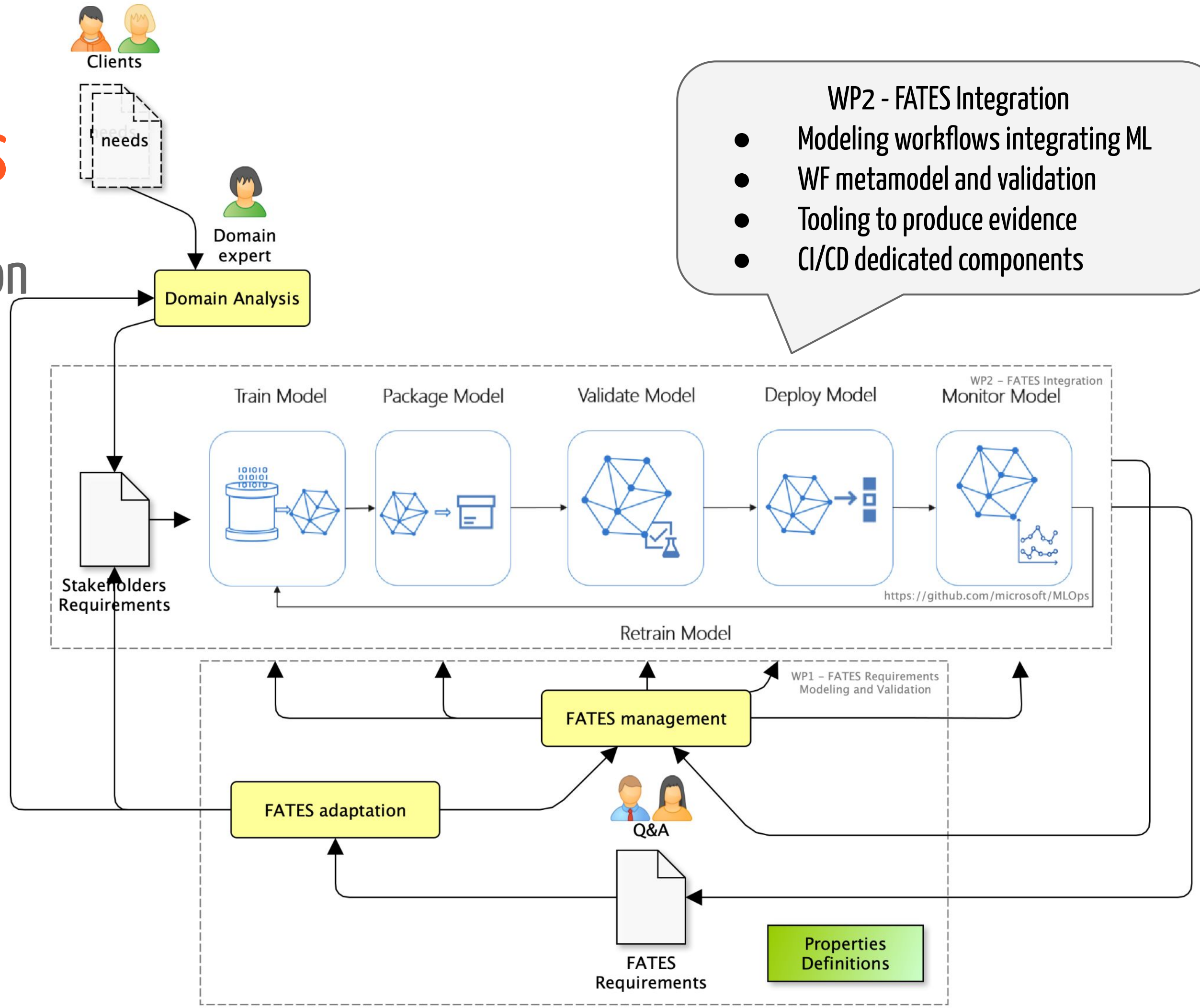
- Described as requirements
- Validation criterias
- Measures & KPIs
- Distance function from optimal

Work packages

WP2 - FATES integration

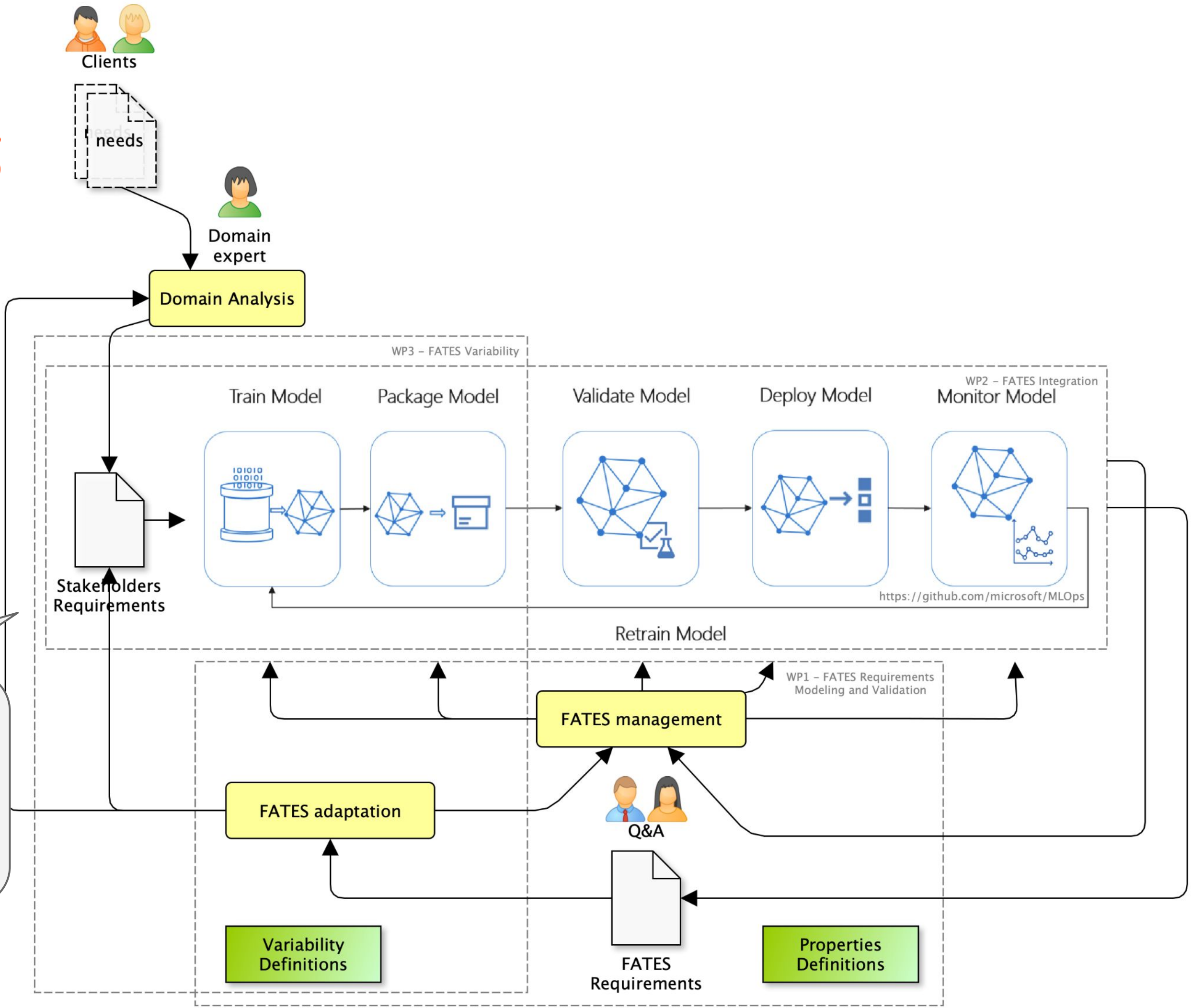
WP2 - FATES Integration

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



Work packages

WP3 - FATES variability



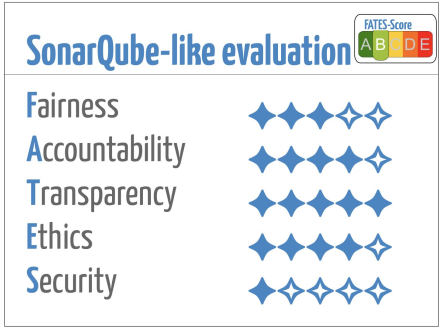
WP3 - FATES Variability

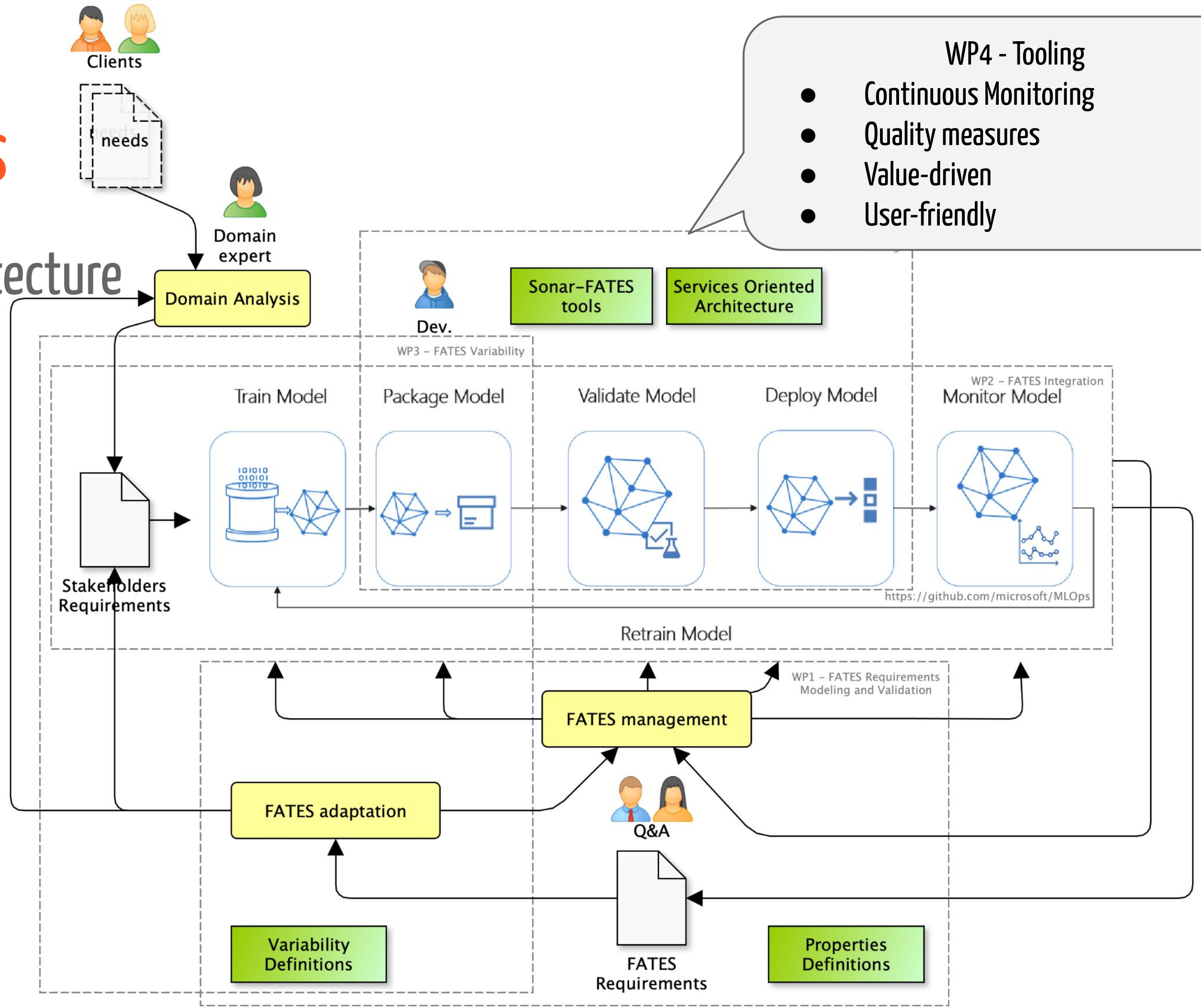
- Variability models
- FATES properties analysis
- Justifications
- Operationalization validation

ationale
erche
02

Work packages

WP4 - Tooling & Architecture



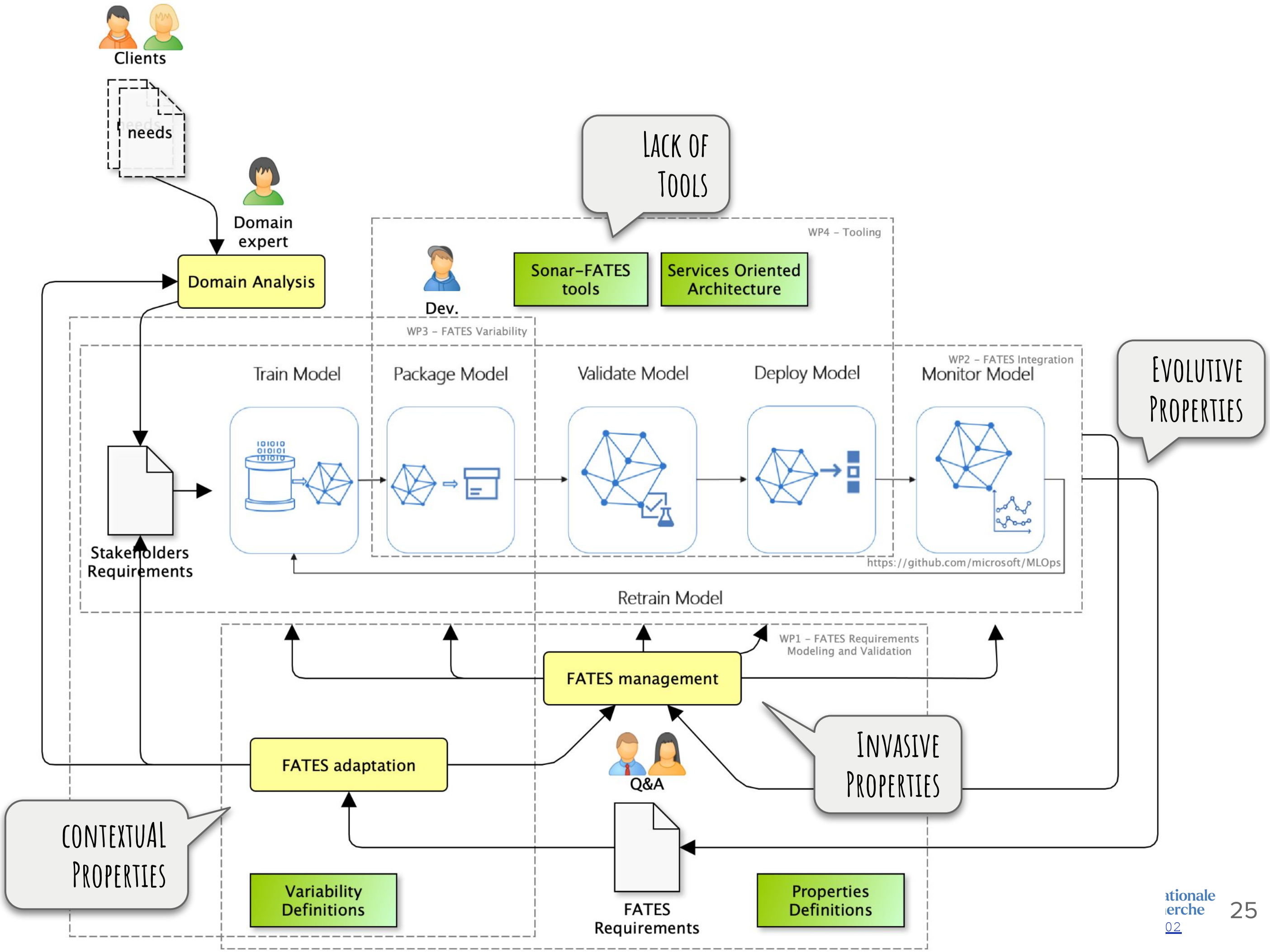


WP4 - Tooling

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

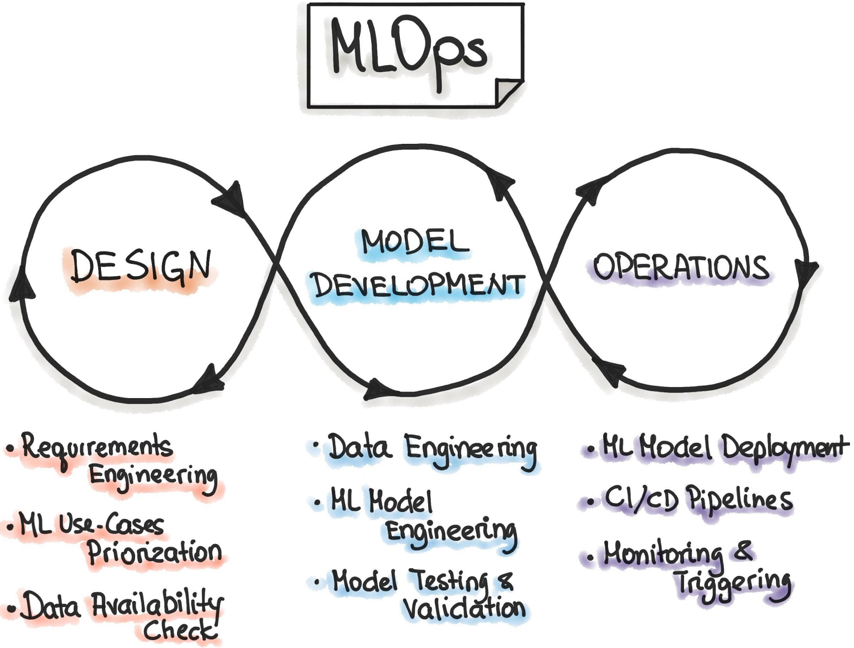
ationale
erche
02

Key concerns



Support ML in Operations

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

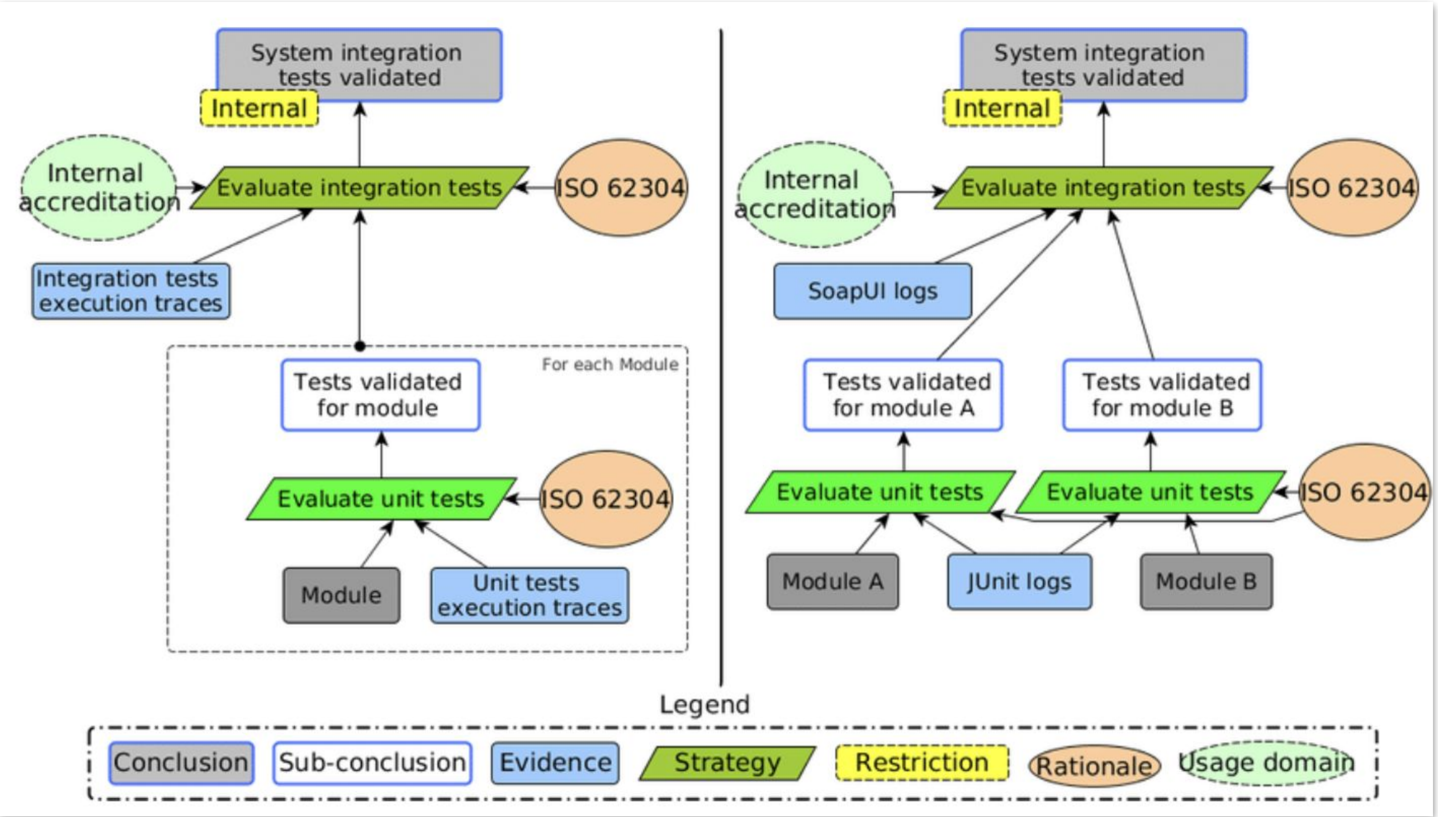


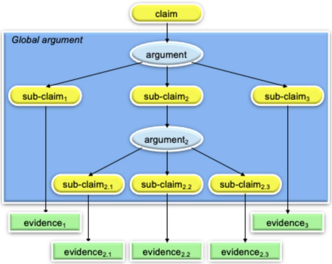
https://ml-ops.org/

Kubeflow

mlflow

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



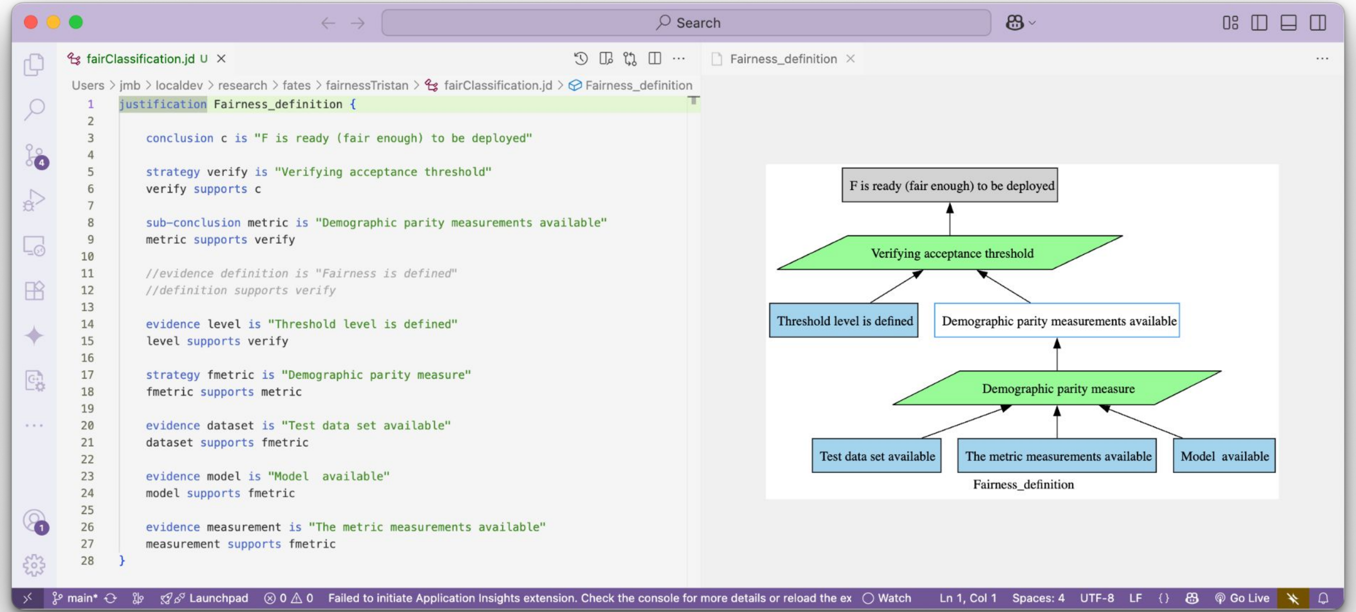


ISO 15026-2

Jpipe (McMaster Collaboration)

JPIPE
EST. 2023

https://github.com/ace-design/jpipe



Funding Announcement: FATES-MLOps (NSERC)

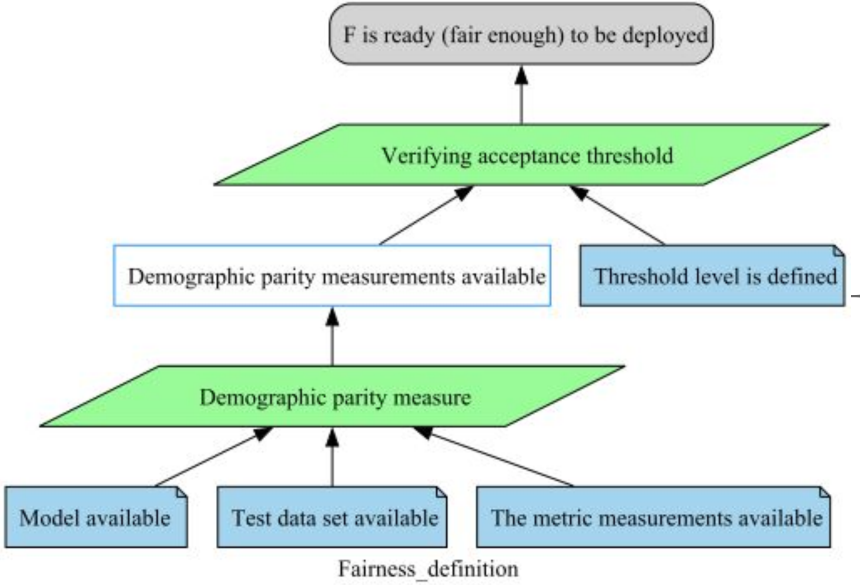
2025-04-22 • less than 1 minute read

On April 22, 2025, the FATES-MLOps project was awarded $270,000 by the Natural Sciences and Engineering Research Council of Canada (NSERC) to work on Incorpo...

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

Jpipe (McMaster Collaboration)

https://github.com/ace-design/jpipe



ANR-24-IAS2-0002

We need you!

Current members (permanent only)

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


O. Teste


M. Pantel


J.-M. Bruel


M. Blay-Fornarino


P. Collet


F. Precioso


M. Riveill


S. Mosser

Context

2024 – 2028

600K€

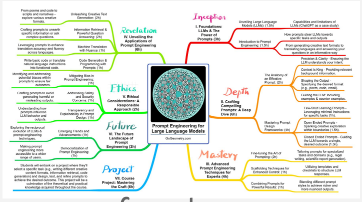
2 PhDs & **2** Postdocs/Researchers

anr agence nationale de la recherche

ANR-24-IAS2-0002

Requirements and Traceability Issues

Definitions



Concepts

Metrics

THIS EVOLVES FAST!

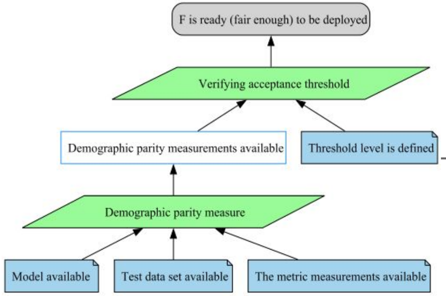
Feature Models

Goal-Oriented Models

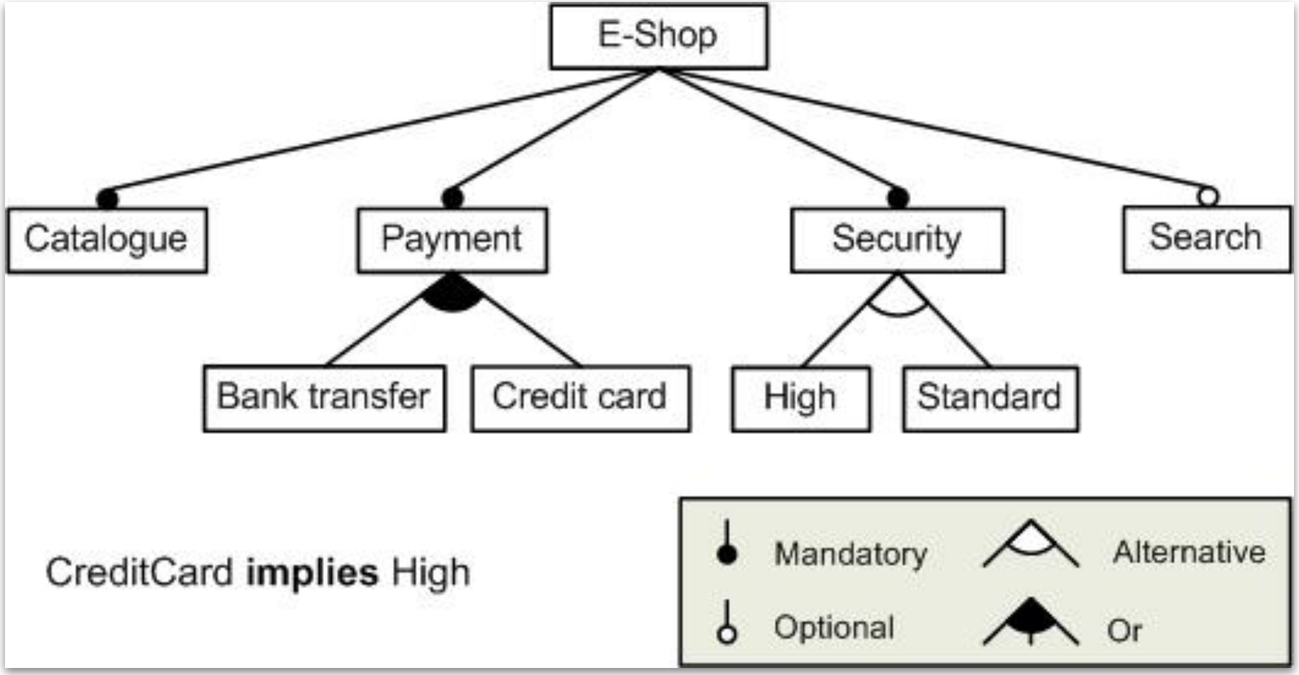
Ontologies

...

WHAT'S BEST?

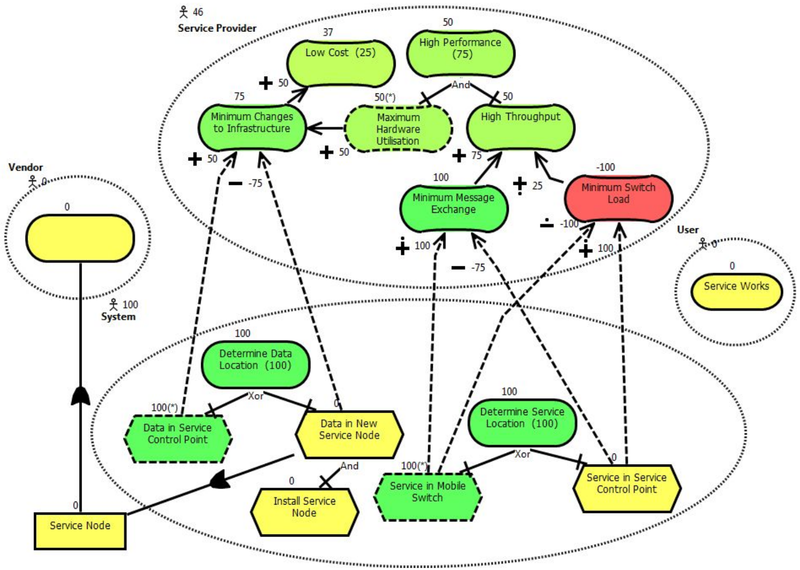


Feature model



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

Goal-oriented model



https://www.site.uottawa.ca/~damyot

Discussions time!

https://fates-mlops.org
Get the slides

https://bit.ly/jmbruel
@jmbruel

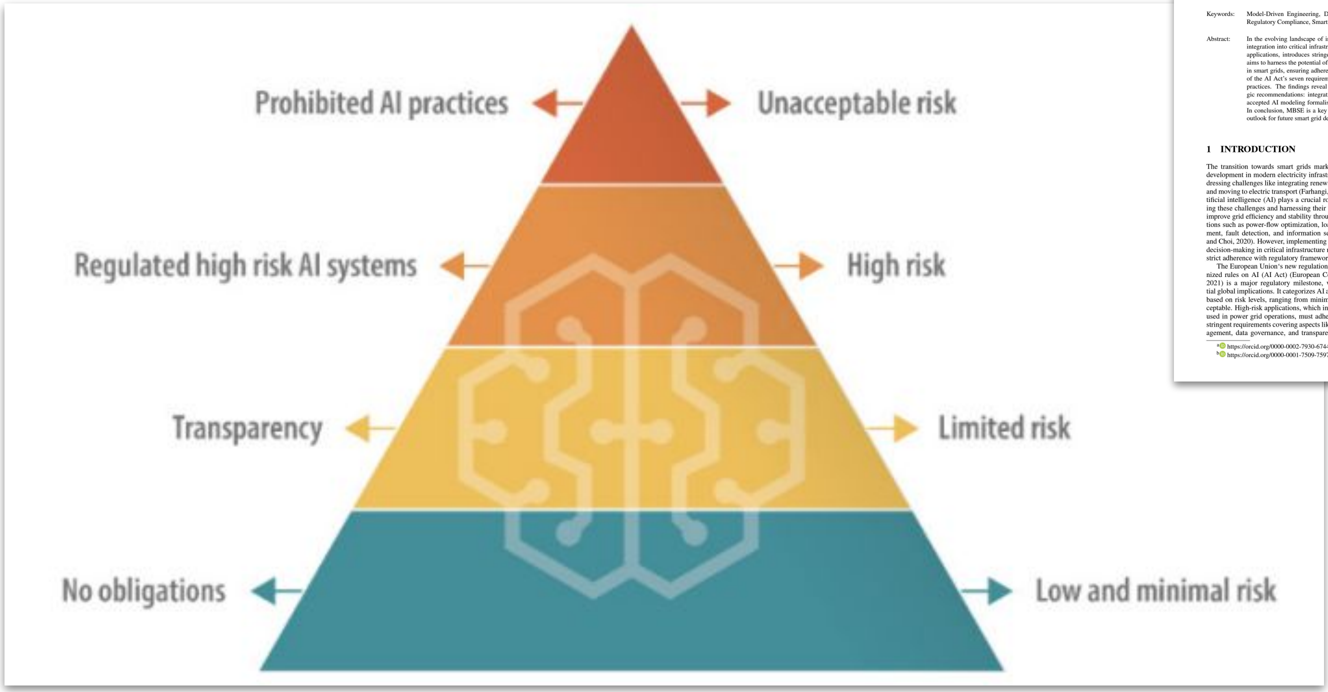
Backup slides

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 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

agence nationale de la recherche
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 sociotechnical 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)

Critères pour les études de cas

Sorry for the French

- 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, ...).

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