No Trust without Regulation!

The European Challenge on Regulation, Liability and Standards

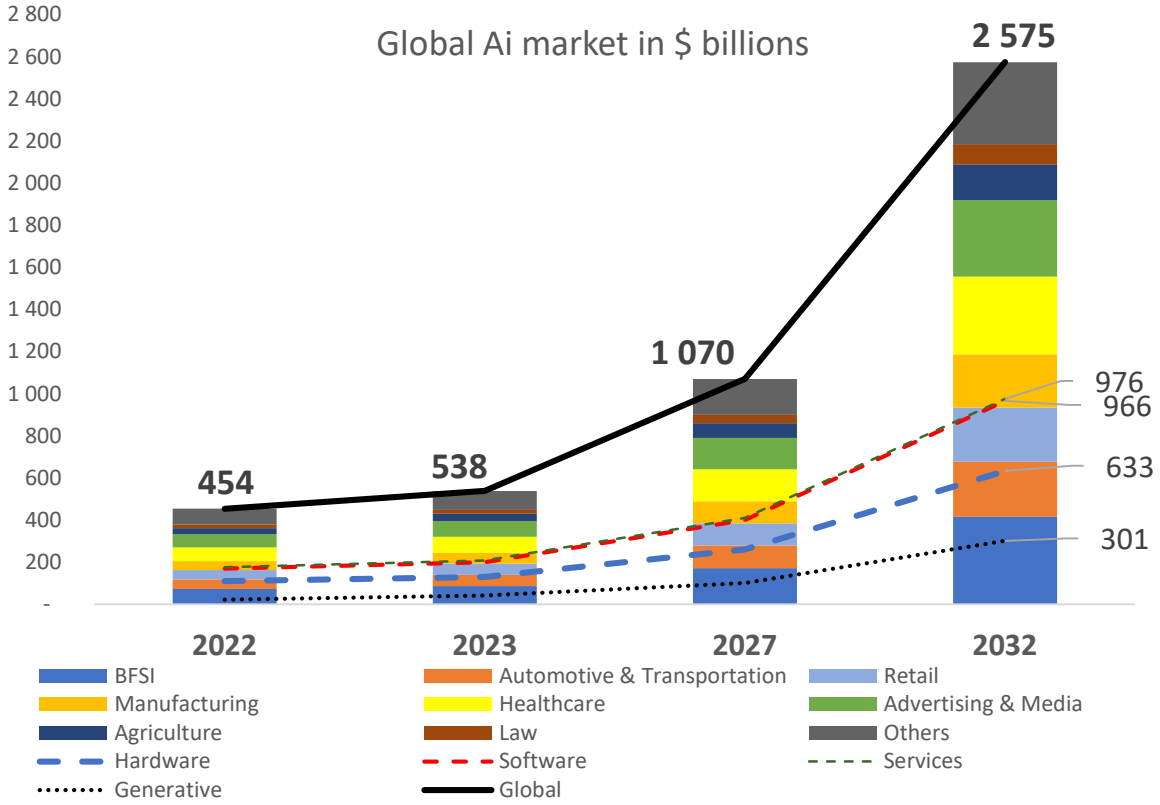
François Terrier

Program Director of List Institute

CEA's AI Senior Fellow

“A well known context…”

→ AI is there… With a very fast growth..



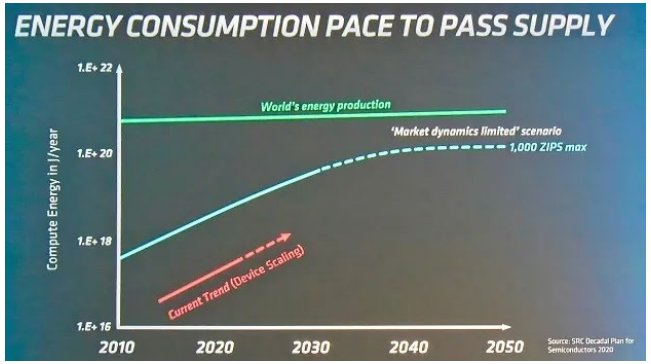
Source: Precedence research + Statista for Gen AI

All word, all domains are impacted

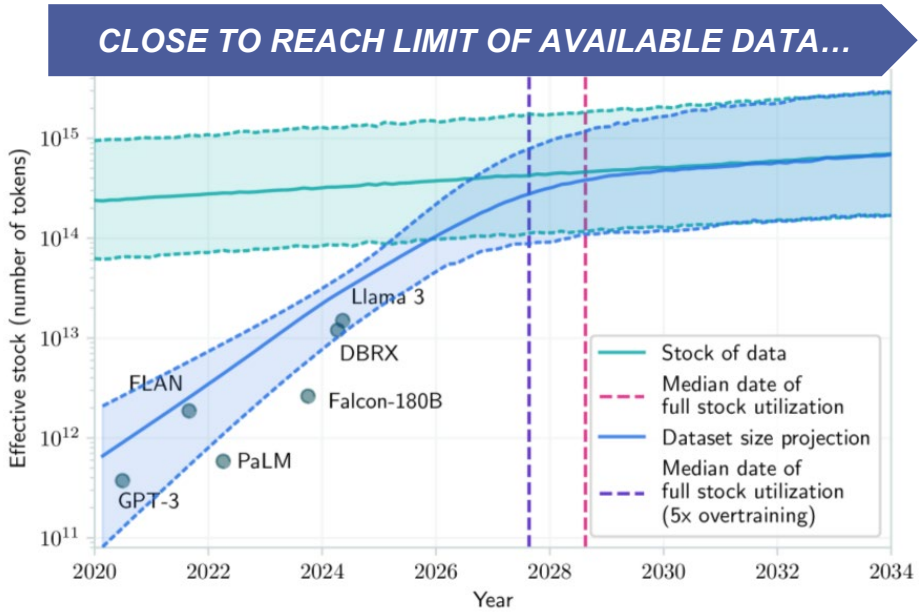
Generative AI (« the buzz ») is and would remain around 15%...

Networks and Data center: very energy and data intensive

- Energy demand will exceed production capacity
  - *Greenhouse gas (GHG) emissions have climbed 48% over the past five years (Google)*
    → Mainly due to the construction of more datacenters and the associated embodied carbon.
  - AI could use up to 1,000TWh (10^{15}) annually by 2026 – *equiv~. to Japan's electricity consumption (IEA) – 450 TWh France*

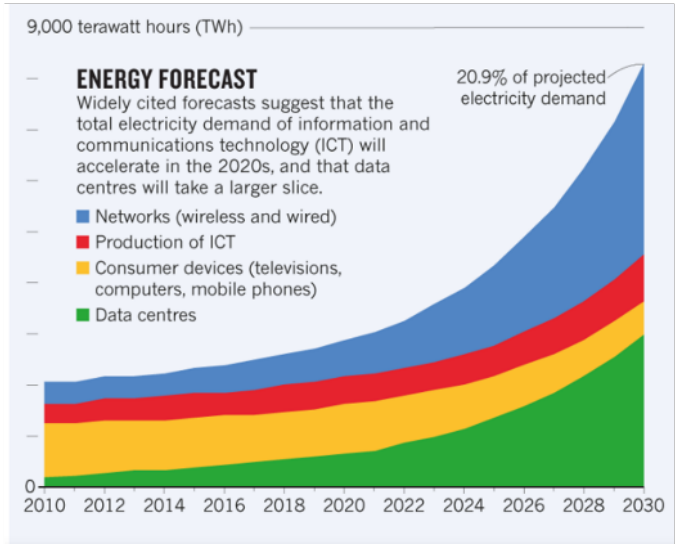


Microsoft in deal for Three Mile Island nuclear power *(837 MW)* to meet AI demand



AI, Generative AI for everythings : a dangerous "buzz"

Frugality, a question of data exchange & computing



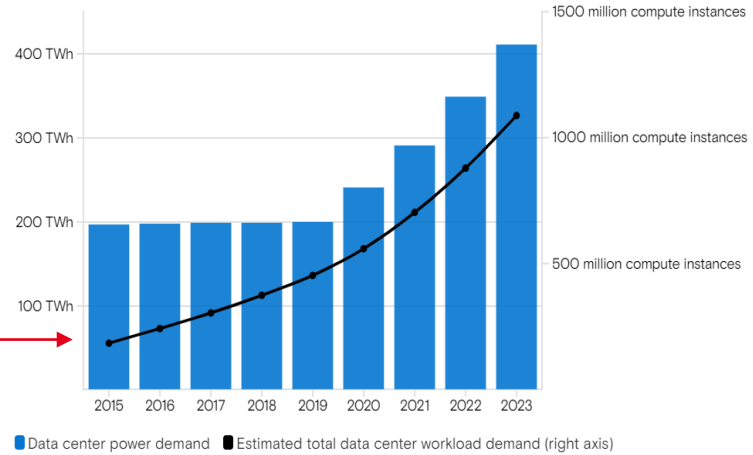
Networks

Production

Devices

Data centres

Source: *How to stop data centers from globing up the world's electricity (Nature)*



Source: Masanet et al. (2020). Cisco IEA Goldman Sachs Research

Training

	GPT-3	GPT-4
Energy consumption	**1,300 MWh**	**7,200 MWh**

Inference

	GPT-3	GPT-4
Energy per Query	0.0003 kWh	0.0005 kWh
Total Queries per Day	10,000,000	10,000,000
Total Annual Energy	**1,095 MWh**	**1,825 MWh**

AI is there with new and huge challenges

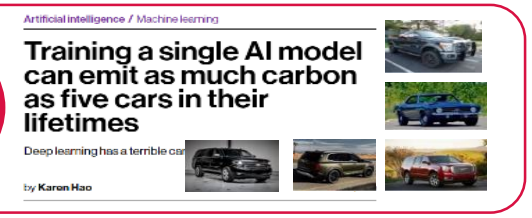
A lot of new applications

Automation

A lot of data, computation

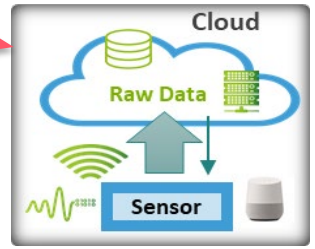
TRUST challenge

FRUGALITY challenge



The user expectations: from Cloud to edge in continuity, AI closer to users

Data today:
- 80 % in the cloud
- 20 % at edge
But within few years:
- 20 % in the cloud
- 80% everywhere

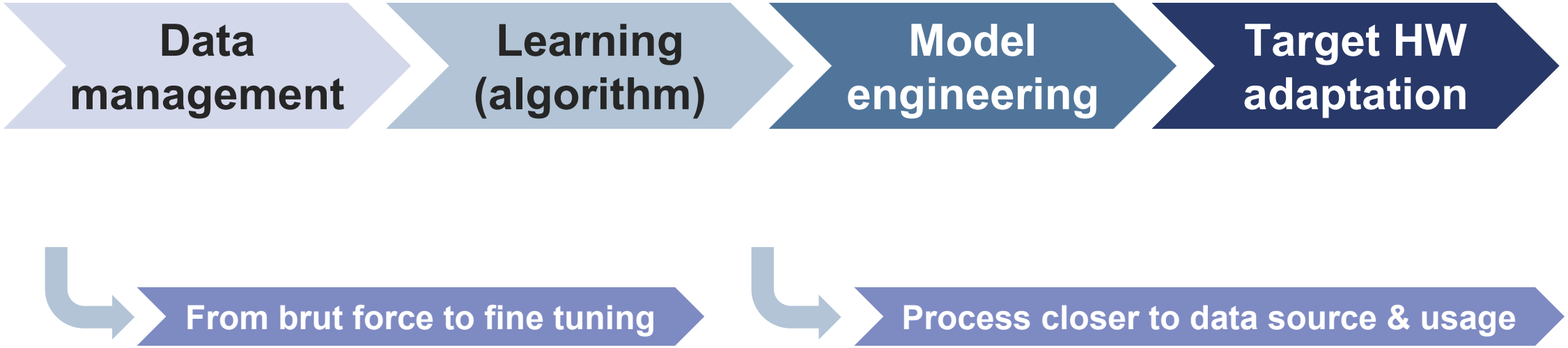


Learning and Inference in the Cloud

Learning and Inference at the edge

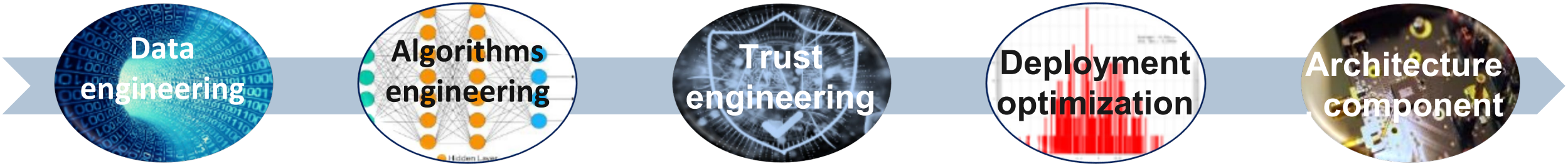
Low Latency *(shortest paths)*
Reliable operations
(even with intermittent connectivity)
Data Privacy *(remain at the edge)*
Energy Efficiency *(less data exchange)*

A continuum of challenges along the whole developpement process



A continuum of AI challenges
A continuum of CEA competencies and technologies

From Data to Implementation



Need of TRUST → Need of SAFETY!

TRUST: the need is there

Ali Rahimi
(Google)

Machine learning becomes alchemy

Yann LeCun
(facebook)

Engineering artefacts preceded the understanding of the theory

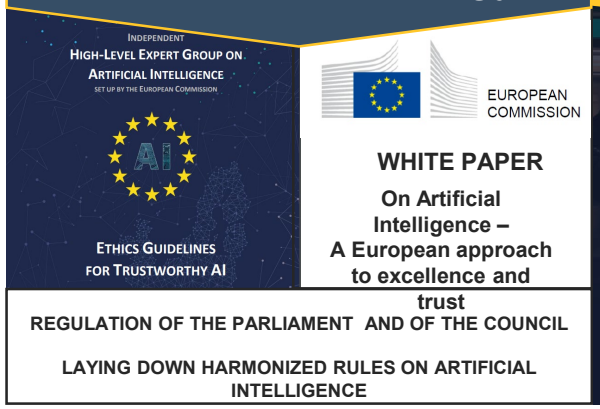
The technology arrives with a poor (industrial) maturity and evident weakness on the definition of the

usages, specifications, design methods, robustness metrics, quality processes…

Breaks all principle of safety certification processes…

AI DEPLOYMENT A GREAT OPPORTUNITY and A STRATEGIC ISSUE

A European strategy



A French strategy

An industrial strategy

From 2017 to 2024: setting up regulations

Need of policy

Outside of Europe: still at stage of recommandations…

NIST National Institute of Standards and Technology U.S. Department of Commerce

www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

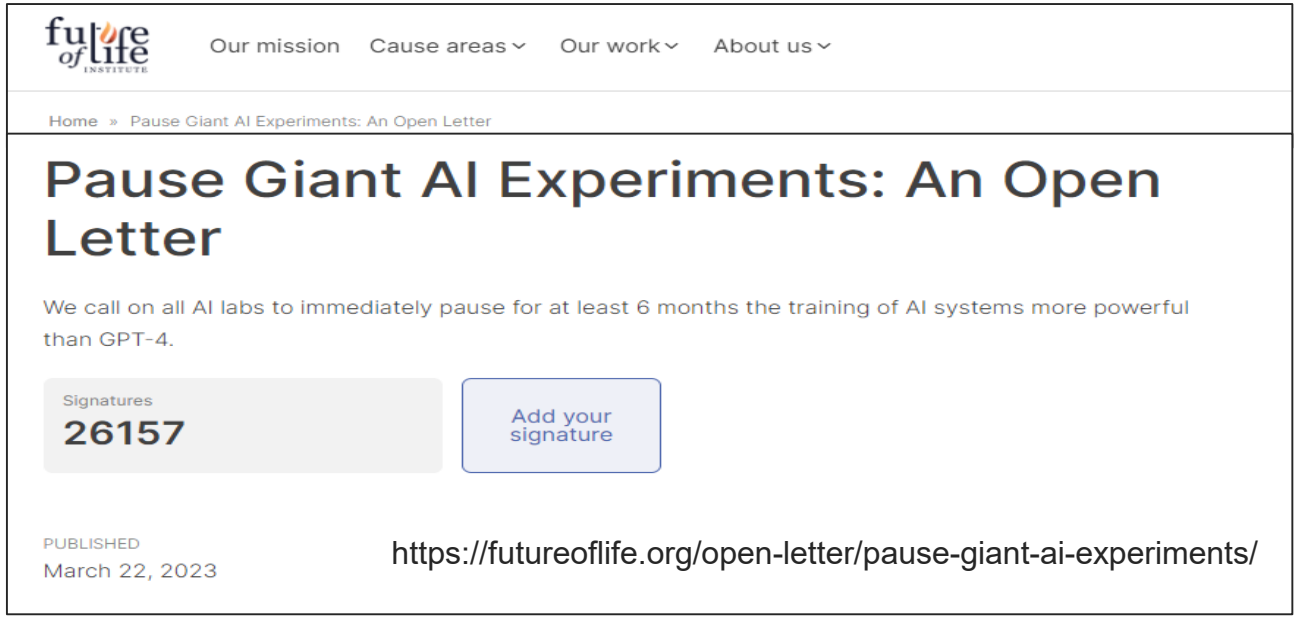
The AI RMF is intended for voluntary use in addressing risks in the design, development, use, and evaluation of AI products, services, and systems.

OECD BETTER POLICIES FOR BETTER LIVES

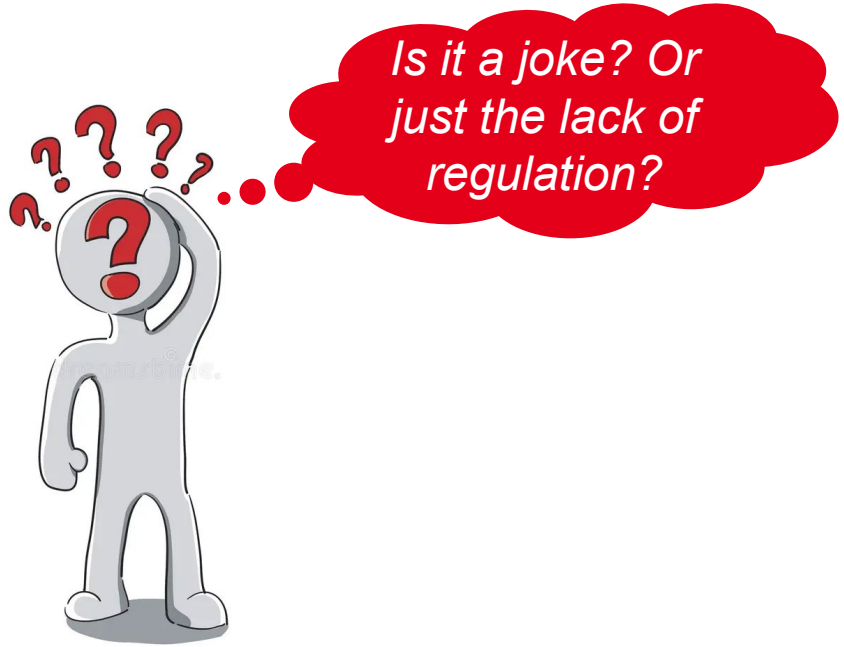
https://oecd.ai/en/ai-principles

… innovative and trustworthy and that respects human rights and democratic values. (May 2019)

And others…



https://futureoflife.org/open-letter/pause-giant-ai-experiments/



European approach to ethics and regulation

Expert group analyse the subject of TRUST

2018

2019: refinement for key sectors

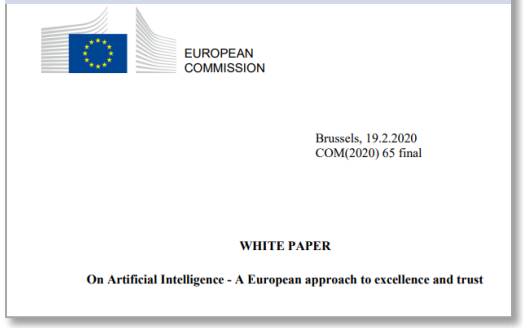
- Manufacturing
- Health
- Justice
- E-Government

Trustworthy AI should be:
(1) lawful - respecting all applicable laws & regulations
(2) ethical - respecting ethical principles and values
(3) robust - both from a technical perspective while taking into account its social environment

7 key requirements:
- Human agency and oversight
- Technical Robustness and safety.
- Privacy and data governance.
- Transparency.
- Societal and environmental well-being.
- Accountability.

European commission vision

2020: Approach for excellence and Trust


EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

Human-centric AI:
- **AI system builder is responsible → robustness, safety, privacy, transparency…**
- **Human right must be respected and not subject to automated decision only**

Toward a european regulation

2021: Proposal to the parliament


EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

Ethics imperatives

Europe makes a step forward!

AI is HUGE and EVERYWHERE!

But ...

TRUST
(Quality)
« How?»

FRUGALITY
(Embedded)
« Where? »

USAGE
« Why? »

Quality & Technical

Standards

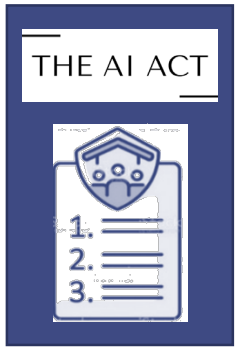
Regulation

Ethics

Toward a GLOBAL POLICY

Standards and regulations: opportunities

- **European regulations**

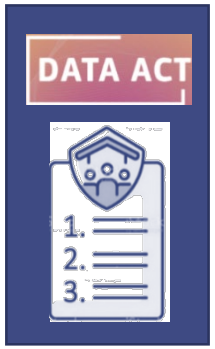


→ **Will secure the uses**

✓ Clear definition of responsibilities

✓ Explicit definition of expectations

« Based on best effort of qualification with state of the art methods »



→ **Develop IoT Data market**

✓ Secure producers

✓ Facilitate data mobility

« Introduce Data market for industry »

Regulations and standards will ease AI deployment and business

→ **A national initiative: AFNOR SPEC 2314**
the first methodology to evaluate AI impact

✓ Definition of the concepts
✓ Methodology: question to ask yourself
✓ Proposition of metrics
✓ Good for communication about frugality

→ A huge work, operationally usable

→ Pushed at European level with the EC

AI Act

A global policy set up by Europe

European Commission

European Parliament

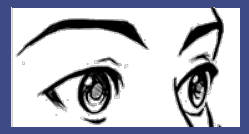
A complete approach with 3 pilars: regulation, liability, conformity

ISO

ETHICS GUIDELINES FOR TRUSTWORTHY AI

CEN CENELEC

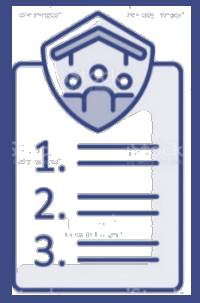
Conformity
(Standards)



Quality & Technical

THE AI ACT

Regulation

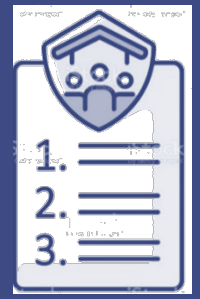


AI liability

Liability

DATA ACT

Regulation



IoT Data market

- **Secure producers**
- **Facilitate data mobility**

- Effective:11/01/2024
- Applicable: 09/2025.

Toward an European regulation for AI deployment respecting the european values

European Parliament presentation: www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792
The act (108 pg) : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

- A strong European legislation,
  (*i.e applicable as it is to any system or service provided in any EU country)*

2021: Proposal to the parliament


EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

- **List of prohibited AI**
- **Risk classification**
- **Rules for high risk AI systems**
- **Transparency obligations**
- **Support to innovation**

Technology neutral

Risk based approach

Market focuss
(preserve research)

AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206&qid=1692668629969
Final amendments: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

Toward an European regulation for AI deployment respecting the european values

- A strong European legislation,
  (*i.e applicable as it is to any system or service provided in any EU country)*

- Technology neutral
- Risk based approach
- Market focuss *(preserve research)*

- **List of prohibited AI**
- **Risk classification**
- **Rules for high risk AI systems**
- **Transparency obligations**
- **Support to innovation**

- Effective: 01/08/2024
- Prohibition of uses: 02/02/2025
- General purpose AI: 02/02/2025
- Low and limited risk: 02/08/2026
- High-risk systems: 02/08/2027

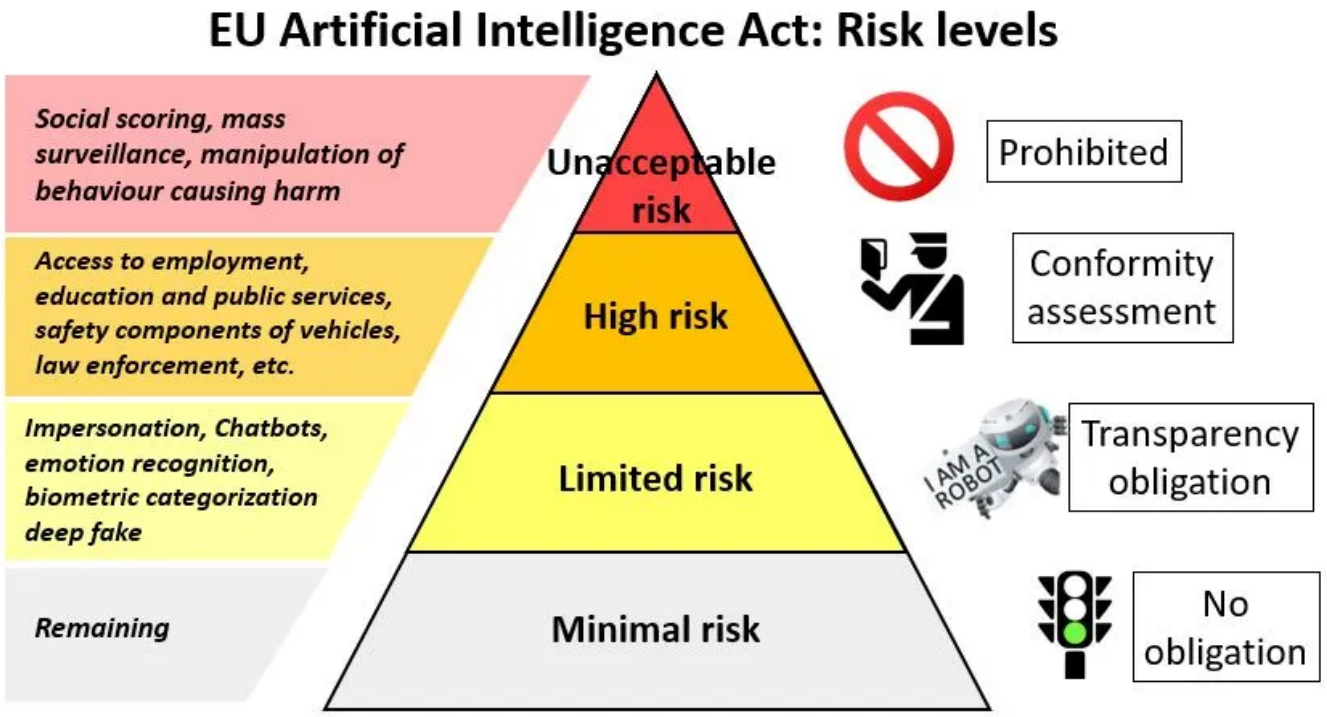
”e.g.: toys, radio equipment, medical devices...”

Remark: **Defense is out of application scope of the AI Act**, but is of strong interest for Defense

→ E.g.: In France, DGA has build a quality guide for AI development.
It will integrate in an incoming update integrating more finely Safety considerations
(*with collaborations with CEA, CNRS, Inria and others academics*)

Toward an European regulation: centered on the usage and risk analysis…

➔A risk based approach



3 levels of risks
depending of the usage domain
+ 1 Specific case

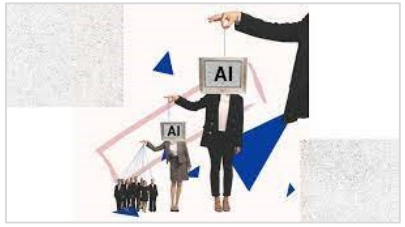
- High risk systems: *Safety obligations*
- Foundational models: *Dedicated Transparency obligations*
- Limited risk: *Transparency obligations*
- Low or minimal risk: *No obligations*

European regulation: centered on the usage

ETHICS GUIDELINES FOR TRUSTWORTHY AI

European Parliament

THE AI ACT



PROHIBITED USAGES

Title II (Article 5) + Annex II

- **COGNITIVE BEHAVIOURAL MANIPULATION leading to possible harm**
  - AI systems that deploy manipulative 'subliminal techniques' (incl. neuro-technology with AI)
  - AI systems that exploit specific vulnerable groups (physical or mental disability)
- **SOCIAL SCORING & Cie by PUBLIC AUTHORITIES**
  - Law enforcement with predictive systems (e.g.: based on profiling, location or past criminal behavior)
  - Emotion recognition in law enforcement, border management, the workplace, educational institutions
- **REAL TIME & REMOTE BIOMETRIC IDENTIFICATION**
  - "Real-time" and "Post" (systematic) remote biometric identification (or analysis of public footage) systems in publicly accessible spaces (except for serious crimes after judicial authorization)
  - Indiscriminate and untargeted scraping of biometric data from social media or CCTV footage to create databases

AI Act, in practice

The main category of interest is « High risk systems »

"risk of harm to the health and safety or a risk of adverse impact on fundamental rights"

(=> system whose use inherently entails some risk, and in which we want to include AI)

➔ They require conformity assesment

1. Against the AI Act requirements (cf Annex IV)
2. Against their own existing regulations (cf. Annex II)

They includes all level of risk of all the safety related systems according to the existing domain regulation (listed in Annex II)

Specific cases have been added (not already regulated bu specific regulation) (cf. Annex III)

Toward an European regulation…

European Parliament

THE AI ACT

High risk systems

Title III (Article 6)

- **FOR ALL DOMAINS WITH EXISTING REGULATIONS:** **Annex II**

  → General principles **+ requirement of EU domain regulations**

  - *Ex. : transportation, energy, medical, toys, lifts, machinery, radio, pressure…*

+ SPECIFIC SOCIETAL AREAS *(requiring registration)* **as:** **Annex III**

- *Management of critical infrastructure; (road, water, energy…)*
- *Education/employment access/assessment;*
- *Access to essential private/public services; (credit, allocation, emergency dispatching…)*
- *Recommendation systems (>45M use*
- *Biometry & person categorization;*
- *Democratic process; Law enforcem*
- *Migration; Administration of justice*

Requirements

European Parliament

High-risk AI based systems

- ***RISK ANALYSIS AND ESTIMATION ACCORDING TO USE***

- *Data set pertinent, representative, free of errors and complete*

- *Technical documentation establishing conformity to requirements*

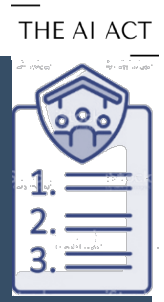
- ***Automatic recording of events ('logs')***

- *Sufficient transparency* ***TO INTERPRET OUTPUT & USE IT APPROPRIATELY***

(→ continuously maintained during the system life)

Requirements

European Parliament



FOUNDATIONAL MODEL based applications

- **Assess and mitigate possible risks** *(regarding to the possible uses and contexts of use)* (to health, safety, fundamental rights, the environment, democracy and rule of law)
- **Register** the models in the EU database before release on the EU market
- Comply with **transparency** requirements:
  - *Disclosing that the content was AI-generated*
  - *Ensure safeguards against generating illegal content*
  - *Provide publically detailed summaries of the copyrighted data used for training*
  - *Provide capabilities measuring and logging resource consumption* *(over their entire lifecycle)*

Should be technology agnostic…

Challenge is how to adopt a risk oriented approach when usages are not known in advance

Requirements

European Parliament

Limited-risk AI based systems *(not based on Generative AI)*

- *Such as systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorisation systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes)* → ***limited set of transparency obligations****:*

  - Interacting with natural persons:
    ensure natural persons are informed of the AI nature
  - Disclose that the content has been artificially generated or manipulated

Title IV (Article 52)

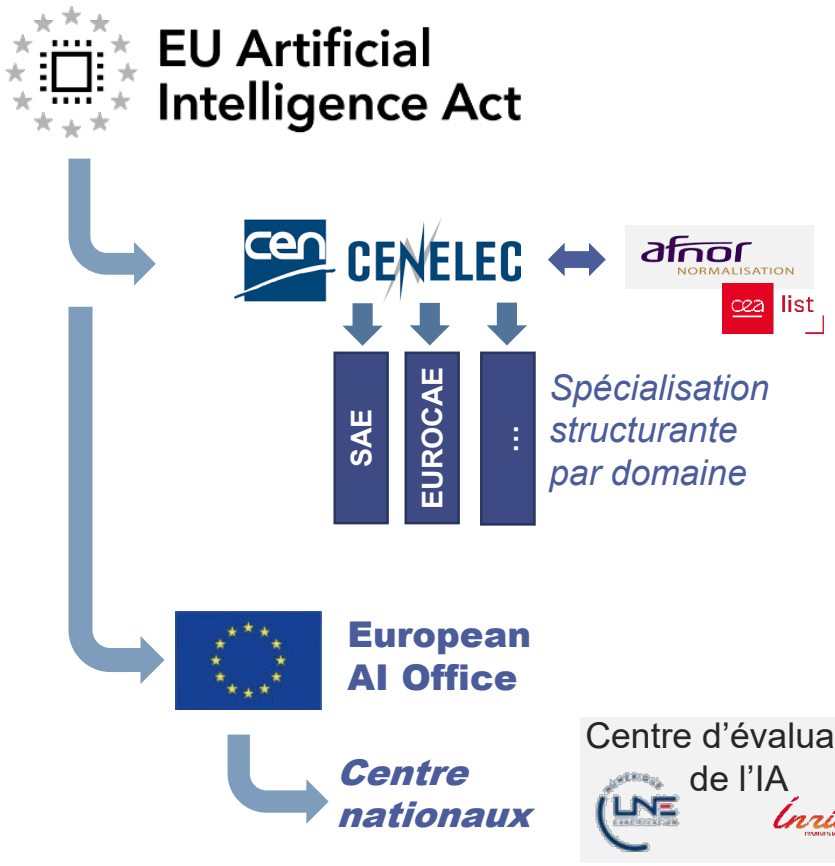
« Certain AI systems », « Limited risk systems », « Low-risk systems »

AI liability: a more protective law

The specific characteristics of AI
make it particularly difficult to meet the burden of proof for a successful claim
(e.g. opacity/lack of transparency, explainability,
autonomous behaviour, continuous adaptation, limited predictability)

- **Adaptation of law to allow for compensation for damages without the need to prove a fault**
  - *Reduce liability rules uncertainty and risk of legal fragmentation*
  - *Causality not mandatory*
    (except for High Risk AI because they have to provide transparency and thus give the means to establish causality links)
- **Responsability to the provider of product and services (depending on context of use)**
- **Will depend to impact analysis depending of the risks entailed by the uses**
- **Will depend to quality of development, transparency, effective oversight by natural persons**

IA Act – une réglementation en marche



La réglementation :

- Adoption Parlement et Conseil européens : 21/05/2024
- Publication journal officiel → juillet 2024
- Interdiction usage prohibés → ~janvier 2025

✓ *Publication des standards de conformité → ~avril 2025*

- Entrée en vigueur opérationnelle de la loi → ~août 2025
  - ➢ Impact fort : usage à « haut risque » *(critiques/sensibles)*
  - ➢ Impact modéré : « risques limités » → processus qualité

Implication CEA

- ❖ Développement expertise et technologie — Confiance ai — THALES
- ❖ Contribution à la normalisation (sujets « risques »)
- ❖ Construction d'une offre outillée support aux processus de développement et de qualification → startup **SafenAI** (11/2024)

Enjeu : quels leviers pour déployer le champs d'action à l'IAG

Standards organization

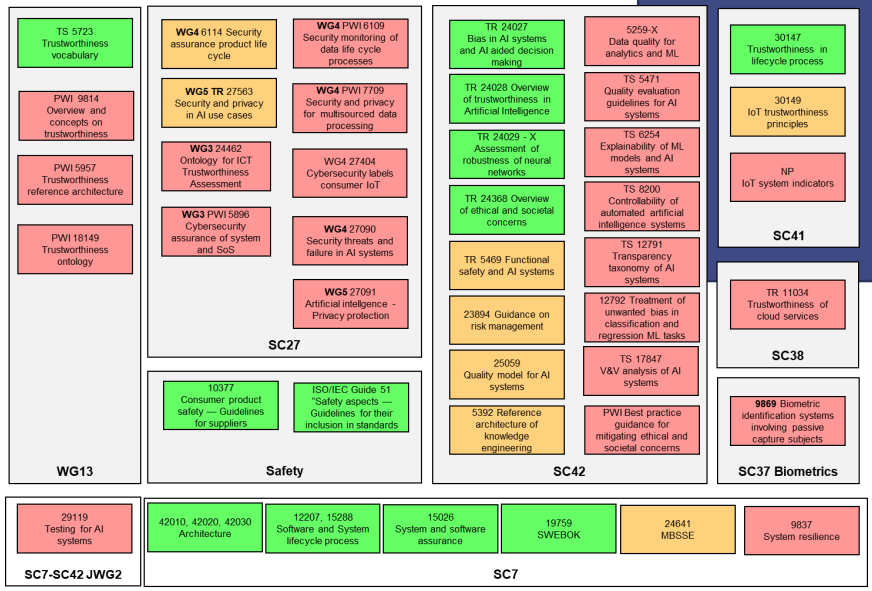
Horizontal and general standard for AI Act

Harmonised with the existing ones

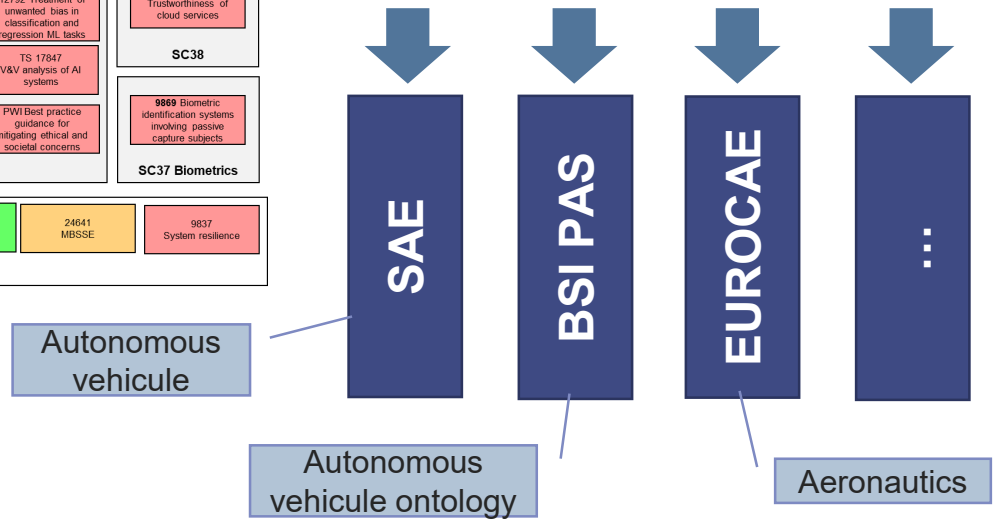
Mainly identify & reuse existing elements

A terminology and conceptual BACKBONE FOR DOMAIN STANDARDS

- *Risk management systems*
- *Governance and quality of datasets*
- *Record keeping through logging capabilities*
- *Transparency and information provisions*
- *Human oversight*
- *Accuracy specifications*
- *Robustness specifications*
- *Cybersecurity specifications*
- *Quality management systems*
- *Conformity assessment*



Credits: Henri Sohier, Co-project leader, "AI Trustworthiness characterization", CEN-CENELEC JTC21 WG4



Defines the « What » = Concepts = *Not the « How »*

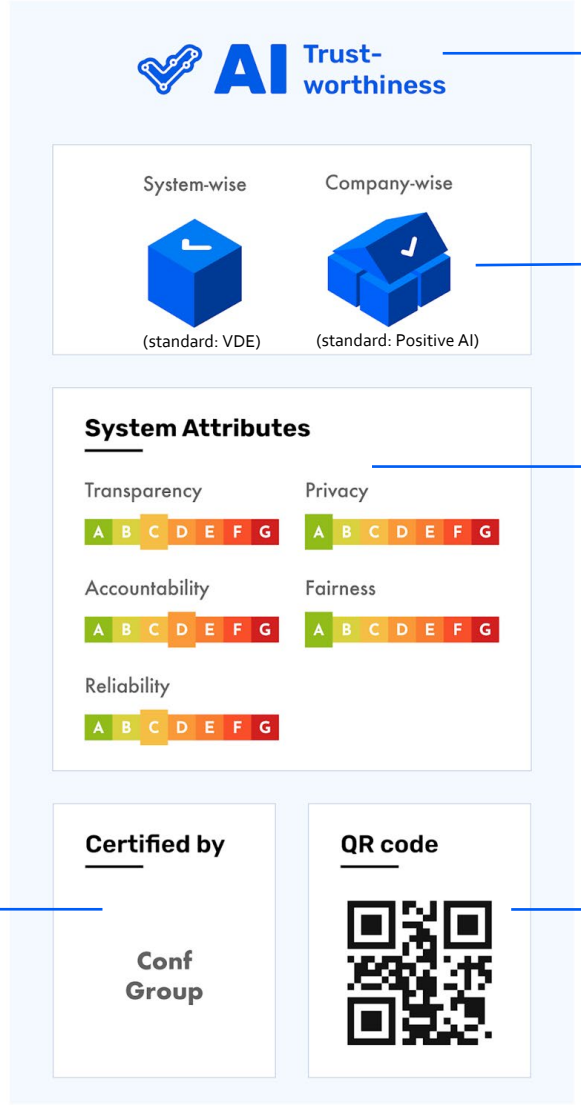
Labelling approach to complete certification

VDE

POSITIVE AI

IEEE

Confiance ai



A virtual example

Label for system trust and/or company trust (or more)
Name of the original labels or standards

System attributes in case of system trust

Optional if it can be a self-declaration

Link to more information (the summary can never show everything)

What about safety?

TRUST challenge: a set of characteristics

ENGINEERING VIEW POINT

USAGE VIEW POINT

Safety community see AI components as (physical) systems

SAFETY, CERTIFICATION,

Quality, Reliability,

Security, Privacy

Robustness, Accuracy

Traceability, Interpretability,

Ethics, Societal Impact,

Accountability,

Fairness, Explainability

TRANSPARENCY

AI community see AI components as SW

AI Act

AI BASED SYSTEM AND SOFTWARE ENGINEERING FOR TRUST

Challenges:

- **Performance**
- **Frugality**
- **Trust**

For

Responsible AI

Some activity examples

- Toward an open tool platform for Responsible AI
- Formal methods and certification
- Robustification against adversarial attacks
- Distributed AI and resilient decentralized decisions

A very large project to federate national efforts:
Open methodoloy, open generic facilitators, Closed use case/business

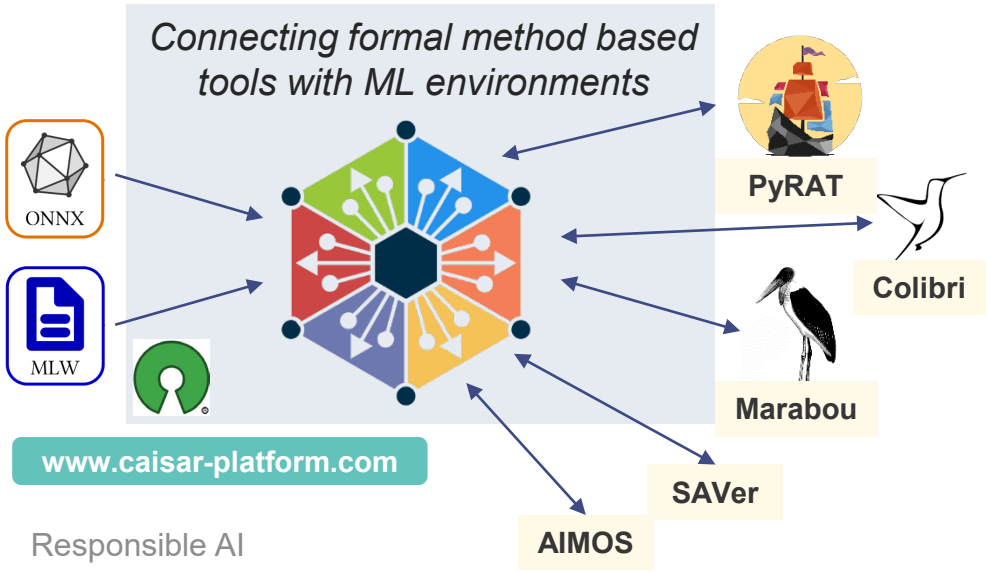
Confiance ai

FTP >300

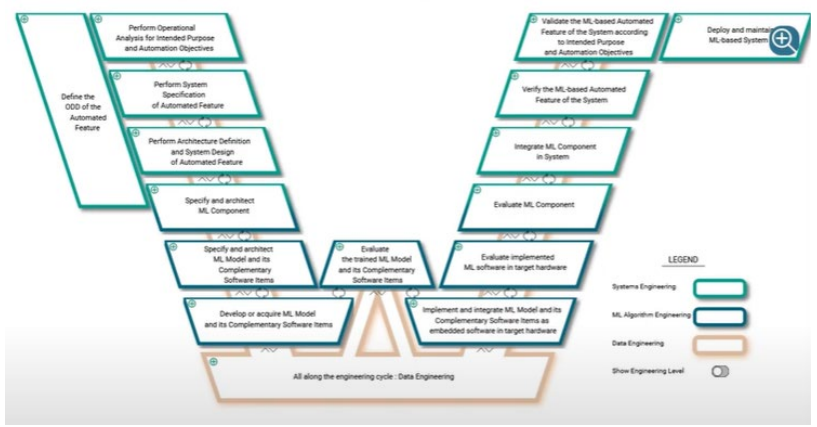
Partners >50

Use cases >20

- A set of **tools** on the whole development cycle
- Example: CAISAR open source platform



- **Evaluation** of methodology: « Body of knowledge »



bok.confiance.ai/

From research to industry: **Trust by safety**

❖ *Safe AI: Research on specification, design and verification to ensure AI safety*

Formal methods a long time ago: The prophecy

1979: "Program verification is bound to fail. We can't see how it's going to be able to affect anyone's confidence about programs", in *Social processes and proofs of theorems and programs*", Com. of ACM.

The prophets By Richard De Millo, Richard Lipton, and Alan Perlis.

- Distinguished Professor of Computing at the Georgia Tech
- VP and CTO of Hewlett-Packard
- Yale, Berkeley, Princeton, Georgia Tech
- Knuth Prize winner
- ACM, Carnegie Mellon, Yale, Purdue
- First Turing Award recipient

Fast forward a few decades

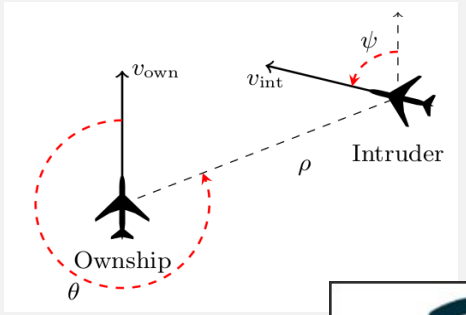
Astrée Software, PolySpace Technologies, bugseng, Modelica, frama C Software Analyzers, ANSYS, Esterel Technologies, Z3 Microsoft Research, CompCert, IRSN, Airbus, DGA, RATP, Alstom, IBM, AMD, Thales, EDF, Intel, Siemens, NASA, Schneider Electric, Safran, ST

What about AI? We have been here before…

Anti-collision control for drone: Formalized specification, but implementation too large for drones

→ NN + safety proof

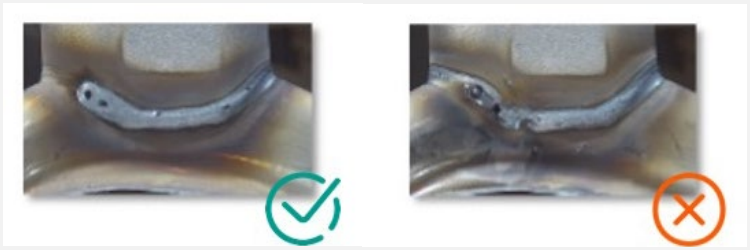
AIRBUS



Welding quality control by vision: Informal specification

→ NN +
proved robustness evaluation

Renault Group



Confiance.ai

Connected strategic programs

PEPR IA → *Priority on Machine Learning fondations*

Starts with 9 Projects, 50 partners

> 150 researchers, > 150 new PhD

73 M€ - *TRL 1-4*

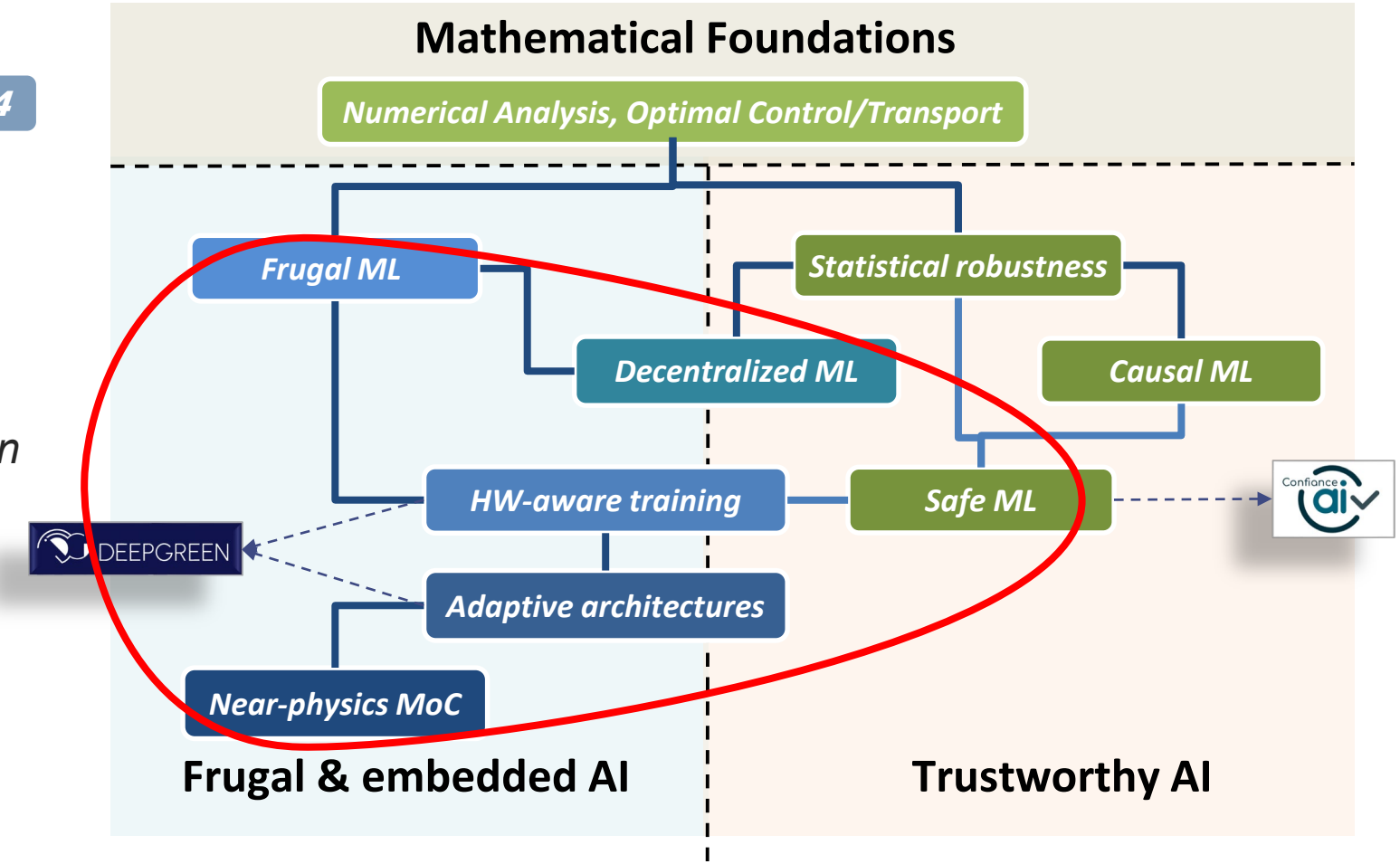
Program directors:
François Terrier, Jamal Atif, Karteek Alahari

CEA focus:

- **Frugal** AI:
  Frugal learning algorithm for Gen-AI

- **Embedded** AI:
  - *Model optimisation, code generation*
  - *Modular, flexible architectures*
  - *Emerging computation models*

- **Trust, Safe an Secure** AI:
  - *Formal method for AI safety*
  - *Secure distributed learning*



A national and open platform for AI deployment: Open the game for new business and technologies

DEEPGREEN aidge

Members 40

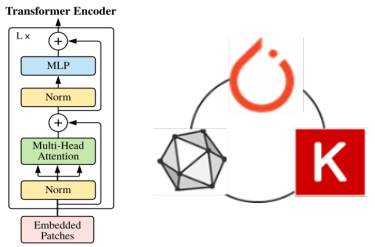
Countries 5

Partners 20

Use cases 12

Modelling

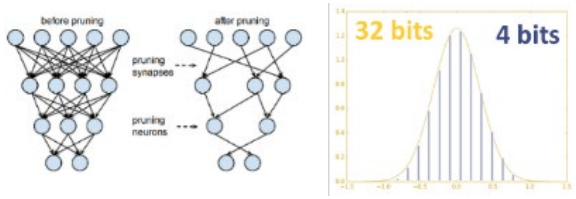
Interoperable data flow graph



State of the art models:
CNN RNN GAN Attention

Optimization

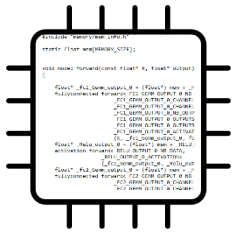
Post and Aware training optimization



Quantization, Pruning, Compression
Innovative QAT based on SAT and LSQ

Deployment

Full control of the deployment



Efficient code generation engine based on template and optimized implementation with the most appropriate language (C, C++, Cuda, ...)

deepgreen.ai/ - gitlab.eclipse.org/eclipse/aidge/

ECLIPSE FOUNDATION

Open data, models and generative AI
From buzz/research to industry: **Data and learning frugality**

❖ ***Frugal learning*** *algorithm for* ***Gen-AI****:*
Research on algorithms and methods to build efficient operational usages.
Use case examples:

Common lab with Thales - cortAIx
→ multimodal Gen AI for crisis analysis

THALES

AMIAD - Agence Ministérielle pour l'IA de... + Suivre
2 207 abonnés
1 mois • Modifié •

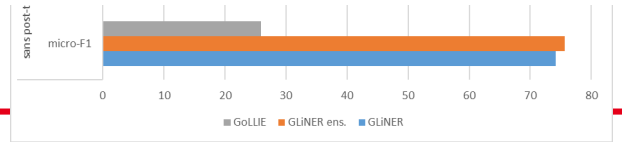
CEA-List @EvalLLM2024 : 1st place at the challenge "Prompting a LLM vs finetune a SLM?"

#JEPTALN2024 qui s'est tenue à #Toulouse les 8-12 juillet.

SLM (GLiNER – 0.3B)
Finetuning on
- few specific data
- and synthetic data

LLM (GoLLIE – 13B)
In-Context-Learning
on provided example

Few days of work



CODE COMMONS: Building a qualified code base for automated SW engineering

Software Heritage Inria INVENTEURS DU MONDE NUMÉRIQUE

OpenLLM-France LINAGORA

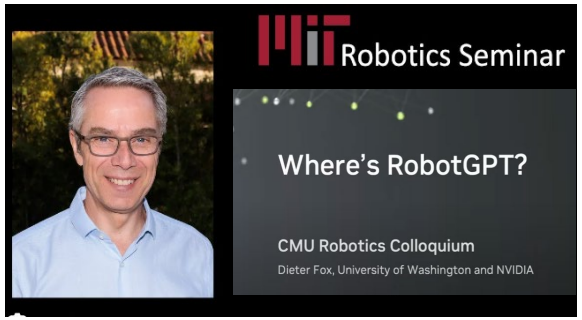
Open and sovereign set of multimodal LLMs – Application for educational assistant

Open data, models and generative AI for Robotics

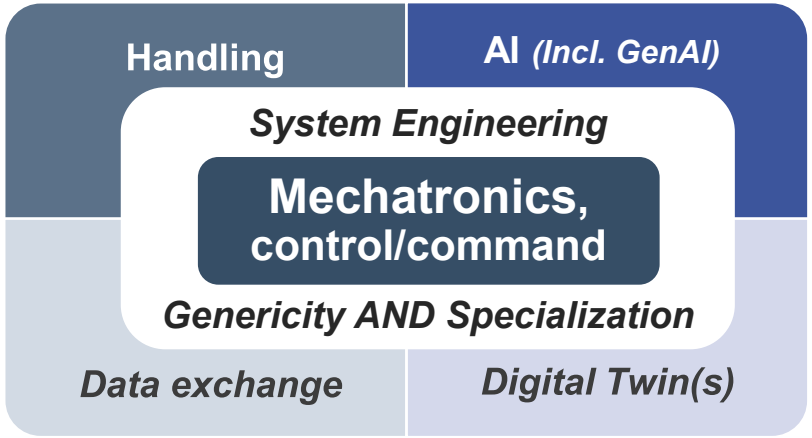
➔ Functionalization of robots is the key for the deployment

Objectives: More autonomy, mobility, analytics

➔ to more productivity, quality, flexibility, simplicity to users



https://robot-ma.github.io/MA_paper.pdf



Robotics need multiple modalities
(language, vision, #physics measures)

➔ *A huge data collection challenge*
➔ *Role of GenAI to be balanced with focused High Perf model*

❖ Open starting models, Open sharing data, could help

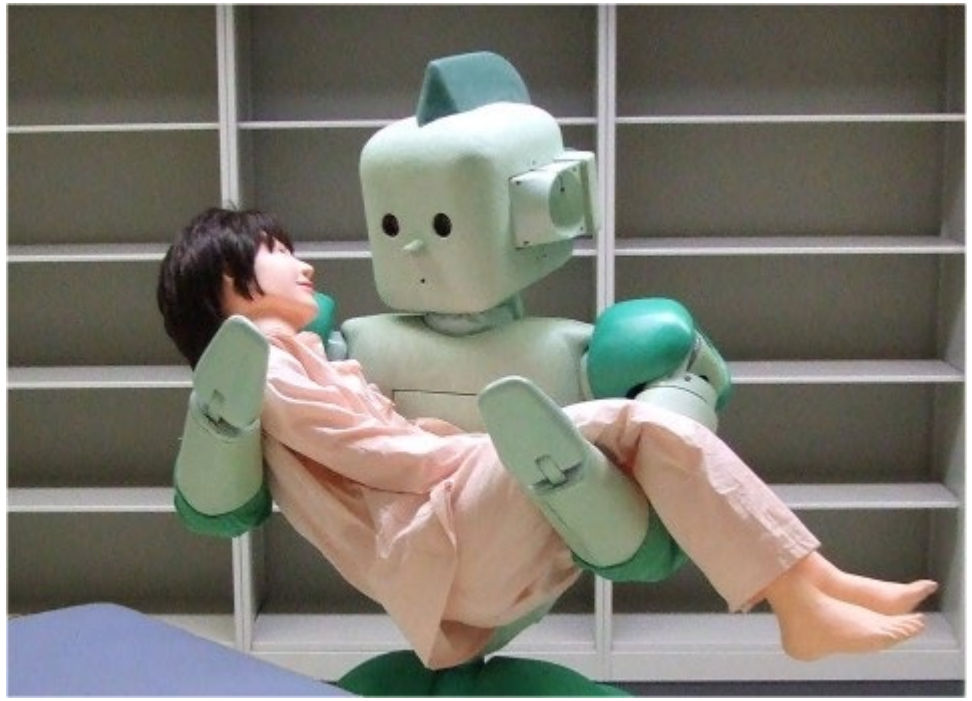
➔ With traceability of data source, qualification and usage

Autolearning robotics to faster robot programing and human interactions

pixano

pixano.cea.fr

LETS HAVE A DREAM



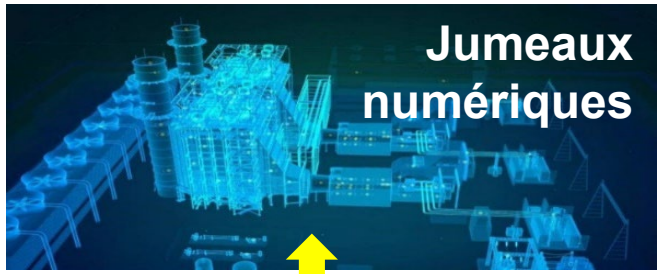
Credits: RI-MAN, RIKEN, Japan - tc.nagoya.riken.jp/RI-MAN/

Toward a kind of INDUSTRIAL METAVERSE

A booster of productivity, competitivity and business development:

Ensure and exploit in real-time the continuity of the interactions among the 3 key digital technologies of

Efficient
Secure
Safe
Frugal

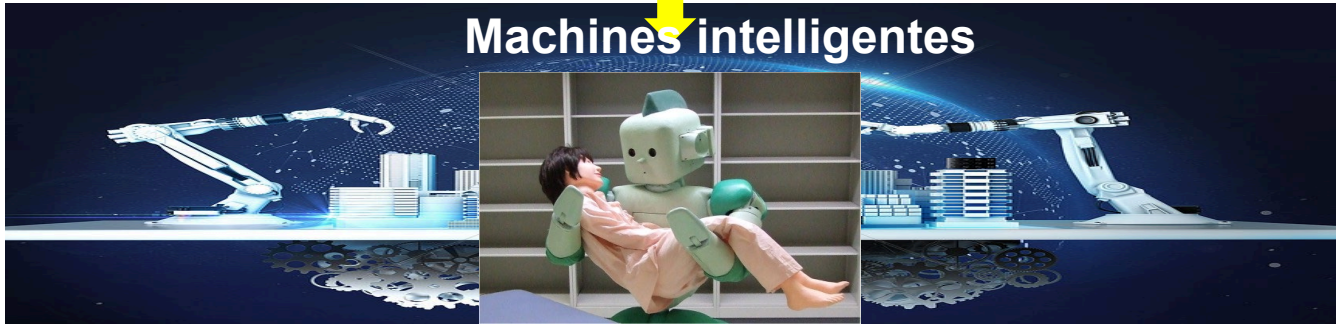


Structure
Physics
Logic

Security, privacy
Semantics, ontology
Real time connection to physical word

Market
Traceability

Intuitive programming
Auto learning



Prehension
Cooperation

Thanks!

AI is the electricity of the future,
but not the electricity fairy!

AI isn't magic:
Beyond data and computation,
AI requires a great deal of skill
and hard work
to achieve the best performance
with good efficiency and quality...

Responsible AI:

TRUST
& FRUGALITY

will make the difference

francois.terrier@cea.fr