CONSEIL DE L'INNOVATION

Grand Défi Program on "Trustworthy & Industrial AI"

Julien Chiaroni, Program Director for "Trustworthy AI for Industry"

General Secretary for Investment, French Innovation Council

(with the support of many contributors, Juliette Mattioli, Roldolphe Gelin, Christophe Guettier and al.)

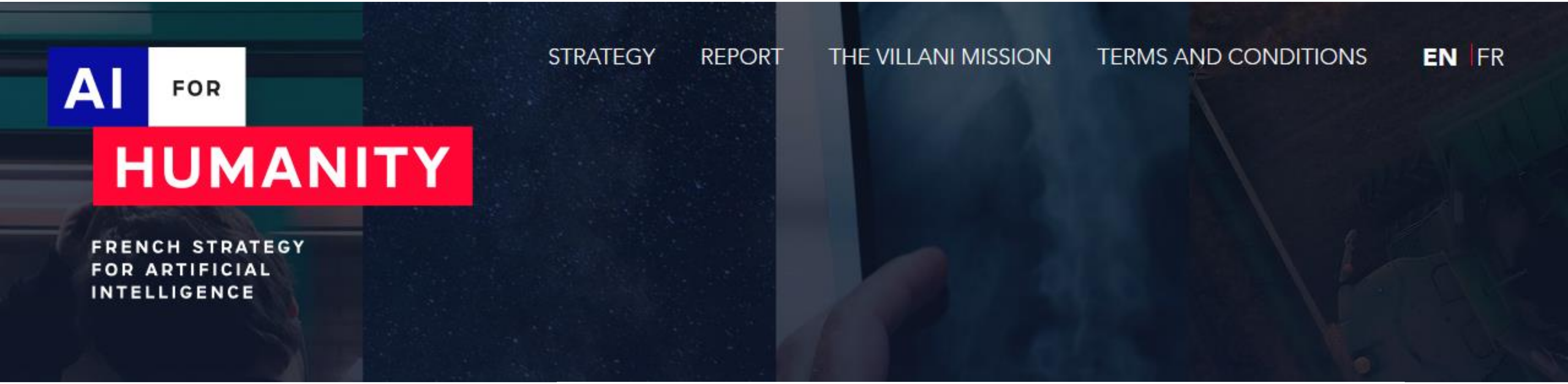
French Innovation Council

- **Define the main priorities of the French innovation policy**, supported by evaluation and prospective work

- **Foster innovation policy coordination and simplify the landscape of innovation's financial support.** In particular, it will ensure their good coordination with regional and European systems, with the aim of preparing our companies and our public research stakeholders to access the calls for projects most suited to their needs.

- **Make recommendations on the financial resources dedicated to innovation policy**, in order to encourage the emergence of breakthrough innovations and their industrialization in France

➔ **5 "Grand Défis" since 2019** (Trustworthy & Certification AI, AI for Health Diagnosis, …)



AI FOR HUMANITY

FRENCH STRATEGY FOR ARTIFICIAL INTELLIGENCE

AI MANIFESTO

Air Liquide
Président Directeur Général
Benoît Potier

Dassault Aviation
Président Directeur Général
Éric Trappier

EDF
Président Directeur Général
Jean- Bernard Lévy

Renault
Expert Leader IA
Jean-Marc David

Safran
Directeur Innovation et R&D
Stéphane Cueille

Thales
Président Directeur Général
Patrice Caine

P. Caine

Total
Président Directeur Général
Patrick Pouyanné

Valeo
Président Directeur Général
Jacques Aschenbroich

Bruno Le Maire
Ministre de l'Economie et des Finances

… and many others industrails

Trustworthy & Privacy : mains risks for AI adoption

2 priorities to support design, deploy and maintain Industrial AI based critical system

Artificial Intelligence Index Report 2021

Stanford University Human-Centered Artificial Intelligence

NUMBER of NEW AI ETHICS PRINCIPLES by ORGANIZATION TYPE, 2015-20

Source: AI Ethics Lab, 2020 | Chart: 2021 AI Index Report

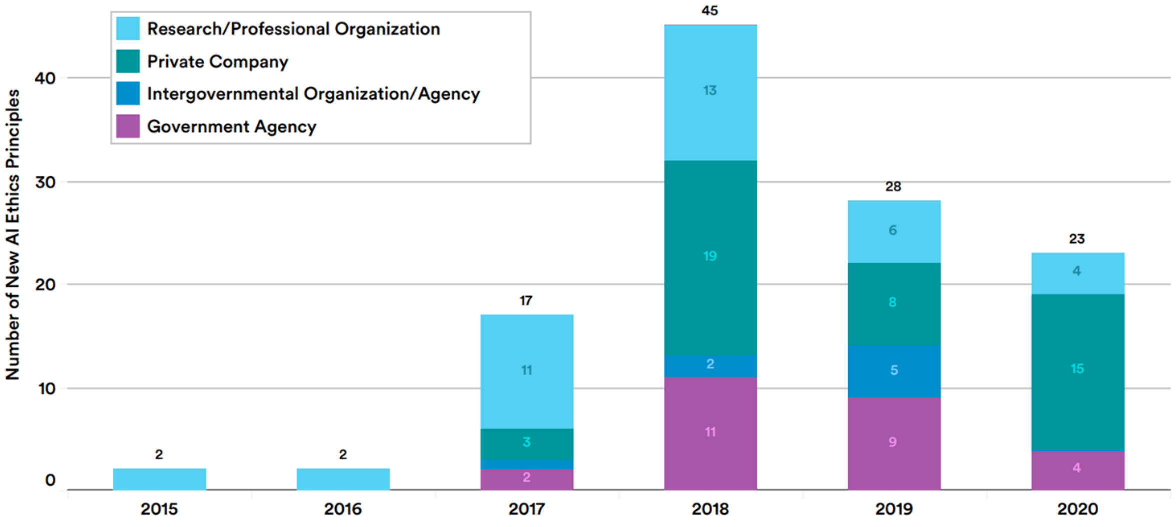
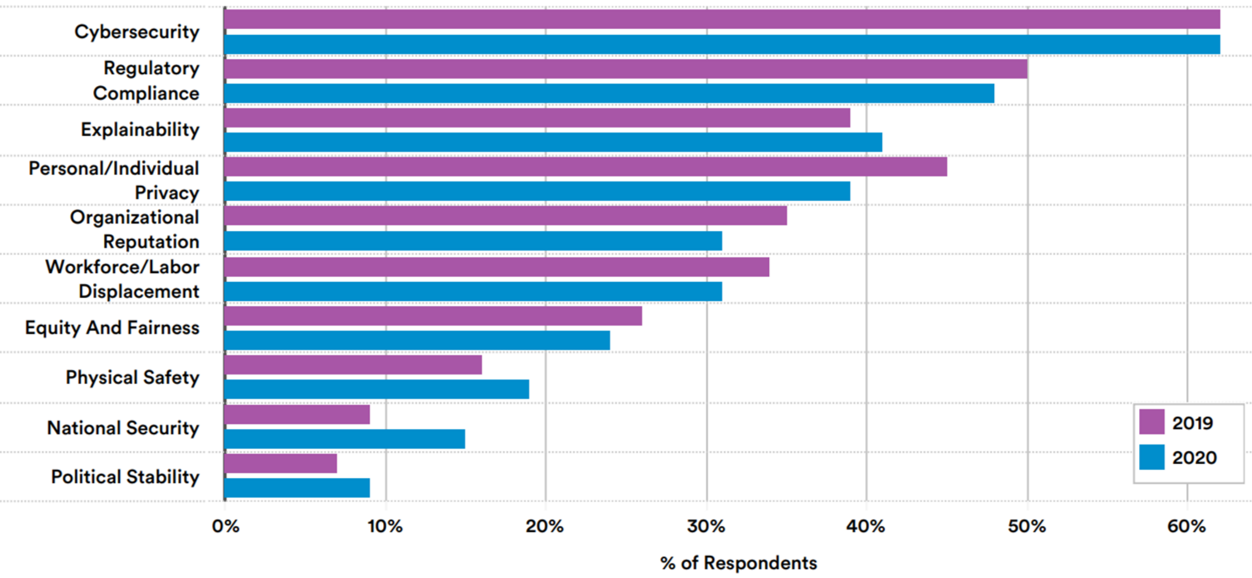


Figure 5.1.1

RISKS from ADOPTING AI THAT ORGANIZATIONS CONSIDER RELEVANT, 2020

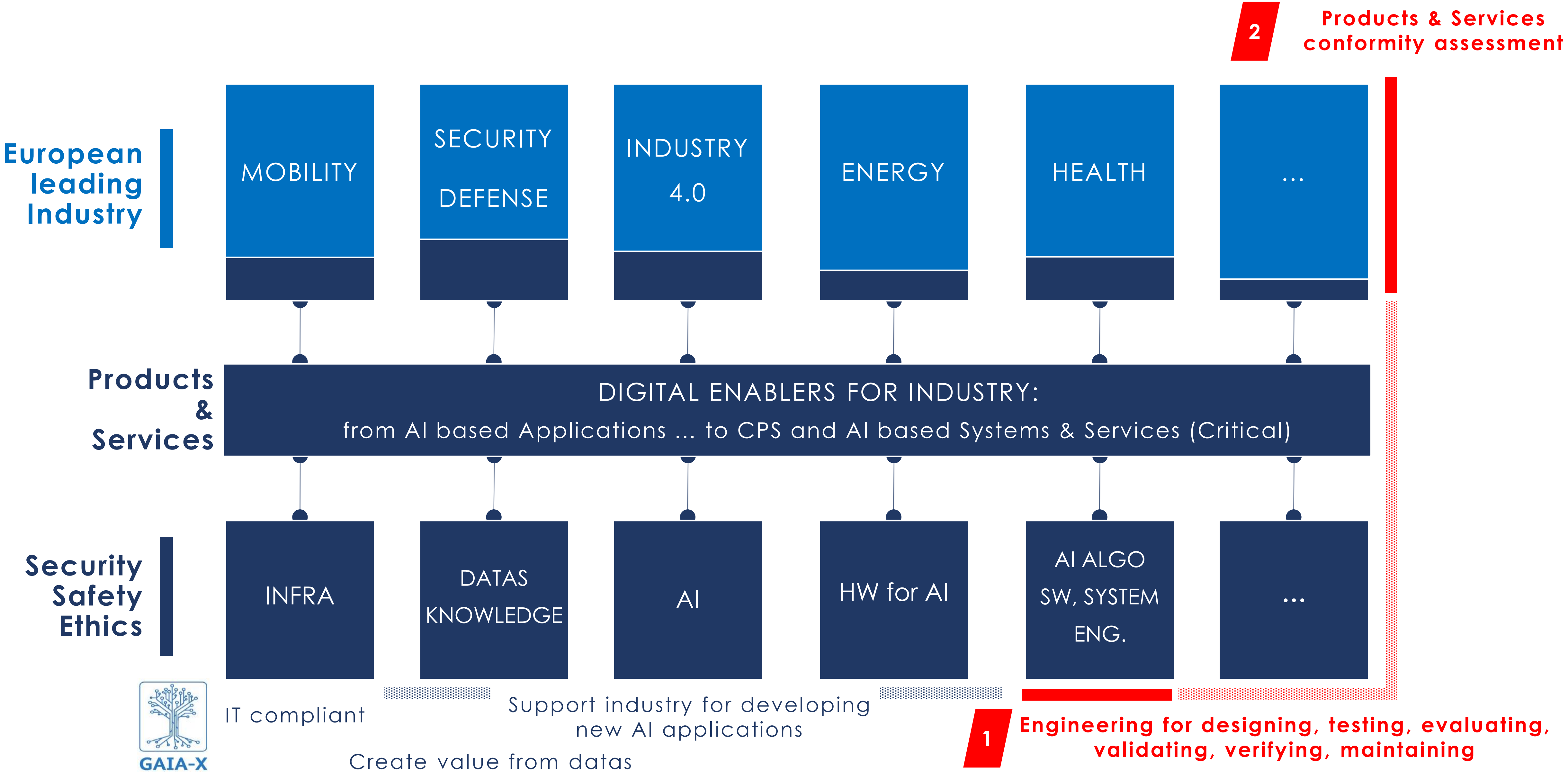
Source: McKinsey & Company, 2020 | Chart: 2021 AI Index Report



Next step : Toward technical solutions to fulfill regulatory compliance (safety, explainability, …)

Trustworthy for Industrial AI : the race for B2B applications

2 majors technical priorities to support AI applications & business in many European leading sectors



2 Products & Services conformity assessment

European leading Industry

MOBILITY

SECURITY DEFENSE

INDUSTRY 4.0

ENERGY

HEALTH

...

Products & Services

DIGITAL ENABLERS FOR INDUSTRY:

from AI based Applications ... to CPS and AI based Systems & Services (Critical)

Security Safety Ethics

INFRA

DATAS KNOWLEDGE

AI

HW for AI

AI ALGO SW, SYSTEM ENG.

...

GAIA-X

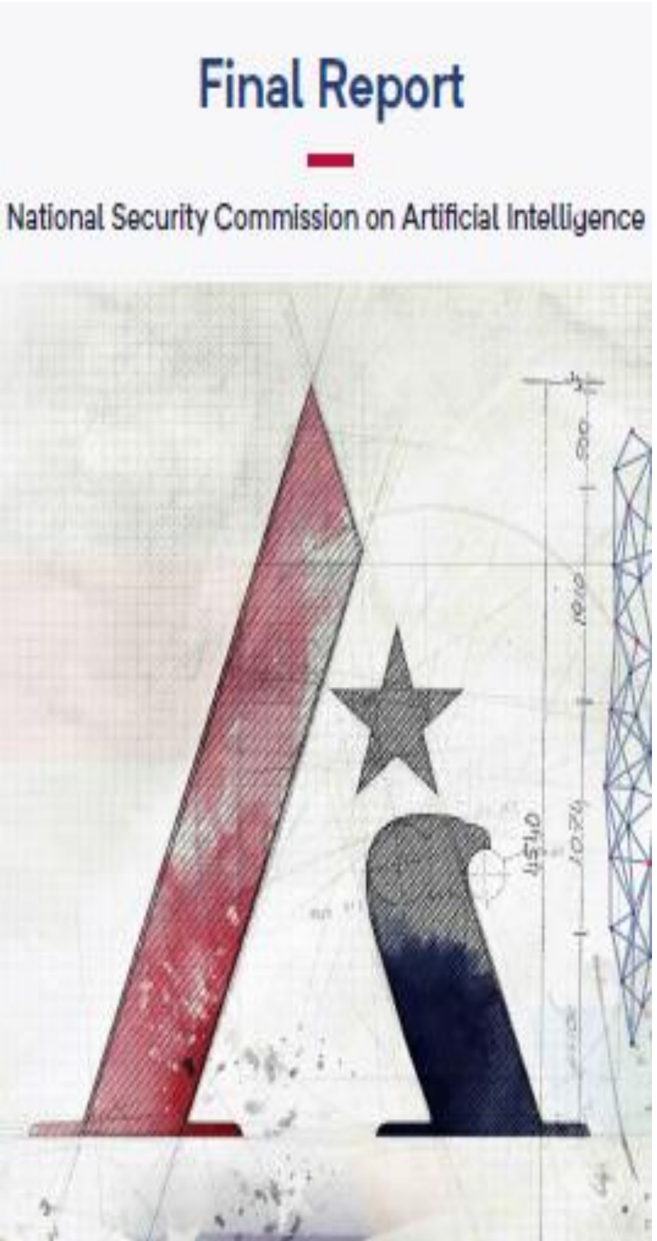
IT compliant

Support industry for developing new AI applications

Create value from datas

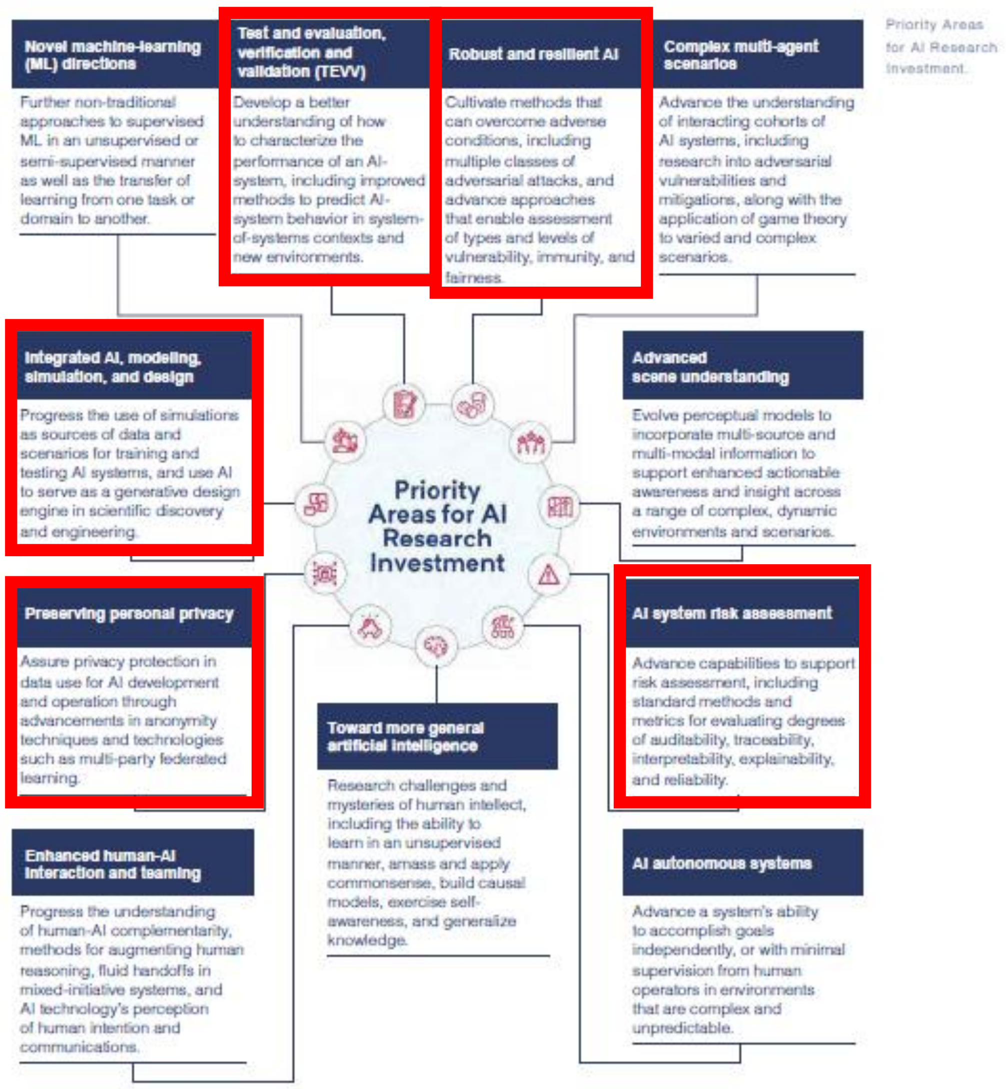
1 Engineering for designing, testing, evaluating, validating, verifying, maintaining

Priorities that are also promoted by the US National Security Commission on AI



Final Report

National Security Commission on Artificial Intelligence



Priority Areas for AI Research Investment.

Novel machine-learning (ML) directions

Further non-traditional approaches to supervised ML in an unsupervised or semi-supervised manner as well as the transfer of learning from one task or domain to another.

Test and evaluation, verification and validation (TEVV)

Develop a better understanding of how to characterize the performance of an AI-system, including improved methods to predict AI-system behavior in system-of-systems contexts and new environments.

Robust and resilient AI

Cultivate methods that can overcome adverse conditions, including multiple classes of adversarial attacks, and advance approaches that enable assessment of types and levels of vulnerability, immunity, and fairness.

Complex multi-agent scenarios

Advance the understanding of interacting cohorts of AI systems, including research into adversarial vulnerabilities and mitigations, along with the application of game theory to varied and complex scenarios.

Integrated AI, modeling, simulation, and design

Progress the use of simulations as sources of data and scenarios for training and testing AI systems, and use AI to serve as a generative design engine in scientific discovery and engineering.

Advanced scene understanding

Evolve perceptual models to incorporate multi-source and multi-modal information to support enhanced actionable awareness and insight across a range of complex, dynamic environments and scenarios.

Priority Areas for AI Research Investment

Preserving personal privacy

Assure privacy protection in data use for AI development and operation through advancements in anonymity techniques and technologies such as multi-party federated learning.

AI system risk assessment

Advance capabilities to support risk assessment, including standard methods and metrics for evaluating degrees of auditability, traceability, interpretability, explainability, and reliability.

Toward more general artificial intelligence

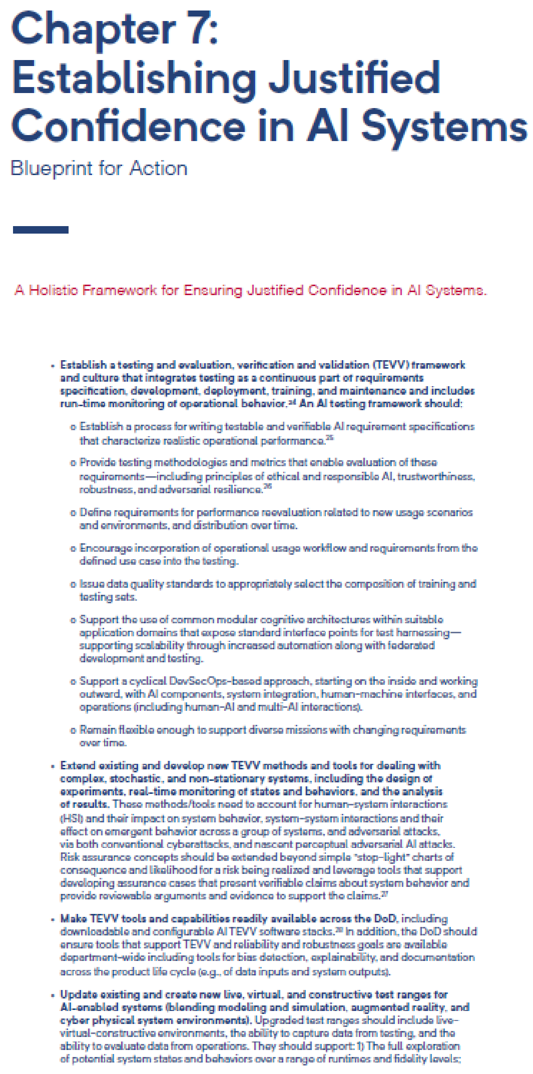
Research challenges and mysteries of human intellect, including the ability to learn in an unsupervised manner, amass and apply commonsense, build causal models, exercise self-awareness, and generalize knowledge.

Enhanced human-AI interaction and teaming

Progress the understanding of human-AI complementarity, methods for augmenting human reasoning, fluid handoffs in mixed-initiative systems, and AI technology's perception of human intention and communications.

AI autonomous systems

Advance a system's ability to accomplish goals independently, or with minimal supervision from human operators in environments that are complex and unpredictable.



Chapter 7: Establishing Justified Confidence in AI Systems

Blueprint for Action

A Holistic Framework for Ensuring Justified Confidence in AI Systems.

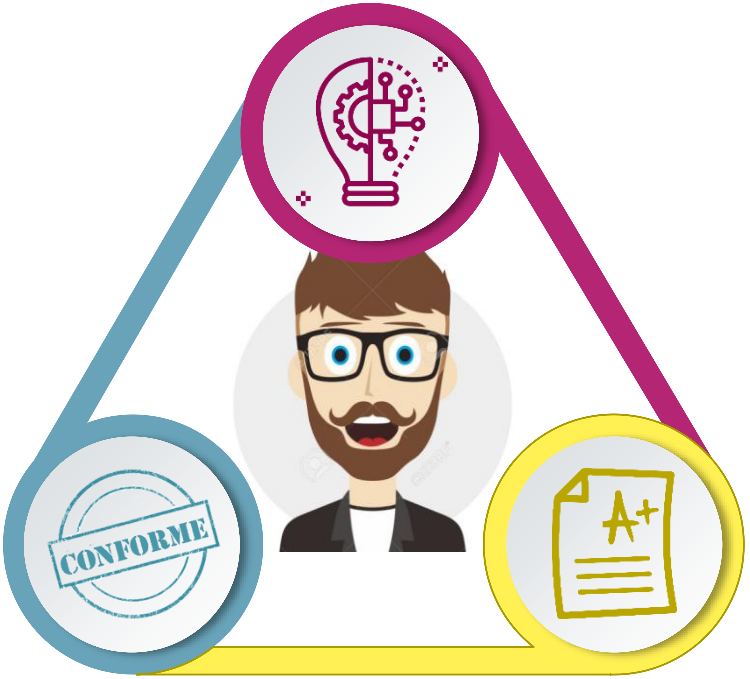
- **Establish a testing and evaluation, verification and validation (TEVV) framework and culture that integrates testing as a continuous part of requirements specification, development, deployment, training, and maintenance and includes run-time monitoring of operational behavior.[24] An AI testing framework should:**
  - Establish a process for writing testable and verifiable AI requirement specifications that characterize realistic operational performance.[25]
  - Provide testing methodologies and metrics that enable evaluation of these requirements—including principles of ethical and responsible AI, trustworthiness, robustness, and adversarial resilience.[26]
  - Define requirements for performance reevaluation related to new usage scenarios and environments, and distribution over time.
  - Encourage incorporation of operational usage workflow and requirements from the defined use case into the testing.
  - Issue data quality standards to appropriately select the composition of training and testing sets.
  - Support the use of common modular cognitive architectures within suitable application domains that expose standard interface points for test harnessing—supporting scalability through increased automation along with federated development and testing.
  - Support a cyclical DevSecOps-based approach, starting on the inside and working outward, with AI components, system integration, human-machine interfaces, and operations (including human-AI and multi-AI interactions).
  - Remain flexible enough to support diverse missions with changing requirements over time.
- **Extend existing and develop new TEVV methods and tools for dealing with complex, stochastic, and non-stationary systems, including the design of experiments, real-time monitoring of states and behaviors, and the analysis of results.** These methods/tools need to account for human-system interactions (HSI) and their impact on system behavior, system-system interactions and their effect on emergent behavior across a group of systems, and adversarial attacks, via both conventional cyberattacks, and nascent perceptual adversarial AI attacks. Risk assurance concepts should be extended beyond simple "stop-light" charts of consequence and likelihood for a risk being realized and leverage tools that support developing assurance cases that present verifiable claims about system behavior and provide reviewable arguments and evidence to support the claims.[27]
- **Make TEVV tools and capabilities readily available across the DoD**, including downloadable and configurable AI TEVV software stacks.[28] In addition, the DoD should ensure tools that support TEVV and reliability and robustness goals are available department-wide including tools for bias detection, explainability, and documentation across the product life cycle (e.g., of data inputs and system outputs).
- **Update existing and create new live, virtual, and constructive test ranges for AI-enabled systems (blending modeling and simulation, augmented reality, and cyber physical system environments).** Upgraded test ranges should include live-virtual-constructive environments, the ability to capture data from testing, and the ability to evaluate data from operations. They should support: 1) The full exploration of potential system states and behaviors over a range of runtimes and fidelity levels;

French Program “Grand Defi” on Trustworthy AI for Industry (Launched In 2019)

How to design, deploy, maintain, certify AI based critical systems ?

… industry strongly involved in programs, especially AI Manifesto members

… Cooperation with French basic research Initiatives, such as Aniti or DataIA, and academic research



Technological pillar 1

DATAS, AI ALGO, SW, SYSTEMS engineering to design, deploy and maintain AI based critical system

Applications conformity assessment Pillar 2

Ensure the right operational exploitation

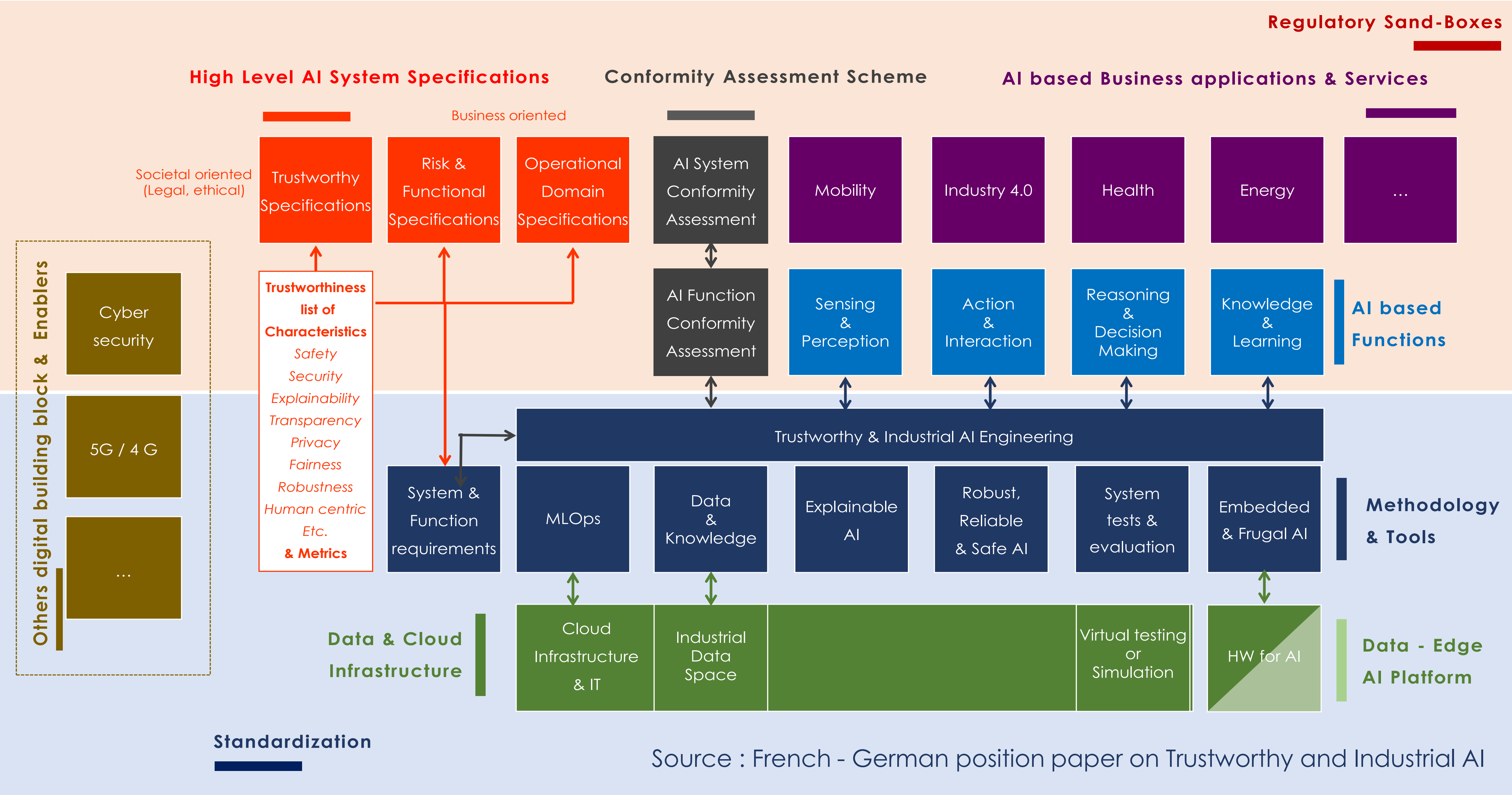
Norms pillar

Norm, standard and regulation environment toward certification

cen CENELEC ETSI
EUROPEAN STANDARDS ORGANIZATIONS

Toward global strategy with coordinated programs and funding (Private, Public)

A standardization and technical framework for Trustworthy and Industrial AI



Regulatory Sand-Boxes

High Level AI System Specifications

Conformity Assessment Scheme

AI based Business applications & Services

Business oriented

Societal oriented (Legal, ethical)

Trustworthy Specifications

Risk & Functional Specifications

Operational Domain Specifications

AI System Conformity Assessment

Mobility

Industry 4.0

Health

Energy

…

Trustworthiness list of Characteristics
Safety
Security
Explainability
Transparency
Privacy
Fairness
Robustness
Human centric
Etc.
& Metrics

AI Function Conformity Assessment

Sensing & Perception

Action & Interaction

Reasoning & Decision Making

Knowledge & Learning

AI based Functions

Others digital building block & Enablers

Cyber security

5G / 4 G

…

Trustworthy & Industrial AI Engineering

System & Function requirements

MLOps

Data & Knowledge

Explainable AI

Robust, Reliable & Safe AI

System tests & evaluation

Embedded & Frugal AI

Methodology & Tools

Data & Cloud Infrastructure

Cloud Infrastructure & IT

Industrial Data Space

Virtual testing or Simulation

HW for AI

Data - Edge AI Platform

Standardization

Source : French - German position paper on Trustworthy and Industrial AI

Technological pillar : Confiance.ai

Confiance.ai

- **Methods, guidelines and interoperable tools for designing, testing, evaluating, validating, deploying and maintaining AI based Critical Systems and Services (safety, business)**

- **4 main applications :** autonomous systems, Supervisory and planning systems, Engineering optimization systems, Optimizing processes and services

- **Technological Road-Map with cross-sectorial expression of needs:**
  - Expression of needs is shared by X-sectors industries (roughly 75% of topics). common project is strategic to share competencies, risks, funding and foster disruptive innovation in this field.
  - List of 20 major industrial issues* (structured by products and services life cycle)
  - List of 22 major technological barriers* (structured by main topics)
  - List of use cases (more than 40) to support R&T

- **Common team** located in Paris-Saclay and Toulouse

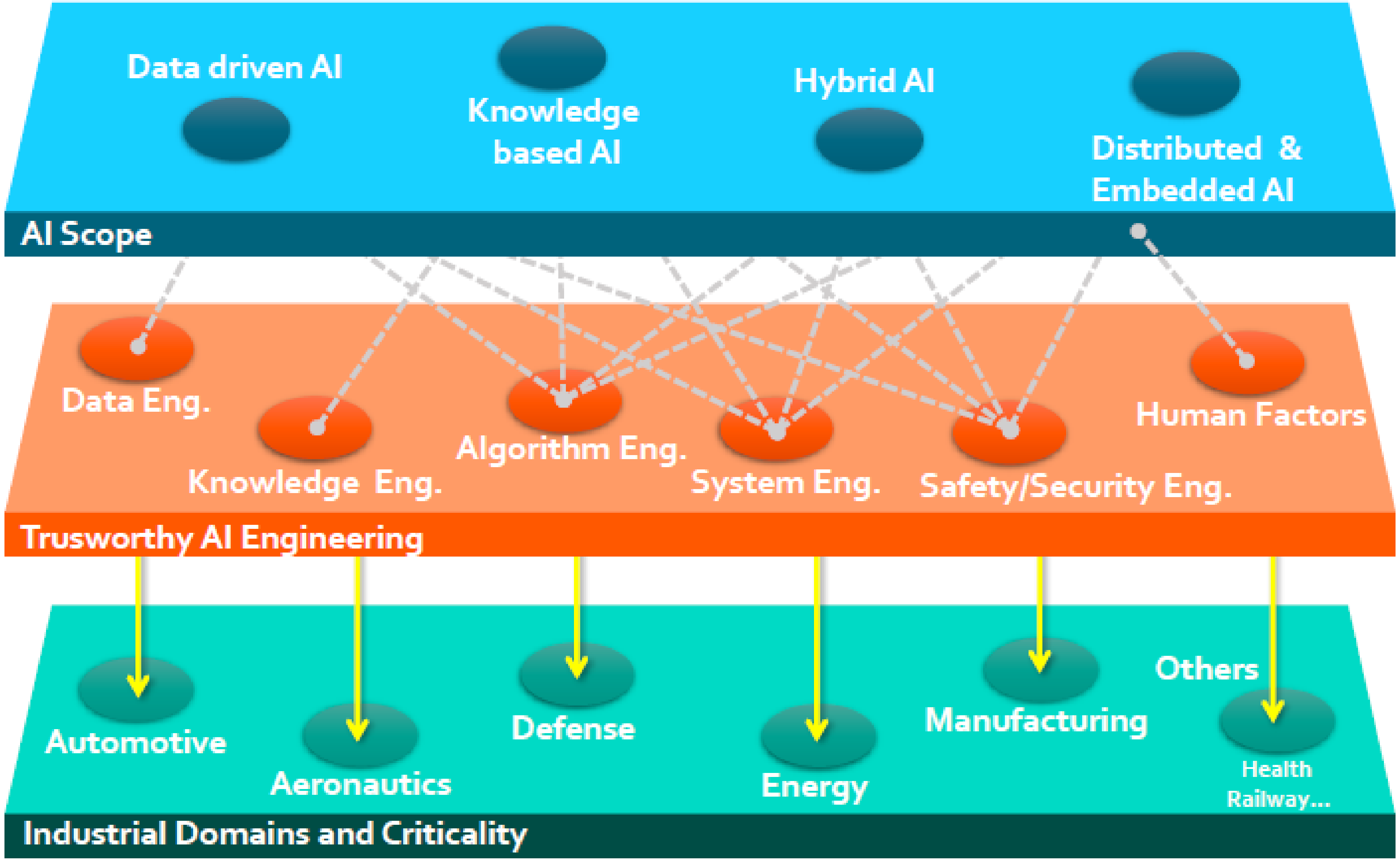
French Program “confiance.ai” : 45 M€ for 4 years duration

Confiance ai

Industrial Partners

AIRBUS
Air Liquide
Atos
eDF
NAVAL GROUP
RENAULT
SAFRAN
sopra steria
THALES Building a future we can all trust
Valeo

Feb, 8th 2021



Academic & RTO partners

cea
Inria
IRT SAINT EXUPÉRY
SystemX

… Cooperation also with French basic research Initiatives and ecosystems

Trustworthy

Trustworthiness list of Characteristics

Safety

Security

Explainability

Transparency

Privacy

Fairness

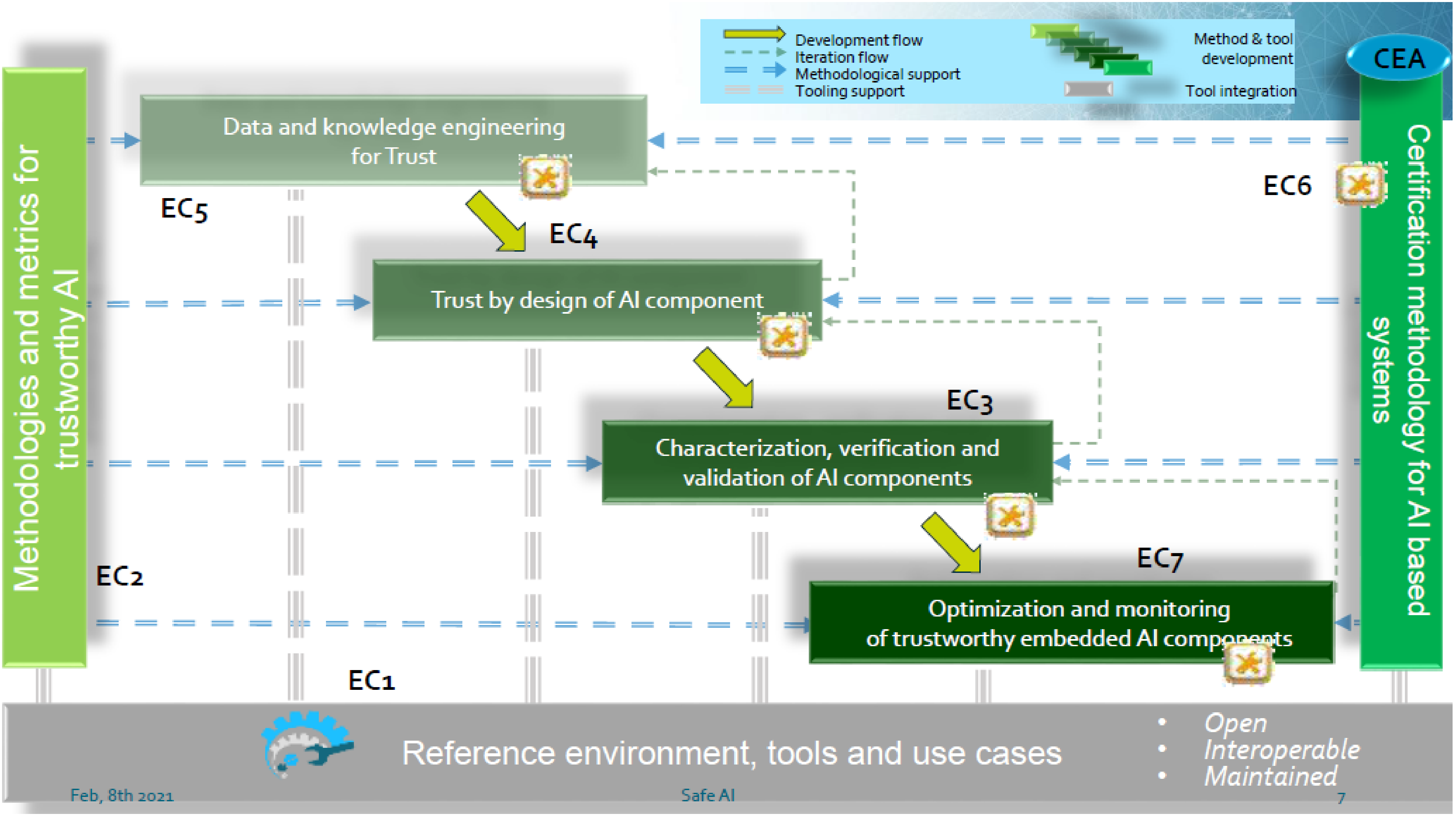
Robustness

Human centric

Etc.

& Metrics

Human Oversight



Development flow

Iteration flow

Methodological support

Tooling support

Method & tool development

Tool integration

CEA

Methodologies and metrics for trustworthy AI

Data and knowledge engineering for Trust

Trust by design of AI component

Characterization, verification and validation of AI components

Optimization and monitoring of trustworthy embedded AI components

Certification methodology for AI based systems

Reference environment, tools and use cases

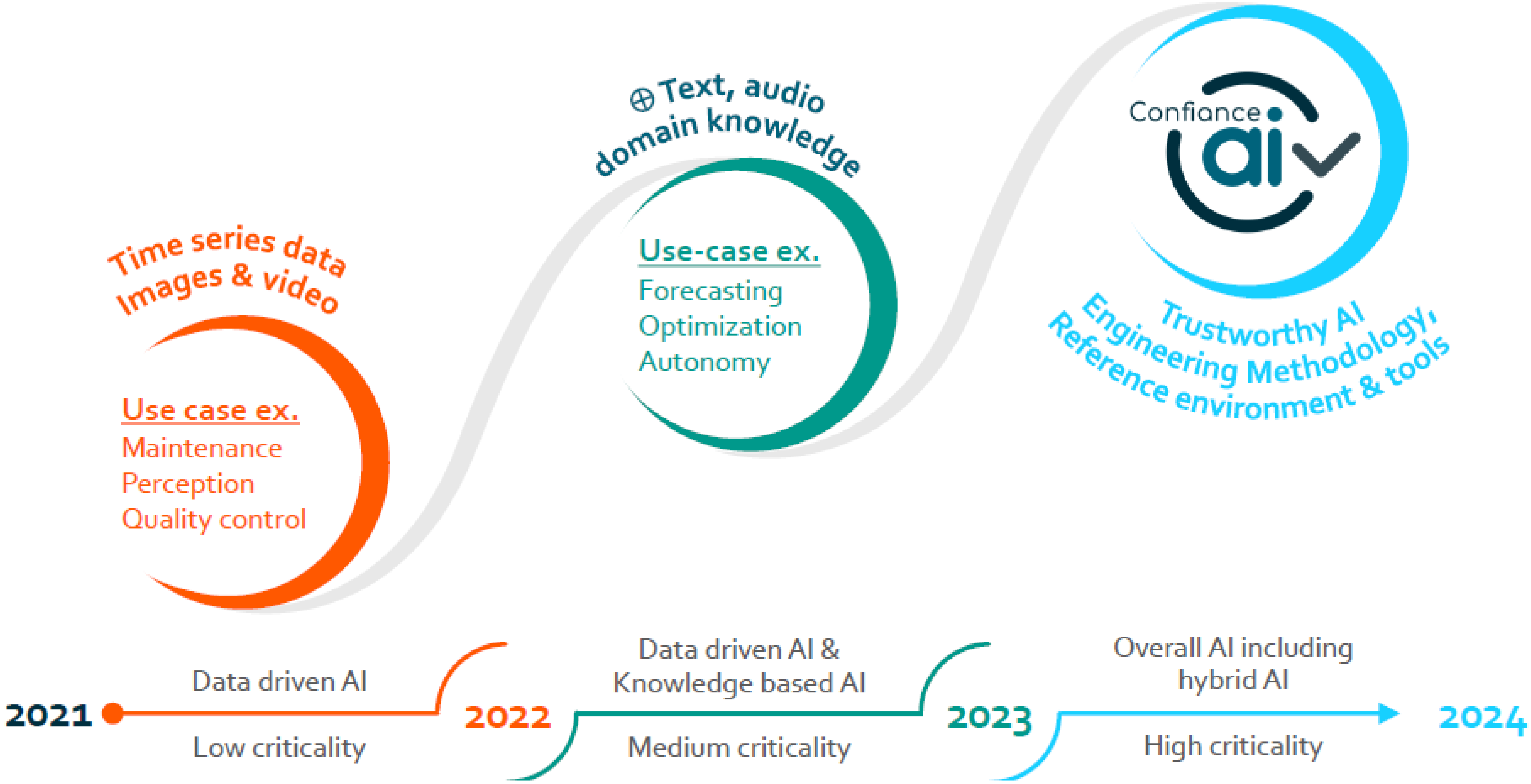
- *Open*
- *Interoperable*
- *Maintained*

Feb, 8th 2021

Safe AI

7

An incremental roadmap supported by various use cases



Time series data
Images & video

Use case ex.
Maintenance
Perception
Quality control

⊕ Text, audio
domain knowledge

Use-case ex.
Forecasting
Optimization
Autonomy

Confiance ai

Trustworthy AI
Engineering Methodology,
Reference environment & tools

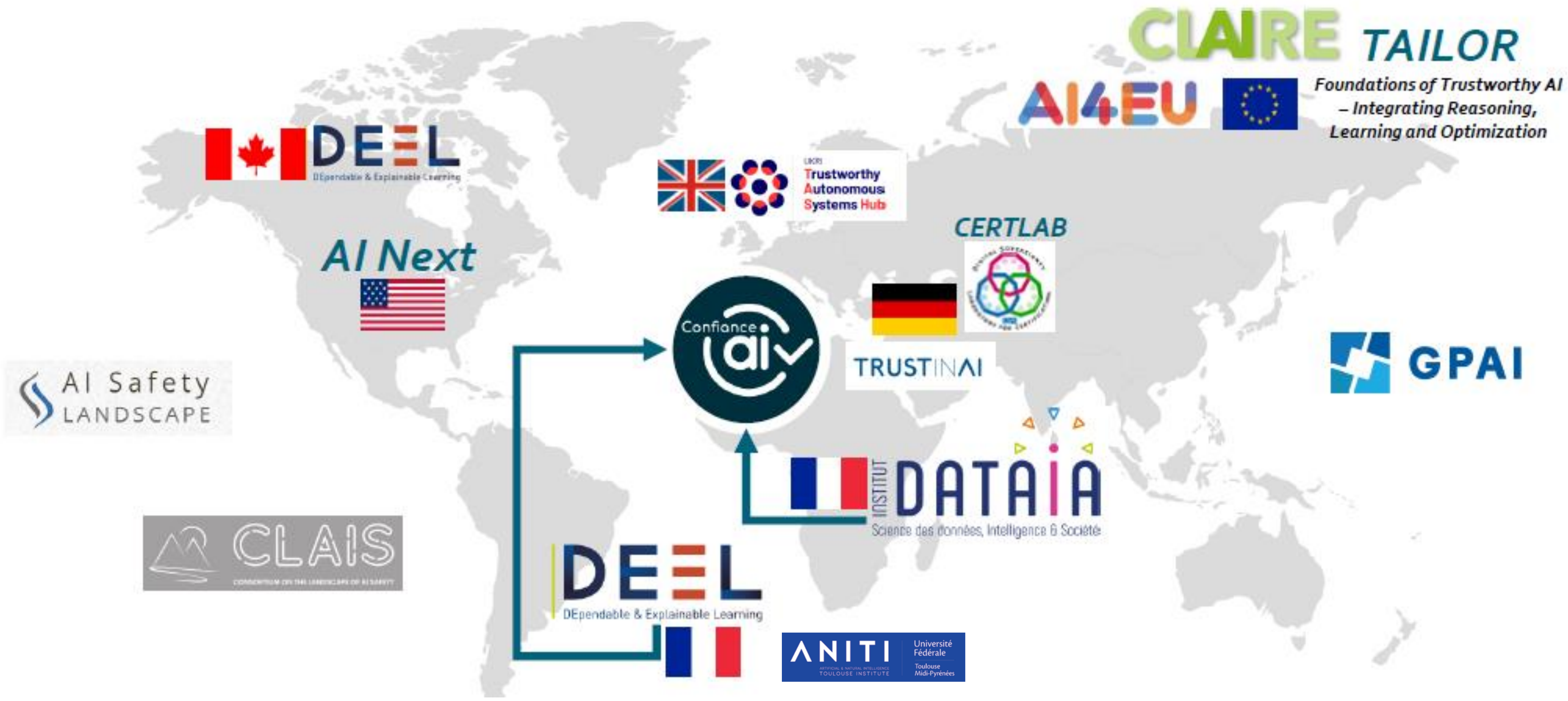
2021 — Data driven AI — Low criticality

2022 — Data driven AI & Knowledge based AI — Medium criticality

2023 — Overall AI including hybrid AI — High criticality

2024

A non exhaustive view of the Trustworthy AI Ecosystem


CLAIRE
TAILOR
AI4EU
Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
DEEL
DEpendable & Explainable Learning
UKRI
Trustworthy Autonomous Systems Hub
CERTLAB
AI Next
Confiance ai
TRUSTINAI
GPAI
AI Safety LANDSCAPE
INSTITUT DATAIA
Science des données, Intelligence & Société
CLAIS
CONSORTIUM ON THE LANDSCAPE OF AI SAFETY
DEEL
DEpendable & Explainable Learning
ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE
Université Fédérale Toulouse Midi-Pyrénées

Applications conformity assessment Pillar:

PRISSMA, a 1st platform on New Autonomous Mobility

Evaluation, homologation and certification Pillar

- **Toward evaluation, homologation and certification:** complementary programs to ensure "end to end" approaches and specificities for AI applications (mobility, …) in cooperation with academic, industrial ecosystems and regulation autorithies

- **Applications oriented Programs (first for Autonomous mobility, on going work on proposal, kick-off beginning of Q4 2020 – 3 years) in accordance to sectorial road map** to face all the challenges and infrastructures required to homologate and certify products or services

- **Common team from Academic laboratories, RTO, Industries and regulation Authorities**

Toward homologation for New Autonomous Mobility

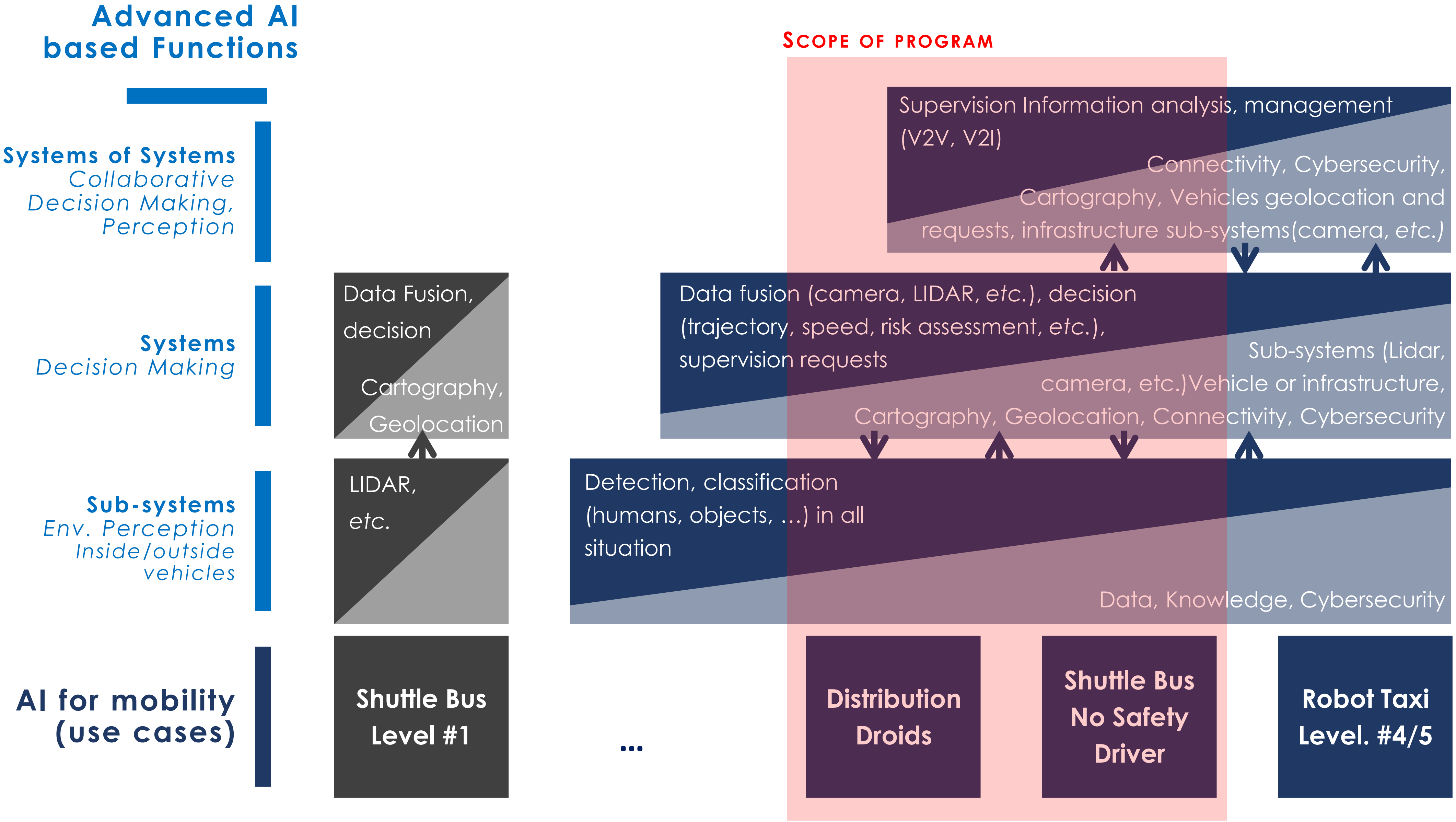
- **Covering "end to end" approach for homologation of systems and integration** with a focus on perception and localization and a focus on collective public transportation (such as bus …)

- **Partners :**
  - Leader
  - Industry and technological Providers
  - Academic

- **Evaluation Road-Map** with an articulation with on-going initiatives : 1) Datas and scenario, 2) simulation and evaluation, 3) tests, 4) cybersecurity (focus on AI issues)

- **Willing to set up a transversal use case with "confiance.ai" technological pillar**

#AI for mobility: Program Today in terms of Applications



Advanced AI based Functions

Systems of Systems
Collaborative Decision Making, Perception

Systems
Decision Making

Sub-systems
Env. Perception Inside/outside vehicles

AI for mobility (use cases)

SCOPE OF PROGRAM

Supervision Information analysis, management (V2V, V2I)

Connectivity, Cybersecurity, Cartography, Vehicles geolocation and requests, infrastructure sub-systems(camera, *etc.*)

Data Fusion, decision

Cartography, Geolocation

Data fusion (camera, LIDAR, *etc.*), decision (trajectory, speed, risk assessment, *etc.*), supervision requests

Sub-systems (Lidar, camera, etc.)Vehicle or infrastructure, Cartography, Geolocation, Connectivity, Cybersecurity

LIDAR, *etc.*

Detection, classification (humans, objects, …) in all situation

Data, Knowledge, Cybersecurity

Shuttle Bus Level #1

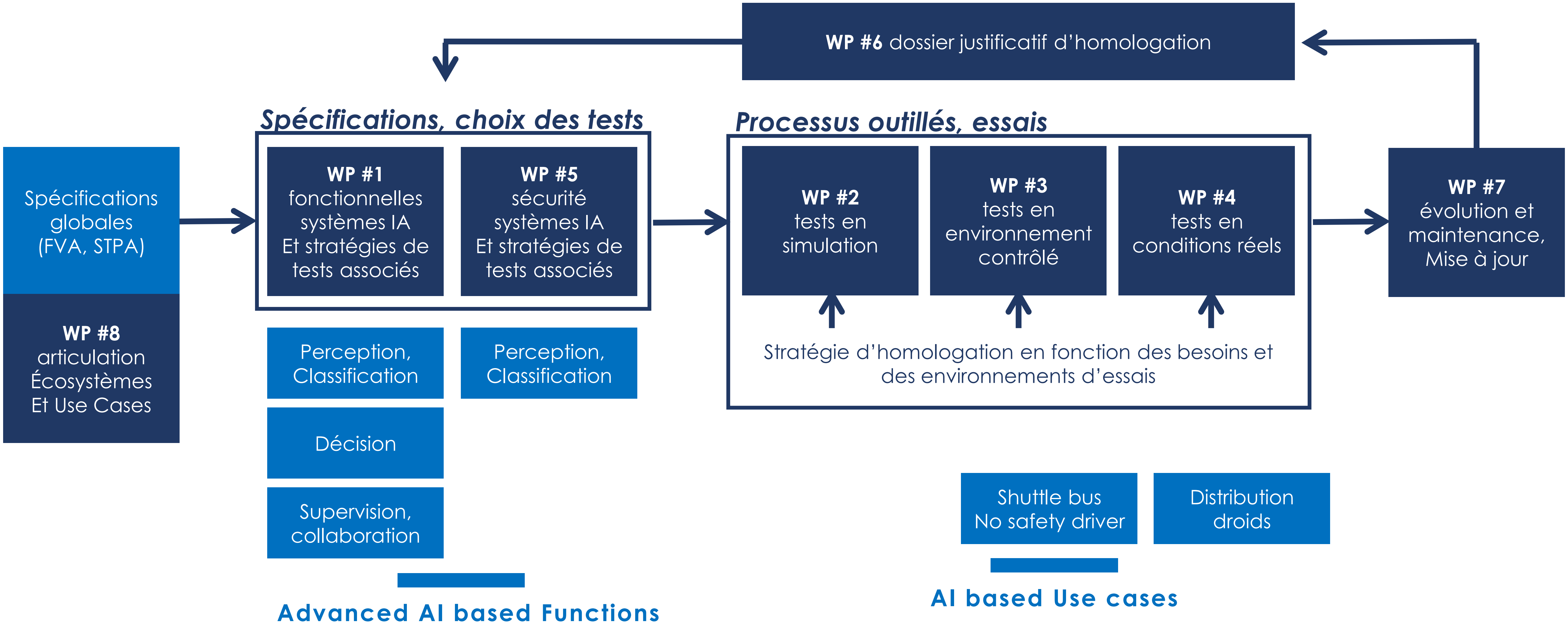
…

Distribution Droids

Shuttle Bus No Safety Driver

Robot Taxi Level. #4/5

#PRISSMA : Project Descrption



WP #6 dossier justificatif d'homologation

Spécifications, choix des tests

Processus outillés, essais

Spécifications globales (FVA, STPA)

WP #8 articulation Écosystèmes Et Use Cases

WP #1 fonctionnelles systèmes IA Et stratégies de tests associés

WP #5 sécurité systèmes IA Et stratégies de tests associés

WP #2 tests en simulation

WP #3 tests en environnement contrôlé

WP #4 tests en conditions réels

WP #7 évolution et maintenance, Mise à jour

Perception, Classification

Perception, Classification

Décision

Supervision, collaboration

Stratégie d'homologation en fonction des besoins et des environnements d'essais

Shuttle bus No safety driver

Distribution droids

Advanced AI based Functions

AI based Use cases

Standardization Pillar:

Standardization pillar (Pilot: AFNOR)

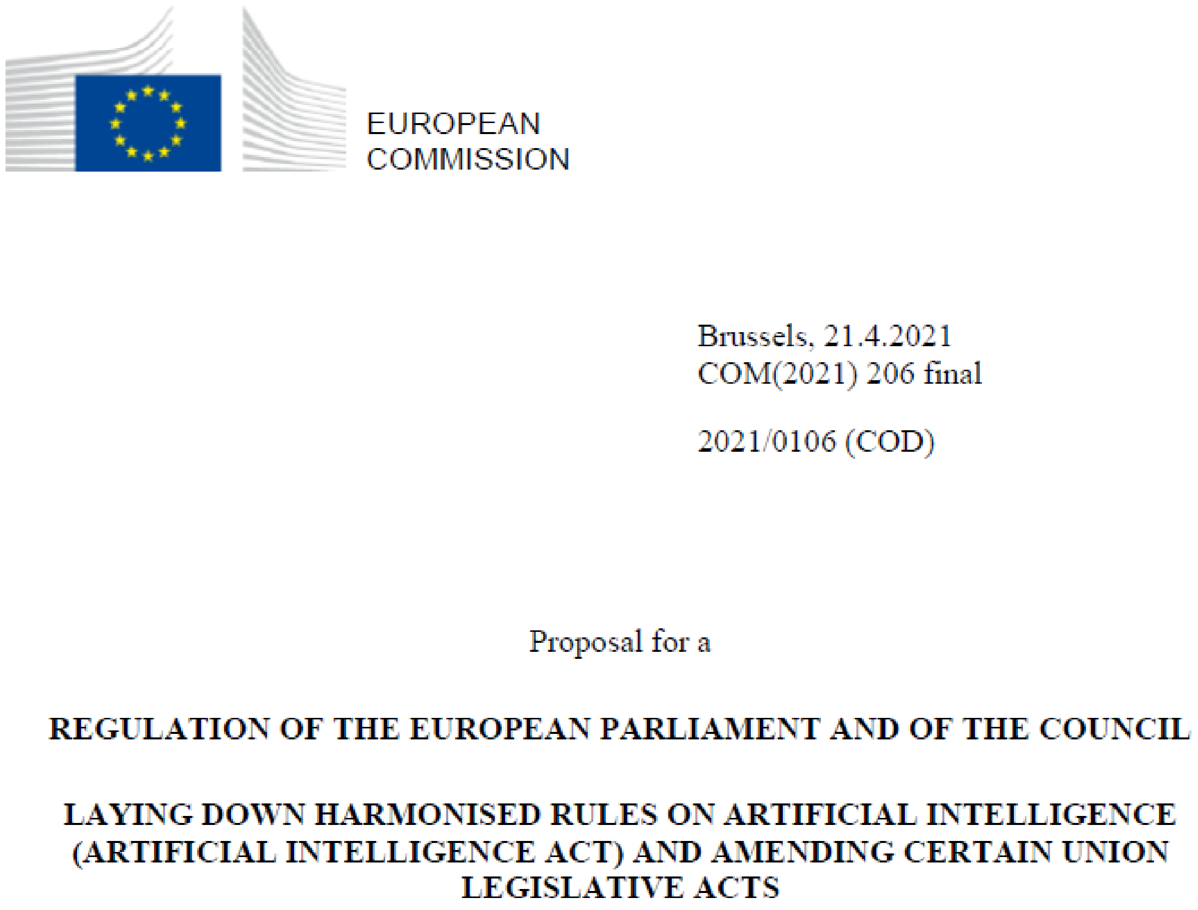
- **Setup a standardization road map (and priorities):** common work between R&T Team (Pillar 1) and Standardization Team (Pillar 3); use R&T road map already available to build a coordinated road map on standardization; in parallel consolidation with ecosystems through request for information, …

- **Set up international cooperation to promote common vision on standardization for AI at EU and international level**

- **Ecosystems & Network:** setup an information sharing plateform with national ecosystems, support start-up, fostering initiatives for Startups, SMEs and academic laboratories in order to involve them on standardization work, …

- **Support standardization works** in accordance with strategy and road-map

Trustworthy and Industrial AI: a proposed regulation in Europe



EUROPEAN
COMMISSION

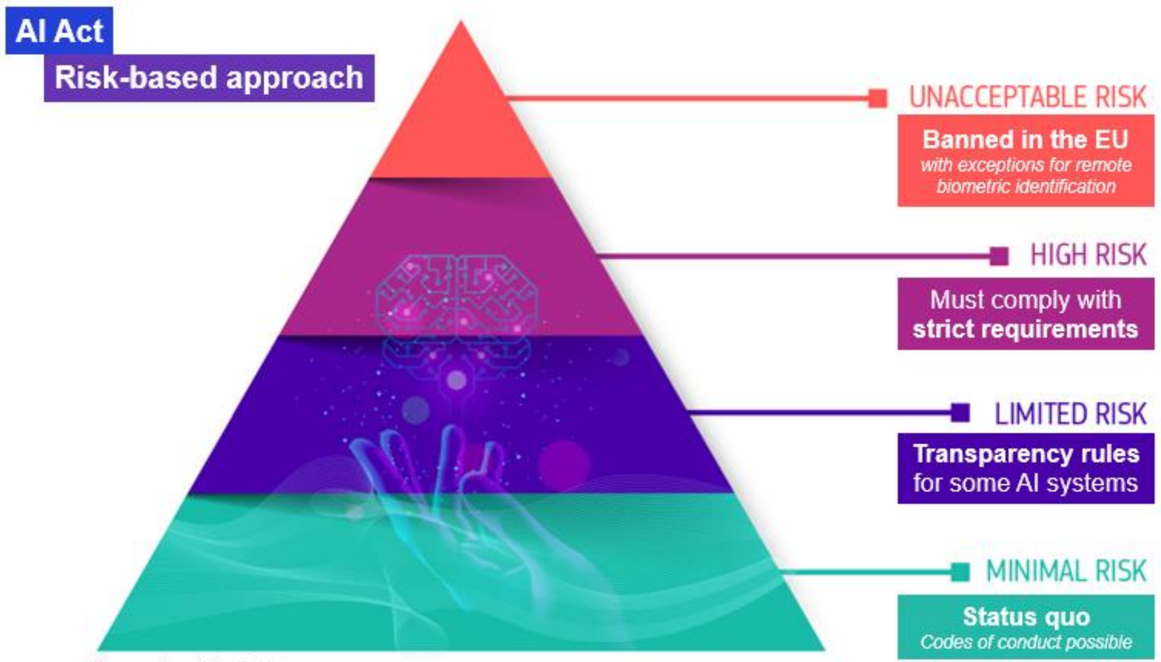
Brussels, 21.4.2021
COM(2021) 206 final

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



AI Act
Risk-based approach

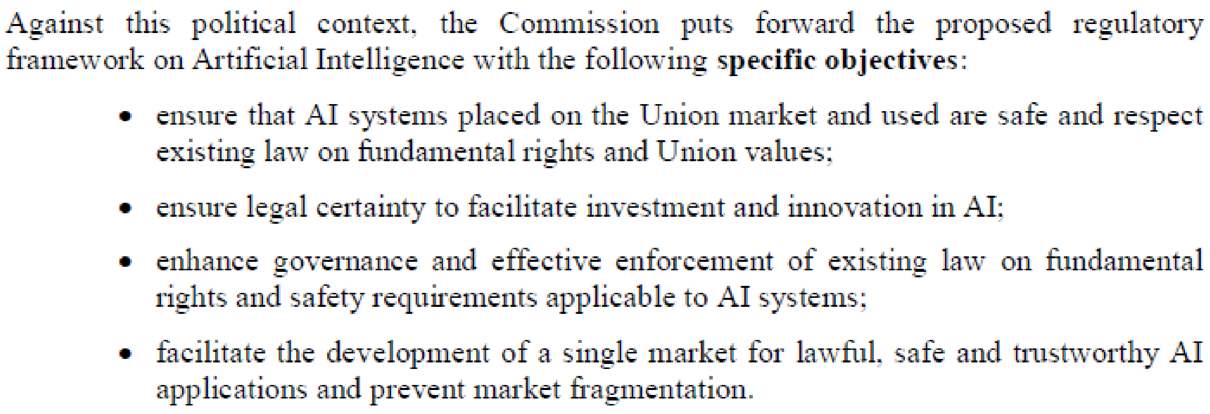
UNACCEPTABLE RISK
Banned in the EU
with exceptions for remote biometric identification

HIGH RISK
Must comply with
strict requirements

LIMITED RISK
Transparency rules
for some AI systems

MINIMAL RISK
Status quo
Codes of conduct possible

European Commission, 2021



Against this political context, the Commission puts forward the proposed regulatory framework on Artificial Intelligence with the following **specific objectives**:

- ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values;
- ensure legal certainty to facilitate investment and innovation in AI;
- enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
- facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

CONSEIL DE L'INNOVATION

Thank You for your Attention