An engineering platform for trustworthy AI

Francesca Rossi's Keynote@AAAI

AI's capabilities, limitations, ethics issues

Capabilities

Data-driven approaches

- Learning from data
- GANs
- Transformers
- Computer vision
- Natural Language Interpretation and Generation

Rule-based, symbolic, and logical approaches

- Explicit procedure to solve a problem
- Reasoning, planning, scheduling, optimization for complex problems

Limitations

- Generalizability and Abstraction
- Robustness and Resiliency
- Contextual awareness
- Multi-agent cooperation
- Resource efficiency (data, energy, computing power)
- Adaptability
- Causality

Ethics issues

- Trust
  - Fairness, robustness, explainability, causality, transparency
- Data governance, privacy, liability, human agency
- Impact on work and society
- AI autonomy vs augmented intelligence
- Real vs online life, metrics of success/goals

IBM

Michael Littman's Keynote@AAAI

Conclusions

The AI has made remarkable progress.

- Leaps forward in language- and image-processing tasks.
- Applications like healthcare and self driving cars.
- Still far short of the field's founding aspirations

Inflection point: Urgent to consider downsides.

- Automating decisions at scale carries risks.
- People misled, discriminated against, physically harmed.
- Historical data can exacerbate biases/inequalities.
- Social sciences part of broader AI conversation.
- Ongoing engagement essential.

Governments need to:

- Recognize the importance of AI, move quickly.
- Keep people informed, support broad education.

AI research community needs to:

- Learn to share findings in informative/actionable ways.
- Avoid hype and discuss dangers and benefits.
- Incorporate AI into community-wide systems.
- Make goal to empower, not devalue, people.

https://ai100.stanford.edu/2021-report/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence

5 walls of AI

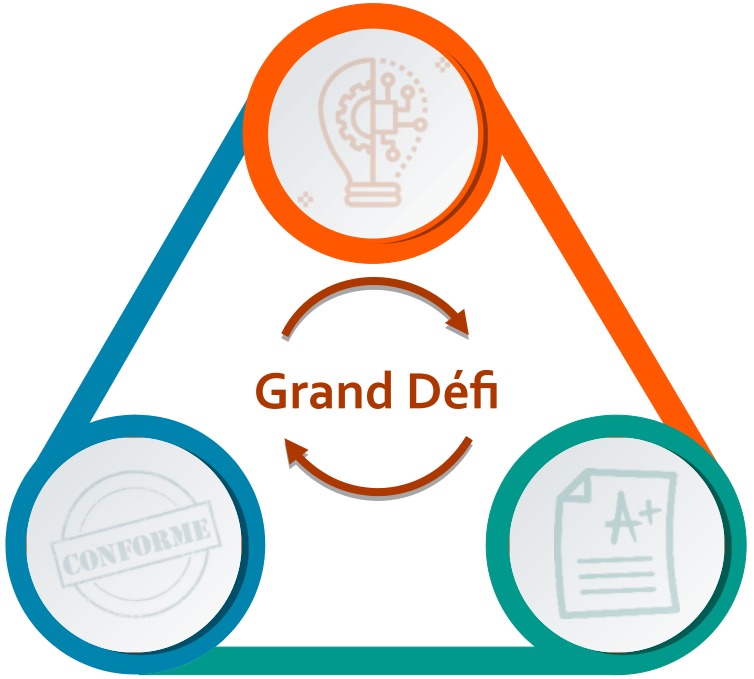
Trust
Energy
Security
Interaction
Inhumanity

Trustworthy & certification AI:
from data/algo to AI SW & Systems Engineering

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

Technological pillar

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



Norms pillar

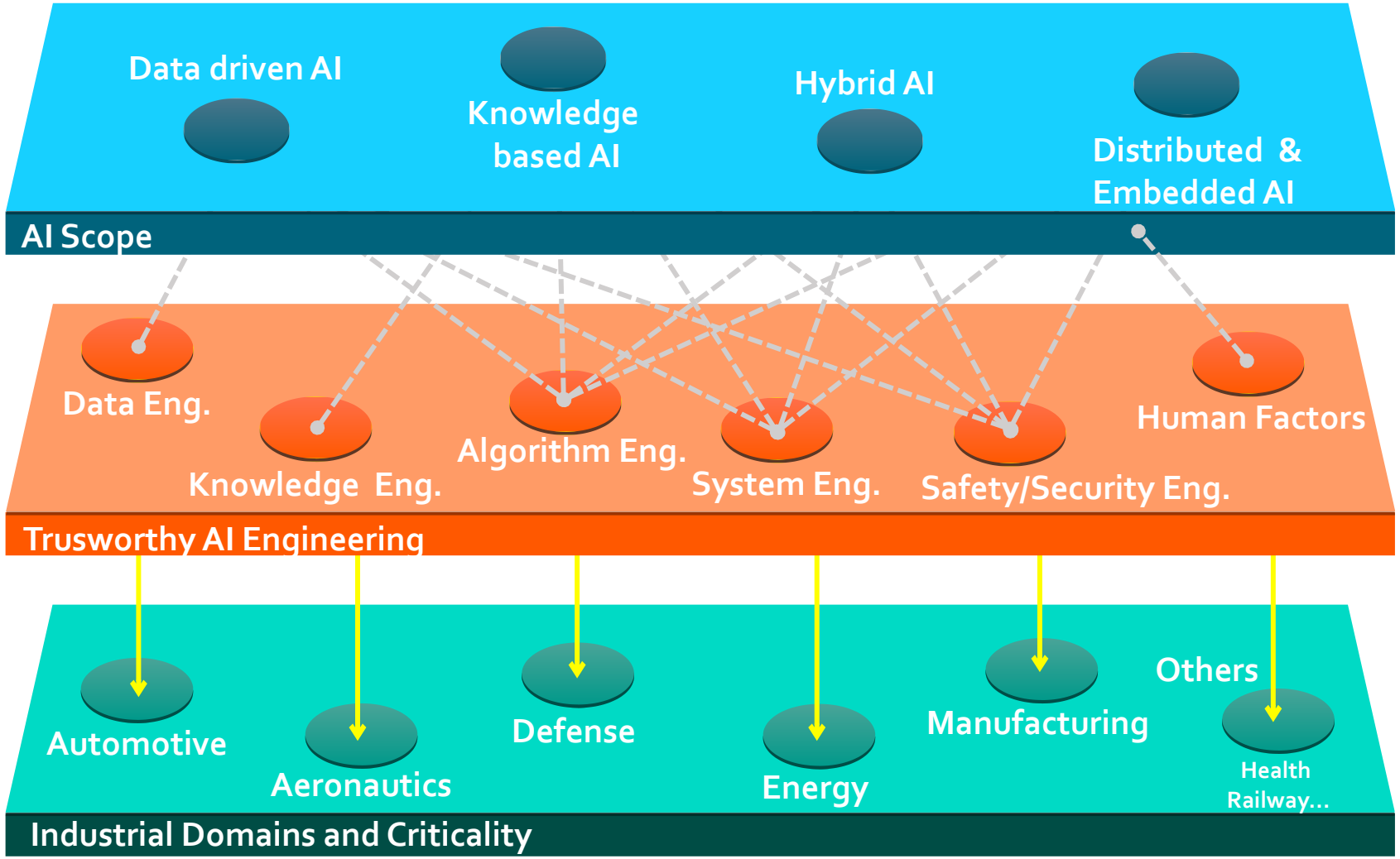
Norm, standard and regulation environment toward certification

Applications Evaluation Pillar

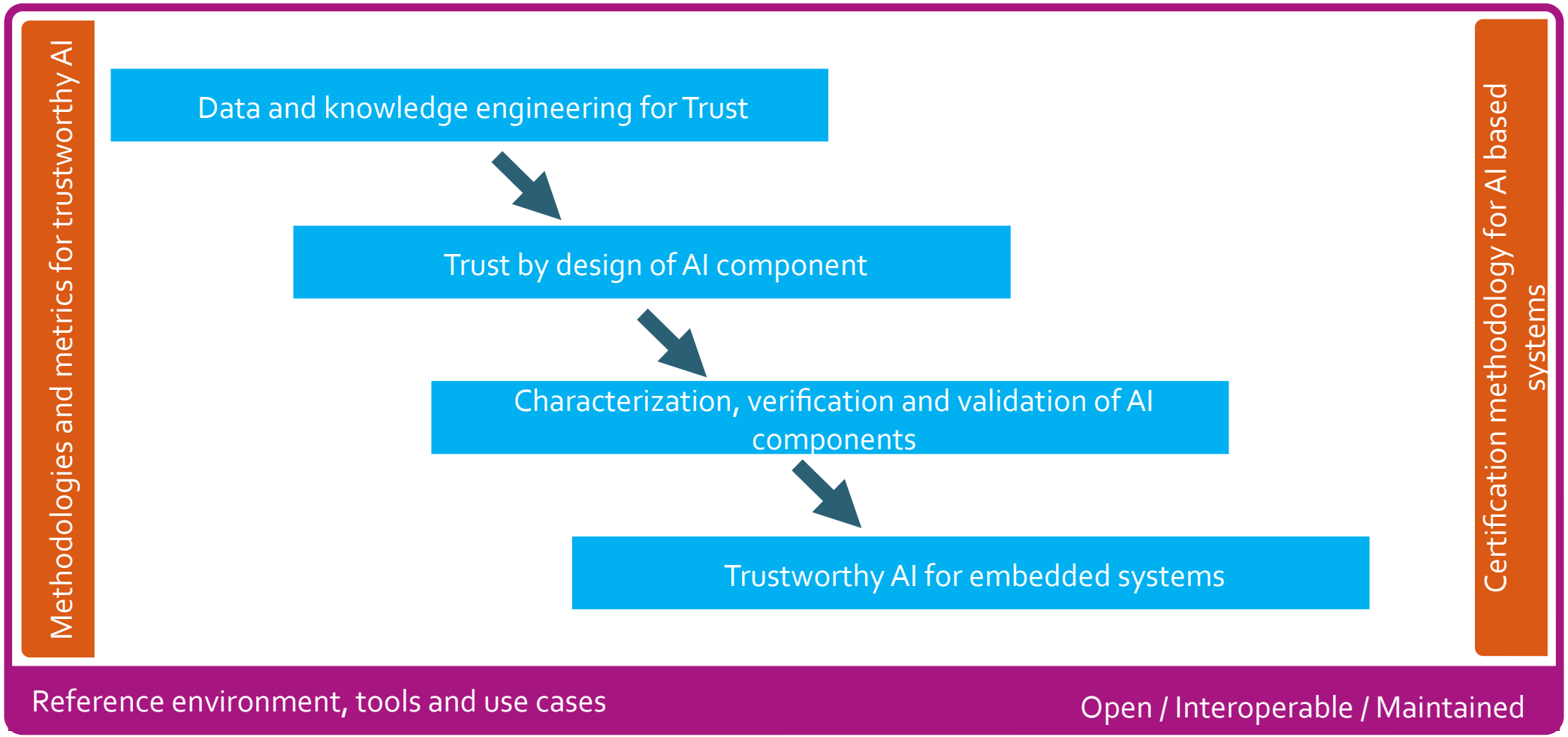
Ensure the right operational exploitation

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

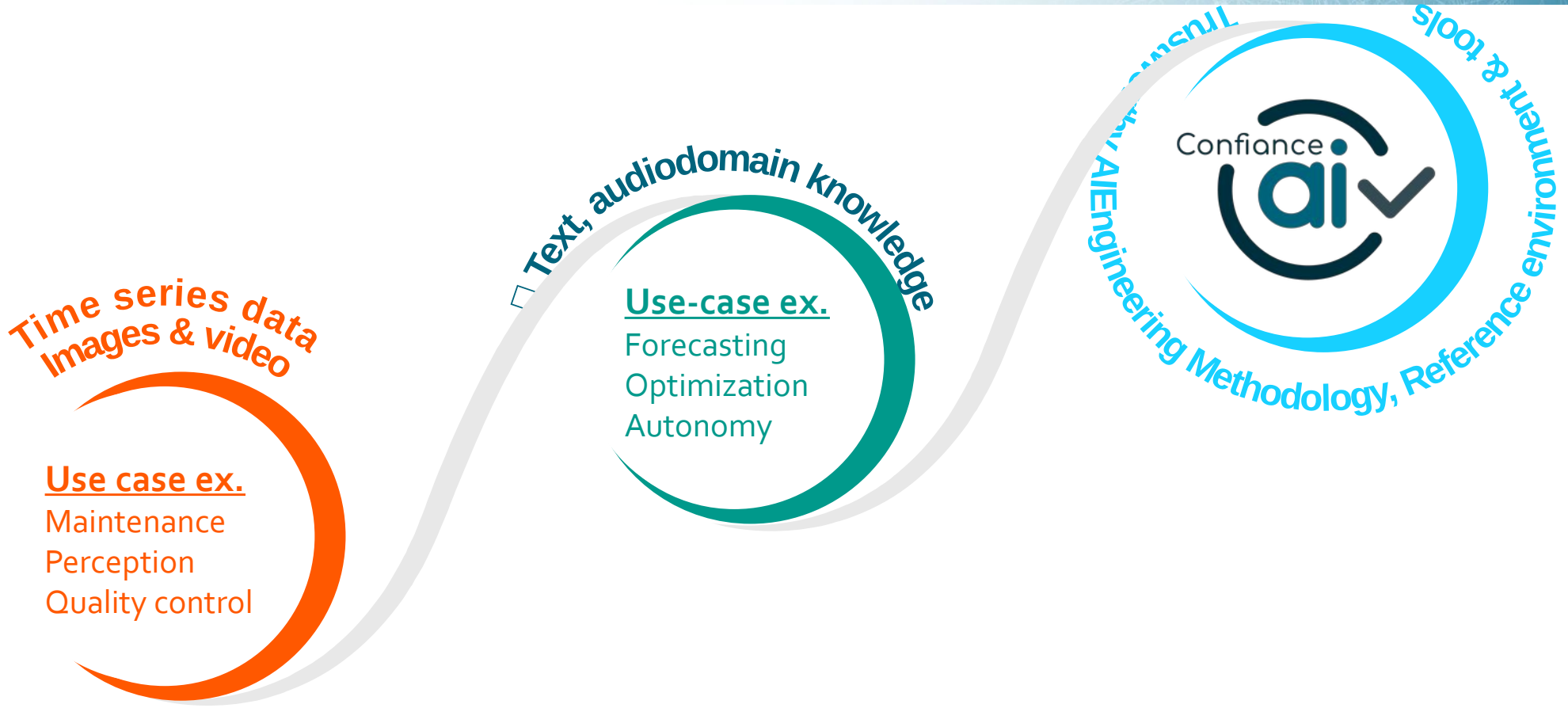
Confiance.ai program
(Global budget: 45M€, Duration: 4 years)



Program architecture



An incremental roadmap validated by various use-cases



2021 — Data driven AI — Low criticality

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

2023 — Overall AI including hybrid AI — High criticality

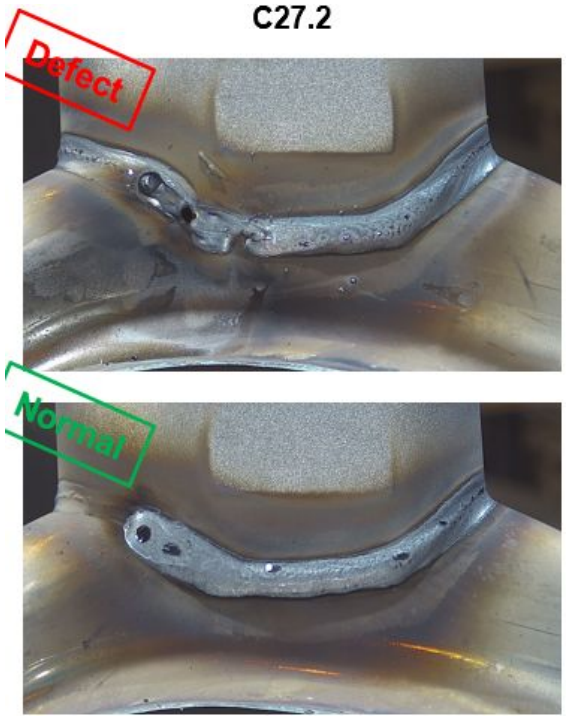
2024

Scientific challenges (overview)

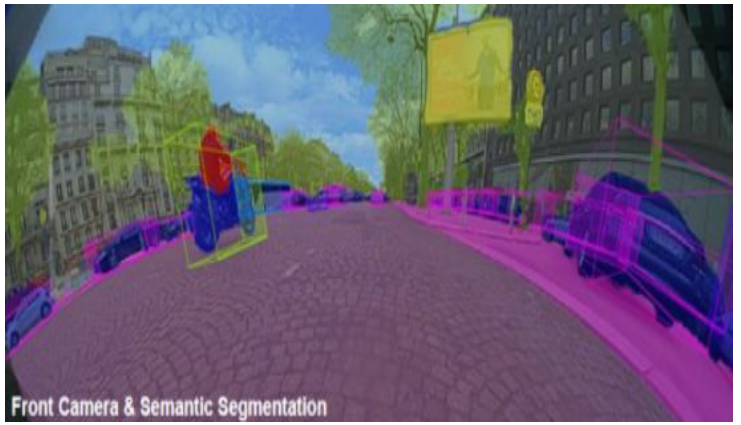
- Trustworthy system engineering with AI components
  - Qualify AI-based components and systems
  - Building AI components with controlled trust
  - Embeddability of trustworthy AI
- Trust and learning data
  - Qualify data/knowledge for learning
  - Building data/knowledge to increase confidence in learning
- Trust and human interaction
  - Trust-generating interaction between users and AI-based system
  - Trust-generating interaction between designer/certifiers and AI-based systems

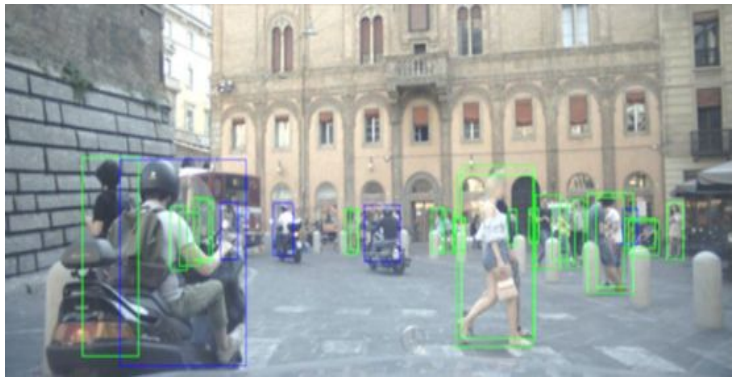
2 examples of use cases

- Renault : welding defaults detection



Valeo : urban scene interpretation





Confiance.ai Ecosystem

afnor

A larger view of the Thrustworthy AI Ecosystem

Workshop Agenda

10:25	Can we measure trust? **Agnès Delaborde (LNE, Laboratoire National de Métrologie et d'Essais)**
10:40	Justifying trust in AI/ML system using Engineering Models and Assurance Cases, **Eric Jenn (IRT Saint-Exupéry), Morayo Adedjouma (CEA List)**
10:55	How to trust your data: challenges to increase confidence in the data lifecycle of critical systems, **Flora Dellinger (Valeo)**, Camille Dupont (CEA), Xavier Perrotton (Valeo)
11:05	Questions
11:10	Building labelled datasets for real-world tasks with active learning, **Thomas Dalgaty (CEA**), **Fritz Poka Toukam (CEA)**, Oriane Simeoni (Valeo), Spyros Gidaris (Valeo), Hedi Ben-Younes (Valeo), Nicolas Granger (CEA), Camille Dupont (CEA)
11:25	An information geometry approach to Randomised smoothing, Hatem Hajri (IRT SystemX), **Pol Labarbarie (IRT SystemX)**
11:35	Uncertainty Quantification for Customers Demand Forecasting, **Marc Nabhan (Air Liquide)**

www.confiance.ai
contact@irt-systemx.fr