THALES
Building a future we can all trust

Trustworthy AI for defense : engineering challenges

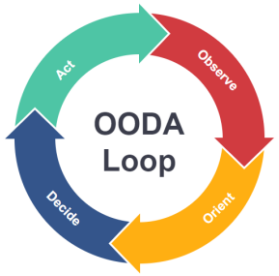
Patricia Besson
Thales/cortAIxLabs France

www.thalesgroup.com

Artificial Intelligence for defense : new capabilities in a constrained environment

> Artificial intelligence ⇔ artificial capabilities

- **Acceleration** and **automatisation** of the **OODA loop** (Observe, Orient, Decide, Act)



- **Autonomous** or **semi-autonomous** systems

> AI in defense needs to operate in an embedded AND safety critical world. Systems should work :

All the time, in the time –
usual product lifespan 10-20 years

Using little/specific data, potentially with controlled access –
Sovereign & classified data requiring specific access rights

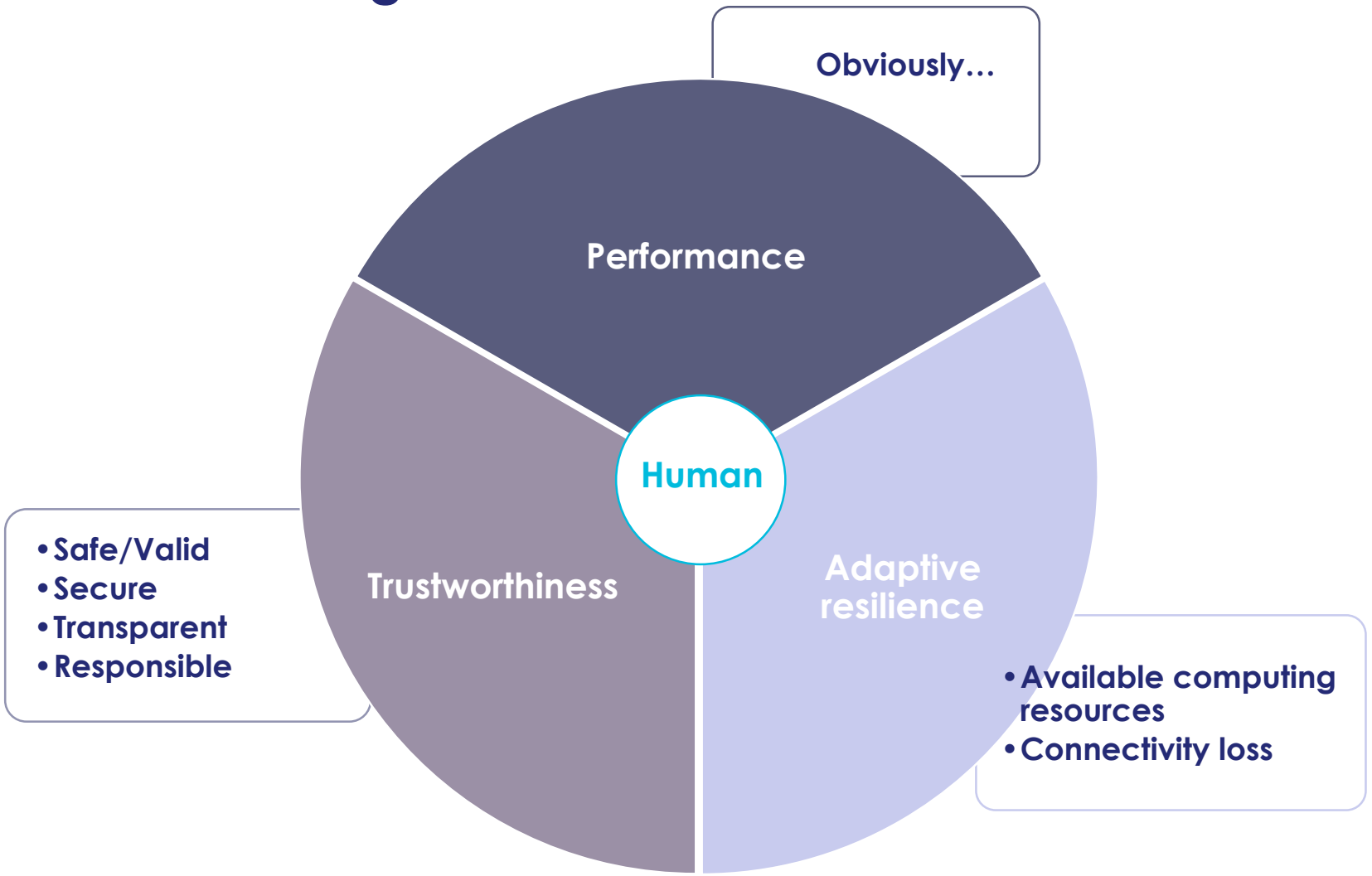
In changing conditions –
including extreme/rare environnement: °C, radiations, vibrations…

Constrained in size, weight, power and cost –
Field & Edge deployment

cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / **Template:** trtp version 9.0.0 based on **Template: 87211168-DOC-GRP-EN-007**

{OPEN}

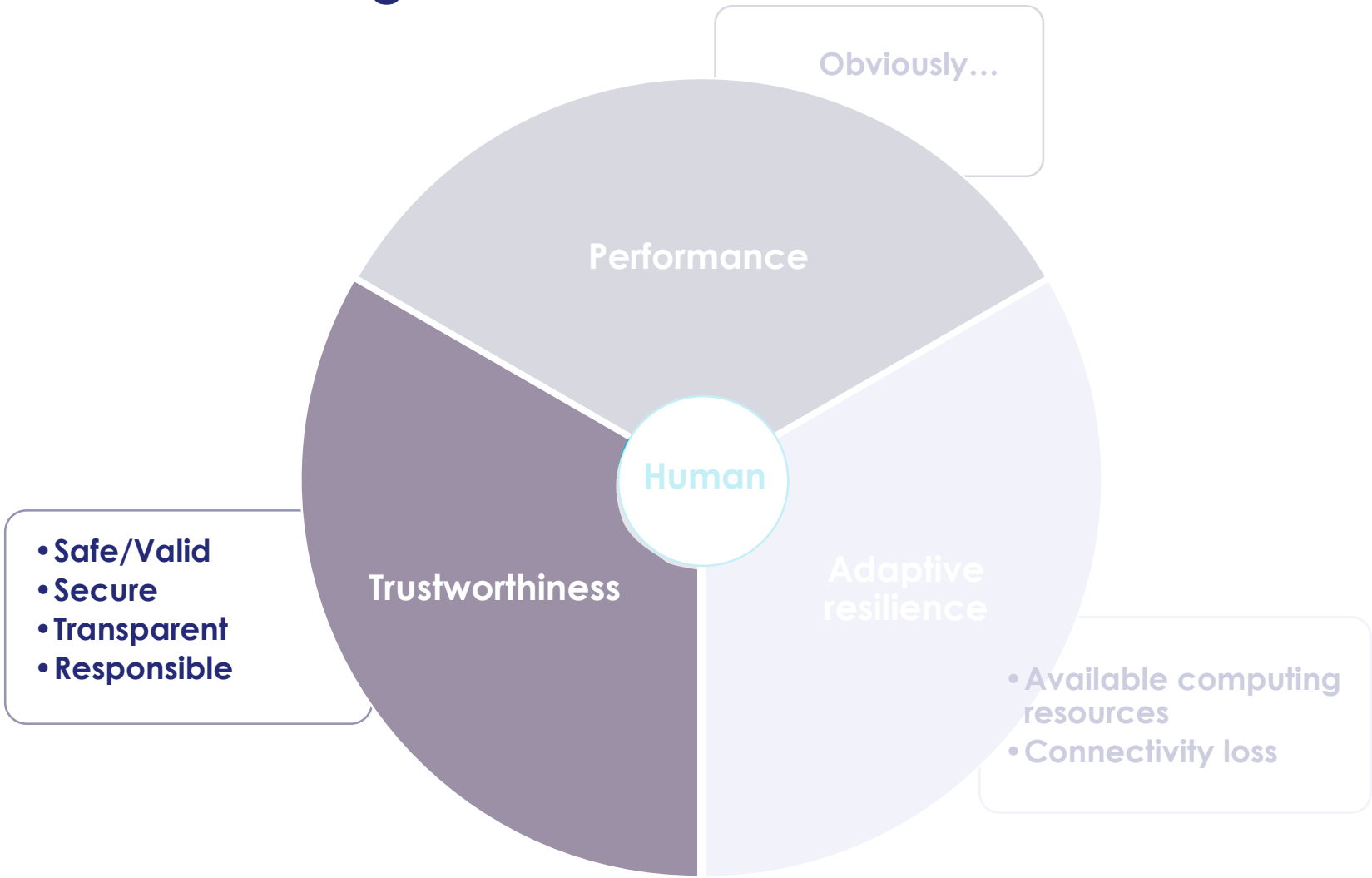
AI in Products : Technological Levers for Critical Defense and Security Systems



cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / **Template:** trtp version 9.0.0 based on **Template: 87211168-DOC-GRP-EN-007**

{OPEN}

AI in Products : Technological Levers for Critical Defense and Security Systems



cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / **Template:** trtp version 9.0.0 based on **Template: 87211168-DOC-GRP-EN-007**

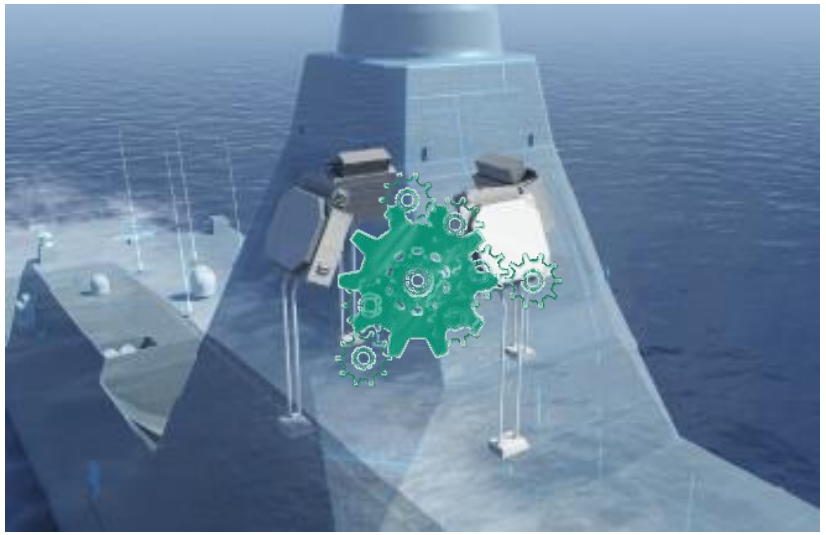
{OPEN}

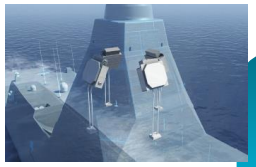
Trustworthy AI

VALID	SECURE	TRANSPARENT	RESPONSIBLE
Doing all and only what it is meant to do	Resilient and robust vs adversarial conditions	Explainable, understable, providing context justifications	Compliant with regulation, legal, ethical frameworks

{OPEN}

Trustworthy AI : it all starts with the system





SYSTEM

IVVQ viewpoint

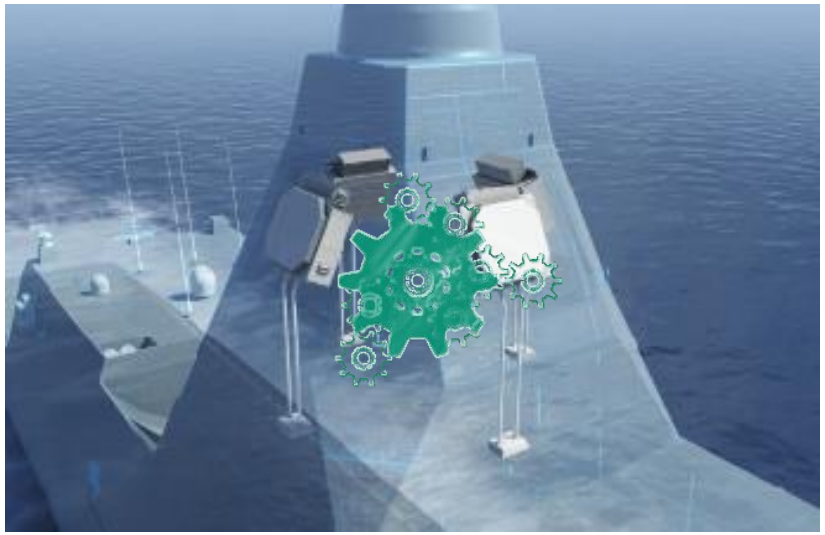
AI MODEL

Data viewpoint

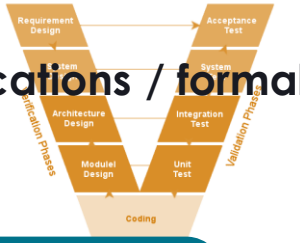
{OPEN}

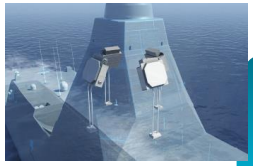
icon ©FreePik

Trustworthy AI : it all starts with the system



Specifications / formalizations





SYSTEM

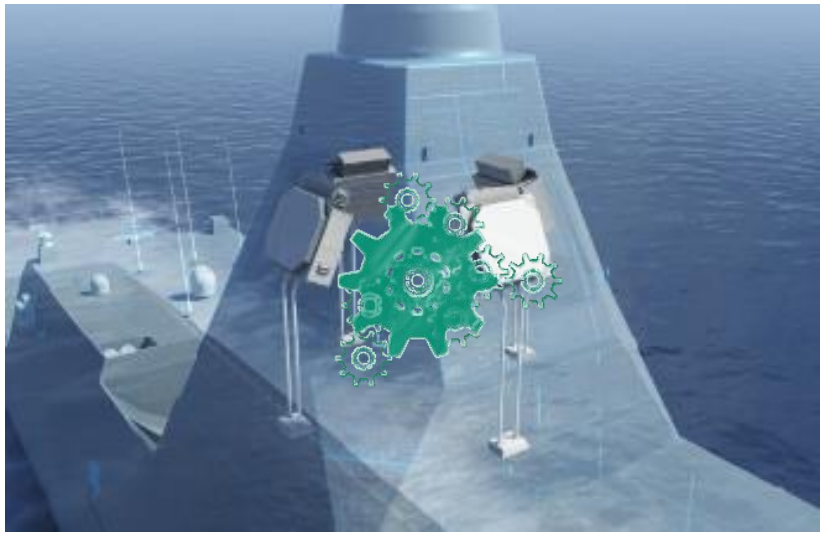
IVVQ viewpoint

AI MODEL

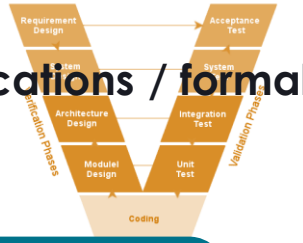
Data viewpoint

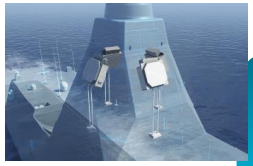
icon ©FreePik

Trustworthy AI : it all starts with the system



Specifications / formalizations





SYSTEM

IVVQ viewpoint

How to reconcile both viewpoints?

AI MODEL

Data viewpoint

icon ©FreePik

Evaluating and testing AI models

> Valid model ⇔ doing all and only what it is meant to do

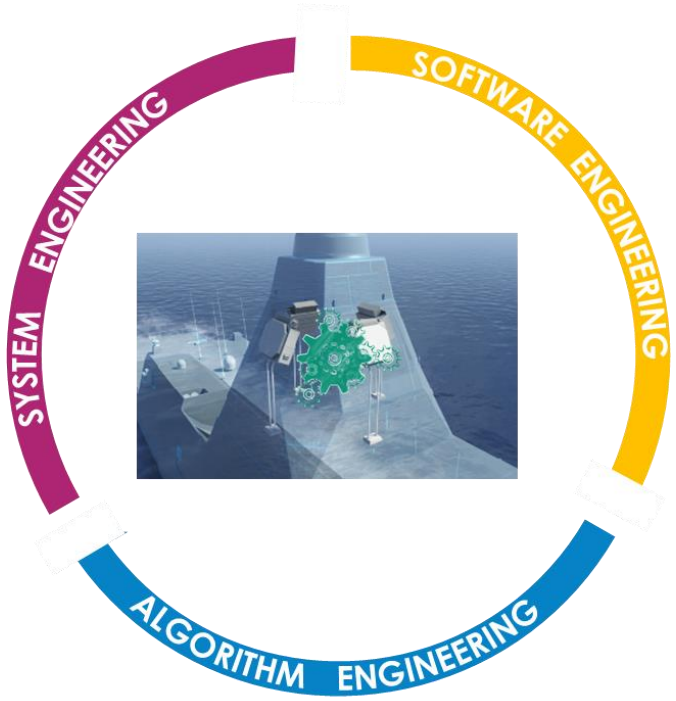
> Approaches to evaluating and testing AI models :

- Against a **pre-defined evaluation data base** ➔ Constituted how and by whom? Representativity?
- **By usage** : in-situ user feedback ➔ only possible for some systems (decision-aiding systems)
- Typical **system engineering approach** : testing and documenting explicit properties and behaviours of the model ➔ mandatory for certification

No universally accepted testing standards to date ➔ inconsistencies in evaluation methods and results

{OPEN}

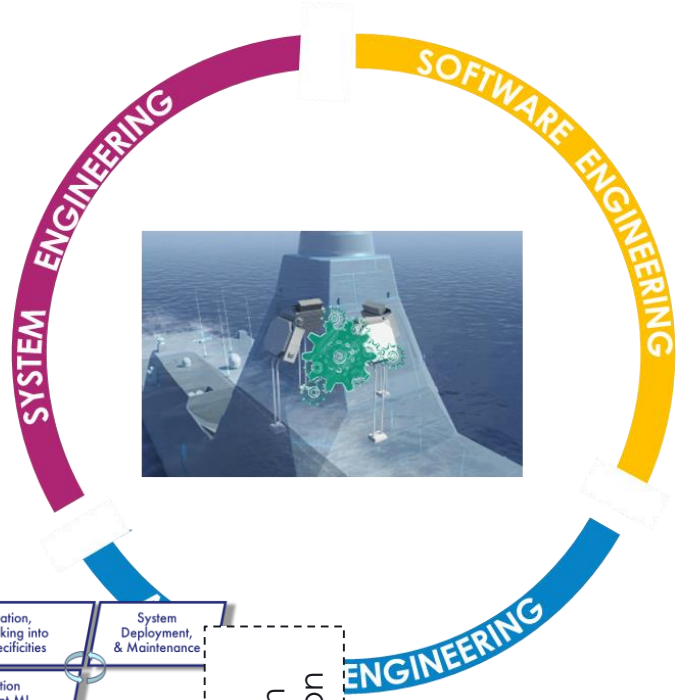
A triple-engineering approach



cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / Template: 87211168-DOC-GRP-EN-007

{OPEN}

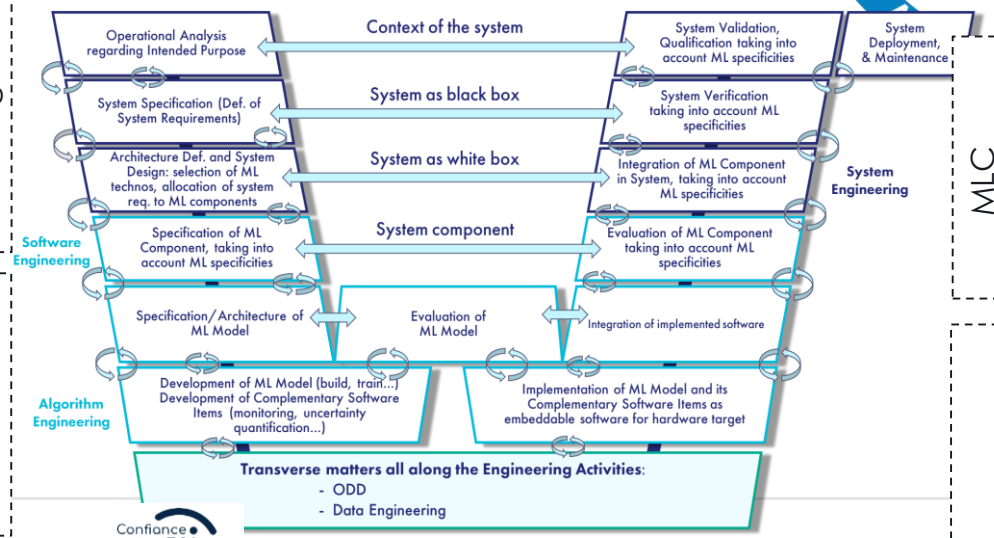
Reconciling the system and the model viewpoints



Standards (ARP6983)
➔ towards **certification** of **ML components**

ML Constituent Design

MLC SW/HW Development

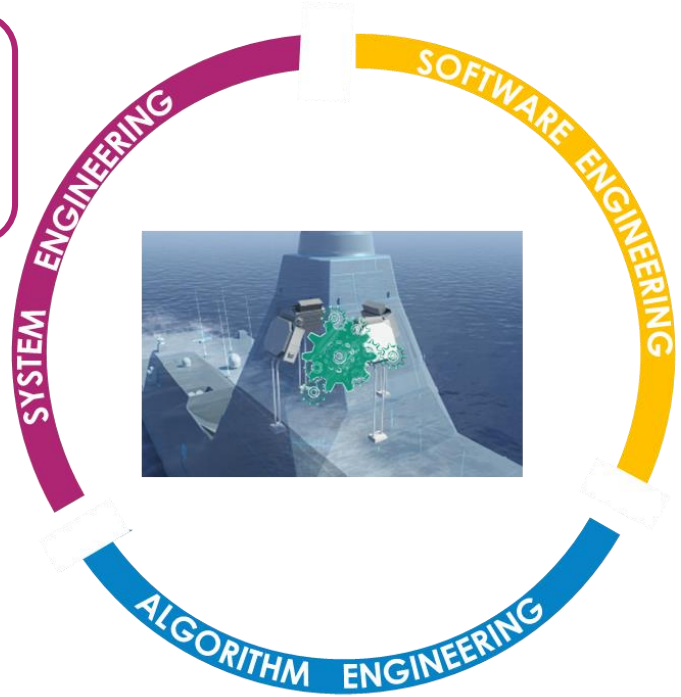


MLC Integration & Verification

MLC SW/HW I & V

Reconciling the system and the model viewpoints

Reconciling **Data** and **Functional Intent**
➔ Operational Design Domain **(ODD)**
➔ Rigorous methodologies and tools (problem formalization, symbolic AI,...)



Standards (ARP6983)
➔ towards **certification** of **ML components**

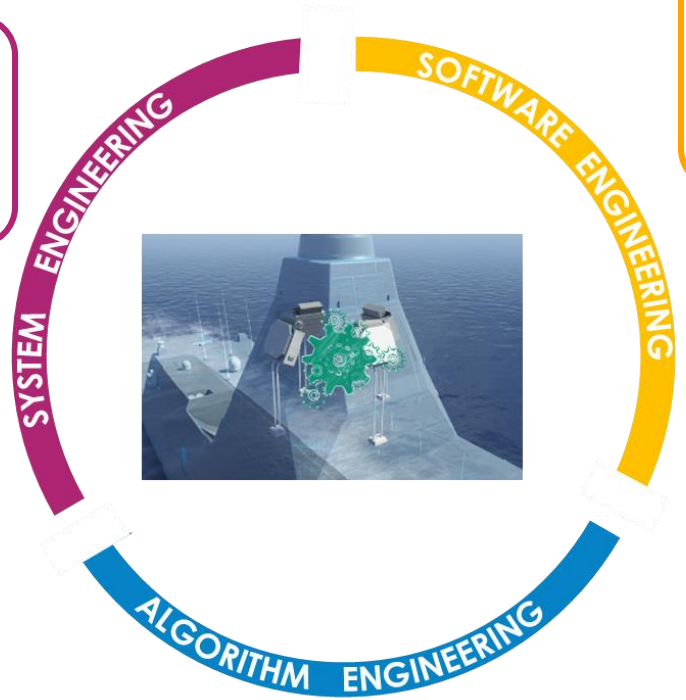
Operational Design Domain **(ODD)**

Voluntary **restriction of the Operational Domain** (specific operating conditions) within which an **AI constituent** within a **given system** is **intended to function**

cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / Template: 87211168-DOC-GRP-EN-007

Anchor the approach in Software Engineering

Reconciling **Data** and **Functional Intent**
➔ Operational Design Domain **(ODD)**
➔ Rigorous methodologies and tools (problem formalization, symbolic AI,...)



- Data & Knowledge **design guidelines**
- Continuous Integration
➔ **ModelOps**

Standards (ARP6983)
➔ towards **certification** of **ML components**

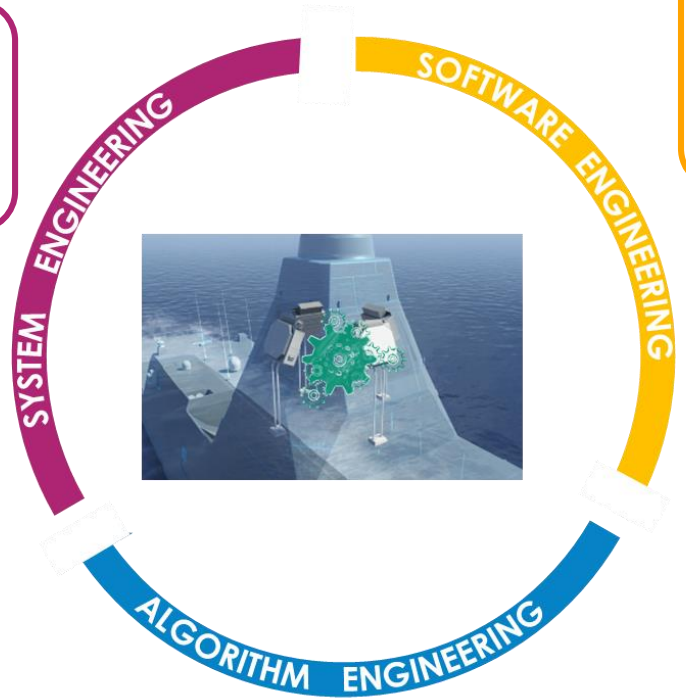
Formal verification/ proof of code

cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / **Template:** trtp version 9.0.0 based on **Template: 87211168-DOC-GRP-EN-007**

Tackle the Algorithm Engineering deadlocks

Reconciling **Data** and **Functional Intent**
➔ Operational Design Domain **(ODD)**
➔ Rigorous methodologies and tools (problem formalization, symbolic AI,...)

- Data & Knowledge **design guidelines**
- Continuous Integration
➔ **ModelOps**



Standards (ARP6983)
➔ towards **certification** of **ML components**

Formal verification/ proof of code

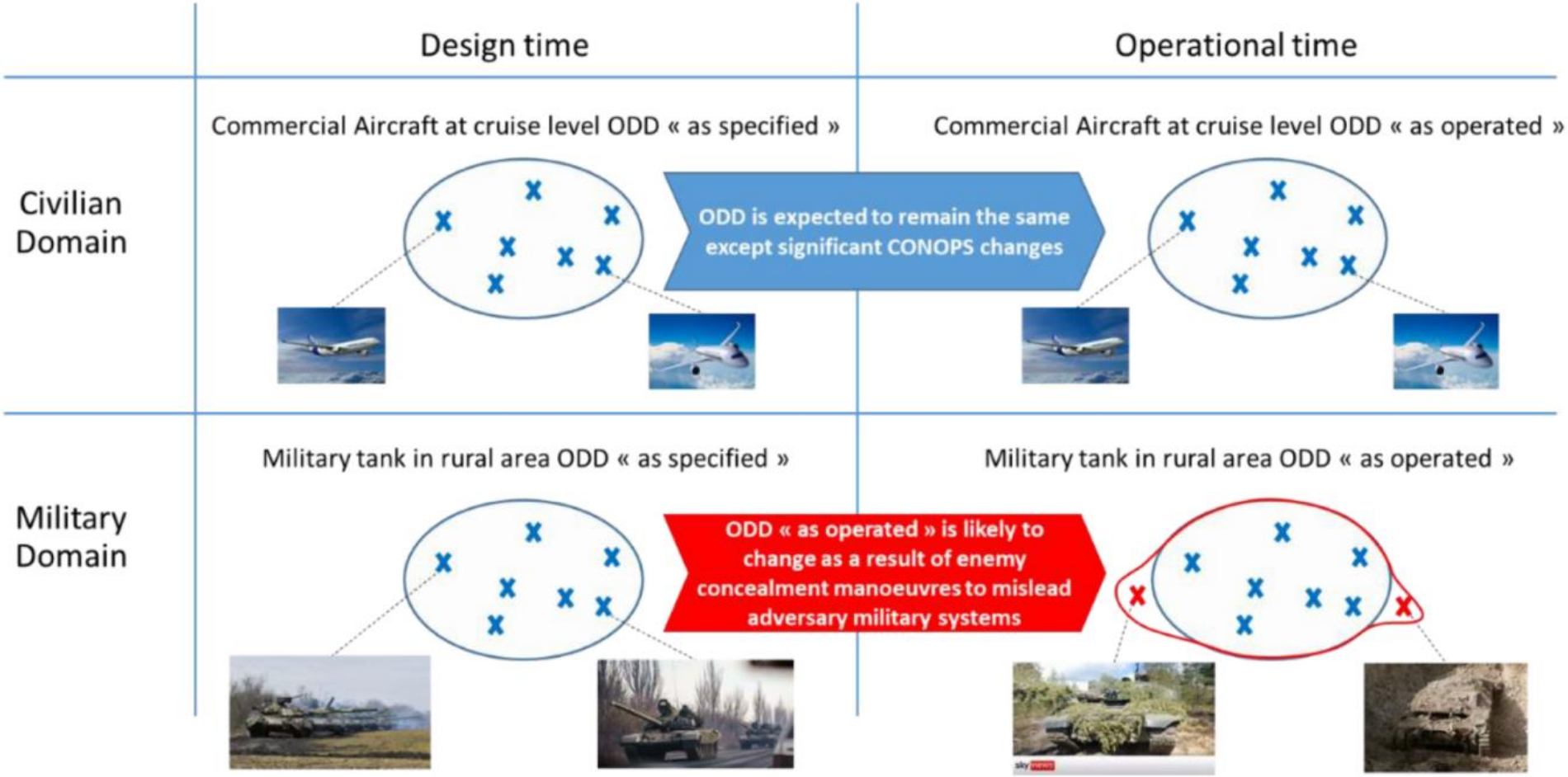
Monitoring
in **operational environment** (abnormality detection, XAI)

Metric Definition
to **assess** and **monitor** algorithm **fit-for-purpose**

Evaluation & tests :

- **Formal** methods
- **Axiomatic** proofs (robust by-design models)
- **Experimental** approach (Acceptable for IVVQ as long as **analyses are complete, traceable** and **quantitative**)
- Explainable AI (**XAI)**
- **Cybersecurity** tests

Adaptive ODD: Incremental Learning & Qualification



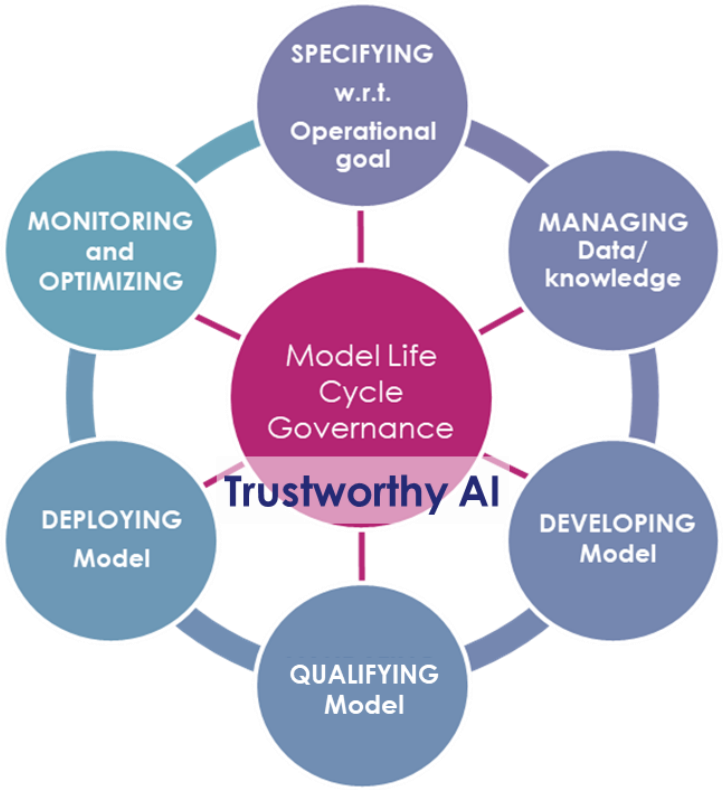
Differences between civilian and military ODDs *[From EDA TAID white Paper]*

End-to-End Trustworthy AI Engineering Lifecycle

Critical needs in defense

At run time

Runtime Assurance (adaptive ODD)



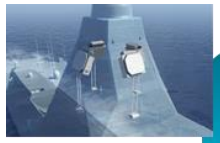
At design time

Development assurance

cortAIx/Labs/ARX/PBE/25,0029 Thales Research & Technology France / Template: trtp version 9.0.0 based on Template: 87211168-DOC-GRP-EN-007

{OPEN}

Conclusion : Trustworthy AI is at hand



SYSTEM

IVVQ viewpoint

AI MODEL

Data viewpoint

Interdisciplinary approach is a key to sucess: system engineering, safety, security, software...

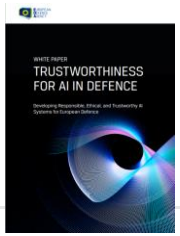
➔ *IA can't be trusworthy per se*
➔ *The goal is to quantify and manage the risk*

Some resources:

- Take a look at the Confiance.AI project outputs
  https://www.confiance.ai/

- We will release soon an **open source soft** for benchmarking neural networks based on **confidence scores** on the ThalesGroup GitHub
  https://github.com/ThalesGroup

- **EUROCAE/SAE ARP6983** to be released in early **2026**



- Take a look at the freshly published **EDA white paper** on **Trustworthy AI for Defense**
  https://eda.europa.eu/docs/default-source/brochures/taid-white-paper-final-09052025.pdf

{OPEN}

THALES
Building a future we can all trust

Patricia Besson

Thales / cortAIx Labs France

patricia.besson@thalesgroup.com