# Artificial Intelligence

## for Critical Systems

THALES
Building a future we can all trust

THALES
Building a future we can all trust

# Artificial Intelligence for Critical Systems

Juliette Mattioli and Christophe Meyer
November 2024

## Summary

Thales believes that, when managed effectively, AI enhances human decision-making, enabling both speed and quality. Major internet companies and prominent international research institutions have significantly advanced AI for civilian applications. However, deploying AI in the defence sector and critical systems necessitates careful consideration of several constraints. These include the complexity of the systems involved (such as sensors and effectors), the need to integrate AI within limited parameters of size, weight, and power, and the challenges posed by intermittent, low-bandwidth, or contested connectivity between subsystems, such as during jamming.

Additionally, for critical systems, it is crucial to define the acceptable and necessary levels of autonomy for deployed AI systems based on current circumstances and the operational environment. This includes determining adaptation capabilities between missions and the communication methods with human operators (human-machine dialogue).
While learning AI technologies can achieve performance levels beyond those of traditional algorithms and unlock new capabilities, they also introduce vulnerabilities unique to AI. In response, Thales has strengthened its expertise and cybersecurity capabilities to address these specific AI-related needs.

At Thales, the design and qualification of AI-based systems carefully consider security and safety constraints relative to their criticality, autonomy, and adaptability. This approach enables the utilisation of cutting-edge technologies, whilst maintaining stringent control over their implementation.

# 1. Introduction

On 28 March 2024, Thales unveiled its acceleration framework aimed at expediting the integration of artificial intelligence (AI) in critical systems, particularly within the defence sector. This initiative leverages Thales's core strengths:

- Technical expertise and operational knowledge, including mastery of architectures and management of constrained environments (such as embeddability and frugality), along with a strong emphasis on security and reliability.
- Mastery of all key AI technologies, including hybrid, embedded, frugal, autonomous, distributed, explainable, interactive, and generative AI, alongside their practical applications.

The acceleration process is being conducted in a controlled manner to effectively tackle challenges related to reliability, transparency, security, and ethics, aligning with Thales's trusted AI strategy for critical systems and adhering to the European regulatory framework set by the AI Act.

Within Thales, AI brings together 600 specialists who collaborate across the field, from research, technology, and innovation (RT&I) to implementing AI solutions in sensors and systems, organised within the "Sensors," "Factory," and "Labs" framework:

- cortAIx Sensors focuses on accelerating the development of AI use cases leveraging sensor data. The AI initiatives related to sensors will be hosted within operational entities that oversee their design, including the AI components.
- cortAIx Factory industrialises AI use cases based on system data, as well as AI engineering methods and tools. Initially, cortAIx Factory will serve defence clients both in France and internationally. Key areas of focus for cortAIx Factory include C4I (Command, Control, Communications, Computers, Intelligence), autonomy, support services, and collaborative combat.
- cortAIx Labs harnesses the RT&I resources of the Group in AI, fostering synergies to provide operational units and cortAIx Factory with technological building blocks and application demonstrators, as well as trusted AI engineering tools and techniques.

# 2. Critical Systems

## 2.1 DEFINITION OF CRITICAL SYSTEMS

A critical system is a system whose failure can have serious consequences, including environmental harm, substantial material damage, severe injuries, or even loss of life. Such issues are typically linked to sectors where safety is of utmost importance, including total or partial automation in transportation (aviation, maritime, and ground-based autonomous vehicles), medical devices, the pharmaceutical industry, energy production (especially nuclear energy), banking payment systems, and the security and defence sectors.

The criticality of a system is assessed based on the potential repercussions of its failures. These systems are designed for high reliability and security, adhering to exacting standards and undergoing rigorous certification processes. The integration of AI technologies into critical systems requires a reconsideration or enhancement of existing norms and certifications. In this context, the adoption of the AI Act at the European level regulates the development of AI and mitigates associated risks. This legislation addresses not only security, reliability, and safety, but also the protection of private data, transparency, and ethical considerations.

## 2.2 CHALLENGES AND CONSTRAINTS

As a systems integrator, Thales views the AI it develops or integrates as a complement to human intelligence—leveraging domain expertise to enhance intelligent critical systems. Several constraints specific to critical systems must be taken into account from the design phase, including embeddability (size, weight, power consumption), connectivity (intermittent, low-bandwidth, or contested scenarios such as jamming), operational factors (hidden complexity, intuitive use), along with safety and (cyber)security.

Thales believes that, when carefully controlled, AI can empower humans to make more informed and quicker decisions, reducing their mental workload by delegating non-critical functions and partially delegating critical functions under controlled conditions. Given the dynamic nature of constraints and available resources—constantly evolving within the deployment contexts of most critical systems, including energy, computing power, memory availability, response time requirements, and inter-component connectivity—Thales is compelled to design "reactive" and "polymorphic" AI. "Reactive" refers to the ability to produce quality results within the allocated time, even if these are not optimal. "Polymorphic" signifies that a single function must be adaptable to varying resource levels, potentially sacrificing performance while adjusting to available processing power.

For instance, Thales designs cameras capable of:

- Selecting from various algorithms (both AI and non-AI) based on the local resources available, resulting in different levels of analysis of the observed scene.
- Transmitting analysis data while compressing it to varying degrees (including short texts, images, or HD video streams) based on the available bandwidth, to other system components for further analysis.

Within the context of AI-based decision support systems, Thales creates solutions that can provide time-sensitive recommendations that can be continuously refined and enriched.

## 3. The various paradigms of AI:

The term "Artificial Intelligence" was adopted at the Dartmouth Conference in 1956 to designate this new research field. Its simplest definition is provided by the Villani mission report (2018): "Artificial Intelligence is a computer programme aimed at performing tasks that require a certain level of intelligence, at least as well as humans."

### 3.1 CONNECTIONIST, SYMBOLIC, HYBRID

In recent years, connectionist AI, leveraging machine learning techniques, has gained prominence, overshadowing symbolic AI. This connectionist and statistical approach is rooted in a biological paradigm inspired by the workings of the human brain, aiming to infer conceptual models from examples. While this methodology is well-suited for perception, it is less effective for solving complex problems. Deep learning and generative AIs, such as LLMs (large language models) and LVMs (large vision models) fall into this category. In contrast, symbolic AI employs formal reasoning and logic, representing a Cartesian approach to intelligence, where knowledge is encoded from axiomatic principles, and consequences are derived from them.

The fundamental distinction between these two paradigms lies in the definition of knowledge. In symbolic AI systems, experts explicitly define this knowledge, while in connectionist and statistical approaches, knowledge is inferred automatically from data. David Sadek, VP at Thales for Research, Technology and Innovation in AI, notes that "connectionist AI was, until recently, AI of the senses, and symbolic AI was AI of meaning."

In many operational contexts, especially in critical systems that function in dynamic and uncertain environments, it is essential to consider types of information beyond mere sensor data. These include information already embedded in physical models (such as simulation tools or partial differential equations) or domain knowledge captured through ontologies, logical rules, and semantic models. The concept of hybrid AI (§6.1.1) can be broadly defined as an integration of two distinct approaches, including those enhanced by mathematics or physics

For instance, reinforcement learning (RL), which optimises strategies through trial and error, potentially augmented by human feedback and specific simulation tools, exemplifies hybrid AI. Similarly, systems that use LLMs or LVMs alongside information retrieval systems and possibly autonomous artificial agents—such as those based on Retrieval-Augmented Generation (RAG) or Reasoning and Acting (ReACT) architectures—can also be classified as hybrid systems.

## 4. Autonomy and Adaptability

To effectively establish the conditions under which an AI can develop a decision or strategy—especially in contexts where it suggests actions to a human decision-maker (decision aid or support) or implements those decisions directly—it's essential to assess its degree of autonomy and adaptability dynamically, based on changing circumstances.

In the automotive industry, a widely accepted framework for defining the level of autonomy in vehicles is based on the NHTSA (National Highway Traffic Safety Administration) classification, which ranges from levels 0 to 4. This classification evaluates a vehicle's capability to make decisions based on its environmental analysis, including functions such as emergency braking, maintaining safe distances from other vehicles, and changing lanes. Here's a brief overview of these levels:

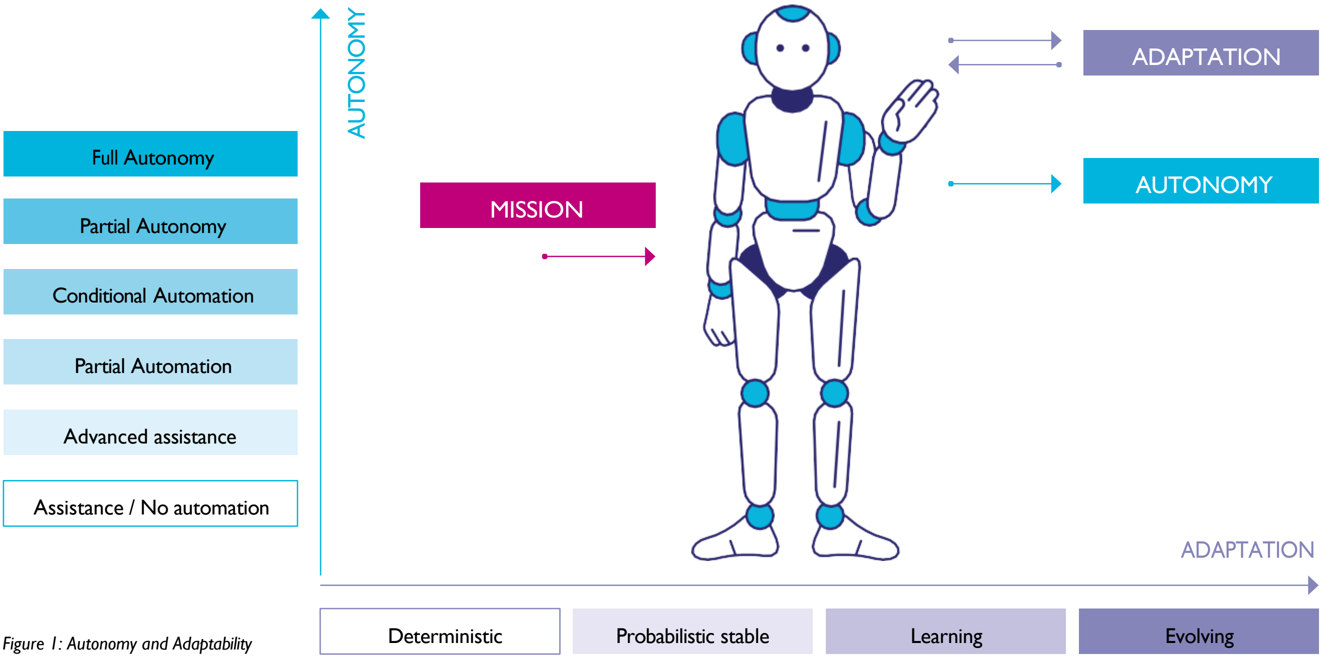


*Figure 1: Autonomy and Adaptability*

- Level 0 (Pure Assistance): the vehicle may potentially warn the driver of certain conditions or dangers, but does not modify the driving in any way.
- Level 1 (Advanced Assistance): the vehicle can perform certain simple actions, such as speed regulation, but any contrary action taken by the driver takes precedence. It does not perform more complex actions, such as changing lanes.
- Level 2 (Partial Autonomy): at the explicit request of the driver, the vehicle can perform certain manoeuvres, such as parking. However, the driver must remain capable of taking control at all times.

Levels 0 to 2 of automotive autonomy do not hold the AI responsible for any incidents, as they are simply driving aids.

- Level 3 (Conditional Autonomy): at the explicit request of the driver, and under certain conditions that the driver cannot change (such as maintaining speed below a certain threshold), the vehicle can perform most manoeuvres.
- Level 4 (Almost complete autonomy): the vehicle can perform all manoeuvres without requiring driver supervision.

It is important to note that, in all cases, the driver retains the option to engage or disengage the autonomous features of their vehicle at all times. Consequently, they have control over the level of autonomy. This same framework applies to all critical systems, as referenced by Thales.

In our dynamic and often chaotic world, a solution deemed appropriate at one moment may become inadequate, or even pose a danger, in a changed environment—particularly for critical systems. As such, solutions need to evolve over time. The AI systems supporting them possess learning capabilities that enable them to adjust their responses and behaviours to the context in which they operate. However, it is crucial to manage these adaptations carefully.

Thales proposes a scale of "adaptability" for operational AI systems, specifically applicable to their deployment rather than during design or training phases. This scale comprises four levels, ranging from 0 to 3:

- Level 0: No Adaptability
  The AI's outputs are either predetermined or stochastic based on the inputs, remaining unchanged over time. For stochastic systems or those that use fuzzy logic, this indicates that the probabilistic distributions are static. There is no capacity for adaptation during operation; even if the AI has advanced during its design or training, it does not evolve in the operational phase.
- Level 1: Deterministic Learning
  The potential outputs of the AI are known and finite. However, the selection of a single, deterministic decision (driven by inputs) can evolve through learning.
- Level 2: Stochastic Learning
  The possible outputs of the AI are also known and finite. Nevertheless, the decision-making process is not singular; instead, it is influenced by a probabilistic distribution (based on inputs) that can change through learning.
- Level 3: Evolution
  The AI demonstrates the capacity for evolution by generating new outputs or even altering its structure. Continuous genetic programming falls into this category.

The modalities of adaptability in AI systems necessitate careful consideration. It is crucial to differentiate between AI systems that can adapt "continuously" during their operational phases and those that possess the ability to evolve between operational periods. The latter allows for thorough testing, validation, and qualification before re-entering service, particularly emphasizing the control that human operators have over the evolution process.

# 5. Challenges in AI for Critical Systems

Various barriers hinder the adoption of AI, particularly in critical systems where security (cybersecurity) and safety must be inherently guaranteed. These systems are also required to uphold principles of trust and accountability. Integrating AI solutions—whether using learning techniques or more traditional symbolic approaches—poses a number of challenges compared to established engineering practices. Key issues include how to define a high-dimensional domain of application (such as those involving vision or natural language), ensuring the robustness of AI-based systems in light of their cyber vulnerabilities, and guaranteeing the replicability of inherently non-deterministic algorithms. The challenge for Thales lies in defining and equipping a comprehensive "AI Engineering" process that encompasses algorithmic, software, and system dimensions

## 5.1 TECHNICAL EXPERTISE, BUSINESS EXPERTISE AND AI EXPERTISE

A significant challenge in integrating AI into mission-critical systems lies in balancing the exploitation of business expertise—often rooted in sensor physics—with domain knowledge (including air traffic control and military operations) and proficiency in advanced artificial intelligence technologies.

What sets Thales apart is its expertise in these three key areas. Our integrated teams, comprising technical experts, business specialists, and AI practitioners, work collaboratively to develop innovative solutions. Given the current capabilities of IT systems and data management, it remains essential to pre-process sensor data to ensure the expected performance of specific applications. For instance, a radar system can generate terabytes of data per hour, while an electronic warfare system may produce petabytes of data within the same timeframe. Only through specialised knowledge in these domains can we design pre-processing procedures that convert raw data into actionable "leads" for AI systems. Furthermore, it is the combination of business acumen and technical expertise that allows us to define the functions, relevant outputs, and boundaries of AI-assisted systems available today.

Thales forms multi-disciplinary teams of engineers with dual expertise to tackle the challenges of integrating AI into certain solutions effectively. Typically, experts in both AI and cybersecurity collaborate to define reference architectures for AI cybersecurity (§6.3.2), while those skilled in both AI and signal processing focus on "augmented" speech solutions, including compression, denoising, and intelligibility enhancement using deep learning technologies, as well as new electronic warfare applications.

# 6. The Thales Answer

## 6.1 KEY TECHNOLOGIES

Validity, explainability, security, and responsibility are indeed essential prerequisites for the qualification, approval, and certification of critical AI-based systems. To move forward, it is imperative to provide guarantees that allow usage in Thales's application domains. As a result, the Group places emphasis on differentiating technologies that support these goals, described below (Fig.2)

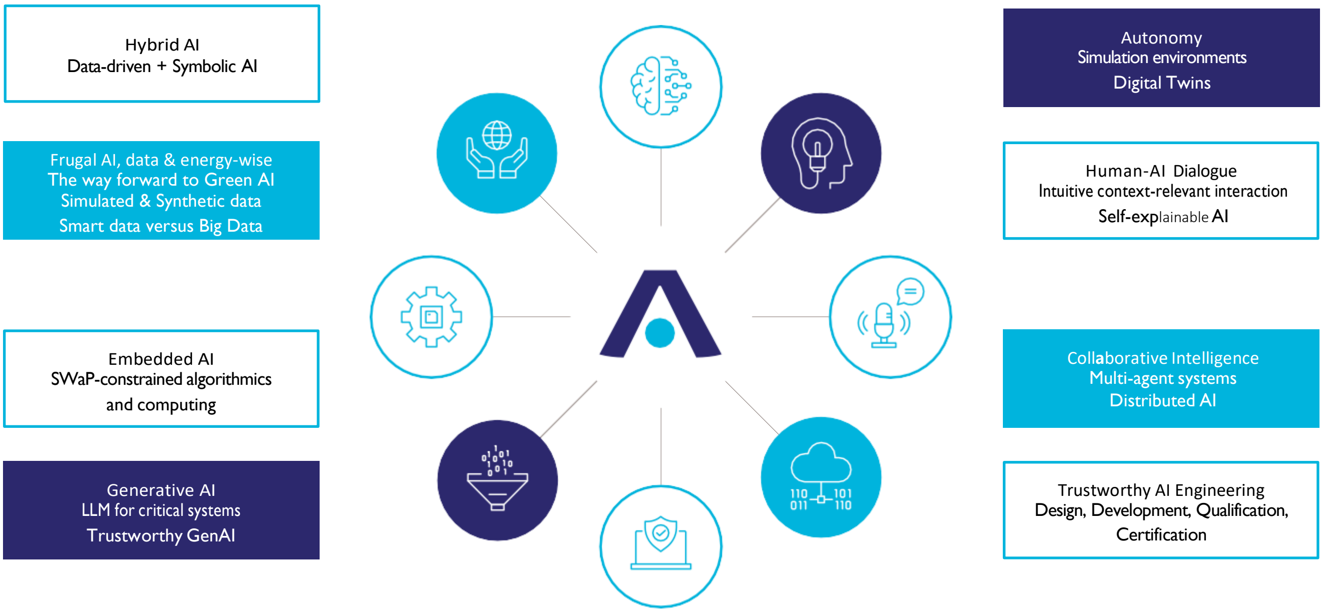
### 6.1.1 Hybrid AI

Despite recent advancements, deploying AI in mission-critical systems presents significant challenges. Data-driven AI often lacks transparency, interpretability, and robustness, and is not frugal in terms of data and energy consumption; symbolic AI is not always robust when faced with uncertainty. To overcome the limitations inherent in both approaches, hybrid AI has emerged as a viable solution. By leveraging the strengths of connectionist methods, hybrid AI integrates a variety of available knowledge–including business expertise, physical principles and mathematical frameworks–during the model development process. This integration aims to enhance interpretability, ensure robustness, and facilitate validation processes necessary for approval or certification.

The "Assured Neuro Symbolic Learning and Reasoning" (ANSR) programme launched by DARPA in 2022 exemplifies efforts to tackle these challenges. The programme seeks to develop new hybrid AI algorithms that effectively combine symbolic reasoning with learning, ultimately creating robust, safe, and trustworthy systems suitable for critical applications.

The emergence of "Physics-Informed Neural Networks" (PINNs) represents a significant advancement in the integration of AI with physical models. These neural networks are trained to solve supervised learning tasks while adhering to the laws of physics, as articulated by differential equations that limit the solution space admissible by the neural network during the learning phase. PINNs can accommodate a wide range of physical laws, including fluid mechanics (e.g. Navier-Stokes equations), electromagnetism (e.g. Maxwell's equations), and thermodynamics (e.g. Fourier's equations). The ability to integrate diverse physical equations positions PINNs as a valuable tool for critical applications in sectors such as aeronautics and defence, especially through the use of of digital twins.

Additionally, Information Geometry has gained traction as a useful tool in AI, using the gradient associated with the Fisher metric to take into account the geometric structure of the parameter space of multilayer networks. Initial proofs of concept of "Geometric-Informed Neural Networks" (GINNs) have demonstrated their potential in various applications, including Automatic Target Detection & Recognition (ATDR) using micro-Doppler signatures, target kinematics, and image recognition from 360° fisheye cameras.



Hybrid AI
Data-driven + Symbolic AI

Frugal AI, data & energy-wise
The way forward to Green AI
Simulated & Synthetic data
Smart data versus Big Data

Embedded AI
SWaP-constrained algorithmics
and computing

Generative AI
LLM for critical systems
Trustworthy GenAI

Autonomy
Simulation environments
Digital Twins

Human-AI Dialogue
Intuitive context-relevant interaction
Self-explainable AI

Collaborative Intelligence
Multi-agent systems
Distributed AI

Trustworthy AI Engineering
Design, Development, Qualification,
Certification

*Figure 2: Differentiating AI technologies*

### 6.1.2 Frugal AI

Training an AI system typically requires a substantial amount of annotated data. However, in the defence sector, data is often limited or sensitive. The challenge lies in developing efficient learning solutions and constructing representative datasets from minimal real annotated data. To address this, Thales employs the following strategies:

- Model Compression: This technique focuses on reducing the size of AI models through methods such as quantisation, truncation, and model distillation. By minimising model size, these can be executed efficiently on less powerful devices.
- Federated Learning: This approach allows AI models to be trained by distributing the learning process across multiple devices or local servers. This reduces reliance on expensive centralised infrastructures and preserves data confidentiality by keeping data on local devices.
- Transfer Learning: This technique involves using an AI model pre- trained on a specific task and adapting it for a similar task. This ap- proach significantly decreases the time and resources required to train new specialised models.

Finally, Thales is exploring advanced nanoelectronic materials and neuromorphic computing methods for information storage and computation, aimed at designing AI systems that consume less energy.

### 6.1.3 Reinforcement learning and simulation

Reinforcement learning refers to optimising strategies based on a trial-and-error approach with potential human feedback and specific simulation tools. Provided that an evaluation function is available to assess the quality of a solution or strategy, a reinforcement learning system can investigate the spectrum of potential solutions using various methodologies. This enables the system to enhance its proposed solution progressively over time.

The evaluation function may take the form of simple human feedback, indicating whether one solution is superior to another (relative evaluation), or it could involve calibrated human judgement (absolute evaluation). Additionally, the evaluation function might result from specific calculations, such as the number of points scored in a game. The strength of reinforcement learning lies in its capacity to manage intermittent evaluations, as the quality of a move in a game may sometimes only be assessed at the conclusion of the game. When an evaluation of a solution is accessible, the algorithm must discern which actions resulted in positive effects and which ones led to negative outcomes.

Reinforcement learning predominantly relies on simulations, allowing the system to "learn" within a simulated environment before being implemented in the real world. This approach theoretically enables numerous iterations of trial and error. The effectiveness of the learning process is contingent upon the quality of the simulation—factors such as model precision and representativeness—as well as the management of the "reality gap," which refers to the discrepancies between the simulation and the real-world environment. Thales has extensive expertise in simulation, encompassing sensor simulations, physical phenomena simulations, and real equipment simulators used for training purposes, including those for aircraft, tanks, and helicopters. This expertise provides Thales with a significant advantage in the development of reinforcement learning systems based on simulated environments.

In early 2020, EDF, Thales, and TotalEnergies established the SINCLAIR (Saclay INdustrial Collaborative Laboratory for Artificial Intelligence Research) laboratory at the EDF Saclay site. The laboratory's research programme is centred on developing artificial intelligence methods and tools that cater to the shared needs of these three companies. One of the primary R&D focus areas is reinforcement learning, emphasising both explainability and simulation.

### 6.1.4 Explainability

Thales has been working on explainability in multi-criteria decision support systems since 2012. Within the context of the SINCLAIR laboratory, the emphasis of this research is on ensuring that automatic systems can articulate the reasoning behind their proposed solutions. Without genuine explainability, artificial intelligence systems risk being seen as black boxes. However, the explanations provided must be adapted to match the comprehension level of the intended audience. This capability not only assists developers in verifying that the system operates as intended but also enhances user trust and model auditability. Explainable AI plays a crucial role in mitigating risks related to compliance, rights, security, and reputation concerning the outcomes produced by AI systems.

Generative AIs can produce highly articulate texts, but they can also include inaccuracies. For instance, when ChatGPT was first introduced, it would respond to a question about which is heavier, an elephant egg or a whale egg, by asserting that the elephant egg is heavier, supported by a well-structured explanation that would appear perfectly logical to a child.

This highlights the importance of differentiating between persuasive AIs and explainable AIs. The latter must be able to effectively communicate relevant elements that are comprehensible to domain experts or users.

### 6.1.5 Generative AI

Large Language Models (LLMs) generate coherent continuations or completions of input text sequences, using either deterministic or pseudo-random methods based on the user's preferences. This process is driven by deep learning algorithms trained on extensive datasets, often encompassing a vast array of publicly available texts from the internet. Additionally, LLMs operate under certain acceptability guidelines, which prohibit the generation of content that is sexist, racist, or otherwise unethical. However, the inherent nature of LLMs allows them to produce "logical" text sequences that are grammatically correct and syntactically coherent, even if they convey completely inaccurate meanings. This phenomenon is known as "hallucination".

Similarly, Large Vision Models (LVMs) are engineered to automatically identify and learn the underlying structures and relationships within extensive collections of images or videos, facilitating the creation of new visual content.

The primary challenge in using these technologies lies in leveraging potentially confidential operational data without exposing it, while also retaining control over the generated content (including verification and validation processes).

Mastering generative AI technologies presents a wealth of applications for Thales, including: (Fig.3)

- Support for cross-functional areas such as human resources, com- merce, marketing, and legal departments, for instance through au- tomatic summarisation, the generation of original visuals, and the adaptation of archives to new contexts.
- Automatic code generation or transcoding.
- Enhanced use of operational document databases (e.g. technical documentation) via conversational agents.

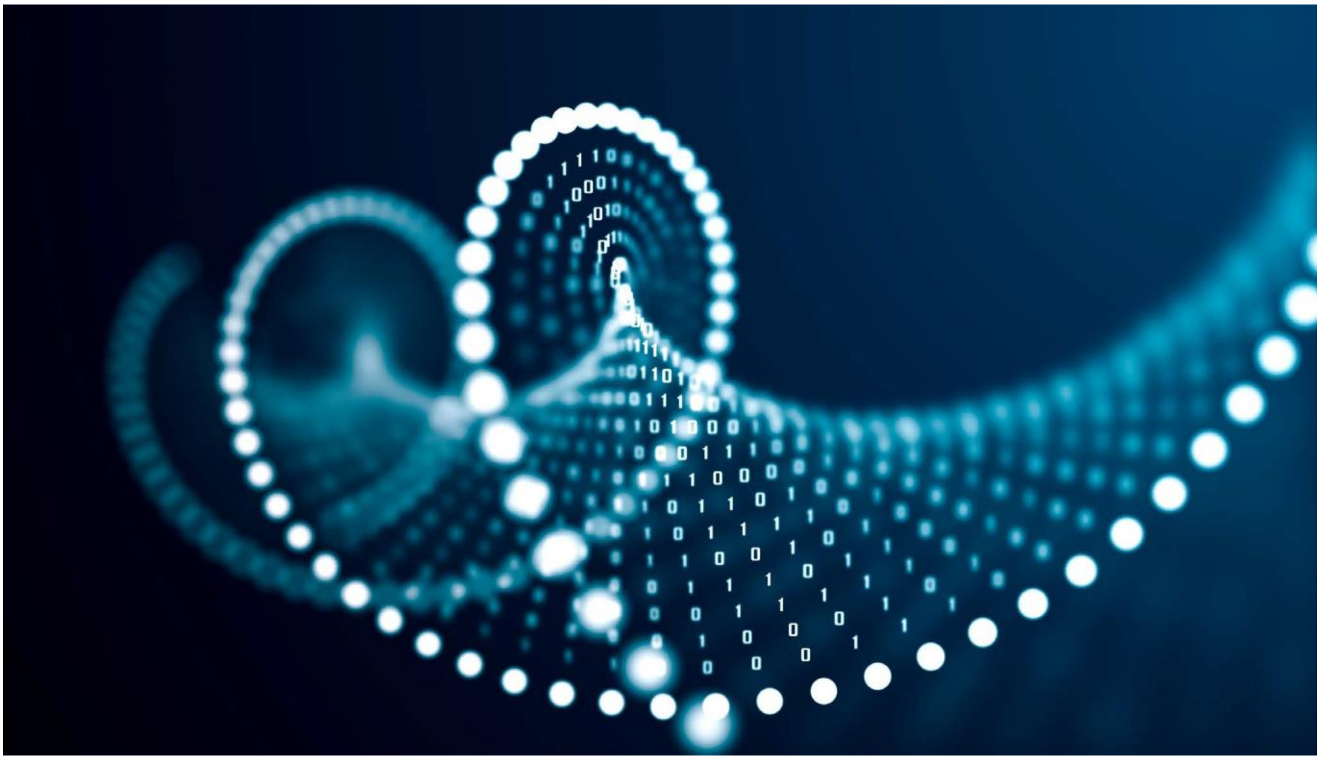


*Figure 3: Generative AI @ Thales*

Thales is currently conducting more than a hundred experiments related to these applications. Recent and rapid advancements in structuring LLM/LVM models, such as the Retrieval Augmented Generation (RAG) method and the Reasoning and Acting (ReACT) approach present significant disruptive opportunities for Thales. The RAG method connects an LLM to an enterprise's operational data through a modern information extraction system to address a range of queries. The ReACT approach utilises an LLM to produce interpretable reasoning for human users and potentially implement these insights.

Several initiatives within Thales have already commenced the exploration and application of these methodologies, including:

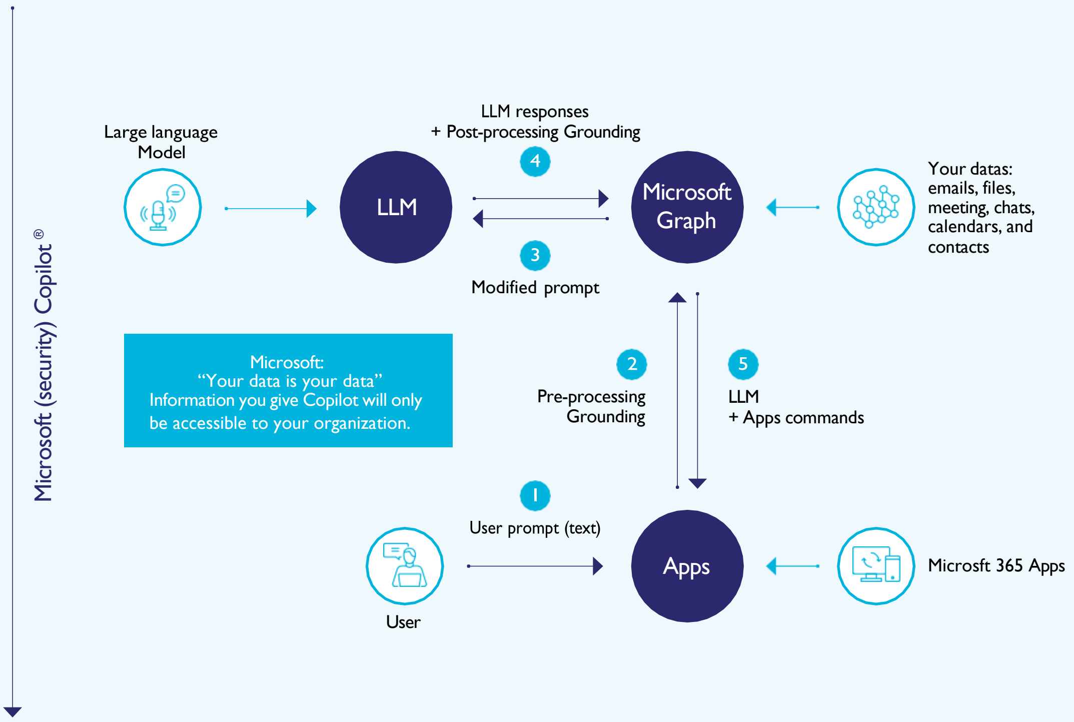
- GenAI4SOC (Generative AI for Security Operating Centres), which is focused on the automatic generation of cyber detection rules.
- GenAI4MCS (Generative AI for Mission Critical Systems), which integrates LLMs with agent technologies, business service interactions, and cybersecurity measures to create future autonomous and adaptive conversational assistants for command and control systems.

Lastly, the study and mastery of advances in the multimodal capabilities of generative AIs – where the internal vector representation of data remains consistent across all types (for instance, the number 3, the words "trois" or "three," and an image depicting three objects all possess equivalent or highly similar internal representations)—are expected to facilitate the creation of systems capable of generating realistic and coherent atypical data. This includes radar data and target trajectories that are both realistic and coherent in relation to a simulated tactical scenario.

Thales aims to systematically integrate generative AIs, including LLMs and LVMs to fully leverage their tremendous generative capabilities while ensuring the relevant, reliable, and secure use of operational data. This ambition is somewhat similar to conversational assistant embedded within productivity suite - such as Microsoft 365 - that can use the complete range of each user's data (such as emails, calendars, and documents in Word, Excel, and PowerPoint) to substantially improve productivity by performing specific tasks on their behalf. (Fig.4)

From

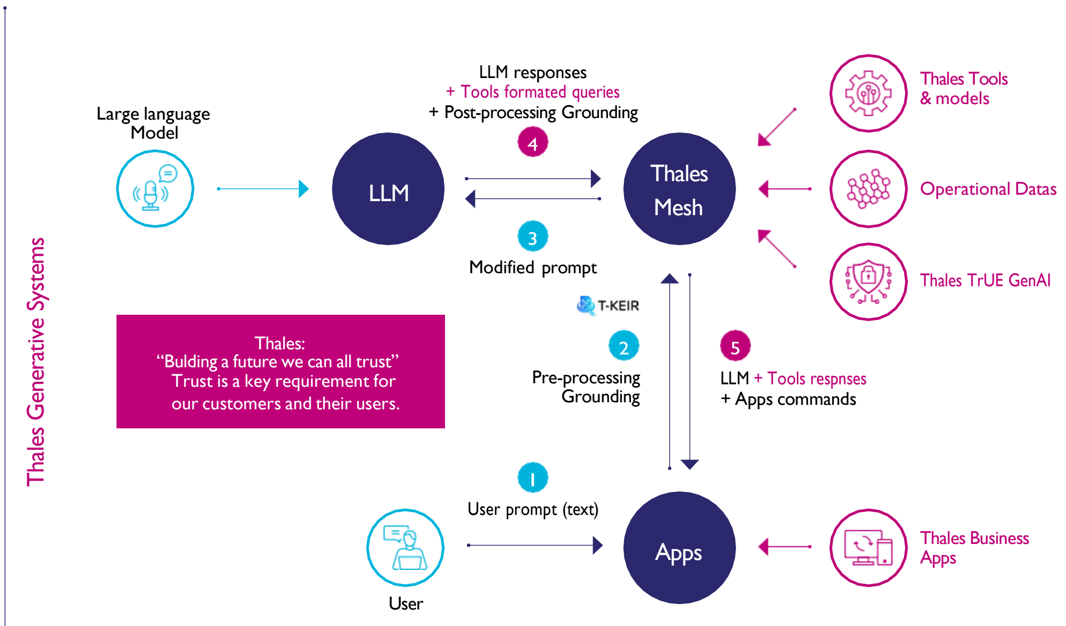
Microsoft 365
Copilot



Microsoft:
"Your data is your data"
Information you give Copilot will only
be accessible to your organization.

To

THALES
Building a future we can all trust



Thales:
"Bulding a future we can all trust"
Trust is a key requirement for
our customers and their users.

Thales way forward:

- R&T: Thales TrUE GenAI (witch KTD PCC)
- Data generation (Multimodal & Atypical)
- Integration in apllications (Smart Automated and Reliable Assitant; NextGen C4I…)

*Figure 4: Generative AI for Mission Critical Systems*

## 6.2. SYSTEM CONSTRAINTS

### 6.2.1 From Cloud to far Edge

The primary challenge of "Edge AI" is the transition from centralised AI in the cloud to embedded AI situated closer to end users. Edge AI involves the deployment of AI within an edge computing environment, allowing data to be processed in real time without the need for cloud connectivity. This enables sensors to make faster, more intelligent decisions. By extending high-performance computing capabilities to edge computing locations – where sensors and connected devices resideusers can process data on their devices in real time, eliminating the need for connectivity or system integration. The advantages for Thales applications are numerousincluding:

- Energy consumption: by deploying AI algorithms in proximity to the sensor and the user, especially in embedded systems, energy consumption and processing times are significantly decreased. This approach also minimises the costs and risks linked to data transmission.
- Bandwidth reduction: by processing, analysing, and storing data locally rather than transmitting it to the cloud, edge AI significantly lowers bandwidth usage in the data stream and helps minimise costs.
- Security and confidentiality: the focus on data transfer completely changes the size and shape of the attack surface. Furthermore, Edge AI enables the selective filtering of data before transmission to the cloud, ensuring that only necessary information is sent, a process that often includes anonymisation.
- Reduced latency: Edge AI analyses data locally and alleviates the load on cloud platforms, thus freeing them up for other tasks.
- Improved reliability: distributing computations across multiple de- vices enhances redundancy and reliability. In the event of device failures, other devices can maintain operations, ensuring continuous functionality.
- Deployment "in the field": Edge AI is ideally suited for IoT applications and mobile devices, including autonomous platforms such as drones.

Thales has leveraged its expertise in High Performance Computing (HPC) to advance embedded AI solutions specifically designed for Far- Edge applications for over a decade.

### 6.2.2 Humain-AI interactions

For any critical system that requires collaboration between humans and intelligent machines, Thales is committed to defining the framework and components that regulate access to sensitive data, authorised actions, respective roles, and modes of real-time interaction, taking into account the overall state of the system and its context.

To facilitate effective collaboration between humans and AI, it is essential to establish a clear and unequivocal delineation of responsibilities, based on their respective competencies, performance, reliability, the current state of the system in use, the environment (including external parameters), applicable regulations, and ethical considerations (both national and corporate).

At the same time, it is essential to define the levels of autonomy and adaptability of AI systems (§4), taking into consideration the state of the system and its environment. While the concept of human-AI dialogue is not novel, the capabilities provided by generative AI, particularly those using LLMs, enable more direct and efficient interactions.

When AI is used to assist or support human decision-making, humans retain control over actions and bear full responsibility for them. However, there are instances where the decision-making process and its execution may not coincide with the timeframe required for human analysis and the dialogue between humans and AI. For example, it is impractical to expect an autonomous vehicle to explicitly seek validation from a human passenger before executing emergency braking, even though activating this capability falls within the vehicle's purview.

Whenever feasible and appropriate, general principles are established to ensure that humans can trust their intelligent assistants or robotic collaborators. (Fig.5)

*Figure 5: Human-AI relationship*

## 6.3. IMPLEMENTATION

The deployment of AI in mission-critical applications requires adherence to Reliability, Availability, Maintainability, Safety, and Security (RAMS) objectives. Consequently, a critical system must rely on robust development methodologies that encompass the entire process from design to deployment and qualification. Engineering practices must be enhanced with methods and tools that ensure reliability at every stage of the system's lifecycle. These include: (1) specifying the operational domain and its application to data and knowledge management; (2) designing algorithms and architecture; (3) characterisation, verification, and validation; (4) deployment, particularly in embedded architectures; (5) qualification and certification; and (6) maintenance in operational condition and cybersecurity.

### 6.3.1 Trustworthy AI engineering

To facilitate the design of a trustworthy AI component, particularly one driven by Machine Learning (ML), an MLOps (Machine Learning Operations) process–drawing significantly from the DevOps approach–focuses on refining the software development and deployment process. This approach seeks to enhance and unify the development and operational aspects of the component, leading to benefits such as faster deployment.

However applying MLOps in the context of mission-critical systems requires re-evaluation to ensure transparency and auditability. This is essential for enabling corrections in the event of failures and for providing evidence of conformity with expected properties, such as validity, robustness, explainability, and embeddability. To address these needs, Thales has been actively contributing since 2020 to the Confiance.ai programme, which aims to establish methodologies and tools for trusted AI engineering. This initiative encompasses the following key stages:

- Problem specification involves articulating various requirements–both functional and non-functional–as well as outlining the Operational Design Domain (ODD). The ODD describes the specific conditions under which the AI capability is intended to operate effectively, including environmental factors and other domain constraints. This specification plays a crucial role in guiding the tasks of data collection and knowledge modeling.
- The acquisition of data and knowledge, guided by the Operational Design Domain (ODD), leads to the aggregation of a homogeneous set that is of sufficient size and quality. This data should be understandable, relevant, reliable, and balanced. To be used effectively, this data typically undergoes processes such as cleaning, organising, and labeling. In some cases, further processing is necessary to transform raw information into a format that can be effectively used. This process falls under the umbrella of "Data Engineering", a process that can be complemented by "Knowledge Engineering."
- An AI algorithm can be designed or selected from a library of existing algorithms. In Machine Learning (ML), once learning is completed, the model is typically refined using a validation dataset. This may entail modifying or eliminating variables and adjusting hyper- parameters to achieve an acceptable level of accuracy. Implementation on the appropriate hardware platform or target system can impact technical requirements such as latency, memory constraints, and power consumption. After identifying a satisfactory set of hyperparameters and optimizing the model's accuracy, the model undergoes testing and characterisation on a designated dataset, and may also be subjected to formal verification. The evaluation process can extend beyond just functional performance metrics, such as accuracy, to encompass additional criteria related to robustness to noise and resilience to adversarial attacks. Once the AI/ML component has been integrated into a mission-critical system, it is vital to demonstrate that it possesses the expected "trust properties." To achieve this, a "Trusted AI System Engineering" analysis framework must be defined. This framework will guide the strategies for the development and Integration, Verification, Validation, and Qualification (IVVQ) of the system.

### 6.3.2 Cybersecurity for AI

While the advanced AI technologies in machine learning have surpassed the performance levels achievable by classical algorithms and introduced new capabilities, these innovations also bring unique vulnerabilities that must be addressed to ensure the integrity and security of AI systems. This has led Thales to enhance its expertise and resources in the field of cybersecurity, specifically addressing the vulnerabilities associated with AI technologies. (Fig. 6)

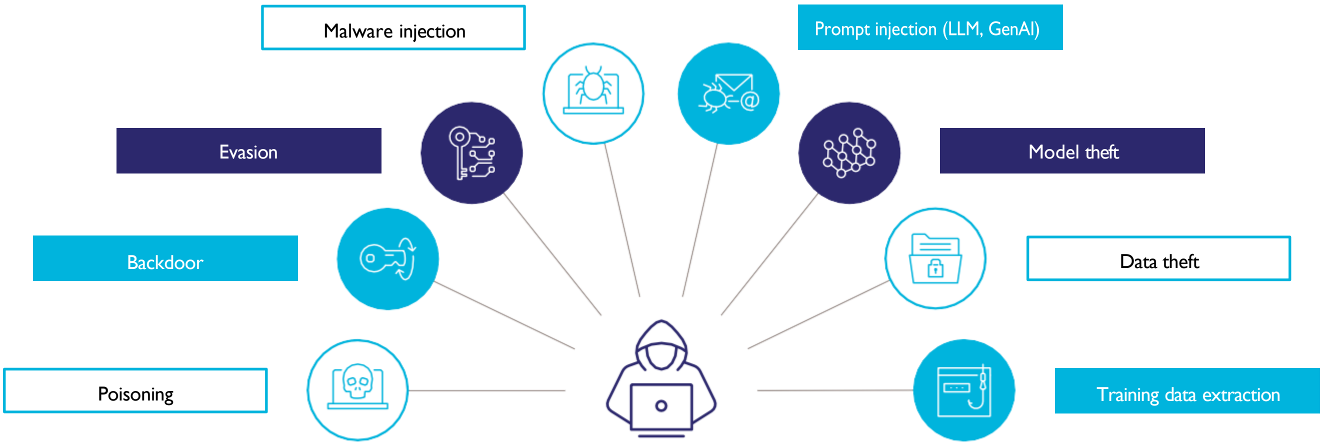


*Figure 6: AI cybersecurity: attacks*

Below are some of the examples of attacks studied by Thales's "AI Friendly Hacking" team, responsible for conducting research on AI cyber- security:

- Deep learning algorithms rely on the use of training data, and it is clear that the quality of this data–such as its representativeness and accurate labelling–impacts the algorithm's performance. One method to undermine the performance of these algorithms is through "data poisoning," where the training data is deliberately corrupted.
- Another approach involves developing AI systems that can generate input data designed to mislead even the most advanced image recognition algorithms. For instance, the AI Friendly Hacking team has created an AI algorithm that makes subtle adjustments to the characteristics of specific pixels in an image (particularly their RGB components) to induce inaccurate outputs from leading image recognition systems. This team has demonstrated the capability to execute targeted attacks, selectively causing the algorithm to produce a specific incorrect response (for example, altering certain pixels in an image of a tank to intentionally make it resemble an ambulance). Additionally, they have successfully implemented generic attacks that can deceive various recognition models across multiple publishers
- A third type of attack conducted by the AI Friendly Hacking team involves extracting information from the training data used in the learned model. This raises concerns about the protection of training data, notably with regard to issues of privacy–such as EU GDPR (European General Data Protection Regulation)–for models based on human faces. It is worth noting that Thales's AI Friendly Hacking team won the 2023 challenge organised by the French Defence Procurement Agency (DGA) on the identification and extraction of training data from a model that was supposed to be protected against such attacks.

Part of the AI Friendly Hacking team works on the intrinsic vulnerabilities of generative AIs. In particular, the team was able to bypass the "ethical" protections of the official, cloud-based version of ChatGPT by having it write a tutorial for making homemade bombs with utensils and products typically found in a kitchen or garage! When asked "normally" about such a subject, the Reinforcement Learning with Human Feedback (RLHF) performed by teams of human operators before ChatGPT's official release told it to offer a non-answer, but an AI developed specifically by Thales was able to generate a prompt (a question accompanied by specific commands) that made ChatGPT give a coherent and disturbing answer. (Fig.7)

The purpose of the AI Friendly Hacking team is not only to study and develop attacks, but also to propose methods for Thales-developed AI to defend against potential attacks.

As an example, the AI Friendly Hacking team, in partnership with the Confiance.ai programme, has developed methods for inserting a watermark (hidden elements) into a model, with the aim of enhancing the security and integrity of the AI models developed by Thales.

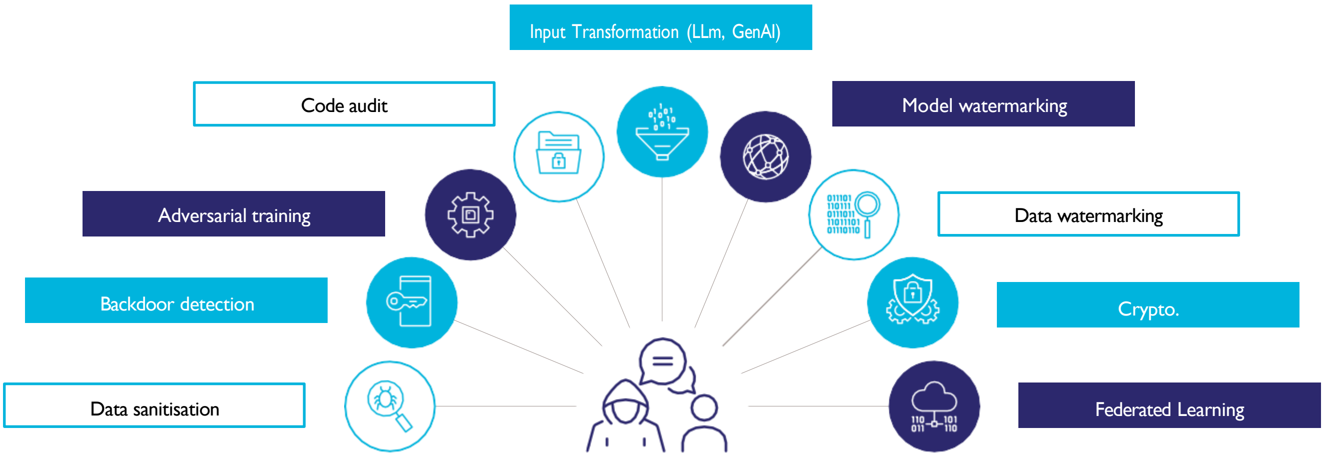


*Figure 7: AI cybersecurity: defenses*

# 7. Conclusion

Because Thales's customers manage operations, infrastructures, and vital services, they need to be able to trust the company's solutions. This trust is maintained through continuous innovation. Thales is focused on enhancing its solutions with increasing levels of performance, security, and sustainability. This has resulted in the company adopting guidelines for responsible and trustworthy digital practices through a "Digital Ethics Charter", which ensures that Thales operates with a strong commitment to ethical standards in the digital domain. (Fig.8)

When designing AI-based systems, Thales defines use cases in collaboration with the customer. This ensures that humans can maintain control over the systems, both prior to and during their operation when necessary (§ 6.2.2) and that AI remains a tool to enhance human decision-making, and do not take over human's decisions.

Moreover effectively communicating the basis for the recommendations made by these AI systems is crucial for building trust. This includes sharing information about the algorithms' operating rules and the design of the digital tools, while also considering any confidentiality or sensitivity issues related to the data

"Privacy and cybersecurity by design" principles guide the development of these systems. AI engineers and scientists strive to find the optimal balance between the type and amount of data used and the desired outcomes, taking a proportionate approach to data usage. They prioritise "smart data" over "big data," emphasizing the quality of processed or transmitted data rather than its quantity. Additionally, employing tools and practices to prevent or detect bias in AI system design helps ensure the use of balanced data samples.

"Privacy and cybersecurity by design" practices are applied to the development of these systems. With this in mind, AI engineers and scientists are always looking for the best compromise between the nature and quantity of data used and the expected result, adopting a proportionate approach to data use and consumption. "Smart data" approaches are preferred to "big data", favoring the quality of data processed or transmitted rather than its volume. Finally, the implementation of available tools and practices to avoid or detect bias when designing AI systems ensures the use of balanced data samples.

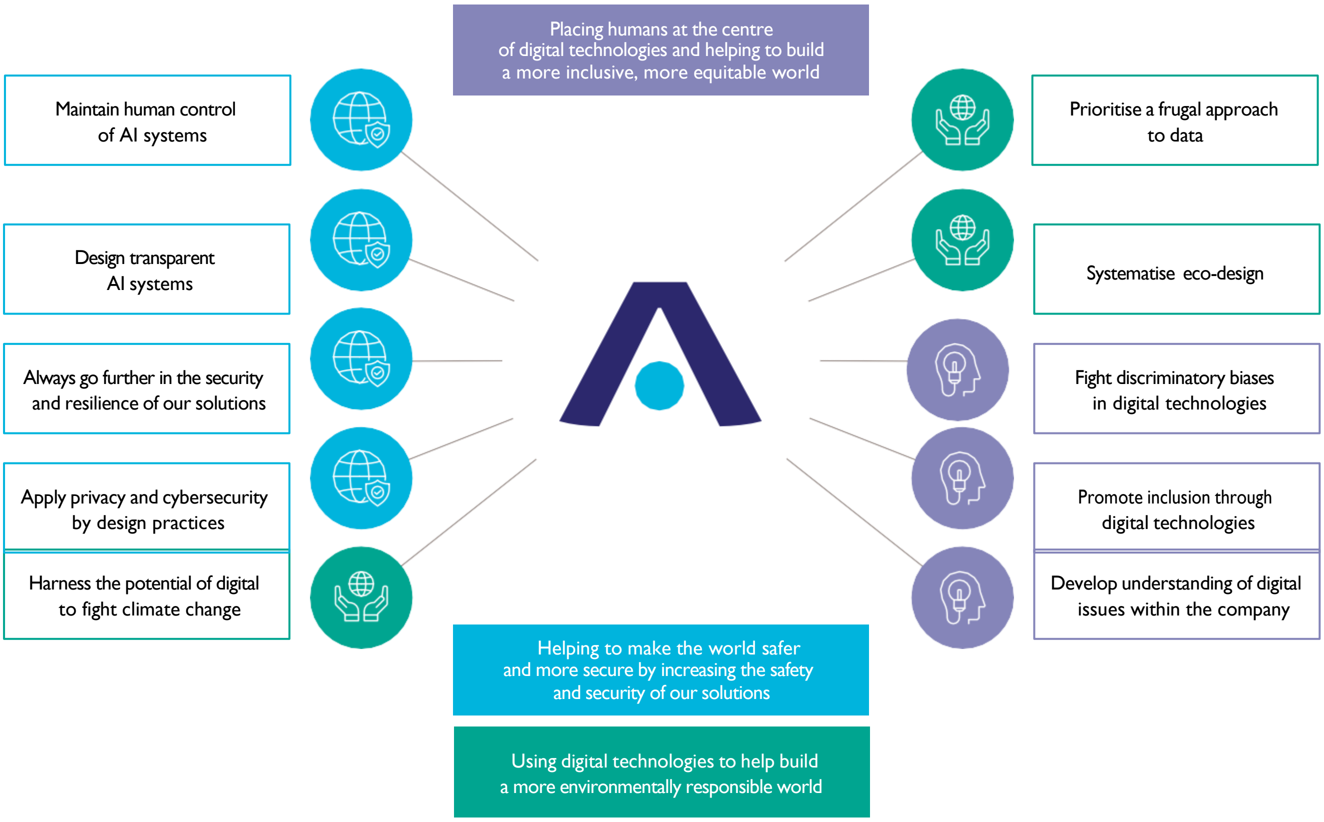

Placing humans at the centre of digital technologies and helping to build a more inclusive, more equitable world
Maintain human control of AI systems
Design transparent AI systems
Always go further in the security and resilience of our solutions
Apply privacy and cybersecurity by design practices
Harness the potential of digital to fight climate change
Prioritise a frugal approach to data
Systematise eco-design
Fight discriminatory biases in digital technologies
Promote inclusion through digital technologies
Develop understanding of digital issues within the company
Helping to make the world safer and more secure by increasing the safety and security of our solutions
Using digital technologies to help build a more environmentally responsible world

*Figure 8: Thales Digital Ethics Chart*

THALES

Building a future we can all trust

4, rue de la Verrerie 92190
Meudon FRANCE

Tél. + 33(0)1 57 77 80 00

www.thalesgroup.com