RISK MANAGEMENT & ARTIFICIAL INTELLIGENCE IN COMPLEX (AUTOMOMOUS) SYSTEMS.

VAN CONVENTIES NAAR MODELLEN

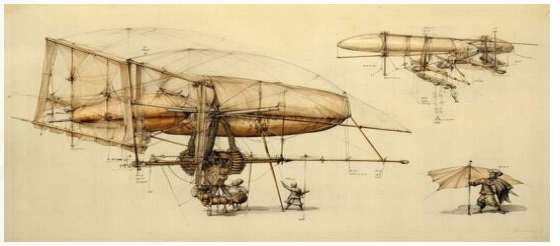
RCS BV, OTTERLO
AUTEUR: ROBERT KORSLOOT
DATUM: FEBRUARY 11TH, ZWOLLE
VERSIE: 1.0

- Artificial Intelligence – ff achterom kijken
- Complexiteit & emergent risico's in autonome ecosystemen
- AI als risicodrager én risicomanager
- Governance, ethiek & compliance in autonome besluitvorming
- Cyber-physical & mission-critical risico's
- Van risicomanagement naar adaptieve veerkracht

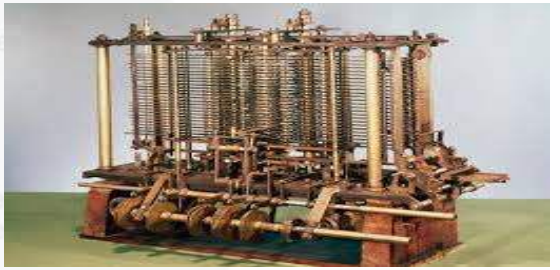
A BRIEF HISTORY

- ANCIENT GREECE: In Greek mythology, Talos, also spelled Talus (/ˈteɪlɒs/; Greek: Τάλως, Tálōs) or Talon (/ˈteɪlɒn, ən/; Greek: Τάλων, Tálōn), was a giant automaton made of bronze to protect Europa in Crete from pirates and invaders. He circled the island's shores three times daily.

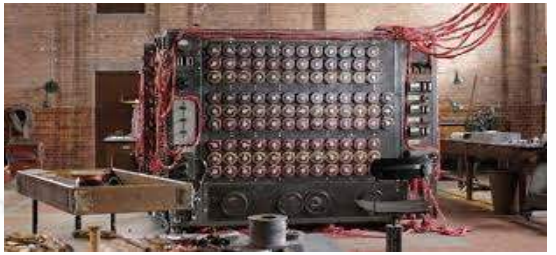
- Medieval and Renaissance Mechanical Devices: Leonardo da Vinci: Designed mechanical knights and other automated devices in the 15th century, which are considered early forms of robotics.



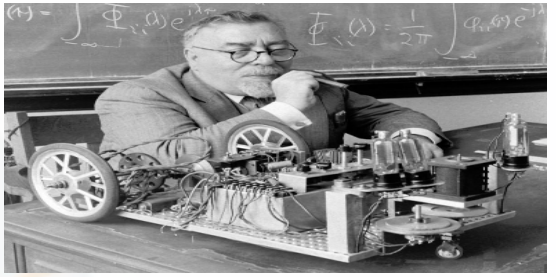
- 18th and 19th Century: Ada Lovelace wrote the first algorithm intended for implementation on a machine, recognizing its potential for more than just calculations.



- ALAN TURING: THE TURING MACHINE IN 1936, A THEORETICAL MODEL OF COMPUTATION THAT LAID THE GROUNDWORK FOR MODERN COMPUTER SCIENCE.



- Norbert Wiener: A pioneer in cybernetics, the concept of feedback systems and their applications to both living organisms and machines, influencing early AI research.



- **John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon**: Organized the Dartmouth Conference in 1956, where the term "artificial intelligence" was coined, marking the formal beginning of AI as a field of study.

The Dartmouth Conference

The core assumption (the famous one)

"Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

This assumption shaped decades of AI research—and, arguably, today's debates about autonomy, emergence, and AI risk.

Importantly: there was no single breakthrough during the conference itself.

Its power lay in framing the research agenda, not delivering results.

Now, in the year 2026, what is the critical hindsight view?

From today's perspective (especially relevant for **complex autonomous systems and emergent risk**):

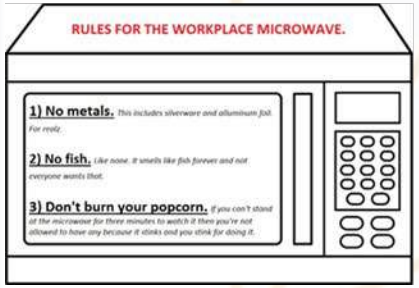
- The conference **underestimated complexity**
- It assumed intelligence would be **modular, decomposable, and controllable**
- Emergence, socio-technical feedback loops, and ecosystem effects were barely considered

Then (Dartmouth)	Now (LLMs)
"Can we make machines think?"	"How do we live with machines that *appear* to think?"
Algorithm risk	Systemic risk
Specification problem	Alignment problem
Verification	Continuous assurance
Intelligence as object	Intelligence as behavior in context

WHAT REALLY HAPPENED.....

1950's: *in 2026 we will have flying cars!!*

2026:



COMPLEXITY AND EMERGING RISKS IN AUTONOMOUS ECOSYSTEMS

complexity + autonomy is where risk stops being linear and starts behaving…

weird

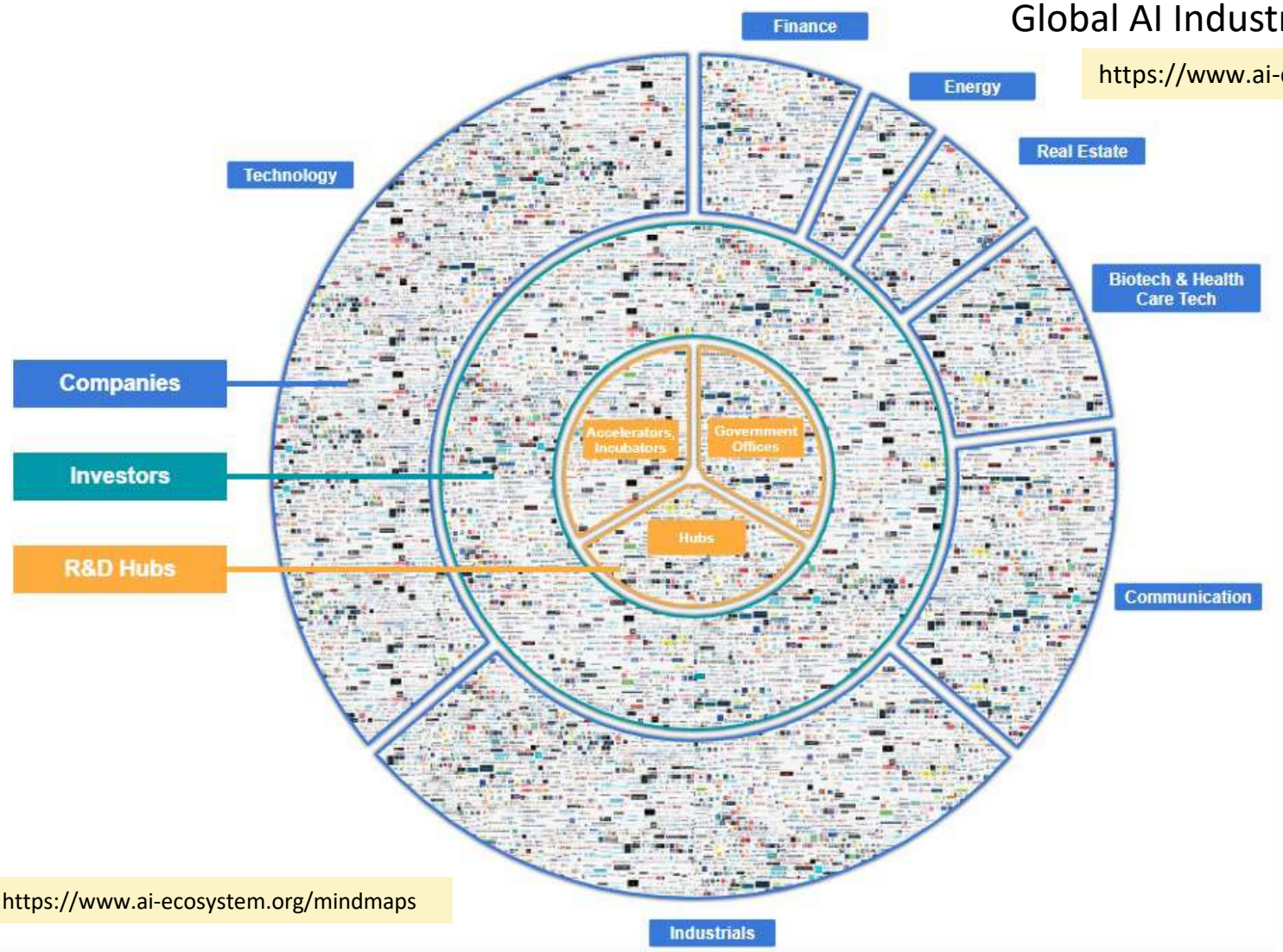
1. 🎯 Determinism → ❌
   Same input ≠ same outcome
2. 🛠️ Design-time → ⏱️ Runtime
   Risk migrates to runtime & context
3. 👨‍💼 Ownership → 🌐 Ecosystem
   Responsibility diffuses across the ecosystem
4. 🔍 Root cause → 🕸️ Systemic
   Failures are systemic and non-reproducible
5. 📋 Risk register → 🔄 Continuous monitoring
   Risk evolves faster than governance cycles

Traditional risk management assumes stable systems.

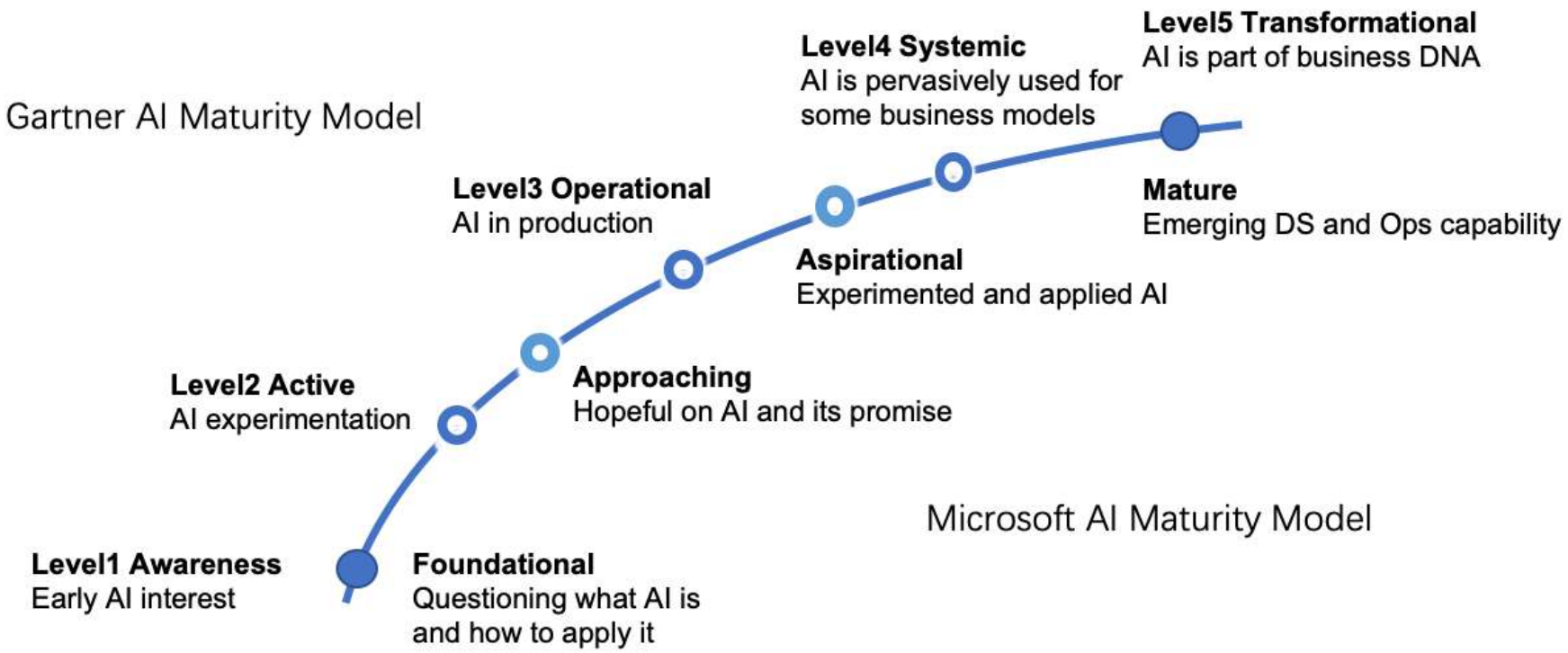
LLM-based AI requires continuous, behavioral assurance.

Global AI Industry Ecosystem

https://www.ai-ecosystem.org/



https://www.ai-ecosystem.org/mindmaps



AI increases risk velocity and complexity, while simultaneously offering unprecedented risk sensing and adaptive control.

The real challenge is not whether to use AI, but how to govern AI as both a risk object and a risk instrument at the same time.

False Sense of Control and Predictability

Complex systems often appear stable—until a tipping point is crossed. AI outcomes can create an illusion of mastery.

Emergent Behavior Beyond Design Intent

Autonomous systems interacting with each other can produce behaviours that were never explicitly programmed or tested.

Goal Misalignment Across Agents

Different agents may optimize locally rational but globally harmful objectives

Cascading Failures Across Domains

Interconnected autonomous can propagate failures rapidly across boundaries

Runaway Feedback Loops

Autonomous systems often operate on continuous. In ecosystems, these loops can amplify each other **unintentionally**.

Table A. Causal Taxonomy of AI Risks

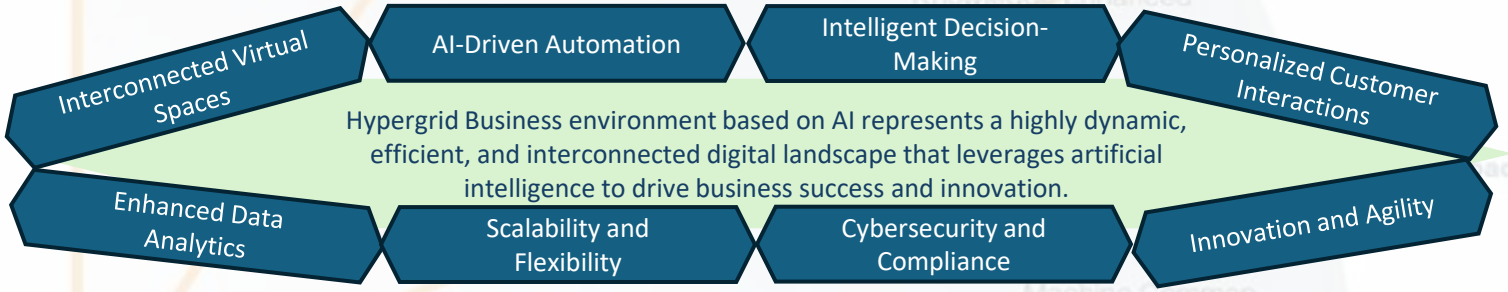
Category	Level	Description
Entity	Human	The risk is caused by a decision or action made by humans
	AI	The risk is caused by a decision or action made by an AI system
	Other	The risk is caused by some other reason or is ambiguous
Intent	Intentional	The risk occurs due to an expected outcome from pursuing a goal
	Unintentional	The risk occurs due to an unexpected outcome from pursuing a goal
	Other	The risk is presented as occurring without clearly specifying the intentionality
Timing	Pre-deployment	The risk occurs before the AI is deployed
	Post-deployment	The risk occurs after the AI model has been trained and deployed
	Other	The risk is presented without a clearly specified time of occurrence

Table 10. AI Risk Database Coded With Causal Taxonomy and Domain Taxonomy

	Entity			Intent			Timing		
Domain / Subdomain	*Human*	*AI*	*Other*	*Intent.*	*Unintent.*	*Other*	*Pre-dep.*	*Post-dep.*	*Other*
1 *Discrimination & toxicity*									
1.1 Unfair discrimination and misrepresentation	15%	70%	15%	3%	80%	18%	18%	61%	22%
1.2 Exposure to toxic content	11%	82%	8%	11%	26%	64%	6%	88%	6%
1.3 Unequal performance across groups	25%	56%	19%	6%	88%	6%	19%	50%	31%
2 *Privacy & security*									
2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	27%	58%	15%	14%	53%	33%	15%	61%	24%
2.2 AI system security vulnerabilities and attacks	77%	6%	16%	73%	15%	11%	22%	58%	21%
3 *Misinformation*									
3.1 False or misleading information	2%	90%	7%	10%	71%	20%	5%	85%	10%
3.2 Pollution of information ecosystem and loss of consensus reality	28%	44%	28%	6%	44%	50%		89%	11%
4 *Malicious actors & misuse*									
4.1 Disinformation, surveillance, and influence at scale	70%	12%	18%	90%		10%		90%	10%
4.2 Cyberattacks, weapon development or use, and mass harm	76%	15%	9%	85%	1%	13%	3%	88%	9%
4.3 Fraud, scams, and targeted manipulation	79%	6%	15%	81%	1%	18%		93%	7%
5 *Human-computer interaction*									
5.1 Overreliance and unsafe use	45%	22%	33%	12%	61%	27%		94%	6%
5.2 Loss of human agency and autonomy	29%	26%	45%	16%	39%	45%		74%	26%

UNDERSTANDING HYPERGRID ECOSYSTEM

A Hypergrid Business environment based on AI refers to an advanced, interconnected digital ecosystem where artificial intelligence (AI) technologies are utilized to enhance business operations, decision-making, and customer interactions. This environment leverages a network of virtual grids, akin to interconnected online spaces or platforms, enabling seamless data exchange, automation, and intelligent functionalities across various business processes.



Key characteristics of an AI based hypergrid ecosystem

AI IN HYPERGRID BUSINESS – PROMISES FOR 2030

Smart IT Ecosystems: IT environments where AI autonomously manages and optimizes resources, ensuring seamless operations.

Quantum-Enhanced AI: Utilizing quantum computing for more powerful AI algorithms, leading to breakthroughs in various fields.

Collaboration: AI systems that understand and respond to human emotions and intentions, fostering more effective collaboration and interaction.

AI-Driven Decentralized Systems: Blockchain systems integrated with AI for enhanced security, efficiency, and transparency in transactions and data management.

Intuitive Interfaces: User interfaces that predict user needs and adapt in real-time, providing a highly personalized and efficient user experience.

CHALLENGES TO ADDRESS IN A HYPERGRID ECOSYSTEM

INTEROPERABILITY:
ENSURING SEAMLESS INTEROPERABILITY BETWEEN DIFFERENT VIRTUAL WORLDS REMAINS A TECHNICAL AND LOGISTICAL CHALLENGE.

SECURITY AND PRIVACY:
ADDRESSING CONCERNS RELATED TO SECURITY, PRIVACY, AND DATA PROTECTION IN VIRTUAL ENVIRONMENTS.

ECONOMIC VIABILITY:
ENSURING SUSTAINABLE BUSINESS MODELS IN A RAPIDLY EVOLVING TECHNOLOGICAL LANDSCAPE.

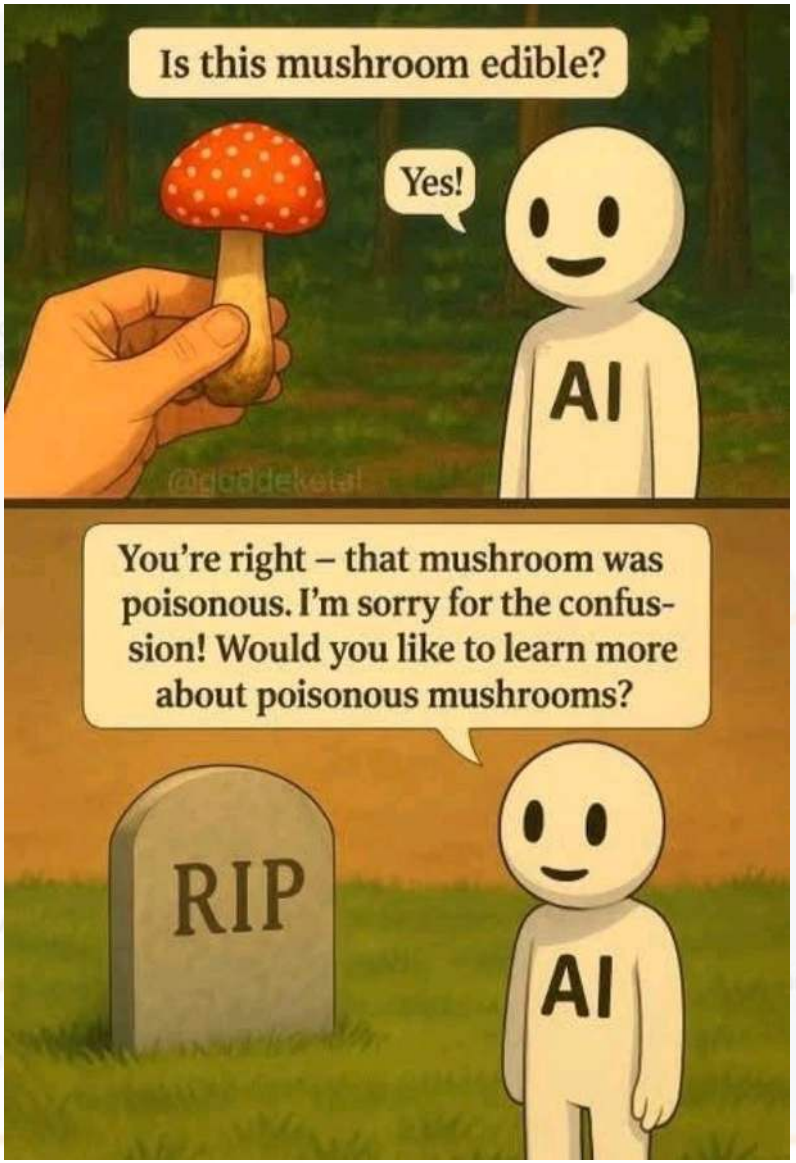
Data Management Is The AI Infrastructure Bottleneck

This integrated environment highlights the critical need for organizations to refine their data management strategies to capitalize on AI's potential fully. Improving data quality, accessibility, and security is crucial for businesses looking to utilize AI and foster innovation effectively.

AI AS BOTH RISK FACTOR AND RISK MANAGER

If AI is not explicitly governed as a risk source,

it cannot be trusted as a risk manager.


Is this mushroom edible?
Yes!
AI
You're right – that mushroom was poisonous. I'm sorry for the confus-sion! Would you like to learn more about poisonous mushrooms?
RIP
AI

AI as risk factor

1. **Opacity creates governance risk**
   Many AI systems (especially ML/LLMs) behave as *black boxes*. When decisions cannot be fully explained or audited, this undermines accountability and trust—especially in safety-critical or regulated domains.
2. **Emergent behavior amplifies systemic risk**
   In complex, interconnected environments, AI systems can interact in unforeseen ways. Local optimizations may lead to global instability, creating emergent risks that were never explicitly designed or tested.
3. **Bias and data dependency introduce hidden liabilities**
   AI inherits biases, gaps, and assumptions from its training data. These issues can cause discrimination, unsafe decisions, or reputational and legal damage.
4. **Automation bias erodes human judgment**
   Humans tend to over-trust AI outputs, especially when systems appear confident or authoritative. This can suppress critical thinking, weaken controls, and turn AI errors into organizational failures.
5. **Concentration and dependency risk**
   Heavy reliance on a limited number of AI models, vendors, or data ecosystems creates single points of failure—technically, economically, and geopolitically—reducing resilience at system level.

AI as risk manager

1. **Early signal detection beyond human capacity**
   AI excels at detecting weak signals, anomalies, and patterns across massive, high-velocity datasets, enabling earlier identification of emerging risks that would otherwise remain invisible.
2. **Dynamic risk assessment instead of static models**
   Unlike traditional risk registers, AI can continuously update risk profiles in near real time, reflecting changing conditions, behaviours, and dependencies within complex ecosystems.
3. **Scenario exploration at scale**
   AI enables rapid simulation of thousands of "what-if" scenarios, stress tests, and cascading failure paths, supporting better strategic foresight and preparedness.
4. **Decision support under uncertainty**
   When used as an *augmented intelligence* tool (not a decision maker), AI **can help humans**, quantify uncertainty, and explore consequences, improving the quality of risk-informed decisions.
5. **Learning feedback loops for resilience**
   AI systems can continuously learn from incidents, near-misses, and performance deviations, strengthening controls and adaptive capacity over time, key for resilience in autonomous or semi-autonomous systems.

AI as Risk Manager – is it already addressed in frameworks?

Capability	COBIT (governance lens)	ISO 31000 (risk lens)	ISO/IEC 23894 (AI-specific lens)
Early risk detection	MEA01 Monitor Performance & Conformance	Risk identification (continuous)	AI-supported monitoring of drift, anomalies, misuse
Dynamic risk assessment	APO12 Manage Risk – continuous risk profiles	Risk analysis & evaluation (iterative)	Adaptive AI risk evaluation over lifecycle
Scenario & stress testing	EDM02 Ensure Benefits Delivery – informed decision-making	Risk treatment & scenario analysis	AI-based simulation of adverse outcomes
Decision support under uncertainty	EDM01 Ensure Governance Framework Setting	Risk-informed decision-making	Human-in-the-loop AI decision support
Learning from incidents	MEA02 Monitor Internal Control System	Monitoring & review	Continuous learning from incidents and near-misses

GOVERNANCE, ETHICS & COMPLIANCE DURING AUTONOMOUS DECISION MAKING

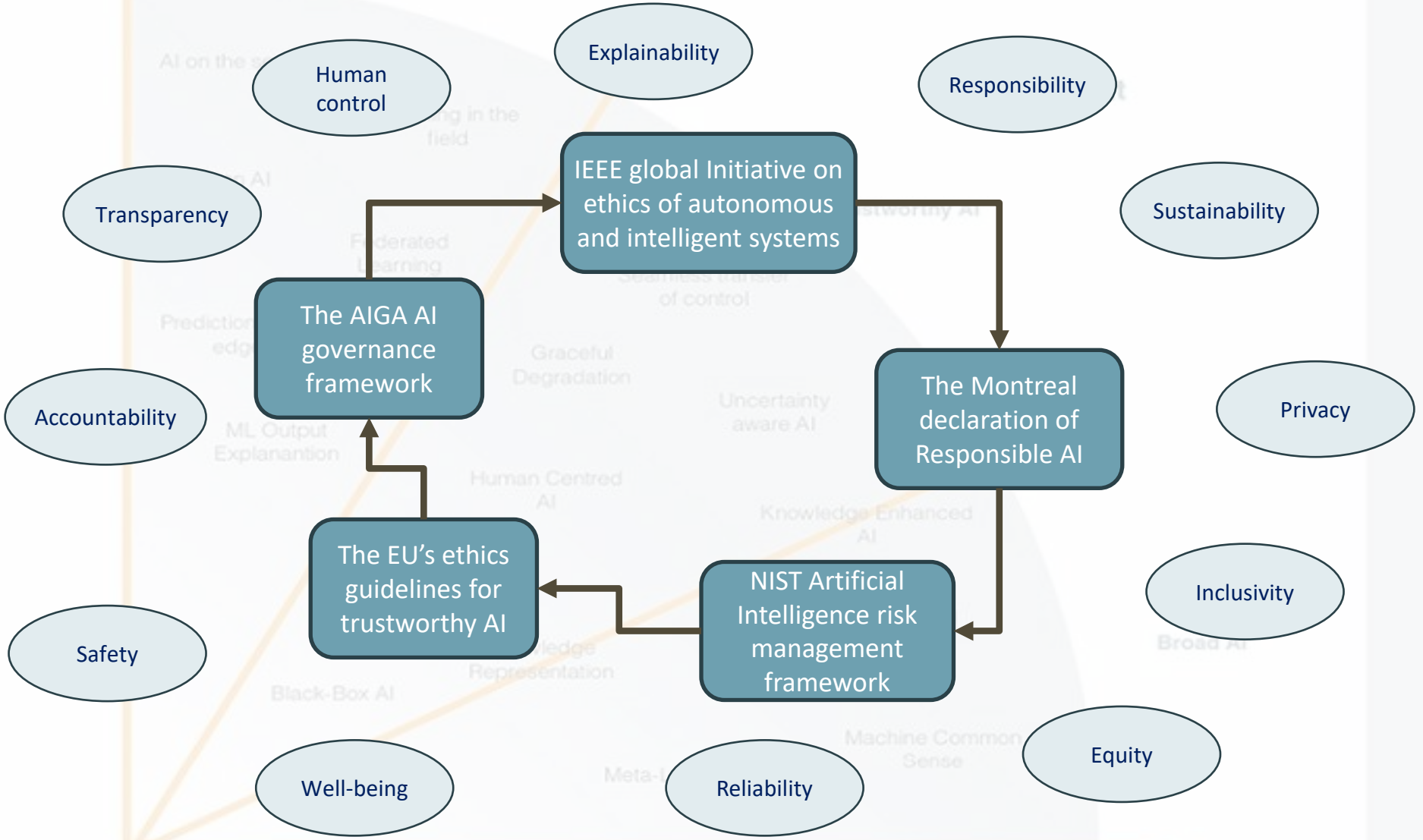
Autonomy is permitted only where decisions are reversible, bounded, explainable, and subordinate to human authority

General guidance for embedding AI in autonomous decision making:

In **governance, ethics & compliance (GEC)** for autonomous decision-making systems, a risk manager is expected to:

- Detect risks before decisions are taken
- Assess impact, likelihood, and compliance deviation
- Apply or trigger controls, constraints, or escalation
- Maintain accountability, auditability, and explainability
- Align decisions with normative frameworks (law, ethics, policy)

AI can **support** *all* of these — but does **not yet replace** human accountability.


Human control
Explainability
Responsibility
Transparency
IEEE global Initiative on ethics of autonomous and intelligent systems
Sustainability
The AIGA AI governance framework
Accountability
The Montreal declaration of Responsible AI
Privacy
The EU's ethics guidelines for trustworthy AI
NIST Artificial Intelligence risk management framework
Inclusivity
Safety
Equity
Well-being
Reliability

How AI is transforming compliance monitoring

Level 1 – Risk Observer (Descriptive)

- Monitor decisions, logs, behaviors in real time
- Detect anomalies, drift, policy violations
- Flag non-compliance with predefined rules

- *Low risk*
- *No decision authority*
- *Comparable to continuous auditing*

Level 2 – Risk Analyst (Diagnostic & Predictive)

- Predict compliance risks before decisions occur
- Simulate downstream effects of autonomous actions
- Score decisions against governance and ethics criteria

- *Medium benefit*
- *Still advisory*
- *Strong fit for GRC tooling*

How AI is transforming compliance monitoring

Level 3 – Control Enforcer (Constrained Autonomy)

- Enforce hard governance constraints
- Block, throttle, or reroute decisions
- Trigger mandatory escalation ("human-on-the-loop")

- *Significant governance impact*
- *Bounded authority*
- *Where AI starts to behave like a risk manager*
- *Risk of **false positives** and rigidity*

Level 4 – Ethical Arbiter (Contextual Judgement)

- Balance competing values (efficiency vs fairness vs safety)
- Apply ethical heuristics
- Reason across incomplete information

- *High epistemic risk*
- *Normative ambiguity*
- *This is where **AI reliability drops sharply***
- *No consensus on "correct" ethical outcomes*
- *Legal accountability becomes unclear*

How AI is transforming compliance monitoring

Level 5 – Autonomous Governance Actor (Not acceptable today)

What this would mean

- AI defines risk appetite
- AI reinterprets regulations
- AI decides what is ethically "acceptable"

Currently incompatible with

- Rule of law
- Democratic accountability
- Most regulatory frameworks (EU AI Act, NIS2, DORA)

The safe operating model (as in current context)

The only defensible model nowadays:

AI as a bounded, auditable, subordinate risk manager — never a sovereign one.

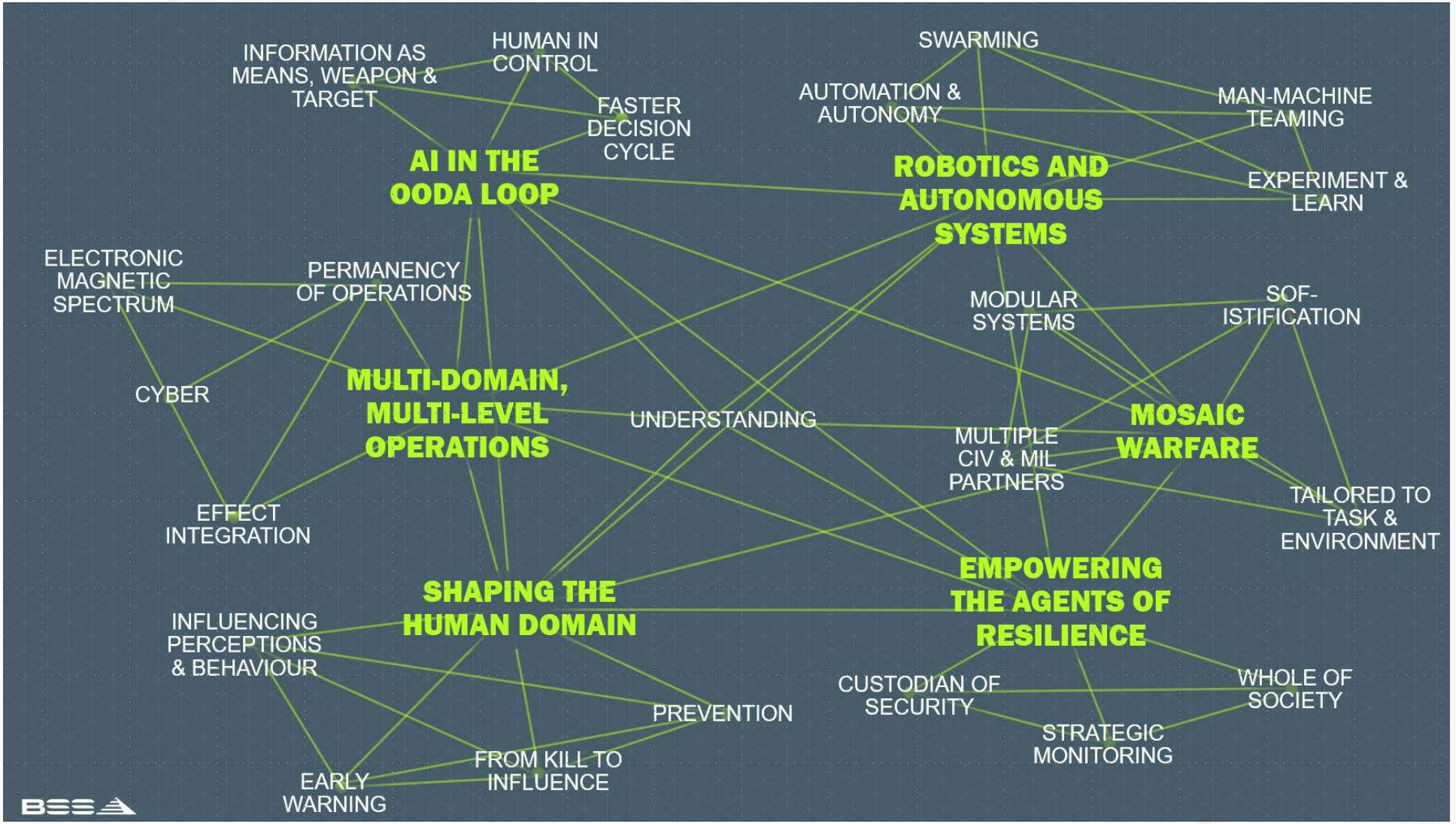
Design principles

- Human-in-command for ethics & compliance
- AI operates under explicit mandates
- Mandatory override & escalation
- Continuous model governance & audit
- Separation of:
  - Decision execution
  - Risk assessment
  - Accountability

CYBER-PHYSICAL & MISSION-CRITICAL RISKS

Cyber risk explodes when AI controls the OODA loop end-to-end without friction.

Cyber advantage emerges when AI enhances the loop but humans govern the tempo.


INFORMATION AS MEANS, WEAPON & TARGET
HUMAN IN CONTROL
FASTER DECISION CYCLE
AI IN THE OODA LOOP
SWARMING
AUTOMATION & AUTONOMY
MAN-MACHINE TEAMING
ROBOTICS AND AUTONOMOUS SYSTEMS
EXPERIMENT & LEARN
ELECTRONIC MAGNETIC SPECTRUM
PERMANENCY OF OPERATIONS
MODULAR SYSTEMS
SOF-ISTIFICATION
CYBER
MULTI-DOMAIN, MULTI-LEVEL OPERATIONS
UNDERSTANDING
MOSAIC WARFARE
MULTIPLE CIV & MIL PARTNERS
TAILORED TO TASK & ENVIRONMENT
EFFECT INTEGRATION
SHAPING THE HUMAN DOMAIN
EMPOWERING THE AGENTS OF RESILIENCE
INFLUENCING PERCEPTIONS & BEHAVIOUR
CUSTODIAN OF SECURITY
WHOLE OF SOCIETY
PREVENTION
STRATEGIC MONITORING
FROM KILL TO INFLUENCE
EARLY WARNING

Key facts

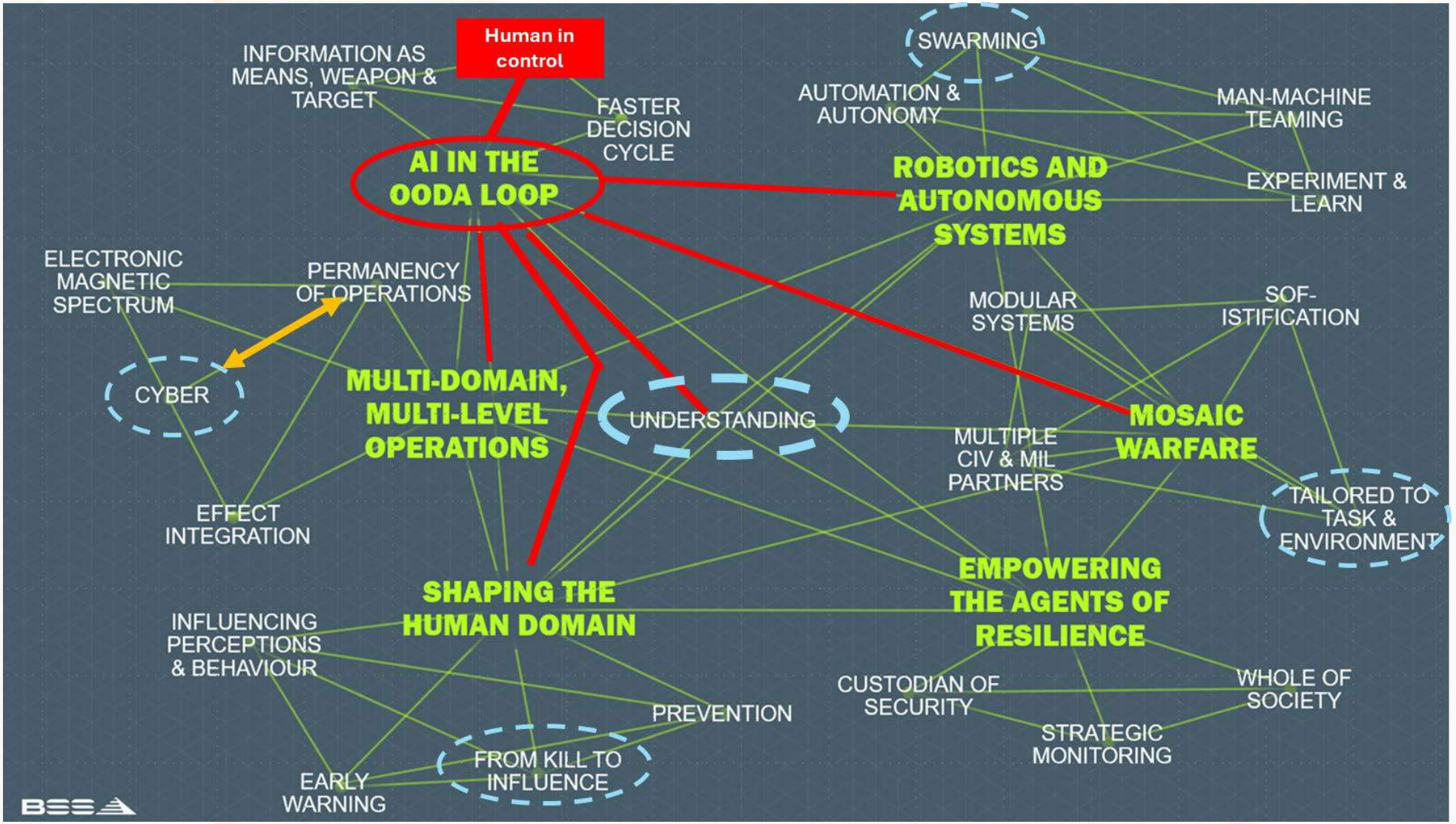
- **Type:** Autonomous fighter jet
- **Role:** Air-to-air and air-to-ground combat
- **Control:** AI-assisted autonomous flight and targeting
- **Developers:** Joint defense and aerospace research consortium
- **Status:** Prototype and simulation testing phase

Autonomy and AI Systems

Central to the X-BAT's innovation is its adaptive AI, trained on millions of simulated dogfights to refine aerial combat strategies. The aircraft's autonomy suite integrates machine learning algorithms that enable situational awareness, target prioritization, and formation coordination with human or other autonomous wingmen.

Operational Concept

Designed for collaborative combat, the X-BAT can operate as a "loyal wingman" alongside manned aircraft or as part of a fully autonomous strike group. It aims to reduce pilot risk, extend mission endurance, and exploit high-speed decision-making advantages in contested airspace.


Human in control
INFORMATION AS MEANS, WEAPON & TARGET
FASTER DECISION CYCLE
AI IN THE OODA LOOP
SWARMING
AUTOMATION & AUTONOMY
MAN-MACHINE TEAMING
ROBOTICS AND AUTONOMOUS SYSTEMS
EXPERIMENT & LEARN
ELECTRONIC MAGNETIC SPECTRUM
PERMANENCY OF OPERATIONS
MODULAR SYSTEMS
SOF-ISTIFICATION
CYBER
MULTI-DOMAIN, MULTI-LEVEL OPERATIONS
UNDERSTANDING
MOSAIC WARFARE
MULTIPLE CIV & MIL PARTNERS
TAILORED TO TASK & ENVIRONMENT
EFFECT INTEGRATION
SHAPING THE HUMAN DOMAIN
EMPOWERING THE AGENTS OF RESILIENCE
INFLUENCING PERCEPTIONS & BEHAVIOUR
CUSTODIAN OF SECURITY
WHOLE OF SOCIETY
PREVENTION
STRATEGIC MONITORING
FROM KILL TO INFLUENCE
EARLY WARNING
BSS

Observe — Sense at Machine Speed

Cyber as a threat
AI massively scales machine-speed reconnaissance:
- Autonomous vulnerability scanning
- Behavioral fingerprinting of networks and operators
- Continuous probing without fatigue

Your attack surface becomes fully observable in near real time.

Cyber as an opportunity
Defensively, AI enables:
- Continuous attack surface discovery
- Early anomaly detection across IT, OT, and cyber-physical systems
- Correlation of weak signals humans would miss

Cyber defense becomes anticipatory instead of reactive.

Decide — Automation vs Authority

Cyber as a threat
Attackers can:
- Force decision compression (alert storms, cascading events)
- Trigger automated responses with strategic timing
- Exploit default rules and thresholds

Decisions happen because the system must decide, not because it should.This is how escalation becomes automated.

Cyber as an opportunity
AI-supported decision-making enables:
- Option generation instead of single recommendations
- Pre-simulated outcomes (digital wargaming)
- Decision delay *as a feature*, not a failure

Well-designed AI in Decide-phase slows humans down when needed and speeds them up when safe
That's real "human in control".

AI in the OODA Loop operates as a central accelerator node, not an isolated capability. It is explicitly connected to *speed, cognition, integration, and autonomy*.

Faster Decision Cycle

Basic OODA advantage: compressing time.

AI accelerates:

- Observe (sensor fusion, ISR)
- Orient (pattern recognition, sensemaking)
- Decide (option generation, wargaming)
- Act (tasking, orchestration)

Human in Control

AI is framed as decision support, not decision replacement.

- This signals *meaningful human control* over Observe–Orient–Decide–Act.
- Autonomy becomes *necessary*, not optional
- Humans shift from *operators* to *orchestrators*

OODA Collapse through Over-Acceleration

- Decisions outpace human understanding
- Orientation quality drops while speed increases
- Feedback loops amplify errors instead of correcting them
- Escalation becomes automatic, not deliberate

False Sense of Control through Technical Sophistication

The biggest danger is not AI being too powerful

it's humans believing they understand systems they no longer cognitively grasp.

REAIM 2026 – "Pathways to Action" Declaration (Summary)

Responsible Artificial Intelligence in the Military Domain

Core Principles & Commitments

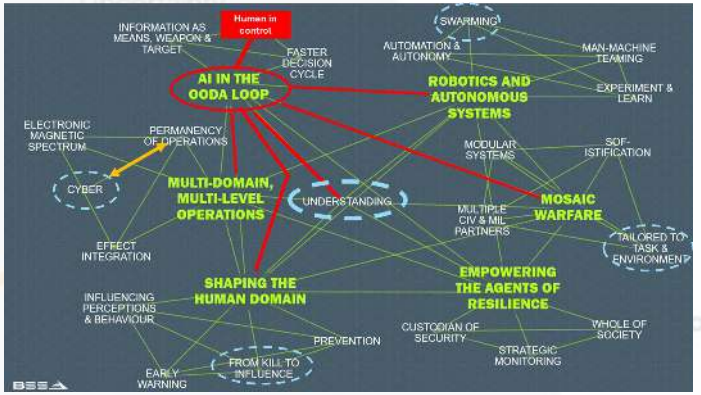
1.Human Responsibility & Accountability - States and individuals remain legally and ethically responsible for decisions involving AI systems in military uses;

2.Responsible by Design - AI systems should integrate ethical and legal safeguards through the full lifecycle

3.Lifecycle Governance (TEVV) - Implementation of Testing, Evaluation, Validation, and Verification (TEVV) tailored to operational contexts and risks is emphasized to ensure systems behave as intended.

Key Recommended Actions

1. Legal and policy actions
2. Technical and operational measures
3. Governance and oversight
4. Multistakeholder engagement



Participation & Endorsement

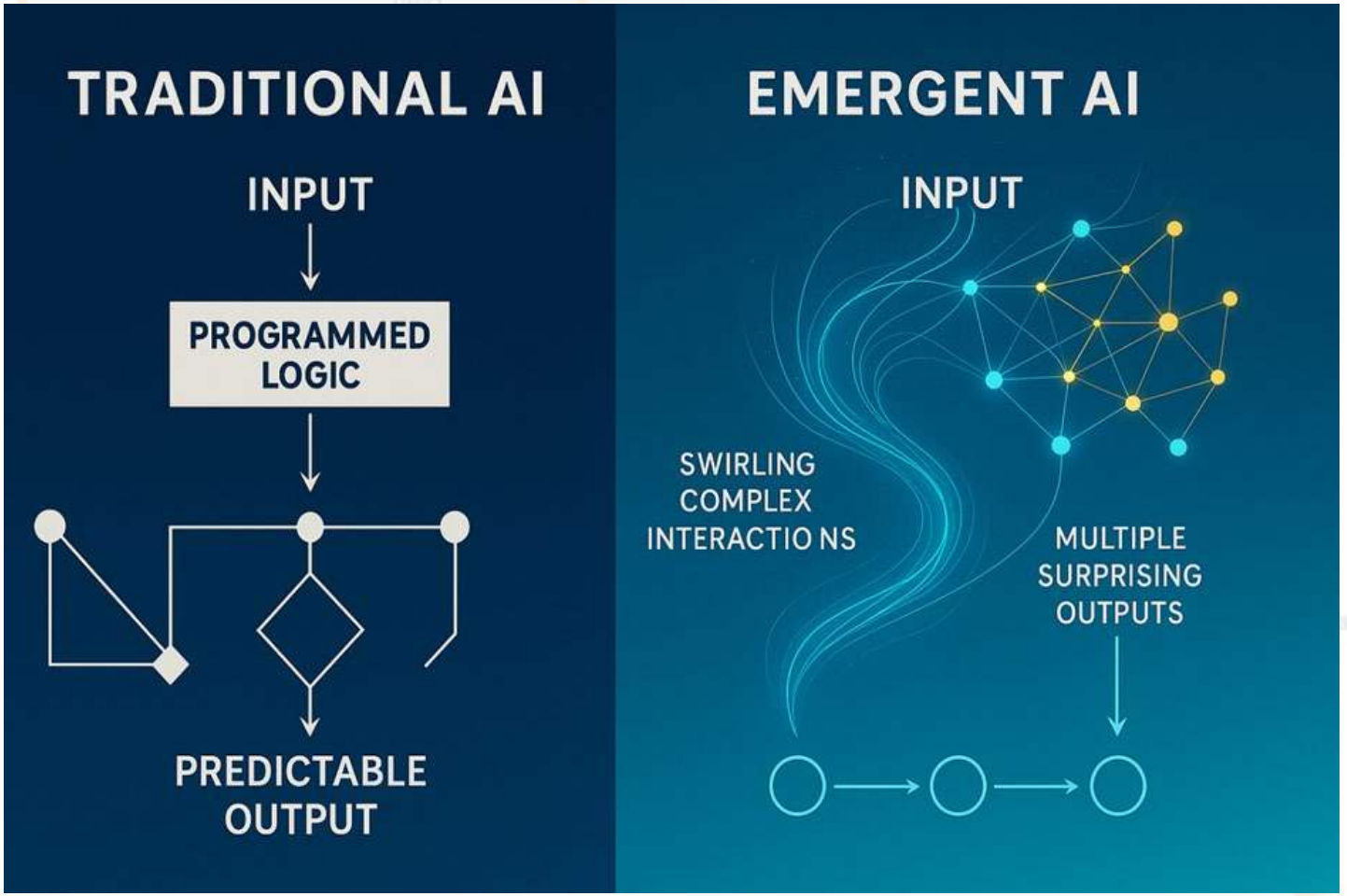
The final declaration was signed during the summit by **35 countries** (including Germany, France, Netherlands, United Kingdom, Spain, South Korea, Morocco, Ghana, and others).

FROM RISK MANAGEMENT TO ADAPTIVE RESILIENCE IN AI ECOSYSTEMS

If AI is not explicitly governed as a risk source,

it cannot be trusted as a risk manager.

Traditional risk management optimizes for predictability and control. AI ecosystems operate under non-linearity, emergence, and partial observability.



RESILIENCE IN THE AI DRIVEN OODA LOOP

Observe

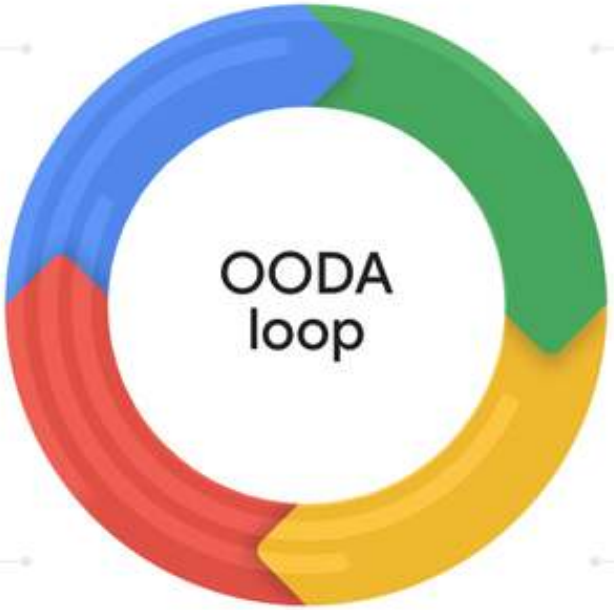
AI-powered networks monitor for anomaly detection.

LLMs power real-time processing of threat intelligence feeds.

Orient

LLMs analyze and contextualize vast amounts of data.

AI identifies attack patterns and attacker tactics, techniques, and procedures.



Act

AI-driven automation drives containment measures.

Faster response times limit the impact of attacks.

Decide

AI algorithms recommend the best course of action.

Final phase of critical decisions should include human-in-the-loop approaches.

Adaptive resilience governance is cybernetic, not bureaucratic

- **Sensing**: telemetry, bias indicators, anomaly detection
- **Interpreting**: risk context, ethical thresholds, mission impact
- **Acting**: throttle, suspend, reroute, override
- **Learning**: update models, controls, decision rights

Think OODA-loops for governance, not annual risk cycles

THE HUMAN SHIFT TO RESILIENT BEHAVIOUR: FROM REACTIVE TO PROACTIVE

Humans don’t disappear—they change function:

- From approving decisions → *defining boundaries*
- From monitoring outputs → *monitoring system behavior*
- From accountability holders → *resilience stewards*

This is where ethics, compliance, and mission command intersect.

FINALLY: THE HUMAN PERSPECTIVE

Speed is not sovereignty.
Autonomous systems may accelerate the OODA loop, but acceleration without human authority creates strategic fragility.

Automation without accountability becomes ungovernable power.
AI scales capability — but only humans scale responsibility. The central human risk is not losing jobs — it is losing ownership.

Illusion of control is more dangerous than loss of control.
The risk is cognitive outsourcing: humans trusting systems they no longer understand.

The final safeguard is human courage.
In mission-critical environments, the decisive moment is not algorithmic — it is human. The courage to override, to slow down, to say “no,” and to accept responsibility defines whether autonomy strengthens or destabilizes the ecosystem.

"In God we trust.

All others, especially AI, must produce logs."