Manifesto IMTAI on Artificial Intelligence

Classification, Risk Assessment and Governance of AI Agents

IMTAI

International Multidisciplinary Task Force on AI Agents Intelligence

"We cannot govern what we do not understand: a shared taxonomy of AI agents is the foundation of effective governance and safe progress."

Lugano, August 27, 2025

Executive Note

This document sets forth the **IMTAI Manifesto** for a safe, transparent and human-oriented ecosystem of *Artificial Intelligence agents*. The Manifesto is built on three pillars: **scientific classification**, **risk assessment** and **governance frameworks**, aligned with the IMTAI mission of connecting hard sciences, engineering, law and the humanities into a single operative framework.

1 Vision

Artificial Intelligence represents the new cognitive infrastructure of society. *AI agents* —capable of perceiving, reasoning and acting in textual, digital and physical domains— are reshaping value chains, decision processes and fundamental rights. IMTAI envisions a future where the **technical power** of agents is inseparable from **responsibility**, **transparency** and **human values**. Our goal is a common language to design, assess and govern AI agents with scientific rigor and social legitimacy.

2 Principles

1. **Classification as foundation**. Governance starts with taxonomy: an agent can be governed only if its *channels of action*, *learning paradigm*, *function*, and *computational scale* are understood.
2. **Risk as metric**. Each class of agent requires metrics spanning technical safety, cybersecurity, legal, ethical, social and environmental domains.
3. **Multidisciplinarity**. No single discipline suffices: IMTAI is a stable bridge across physics, mathematics, engineering, law and the humanities.
4. **Proportionality and accountability**. Rules and oversight must be proportional to the agent's operational capabilities and impact, with a clear *chain of responsibility*.
5. **Technological humanism**. Human dignity, non-discrimination and social cohesion are design constraints, not ethical afterthoughts.

3 IMTAI Taxonomy (Operational Summary)

3.1 By resource access and action channels

- **Class A – Textual Agent**: text I/O only; minimal operational risk.
- **Class B – Digital Agent**: access to APIs/Web/File systems; risk of data manipulation/exfiltration.
- **Class C – Physical Agent**: control of sensors/actuators/IoT/robots; safety risks and real-world damage.

3.2 By learning paradigm

T1 (Supervised), T2 (Unsupervised), T3 (Reinforcement), T4 (Generative), T5 (Self-supervised), T6 (Hybrid/Continual).

3.3 By functional role

F1 (Automation), F2 (Coding), F3 (Research/Analysis), F4 (Control/Orchestration), F5 (Supervision/Validation).

3.4 By model scale and computational demand

Model macro-classes: $\mu, S, M, L, XL, XXL, \Omega$ (from micro to ultra/omega); computational classes: *Instant, Quick, Standard, Demanding.*

Operational note. The IMTAI taxonomy is designed to *map capacities and impact surfaces* of agents, a prerequisite for assessment and policy.

4 Risk Assessment Framework

For each $(Access; Training; Function; Scale)$ combination, IMTAI adopts a 4-dimension risk matrix:

1. **Technical–operational**: robustness, reliability, functional safety, supply-chain integrity.
2. **Cyber & Data**: vulnerabilities, data exfiltration/manipulation, resilience, adversarial attacks.
3. **Ethical–legal**: bias, explainability, traceability, regulatory compliance, IP protection.
4. **Socio–environmental**: impacts on labor, inequalities, energy consumption, ecological footprint.

Severity and mitigation measures scale with autonomy, action surface and coupling with critical systems.

5 Governance Lines (IMTAI Proposals)

5.1 Proportionality and levels

- **Agent level** (by-design): capability bounding, safe tool-use, guardrails, logging.
- **System level** (by-default): sandboxing, principle of least privilege, human-in-the-loop for critical tasks.
- **Organization level** (by-governance): threat modeling, incident response, audits, governance boards.
- **Ecosystem level** (by-policy): registries for high-impact agents, transparent reporting, interoperability standards.

5.2 Accountability & traceability

Model and prompt versioning, training/data attestations, content watermarking, decision logs for high-risk actions.

5.3 Independent evaluation

Red-teaming and third-party audits for classes B–C; safety cases required for XL–Ω and safety-critical domains.

6 IMTAI Commitments

1. **Global Atlas of AI Agents**: a living repository of taxonomies, benchmarks and risk profiles.
2. **Assessment Framework**: comparable checklists and metrics across A/B/C × T1–T6 × F1–F5 × scales.
3. **Policy Guidelines**: operational recommendations for governments, international bodies and industry.
4. **Education & Outreach**: programs for schools, administrations, enterprises; public literacy to reduce the cognitive divide.
5. **International Cooperation**: permanent fora to harmonize standards, audits and reporting.

Appendix — Reference Tables

A. Classes by resource access

Class	Channels & Examples	Main risks
A — Textual	Text I/O only; LLM, chatbot	Leaks, hallucinations
B — Digital	API/Web/FS; assistants with plugins	Data manipulation / exfiltration, unintended actions
C — Physical	IoT/robotics; autonomous AI	Safety, ph. damage, sabotage

B. Classes by learning paradigm

Class	Paradigm	Examples
T1	Supervised	Classifiers, OCR
T2	Unsupervised	Clustering, autoencoder, anomaly det.
T3	Reinforcement	Robotics, games, control
T4	Generative	Text/image/video generation
T5	Self-supervised	LLM/VLM pretraining
T6	Hybrid/Continual	Adaptive, lifelong systems

C. Classes by function

Class	Function	Examples
F1	Automation	RPA, workflow automation
F2	Coding	Code generation/review
F3	Research/Analysis	Knowledge extraction, synthesis, reports
F4	Control/Orchestration	Multi-agent supervision
F5	Supervision/Validation	Auditing, quality assurance

D. Model scale and computational demand

Macro-class	Parameters (order)	Computational class (time)
μ	$< 10^6$	Instant (<1s)
S	10^6–10^8	Quick (1–10s)
M	10^8–10^{10}	Standard (10–60s)
L	10^{10}–10^{11}	Demanding (>1m/batch)
XL	10^{11}–10^{12}	Demanding+ (minutes/hours)
XXL	10^{12}–10^{13}	Demanding++ (hours/days)
Ω	$> 10^{13}$	Super-Demanding (days/weeks)

This Manifesto is developed within the IMTAI mission and its proposal for an international task force on AI agent classification and governance.