IMTAI AI Manifesto

Classification, Risk Assessment and Governance of AI Agents

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."*

IMTAI
International
Multidisciplinary
Task Force on AI

Lugano, August 27, 2025

www.imtai.org | info@imtai.org

IMTAI AI Manifesto

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.

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 bridges 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.

Taxonomy Overview

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.

By Learning Paradigm: T1 (Supervised), T2 (Unsupervised), T3 (Reinforcement), T4 (Generative), T5 (Self-supervised), T6 (Hybrid/Continual).
By Functional Role: F1 (Automation), F2 (Coding), F3 (Research/Analysis), F4 (Control/Orchestration), F5 (Supervision/Validation).
By Model Scale:

μ ($<10^6$ params), S (10^6–10^8), M (10^8–10^{10}), L (10^{10}–10^{11}), XL (10^{11}–10^{12}), XXL (10^{12}–10^{13}), Ω ($>10^{13}$).

Risk Assessment and Governance

For each taxonomy combination, IMTAI adopts a 4-dimension risk matrix: Technical–operational, Cyber & Data, Ethical–legal, Socio–environmental. Governance includes proportionality at agent, system, organization, and ecosystem levels; accountability via versioning and logging; and independent evaluations through red-teaming and audits.

Commitments

- **Global Atlas of AI Agents**: A living repository of taxonomies, benchmarks, and risk profiles.
- **Assessment Framework**: Comparable checklists and metrics across classes.
- **Policy Guidelines**: Operational recommendations for governments, bodies, and industry.
- **Education & Outreach**: Programs for literacy and to reduce the cognitive divide.
- **International Cooperation**: Fora to harmonize standards, audits, and reporting.