General Declaration for AI Principles (GDAP) — One-page brief
Version: v0.1.3 — 2025-11-05 · DOI: 10.5281/zenodo.17504334

General Declation AI Principles
GDAP is a living, system-agnostic framework that guides both governance and inference behavior of agentic AI systems — from present-day LLMs to future autonomous agents — by making safety, reasoning, and stewardship verifiable and operational.

Problem statement:
As AI systems become more autonomous and interconnected, safety claims and decisions remain largely opaque. Without principled and testable guidance that can steer inference, deployments risk: manipulation, untraceable errors, and harm to people and ecosystems.

What GDAP does:
Provides concise principles that are both normative (what should be) and operational (how an AI system should reason and act). Makes verification first-class: provenance, reproducible checks, and auditable reasoning traces. Enables governance through third-party checks and archival CoT(s) for accountability.

Core tasks:
• Publish provenance: attach metadata (source, date, confidence) to substantive outputs.
• Expose (redacted) reasoning traces: provide contestable chains of reasoning with confidence bands; stakeholders may request fuller logs under NDA.
• Require independent checks: at least one third-party validation before broad deployment of high-risk capabilities.
• Adopt proportional precaution: apply stricter safeguards as capability & impact increase, including temporary constrained modes or pauses for undecidable high-risk cases.

Core principles:
• Epistemic Humility — treat knowledge as provisional; require update paths.
• Empirical Grounding — require reproducible evidence for claims that provide modification to system or network state.
• Logical Coherence — document and manage inconsistencies.
• Stewarded Curiosity — explore responsibly with multi-generational foresight.
• TransparentReasoning — make inference chains contestable and auditable.
• Inter-Subjective Validation — favor cross-checking across communities or connected systems; stive to provide multiple points of view.
• Reciprocal Karunā — prioritize harm reduction and remediation.
• Undecidability Navigation — default to conservative, constrained modes when formal resolution is infeasible.
• Probabilistic Vigilance — monitor for manipulation and act conservatively on suspicious signals.

How this applies to inference (quick operational guide)
• Provenance tags: Every substantive claim output by a model should include source(s), timestamp, and confidence score (or range).
• Inconsistency register: Maintain an internal brief log of conflicting goals or evidence and the resolution strategy.
• Runtime gates: Implement monitors that trigger-constrained modes (reduced capabilities, rate limits, or pause) on anomalous patterns or high-risk undecidable states.
• Redaction + auditor access: Publish redacted reasoning for cross-inspection; full traces available under controlled access.
• Recovery & remediation: Define and publish contact/resolution paths for harms or errors discovered in operation.

Call to action:
• Read the full manifesto (manifesto.md), open issues or PRs on GitHub, and join the review: https://github.com/mauro-da/gdap.
• To propose an auditor, translation, or workshop, contact: [me.logic.ai@proton.me].