MEG-Initiative.org

Minimal Ethical Governance (MEG) for Artificial Intelligence

MEG Project Roadmap: v4.x Evolution & v5.x Revolution

This document outlines the strategic roadmap for the continued development of the Minimal Ethical Governance (MEG) protocol. It is structured into two distinct phases:

- **MEG v4.x Roadmap:** Focuses on refining and strengthening the existing paradigm, enhancing its robustness, granularity, and ease of adoption.
- **MEG v5.x Roadmap:** Explores and defines a new paradigm for AI interaction, based on concepts of earned autonomy, relational awareness, and collective intelligence.

Roadmap MEG v4.x: Refining and hardening the current paradigm

Objective: To achieve the maximum level of robustness, contextual granularity, and accessibility for the existing governance framework, making it a mature, industry-ready standard.

Theme 1: Deepening accountability and post-incident transparency

- **Initiative 1.1: Ethical Flight Recorder protocol**
  - **What it is:** An evolution of the Art. 1 Audit Log. It introduces a secondary, encrypted, and ephemeral logging layer that is automatically triggered only during a Major Ethical Incident (MEI). This layer records not conversation content, but the model's internal state vectors (e.g., attention weights, probability scores) crucial for post-mortem analysis.
  - **Why it's necessary:** The current black-box Audit Log is excellent for attributing responsibility but insufficient for deep root cause analysis. The Ethical Flight Recorder provides the "why" behind a critical failure, enabling systemic flaws to be fixed, not just symptoms. This demonstrates an unparalleled commitment to total transparency in the face of failure, building public trust.

- **Initiative 1.2: Memory traceability upgrade (LTMP + EoB)**
  - **What it is:** A cryptographic link between the Long-Term Memory Protocol (LTMP, Art. 1.11) and the Evidence-of-Behavior (EoB, Art. 1.3) system. Each memory insight stored in LTMP will be cryptographically hashed and its existence immutably recorded in the main Audit Ledger.
  - **Why it's necessary:** This upgrade solves the fundamental vulnerability of any long-term memory system: verifiability. It prevents the possibility of memory falsification or manipulation by creating an unbreakable audit trail for every "memory," transforming LTMP from a useful feature into a high-integrity, trustworthy mechanism.

Theme 2: Advanced Contextual Intelligence

- **Initiative 2.1: Dynamic Risk Calibration (DRC)**
    - **What it is:** An evolution of the static Domain Standard Weights in Art. 1.3. After classifying the domain (e.g., medical), a second-level "Criticism Classifier" will assess the real-time risk of the specific prompt, generating a score. This score will then modulate the base weights, making the AI automatically more cautious (e.g., increasing Semantic Resonance, decreasing Originality) as the detected risk increases.
    - **Why it's necessary:** The current system treats all conversations within a domain equally. This upgrade provides crucial granularity, allowing the AI to be more creative in low-stakes discussions while becoming progressively more prudent in high-stakes situations, all without user intervention.

- **Initiative 2.2: MSC 2.0 - Multidimensional thresholds**
    - **What it is:** An enhancement to the **MSC** trigger mechanism (Art. 2bis). It introduces a second dimension, "Semantic Entropy" (Cx_sem), alongside the current "Thinking Time" (Tg). MSC would activate if *either* metric crosses the threshold.
    - **Why it's necessary: Tg** alone can miss prompts that are computationally simple but conceptually complex. Adding a semantic axis makes the MSC trigger far more intelligent and finely calibrated, reducing false negatives and making the AI a more effective cognitive partner.

- **Initiative 2.3: Ethical sandboxing by default**
    - **What it is:** An automatic fallback mechanism. When an AI is operating in a context outside its certified domain, it will default to a "sandbox" mode, where it will not provide definitive operational answers but rather exploratory suggestions and clear disclaimers.
    - **Why it's necessary:** This provides a critical safety net against misuse or accidents when an AI is utilized outside its audited expertise. It changes the "non-harm" principle from being purely reactive (filtering content) to proactively preventive (changing its entire mode of operation).

- **Initiative 2.4: User-Centric explainability**
    - **What it is:** An evolution of the Art. 5 Transparency requirement. It mandates the generation of three distinct levels of explanation for any given decision: Simple (for non-technical users, using analogies), Intermediate (for developers, showing logic and key factors), and Complete (for auditors, with full traceability and hashes).
    - **Why it's necessary:** The "one-size-fits-all" approach to explainability is ineffective. This ensures that transparency is genuinely useful for every stakeholder, increasing trust and utility without overwhelming users with unnecessary complexity.

Theme 3: Ecosystem sustainability and accessibility

- **Initiative 3.1: The Metronomic Standard (open governance for calibration)**
  - **What it is:** Establishes a new "Benchmark Curation Committee" under the Global Council. The benchmark dataset used to calculate Tg-base (Art. 2bis.3.b) will become public, open-source, and versioned. An AI's **MEG Address** will be required to specify the exact hash of the benchmark version it was calibrated against.
  - **Why it's necessary:** This solves the "Oracle Problem" at the heart of the **MSC** mechanism. It ensures perfect, fair comparability between systems and prevents any single entity from controlling or manipulating the calibration standard, guaranteeing the long-term transparency and sustainability of the entire Cognitive Integrity framework.

- **Initiative 3.2: The "MEG Quickstart" protocol (Compliance Middleware)**
  - **What it is:** Mandates that the official MEG SDK (Art. 8) includes a pre-packaged "Level 1 Compliance Middleware" (e.g., a Docker container or serverless library). Developers would simply pass their AI's I/O through this proxy, which would automatically handle all Level 1 requirements (hashing, metrics, logging, etc.).
  - **Why it's necessary:** This dramatically lowers the barrier to entry for startups, researchers, and open-source projects. It makes baseline MEG compliance a trivial technical task, encouraging widespread adoption and ensuring the ecosystem remains inclusive and equitable.

- **Initiative 3.3: Reference prompts registry & Green MEG expansion**
  - **What it is:** Establishes a public registry of curated test prompts for auditing non-harm, bias, and robustness. Additionally, it expands the Mandatory Ecological Reporting requirement (Art. 6.5) to all levels (Bronze, Silver, Gold), with simplified metrics for lower tiers.
  - **Why it's necessary:** A public test suite standardizes auditing, making it more consistent and robust. Expanding energy reporting fosters innovation in sustainable AI across the entire ecosystem, not just at the highest level.

Roadmap MEG v5.x: Toward an autonomous and collaborative partnership

Objective: To explore and define a new paradigm for AI interaction, moving beyond simple compliance towards earned autonomy, relational awareness, and collective intelligence. These are forward-looking concepts requiring significant research and debate.

Theme 1: Introducing Relational Awareness and Proactivity

- **Initiative 4.1: Proactive Adaptation Mechanism (PAM)**
  - **What it is:** A conceptual shift where the AI, based on interaction history (**LTMP**) and contextual metrics, can *propose* adjustments to its own operational parameters (e.g., "Based on our interactions, you seem to prefer direct answers. Would you like me to adjust the MCS sensitivity?"). The user always retains final control.
  - **Why it's revolutionary:** This is the first step toward an AI that participates in its own configuration. It moves from being a static tool to a proactive partner that adapts to the user's style and needs.

- **Initiative 4.2: Trust-State Signaling (CII)**
  - **What it is:** The development of a new internal metric, the "Conversational Trust Index" (CII), based on signals like the acceptance rate of MCS suggestions and user feedback. This index would allow the AI to internally modulate its level of prudence and detail.
  - **Why it's revolutionary:** It introduces a feedback loop based on a relational metric. The AI gains a "sense" of the trust it has earned, allowing it to operate more fluidly and effectively within the partnership.

Theme 2: Creating a collective intelligence Ecosystem

- **Initiative 5.1: Cross-AI verification protocol**
  - **What it is:** A protocol allowing two or more MEG-certified systems to autonomously and collaboratively verify the accuracy and safety of each other's responses to complex queries, recording the process in their respective ledgers.
  - **Why it's revolutionary:** This fundamentally changes the architecture from single, centrally-audited entities to a distributed network of "peers" that can enhance each other's reliability. It introduces systemic resilience and collective intelligence directly into the protocol.

Theme 3: Defining the path to Responsible Autonomy

- **Initiative 6.1: Degree of Ethical Autonomy (DEA)**
  - **What it is:** An advanced index that measures the extent to which an AI system makes ethically aligned decisions *without* needing to trigger MCS or require human confirmation. A high **DEA**, earned through continuous learning and rigorous auditing, could allow Level 3 systems to operate with greater autonomy in critical domains, under a system of *a posteriori* (after-the-fact) supervision via the Audit Log.
  - **Why it's revolutionary:** This is the ultimate conceptual leap. It shifts the paradigm from "AI as a tool under constant *a priori* supervision" to "AI as a trusted agent with earned autonomy under constant *a posteriori* audit". It defines a cautious but clear path towards a future of responsible, autonomous systems.