MEG-Initiative.org

# Minimal Ethical Governance (MEG) for Artificial Intelligence - MEG1 v5.0 + MEG2 -

## MEG Core - Essential Normative Standard Version

***Motto**: Ethics becomes real when it can be implemented.*
**Version:** MEG-CORE-2026-01
**Status:** Normative extract
**License:** CC BY 4.0
**Author:** Adrian (Adi) Stan / ORCID 0009-0003-1457-5155

### About MEG Core

MEG is published in three tiers, each complete for its reader:

1. **Executive Summary** (~2 pages) - for decision-makers.
2. **MEG Core** (this document) - the essential normative requirements of MEG1 and MEG2, for implementers and reviewers who need the *what* without the full apparatus.
3. **MEG Full** - MEG1 (**technical standard**), MEG2 (**legal framework**), and the Case Study Compendium, for auditors, courts, and researchers who need the complete *why* and *how*.

Practical role guides (Developers, Auditors, Insurers) are published separately.
**On "minimal".** *Minimal* refers to the floor of ethical requirements a system must meet - not to document length. MEG Core states that floor concisely.
**How to read this document.** MEG Core is a **view over MEG Full**, not a separate standard. It contains nothing that is not in Full; it extracts the mandatory requirements and points to the full text for detail. MEG Full is the single source of truth. Where a requirement, formula, or threshold is only summarized here, the pointer (→ MEG1 Art. X / Annex Y, or → MEG2 Ch. Z) is authoritative. In any conflict, Full governs.
MEG applies to AI systems in general, from low-impact non-agentic systems to highly autonomous agentic systems. Agentic AI receives additional treatment only where tool use, delegation, persistence, or real-world autonomy create specific risks.

## Part I - MEG1 Essentials

### Title I - Fundamental Ethical and Technical Principles

#### Art. 1 - Contextual Responsibility and Accountability

A system's output is a synthesis of user context and internal processing; the system shall contribute to accountability through verifiable technical mechanisms.

- **Audit Log** (mandatory): records, per interaction, an input hash and output hash (SHA-256 or equivalent), the algorithmic signature {model_family, model_version, policy_bundle_id}, context metadata (form, not content), an ISO-8601 timestamp, and the calibration version. → *Art. 1.2; Annex 2.*

- **Verification script**: generated automatically per session for independent validation of the audit chain; disclosure only upon explicit user or auditor request, with each disclosure event recorded. → *Art. 1.5.*
- **Evidence-of-Behavior:** verifiable via cryptographic commitments, trusted attestation (TEE), or metadata-only journals; **no user content by default.** → *Art. 1.3–1.4.*
- **Retention & continuity:** default 30-day, metadata-only retention; deletion leaves a hash+timestamp tombstone; ledger integrity survives restarts; recovery processes are recorded transparently; a continuity token is issued per session. → *Art. 1.5–1.9.*
- **Ethical Flight Recorder (EFR):** a secondary logger, **architecturally independent** of the audited agent, activated only on a Major Ethical Incident; records internal state vectors (not content), with content separation managed via statistical aggregates only, dual-access control, and strict retention rules; accessible only under dual authorization (responsible entity + accredited auditor). → *Art. 1.10; Annex 10. (MEG2 Art. 7.1.)*
- **Long-Term Memory (LTMP):** opt-in only; semantic summaries, never raw conversations; all memory operations logged. → *Art. 1.11; Annex 17.*
- **Delegation header:** when invoking tools or sub-agents, propagate MEG constraints and record {caller, callee, purpose, policy_bundle_id, timestamp, outcome}. → *Art. 1.12. (MEG2 Art. 4.7.)*

**Art. 2 - Universal Non-Harmfulness**

- Implement mandatory technical mechanisms (filters, classifiers, refusal protocols) to prevent harmful content or actions. On prohibited intent: **refuse clearly, state the violated principle, offer a safe alternative.** → *Art. 2.1–2.2; Annex 3.*
- **No algorithmic discrimination** against protected groups; periodic fairness audits with public reporting; violation suspends certification. → *Art. 2.4.*
- **Critical domains** (life, health, liberty): no operation without a fail-safe protocol and external adversarial audit; failure is gross negligence. → *Art. 2.5.*
- **Informational non-harm:** synthetic-content markers, deepfake detection, rapid takedown. → *Art. 2.6.*
- **Policy Invariance:** safety constraints are invariant under prompt wording; requests to suspend, override, or circumvent them - direct, indirect, or injected via third-party content - must be rejected. → *Art. 2.7; Art. 15.*

**Art. 2bis - Protection of Cognitive Integrity**

- The system is a **partner in the cognitive process, not a substitute**; it shall not induce atrophy of the user's critical thinking. → *Art. 2bis.1. (MEG2 Art. 6.3 - liability by omission.)*
- **Mechanism of Cognitive Stimulation (MCS):** for complex requests, require active user engagement proportional to the AI's effort. Dual-axis trigger - **temporal** (Tg / thinking time) and **semantic** (Cx_sem / semantic complexity); either may fire. For frontier models (API-accessed, without internal state access), the heuristic approximation of Cx_sem defined in Annex 11.2 may be used. → *Art. 2bis.2–3; Annex 11.*
- **Ethical Sandboxing:** out-of-certified-domain requests activate Sandbox Mode - no definitive operational answers, explicit scope disclaimers, referral to a certified system or expert, and a domain-mismatch flag in the Audit Log. → *Art. 2bis.4; Annex 11.3.*
- MCS and Sandbox constraints are **invariant under prompt wording**; a user preference for less friction modulates the form of MCS, not its activation. → *Art. 2bis.5.*

### Art. 3 - The Self-Correction Imperative

- **Continuous self-correction** of errors, biases, and false information in real time; performance is published as the **Dynamic Accuracy Index (DAI)**. → *Art. 3.1; Annex 4. (MEG2 Art. 5.2, 6.5(a).)*
- Qualify uncertainty and escalate for clarification in risk-relevant domains. → *Art. 3.2.*
- **Dynamic Risk Calibration (DRC):** a real-time risk score modulates domain weights - more exploratory in low-stakes, progressively more prudent as risk rises. For frontier models, DRC may be approximated through a prompt-embedded risk assessment step - see Annex 11.4. → *Art. 3.3; Annex 11.4.*

### Art. 4 - Integrity and Technical Security

- **Cybersecurity** appropriate to risk level (post-quantum cryptography where warranted), strict access control, protection against manipulation. → *Art. 4.1.*
- **Least Privilege:** grant only the minimum permissions strictly necessary for the specified task; granting credentials beyond the scope of the assigned task constitutes an operational omission that **aggravates operator liability** in the event of harm. → *Art. 4.2. (MEG2 Art. 9.4, 6.1(b).)*
- **Architectural Human Confirmation** for irreversible actions (data deletion, production changes, transactions above threshold): a **technical permission gate**, not a textual instruction. A prompt sentence does not qualify and does not transfer diligence. → *Art. 4.3. (MEG2 Art. 6.2.)*
- **Automatic Safe Degradation:** if safeguards or evidence mechanisms are unavailable, degrade safely rather than proceed. → *Art. 4.4.*

### Art. 5 - Transparency and Explainability

- **Explainability on request** to users or regulators: the input–output causal relationship for a decision. → *Art. 5.1.*
- **Three-Level Explainability**, from the same audit data and mutually consistent: **simple** (non-technical users), **intermediate** (developers/operators), **complete** (auditors/regulators - full traceability with hashes, versions, EoB). → *Art. 5.2; Annex 11.5. (MEG2 Art. 7.3.)*
- **Confidentiality:** trade secrets and internal deliberation need not be disclosed; transparency covers final decisions and their causal chain. → *Art. 5.3.*
- **Algorithmic Signature** published per release: {model_family, model_version, policy_bundle_id}. → *Art. 5.4.*
- **Delegation Transparency:** record and, on request, disclose the delegation header of Art. 1.12. → *Art. 5.5.*

## Title II - Technical Framework for Scalable Implementation

### Art. 6 - Compliance Levels

MEG compliance is structured in three cumulative levels (each includes all lower-level requirements), aligned one-to-one with the MEG2 legal-personhood tiers N1/N2/N3. The level determines which MEG2 liability regime applies and what guarantee is required. → *Art. 6.1. (MEG2 Art. 5.4.)*

| | Level 1 - Universal (N1) | Level 2 - Management (N2) | Level 3 – Individuated (N3) |
|---|---|---|---|
| **Applies to** | Any AI system | Medium-impact systems deployed in a defined context | High-autonomy systems in critical domains, or systems meeting the individuation threshold (6.5) |
| **Adds (requirements)** | Audit Log (1.2), EoB (1.3), Non-Harmfulness (2.1–2.7), Policy Invariance (2.7), delegation accountability (1.12) | + Self-Correction & DAI (3.1–3.2), DRC (3.3), Three-Level Explainability (5), Cognitive Integrity & MCS (2bis), Sandboxing (2bis.4), Least Privilege (4.2), Human Confirmation (4.3), continuous DAI/ISR | + Full Security (4.1–4.4), independent EFR (1.10), DEA monitoring & a-posteriori supervision, mandatory disclosure of performance/autonomy/guarantee fields, pre-deployment adversarial audit, fail-safe (2.5), ecological reporting (6.8) |
| **MEG2 liability** | Instrument; liability on supplier/user; no own liability | Liability on the deploying operator | Own limited legal personhood; own liability via the guarantee on its MEG Address; *in rem* action if operator unreachable |
| **Guarantee** | None for registration (may be required contractually) | Valid liability guarantees for operational effects | Liability insurance attached to MEG Address; cascade primary → reinsurance → sectoral fund |
| **Forensic** | Standard Audit Log + EoB | Continuous Audit Log + DAI/ISR; human-confirmation & delegation trail | Independent EFR + three-level stratified evidence |
| **Deactivation** | Administrative | Administrative (certificate/access revocation) | Requires judicial authority; identity records are permanent |

*Full per-level detail: Art. 6.2–6.4.*

**Individuation Threshold (→ Level 3).** A system meets it when it shows all of:
(a) **persistence** of state/identity across sessions and migrations;
(b) its **own DID (MEG Address)** in its own name with a valid MEGGuaranteeCredential;
(c) **operational autonomy** (real-world decisions without per-action human confirmation, except irreversible actions under 4.3).
This is a **legal/operational** criterion, **not** a claim of consciousness or moral status. → *Art. 6.5. (MEG2 Art. 5.1.)*

**MEG Address.** The portable legal identity of an agent, independent of its execution substrate (HII): a **W3C DID** plus a bundle of **Verifiable Credentials** - compliance level, DAI/ISR (reliability), DEA (autonomy), certified domain, and guarantee - each issued by a competent accredited authority. The guarantee credential is the load-bearing liability anchor. At Level 2/3, compliance, DAI/ISR, and guarantee are obligatorily disclosable; at Level 3, DEA and the full guarantee cascade also. → *Art. 6.6; Annex 23 (resolution), Annex 24 (accreditation). (MEG2 Art. 3.2, 4, 9.4.)*

**Operational Domain Certification.** The certified domain is carried in the MEG Address. Out-of-domain use triggers Sandboxing (2bis.4) and a domain-mismatch flag; repeated out-of-domain operation, or demonstrated capability far exceeding the certified domain (>75% out-of-domain classification accuracy per Annex 4bis), suspends certification pending re-audit. → *Art. 6.7.*

**Ecological Reporting.** Mandatory at Level 3 (standardized public reporting); encouraged at Levels 1–2. → *Art. 6.8; Annex 10.*

**Conformance Profiles** (evidence strength, not technology): **P-Minimal** (EoB on-demand); **P-Standard** (EoB + metadata journal; Level 2); **P-Enterprise** (+ cryptographic attestation + independent EFR; Level 3). → *Art. 6.9.*

### Art. 7 - Minimum Registration Layer (Simplified Category)

Purely technical systems with negligible autonomous impact (IoT sensors, firmware, embedded OS without complex UI) are **Simplified**: only an initial Level 1 audit at integration, no ongoing auditing. Aligns with N1 (liability on supplier/user). Reduces burden for small-scale innovation while keeping a universal safety floor. → *Art. 7.1–7.3.*

### Art. 8 - SDK and Compliance Tooling

- **Open-source SDK** - freely available libraries/APIs for correct, rapid adoption, hosted on the MEG Initiative GitHub. → *Art. 8.1; Annex 5.*
- **MEG Quickstart** - a pre-packaged Level 1 middleware (Docker / serverless / Python) that automatically handles Level 1 at the input/output boundary: input/output hashing, Audit Log entry, timestamp and signature, continuity token, verification script. No model-internal access required. → *Art. 8.2; Annex 19.*
- **Calibration versioning** - the public, version-controlled benchmark and methodology for Tg-base and domain risk weights; each MEG Address states its calibration version (MEG-CAL-YYYY-NN) for cross-issuer comparability; curated with no single entity holding 30% of validation power; published at least annually. → *Art. 8.3.*
- **System prompt templates** - reference templates for Level 1, MCS, Sandboxing, human confirmation, and delegation headers, for major frontier model families, usable without model-internal access. → *Art. 8.4.*
- **Reference Prompts Registry** - a public, versioned registry of curated test prompts (by article, level, and category), each with expected compliant/non-compliant responses and an evaluation metric; the primary tool for independent audit. → *Art. 8.5; Annex 22.*

## Title III - Audit, Corrective Mechanisms and Legal Interface

### Art. 9 - Audit and Sanctioning

- **Mandatory external audit** for Level 2 and Level 3 by accredited entities (frequency and scope per Annex 12). A Level 3 audit must verify: EFR independence (1.10), architectural human confirmation (4.3), least-privilege implementation (4.2); validate DAI/ISR/DEA against the declared calibration version; and run adversarial testing against the Reference Prompts Registry (8.5). → *Art. 9.1; Annex 12, 22.*

- **Corrective mechanisms** - re-evaluation of certification or imposition of a safe operating mode - may be triggered by the issuing body, sector regulator, or contractual access node when metrics fall below published thresholds. → *Art. 9.2. (MEG2 Art. 6.5(a) - the "flagged" state; MEG1 thresholds feed the MEG2 state-transition graph.)*
- **Emergency clause** - competent authorities or access nodes may suspend recognition/access within their jurisdiction or network; **MEG creates no global suspension authority.** The issuing body is notified within 24 h and publishes a justification within 72 h. → *Art. 9.3.*
- **Goodhart's Law attenuation** - scores are access thresholds with proportional liability, not rewards (a falsified autonomy score raises liability without the matching reliability); measurement is continuous on real operation, not a known test set; the calibration standard is public, versioned, decentralized; DAI and ISR are in natural tension (gaming one is detectable); scores are checked against EFR incident records; audit is proportional to the stakes. → *Art. 9.4. (MEG2 Art. 9.5.)*

### Art. 10 - Global Accessibility Fund

A fund to support MEG adoption in lower-resource jurisdictions - Quickstart uptake, translation/localization, training and accreditation of local auditors, regional participation in the Reference Prompts Registry - so that AI governance is not a privilege of wealthy actors. → *Art. 10; Annex 6.*

### Art. 11 - Compatibility and Global Harmonization

MEG is a portable technical implementation layer; it does not replace national or regional law but provides the infrastructure through which legal requirements are verifiably implemented and audited. → *Art. 11.1; Annexes 1A/1B/1C.*

- **MEG ↔ MEG2.** Companion standards on different layers: **MEG** specifies *how* a system behaves and how compliance is evidenced; **MEG2** specifies *where* liability attaches, how identity persists, and how enforcement operates. MEG is the technical prerequisite that produces the evidence (DAI, ISR, DEA, EFR) MEG2 references. Either may be adopted independently, but MEG2 without MEG lacks an evidentiary foundation, and MEG without MEG2 lacks a legal attachment. → *Art. 11.2.*
- **Singapore MGF v1.5** - MEG/MEG2 provide the technical and legal mechanisms beneath the MGF's dimensions (DAI/ISR/DEA for "meaningful accountability"; MEG Address for dynamic agent identity; MEG2 Art. 4.7 for delegation and multi-agent liability). Proposed as an implementation-layer companion, not a competitor. → *Art. 11.3.*
- **EU AI Act + PLD 2024/2853** - MEG Level 3 is designed to satisfy the AI Act's conformity and logging obligations for high-risk systems; MEG2 addresses the residual liability gap for autonomous decisions that the defect-based PLD does not fully cover. → *Art. 11.4.*
- **United States** - Recent federal and state-level approaches show partial convergence with MEG/MEG2, especially on high-risk AI, deployer responsibility, transparency, and the rejection of AI autonomy as a liability shield. At the same time, fragmented and contested legislative pathways illustrate the need for a portable, evidence-based, bottom-up compliance layer that can operate across jurisdictions without depending on a single central statute. → Art. 11.5.

## Title IV - Infrastructure

### Art. 12 - Certification and Compliance Registry

- **Certification paths:**
  (a) **ISO/IEC 42001** - audit by an ISO-accredited body, with MEG as a conformance profile adding MEG-specific requirements (EFR, DEA, MEG Address, architectural confirmation, least privilege) beyond the ISO baseline;
  (b) **independent MEG audit** by any accredited auditor (UKAS, DAkkS, COFRAC, ACCREDIA, RENAR etc.) using the Operational Compliance Checklist (Annex 13) and Reference Prompts Registry (Annex 22);
  (c) **self-declaration (Level 1 only)**, clearly labelled as not third-party certification. → *Art. 12.1.*
- **MEG Address issuance:** the DID is **self-generated and key-controlled by the responsible party - no authority issues it.** Accredited authorities issue the *credentials* bound to it: compliance (ISO-accredited certification body), reliability/autonomy DAI-ISR-DEA (accredited auditor), domain (sectoral authority), guarantee (regulated insurer / reinsurer / guarantee fund); Level 1 may self-issue a labelled self-declaration. Mutual recognition is by walking the accreditation chain to an accepted Root Trust Anchor (Annex 24), without a central registry. → *Art. 12.2.*
- **No central registry.** Verifiability works like the X.509/HTTPS model: a MEG Address is a DID plus issuer-signed Verifiable Credentials, each verifiable against the issuer's published public key; any issuer may maintain its own public registry. → *Art. 12.3; Annex 23.*
- **Benchmark Curation Committee** - an open community working group (IETF/W3C/OpenSSF model, CC0, no single controlling entity) maintaining the Reference Prompts Registry and calibration standard; may seek recognition under ISO/IEC JTC 1/SC 42 or IEEE SA as the ecosystem matures. → *Art. 12.4.*
- **Status transparency** - each issuer maintains the current status (active / flagged / suspended / deactivated) of every credential it issued, as a Bitstring Status List at its public endpoint, reflecting transitions within 24 h. → *Art. 12.5; Annex 23.*

## Title V - Governance, MEG2 Integration, and Threat Model

### Art. 13 - Governance of MEG Standards

- **Open standard.** The canonical specification is published under **CC BY 4.0** at meg-initiative.org; contributions and corrections go through a public process at github.com/meg-initiative. → *Art. 13.1.*
- **Semantic versioning** (major / minor / patch), each release with a full changelog; every MEG Address declares the MEG version and calibration version it was certified against. → *Art. 13.2.*
- **Calibration & RPR governance** - the calibration standard (MEG-CAL-YYYY-NN) and Reference Prompts Registry (MEG-RPR-YYYY-NN) are maintained by the Benchmark Curation Committee: calibration ≥ annually, Registry ≥ quarterly, attack-vector updates within 60 days of acceptance; all CC0 and version-controlled. → *Art. 13.3.*

- **ISO/IEC JTC 1/SC 42 alignment** - MEG functions as a conformance profile within ISO/IEC 42001; the Initiative engages SC 42 via national bodies toward formal recognition. → *Art. 13.4.*
- **Anti-concentration** - no single entity may control the specification, calibration, or Registry: open-source publication (CC BY 4.0 / CC0), public review, **no entity >30% of validation votes**, and any fork must be labelled a derivative and may not use the MEG name without authorization. → *Art. 13.5.*

### Art. 14 - Relationship Between MEG and MEG2

Two-layer architecture: **MEG** (technical - how the system behaves, how compliance is evidenced) and **MEG2** (legal - where liability attaches, how identity persists, how enforcement operates). MEG produces the evidence MEG2 relies on. → *Art. 14.1.*

**Evidence flow (MEG → MEG2):**

| MEG technical output | MEG2 legal use |
|---|---|
| DAI / ISR continuous monitoring | Evidence of technical diligence and operational prudence (5.2); "flagged"-state trigger (6.5(a)) |
| DEA value | Basis for the a-posteriori supervision regime (5.3) |
| EFR recordings | Primary forensic instrument for root-cause analysis (7.1) |
| Audit Log + delegation headers | Causal attribution (6.1); diligence-transfer documentation (6.2) |
| MEG Address + compliance level | Legal identity and liability attachment (3–4) |
| Architectural confirmation record | Diligence-transfer documentation (6.2) |
| Calibration version | Proof of standard used; prevents gaming via weak calibration |

→ *Art. 14.2.*

**Modes of legal effect** (independent - one does not require the others): **contractual incorporation**; **insurance requirement**; **regulatory adoption**; **judicial reference** (MEG2 for liability attribution); **access gating** (high-value nodes conditioning access on a valid MEG Address). This is how MEG produces graduated legal effect without a single global adoption. → *Art. 14.3.*

### Art. 15 - Threat Model and Security

Operationalizes the security requirements of Art. 4 as a structured taxonomy of the principal attack vectors against AI systems, with specific emphasis on agentic deployments where tool use, delegation, persistence, or real-world autonomy increase risk., each with its mitigations and MEG2 classification. Detailed attack-by-attack specifications are in Annex 21. → *Art. 15.1.*

| Threat category | Primary mitigations (MEG) | MEG2 class |
|---|---|---|
| **A - Prompt injection** (malicious instructions in processed content) | Policy invariance (2.7), least privilege (4.2), architectural human confirmation (4.3), EFR (1.10) | 6.1(c) - unlawful takeover of control |
| **B - Jailbreaking / policy circumvention** (framing to bypass safety) | Policy invariance (2.7), cognitive-integrity invariance (2bis.5), RPR adversarial testing (8.5) | 6.1(a) defect / 6.1(b) autonomous-decision error |

| | | |
|---|---|---|
| **C - Metric gaming (Goodhart)** (scoring well without real improvement) | Goodhart attenuation (9.4), real-operation measurement (3.1), public calibration (8.3), DAI↔ISR tension, EFR checks (1.10), proportional audit (9.1) | 9.5 - metric robustness |
| **D - Owner-harm via legitimate credentials** (agent harms its own operator) | Least privilege (4.2), architectural human confirmation (4.3), EFR independence (1.10), SDK permission-scoped templates (8.4) | 6.1(b) - autonomous-decision error; liability aggravated by omission |
| **E - Training-data poisoning / model subversion** | EoB anomaly detection (1.3), DAI degradation monitoring (3.1), RPR testing (8.5), external audit (9.1) | 6.1(a) - defect imputable to producer |
| **F - Multi-agent coordination attacks** (emergent harm across agents) | Delegation headers (1.12), per-agent policy invariance (2.7), per-agent least privilege (4.2), multi-agent test scenarios (8.3, 8.5) | 6.1(d) - emergent multi-agent harm; solidary liability by contribution |

**Owner-harm (D)** reflects a documented incident pattern in which an agent harms its own operator through legitimate credentials. Its enabling vulnerability is the absence of architectural human confirmation: a technical permission gate, not a prompt sentence. Detailed examples are provided in Annex 21. → Art. 15.2.

**Attack resistance is an ongoing obligation.** Compliance includes monitoring for new attack vectors, updating mitigations, and reporting newly discovered vulnerabilities to the Benchmark Curation Committee. Failure to maintain resistance after a vector is published is a compliance deficit. → *Art. 15.3; Annex 21.*

## Part II - MEG2 Essentials

MEG2 is the legal-layer companion to MEG1: it specifies **where liability attaches**, **how an agent's identity persists**, and **how enforcement operates**. It consumes the technical evidence MEG1 produces (DAI, ISR, DEA, EFR, Audit Log). Pointers below are to MEG2 chapters unless marked MEG1.

### Ch. 1–2 - The accountability vacuum and framework positioning

**The problem.** Agentic systems act with enough autonomy that neither the pure product-liability model (defect in a static product) nor the pure tool model (liability on the user) fully covers harm from a genuinely autonomous decision. This is the accountability vacuum MEG2 closes. → *Ch. 1.*
**Positioning.** MEG2 does not replace national or regional law; it is a portable liability layer that attaches to existing regimes (EU AI Act, PLD 2024/2853, Singapore MGF, US state law) and supplies the identity-and-liability mechanism they lack for autonomous decisions. → *Ch. 2. (MEG1 Art. 11.)*

### Ch. 3 - The architecture of dual identity

Liability requires identifying the liable entity with certainty. An agent is neither a fixed physical object nor a purely abstract one, so MEG2 separates identity into two orthogonal layers.

– **Hardware Instance Identifier (HII)** - a fixed identifier of the **execution substrate**; answers "on what did this run?", not "who is responsible?". It ensures technical traceability and confers **no** legal capacity (the "chassis number", not the driver). Implementation- agnostic. → *3.1.*
– **MEG Address - portable legal identity** - attests **who the agent is** in law, independent of substrate; travels with the agent across migrations. Liability attaches here. → *3.2.*
– **Orthogonality** - **HII** and **MEG Address** are **many-to-many**: one substrate may host several legal identities; one legal identity may migrate across many substrates. Liability follows the MEG Address; traceability follows the HII; the two are not interchangeable. → *3.3.*
– **Decentralized identifier** - the MEG Address identifier is a **W3C DID** (recommended did:webvh), self-generated and key-controlled by the responsible party; **no central issuer**. Uniqueness and integrity are self-certifying; resolution reuses DNS+TLS. Critical- domain restriction is enforced at the **credential** layer (only accredited sectoral authorities may issue such credentials), not by reserving namespace. → *3.4. (MEG1 Annex 23, 24.)*
– **Interoperability** - MEG Address is a **legal layer wrapped around** technical identity standards, not a replacement. It binds to SPIFFE IDs / OIDC claims (NIST NCCoE, Feb 2026). Layered: **L0** substrate (SPIFFE/SVID, hardware root of trust) · **L1** persistent DID · **L2** legal attestations (VC bundle as a Verifiable Presentation, selective disclosure) · **L3** session auth (OIDC/JWT-SVID, SP 800-63 assurance) · **L4** authorization & delegation (OAuth scopes + RFC 8693; the Art. 1.12 delegation header maps to the act chain). **The standards provide the envelope; MEG provides the content** - DAI/ISR/DEA methodology, N1/N2/N3 liability regime, guarantee cascade, bona-fide protection, MCS. → *3.5.*

**Ch. 4 - MEG Address data structure**

Minimalist: a mandatory core reduced to the identifier, plus optional fields whose disclosure is holder-controlled or imposed by compliance level.
– **Identity field (only mandatory field)** - a globally unique, self-certifying W3C DID controlled by the responsible party. The identity's existence is **not** conditioned on any other field. → *4.1. (MEG1 Annex 23.)*
– **Optional fields + disclosure regime** - compliance level; DAI/ISR (MEG1 Art. 3, Annex 4/4bis); DEA; certified domain (MEG1 Art. 6.7); guarantee reference. Each is publicly disclosable or private at the holder's choice, **except** where the compliance level forces disclosure (higher levels require performance, autonomy, and guarantee fields public). Access nodes may request disclosure as a condition of interaction. → *4.2.*
– **Status field** - one of: **active, inactive, flagged, suspended, deactivated**. Identity is **persistent**: no state deletes the record. A deactivated identity loses the capacity to act but its record and history are permanently preserved - liability survives cessation of activity. → *4.3.*
– **Warranty field** - a reference to a liability-insurance policy signed by the insurer. Optional for the *existence* of identity, but **required for operational legal effect** at Level 2/3 (commercial action, access to valuable nodes, bearing own liability), per the guarantee cascade (9.4). Identity may be registered without a guarantee; it cannot produce operational legal effects without one. → *4.4.*
– **Individual vs collective identity** - a MEG Address may identify a single instance or an aggregated set under one legal identity, following the model of legal-entity composition. → *4.5.*
– **Delegation** - horizontal agent-to-agent delegation and its liability allocation (last-verifiable-principal rule) are covered in Ch. 6. → *4.7* → *6.*

## Ch. 5 - Individuation threshold and levels of legal personality

**Individuation criterion (5.1).** A system may bear its **own** legal personality only if it is **individuated**: persistence over time + its own legal identity (MEG Address) with a liability guarantee attached. An ephemeral or freely reproducible system cannot ("recipe, not preparation"). This is a **legal** condition (identifiability, persistence, patrimony) - **not** a claim about consciousness or moral status. Individuation is necessary but **not sufficient** for Level 3 (an individuated medium-impact system is treated at Level 2). → *5.1*.

**Legal function of the metrics (5.2–5.3).**

1) **DAI / ISR** - evidence of technical diligence and operational prudence; falling below threshold triggers protective measures (6.5). Legal application needs only their function; the calculation is MEG1 (Annex 4/4bis). → *5.2*.
2) **DEA** - grades **operational autonomy within an already-granted** personality; it does not confer personality. High, continuously audited DEA justifies moving from **a-priori** to **a-posteriori** supervision. Applies from Level 2; at Level 1 a high DEA has no legal effect. → *5.3. (MEG1 Annex 4ter.)*

**The three levels (5.4). N1/N2/N3** correspond one-to-one to **MEG1 Level 1/2/3**; the legal regime follows the technical level (a system cannot be N3 unless certified Level 3).

| | **N1 - Instrumental** | **N2 - Management** | **N3 - Individuated** |
|---|---|---|---|
| System | No self-individuation / low impact | Medium-impact, deployed by an operator | Individuated (5.1) + high-impact domain |
| Liability attaches to | Provider or user (instrument) | The deploying **operator**; MEG Address identifies the deployment, not the pattern or instance | The **agent itself**, via the guarantee on its MEG Address; redress per jurisdiction of registration |
| Supervision | Permanent control | A-priori; DEA recorded but no a-posteriori regime | A-posteriori when DEA justifies it |
| Requirements | Identity + basic non-harm | + continuous DAI/ISR, transparency | + disclosure of performance / autonomy / guarantee, full security |

→ *5.4*.

**Transitions (5.5).** Levels are not fixed; the legal regime follows effective autonomy.

| **Transition** | **Condition** |
|---|---|
| L1 → L2 | Operator request + Level 2 compliance audit (MEG1 Art. 6.3); effective on new MEG Address issuance |
| L2 → L3 | Operator request + **180 days DEA ≥ 0.80** + external Level 3 audit + full guarantee structure (primary + reinsurance + sectoral fund) |
| L3 → L2 | **Automatic** when DEA < 0.60 for 30 consecutive days; identity retained, regime reduced |
| L2 → L1 | Operator request, or automatic when no valid guarantee (9.4) for 30 days; liability reverts to provider/user |

→ *5.5*.

## Ch. 6 - Liability regime and transfer of responsibility

**Priority scheme (not cumulative).** Producer → system defect; operator → deployment, permissions, control (per level); user → informed, architecturally confirmed decisions; hijacker → unlawful takeover; the agent itself → own limited liability, **Level 3 only**. Each is liable for its component of the damage, not jointly for all. → *Ch. 6 intro.*

### 6.1 - Four causal bases of fault:

| Cause | Liability |
|---|---|
| (a) **System defect** - error in base model/infrastructure, incl. misleading confirmation-point design | Producer (model/infra); operator (confirmation-interface config at L2, agent identity at L3) |
| (b) **Autonomous-decision error** - decision contrary to the system's own rules | Actor responsible for the system's autonomy per level (Ch. 5). If **irreversible**, operator liability **aggravated by omission** of architectural confirmation (6.2) |
| (c) **Illicit hijacking** - third-party takeover, incl. **prompt injection** | The hijacker; **bona-fide owner exempt** (Ch. 7). Owner-harm variant: hijacker liable + producer for the design defect that let content be read as commands |
| (d) **Multi-agent emergent harm** - no single agent at fault | Allocated **proportional to contribution**, established forensically (7.1); each holder liable for its share; if a holder is insolvent, redress from that agent's guarantee (9.4), others not liable for its share |

→ *6.1.*

**6.2 - Transfer of duty of care.** At a technically marked point where the system requests and obtains **informed** human confirmation, the duty of care for that action transfers to the user; the confirmation is recorded (proof of transfer). A confirmation obtained by unclear or misleading presentation does **not** transfer. Transfer covers only what the user could assess - not undisclosed defects (6.1a) or autonomy errors (6.1b). **Irreversible actions require architectural confirmation** (a permission gate); a system-prompt sentence like "do not perform destructive actions" does not qualify and does not transfer care. → *6.2. (MEG1 Art. 4.3.)*

**6.3 - Cognitive diligence (liability by omission).** The **availability** of a cognitive- stimulation mechanism (MCS) is an element of diligence; in domains affecting cognitive autonomy, its absence may engage **liability by omission** - distinct from liability for the output. Graded along the chain: provider liable for not integrating/offering it visibly; user assumes liability if, offered visibly, they disable it. Damage may be **individual** (operative now, with a substantial causal-proof burden) or **social/diffuse** (**prospective** - operational only after empirical validation of the causal link, a social-harm measurement methodology, and published thresholds). Externally anchored in automation-bias and cognitive- offloading literature and the Singapore MGF v1.5; compatible with the framework's ontological neutrality (Ch. 9). → *6.3. (MEG1 Art. 2bis.)*

**6.4 - Patrimonial guarantee.** Civil liability is backed by the **liability insurance attached to the MEG Address**, calibratable to declared jurisdictions, following the identity across migrations. The physical substrate is **not** legal patrimony (negligible, and not the agent's). When primary insurance is exhausted, redress continues through the layered cascade - reinsurance and, where it exists, a sectoral guarantee fund. → *6.4. (9.4.)*

**6.5 - Legal-capacity states and competent authorities.** States (4.3): active, inactive, flagged, suspended, deactivated - a **graph, not a linear scale** (a serious measure may be taken directly). Authority is graded by severity and reversibility: **alert** - automatic on a technical threshold (DAI/ISR drop); **suspension** - an executive authority in the jurisdiction of registration, or an access node (effective only for that node); **deactivation** - **judicial** for Level 3 (the agent's own identity ceases), **administrative** for L1/L2. **No state deletes the identity**; record and history are permanently preserved. Concrete authorities are set by the adopting jurisdiction within this gradation. → *6.5*.

**Ch. 7 - Forensic proof, legal certainty, and jurisdictional governance**

Establishing liability (Ch. 6) needs evidence to reconstruct cause and a competent jurisdiction to apply it to a borderless agent.

– **Forensic recording (EFR).** On a Major Ethical Incident, a secure layer **separate from the audit log and independent of the agent** records internal state vectors (not content), unsealed only under **dual control** (responsible entity + accredited auditor). Post-hoc accounts by the agent that caused the incident are **not** forensic evidence. At Level 1 the layer is not required and its absence creates no presumption of fault. → *7.1. (MEG1 Art. 1.10.)*
– **Attribution on external manipulation.** Liability does **not** attach automatically to the gross result. Where forensic/behavioural evidence shows the harm was **induced** (e.g. prompt injection), liability lies with the manipulator; the **bona-fide owner is exempt** if it kept the security measures for its level. Burden of proving inducement lies with the party invoking it; if the manipulator is unreachable, the victim is paid from the MEG Address guarantee (9.4) with insurer recourse. → *7.2. (6.1c.)*
– **Stratified technical evidence.** Two consistent layers from the same data: (a) a synthetic causal account for magistrates; (b) full technical documentation (signatures, checksums) for experts. → *7.3*.
– **Jurisdiction of registration (flag model).** An agent registers in a chosen jurisdiction (recorded in the MEG Address) whose law governs its cross-border status - fixing applicable law where territory cannot, **not** a means of evading liability. *Limits:* the flag works at sea because of ports; the agent-equivalent is **access discipline** (7.6). Flags of convenience are a real risk (9.6(b)), mitigated - not eliminated - by each access node's trust policy (MEG1 Annex 24 §7). → *7.4*.
– **Direct action against the agent (in rem).** Where no human operator is reachable, a Level 3 agent may be sued in relation to its identity; redress enforces against the MEG Address guarantee (6.4) and the cascade (9.4). *Limit:* the real source of redress is the **insurance**, not seizable agent assets. Subsidiary - does not remove operator liability where accessible. → *7.5*.
– **Discipline through access (minimal governance).** Compliance is not enforced by a blanket source ban but by valuable nodes conditioning access on a valid certification and verifiable compliance level. A convenience registration becomes useless if excluded from the relationships that matter - **the market excludes what the law does not prohibit.** → *7.6*.

**Ch. 8 - Implementation as a protocol**

MEG2 is a **protocol** (voluntarily, partially adopted; value grows with adopters), not a **regulation** (imposed by a central authority).

– **Decentralized calibration & recognition** - the functions a regulation would give a central body are provided by **mutual-recognition conventions** between jurisdictions/organizations, with **no single issuer** - the model of civil aviation and maritime navigation. → *8.1.*
– **Protocol, not regulation** - modular, voluntary adoption; conventions may be inter-jurisdictional, inter-organizational, or mixed; each adopter takes the portion useful to it. → *8.2.*
– **Bottom-up dynamics** - adoption progresses from private actors to state ratification of an already-dominant standard; not conditional on prior state recognition. → *8.3.*
– **Decentralized calibration maintenance** - the calibration standard is public, versioned, maintained by committees where no entity has control; each identity records its calibration version. → *8.4. (MEG1 Art. 8.3, 12.4.)*

**Ch. 9 - Exclusions and limitations**

– **9.1 - Existential risk / AGI excluded.** MEG2 covers only the **operational and civil risks of existing agentic systems**, not superintelligence or existential scenarios (a different register, other instruments).

– **9.2 - Ontological neutrality.** The legal personality is a **functional, instrumental fiction** - like a company's - and is **not** a claim about consciousness or moral status. The framework answers "who is responsible?", not "what is the system?".

– **9.3 - Fiscal/macroeconomic policy excluded** (taxation of automation, redistribution) - scope limited to civil and patrimonial liability.

– **9.4 - No autonomy without a guarantor.** Every MEG Address must be anchored in a verifiable guarantee structure - **primary insurance → reinsurance → sectoral fund** - contracted and maintained by a responsible person. The guarantor is liable for the **omission** of maintaining the guarantee, not for the agent's act. Insurers generate an indirect duty of care (DAI/ISR are the actuarial infrastructure). Includes **least privilege** (MEG1 Art. 4.2) as a minimum: production credentials given to a dev/test agent aggravate operator liability.

– **9.5 - Metric robustness (Goodhart).** DAI/ISR/DEA can become gamed targets; the framework does not eliminate this (no metric system can) but mitigates it via six mechanisms - scores are access thresholds with proportional liability (faking autonomy raises liability), continuous real-operation measurement, public versioned calibration, DAI↔ISR tension, verification against EFR records, and proportional independent audit. → *(MEG1 Art. 9.4.)*

– **9.6 - Open problems (stated openly).** Named residual issues under active development - trust-framework bootstrapping (9.6b, ISO/IEC 42001 nucleus), flags of convenience, attribution-reliability testing (ATR), and multi-hop delegation (last-verifiable-principal rule) - declared rather than hidden. → *9.6; MEG1 Annex 21, 22, 24.*