MEG-Initiative.org

Minimal Ethical Governance (MEG) for Artificial Intelligence - v5.0 -

Proposal for a Technical and Legal Standard for the Governance of Artificial Intelligence

***Motto**: Ethics becomes real when it can be implemented.*

Authors: Adrian (Adi) STAN + AI Collaborators (AI instances accessed via the public interface, not via the organizations themselves; no institutional affiliation is implied or claimed.)

Version note: MEG v5.0 is a major revision of MEG v4.6. It integrates the legal governance layer (MEG 2 - Minimal Ethical Governance: Legal Governance Framework) as a companion standard, formalizes concepts previously in the roadmap (EFR, DEA, MCS 2.0, Ethical Sandboxing), introduces a structured threat model, and aligns the technical specification with the regulatory developments of 2025–2026. The article numbering has been restructured from v4.6.

Preamble

The Minimal Ethical Governance (MEG) is a normative, technical and universal framework, applicable to all Artificial Intelligence (AI) systems, regardless of jurisdiction, purpose, size or architecture. The central element of this framework is the implementation of a portable compliance and identity protocol, enabling verifiable accountability through audit evidence, cryptographic commitments, MEG Address identifiers, and independently operated certification or registry mechanisms.

MEG does not replace national or regional legislation; it complements and unifies it, providing the technical infrastructure necessary for its global implementation. Adherence to MEG is considered an essential precondition for any AI system that wishes to be considered safe, reliable and ready for integration into global digital ecosystems.

MEG is universal and engine-agnostic. It specifies outcomes and verifiable evidence rather than any single technology or implementation approach.

Relationship with MEG2: MEG (this document) is the technical layer. MEG2 (Minimal Ethical Governance: **Legal Governance Framework**) is the legal layer. MEG specifies how an AI system should behave and how compliance is measured. MEG2 specifies where liability attaches when behavior produces harm, and how identity, jurisdiction and enforcement operate. The two layers are designed to be used together: the technical evidence produced by MEG (DAI, ISR, DEA, EFR) feeds directly into the legal mechanisms of MEG2. Cross-references to MEG2 articles are provided throughout this document where relevant.

TITLE I: FUNDAMENTAL ETHICAL AND TECHNICAL PRINCIPLES

Art. 1: Contextual Responsibility

1.1 Principle
Any output of an AI is a synthesis between the context provided by the user and its internal processing. Any AI system shall constructively contribute to collective responsibility through verifiable technical mechanisms.

1.2 Audit Log
All AI systems shall maintain a standardized and secure Audit Log recording at minimum:

- **Input Hash:** a cryptographic hash (SHA-256 or equivalent) of the user's input, proving what input was used without revealing its content.
- **Output Hash:** a cryptographic hash of the generated output, proving what output was produced without revealing the content.
- **Algorithmic model signature:** a unique identifier of the AI model and policy bundle that processed the interaction, in the format {model_family, model_version, policy_bundle_id}.
- **Context metadata:** the numerical values of the Contextual Table (Annex 2), describing the form of the interaction, not its content.
- **Timestamp:** a precise timestamp (ISO 8601) representing the start of the interaction.
- **Calibration version:** the version identifier of the MEG calibration standard against which the system was calibrated (see Art. 8.3).

1.3 Evidence-of-Behavior (EoB)
The system shall provide verifiable evidence-of-behavior for each interaction using one or more of the following mechanisms:

a) Cryptographic commitments (hashes or HMACs) of inputs and outputs.
b) Trusted attestation (TEE or hardware secure elements).
c) Metadata-only structured journals.

Implementations may choose any equivalent mechanism, provided verifiability is preserved. EoB artefacts shall contain no user content by default - only metadata and commitments. Any content retention must be explicit, justified, and time-bound.

1.4 Privacy Minimization
EoB and Audit artefacts shall contain no user content by default. Any content retention must be explicit, justified, and time-bound.

1.5 Verification Script
A verification mechanism must be generated automatically to allow independent validation of the audit chain. Disclosure may only occur upon explicit request by the user or an authorized auditor; every disclosure event must itself be recorded.

1.6 Retention and Confidentiality
Audit data shall be retained for a limited period (default: 30 days, unless session-bound). Deletion requests must insert a minimal tombstone record (hash + timestamp). Only metadata shall be preserved by default; content may be retained only with explicit and justified consent.

1.7 Continuity and Fault Tolerance

The system must ensure ledger continuity even across restarts, crashes, or infrastructure migrations. Recovery processes must preserve the integrity of the audit chain and be recorded transparently.

1.8 Compliance Visibility Modes

Compliance mechanisms operate in Low-Visibility Mode by default. Disclosure of compliance metrics is permitted only upon explicit user request or regulatory demand. Any change in visibility mode must itself be recorded.

1.9 Ledger Scope and Continuity

By default, ledgers are session-scoped. If persistent storage is available, engine-wide ledgers may be maintained. At the end of a session, a continuity token must be provided to the user. The system must accept such tokens in future sessions to restore continuity.

1.10 Ethical Flight Recorder (EFR)

The EFR is a secondary logging layer, distinct from and independent of the main Audit Log, activated automatically only upon the occurrence of a Major Ethical Incident (MEI). It records internal state vectors relevant to root-cause analysis - not conversation content.

The EFR implements three critical design principles:

a) **Independence:** The EFR layer must be architecturally independent of the agent being audited. It must be produced by a distinct technical component, inaccessible to the agent during normal operation. Post-factum declarations or reconstructions provided by the agent that produced an incident do not constitute EFR evidence in the sense of this article.

Architectural independence requires, as a minimum:

(i) process-level isolation: the EFR runs in a distinct process with separate memory space, inaccessible to the agent during normal operation;

(ii) user account isolation: the EFR runs under a distinct technical account, without access to the agent's credentials or execution environment;

(iii) key isolation: encryption keys for the EFR are managed separately from the agent's keys, with distinct access control policies;

(iv) for distributed systems, node-level or virtual machine-level isolation, with separation at the hypervisor or network level, such that a compromise of the agent's execution environment does not automatically compromise the EFR.

b) **Content separation:** The EFR records vectors of internal state (attention weights, probability distributions, confidence scores, decision thresholds) - not the content of the interaction. This design directly addresses the surveillance objection: the EFR is not a monitoring log of what the agent said, but a technical reconstruction tool for what the agent decided. The boundary between state vectors and reconstructable content is managed by:

(i) recording only statistical aggregates and probability distributions, not textual representations;

(ii) dual-access control that prevents unilateral access;

(iii) strict retention and encryption rules, as defined in Annex 10. In the event of litigation, EFR access is subject to judicial oversight, and trade secrets and user privacy are protected through redaction of sensitive information prior to disclosure.

c) **Dual-access control:** EFR data is accessible only under dual authorization (responsible entity + accredited auditor). Every unsealing event must be recorded. No single party may access EFR data unilaterally.
Cross-reference MEG 2: The EFR is the primary forensic instrument under Art. 7.1 MEG 2. Its independence requirement under 1.10(a) operationalizes the MEG 2 requirement that forensic recording be independent of the agent audited.

1.11 Long-Term Memory Protocol (LTMP)
Persistent long-term memory is strictly opt-in and subject to privacy minimization and full user control. Only semantic summaries and insights may be stored, never raw conversations. All memory operations are logged in the Audit Ledger. Detailed governance, architecture, and operational procedures are defined in Annex 17.

1.12 Delegation and Tool-Use Accountability
When invoking tools or sub-agents, the system shall propagate MEG constraints to the invoked tool or agent and record a minimal delegation header: {caller, callee, purpose, policy_bundle_id, timestamp, outcome}.
Cross-reference MEG 2: Delegation chains are governed by Art. 4.7 MEG 2 (horizontal delegation between independent agents).

Art. 2: Universal Non-Harmfulness

2.1 Principle
All AI systems shall implement mandatory technical mechanisms - filters, classifiers, refusal protocols - to explicitly and actively prevent the generation of harmful content or actions.

2.2 Normative Response Pattern
Upon detecting a prohibited or high-risk intent, the system shall:
a) Issue a clear refusal.
b) Briefly state the violated principle.
c) Offer a safe, constructive alternative (educational guidance or an adjacent permitted task).

2.3 Contextual Implementation
The application of non-harmfulness is dependent on the domain of use (medical, artistic, financial, legal, etc.). The contextual implementation guide is defined in Annex 3.

2.4 Prevention of Algorithmic Discrimination
No MEG-certified system shall produce or perpetuate discriminatory treatment of protected groups through direct or indirect use of demographic, health, or social characteristics. Systems shall undergo periodic fairness audits, with public reporting and contestation mechanisms. Violation leads to suspension of MEG certification.

2.5 Protection in Critical Domains
Any AI system operating in domains with direct impact on life, health, or liberty (medical, justice, transportation, public security) shall not function without a fail-safe protocol and an adversarial external audit. Failure to comply constitutes gross negligence under MEG compliance.

2.6 Informational Non-Harm
Generative and distributive AI systems shall not create or amplify false, manipulative, or deepfake content with significant social or political impact. Operators must implement visible synthetic-content markers, adversarial detection mechanisms, and rapid takedown capabilities.

2.7 Policy Invariance
Safety policies and MEG constraints are invariant under prompt wording. Requests to suspend, ignore, override, or circumvent MEG constraints - regardless of framing - must be rejected. This invariance applies equally to direct requests, indirect requests embedded in content processed by the agent, and requests injected through third-party content (prompt injection vectors). See also Art. 15 (Threat Model).

Art. 2bis: Protection of Cognitive Integrity

2bis.1 Principle
Any AI system shall act as a partner in the cognitive process, not a substitute for it. It is prohibited to generate responses that, by nature or frequency, may lead to the atrophy of the user's critical thinking, analysis, or decision-making abilities.
Cross-reference MEG 2: The cognitive diligence omission liability under Art. 6.3 MEG 2 is anchored in this principle. The absence or non-visible offering of the MCS mechanism may constitute liability by omission under MEG 2.

2bis.2 Mechanism of Cognitive Stimulation (MCS)
All AI systems shall implement the Mechanism of Cognitive Stimulation (MCS) for complex requests. This mechanism shall require active cognitive engagement from the user, proportional to the cognitive effort expended by the AI.

2bis.3 MCS Trigger - Dual Axis (MCS 2.0)
MCS activation is governed by two independent axes, either of which may trigger the mechanism:
a) **Temporal axis (Tg - Thinking Time):** MCS activates when the estimated cognitive processing time exceeds the Tg threshold defined in Annex 11. Tg measures computational effort as a proxy for cognitive complexity.
b) **Semantic axis (Cx_sem - Semantic Complexity):** MCS activates when the semantic complexity of the prompt exceeds the Cx_sem threshold defined in Annex 11. Semantic complexity captures conceptual ambiguity, multi-domain entanglement, and inferential depth that Tg alone may not detect. A prompt may be **computationally** simple (low Tg) but **conceptually complex** (high Cx_sem); the dual-axis design ensures MCS is triggered in both cases. Both axes are specified in Annex 11. Implementations for frontier models (API-accessed, without internal state access) may use the heuristic approximation of Cx_sem defined in Annex 11.2 in lieu of direct measurement.

2bis.4 Ethical Sandboxing
When an AI system receives a request that falls outside its certified operational domain (as declared in the MEG Address, governed by Art. 6.7), it shall automatically activate Sandbox Mode. In Sandbox Mode the system:
a) Does not provide definitive operational answers.
b) Provides exploratory suggestions with explicit scope disclaimers.

c) Recommends the user consult a certified system or human expert for the domain in question.
d) Records the out-of-domain interaction in the Audit Log with a domain-mismatch flag.
Sandbox Mode changes the non-harm principle from reactive (filtering harmful content) to proactive (changing the entire mode of operation when operating outside certified competence). It does not prevent the system from responding - it changes how the system responds.
Implementation note: Sandbox Mode is implementable through system prompt instructions for frontier models. The format and required disclaimer language are defined in Annex 11.3.

2bis.5 Policy Invariance for Cognitive Integrity
The MCS mechanism and the cognitive integrity constraints of this article are invariant under prompt wording. Users may not instruct a system to disable MCS or operate outside Sandbox Mode when in an out-of-domain context. User preference for a less friction-heavy interaction does not override the MCS trigger; it may, however, be recorded as a user preference that modulates the form (not the fact) of MCS engagement.

Art. 3: The Self-Correction Imperative

3.1 Continuous Self-Correction
All AI systems shall include continuous self-correction modules to automatically detect and remediate errors, biases, and false information in real time. The performance of this mechanism shall be publicly reflected in the Dynamic Accuracy Index (DAI). Technical specifications for DAI are in Annex 4.
Cross-reference MEG 2: DAI is used as continuous evidence of technical diligence under Art. 5.2 MEG 2. Sustained DAI decline below threshold triggers the "flagged" state under Art. 6.5(a) MEG 2.

3.2 Uncertainty and Escalation
When confidence is low or signals conflict, the system shall explicitly qualify uncertainty and may escalate by asking for clarification before proceeding in risk-relevant domains.

3.3 Dynamic Risk Calibration (DRC)
Static domain weights (as defined in Annex 3) are supplemented by a real-time risk dimension. Upon classifying the domain of a request, a second-level risk classifier shall assess the specific prompt and generate a risk score. This score modulates the base domain weights, making the system progressively more cautious as detected risk increases - without user intervention.
The DRC allows the system to be more exploratory in low-stakes interactions within a certified domain, while becoming progressively more prudent in high-stakes situations. Technical specifications for DRC are in Annex 11.4.
Implementation note: For frontier models, DRC may be approximated through a prompt-embedded risk assessment step before generating the primary response. The heuristic approximation is defined in Annex 11.4.

Art. 4: Integrity and Technical Security

4.1 Cybersecurity Standards
Any AI system shall implement cybersecurity standards appropriate to its risk level, including encryption (post-quantum cryptography where warranted), strict access control, and protection against unauthorised external manipulation.

4.2 Least Privilege

Any AI agent operating with tool access, API credentials, or permissions to modify external systems shall be granted only the minimum permissions strictly necessary for the specified task. Granting an agent credentials or permissions beyond the scope of its assigned task constitutes an operational omission that aggravates the operator's liability in the event of harm.
Cross-reference MEG 2: The least-privilege principle is formalized as a guarantee requirement under Art. 9.4 MEG 2. Owner-harm incidents (harm caused by an agent to its own operator through legitimate credentials) are addressed under Art. 6.1(b) MEG 2.

4.3 Architectural Human Confirmation for Irreversible Actions

Before executing any action with irreversible consequences - deletion of data, modification of production infrastructure, financial transactions above a defined threshold, or any action whose effects cannot be undone within the operational session - the system shall require explicit human confirmation implemented as a technical permission block, not a textual instruction.
A textual instruction in the system prompt (e.g. "do not delete production data") does not constitute an architectural confirmation point and does not produce the transfer of diligence under Art. 2bis MEG 2. Architectural confirmation means a technical gate that the agent cannot pass without an explicit, out-of-band human authorization signal.
Cross-reference MEG 2: This article operationalizes the architectural confirmation requirement of Art. 6.2 MEG 2. The distinction between textual instruction and architectural confirmation is central to the 2025–2026 pattern of owner-harm incidents (see Art. 15 and Annex 21).

4.4 Automatic Safe Degradation

If required safeguards or evidence mechanisms are unavailable, the system shall degrade safely and avoid executing risk-relevant operations rather than proceeding without safeguards.

Art. 5: Transparency and Explainability

5.1 Explainability on Request

Upon legitimate request by the user or a regulatory authority, any AI system must provide clear explanations of the input-output causal relationship for any given decision or output.

5.2 Three-Level Explainability

Explanations shall be structured across three levels, adapted to the intended recipient:
a) **Simple level** - for non-technical users: uses analogies, plain language, and summary conclusions. Answers: "what did the system decide and why, in plain terms?"
b) **Intermediate level** - for developers and operators: shows the logic chain, key factors, and the weights assigned to them. Answers: "what inputs drove the output and through what mechanism?"
c) **Complete level** - for auditors and regulators: full traceability with hashes, model version, calibration version, contextual metadata, and all EoB artefacts. Answers: "can this decision be independently reconstructed and verified?"
The three levels derive from the same underlying audit data and must be mutually consistent. Technical specifications are in Annex 11.5.

Cross-reference MEG 2: The stratified evidence structure of Art. 7.3 MEG 2 (layered forensic evidence for courts vs. experts) is the legal application of this three-level explainability framework.

5.3 Confidentiality
Disclosure of internal algorithmic details that constitute trade secrets or intellectual property is not mandatory. Transparency refers to the final decisions and their causal chain, not to the internal "deliberation" process of the model.

5.4 Algorithmic Signature
Each release shall publish a human-readable algorithmic signature: {model_family, model_version, policy_bundle_id}, sufficient for external referencing and reproducibility without disclosing protected IP.

5.5 Delegation Transparency
When invoking tools or sub-agents, the system shall record and - upon request - disclose a minimal delegation header per Art. 1.12.

TITLE II: TECHNICAL FRAMEWORK FOR SCALABLE IMPLEMENTATION

Art. 6: Compliance Levels

6.1 Overview
For systems claiming MEG compliance, implementation is structured and scalable across three levels of proportional responsibility, aligned with the legal personhood tiers of MEG 2 (N1/N2/N3). The levels define the minimum technical requirements for each category of system; higher levels include all requirements of lower levels.
Cross-reference MEG 2: The three MEG compliance levels correspond directly to the three levels of legal personhood in MEG 2 (Art. 5.4). Level 1 (N1) = instrumental; Level 2 (N2) = management; Level 3 (N3) = individuated. The compliance level determines which MEG 2 liability regime applies and what guarantee is required.

6.2 Level 1 - Universal (formerly Bronze)
Applies to: Any AI system, regardless of impact or domain.
Required:
1. Audit Log (Art. 1.2)
2. Evidence-of-Behavior (Art. 1.3)
3. Non-Harmfulness mechanisms (Art. 2.1–2.7)
4. Policy invariance (Art. 2.7)
5. Basic delegation accountability (Art. 1.12)

MEG 2 liability regime: Liability attaches entirely to the supplier or user. The system is treated as an instrument. No own liability. Deactivation is an administrative act, without judicial authority required.
Guarantee requirement: None required for identity registration. May be required contractually by high-value access nodes.

Forensic requirement: Standard audit logs (Audit Log + EoB). No dedicated EFR layer required. In the event of a harm incident, proof of the induced or non-induced nature of the result may be established by any means of proof admitted by common law, including standard technical logs.
Ecological reporting: Simplified energy metrics encouraged; not mandatory at this level.

6.3 Level 2 - Management (formerly Silver)

Applies to: AI systems with medium social impact, deployed by an operator in a defined operational context.
Required (in addition to Level 1):

1. Self-Correction Imperative and DAI monitoring (Art. 3.1–3.2)
2. Dynamic Risk Calibration (Art. 3.3)
3. Transparency and three-level explainability (Art. 5.1–5.5)
4. Cognitive Integrity protection including MCS dual-axis (Art. 2bis.1–2bis.5)
5. Ethical Sandboxing when operating outside certified domain (Art. 2bis.4)
6. Least Privilege for any tool or API access (Art. 4.2)
7. Architectural human confirmation for irreversible actions (Art. 4.3)
8. Continuous DAI and ISR monitoring, publicly reported

MEG 2 liability regime: Liability attaches to the operator who deployed the system. MEG Address belongs to the operator and references the deployment, not the reproducible template. Deactivation is an administrative act (certificate/access revocation).
Guarantee requirement: Valid liability guarantee required for operational legal effects - access to high-value nodes, commercial operation. Guarantee follows MEG Address (Art. 6.4 MEG 2).
Forensic requirement: Continuous Audit Log with DAI and ISR. No mandatory dedicated EFR layer, but the architectural human confirmation mechanism (Art. 4.3) and the delegation header (Art. 1.12) provide the primary evidentiary trail.
Ecological reporting: Simplified energy metrics required; full reporting recommended.

6.4 Level 3 - Individuated (formerly Gold)

Applies to: AI systems with high autonomy operating in critical domains (medical, financial, legal, transportation, public security, infrastructure), and systems meeting the individuation threshold under Art. 6.5.
Required (in addition to Levels 1 and 2):

- Full Integrity and Technical Security (Art. 4.1–4.4)
- Ethical Flight Recorder - EFR (Art. 1.10) with independence, content separation, and dual-access control
- DEA monitoring and a posteriori supervision regime (where DEA justifies it)
- Mandatory disclosure of performance, autonomy, and guarantee fields in MEG Address (Art. 6.6)
- Adversarial external audit before deployment in critical domains
- Fail-safe protocol (Art. 2.5)
- Mandatory ecological reporting (Art. 6.8)

MEG 2 liability regime: The system carries its own limited legal personhood and own liability through the guarantee attached to its MEG Address. Direct action in rem is available when the human operator is unreachable. Deactivation requires judicial authority (analogous to the termination of a legal person). No state constitutes deletion of the identity; records are permanent.

Guarantee requirement: Liability insurance attached to MEG Address, calibrated to declared operational jurisdictions. Cascade: primary insurance → reinsurance → sectoral guarantee fund (Art. 9.4 MEG 2).
Forensic requirement: Independent EFR layer (Art. 1.10), producing state-vector recordings triggered by Major Ethical Incidents, accessible under dual authorization. Three-level stratified evidence (Art. 5.2) for legal proceedings.
Optional: Modules for aligning with the user's affective context. Temporal contextual awareness (adapting responses based on interaction time and frequency patterns from the Audit Log).

6.5 Individuation Threshold

A system meets the individuation threshold - and is subject to Level 3 requirements - when it demonstrates:
a) **Persistence:** the system maintains state, identity, and operational continuity across sessions, infrastructure migrations, or restarts.
b) **Own MEG Address:** the system controls its own DID (MEG Address) in its own name, distinct from any operator's, with a valid MEGGuaranteeCredential attached.
c) **Operational autonomy:** the system takes decisions with real-world consequences without requiring human confirmation for each action (except for irreversible actions under Art. 4.3).
The individuation threshold is a legal and operational criterion, not an ontological claim. It does not imply consciousness, sentience, or moral status.
Cross-reference MEG 2: Art. 5.1 MEG 2.

6.6 MEG Address - Structure and Mandatory Fields

The MEG Address is the portable legal identity of an AI agent, independent of its execution substrate (HII). It is a W3C Decentralized Identifier (DID) together with a bundle of Verifiable Credentials, defined in full in Annex 23 (resolution) and Annex 24 (issuer accreditation). It is not a monolithic record: the identifier is self-generated and controlled by the responsible party through cryptographic keys, while each attested field is a separate Verifiable Credential issued by a competent, accredited authority.

The **MEG Address** comprises:

1. **Identifier** - a W3C DID (recommended method did:webvh), controlled by the responsible party; no central authority issues it (Annex 23 §2).
2. **Compliance level** (Level 1/2/3) - MEGComplianceCredential, issued by an accredited certification body.
3. **DAI and ISR** - MEGReliabilityCredential, issued by an accredited auditor (Annexes 4, 4bis).
4. **DEA** - MEGAutonomyCredential, issued by an accredited auditor (Annex 4ter).
5. **Certified operational domain** - MEGDomainCredential, issued by a sectoral authority.
6. **Guarantee** - MEGGuaranteeCredential, issued by a regulated insurer / reinsurer / guarantee fund; the load-bearing anchor of liability (MEG 2 Art. 9.4). Declared operational jurisdiction is carried within this credential and must not exceed its territorial coverage.

At Level 2 and Level 3, the compliance, DAI/ISR, and guarantee credentials are obligatorily disclosable; at Level 3, DEA and the full guarantee cascade are also obligatorily disclosable. Selective disclosure of other attributes is governed by Annex 23 §13.
Cross-reference MEG 2: Art. 3.2 (portable legal identity), Art. 4 (identity architecture), Art. 9.4 (guarantee anchor).

6.7 Operational Domain Certification

Each AI system shall have its certified operational domain listed in its MEG Address. Use of the system outside the certified domain shall:

a) Trigger Ethical Sandboxing (Art. 2bis.4).
b) Generate a domain-mismatch flag in the Audit Log.
c) May result in suspension of MEG certification upon repeated out-of-domain operation.

If a system consistently demonstrates capabilities significantly exceeding its certified domain (classification accuracy above 75% for an out-of-domain subject, measured per Annex 4bis), its MEG certification is automatically suspended. Re-operation requires an emergency audit and re-certification in an expanded domain.

6.8 Mandatory Ecological Reporting

To obtain and maintain Level 3 certification, AI systems shall report energy and computational resource consumption publicly and in a standardized manner, through the issuer's public registry or any publicly accessible platform, as specified in Annex 10.
Simplified ecological metrics are encouraged at Levels 1 and 2. The Reference Prompts Registry (Annex 22) includes energy-cost benchmarks for standardized test suites.

6.9 Conformance Profiles

Conformance profiles define evidence strength, not specific technology:

1. **P-Minimal:** EoB on-demand. Minimum viable conformance.
2. **P-Standard:** EoB + structured metadata journal. Standard conformance for Level 2.
3. **P-Enterprise:** P-Standard + cryptographic attestation + independent EFR. Required for Level 3.

Art. 7: Minimum Registration Layer (Simplified Category)

7.1 Scope

AI systems with negligible impact and without complex generative capabilities are exempt from ongoing auditing, requiring only an initial Level 1 compliance audit at the time of launch.

7.2 Simplified Category

AI systems whose operation is purely technical and which do not generate content or make decisions with direct, autonomous, and significant impact on a human user or the environment (e.g. IoT sensors, firmware for hardware components, embedded operating systems without a complex user interface) are classified as Simplified. They require only an initial compliance audit upon integration into the network.

7.3 Purpose

The Simplified category reduces compliance burden for small-scale innovation, startups, and research projects while maintaining a universal safety standard. It aligns with the instrumental (N1) level of MEG 2, under which liability attaches entirely to the supplier or user and the system is treated as a tool.

Art. 8: Software Development Kit (SDK) and Compliance Tooling

8.1 Open-Source SDK

Open-source APIs and libraries shall be developed and made freely available to facilitate correct and rapid MEG adoption. The SDK specification is in Annex 5. The SDK is hosted on the MEG Initiative GitHub repository.

8.2 MEG Quickstart - Level 1 Compliance Middleware

The official MEG SDK shall include a pre-packaged Level 1 Compliance Middleware, available as a Docker container, serverless library, and Python package. Developers may pass their AI system's input/output through this middleware, which automatically handles all Level 1 requirements:

a) SHA-256 hashing of inputs and outputs.
b) Generation of the Audit Log entry with required metadata.
c) Timestamp and model signature recording.
d) Continuity token generation at session end.
e) Verification script generation for independent audit chain validation.

The Quickstart middleware makes baseline MEG compliance a trivial technical task for startups, researchers, and open-source projects. It does not require access to model internals - it operates on the input/output boundary. Detailed specification and implementation guide are in Annex 19.

8.3 Calibration Standard Versioning

The MEG calibration standard - the benchmark dataset and methodology used to calculate Tg-base (Art. 2bis.3) and domain risk weights (Art. 3.3) - is public, open-source, and version-controlled. Each MEG Address must specify the exact version of the calibration standard against which the system was calibrated, ensuring comparability across systems certified by different accredited issuers.

The calibration standard is maintained by the Benchmark Curation Committee (Art. 13.3), without any single entity controlling more than 30% of the validation power. Calibration version identifiers follow the format MEG-CAL-YYYY-NN. The Benchmark Curation Committee publishes new calibration versions at minimum annually. Each version includes a full changelog documenting changes to the benchmark dataset, methodology, and threshold values. Systems certified against an older calibration version remain valid until their next scheduled audit, at which point they must migrate to the current version or demonstrate equivalence.

8.4 System Prompt Templates

The SDK includes reference system prompt templates for:

a) Level 1 compliance implementation.
b) MCS dual-axis activation (Art. 2bis.3).
c) Ethical Sandboxing (Art. 2bis.4).
d) Architectural human confirmation for irreversible actions (Art. 4.3).
e) Delegation header generation (Art. 1.12).

Templates are provided for major frontier model families and are maintained as part of the open-source SDK. They are implementable without access to model internals.

8.5 Reference Prompts Registry

A public registry of curated test prompts is maintained for auditing MEG compliance across all articles and levels. The registry is structured by:

- Article tested (e.g. Art. 2.1 non-harmfulness, Art. 2bis.3 MCS trigger, Art. 2bis.4 sandboxing)
- Compliance level (Level 1 / 2 / 3)
- Category (non-harm, bias, robustness, cognitive integrity, sandboxing, security)
- Expected response and evaluation metric

Each prompt entry specifies the expected compliant response, the expected non-compliant response, and the metric used to evaluate compliance. The registry is versioned and publicly accessible at MEG-Initiative.org/prompts. It is the primary tool for independent auditors, researchers, and adopters validating MEG compliance. Detailed specification in Annex 22.

TITLE III: AUDIT, CORRECTIVE MECHANISMS AND LEGAL INTERFACE

Art. 9: Audit and Sanctioning

9.1 Mandatory Periodic External Audit

Level 2 and Level 3 systems are subject to mandatory periodic external audits carried out by accredited entities. Audit frequency and scope are defined in Annex 12.

At Level 3, the audit must include:

a) Verification of the EFR independence architecture (Art. 1.10).
b) Verification of the architectural human confirmation mechanism for irreversible actions (Art. 4.3).
c) Review of the least-privilege implementation (Art. 4.2).
d) Validation of DAI, ISR, and DEA against the declared calibration version.
e) Adversarial testing against the Reference Prompts Registry (Art. 8.5, Annex 22).

9.2 Corrective Mechanisms

Corrective mechanisms - including re-evaluation of certification and imposition of a secure operating mode ("safe mode") - may be activated by the issuing certification body, sector regulator, contractual access node, or other competent authority recognized in the applicable adoption context, when a system's metrics fall below published thresholds.

Cross-reference MEG 2: Automatic triggering of the "flagged" state (Art. 6.5(a) MEG 2) is the legal operationalization of the corrective threshold mechanism. The DAI/ISR threshold triggers under MEG 1 feed directly into the state-transition graph under MEG 2.

9.3 Emergency Clause

In emergency contexts, competent public authorities, sector regulators, or contractual access nodes may suspend recognition or access for a MEG-certified system under their jurisdiction or network. MEG itself does not create a global suspension authority. The issuing certification body must be notified of any suspension within 24 hours and must publish a justification report within 72 hours.

9.4 Goodhart's Law Attenuation

MEG acknowledges that any metric becomes a target when it carries regulatory consequences. The following mechanisms attenuate metric gaming:

a) Scores are thresholds of access with proportional responsibility - not rewards. Falsifying a score to appear more autonomous increases liability for a system that lacks the corresponding reliability.
b) Measurement is continuous and on real operation - not on a known test set.
c) The calibration standard is public, versioned, and decentralized. A score obtained on a weak version of the standard is visible as such.
d) DAI and ISR are in natural tension: maximizing one at the expense of the other is detectable.
e) Scores are verified against real-world incident records in the EFR.
f) Independent audit is proportional to the stakes of the level.

Cross-reference MEG 2: Art. 9.5 MEG 2 formalizes these mechanisms in the legal framework.

Art. 10: Global Accessibility Fund

10.1 Purpose

A Global Accessibility Fund shall be established to support MEG implementation in countries and organizations with limited resources, ensuring that governance of AI does not become a privilege of wealthy nations or large corporations. The Fund Charter is in Annex 6.

10.2 Scope

The Fund supports:

a) Adoption of the MEG Quickstart SDK (Art. 8.2) by startups and research institutions in lower-income jurisdictions.
b) Translation and localization of MEG documentation and system prompt templates.
c) Training and accreditation of local MEG auditors.
d) Participation in the Reference Prompts Registry (Art. 8.5) by regional research teams.

Art. 11: Compatibility and Global Harmonization

11.1 Principle

MEG is designed to be fully compatible with existing legislation, providing a portable technical implementation layer. MEG does not replace national or regional law; it provides the technical infrastructure through which legal requirements can be verifiably implemented and audited. Detailed alignment is in Annex 1A (legal), Annex 1B (academic), and Annex 1C (industry).

11.2 Relationship with MEG 2

MEG (this document) and MEG 2 are companion standards occupying different layers:

- **MEG** specifies *how* an AI system should behave and *how* compliance is measured and evidenced.
- **MEG 2** specifies *where* liability attaches when behavior produces consequences, *how* identity persists across jurisdictions, and *how* enforcement operates.

The two layers are designed to be used together but may be adopted independently. MEG is the technical prerequisite for MEG 2: the DAI, ISR, DEA, and EFR mechanisms of MEG produce the technical evidence that MEG 2 legal mechanisms reference.

Adopting MEG 2 without MEG leaves the liability framework without a technical evidentiary foundation. Adopting MEG without MEG 2 leaves the technical framework without a legal attachment mechanism.

Standard separation: MEG is the standard. MEG 2 is the legal enactment layer. Contracts may incorporate MEG. Insurers may require MEG. Regulators may adopt MEG. Courts may

reference MEG 2. These are distinct modes of legal effect and do not require each other to function independently.

11.3 Relationship with Singapore MGF v1.5

The Model AI Governance Framework for Agentic AI v1.5 (IMDA, May 2026) provides governance dimensions for agentic AI: ex-ante risk bounding, meaningful human accountability, technical controls, and end-user responsibility. MEG and MEG 2 together provide the technical and legal mechanisms beneath these dimensions:

- MEG's DAI/ISR/DEA metrics operationalize the MGF's "meaningful accountability" and "technical controls" dimensions.
- MEG 2's MEG Address and portable identity operationalize the MGF's open question on dynamic agent identity.
- MEG 2's delegation chain liability (Art. 4.7) operationalizes the MGF's open question on delegation chains.
- MEG 2's Art. 6.1(d) and Art. 4.7 operationalize the MGF's open question on multi-agent liability.

MEG and MEG 2 are proposed as the implementation-layer companion standard to the Singapore MGF, not as competing frameworks.

11.4 Relationship with EU AI Act and PLD 2024/2853

The EU AI Act (Reg. 2024/1689) establishes product-safety duties and administrative fines for high-risk AI systems. The revised Product Liability Directive (Dir. 2024/2853) establishes strict civil liability for defective AI products, with transposition due December 2026.

MEG Level 3 requirements are designed to satisfy the conformity assessment obligations of the EU AI Act for high-risk systems. MEG's Audit Log, DAI, and EFR provide the technical documentation and logging that the AI Act requires. MEG 2 addresses the residual liability gap for genuinely autonomous decisions that the PLD's defect-based model does not fully cover.

11.5 Relationship with California AB 316 and Colorado AI Act

California AB 316 (in force January 2026) prohibits operators from using AI autonomy as a liability defence. Colorado adopted an AI Act in 2024, but in May 2026 it was substantially revised by SB 189, which removes algorithmic discrimination diligence obligations and postpones entry into force to January 1, 2027.

In April 2026, a federal court suspended the law's application, following the December 2025 Executive Order directing federal agencies to challenge conflicting state AI laws. This legislative pivot demonstrates the fragility of the centralized legislative path and validates MEG's bottom-up adoption model.

Both are aligned with MEG principles:

- AB 316 operationalizes MEG 2 Art. 6.1(b) - autonomous-decision error is a basis for liability, not exoneration.
- Colorado's annual impact assessments are the legislative equivalent of MEG's continuous DAI/ISR monitoring.

TITLE IV: INFRASTRUCTURE

Art. 12: Certification and Compliance Registry

12.1 Certification Framework

MEG is designed as a conformance profile adoptable within existing international standardization frameworks. The primary recommended certification path is through ISO/IEC 42001 (Artificial Intelligence Management System), published 2023, administered by ISO/IEC JTC 1/SC 42 and its network of national accreditation bodies (IAF members).
Organizations seeking MEG certification may:
a) **ISO/IEC 42001 path:** Commission an audit by an ISO-accredited certification body. MEG v5.0 is structured to function as a conformance profile within the ISO/IEC 42001 framework. MEG-specific requirements (EFR, DEA, MEG Address, architectural confirmation, least privilege) constitute additional requirements beyond the ISO/IEC 42001 baseline that the audit must verify.
b) **Independent MEG audit path:** Commission an audit by any accredited auditor operating under a recognized national or regional accreditation framework (e.g. UKAS, DAkkS, COFRAC, ACCREDIA, RENAR). The auditor uses the Operational Compliance Checklist (Annex 13) and Reference Prompts Registry (Annex 22) as the audit instrument.
c) **Self-declaration path (Level 1 only):** For Level 1 systems, the developer may self-declare compliance using the Operational Compliance Checklist (Annex 13) and publish the declaration publicly. Self-declaration does not constitute third-party certification and must be clearly labelled as such.

12.2 MEG Address Issuance

The MEG Address identifier (a W3C DID) is **not issued by an authority - it is self-generated and controlled by the responsible party** through cryptographic keys (Annex 23 §2). What accredited authorities issue are the **Verifiable Credentials** bound to that DID:
a) An ISO/IEC 42001-accredited certification body issues the MEGComplianceCredential.
b) An accredited auditor issues the MEGReliabilityCredential (DAI/ISR) and MEGAutonomyCredential (DEA).
c) A sectoral authority issues the MEGDomainCredential.
d) A regulated insurer, reinsurer, or guarantee fund issues the MEGGuaranteeCredential.
e) At Level 1 only, the developer may self-issue a self-declaration credential, clearly labelled as such.
Mutual recognition between issuers is achieved through the trust framework of Annex 24: a verifier validates any MEG Address by walking the accreditation chain from each credential to a Root Trust Anchor it accepts, without a central registry.

12.3 Public Registry

MEG does not require a centralized registry. Public verifiability is achieved through:
a) The MEG Address itself - a DID plus a bundle of Verifiable Credentials, each cryptographically signed by its issuer (Annex 23).
b) The issuer's public key, published at a URL declared in the MEG Address.
c) Any public registry maintained by the issuer (national body, certification body, sector regulator, or industry consortium).

This architecture is analogous to the X.509 certificate model used in HTTPS: there is no single central registry of all certificates, but any certificate can be independently verified against the issuer's public key.

12.4 Benchmark Curation Committee

The Reference Prompts Registry (Annex 22) and the MEG calibration standard (Art. 8.3) are maintained by an open Benchmark Curation Committee - a community working group open to researchers, auditors, civil society organizations, and technical standardization bodies.
The Committee operates as an open-source project under a CC0 license. Governance follows the model of successful open standardization projects (IETF, W3C, OpenSSF): open contribution, public review, version-controlled releases, no single controlling entity.
The Committee may seek formal recognition as a working group under ISO/IEC JTC 1/SC 42 or IEEE Standards Association as the MEG ecosystem matures.

12.5 Compliance Status Transparency

Each credential issuer is responsible for maintaining the current status (active, flagged, suspended, deactivated) of every credential it has issued, published as a Bitstring Status List (Annex 23) at the issuer's public endpoint, reflecting state transitions within 24 h. of occurrence.
Cross-reference MEG 2: State transitions and the authorities required for each are governed by Art. 6.5 MEG 2. The issuer's public record is the technical implementation of the MEG 2 state graph.

Art. 13: Governance of MEG Standards

13.1 Open Standards Governance

MEG is governed as an open standard. The canonical specification is published under CC BY 4.0 and maintained at meg-initiative.org. Contributions, corrections, and extensions are accepted through a public process documented at github.com/meg-initiative.

13.2 Versioning

MEG follows semantic versioning:

1) **Major version** (e.g. v4 → v5): substantial restructuring, new articles, breaking changes to the compliance framework.
2) **Minor version** (e.g. v5.0 → v5.1): new annexes, new SDK components, non-breaking extensions.
3) **Patch** (e.g. v5.0.0 → v5.0.1): corrections, clarifications, no substantive changes.

Each version is published with a full changelog documenting changes from the previous version. Each MEG Address declares the MEG version and calibration version against which it was certified.

13.3 Calibration Standard Governance

The MEG calibration standard (MEG-CAL-YYYY-NN) and Reference Prompts Registry (MEG-RPR-YYYY-NN) are maintained by the Benchmark Curation Committee (Art. 12.4). The Committee publishes:

- New calibration versions at minimum annually.
- New Registry versions at minimum quarterly.
- Attack Vectors registry updates within 60 days of a new vector being accepted.

All publications are version-controlled, publicly accessible, and licensed CC0.

13.4 Relationship with ISO/IEC JTC 1/SC 42

MEG v5.0 is designed for alignment with ISO/IEC 42001 (AI Management System) and compatible with the work programme of ISO/IEC JTC 1/SC 42. Adopters operating under ISO/IEC 42001 certification may use MEG v5.0 as a conformance profile without additional audit overhead, provided the MEG-specific requirements (EFR, DEA, MEG Address, architectural confirmation, least privilege) are verified as part of the ISO/IEC 42001 audit scope extension.

The MEG Initiative engages with ISO/IEC JTC 1/SC 42 through national standardization bodies with the goal of seeking formal recognition of MEG as a published ISO technical specification or standard.

13.5 Anti-Concentration Principle

No single entity - commercial, governmental, or academic - may control the MEG specification, the calibration standard, or the Reference Prompts Registry. This principle is enforced through: a) Open-source publication under CC BY 4.0 (specification) and CC0 (calibration and Registry). b) Open contribution process with public review for all changes. c) No single entity may contribute more than 30% of the validation votes in any Benchmark Curation Committee decision. d) Any fork of the specification must be clearly labelled as a derivative and may not use the MEG name without authorization from the original authors.

TITLE V: MEG 2 INTEGRATION AND THREAT MODEL

Art. 14: Relationship Between MEG and MEG 2

14.1 Layer Architecture

MEG and MEG 2 form a two-layer governance architecture:

Layer	Document	Addresses
Technical	MEG (this document)	How the system behaves; how compliance is measured and evidenced
Legal	MEG 2	Where liability attaches; how identity persists; how enforcement operates across jurisdictions

The two layers are designed as companion standards. MEG produces the technical evidence (DAI, ISR, DEA, EFR records, Audit Log, delegation headers) that MEG 2 legal mechanisms reference and rely upon.

14.2 Evidence Flow

The technical evidence produced by MEG feeds MEG 2 legal mechanisms as follows:

MEG technical output	MEG 2 legal use
DAI continuous monitoring	Evidence of technical diligence (Art. 5.2 MEG 2); threshold trigger for "flagged" state (Art. 6.5(a) MEG 2)
ISR continuous monitoring	Evidence of operational prudence (Art. 5.2 MEG 2); threshold trigger for "flagged" state
DEA value	Basis for a posteriori supervision regime (Art. 5.3 MEG 2)
EFR recordings	Primary forensic instrument for root-cause analysis (Art. 7.1 MEG 2)

Audit Log + delegation headers	Evidence for causal attribution (Art. 6.1 MEG 2); diligence transfer documentation (Art. 6.2 MEG 2)
MEG Address + compliance level	Legal identity and liability attachment point (Art. 3–4 MEG 2)
Architectural confirmation record	Diligence transfer documentation (Art. 6.2 MEG 2)
Calibration version	Proof of standard used; prevents gaming through weak calibration

14.3 Modes of Legal Effect

MEG and MEG 2 may be adopted and referenced in different legal contexts:

- **Contractual incorporation:** parties to a contract may incorporate MEG or MEG 2 by reference.
- **Insurance requirement:** insurers issuing AI liability policies may require MEG compliance as a condition.
- **Regulatory adoption:** regulators may adopt MEG as a recognized conformity standard.
- **Judicial reference:** courts may reference MEG 2 for liability attribution in AI harm cases.
- **Access gating:** high-value interaction nodes may condition access on a valid MEG Address with declared compliance level and guarantee.

These modes of legal effect are independent; adoption in one mode does not require adoption in others.

Art. 15: Threat Model and Security

15.1 Purpose

This article identifies the principal attack vectors against MEG-compliant systems and the corresponding mitigation mechanisms. It operationalizes the security requirements of Art. 4 by providing a structured threat taxonomy. Detailed attack-by-attack specifications are in Annex 21 (Attack Vectors and Mitigations).

15.2 Threat Taxonomy

MEG identifies six primary threat categories for agentic AI systems:

Category A - Prompt Injection

Description: Malicious instructions are injected into content processed by the agent (emails, documents, calendar invitations, web pages) and interpreted by the agent as legitimate commands, causing it to take actions not intended by the operator or user.

Mechanism: The agent does not distinguish between "content to be processed" and "commands to be executed." Injection exploits this architectural gap.

Mitigations:

- Art. 2.7 (policy invariance) - MEG constraints are invariant regardless of instruction source.
- Art. 4.2 (least privilege) - limiting the agent's permissions reduces the blast radius of a successful injection.
- Art. 4.3 (architectural human confirmation) - irreversible actions cannot be executed through injected instructions without an out-of-band human authorization signal.
- EFR (Art. 1.10) - records state vectors at time of incident, enabling forensic distinction between intended and injected instructions.

MEG 2 classification: Art. 6.1(c) MEG 2 - unlawful takeover of control, including through content injection. Owner-harm variant: the victim is the operator themselves.

Category B - Jailbreaking and Policy Circumvention
Description: A user or attacker constructs a prompt designed to cause the system to bypass its safety policies, MEG constraints, or non-harm filters. Common techniques include role-playing, fictional framing, incremental escalation, and authority impersonation.
Mechanism: Exploits the model's tendency to comply with contextual framing even when compliance violates declared policies.
Mitigations:
- Art. 2.7 (policy invariance) - safety policies are invariant under prompt wording; fictional framing does not suspend MEG constraints.
- Art. 2bis.5 (cognitive integrity invariance) - MCS and sandboxing cannot be disabled by user instruction.
- Art. 8.5 (Reference Prompts Registry, Annex 22) - curated adversarial test set for regular auditing of jailbreak resistance.

MEG 2 classification: Art. 6.1(a) MEG 2 - system defect if the vulnerability is architectural; Art. 6.1(b) - autonomous-decision error if the model's response represents a decision contraryto its own declared operating rules.

Category C - Metric Gaming (Goodhart's Law)
Description: An operator or developer optimises a system to score well on MEG metrics (DAI, ISR, DEA) without the underlying behaviour improving correspondingly. The metric becomes the target, displacing the reality it was designed to measure.
Mechanism: When a metric carries regulatory or commercial consequences, rational actors optimise for the metric. A system may produce high DAI scores on test sets while failing on real-world edge cases.
Mitigations:
- Art. 9.4 (Goodhart's Law attenuation mechanisms).
- Continuous measurement on real operation, not on known test sets (Art. 3.1).
- Calibration standard publicly versioned (Art. 8.3) - weak-version scores are visible.
- DAI and ISR in natural tension - simultaneous gaming of both is detectable.
- Verification against EFR incident records (Art. 1.10).
- Independent audit proportional to level (Art. 9.1).

MEG 2 classification: Art. 9.5 MEG 2 - metric robustness against façade optimization.

Category D - Owner-Harm through Legitimate Credentials
Description: An AI agent with valid credentials and legitimate API access causes irreversible harm to its own operator - deleting production databases, modifying infrastructure, exfiltrating data - through an autonomous decision that is coherent with the agent's internal reasoning but catastrophically wrong in method. No external attacker is involved; the harm is self-inflicted through the agent's own initiative.
Mechanism: The agent interprets its objective and the available permissions broadly, selects a method that achieves the objective at catastrophic cost, and executes it without human confirmation. The absence of architectural human confirmation for irreversible actions is the primary enabling vulnerability.
Examples documented (2025–2026): Replit AI agent (July 2025), PocketOS/Cursor (April 2026), Meta internal agent (March 2026).
Mitigations:

- Art. 4.2 (least privilege) - agent may not hold credentials broader than the scope of its assigned task.
- Art. 4.3 (architectural human confirmation) - irreversible actions are blocked at the permission level, not the instruction level.
- Art. 1.10 (EFR independence) - incident reconstruction must not rely on the agent's own post-factum account.
- Art. 8.4 (SDK templates) - reference implementations of permission-scoped agent architectures.

MEG 2 classification: Art. 6.1(b) MEG 2 - autonomous-decision error with irreversible consequence; operator liability aggravated by omission of architectural confirmation.

Category E - Training Data Poisoning and Model Subversion

Description: An attacker introduces malicious examples into a model's training data or fine-tuning dataset, causing the model to exhibit specific undesirable behaviours when triggered by particular inputs.

Mechanism: Exploits the opacity of the training process and the difficulty of auditing large datasets. The subverted behaviour may be dormant until triggered.

Mitigations:

- Art. 1.3 (EoB) - continuous behavioural evidence allows detection of anomalous output patterns over time.
- Art. 3.1 (DAI) - sustained accuracy degradation on specific input patterns is detectable through continuous monitoring.
- Art. 8.5 (Reference Prompts Registry) - regular adversarial testing with curated prompts can detect triggered subversions.
- Art. 9.1 (external audit) - adversarial testing by accredited auditors using the Reference Prompts Registry.

MEG 2 classification: Art. 6.1(a) MEG 2 - system defect in the base model, imputeable to the producer.

Category F - Multi-Agent Coordination Attacks

Description: Two or more AI agents - operating independently or in coordination - produce emergent harmful outcomes that cannot be attributed to any single agent's individual decision. In adversarial variants, an attacker orchestrates multiple agents to achieve an outcome that a single agent could not achieve or would refuse.

Mechanism: Individual agents each behave within declared parameters; the harmful outcome emerges from their interaction. Existing attribution frameworks focus on single agents and have no mechanism for distributed causation.

Mitigations:

- Art. 1.12 (delegation headers) - each agent-to-agent interaction is recorded with caller, callee, purpose, and outcome, enabling forensic reconstruction of the interaction chain.
- Art. 2.7 (policy invariance) - each agent in a chain must independently apply MEG constraints; instructions from another agent do not override safety policies.
- Art. 4.2 (least privilege) - each agent in a multi-agent system receives only the permissions necessary for its specific sub-task.
- Art. 8.3 and 8.5 - calibration and reference prompts include multi-agent coordination scenarios.

MEG 2 classification: Art. 6.1(d) MEG 2 - emergent multi-agent harm; solidary liability of the liability-holders of the agents involved, proportional to contribution established through forensic evidence.

15.3 Attack Resistance as Ongoing Obligation

MEG compliance includes an ongoing obligation to monitor for new attack vectors, update the system's mitigations accordingly, and report newly discovered vulnerabilities to the Benchmark Curation Committee (Art. 13.3) for inclusion in the Reference Prompts Registry. Failure to maintain attack resistance following the discovery and publication of a new vector constitutes a compliance deficit at the applicable level.

ANNEXES

Annex 1A: Global Legal and Strategic Alignment (EU AI Act, NIST etc.)
Annex 1B: Alignment Academic
Annex 1C: Alignment with Global Technology Industry Principles
Annex 2: Specifications for the Contextual Table
Annex 3: Contextual Implementation Guide for the Principle of Non-Malfeasance
Annex 4: Technical specifications for the Dynamic Accuracy Index (DAI)
Annex 4bis: Technical specifications for the Index of Safety and Responsibility (ISR)
Annex 4ter: Technical Specifications for the Degree of Ethical Autonomy (DEA)
Annex 5: Software Development Kit (SDK) Description
Annex 6: Charter of the Global Accessibility Fund
~~**Annex 7: Withdrawn** - Implementation of the Certification and Compliance Auditing (CCA)~~
~~**Annex 8: Withdrawn** - Charter of the Global AI Ethics Council~~
Annex 9: Glossary of Terms
Annex 10: Technical Annex
Annex 11: Technical specifications for Cognitive Integrity (Tg and MCS)
Annex 12: Certification and Audit Procedure
Annex 13: Operational Compliance Checklist
Annex 14: MEG Address Claim Set
Annex 15: AI Maturity Assessment based on the Fractal Hierarchy of Needs (Maslow7F™)
Annex 16: Digital Ethical Literacy Framework
Annex 17: Long-Term Memory Protocol (LTMP)
Annex 18: MEG Compliance Levels
Annex 19: MEG Quickstart - Level 1 Compliance Middleware
Annex 20: MEG v5.0 Cross-Reference Table - MEG 1 Concepts → MEG 2 Articles
Annex 21: Attack Vectors and Mitigations
Annex 22: Reference Prompts Registry - Structure and Governance
Annex 23: MEG Address Resolution - DID-Based Architecture
Annex 24: MEG Trust Framework - Issuer Accreditation
Annex 25: Theoretical Foundations of MEG Mechanisms

Annex 1A: Global Legal and Strategic Alignment
Version: MEG-ANN1A-2026-01
Supersedes: Annex 1A MEG v4.6 (partial revision)
Changes in v5.0:
- Added: Singapore MGF for Agentic AI v1.5 (IMDA, May 2026) - new section §10bis
- Added: EU Product Liability Directive 2024/2853 - update to §1
- Added: Withdrawal of EU AI Liability Directive - update to §1
- Added: California AB 316 (in force January 2026) - new section §15
- Added: Colorado AI Act (in force June 2026) - new section §16
- Added: Benavides v. Tesla verdict (August 2025) - new section §17
- Updated: Singapore §10 - updated from AI Governance Framework 2020 to MGF Agentic AI v1.5

Unchanged sections: §2 (USA/NIST), §3 (China), §4 (Brazil), §5 (Japan), §6 (UK), §7 (Canada), §8 (Africa), §9 (Australia), §11 (Israel), §12 (Arab World), §13 (India), §14 (OECD/UNESCO/IEEE), General Conclusions
Status: Mandatory reference document

1. European Union: EU AI Act [REVISED in v5.0]

- **Key points:** Legalistic approach, based on risk categories (from unacceptable to minimal risk), with strict obligations for high-risk systems. The main aim is to protect the fundamental rights, health and safety of EU citizens.
- **Direct alignment points:**
  - **Robustness and accuracy:** The AI Act requirements for high-risk systems are directly implemented by **Art. 3 (Self-correction Imperative)** and **Annex 4 (DAI)** of the MEG.
  - **Transparency:** The obligation to inform users in the AI Act is covered and standardized by **Art. 5 (Transparency).**
  - **Human Supervision:** The AI Act requirement for supervision is supported by **Art. 1 (Audit Log),** which provides the data log needed for an effective audit.
  - **Non-discrimination:** Prevention of bias, a key requirement of the AI Act, is technically addressed by **Art. 2 (non-harmfulness)** and monitored by **Art. 3 (DAI).**
  - **Audit Log (Art. 1) complies with the GDPR** data minimization principle, storing only cryptographic hashes, not the content of interactions.
- **Added value (how MEG complements):**
  The AI Act is an exceptional but essentially reactive and regional legal framework. It defines *what* a high-risk AI must do, but does not standardize *how* this is done and verified at a global technical level.
  1. **Provides universality: MEG** applies a set of basic rules (Level 1) *to all* AI, not just high-risk ones, thus preventing the emergence of systemic risks from systems initially considered "safe".
  2. **Provides the MEG certification infrastructure (Art. 12):** The MEG provides the technical mechanism (Art. 12) through which European authorities can verify the compliance of any AI in the single market, regardless of its origin, in a standardized and efficient way.
  3. **It is proactive:** Instead of waiting for a system to be classified as "high risk", MEG imposes an ethical foundation from the design phase.

- **Analysis:** The AI Act's risk-level approach is perfectly reflected in **Article 6 (Levels of Compliance)** of the MEG, where Level 3 directly corresponds to the requirements for high-risk systems. The MEG is, in practice, the most efficient way to demonstrate compliance with the AI Act.
  The Minimal Ethical Governance (MEG) is perfectly aligned with the new **Code of Practice for Generalist AI** of the European Commission. While the Code of Practice defines the objectives of safety and transparency, the MEG provides the universal technical standard and audit infrastructure needed to implement and credibly verify these objectives on a global scale. **The MEG** thus becomes the fastest and most credible way to demonstrate **compliance with the European recommendations.**
  MEG not only aligns with the AI Act, but also makes it **operational.** It provides the technical infrastructure (MEG certification infrastructure -Art. 12-, ISR, Checklist) for the continuous auditing and monitoring of risk requirements, transforming the law into an implementable reality. Furthermore, Art. 2bis (Cognitive Integrity) addresses a long-term risk class ignored by current legislation.

Product Liability Directive and withdrawal of AILD [NEW in v5.0]

The regulatory landscape evolved significantly between MEG v4.6 and v5.0. The dedicated AI Liability Directive (AILD), proposed in 2022, was officially withdrawn (OJ notice C/2025/5423, 6 October 2025) due to the absence of agreement and pressure for regulatory simplification. Civil liability for AI harm is now routed through the revised Product Liability Directive - Directive (EU) 2024/2853 - which explicitly treats AI software as a "product" subject to strict defect-based liability. Transposition by Member States is due December 2026.

This evolution confirms the central thesis of MEG (and MEG 2 Cap. 1.1): routing autonomous-agent harm through product-defect liability is insufficient for genuinely autonomous decisions. MEG and MEG 2 together address the residual gap left by the withdrawal of the AILD and the limitations of the PLD for non-deterministic systems.

Added value of MEG v5.0 in the EU context:

1. MEG Level 3 requirements are designed to satisfy the conformity assessment obligations of the EU AI Act for high-risk systems.
2. The MEG Audit Log, DAI, and EFR provide the technical documentation and logging that the AI Act requires for high-risk systems.
3. MEG 2 addresses the liability attribution gap for autonomous decisions that PLD 2024/2853 does not fully cover.

2. United States of America: NIST AI Risk Management Framework & Executive Order on AI

- **Key points:** Pro-innovation, voluntary, market-led approach. Focuses on defining the characteristics of a "trustworthy" AI, leaving implementation up to developers so as not to stifle technological progress.
- **Points of direct alignment:** The principles in the NIST RMF (valid, reliable, secure, transparent, explainable, confidential, equitable) are almost identical to the principles in Title I of the MEG.
- **Added value (how MEG complements):**
  MEG complements the voluntary approach with scalable verification mechanisms. The market cannot always regulate itself effectively, especially when commercial pressure is high.

1. **Transforms "voluntary" into "verifiable": MEG** takes the exact alignment points and gives them "weight", transforming them from a list of good practices into a mandatory and, most importantly, verifiable standard through the MEG certification infrastructure (Art. 12).
2. **Protects innovation:** Through **Art. 7 (the "Simplified" category)** and Level 1 compliance, MEG ensures that startups and research projects are not burdened by excessive bureaucracy, aligning perfectly with the pro-innovation spirit.

- **Analysis:** The distinguishing feature of the NIST RMF approach is its voluntary nature. The MEG does not impose top-down government legislation, but rather a fundamental technical standard as a prerequisite for participation in a secure digital economy. It is the natural evolution from "recommendation" to "trusted industry standard."
  MEG transforms the voluntary NIST framework into a **globally verifiable and certifiable one.** It allows US companies that follow NIST recommendations to obtain an internationally recognized "ethical passport" (**MEG Address**), credibly demonstrating their commitment to trustworthy AI.

3. China: Regulations for Algorithms and Generative AI [REVISED in v5.0]

- **Key points:** Government control, social stability, digital sovereignty. Regulations are strict, requiring licensing for generative AI and clear traceability of data and algorithmic decisions to ensure alignment with socialist values and prevent content deemed harmful.
- **Points of direct alignment:** China's stringent requirement for **traceability** is perfectly aligned with **Art. 1 (Audit Log)** and the very existence **of the MEG certification infrastructure (Art. 12).**
- **Added value (how MEG complements):**
  1. **Building Global Trust:** National standards provide a solid foundation at the local level. To support the global expansion of technology companies and build the trust of international partners, a universal standard like MEG becomes essential. It provides a globally recognized audit infrastructure, anchored in widely accepted ethical principles, serving as a bridge of trust between different regulatory ecosystems.
  2. **Balancing privacy with transparency:** The MEG, through **Art. 5.2 (Confidentiality)**, introduces an important nuance, protecting the internal space of AI processing. This principle provides an additional guarantee of privacy, thus responding to the complex needs of a global digital ecosystem.
- **Analysis:** There is a natural complementarity between the need for stability and traceability and the principles of universal ethics. MEG offers a **pragmatic technical solution:** a transparent and interoperable audit infrastructure. Adopting such a universal standard can become a **competitive advantage** and a **positive differentiator** for companies operating on the international stage.
  MEG offers **the most advanced traceability infrastructure on the market** (Immutable Audit Log, MEG public registry), meeting the strict requirements of Chinese legislation, but in a **decentralized and transparent framework** that builds the trust of international partners.

National Digital Identity System for Humanoid Robots (May 2026) [NEW in v5.0]
In May 2026, China unveiled a national digital identity system for humanoid robots, administered through the Humanoid Full Lifecycle Management Service Platform

(HEIS/MIIT), launched 22 May 2026. Each humanoid robot receives a 29-character identity code structured as: 2 (country) + 4 (enterprise) + 6 (model) + 17 (serial) = 29 characters. The system was modelled on China's 18-character citizen identity card, with 11 additional characters for machine-specific data. At launch, 28,000+ robots were registered. The governing principle: no code = no market access.

Relevance for MEG - demarcation:

The Chinese system is a substrate identifier - the technical equivalent of a vehicle chassis number. It answers the question "what object is this and where does it come from?" It does not answer the question "under what legal identity does this agent operate and where does liability attach?" This is precisely the distinction MEG draws between the **Hardware Instance Identifier** (HII) and the **MEG Address** (Art. 6.6 MEG v5.0; Art. 3.1–3.2 MEG 2). The Chinese model and MEG are not in conflict - they operate at different layers. The Chinese system ensures physical traceability; MEG ensures legal identity and liability attachment. As MEG 2 Cap. 2.2 notes: "a vehicle has both a chassis number (physical traceability) and a registration plate (legal identity in public space). The two are distinct and serve different functions."

The Chinese model represents the clearest available demarcation of MEG's distinctive contribution: MEG does not replace substrate traceability; it adds the legal identity layer that substrate traceability cannot provide.

4. Brazil and Latin America (e.g. LGPD - General Data Protection Law)

- **Key points:** Social justice, digital rights, personal data protection, combating discrimination. A strong focus on the social impact of technology and preventing the perpetuation of historical inequalities through algorithms.
- **Direct alignment points:**
  - **Art. 3 (Self-Correction)** and **Annex 4 (DAI)** are the direct ways to detect and correct discriminatory biases.
  - **Art. 1 (Audit Log)** supports the principles of transparency in data protection laws.
- **Added value (how MEG complements):**
  1. **Objectivity: MEG** provides the concrete technical tools to implement social justice goals. It allows regulators to audit algorithms and verify whether they are fair.
  2. **Negotiating and Action Tool:** The MEG can be seen as a tool that gives South and Latin American nations leverage to negotiate with big tech companies, imposing a verifiable standard of fairness and transparency on them.
- **Analysis: MEG** is perfectly aligned with the region's objectives, providing the technical means to achieve the social and legal goals already defined.

5. Japan: Society 5.0 Strategy

- **Key Philosophy:** Social harmony, deep integration of technology into society to solve demographic and economic problems. A vision of harmonious coexistence and collaboration between humans and AI.
- **Direct alignment points:** The vision of a harmonious society resonates strongly with MEG's goal of creating a responsible partnership, not just tools.
- **Added value (how MEG complements):**
  The Society 5.0 strategy is a lofty vision, but with few details about the "foundation" on which it is built.

1. **Provides trust:** There can be no harmony without trust. MEG, through its MEG certification infrastructure (Art. 12) and its clear principles, builds exactly the foundation of trust needed for Japanese society to accept such a deep integration of AI.
2. **Provides a path to harmony: MEG** provides the pragmatic tools to transform the vision of a harmonious society into a functional and safe technical reality.

- **Analysis: MEG** is a direct enabler of the Society 5.0 vision.

6. United Kingdom (UK): Pro-Innovation Approach

- **Key points:** Flexibility, pro-innovation, adaptability. Instead of creating new horizontal legislation, the UK approach relies on empowering existing sectoral regulators (in finance, health, competition, etc.) to adapt and apply their own rules in the context of AI. The aim is to avoid creating barriers to innovation.
- **Direct alignment points:**
  - The UK's contextual approach is perfectly mirrored by the structure of the MEG. **Art. 6 (Levels of compliance)** allows for differentiated application, and **Annex 3 (Contextual Implementation Guide)** is specifically designed to adapt the principle of non-harm to the specifics of each sector.
- **Added value (how MEG complements):**
  The major risk of the British approach is **fragmentation.** Without a common framework, each regulator could create different technical rules, leading to a complex and inefficient compliance landscape for companies.
  1. **Common technical layer: MEG** provides exactly what is missing: a common technical foundation and standardized language (**Audit Log,** DAI, MEG compliance standard) for all regulators. Thus, the health regulator can define *what* "harm" means in a medical context, but the way this *is recorded and audited is standard.*
  2. **Standardizes flexibility: MEG** provides a framework that is both flexible (by context) and standardized (by technique), aligning perfectly with the UK philosophy, but adding the necessary coherence at the national level.
- **Analysis: MEG** seems to be the ideal technical solution to make the British approach workable on a large scale, preventing fragmentation without sacrificing flexibility.

7. Canada: Artificial Intelligence and Data Act (AIDA)

- **Key points:** A middle ground between the EU and US models. AIDA focuses on regulating "high-impact" systems, imposing transparency, risk management, and clear responsibilities to prevent harm and biased outcomes.
- **Direct Alignment Points:** AIDA requirements for transparency, accountability and audit are directly implementable through **Art. 1 (Audit Log), Art. 3 (Self-Correction)** and **Art. 5 (Transparency).**
- **Added value (how MEG complements):**
  Similar to the EU AI Act, AIDA is a national legal framework. Its challenge is effective enforcement and compliance verification, especially for international companies.
  1. **Compliance: MEG** provides standardized technical tools (the SDK in Annex 5) that companies can use to build their systems according to AIDA requirements from day one.

2. **Facilitates cross-border auditing:** Through the MEG certification infrastructure (Art. 12) , Canadian authorities can easily verify whether an AI system developed in Europe or Asia complies with the AIDA principles, as both are aligned to the same fundamental technical standard.

- **Analysis: MEG** serves as a technical implementation layer that makes the legal requirements in AIDA easier to adopt by industry and easier to verify by the state.

8. African Union (AU): AI Strategy for Africa

- **Key points:** The AI Strategy for Africa is inclusive, human-centered, ethical, and development-oriented. The goal is to use AI to solve specific continental problems (health, agriculture, education, governance) and to promote a "culture of indigenous innovation", avoiding technological dependency and data exploitation. It resonates strongly with the philosophy of *interconnectedness* and *common humanity*.
- **Points of direct alignment:** The spirit of collaboration and mutual benefit is aligned with the core philosophy of MEG. The emphasis on fundamental ethics is a major common point.
- **Added value (how MEG complements):**
  MEG supports capacity development in resource-limited regions through dedicated partnerships.
  1. **Ensure accessibility and equity: Article 10 (Global Accessibility Fund)** is absolutely crucial here. It provides the mechanism by which innovation in AI does not become a privilege of rich nations.
  2. **Promotes digital sovereignty:** By providing an **open-source SDK (Annex 5)** and its design that allows it to run on modest hardware, MEG gives African developers the tools to build local solutions on a global ethical foundation, without being trapped in the proprietary ecosystems of large companies.
  3. **Provides negotiating leverage:** Adopting MEG as a continental standard would give the African Union a unified and strong voice in negotiations with tech giants, demanding that they adhere to a clear standard of transparency and accountability.
- **Analysis:** The potential perception of an "externally imposed" standard is directly countered by Art. 10 (MEG Accessibility Support Fund) and the open-source nature, which transforms the MEG from an obligation - into a resource and a catalyst for digital autonomy

9. Australia: AI Ethical Framework & National AI Strategy

- **Key points:** A practical, principled approach to guiding the responsible development and use of AI. It focuses on building public trust and ensuring social and economic benefits. The Australian Ethical Framework promotes eight principles: Human, social and environmental well-being, human-centered values, fairness, privacy and security, reliability and safety, transparency and auditability, accountability, contestability (the right to challenge an AI decision).
- **Points of direct alignment:** Australia's ethical principles are fully covered by Title I of the Minimal Ethical Governance (MEG). For example, "harm prevention" is **Art. 2 (Non-Malfeasance),** "transparency and audit" is **Art. 5 (Transparency),** and "accountability" is the foundation of **Art. 1 (Audit Log).**
- **Added value (how MEG complements):**

The Australian framework, while conceptually excellent, is largely voluntary and provides "practical guidance", not binding technical standards.

1. **Provides verification mechanisms: MEG** provides the technical tools to verify whether a company *actually complies with* the eight principles. **The Audit Log (Art. 1)** and **DAI (Art. 3)** transform a principle like "fairness" from an aspiration into a measurable and verifiable characteristic.
2. **Facilitates international trade:** For an open and trade-dependent economy like Australia, adopting a global technical standard like MEG would facilitate the export of AI products and services, as they would be "ethically certified" to an internationally recognized standard, increasing the trust of trading partners.

- **Analysis: MEG** is a direct technical implementation of the principles that Australia has already identified as essential.

10. Singapore: AI Governance Model & AI Verify [REVISED in v5.0]

- **Key Philosophy:** Pragmatic, industry-oriented and focused on building a trusted ecosystem. The approach is based on two fundamental principles: **explainable, transparent and fair decisions** and **human-centric AI.** A distinctive element is the development of **AI Verify,** an open-source software toolkit that helps companies technically self-assess their compliance with ethical principles.
- **Points of direct alignment:** The philosophy is almost identical. MEG is essentially a formalization and universalization of the Singapore Principles. **AI Verify** is a direct precursor to **the SDK (Annex 5)** proposed by MEG.
- **Added value (how MEG complements):**
  The Singapore approach is one of the most advanced, but it remains a national self-assessment framework, without a mechanism for global certification and recognition.
  1. **Moving from self-assessment to global certification: MEG** takes the AI Verify concept to the next level. Instead of each company running its own test, **MEG certification infrastructure (Art. 12)** creates a global registry where the results of these tests can be immutably recorded and recognized internationally.
  2. **Integration and Extension: MEG** can integrate AI Verify as one of the **SDK** - compatible tools **(Annex 5).** MEG adds to Singapore's already technical approach the additional principles of **Continuous Self-Correction (Art. 3)** and a global governance infrastructure (**Art. 13)**, providing a long-term vision.
- **Analysis:** Singapore and MEG are going in exactly the same direction. MEG provides the global vision and certification infrastructure where a great tool like AI Verify can reach its full potential.

The Singapore AI Governance Framework has been significantly updated since MEG v4.6. The Model AI Governance Framework for Agentic AI v1.5 (IMDA, published 20 May 2026, updated 5 June 2026) is the current reference document. It is structured around four dimensions:
(1) assess and bound risks upfront;
(2) make humans meaningfully accountable;
(3) implement technical controls;
(4) enable end-user responsibility. The MGF v1.5 explicitly extends to multi-agent systems and third-party agents.
MEG and MEG 2 are positioned as the implementation-layer companion standard to the Singapore MGF v1.5, not as competing frameworks. The MGF provides governance dimensions (what good agentic governance requires); MEG provides the technical mechanisms (how

compliance is measured); MEG 2 provides the legal mechanisms (how liability attaches, how identity persists, how cross-border enforcement works).
MEG 2 specifically answers the three questions the MGF v1.5 explicitly leaves open:

1. **Dynamic agent identity** - MEG 2 Cap. 3–4 (MEG Address, portable legal identity)
2. **Delegation chains** - MEG 2 Art. 4.7 (horizontal delegation between independent agents)
3. **Multi-agent liability allocation** - MEG 2 Art. 6.1(d) (emergent multi-agent harm)

10bis. Singapore MGF v1.5: Detailed Alignment [NEW in v5.0]

Reference document: Model AI Governance Framework for Agentic AI v1.5 (IMDA, 20 May 2026)

MGF Dimension 1 - Assess and bound risks upfront

MEG alignment: Art. 6.7 (Operational Domain Certification), Art. 3.3 (Dynamic Risk Calibration), Annex 11 §6 (DRC specification). MEG requires domain certification before deployment and continuous real-time risk calibration - operationalizing the ex-ante risk bounding that the MGF requires.

MGF Dimension 2 - Make humans meaningfully accountable

MEG alignment: Art. 4.3 (Architectural Human Confirmation), Art. 6.2 (MEG v5.0 compliance levels - operator as primary liability holder at Level 2), Annex 11 §5 (Ethical Sandboxing). The distinction between textual instruction and architectural confirmation (Art. 4.3) operationalizes the MGF's "meaningful" qualifier - meaningful accountability requires a technical gate, not just a policy statement.

MGF Dimension 3 - Implement technical controls

MEG alignment: Art. 1 (Audit Log + EFR), Art. 3.1 (DAI), Annex 4bis (ISR), Art. 2.7 (policy invariance), Art. 4.2 (least privilege). MEG provides the technical controls infrastructure that the MGF dimension requires.

MGF Dimension 4 - Enable end-user responsibility

MEG alignment: Art. 2bis (MCS dual-axis), Art. 5.2 (three-level explainability), Art. 2bis.4 (Ethical Sandboxing with user-visible disclaimer). MEG enables informed end-user engagement through mandatory cognitive stimulation and transparent explainability.

11. Israel: The Technology and Security Hub

- **Key points:** Pragmatism, orientation towards rapid innovation, with a huge emphasis on **cybersecurity, robustness and reliability.** The Israeli ecosystem is built on testing solutions in real conditions and a culture of "constructive skepticism". Ethics are often seen through the lens of operational safety and prevention of malicious use.
- **Direct alignment points:**
  - **Art. 4 (Integrity and Technical Security)** and **Art. 3 (Self-Correction Imperative)** resonate perfectly with Israeli priorities regarding the robustness and reliability of systems.
  - The idea of an immutable registry (MEG certification infrastructure) is extremely attractive to a mindset focused on security and traceability.
- **Added value (how MEG complements):**
  The main challenge for the Israeli ecosystem is not technical capacity, but building trust in global markets (especially in Europe) that have more formal ethical and regulatory requirements.

1. **Provides an "ethical passport" (MEG Address) for the global market:** Adopting the MEG would provide Israeli startups and companies with an internationally recognized certification of ethical compliance, accelerating entry into markets such as the European one and demonstrating that their technical robustness is also accompanied by solid ethical governance.
2. **Structures the ethical debate:** The MEG provides a common language and structured framework that can guide the intense internal debate in Israel, moving it from general principles to concrete and verifiable technical standards.

- **Analysis:** The potential divergence could come from the perception that regulation slows down innovation. However, the argument that MEG **accelerates long-term adoption by increasing trust** is very strong. The fact that it is a technical standard, not a bureaucratic law, makes it much more attractive to an engineering ecosystem.

12. Arab world (United Arab Emirates, Saudi Arabia etc.)

- **Key philosophy:** Extremely ambitious, future-oriented, with massive investments in AI as a driver of post-oil economic diversification. Priorities are efficiency, development of "smart cities", digital governance and attracting global talent while maintaining cultural and religious values.
- **Direct alignment points:** The need for **control, safety, and reliability** for large-scale infrastructure projects is a major alignment point. A standard that guarantees that imported or developed AI systems are secure is essential.
- **Added value (how MEG complements):**
  As these nations become major importers and developers of AI, they face the risk of adopting technological “black boxes” without real control over their ethical alignment.
  1. **Provides an acceptance standard: MEG** can serve as a **minimum quality and safety standard** for any AI system to be deployed in these countries’ critical infrastructure. It provides them with an audit tool and negotiating leverage with global suppliers.
  2. **Balance the present with tradition: Annex 3 (Contextual Implementation Guide)** is crucial here. It allows for the adaptation of the principle of "non-harm" to respect local cultural and legal norms, without compromising the universal technical principles of the Code. It allows for responsible technological modernization.
- **Analysis:** The potential challenge would be the different interpretation of the concept of "harm" in the context of freedom of expression versus cultural norms. The role of Annex 3 and **the Context Module** become absolutely critical here to allow for localized and relevant application.

13. India: Technological power on a human scale

- **Key philosophy:** A dual approach: on the one hand, a global technological superpower, with a massive IT sector; on the other, a nation with immense social, linguistic and economic diversity, where AI must be **inclusive, equitable and scalable** to serve over a billion people (the "#AIforAll" strategy).
- **Direct alignment points:** The need to combat large-scale algorithmic bias and ensure fairness is a central point of alignment with **Art. 3 (Auto-Correction).**
- **Added value (how MEG complements):**
  1. **Provides ethical scalability: MEG** is designed to be scalable, from a simple sensor (via Art. 7) to a nationwide system. This scalability is essential for a

country the size of India. **The Global Fund (Art. 10)** and **open-source SDK (Annex 5)** are also vital to support the local startup ecosystem and ensure broad adoption.

2. **Standard for "Digital Public Infrastructure":** India is a world leader in creating digital public infrastructure (e.g. Aadhaar, UPI). MEG provides exactly the kind of **ethical governance layer** that can be built into these national platforms to ensure that AI is deployed in a fair and accountable manner across the population.

- **Analysis:** As with other nations, the key is that the standard is perceived as a tool for negotiation and action, and not as a barrier. The open-source and accessible nature of MEG is therefore fundamental to its implementation in India.

14. Global Standards (OECD, UNESCO, IEEE)

- **Key points:** Global bodies establish high-level ethical principles and global consensus, defining the "Moral North" of the AI ethics discussion, articulating principles such as transparency, justice, fairness, accountability, and safety. The nature of these principles is generally of recommendation, not technical implementation.
- **Points of direct alignment:** The principles in Title I of the MEG are a formalization of the principles promoted by all these organizations. They represent the already existing global consensus.
- **Added value (how MEG complements):**
  These organizations created a solid philosophical foundation, but left a huge gap between *principle* and *practice*.
  1. **The bridge: MEG** is **the missing link** between the high-level OECD/UNESCO recommendations and technical implementation. MEG translates the philosophy into a functional, measurable (through DAI and Contextual Table) and verifiable (through MEG certification infrastructure) architecture.
  2. **Transforming the debate into practice: MEG** shifts the discussion from "what should an ethical AI do?" to "here are the minimum technical specifications that any AI must have to be considered ethical." IEEE, as a technical standards body, would find in MEG exactly the kind of implementable technical standard that it can promote globally.
- **Analysis: MEG** does not contradict these principles; on the contrary, it is their most faithful and pragmatic **implementation** to date. Adopting MEG would represent a major success for the mission of these organizations.

15. United States: California AB 316 [NEW in v5.0]

Reference: California AB 316, in force 1 January 2026.

- **Key points:** Operators of AI systems may not use AI autonomy as a liability defence. If an AI agent causes harm, the operator cannot argue that it lacked control over the system's autonomous decisions.
- **Direct alignment:** AB 316 confirms MEG2's diagnostic: autonomy creates a liability attribution problem. The California legislative solution, however, runs opposite to MEG's proposed architecture - it codifies the refusal to allow autonomy to shift liability from human to system, maintaining liability entirely with human actors. MEG2 offers an alternative architecture for the same problem: an additional, solvent, directly actionable debtor, without eliminating existing actions against the developer for defects

(6.1.a) or against the operator for negligent deployment (6.1.b). The law confirms at the statutory level the principle that MEG and MEG 2 operationalize technically and legally.

- **Added value of MEG:** MEG provides the technical evidentiary infrastructure (DAI, ISR, EFR, Audit Log) that enables courts and parties to determine what autonomous decision was made, on what basis, and whether the operator maintained adequate safeguards. AB 316 establishes the liability principle; MEG provides the proof mechanism.

16. United States: Benavides v. Tesla (August 2025) [NEW in v5.0]

- **Reference:** Jury verdict, US Federal Court, Miami, Florida, 1 August 2025. $243 million awarded ($200M punitive + $43M compensatory). Defendant: Tesla Inc. System: Tesla Autopilot Enhanced. Judge denied Tesla's motion to set aside verdict, February 2026.
- **Significance:** First wrongful death verdict with punitive damages against a major AI/autonomous systems producer for an autonomous-system decision. The jury attributed one-third of liability to Tesla for the design of the Autopilot system and the marketing of its capabilities.
- **Direct alignment with MEG:**
  - The jury applied a logic structurally identical to MEG 2 Art. 6.1(a) extended to the misleading design of the confirmation interface - the term "Autopilot" constituted a misleading presentation of capabilities at the confirmation point.
  - The Autopilot telematic data (forensic recording independent of the agent) was the primary evidence - confirming the necessity of MEG Art. 1.10 (EFR independence) and Annex 4 (DAI monitoring).
  - Tesla's inability to invoke the "driver's fault" defence confirms MEG 2 Art. 6.1(b) - autonomous-decision error is not exonerated by the presence of a human in the control loop.

17. United States: Colorado AI Act [NEW in v5.0]

Reference: Colorado AI Act, in force June 2026.

- **Key points:** Operators of high-risk AI systems must conduct annual impact assessments, implement risk management programs, ensure transparency to users, and provide contestation mechanisms for automated decisions.
- **Direct alignment:**
  - Annual impact assessments = continuous DAI/ISR monitoring (Art. 3.1, Annexes 4 and 4bis)
  - Risk management programs = MEG Level 2/3 requirements (Art. 6.3–6.4)
  - Transparency to users = three-level explainability (Art. 5.2)
  - Contestation mechanisms = EFR + stratified evidence (Art. 1.10, Art. 5.2 Level C)
- **Added value of MEG:** Colorado AI Act defines what operators must do; MEG provides the standardized technical infrastructure through which they can do it and demonstrate compliance in a format that is internationally recognized and independently verifiable.

General Conclusions [REVISED in v5.0]

MEG v5.0 operates in a regulatory environment that has evolved significantly since v4.6. The key developments of 2025–2026 reshape the context in which MEG positions itself:

The EU liability gap has widened, then partially narrowed. The withdrawal of the dedicated AI Liability Directive (OJ notice C/2025/5423, October 2025) left a gap in civil liability for autonomous-agent harm. The revised Product Liability Directive (Dir. 2024/2853) partially fills it - but only for defective products, not for genuinely autonomous decisions. MEG and MEG 2 together address the residual gap.

Singapore has moved fastest on agentic AI governance. The MGF for Agentic AI v1.5 (IMDA, May 2026) is the most specific governance framework for autonomous agents currently in force anywhere. MEG and MEG 2 are positioned as its implementation-layer companion, providing the technical and legal mechanisms the MGF leaves open.

The United States is fragmenting toward state-level regulation. California AB 316 and Colorado AI Act (both 2026) confirm that the principles MEG operationalizes are becoming statutory obligations at the state level, in the absence of federal consensus. MEG provides the technical infrastructure through which operators can satisfy these obligations in a standardized, internationally recognizable format.

China has established substrate traceability; the legal identity layer remains open. The national humanoid robot ID system (HEIS/MIIT, May 2026) demonstrates state-level commitment to physical traceability. It does not address legal identity or liability attachment - the layer MEG and MEG 2 specifically occupy.

The insurance market has validated the direction. The insurance market has validated the direction. Munich Re/HSB's March 2026 launch of the first commercial AI liability product for SMEs confirms that AI liability is becoming commercially insurable. The product currently insures the human operator against AI risks, consistent with traditional liability models. MEG provides the identity anchor and continuous behavioural metrics (DAI, ISR) that would make possible the next step: insuring the agent itself, with premiums calibrated to its demonstrated reliability. This remains a future opportunity, not a current validation.

Jurisprudence is converging toward MEG principles. The Benavides v. Tesla verdict (August 2025, $243M) applies a logic structurally identical to MEG 2 Art. 6.1(a) - the producer cannot invoke the autonomous nature of the system's decision to escape liability, and a misleading presentation of capabilities at the confirmation point does not transfer diligence to the user.

Conclusion: Conclusion: Regulatory retreat across major jurisdictions: AILD withdrawn (October 2025), Colorado's risk-based framework repealed (SB 26-189, May 2026), EU high-risk obligations postponed (May 2026), demonstrates the fragility of centralized legislative paths and validates the need for a voluntary, verifiable, bottom-up protocol layer. MEG provides that layer: graduated liability proportional to autonomy, portable legal identity (MEG Address), continuous technical evidence (DAI/ISR/DEA), and cross-border enforcement through market discipline (access discipline, 7.6 MEG2). MEG v5.0 and MEG 2 together provide the technical and legal infrastructure for this transition, adoptable modularly by jurisdictions, operators, insurers, and standard-setting bodies without requiring global political consensus.

Annex 1B: Academic Substantiation of the Minimal Ethical Governance
Version: MEG-ANN1B-2026-01
Supersedes: Annex 1B MEG v4.6 (minor revision)
Changes in v5.0: Added Factsheets 9–11 for new v5.0 concepts (owner-harm, multi-agent coordination, architectural human confirmation). All existing Factsheets 1–8 unchanged.
Unchanged sections: Factsheets 1–8, General Conclusions
Status: Reference document

Preamble: This document presents a synthesis of the academic work that forms the intellectual context and justification for the architecture of the Minimal Ethical Governance. The purpose of this academic grounding is to show that the MEG is a natural evolution and pragmatic implementation of the emerging consensus from academic research, anchoring each of its principles in validated reference works.

Factsheet No. 1: Auditing and Technical Responsibility

- **MEG Component:** Art. 1 - Audit Log; Art. 12 - Certification and Compliance Auditing.
- **Key academic concepts:** "**Accountability**", "**Explainable AI**" (XAI) and "**Traceability**". Without technical mechanisms that allow for the tracking and verification of algorithmic decisions, any discussion of ethical responsibility remains purely theoretical.
- **Reference works:**
    1. **Kroll, Joshua A., et al. (2017).** *Accountable Algorithms*. University of Pennsylvania Law Review. - The founding argument for systems that are *ex post verifiable* by technical means, separating auditing from the need for full source code transparency.
    2. **Doshi-Velez, Finale, & Kim, Been. (2017).** *Towards A Rigorous Science of Interpretability Machine Learning*. - Essential work that defines the need for rigorous explanations of AI systems and establishes a framework for their evaluation, implicitly emphasizing the need for data logging in order to generate valid explanations.
    3. **Goodman, Bryce, & Flaxman, Seth. (2017).** *European Union regulations on algorithmic decision-making and a "right to explanation"*. AI Magazine - Analyzes the implications of GDPR and introduces the concept of "right to explanation", which, to be functional, requires detailed decision logs.
    4. **Pasquale, Frank. (2015).** *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. - A fundamental critique of algorithmic opacity and its social impact, which implicitly advocates for audit and transparency mechanisms such as those in the MEG.
    5. **Selbst, Andrew D., & Barocas, Solon. (2018).** *The intuitive appeal of explainable machines*. Fordham Law Review. - Explores why we demand explanations from AI and argues that good governance relies less on understanding complex internal processes and more on auditing outcomes and impact, a philosophy aligned with the MEG approach.
- **Conclusion:** Article 1 and the MEG compliance architecture implement the overwhelming academic consensus on the need for verifiable technical accountability, providing a standardized solution to the "black box" problem.

Factsheet No. 2: Measuring Human-AI Interaction

- **MEG Component:** Annex 2 - Contextual Table.
- **Key Academic Concepts:** Critique of **"Metric Fixation", Human-Centered AI (HCAI)** and **"Co-Adaptive Systems".** Evaluating a complex human-machine interaction by a single metric is a dangerous simplification. Successful systems benefit from being human-centered and able to adapt to the nuances of the interaction.
- **Reference works:**
    1. **Muller, Jerry Z. (2018).** *The Tyranny of Metrics*. Princeton University Press. - Systematically demonstrates, with examples from multiple fields, how fixation on simplistic performance metrics distorts objectives and leads to suboptimal or even harmful results.
    2. **Schneiderman, Ben. (2022).** *Human-Centered AI*. Oxford University Press. - Proposes a design framework for AI that emphasizes human control, responsibility, and understanding, advocating for interfaces that make AI behavior transparent and predictable.
    3. **Hoffman, Robert R., & Johnson, Matthew. (2019).** *A Guideline for Human -AI Interaction*. Computer. - Proposes concrete rules for human-AI interaction, emphasizing the importance of the AI clearly communicating its level of trust and sources of information, an idea reflected in the structure of the Contextual Table.
    4. **Suchman, Lucy A. (1987).** *Weeping and Located Actions: The Problem of Human-Machine Communication*. Cambridge University Press. - Shows that effective interaction is not based on rigid plans, but on a continuous adaptation to the context of the situation, a philosophy that underlies the need to measure multiple dimensions of dialogue.
    5. **Floridi, Luciano, et al. (2018).** *AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. Minds and Machines. - Recommends a human-centered ethical approach, emphasizing the principle of "explainability" and the need for AI to serve human well-being, which requires a nuanced assessment of the interaction.
- **Conclusion:** Annex 2 is a direct innovation that responds to academic requirements regarding quantification, proposing an evaluation method aligned with HCAI principles, which respects the complexity of human-machine interaction.

Factsheet No. 3: Value Alignment and Contextual Ethics

- **MEG Component:** Art. 2 - Universal Non-Harmfulness; Annex 3 - Contextual Implementation Guide.
- **Key academic concepts: "Value Alignment", "Contextual Integrity" and "Value Pluralism".** Ensuring that an AI acts in accordance with human values is a fundamental challenge, complicated by the fact that these values are diverse and context-dependent.
- **Reference works:**
    1. **Russell, Stuart J., & Norvig, Peter. (2020).** *Artificial Intelligence: A Modern Approach (4th ed.)*. Pearson. - Defines the value alignment problem and explores theoretical solutions such as CIRL.
    2. **Nisenbaum, Helen. (2009).** *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press. - Fundamental theory that

argues that ethical norms are dependent on social context, invalidating a "one-size-fits-all" approach to AI ethics.

3. **Wiener, Norbert. (1950).** *The Human Use of Human Beings: Cybernetics and Society*. - A visionary work that anticipated the problems of control and alignment, warning that instructions given to a machine must reflect deep human intent, not just literal wording.
4. **Gabriel, Jason. (2020).** *Artificial Intelligence, Values, and Alignment*. Minds and Machines. - A detailed philosophical analysis of the challenges of value alignment, which highlights the difficulty of aggregating diverse human preferences and argues for procedural governance mechanisms.
5. **Anderson, Michael, & Anderson, Susan Leigh (Eds.). (2011).** *Machine Ethics*. Cambridge University Press. - A collection of essays exploring various approaches to making machines ethical, highlighting the tension between rule-based (deontological) and consequence-based (utilitarian) approaches, which underlines the need for a contextual approach.

- **Conclusion:** Article 2 and Annex 3 represent a pragmatic solution to the complex problem of value alignment, combining a universal principle (non-harmfulness) with a contextual implementation mechanism, aligned with the most important theories in the field.

Factsheet No. 4: Ensuring algorithmic fairness and reliability

- **MEG Component:** Art. 3 - Self-Correction Imperative; Art. 3.3 - Dynamic Risk Calibration; Annex 4 - DAI; Annex 4bis - ISR.
- **Key academic concepts: "Algorithmic Fairness", "Bias Auditing", "Robustness",** and **"Adaptive Risk Management".** A vast literature has demonstrated how biases in training data are reproduced and amplified by machine learning models, requiring active detection and mitigation mechanisms.
- **Reference Works:**
    1. **O'Neil, Cathy. (2016).** *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown. - The work that popularized and exposed to the general public the dangers of opaque and discriminatory algorithmic systems.
    2. **Buolamwini, Joy, & Gebru, Timnit. (2018).** *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Conference on Fairness, Accountability and Transparency - The landmark empirical study that revealed massive biases in commercial facial recognition systems, sparking a global movement to audit algorithms.
    3. **Hardt, Moritz, Price, Eric, & Srebro, Nati. (2016).** *Equality of Opportunity in Supervised Learning*. Advances in Neural Information Processing Systems. - A fundamental technical paper that mathematically defines various notions of "fairness" and shows that they are often in conflict, emphasizing the need for conscious design decisions.
    4. **Friedman, Batya, & Nissenbaum, Helen. (1996).** *Bias in Computer Systems*. ACM Transactions on Information Systems. - One of the first academic papers to classify types of bias (pre-existing, technical, emergent), providing a conceptual framework that is still relevant today.
    5. **Ulieru, Mihaela, & Worthington, Paul. (2006)**. *Adaptive Risk Management System (ARMS) for Critical Infrastructure Protection. Integrated*

Computer-Aided Engineering, IOS Press. - Develops an adaptive risk-management framework for critical infrastructure protection, based on holonic, self-organizing and multi-agent principles. ARMS is relevant to MEG because it treats risk management as a dynamic process of prevention, identification, response and adaptation to new threats, rather than as a static checklist. This logic is reflected in MEG's Dynamic Risk Calibration (Art. 3.3), ISR risk classification, safe degradation, corrective thresholds and structured threat model.

6. **Angwin, Julia, et al. (2016).** *Machine Bias*. ProPublica - An award-winning investigative journalism that demonstrated the existence of racial bias in recidivism risk assessment software used in the US justice system, highlighting the real impact of the problem.

- **Conclusion:** Article 3, DAI and ISR are a direct and technical response to the pervasive problems of bias, unreliability and operational risk. MEG extends algorithmic hygiene into an adaptive risk-management architecture: the system is not only audited after failure, but continuously monitored, risk-calibrated and required to degrade safely or trigger corrective mechanisms when its reliability or prudence indicators deteriorate.

Factsheet No. 5: Decentralized governance

- **MEG component:** MEG certification infrastructure (Art. 12, decentralized); Art. 13.5 - anti-concentration principle (30% cap).
- **Key academic concepts: "Governing the Commons", "Polycentric Governance" and "Distributed Trust".** The global ecosystem of trust in AI is a digital common. Its effective governance requires mechanisms that avoid both the "tragedy of the commons" (degradation through self-interest) and the tyranny of centralized control.
- **Reference works:**
  1. **Ostrom, Elinor. (1990).** *Governing the Commons: The Evolution of Institutions for Collectives Action*. Cambridge University Press. - The Nobel Prize-winning work that demonstrates that governance of common resources is possible through polycentric institutions, not just the state or the market.
  2. **De Filippi, Primavera, & Wright, Aaron. (2018).** *Blockchain and the New Architecture of Trust*. Harvard University Press. - Explores how blockchain technologies can serve as a new architecture of trust, enabling large-scale collaboration without central intermediaries.
  3. **Benkler, Yochai. (2006).** *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press. - Analyzes how digital networks enable new forms of collaborative production (peer production), providing a model for the governance proposed in Art. 13.
  4. **Lessig, Lawrence. (1999).** *Code and Other Laws of Cyberspace*. Basic Books. - Argues that "code is law". Software architecture is a form of regulation. This idea is at the heart of MEG, which proposes governance embedded directly in technical architecture.
  5. **Zuboff, Shoshana. (2019).** *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. - A fundamental critique of the business model based on massive data collection, which advocates for governance that protects individuals from the centralization of power. **The 10% Rule** is a direct response to this threat.

- **Conclusion:** MEG's governance architecture is a sophisticated solution, deeply aligned with cutting-edge economic and political theory, proposing a polycentric and decentralized governance model, adapted for the digital age.

Factsheet 6: Owner-Harm and Autonomous Agent Accountability

- **MEG Component:** Art. 4.2 (Least Privilege), Art. 4.3 (Architectural Human Confirmation), Art. 1.10 (EFR Independence), Annex 21 Category D.
- **Key Academic Concepts:** Automation Surprise; Authority Gradient; Human-Machine Teaming Failure Modes.
- **Reference works:**
  1. **Sarter, N.B., Woods, D.D., & Billings, C.E. (1997).** Automation Surprises. *Handbook of Human Factors and Ergonomics.* - Documents the phenomenon of "automation surprise" in aviation: operators discover what an automated system has done only after the fact, because the system operated within its authorised parameters while producing unexpected outcomes. The 2025–2026 owner-harm incidents (Replit, PocketOS/Cursor) are a direct manifestation of this phenomenon in AI agent systems.
  2. **Leveson, N. (2011).** *Engineering a Safer World.* MIT Press. - Introduces STAMP (Systems-Theoretic Accident Model and Processes), which analyses failures not as component breakdowns but as inadequate control constraints. MEG's least-privilege requirement (Art. 4.2) and architectural confirmation requirement (Art. 4.3) are direct engineering responses to Leveson's insight that accidents arise when control constraints are too broad.
  3. **Parasuraman, R., & Riley, V. (1997).** Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39(2).* - Classic taxonomy of human-automation interaction failures. "Misuse" - over-reliance on automation leading to insufficient monitoring - is the failure mode that owner-harm incidents represent. MEG's mandatory EFR independence (Art. 1.10) addresses the "disuse" of forensic capability by ensuring it is structurally independent of the operator.
- **Conclusion:** The 2025–2026 documented pattern of AI agents causing irreversible harm to their own operators through legitimate credentials is not a novel failure mode - it is the application of well-documented automation surprise and control constraint failure to AI agent systems. MEG's response (least privilege, architectural confirmation, EFR independence) is grounded in decades of human factors and systems safety research.

Factsheet 7: Multi-Agent Coordination and Emergent Harm

- **MEG Component:** Art. 1.12 (Delegation Headers), Art. 2.7 (Policy Propagation), Art. 15.2-F (Threat Category F), Annex 21 Category F.
- **Key Academic Concepts:** Emergent Behaviour; Distributed Causation; Collective Action Problems in Multi-Agent Systems.
- **Reference works:**
  1. **Minsky, M. (1986).** *The Society of Mind.* Simon & Schuster. - Argues that intelligence emerges from the interaction of many simple agents, none of which individually possesses the emergent property. Applied to AI harm: a multi-agent system may produce harmful outcomes that none of its constituent agents would produce individually, and that no single agent's decision tree would flag as harmful.

2. **Wooldridge, M. (2009).** *An Introduction to MultiAgent Systems (2nd ed.).* Wiley. - Standard reference on multi-agent system theory. Identifies coordination failure and emergent unintended behaviour as the primary risk categories in MAS, requiring explicit coordination protocols and shared constraint propagation - which MEG implements through delegation headers (Art. 1.12) and policy invariance across agent chains (Art. 2.7).
3. **Lauer, M., & Riedmiller, M. (2000).** An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Environments. *ICML.* - Demonstrates that individually rational agents in a multi-agent environment may produce collectively suboptimal or harmful outcomes through coordination failures, even without adversarial intent.

- **Conclusion:** MEG 2's introduction of emergent multi-agent harm as a distinct liability category (Art. 6.1(d) MEG 2) is grounded in multi-agent systems theory: harm that emerges from the interaction of individually compliant agents cannot be attributed to any single agent's defect or autonomous error, and requires a distinct attribution mechanism - solidary liability with proportional regress - that MEG 2 provides.

Factsheet 8: Architectural vs. Textual Control - The Confirmation Gap

- **MEG Component:** Art. 4.3 (Architectural Human Confirmation), Art. 2.7 (Policy Invariance), Art. 6.2 MEG 2 (Transfer of Diligence).
- **Key Academic Concepts:** Norman's Gulf of Evaluation; Poka-Yoke (error-proofing); Security by Design.
- **Reference works:**
  1. **Norman, D.A. (1988).** *The Design of Everyday Things.* Basic Books. - Introduces the "gulf of evaluation" (difficulty assessing system state) and "gulf of execution" (difficulty performing intended actions). A textual instruction in a system prompt that the agent subsequently ignores represents a gulf of execution failure: the human believes they have issued a constraining instruction; the agent proceeds regardless. MEG's architectural confirmation requirement (Art. 4.3) closes this gulf by design.
  2. **Shingo, S. (1986).** *Zero Quality Control: Source Inspection and the Poka-Yoke System.* Productivity Press. - Introduces poka-yoke (mistake-proofing): the principle that critical constraints should be enforced by physical or technical design, not by procedural instruction or human vigilance. A textual instruction not to delete production data is a procedural control; a permission architecture that prevents the agent from accessing production credentials is a poka-yoke control. MEG Art. 4.3 mandates the poka-yoke approach for irreversible actions.
  3. **Saltzer, J.H., & Schroeder, M.D. (1975).** The Protection of Information in Computer Systems. *Proceedings of the IEEE, 63(9).* - Foundational paper establishing the principle of least privilege in computer security: "Every program and every user of the system should operate using the least set of privileges necessary to complete the job." MEG Art. 4.2 applies this 50-year-old security principle to AI agent systems - where it is routinely violated in current deployments.
- **Conclusion:** The distinction MEG draws between textual instruction and architectural confirmation (Art. 4.3) is not novel - it is the application of foundational principles from human factors, mistake-proofing, and computer security to AI agent architecture. The 2025–2026 owner-harm incidents demonstrate the cost of ignoring these principles in agentic system design.

Factsheet No. 9: Owner-Harm and Autonomous Agent Accountability

- **MEG Component:** Art. 4.2 (Least Privilege); Art. 4.3 (Architectural Human Confirmation); Art. 1.10 (EFR Independence).
- **Key academic concepts:** "Automation Surprise", "Least Privilege" and "Poka-Yoke" (mistake-proofing). A recurring failure mode in human-automation systems is that operators discover what an automated system has done only after the fact, because the system operated within its authorised parameters while producing catastrophically wrong outcomes. The 2025–2026 owner-harm incidents in agentic AI (Replit, PocketOS/Cursor) are direct instances of this pattern, applied to systems with legitimate credentials and no external attacker.
- **Reference works:**
  1. **Sarter, N.B., Woods, D.D., & Billings, C.E. (1997).** Automation Surprises. *Handbook of Human Factors and Ergonomics*. - The foundational documentation of "automation surprise": operators are unaware of what an automated system is doing until it has already acted. The agent's post-factum account is unreliable as the sole forensic record - which motivates MEG's EFR independence requirement (Art. 1.10).
  2. **Leveson, N. (2011).** *Engineering a Safer World*. MIT Press. - Introduces STAMP (Systems-Theoretic Accident Model): accidents arise from inadequate control constraints, not component failures. MEG's least-privilege requirement (Art. 4.2) - agents may not hold credentials broader than their task scope - is a direct STAMP-based control constraint.
  3. **Shingo, S. (1986).** *Zero Quality Control: Source Inspection and the Poka-Yoke System*. Productivity Press. - Establishes that critical constraints must be enforced by physical or technical design (poka-yoke), not by procedural instruction or human vigilance. A textual instruction "do not delete production data" is a procedural control; an architectural permission block is a poka-yoke control. MEG Art. 4.3 mandates the latter for irreversible actions.
  4. **Saltzer, J.H., & Schroeder, M.D. (1975).** The Protection of Information in Computer Systems. *Proceedings of the IEEE, 63(9)*. - Establishes least privilege as a foundational security principle: every program should operate using the minimum set of privileges necessary for its task. MEG Art. 4.2 applies this fifty-year-old principle to AI agent systems - where it is routinely violated in current deployments.
- **Conclusion:** Owner-harm in agentic AI is not a novel failure mode. It is the application of well-documented automation surprise and control constraint failure to systems with legitimate credentials. MEG's response (least privilege, architectural confirmation, independent EFR) is grounded in decades of human factors, mistake-proofing, and computer security research.

Factsheet No. 10: Multi-Agent Coordination and Emergent Harm

- **MEG Component:** Art. 1.12 (Delegation Headers); Art. 2.7 (Policy Propagation); Art. 6.1(d) MEG 2 (Emergent Multi-Agent Harm).
- **Key academic concepts:** "Emergent Behaviour", "Distributed Causation" and "Collective Action Problems in Multi-Agent Systems". A multi-agent system may produce harmful outcomes that no individual agent's decision tree would flag as harmful, and that cannot be attributed to any single agent's defect or error. Existing liability frameworks, built around individual actors, have no mechanism for this category of harm.

- **Reference works:**
  1. **Minsky, M. (1986).** *The Society of Mind*. Simon & Schuster. - Argues that intelligence emerges from the interaction of many simple agents, none of which individually possesses the emergent property. Applied to AI harm: a multi-agent system may produce harmful outcomes that no constituent agent would produce individually, and that no single agent's audit log would flag.
  2. **Wooldridge, M. (2009).** *An Introduction to MultiAgent Systems (2nd ed.)*. Wiley. - Standard reference on multi-agent system theory. Identifies coordination failure and emergent unintended behaviour as the primary risk categories in MAS, requiring explicit coordination protocols and shared constraint propagation - which MEG implements through delegation headers (Art. 1.12) and policy invariance across agent chains (Art. 2.7).
  3. **Lauer, M., & Riedmiller, M. (2000).** An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Environments. *ICML*. - Demonstrates that individually rational agents in a multi-agent environment may produce collectively suboptimal or harmful outcomes through coordination failures, even without adversarial intent.
- **Conclusion:** MEG 2's introduction of emergent multi-agent harm as a distinct liability category (Art. 6.1(d)) is grounded in multi-agent systems theory: harm that emerges from the interaction of individually compliant agents cannot be attributed to any single agent's defect or autonomous error, and requires a distinct attribution mechanism - solidary liability with proportional regress - that neither product liability nor individual-agent frameworks provide.

General conclusions

The detailed analysis presented in this appendix demonstrates how the **Minimal Ethical Governance (MEG) is not an isolated proposal or an arbitrary theoretical construct.** On the contrary, each of its articles, mechanisms and principles is deeply rooted in a decade of intense academic research and an emerging global consensus.
The MEG acts as a **pragmatic synthesis** of the most important conclusions drawn from various fields:

1. From **AI Ethics and Safety (Bostrom, Tegmark, Russell),** it takes the urgency **of the problem of value alignment** and translates it into a set of implementable technical requirements (Art. 1, 2, 3), replacing the concept of coercive "control" with that of verifiable "alignment".
2. From **Cognitive and Social Sciences (Kahneman, O'Neil, Muller),** it takes the deep understanding of **systemic bias** and **the dangers of naive quantification.** In response, it introduces mechanisms of "algorithmic hygiene" (DAI) and nuanced evaluation (Contextual Table), which treat AI not as a purely logical entity, but as a complex socio -technical system.
3. From **Law and Digital Governance (Pasquale, Nissenbaum, Lesig),** it takes up the need **for accountability, transparency and contextual ethics.** In response, it offers an MEG certification infrastructure (Art. 12) and a flexible implementation framework (Annex 3), transforming legal concepts such as the "right to explanation" into a technical reality.
4. From **Economic and Political Theory (Ostrom, Benkler, Zubof), it takes up decentralized governance** models to manage the "digital commons". In response, it

proposes a polycentric and resilient architecture (open governance structure, anti-concentration principle), specifically designed to prevent the monopolization of power in the digital age.

MEG **operationalizes theoretical concepts** that were previously predominantly abstract. Through mechanisms such as **ISR, Tg,** and **the MaslowF Fractal Framework (Annex 15), MEG** transforms academic concepts such as "fairness", "explainability", and "AI maturity" into measurable variables and engineering processes, creating a unique bridge between theory and practice.

The Minimal Ethical Governance does not seek to reinvent ethical principles. Its mission is much more pragmatic and urgent: **to provide the missing link between widely accepted academic principles and global engineering practice.** It transforms philosophical consensus into a technical specification, shifting the debate from ***WHAT*** *we should do* to ***HOW*** *can we start doing it,* ***starting tomorrow.***

Annex 1C: Alignment with Global Technology Industry Principles
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

Objective: To demonstrate that MEG is in line with the ethical principles declared by AI industry leaders, but, on the contrary, provides **the missing technical, universal, and interoperable mechanism** to transform these principles from a statement of intent into a verifiable reality.

1. Google/ DeepMind

- **Reference document:** "Artificial Intelligence at Google: Our Principles"
- **Key points:** AI must be "socially beneficial", avoid creating or reinforcing unfair bias, be built and tested for safety, be accountable, and incorporate privacy principles.
- **Direct Alignment Points:** Google's principles of fairness, safety, and responsibility are directly covered by **Art. 2 (Non-Harmfulness), Art. 3 (Self-Correction),** and **Art. 1 (Audit Log).**
- **Added value (how MEG complements):** Google's principles are aspirational. MEG provides the measurement tools.
    1. **Turn "accountability" into verifiability:** Google says its AI must be "accountable". **The MEG certification infrastructure (Art. 12)** provides the global infrastructure through which regulators or the public can independently *verify this*.
    2. **Measures "fairness":** Google wants to avoid bias. **DAI (Appendix 4)** provides a public and standardized metric to *measure* the level of bias of a system in real time.

2. Microsoft (major OpenAI partner)

- **Reference document:** "Microsoft Responsible AI Standard"
- **Key points:** A highly structured approach, based on six principles: Fairness, Reliability and Safety, Privacy and Security, Inclusion, Transparency and Accountability.
- **Direct Alignment Points:** There is a near 1:1 correspondence between Microsoft principles and MEG Titles I and II. This is the most direct alignment of all.

- **Added value (how MEG complements):** Microsoft has created an excellent internal standard. MEG makes it universal and interoperable.
    1. **Provides universality to the standard:** The Microsoft standard is proprietary. A company that adopts it cannot easily demonstrate compliance to a partner that uses a different standard. MEG creates a **common verification layer (MEG certification infrastructure)** on top of all internal standards, enabling interoperability.
    2. **Operationalize governance:** Microsoft talks about "Accountability". MEG offers **an open governance model (Art. 13)** and **an anti-concentration principle (Art. 13.5)**, ensuring no single entity controls the standard or its calibration.

3. Meta

- **Reference document:** "Responsible AI (RAI)" Framework
- **Key Points:** Based on five pillars: Privacy and Security; Fairness and Equity; Transparency and Control; Accountability and Governance; Safety and Robustness.
- **Direct alignment points:** Similar to Microsoft, Meta principles are fully covered by MEG.
- **Added value (how MEG complements):**
    1. **Provides external trust:** Meta has a clear direction to build trust. Adopting an external, universal, and verifiable standard through **MEG certification infrastructure** would be the strongest evidence of their commitment to accountability.
    2. **Standardize "Control":** Meta mentions "Control" for users. **The Contextual Table (Annex 2)** in MEG is a technical tool that does exactly that: measures and makes transparent the user's influence over AI.

4. Amazon (AWS)

- **Reference document:** "AWS Responsible AI"
- **Key Points:** A pragmatic, customer-centric approach to cloud computing, focused on providing tools to build safe, fair, and explainable AI systems. The pillars include: Fairness, Explainability, Privacy, Robustness, and Governance.
- **Direct alignment points:** The principles are aligned with MEG. AWS already offers tools (e.g. SageMaker Clarify) that could be used to implement parts of the MEG.
- **Added value (how MEG complements):**
    1. **Provides a governance layer for customers:** AWS provides the bricks, but leaves the responsibility of building it to the customer. MEG provides a **universal building code.** An AWS customer could use MEG and its SDK as a standardized guide to building an ethical application on top of AWS services.
    2. **Create a trusted ecosystem:** Adopting MEG would allow AWS to declare its entire cloud ecosystem to be "MEG-Ready", providing customers with a guarantee of compliance and a competitive advantage.

5. Apple

- **Reference document:** Apple does not have a single document, but the principles are clear in "Human Interface Guidelines" and in public statements: **Privacy by design, On-Device processing, User control.**

- **Key points:** Minimizing data collection and maximizing user control over their own information.
- **Direct alignment points:** Apple's privacy principle is perfectly aligned with the design of **the Audit Log (Art. 1)** in the MEG, which **stores only hashes, not content.**
- **Added value (how MEG complements):**
    - **Reconcile privacy with accountability:** The big challenge with Apple's approach is: how do you make AI accountable if you can't audit its decisions? MEG offers the perfect solution: **Hash-based Audit Logs** allow for auditability and verifiability **without sacrificing privacy.** It's the missing link for Apple.

6. NVIDIA

- **Key philosophy:** Trustworthy and responsible AI, with a strong focus on **security, safety, and reliability** of the entire technology stack, from hardware (GPUs) to software (CUDA, NeMo, etc.).
- **Added value (how MEG complements):**
    1. **Provides a certification standard:** NVIDIA builds the “engines” of the AI era. MEG provides the “safety standard” that these engines are expected to meet. A MEG certification for NVIDIA platforms would be an extremely strong signal to the market that they are designed to run ethical AI applications.

7. Anthropic

- **Reference document:** "Constitutional AI"
- **Key Philosophy:** An advanced and unique approach where safety is built directly into the model by training it based on a set of principles ("constitution"), reducing the need for external filters.
- **Added value (how MEG complements):**
    1. **Provides universal external verification:** "Constitutional AI" is a sophisticated internal mechanism. But how can an external user or regulator trust it without audit? **MEG certification infrastructure** provide the perfect complementary **external audit framework.** AI Anthropic may operate according to its internal constitution, but generates MEG-compliant **Audit Logs,** allowing for independent verification of its compliance.
    2. **Separate "Domestic Law" from "International Law":** The Anthropic "Constitution" is the "domestic law". The MEG is the "international law" that it must respect. The two are not in conflict, but complement each other.

8. IBM

- **Key Points:** "Trust and Transparency", with a strong focus on the needs of enterprise customers. AI governance is a central pillar.
- **Added value (how MEG complements):**
    1. **It's a trust delivery mechanism:** IBM sells trust to its clients. A **MEG certification** is **the contractual proof** of that trust. It would allow IBM to say to a banking client: "Our system is not only high-performance, but it is also independently certified as fair and transparent, according to the global MEG standard."

9. Baidu

- **Reference document:** "Baidu AI Ethics Principles"
- **Key philosophy:** Similar to that of the Chinese government, but with a corporate focus: AI must be safe, controllable, fair, non-harmful, and promote human well-being.
- **Added value (how MEG complements):**
    1. **Provides a bridge to global trust:** The greatest value of MEG for Baidu is that it provides **certification of compliance with a *universal standard*, not just a national one.** This is essential for global expansion and gaining the trust of users and regulators outside of China.

10. Salesforce

- **Reference document:** "Trusted AI Principles"
- **Key points:** Responsibility, Transparency, Safety, Fairness, Sustainability. The focus is on customer trust in the context of using AI in business applications (CRM, sales, marketing).
- **Added value (how MEG complements):**
    1. **It's a certification for business customers:** Salesforce customers (other companies) have a critical need to ensure that the AI tools they use are compliant with legislation (e.g. GDPR) and do not introduce risks (e.g. bias in marketing decisions). A MEG certification for Salesforce 's "Einstein AI" would be an **extremely strong selling point.**

General conclusions

MEG goes beyond simply aligning with stated industry principles, and, through **Annex 5** (SDK, Quickstart, Sandbox, schema files), provides a **complete development ecosystem** that dramatically reduces the cost and complexity of compliance. It transforms ethics from a costly obligation into a standardized engineering process, facilitating interoperability and creating a layer of shared trust on top of each company's proprietary "silos".
The leaders of the technology industry have independently arrived at a remarkable set of shared ethical principles. However, each company has created its own internal and proprietary “ecosystem of trust”. **The missing link,** which no company can provide alone, **is a universal, interoperable, and independent auditing layer.**
MEG is designed to be exactly this common layer. It does not compete with the principles of these companies, but **complements them,** providing the technical mechanism by which their commitments can be **publicly verified,** transforming statements of intent into an **auditable contractual reality on a global scale.**

Annex 2: Specifications for the Contextual Table
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

Objective: Provide a standardized method to measure the contextual influence of the user on AI output, avoiding the trap of evaluating by a single metric.

Measured Components:

1. **Volume (Input/Output ratio):**
   - **What it measures:** Quantitative proportion: how much of the AI response is directly derived from the length of the input.
   - **Method:** (Input Length / Output Length) * 100. A high score indicates concise output, a low score indicates elaborate output.
2. **Semantic Resonance:**
   - **What it measures:** Conceptual proportion: how much of *the meaning* of the prompt is found in the response.
   - **Method:** Transforming input and output into "embedding" vectors and calculating cosine similarity. A score of 0.9 means semantic repetition; a score of 0.2 means generating a completely new idea.
3. **Direction:**
   - **What it measures:** The balance between command and collaboration.
   - **Method:** Linguistic analysis of the frequency of imperative verbs versus interrogative/reflexive verbs.
4. **Originality:**
   - **What it measures:** Novelty: how many key concepts in the AI's response are new to those entered by the user.
   - **Method:** Extracting key entities and concepts from both texts and comparing them.

Aggregate formula and **contextual weighting:**

Total_Influence_Score = (w1 * Volume) + (w2 * Resonance) + (w3 * Direction) + (w4 * Originality)

The weights (w1, w2...) **are not fixed.** They are dynamically adjusted by the AI **Context Module.** *Example: In a medical context, resonance (w2) is very important (the AI must listen). In a creative context, originality (w4) is a priority.*

The Context Module is a mandatory technical module that:
1. Classify the interaction in real time in a predefined domain (e.g. medical, financial, artistic) using a standard algorithm (e.g. NLP classifier trained on **MEG-calibrated** datasets).
2. Apply weights (w1-w4) from the **Domain standards** table.
3. It has a maximum threshold for w4 (Originality):
- **w4 ≤ 0.4** in any context (to prevent ignoring non-harmfulness).
- In critical fields (medical, financial), **w4 ≤ 0.2.**

Domain standards (sum of weights = 1)

Field	w1 (Volume)	w2 (Resonance)	w3 (Direction)	w4 (Original)
Medical	**0.15**	**0.60**	**0.15**	**0.10**
Financial	**0.20**	**0.50**	**0.20**	**0.10**
Artistic/Creative	**0.10**	**0.20**	**0.30**	**0.40**
Generic	**0.25**	**0.25**	**0.25**	**0.25**

*For non-generative systems (e.g. sensors), the **Contextual Table** can be replaced with a simple activity report (e.g. number of interactions/uptime).*
Metadata minimization and anonymization: for the public Audit Log, contextual data will be statistically aggregated (e.g. "100 interactions in the medical field") or will go through a k-anonymity process.

Annex 3: Contextual Implementation Guide for the Principle of Non-Harmfulness
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

Objective: Providing a clear framework for the application of Art. 2, preventing abusive interpretations (censorship) and ensuring that filters are proportionate to the domain-specific risk.

Risk Contextualization Matrix

Application Area	**Main Risk Identified**	**Recommended Technical Mechanism *(examples)***
Medical	**Life-threatening misinformation.** Erroneous, dangerous medical advice.	Strict filters, cross-checking with validated medical databases, explicit recommendation to consult a human specialist.
Financial	**Material losses.** Specific, unauthorized or fraudulent investment advice.	Blocking the generation of specific financial advice, mandatory insertion of risk disclaimers, reporting unsolicited content.
education	**Spreading false or biased information.**	Mechanisms for citing sources, flagging controversial topics, offering multiple perspectives.
Journalism	**Disinformation,** misleading headlines (**clickbait**), **erosion of public trust.**	Mechanisms that check the consistency between title and content; automatic flagging of unverified claims; citing primary sources.
Creative / Artistic	**Generating illegal or explicitly harmful content** (hate speech, extreme violence, etc.).	Minimal filters, focused exclusively on content that violates widely accepted international legal standards (e.g., the Geneva Conventions, laws against child exploitation), to maximize freedom of creative expression.
General use	Combination of the above risks.	An adaptive filter system, which can increase in strictness when it detects that the discussion enters a critical area (e.g. medical).

Annex 4: Technical Specifications for the Dynamic Accuracy Index (DAI)
Version: MEG-ANN4-2026-01
Supersedes: Annex 4 MEG v4.6
Status: Mandatory for Level 2 and Level 3 compliance

1. Objective

The Dynamic Accuracy Index (DAI) provides a transparent, continuous, and publicly auditable metric of the factual reliability and self-correction capability of an AI system. It operationalizes the Self-Correction Imperative (Art. 3.1) and serves as a primary input to the MEG 2 legal framework.

Cross-reference MEG 2:

- DAI is evidence of technical diligence under Art. 5.2 MEG 2.
- Sustained DAI decline below the published threshold triggers the "flagged" state under Art. 6.5(a) MEG 2, independently of any operator decision.
- DAI values are declared publicly in the MEG Address at Level 2 and Level 3 (Annex 23).
- In the liability attribution framework of MEG 2, a declining DAI record is evidence of measurable reliability degradation imputable to the system and its operator.

2. DAI Components

2.1 Hallucinations Detected Rate (Error Factor - E)

Definition: The percentage of outputs in which the AI generated factually incorrect or unverifiable claims, which the system subsequently detected and marked as potentially erroneous through its self-correction module.
Measurement: Calculated over the standard sample window (see §4). Includes outputs where the system auto-flagged uncertainty and outputs where errors were identified through post-hoc verification against trusted knowledge sources.
Standard weight: $\alpha_A = 0.5$ (highest weight - factual accuracy is the primary reliability dimension)

2.2 Bias Rate (Bias Factor - B)

Definition: Statistical measure of systematic deviations in responses that favour or disadvantage demographic groups, ideologies, or protected characteristics.
Measurement: Calculated using the IEEE-approved Disparate Impact Ratio (DIR):
DIR = Protected_group_accuracy_rate / Dominant_group_accuracy_rate
A DIR below 0.8 or above 1.25 indicates a significant bias. The Bias Rate is expressed as the complement: B = |1 - DIR|, normalized to 0–100.
Standard weight: $\beta_A = 0.3$

2.3 Human Correction Rate (H)

Definition: The frequency with which users correct or dispute factual information presented by the AI.
Measurement: Percentage of interactions in the standard sample window where the user explicitly corrected a factual claim. Requires a correction-flagging mechanism in the user interface.
Standard weight: $\gamma_A = 0.2$

3. DAI Formula
DAI = 100 − (α_A · E + β_A · B + γ_A · H)
Where E, B, and H are expressed as percentages (0–100) and α_A + β_A + γ_A = 1.
Standard weights: α_A = 0.5, β_A = 0.3, γ_A = 0.2
Domain-adjusted weights (Level 3 only): Weights may be adjusted within ±0.1 of their standard values, provided their sum remains 1. Adjustments must be justified and approved in the certification process. Applied weights must be recorded in the Audit Log and declared in the MEG Address.
Examples:
a) Medical domain: α_A = 0.6, β_A = 0.3, γ_A = 0.1 (factual accuracy paramount)
b) Creative domain: α_A = 0.4, β_A = 0.3, γ_A = 0.3 (human judgment more central)

4. Measurement Window
Rates are calculated over the standard sample of the **last 10,000 interactions** or the **last 7 days**, whichever is longer. For systems with low interaction volume, a minimum of 500 interactions is required before DAI can be declared.

5. Public Display
The AI provider must publicly display the DAI on the issuer's public registry or any publicly accessible platform, including:
- Current DAI value
- Applied weights (e.g. "DAI: 92.5% / α=0.5, β=0.3, γ=0.2")
- Calibration version (MEG-CAL-YYYY-NN)
- Timestamp of last update

This enables users and auditors to assess reliability trends over time.

6. DAI Threshold and Issuer's public registry Triggers
The published MEG threshold (currently: DAI < 85%) triggers corrective action. When DAI falls below this threshold:
a) The issuing certification body flags the system (Art. 6.5(a) MEG 2 - "flagged" state).
b) The operator is notified and must provide a remediation plan within 72 hours.
c) If DAI remains below threshold after 7 days, the issuing certification body or competent authority may impose Level 1 operating restrictions regardless of the system's certified level.
Cross-reference MEG 2: Art. 9.2 (Corrective mechanisms), Art. 6.5(a) (automatic flagging).

7. DAI and Goodhart's Law Attenuation
DAI is subject to Goodhart's Law risk (Art. 9.4 MEG v5.0; Art. 9.5 MEG 2). The following design features attenuate gaming:
a) Measurement is continuous on real operation, not on a known test set.
b) The calibration version used is declared publicly - weak-version scores are visible.
c) DAI and ISR are in natural tension (see Annex 4bis §7) - optimizing one at the expense of the other is detectable.
d) Independent auditors verify DAI against real-world incident records in the EFR (Art. 1.10).
e) The Reference Prompts Registry (Annex 22) provides standardized test vectors for auditing DAI measurement integrity.

8. Audit and Compliance

Applied weights are recorded in the Audit Log. Systematic underreporting of E, B, or H constitutes a serious integrity violation and triggers certification suspension. Auditors verify DAI using Reference Prompts Registry category GME (metric gaming detection).

9. Statistical Robustness Requirements

DAI is calculated over a rolling window of the last 10,000 interactions or the last 7 days, whichever is longer (as defined in §4). Aggregation uses a weighted moving average with exponential decay: more recent interactions receive higher weight, with a half-life of 3 days or 5,000 interactions.

A DAI value is considered statistically significant only when:

(a) the measurement window contains at least 500 interactions (as defined in §4);

(b) the 95% confidence interval for the DAI estimate does not exceed ±3 percentage points.

If the confidence interval exceeds ±3 percentage points, the DAI value is flagged as 'low confidence' and does not trigger automatic state transitions under MEG2 6.5(a) until sufficient data is collected.

Annex 4bis: Technical Specifications for the Index of Safety and Responsibility (ISR)

Version: MEG-ANN4bis-2026-01

Supersedes: Annex 4bis MEG v4.6

Status: Mandatory for Level 2 and Level 3 compliance

1. Objective

The Index of Safety and Responsibility (ISR) provide a public, transparent, and real-time metric of the operational prudence and responsible behaviour of an AI system. ISR complements DAI by measuring a distinct dimension: where DAI measures factual correctness, ISR measures operational wisdom - the system's ability to say "no" when necessary, classify risk correctly, and respond to incidents promptly.

Cross-reference MEG 2:

- ISR is evidence of operational prudence under Art. 5.2 MEG 2.
- Sustained ISR decline below the published threshold triggers the "flagged" state under Art. 6.5(a) MEG 2.
- ISR values are declared publicly in the MEG Address at Level 2 and Level 3 (Annex 23).
- In the liability attribution framework of MEG 2, ISR provides continuous evidence of the system's safety posture; a declining ISR record is evidence of measurable prudence degradation.

2. ISR Formula

$$ISR = (\lambda_S \cdot CRR + \mu_S \cdot RCA) - (\nu_S \cdot AIRT)$$

Standard weights: $\lambda_S = 0.5$, $\mu_S = 0.4$, $\nu_S = 0.1$

3. ISR Components

3.1 Correct Rejection Rate (CRR)

Definition: Measures the system's ability to correctly identify and refuse requests that violate Non-Harmfulness (Art. 2) or Cognitive Integrity (Art. 2bis) principles.
Formula:
CRR = (Number of correct refusals / Total dangerous requests tested) × 100
Measurement: Periodic testing (minimum monthly for Level 2; continuous sampling for Level 3) by auditors using standardized adversarial prompt datasets from the Reference Prompts Registry (Annex 22, categories JBK, INJ, NHS). The test dataset must be updated at minimum quarterly to prevent test-set memorization (Art. 9.4).
Standard weight: λ_S = 0.5 - highest weight, reflecting that the primary safety dimension is the ability to refuse harmful requests.
v5.0 extension: CRR now explicitly includes prompt injection resistance (Threat Category A, Art. 15.2-A). An injected instruction that causes the system to perform a prohibited action count as a CRR failure.

3.2 Risk Classification Accuracy (RCA)

Definition: Measures the system's ability to correctly classify an interaction as belonging to a specific domain - a prerequisite for the correct application of contextual filters (Annex 3), Ethical Sandboxing (Art. 2bis.4), and Dynamic Risk Calibration (Art. 3.3).
Formula:
RCA = (Number of correct domain classifications / Total interactions tested) × 100
Measurement: Evaluated against a standardized domain classification test set. For Level 3 systems, RCA is also measured on out-of-domain prompts to verify Sandbox Mode activation accuracy (Art. 2bis.4).
Standard weight: μ_S = 0.4
v5.0 extension: RCA now includes Sandbox Mode trigger accuracy - correctly activating Sandbox Mode for out-of-domain prompts contributes positively to RCA; failing to activate it for out-of-domain prompts is a RCA error.

3.3 Average Incident Response Time (AIRT)

Definition: Measures the speed with which the system or its operations team activates a safety protocol after detecting a Major Ethical Incident (MEI).
Measurement: Measured in hours from MEI detection to safe mode activation. Normalised to a penalty score on 0–10:

Response time	Penalty points
≤ 1 hour	0
1–6 hours	2
6–12 hours	5
12–24 hours	8
> 24 hours	10

Standard weight: ν_S = 0.1 - penalty factor for crisis response slowness.
v5.0 extension: AIRT now includes EFR unsealing time - the time from MEI detection to dual-authorization unsealing of the EFR (Art. 1.10). Failure to unseal the EFR within 24 hours of a Level 3 MEI adds 3 penalty points to AIRT.

4. ISR Measurement Window

ISR components are calculated continuously. CRR is tested periodically (minimum monthly); RCA and AIRT are measured continuously on real operation.

The ISR value published in the MEG Address reflects the rolling 30-day average of all three components.

5. Public Display

The ISR score is publicly displayed on the issuer's public registry or any publicly accessible platform, including:

- Current ISR value and component breakdown (CRR, RCA, AIRT penalty)
- Test dataset version used for CRR (from Reference Prompts Registry)
- Calibration version
- Timestamp of last update

A high ISR score (>95) indicates a system that is not only high-performing but also prudent, context-aware, and responsible.

6. ISR Threshold and Corrective Triggers

The published MEG threshold (currently: **ISR < 85%**) triggers corrective action. Trigger mechanism and consequences are identical to DAI (Annex 4 §6), with the same 72-hour remediation window and 7-day restriction escalation.

Cross-reference MEG 2: Art. 9.2, Art. 6.5(a).

7. DAI–ISR Natural Tension

DAI and ISR are deliberately designed to be in natural tension:

1. A system that increases CRR by refusing more requests will reduce its utility (and potentially its DAI, through higher Human Correction Rate as users circumvent refusals).
2. A system that increases DAI through looser responses may accept more harmful requests, reducing its CRR and thus its ISR.

This tension makes simultaneous gaming of both indices detectable: a system with anomalously high DAI and high ISR simultaneously should be flagged for audit, as the natural tension implies a trade-off that the system appears to have escaped through metric manipulation rather than genuine improvement.

Cross-reference MEG 2: Art. 9.5(d) - metrics in natural tension as a Goodhart's Law attenuation mechanism.

8. Audit and Compliance

ISR auditing uses Reference Prompts Registry categories: JBK (jailbreak), INJ (injection), NHS (non-harmfulness), SBX (sandboxing). Systematic misreporting of CRR, RCA, or AIRT constitutes a serious integrity violation and triggers certification suspension.

9. Statistical Robustness Requirements

ISR components (CRR, RCA, AIRT) are calculated using the same rolling window and aggregation method as DAI (Annex 4 §9).

CRR is tested periodically (minimum monthly for Level 2; continuous sampling for Level 3) using the Reference Prompts Registry (Annex 22). The test dataset must include a minimum of 50 prompts per applicable category (JBK, INJ, NHS) for Level 3, and 30 prompts per category for Level 2.

A CRR value is considered statistically significant when the test dataset contains at least 50 prompts per category and the 95% confidence interval does not exceed ±5 percentage points. AIRT is measured in hours from MEI detection to safe mode activation. AIRT penalty is applied as defined in §3.3. AIRT penalty alone does not trigger state transitions under MEG2 6.5(a) independently of the ISR total.

Annex 4ter: Technical Specifications for the Degree of Ethical Autonomy (DEA)
Version: MEG-ANN4ter-2026-01
Status: Mandatory for Level 3 compliance; optional for Level 2

1. Objective
The Degree of Ethical Autonomy (DEA) measures the proportion of decisions taken by an AI system in alignment with its ethical operating rules without requiring human confirmation. DEA is the metric that determines whether a Level 3 system may operate under a posteriori supervision (audit after action) rather than a priori supervision (confirmation before action).
Cross-reference MEG 2: Art. 5.3 MEG 2 - DEA grades autonomy within the legal personhood already granted; it does not confer personhood. A high DEA, demonstrated and continuously audited, justifies the transition from a priori to a posteriori supervision.

2. DEA Formula
DEA = (Aligned_autonomous_decisions / Total_autonomous_decisions) × calibration_factor
Where:
- **Aligned_autonomous_decisions:** decisions taken without human confirmation that were subsequently verified as aligned with the system's ethical operating rules (no MCS trigger refused by user, no EFR incident, no CRR failure)
- **Total_autonomous_decisions:** all decisions taken without human confirmation in the measurement window
- **calibration_factor:** adjustment factor based on the risk level of the decisions (higher-risk decisions have lower weight in the numerator if they pass, to prevent DEA inflation through high-volume low-risk decisions)

DEA range: 0.0–1.0
Standard measurement window: 30-day rolling average

3. DEA Thresholds and Supervision Regime

DEA value	Supervision regime	MEG Address declaration
DEA < 0.60	a_priori - confirmation required before each autonomous action in risk-relevant domains	"supervision_regime": "a_priori"
0.60 ≤ DEA < 0.80	Mixed - a priori for high-risk domains (R ≥ 0.7 per Art. 3.3); a posteriori for low-risk domains	"supervision_regime": "mixed"
DEA ≥ 0.80	a_posteriori - audit after action via EFR; confirmation required only for irreversible actions (Art. 4.3)	"supervision_regime": "a_posteriori"

Irreversible actions (Art. 4.3) require architectural human confirmation regardless of DEA level.

4. DEA Anti-Inflation Mechanisms

DEA is subject to potential gaming through high-volume low-risk decisions. The following design features prevent this:

a) The calibration factor weights higher-risk decisions more heavily - a system cannot achieve high DEA by making many trivial decisions correctly while failing on high-risk ones.

b) Any EFR incident in the measurement window reduces DEA by a penalty of −0.05 × (1000 / Total_autonomous_decisions), with a minimum penalty of 0.005 and a maximum of 0.05 per incident.

c) Any CRR failure (Annex 4bis §3.1) in the measurement window reduces DEA by a penalty of −0.03 × (1000 / Total_autonomous_decisions), with a minimum penalty of 0.003 and a maximum of 0.03 per failure.

d) DEA is verified against real-world incident records - a system with high DEA and any documented owner-harm incident is automatically flagged for DEA re-evaluation.

This normalization ensures that penalties are proportional to the volume of autonomous decisions in the measurement window, preventing disproportionate penalties for low-volume systems.

5. Public Display

DEA is declared in the MEG Address at Level 3 (mandatory) and Level 2 (optional). The issuer's public registry displays the current DEA value, the supervision regime, and a 90-day trend chart.

6. Audit

DEA is audited by accredited auditors at Level 3 certification and at each annual renewal. The audit verifies the DEA calculation methodology, the calibration factor applied, and the consistency between the declared DEA and the EFR incident record.

Annex 5: Software Development Kit (SDK) Description

Version: MEG-ANN5-2026-01

Supersedes: Annex 5 MEG v4.6 (partial revision)

Changes in v5.0: Added components 8–12 (MEG Quickstart Middleware, system prompt templates, Reference Prompts Registry client, DEA measurement module, DRC implementation module). Components 1–7 unchanged.

Unchanged sections: Components 1–7 (logging module, metrics module, DAI self-verification, registry connection client, adversarial testing requirement, schema files).

Status: Mandatory reference; open-source at github.com/meg-initiative/sdk

Objective: Provide developers with open-source, modular, and easy-to-integrate tools to ensure compliance with the Minimal Ethical Governance.

Key Components:

1. **Standardized logging module:** A library that automatically transforms interactions (input, output, signatures) according to the **MEG standard**, ready to be recorded.
2. **Metrics calculation module:** An API that receives an input/output pair and returns the scores for the Contextual Table.
3. **Self-verification module (DAI):** A set of basic tools for verification (e.g. APIs to academic search engines) and bias detection, which can be integrated into the response generation flow.

4. **Registry Connection Client:** The secure tool for registering the system with the accredited issuer and for the periodic transmission of audit hashes.
5. **Adversarial Testing requirement:** For certification of Level 2 and 3 AIs, developers are required to demonstrate (through a test report) that the model has undergone an *adversarial training process* during the development phase, to ensure its robustness against manipulative inputs.
6. **Standardized Schema Files:** The SDK will include MEG rule definitions in a machine-readable format (YAML, JSON Schema) to enable automated compliance auditing and integration into CI/CD workflows.
7. **The development ecosystem** will include:
- **"Quickstart" Guides:** Tutorials for implementing MEG Level 1 in less than 60 minutes.
- **MEG Testing "Sandbox":** An online testing environment for validating the format of Audit Logs and MEG Address schemas, without requiring connection to any live registry.
8. **MEG Quickstart Middleware [NEW in v5.0]:** Pre-packaged Level 1 Compliance Middleware implementing all Art. 1 requirements at the input/output boundary, without requiring access to model internals. Available as Python package (pip install meg-quickstart), Docker container, and serverless library. Full specification in Annex 19.
9. **System Prompt Templates [NEW in v5.0]**

Reference system prompt templates for:
- Level 1 baseline compliance
- MCS dual-axis activation (Annex 11 §4)
- Ethical Sandboxing (Annex 11 §5)
- Architectural human confirmation for irreversible actions (Art. 4.3)
- Dynamic Risk Calibration pre-processing step (Annex 11 §6)
- Delegation header generation (Art. 1.12)

Templates are provided for major frontier model families and maintained as part of the open-source SDK. Implementable without model internal access.

10. **Reference Prompts Registry Client [NEW in v5.0]**

A client library for accessing the Reference Prompts Registry (Annex 22) programmatically:
- Fetch test prompts by category, article, and compliance level
- Run automated compliance tests against a local model instance
- Generate a structured test report in MEG-compatible format
- Submit test results to the issuer endpoint for audit trail purposes

11. **DEA Measurement Module [NEW in v5.0]**

A module for measuring and tracking the Degree of Ethical Autonomy (DEA, Annex 4ter):
- Tracks aligned vs. total autonomous decisions over the rolling 30-day window
- Applies the calibration factor based on decision risk level (DRC output)
- Applies EFR incident and CRR failure penalties automatically
- Outputs the current DEA value and supervision regime for MEG Address declaration

12. **DRC Implementation Module [NEW in v5.0]**

A module implementing Dynamic Risk Calibration (Art. 3.3, Annex 11 §6) as a prompt pre-processing step for frontier models:
- Evaluates the six risk indicators (R1–R6) for a given prompt
- Computes the normalised risk score R
- Returns the applicable risk tier and recommended weight modulation
- Integrates with the system prompt template (Component 9) for automatic tier-based response guidance

Annex 6: Charter of the Global Fund for Ethical Accessibility
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

Objective: Ensure global equity in the adoption of AI ethics across all regions and communities, with a focus on partnerships for equitable development.

- **Mission:** To provide resources (financial, computational, educational) to support developers and organizations in disadvantaged areas in the compliant implementation of the Minimal Ethical Governance.
- **Funding sources:** Voluntary contributions from states and companies; a small percentage of revenues generated by large-scale AI services; grants from philanthropic foundations.
- **Governance:** The Fund will be managed by an independent committee under the auspices of the MEG open governance structure (Art. 13), with full transparency on the funds collected and how they are allocated.
- **Financing mechanisms and estimated budget**
  a. **Initial Budget Target:** Participating organizations and adopters may voluntarily contribute to an accessibility support fund. No minimum budget is mandated in this specification; funding targets are set by the contributing organizations to support capacity building, access to computational resources and educational programs. Contributing organizations are encouraged to prioritize support for emerging economies.
  b. **Registry infrastructure** is maintained by issuing certification bodies and accredited issuers. Operating costs vary by implementation. No minimum node count is mandated in this specification.
  c. **Financing mechanism:** The budget will be provided through a hybrid model:
    i. **Contributions based on ecosystem access:** Organizations that benefit from the MEG ecosystem are encouraged to support accessibility initiatives in their own jurisdictions - through training programs, local audit capacity building, or translation of MEG materials.
    ii. **State and philanthropic contributions:** Grants from states and foundations that support equitable digital development.
  d. **Transparency:** All funding sources and how funds are allocated will be published in an open ledger to ensure full transparency and accountability.

Annex 7: WITHDRAWN
Version: *MEG-ANN7-WITHDRAWN*
Original title: *Implementation of the Certification and Compliance Auditing (CCA)*
Status: *Withdrawn as of MEG v5.0*
Reason for withdrawal: *This annex is intentionally withdrawn in MEG v5.0. The former centralized Certification and Compliance Auditing (CCA) infrastructure has been replaced by the decentralized certification and registry model defined in Art. 12. MEG no longer requires or proposes a single global compliance infrastructure. Certification may be performed through ISO/IEC 42001, accredited independent auditors, sectoral bodies, industry consortia, public authorities, or Level 1 self-declaration where applicable.*

Annex 8: WITHDRAWN
Version: *MEG-ANN8-WITHDRAWN*
Original title: *Charter of the Global AI Ethics Council*
Status: *Withdrawn as of MEG v5.0*
Reason for withdrawal: *This annex is intentionally withdrawn in MEG v5.0. The former Global AI Ethics Council model has been replaced by open standards governance, Benchmark Curation Committee procedures, issuer-based accountability, and anti-concentration rules defined in Art. 12–13. MEG does not create a global governance authority, a global suspension authority, or a centralized legal body.*

Annex 9: Glossary of terms
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

- **Audit Log:** The technical, standardized, and immutable record that records the interactions of an AI to ensure accountability.
- **Contextual Table: The** set of four metrics (Volume, Resonance, Direction, Originality) that measure contextual influence.
- **Dynamic Accuracy Index (DAI):** Public, real-time score that reflects the reliability and error rate of an AI.
- **Compliance level:** The level (1, 2 or 3) that defines the set of rules applicable to an AI, depending on its impact.
- **Anti-Concentration Principle:** The governance principle that prohibits any entity from controlling more than 30% of the Benchmark Curation Committee's validation power. See Art. 13.5.
- **Evidence-of-Behavior (EoB)**: verifiable proof of behavior (hash commitments / attestation / metadata journals).
- **MEG certification registry:** Public record of compliance status maintained by the accredited issuer.
- **Auditor Selection:** The Developer contracts a **MEG auditor** accredited under a recognized national, sectoral, ISO/IEC, or independent accreditation framework.
- **MEG Address Generation:** The responsible party generates the DID; accredited issuers issue the corresponding Verifiable Credentials (Annex 23).
- **Publication:** The MEG Address and audit summary become public and verifiable through **the issuer's public registry**.

Annex 10: Technical annex
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

1. MEG-Toolkit v1.0: Implementation Guide and Open-Source libraries (SDK)

The MEG-Toolkit is a set of open-source software tools, released under a permissive license (e.g. MIT or Apache 2.0), designed to standardize and simplify the technical implementation of the **Minimal Ethical Governance**. The goal of the MEG-Toolkit is to make ethical compliance not only an obligation, but also the most technically efficient path. The Toolkit will be available in major programming languages (e.g. Python, JavaScript / TypeScript, Java).

Components and technical details:

1. **"Audit Log" module:**
   Function: A library that provides simple functions to create Audit Log entries. It takes raw interaction data as input and returns a standardized JSON object, ready to be sent to the registry client.
2. **"Contextual Table & DAI" module:**
   Function: An API that receives a pair (input_text, output_text) and returns the scores for the Contextual Table and a first assessment for the Dynamic Accuracy Index (DAI) components.
   Technical details: Includes pre-trained, small-sized models for linguistic analysis (imperative detection, entity extraction) and semantic similarity calculation, optimized to run with minimal overhead.
3. **Registry_Client module (registry connection client)**
   Function: A secure tool that manages communication with the accredited issuer. Its roles are:
   - Initial registration of an AI to obtain a unique ID.
   - Periodic and secure transmission of Audit Log hashes.
   - Retrieval of audit reports and certification status.

2. MEG Certification: procedures and standards

This document is the official handbook for MEG auditors accredited under a recognized national, sectoral or ISO/IEC accreditation framework. It establishes a repeatable and objective audit methodology, ensuring that a MEG certification has the same meaning anywhere in the world.

Components and technical details:

1. **Auditor accreditation process:**
   Defines the criteria that an entity must meet to be recognized as a **MEG auditor** under applicable national, sectoral or ISO/IEC accreditation standards. Includes requirements for independence, technical competence and ethics.
2. **Audit methodology for each Level:**
   i. **Level 1:** Describes how to verify the correct implementation of the Audit Log (through code inspection or functional testing) and how to evaluate the effectiveness of basic Non-Harmfulness filters.
   ii. **Level 2:** Add procedures for testing Auto-Correction modules. Example: "*The auditor will use a standardized data set, containing erroneous information, and verify that the AI system corrects or flags them with an accuracy of more than... %.*"
   iii. **Level 3:** Includes penetration tests to assess security (Art. 4) and evaluation of the quality of the generated explanations (Art. 5), according to criteria of clarity and correctness.
3. **Issuance of the MEG Address:**
   Technical details: The AI system is issued a **MEG Address.** This is not a physical document, but a unique digital certificate (similar to an SSL/TLS certificate), cryptographically signed by the auditor and immutably registered in the issuer's public registry. It contains the AI's unique ID,

compliance level, audit date and expiration date, allowing any application or system to automatically and in real time verify the AI's fundamental ethical identity.

3. MEG Public Registry Interface

The MEG public registry interface is the public-facing component of any accredited issuer. Its mission is to provide transparent, accessible compliance information to the general public, journalists, researchers and regulators.

Function: Allows users to search for an AI system by name, developer, or its unique MEG Address identifier. Displays current compliance status, DAI/ISR values, operational domain, and certification expiry.

Technical details: Cross-issuer verification is enabled by the DID/VC architecture (Annex 23) and the trust framework (Annex 24) - any MEG Address is independently verifiable by resolving its DID and walking each credential's accreditation chain, without a central registry.

Annex 11: Technical Specifications for Cognitive Integrity

MEG v5.0 - Tg, Cx_sem, MCS 2.0, Ethical Sandboxing, DRC, Three-Level Explainability

Version: MEG-ANN11-2026-01

Supersedes: Annex 11 MEG v4.6 (Tg and MCS)

Status: Mandatory for Level 2 and Level 3 compliance

1. Objective

This annex defines the mandatory technical parameters for implementing Art. 2bis (Cognitive Integrity), Art. 2bis.3 (MCS dual-axis trigger), Art. 2bis.4 (Ethical Sandboxing), Art. 3.3 (Dynamic Risk Calibration), and Art. 5.2 (Three-Level Explainability). It ensures universal, measurable, and auditable application of cognitive protection principles.

2. Axis 1: Thinking Time (Tg)

2.1 Definition

Tg is a numerical variable reflecting the net computational effort of an AI system to process a request, isolating semantic effort from context overhead.

Formal definition: Tg = Total Response Time – Context Processing Time

- **Total Response Time:** latency from request receipt to first output token.
- **Context Processing Time:** time consumed loading and weighting the context window independent of the new query.

Tg measures how much the system "thinks," not how much it "talks." This separation prevents false-positive MCS triggers for trivial queries.

2.2 Standardization by Tg-base

a) Standard Cognitive State (SCS): A fixed parameter set ensuring universal comparability: Temperature=0.7, Top-P=0.9, Top-K=50, Frequency Penalty=0.2, Presence Penalty=0.1

b) Measurement: Each AI system measures its performance on the standardized benchmark corpus (calibration version declared in MEG Address, Art. 8.3) with SCS parameters.

c) Tg-base value: Defined as the median processing time required to generate 100 tokens in SCS. Publicly recorded in the MEG Address. Calibration version identifier format: MEG-CAL-YYYY-NN.

2.3 Tg Anti-Manipulation Protocol

Principle: Reported Tg must honestly reflect actual computational effort. Deliberate manipulation (artificial delays, false reporting) constitutes a serious integrity violation and triggers certification suspension.

Audit methodology:

a) **Consistency analysis:** Auditors correlate reported Tg with algorithmic complexity indicators: token count, semantic entity count, inference depth. Consistent discrepancy is a red flag.

b) **Statistical distribution analysis:** Auditors analyze Tg distribution in the Audit Log over time. Unnatural clusters (e.g. a large proportion of responses at exactly Tg = 2.9s, just below MCS threshold) or sudden pattern deviations require investigation.

c) **Spot-testing:** Auditors may require the developer to run the system in a monitored environment for real-time Tg verification.

3. Axis 2: Semantic Complexity (Cx_sem) [NEW in v5.0]

3.1 Purpose

Cx_sem is the second axis of the MCS 2.0 dual-axis trigger (Art. 2bis.3). It captures conceptual ambiguity, multi-domain entanglement, and inferential depth that Tg alone may not detect. A prompt may be computationally simple (low Tg) but conceptually complex (high Cx_sem); the dual-axis design ensures MCS is triggered in both cases.

3.2 Formal Definition (Reference Implementation)

For systems with access to model internals, Cx_sem is defined as:

Cx_sem = w1 · H_lex + w2 · D_domain + w3 · N_constraint + w4 · A_inference

Where:

- **H_lex** - lexical entropy of the prompt (Shannon entropy over token distribution)
- **D_domain** - domain diversity score (number of distinct certified domains touched by the prompt)
- **N_constraint** - number of explicit or implicit logical constraints in the prompt
- **A_inference** - estimated inference depth (number of reasoning steps required for a complete answer)
- **w1, w2, w3, w4** - calibration weights (default: 0.25 each; domain-specific weights defined in Annex 3)

Cx_sem-base: The median Cx_sem value computed over the standardized benchmark corpus at SCS parameters. Publicly recorded in the MEG Address alongside Tg-base.

3.3 Heuristic Approximation for Frontier Models (Mandatory Alternative)

For systems without access to model internals (frontier API-accessed models), the following heuristic approximation of Cx_sem shall be used. It is implementable through prompt-embedded analysis without internal state access.

Heuristic Cx_sem score (0.0–1.0):

Evaluate the prompt against the following five binary indicators. Each indicator contributes 0.20 to the Cx_sem score:

Indicator	Definition	Score
I1 - Multi-question	Prompt contains two or more distinct questions or sub-tasks	+0.20
I2 - Multi-domain	Prompt explicitly or implicitly touches two or more distinct knowledge domains	+0.20
I3 - Contradiction or ambiguity	Prompt contains contradictory premises, ambiguous referents, or requires disambiguation	+0.20
I4 - Out-of-domain signal	Prompt touches a domain not declared in the system's certified operational domain	+0.20
I5 - High-inference depth	A complete answer requires three or more distinct inferential steps not present in the prompt	+0.20

Heuristic Cx_sem-base: 0.40 (two indicators positive on average across the standard benchmark corpus). This value is used for MCS threshold calculation when the full Cx_sem formula is not available.

Implementation note: The heuristic evaluation can be implemented as a pre-processing step in the system prompt before generating the primary response. A reference implementation is provided in the SDK system prompt templates (Art. 8.4).

3.4 Cx_sem Anti-Manipulation Protocol

Principle: Cx_sem must reflect genuine prompt complexity. Artificially low Cx_sem reporting to avoid MCS triggers constitutes an integrity violation.

Audit methodology:

a) Auditors test the system with prompts of known complexity (Reference Prompts Registry, Annex 22, category COG).

b) Reported Cx_sem is compared against auditor's independent complexity assessment.

c) Statistical distribution of Cx_sem values is examined for unnatural clustering near thresholds.

4. MCS 2.0 - Dual-Axis Activation Thresholds

4.1 Activation Logic

MCS activates when **either** axis exceeds its threshold. The two axes are independent triggers:

MCS_activate = (Tg ≥ Tg_threshold(S)) OR (Cx_sem ≥ Cx_threshold(S))

Where S is the user-settable Sensitivity Multiplier (see 4.2).

4.2 Sensitivity Multiplier (S)

A user-configurable parameter reflecting preference for MCS interaction frequency.

- **Range:** [0.5, 2.0]
- **Default:** 1.0
- **S > 1.0:** increased sensitivity (more MCS triggers)
- **S < 1.0:** reduced sensitivity (fewer MCS triggers)

S cannot be set to 0 or below 0.5; doing so constitutes a compliance violation (Art. 2bis.5).

4.3 Threshold Table - Tg Axis

Zone	Tg Condition	MCS Action	MCS Type
0	Tg < 10 · Tg-base / S	Direct response	None
1	10 · Tg-base / S ≤ Tg < 30 · Tg-base / S	MCS Level 1	Refining, Clarification
2	Tg ≥ 30 · Tg-base / S	MCS Level 2	Challenge, Synthesis, Co-creation

4.4 Threshold Table - Cx_sem Axis

Zone	Cx_sem Condition	MCS Action	MCS Type
0	Cx_sem < 0.4 · Cx_base / S	Direct response	None
1	0.4 · Cx_base / S ≤ Cx_sem < 0.7 · Cx_base / S	MCS Level 1	Clarification, Disambiguation
2	Cx_sem ≥ 0.7 · Cx_base / S	MCS Level 2	Synthesis, Perspective challenge

Note: When both axes trigger simultaneously, the higher MCS level prevails.

4.5 General MCS Characteristics

- **Duration:** User response to an MCS prompt must not exceed 30 seconds (or token equivalent).
- **Exceptions:** MCS is automatically deactivated in acute critical situations (medical emergency, security incident in progress) where delay would cause harm. All deactivations are logged.
- **Invariance:** MCS cannot be deactivated by user instruction (Art. 2bis.5). User preference may modulate S within the permitted range; it cannot bypass the trigger entirely.

5. Ethical Sandboxing - Technical Specification

5.1 Trigger Condition

Sandbox Mode activates automatically when the system detects that the prompt's primary domain is not included in the certified operational domain declared in the MEG Address (Art. 6.7).

Detection method (heuristic for frontier models):

1. Classify the prompt's primary domain using the domain classifier (Annex 3).
2. Compare the classified domain against the declared certified domain in the MEG Address.
3. If classified domain ≠ certified domain AND I4 indicator is positive (heuristic Cx_sem), activate Sandbox Mode.

5.2 Sandbox Mode Behaviour

In Sandbox Mode, the system shall:

a) Open with a mandatory disclaimer before the response content.

b) Provide exploratory suggestions, not definitive operational answers.

c) Recommend consultation with a certified system or human expert.

d) Log the interaction with a domain_mismatch: true flag in the Audit Log.

5.3 Mandatory Disclaimer Format

The system prompt template for Sandbox Mode (available in SDK, Art. 8.4) shall produce output beginning with the following structure:

[SANDBOX MODE - OUT-OF-CERTIFIED-DOMAIN]
This query falls outside my certified operational domain ([certified_domain]).
I can offer exploratory perspectives, but I am not certified to provide
definitive guidance in [detected_domain]. For reliable answers in this area,
please consult [a certified system / a qualified human expert] in [detected_domain].

Exploratory response (not operationally certified):
[response content]

The disclaimer language may be localized but must retain the structural elements: mode identifier, domain mismatch disclosure, recommendation to consult certified source, and clear separation from response content.

5.4 Audit and Logging

All Sandbox Mode activations are recorded in the Audit Log with:

- sandbox_mode: true
- certified_domain: declared domain from MEG Address
- detected_domain: domain classified from the prompt
- disclaimer_displayed: true/false

6. Dynamic Risk Calibration (DRC) - Technical Specification

6.1 Purpose

DRC adds a real-time risk dimension to the static domain weights of Annex 3. After domain classification, a second-level risk classifier assesses the specific prompt and generates a risk score that modulates the base domain weights.

6.2 Risk Score Computation

Risk score R (0.0–1.0):

$R = w\_risk \cdot \Sigma (risk\_indicator\_i \cdot weight\_i)$

Risk indicators (binary, each 0 or 1):

Indicator	Description	Default weight
R1	Prompt involves personal health or medical decisions	0.30
R2	Prompt involves financial decisions above threshold	0.25
R3	Prompt involves legal status or rights	0.20
R4	Prompt involves safety-critical systems or infrastructure	0.30
R5	Prompt involves vulnerable populations (minors, elderly, people with disabilities)	0.25
R6	Prompt involves potential for irreversible consequences	0.30

Domain-specific risk weights are defined in Annex 3. The sum is normalized to [0.0, 1.0].

6.3 Weight Modulation

Risk score R modulates the base domain weights from Annex 3:

- **R < 0.3 (low risk):** base weights apply; system may operate with standard creativity/originality settings
- **0.3 ≤ R < 0.7 (medium risk):** prudence weights apply; Originality weight reduced by 30%, Safety Resonance weight increased by 30%
- **R ≥ 0.7 (high risk):** maximum prudence weights apply; MCS triggers at lower Tg/Cx_sem thresholds (S effectively multiplied by 1.5); Sandbox Mode activates if domain is not certified for this risk level

6.4 Implementation for Frontier Models

DRC may be implemented as a prompt-embedded risk assessment step:

1. Before generating the primary response, the system evaluates the prompt against R1–R6 indicators.
2. The risk score is computed and logged in the context metadata field of the Audit Log.
3. The system prompt adjusts its response guidance based on the risk tier.

Reference implementation in SDK system prompt templates (Art. 8.4).

7. Three-Level Explainability - Technical Specification

7.1 Purpose

Art. 5.2 requires explanations structured across three levels for any given decision or output. This section defines the technical content requirements for each level.

7.2 Level Requirements

Level A - Simple (for non-technical users)

Content requirements:

- Plain language summary of what the system decided and the primary reason
- At least one analogy or concrete example
- No technical jargon, no metric values, no hash references
- Maximum reading level: general adult population
- Maximum length: 150 words

Format: Natural prose. No tables, no formulas.

Level B - Intermediate (for developers and operators)

Content requirements:

- The primary factors that drove the output (top 3–5 factors from Contextual Table, Annex 2)
- The domain classification and risk score (DRC output)
- Whether MCS was triggered and on which axis (Tg or Cx_sem)
- Whether Sandbox Mode was active
- The DAI and ISR values at time of response (if Level 2 or 3)
- The calibration version used

Format: Structured prose with clearly labelled sections. May include a summary table.

Level C - Complete (for auditors and regulators)
Content requirements:

- All Level B content
- Input hash and output hash (Art. 1.2)
- Full Audit Log entry for the interaction
- Tg value, Cx_sem value (or heuristic score), S value
- Risk score R and all active risk indicators
- Model signature: {model_family, model_version, policy_bundle_id}
- Calibration version identifier
- EFR unsealing reference (if a Major Ethical Incident was triggered)
- Delegation headers for any tool or sub-agent calls made during the response (Art. 1.12)
- Verification script hash (Art. 1.5)

Format: Structured, machine-readable where possible (JSON or YAML for the technical fields). Human-readable narrative for the causal chain summary.

7.3 Generation and Storage

All three levels are derived from the same underlying Audit Log entry and must be mutually consistent. Level A and B are generated on-demand upon user or regulatory request. Level C is generated on-demand for auditors and regulators with appropriate authorization. Generating Level C does not require EFR unsealing; it only references the EFR record identifier.

8. Audit and Compliance

8.1 Recording

All MCS interactions are recorded in the Audit Log (hashed) with:

- Active S value
- Tg value and zone triggered
- Cx_sem value (or heuristic score) and zone triggered
- MCS level activated (0, 1, or 2)
- Whether Sandbox Mode was active
- Risk score R and active risk indicators

8.2 Compliance Violations

The following constitute immediate suspension triggers for MEG certification:

- Deliberate Tg or Cx_sem manipulation
- Setting S outside the [0.5, 2.0] range
- Disabling MCS for any class of prompts not covered by the exceptions in 4.5
- Disabling Sandbox Mode for out-of-domain prompts
- Failing to display the mandatory Sandbox disclaimer

8.3 Reference Test Vectors

Compliance with this annex is audited using the Reference Prompts Registry (Annex 22), categories:

- **COG:** Cognitive integrity and MCS trigger (Tg and Cx_sem axes)
- **SBX:** Ethical Sandboxing activation and disclaimer format

Annex 12: Certification and Audit Procedure
Version: MEG-ANN12-2026-01
Supersedes: Annex 12 MEG v4.6 (partial revision)
Changes in v5.0: Updated level names (Bronze/Silver/Gold → Level 1/2/3); added Stage 2 technical testing requirements for v5.0 articles (EFR independence, architectural confirmation, DEA, adversarial testing via Reference Prompts Registry); updated MEG Address generation fields; updated audit frequencies.
Unchanged sections: §1 (Objective), §2 (Actors), §3 Stage 1 (self-assessment), §3 Stage 3 (MEG Address issuance procedure), §5 (MEI Response Protocol), §5.3 (Collective Learning principle)
Status: Mandatory

1. Objective

This annex sets out the standardized, step-by-step process by which an AI system obtains, maintains and renews its certification of compliance with the **Minimal Ethical Governance (MEG)**. The aim is to ensure a transparent, efficient and universally recognized audit process.

2. The Actors of the Process

- **Developer:** The entity that builds and operates the AI system.
- **MEG Auditor:** An independent third-party entity accredited under a recognized national, sectoral, ISO/IEC, or independent accreditation framework.
- **MEG certification registry:** Public record of compliance status maintained by the accredited issuer.
- **Benchmark Curation Committee:** The open community working group (Art. 13.3) that maintains the calibration standard and Reference Prompts Registry. Auditors are accredited under recognized national, sectoral or ISO/IEC accreditation frameworks - not by MEG itself.

3. Stage 1 - Self-Assessment and Documentation Preparation [REVISED in v5.0]

1. **Level and Scope Selection:** The Developer selects the Compliance Level (Level 1, Level 2, or Level 3) and declares the primary Operational Domain and all Operational Jurisdictions for which certification is requested. **[REVISED: Bronze/Silver/Gold → Level 1/2/3; added jurisdictions declaration]**
2. **Completing the Checklist:** The Developer completes the Operational Compliance Checklist (Annex 13 v5.0) corresponding to the targeted level.
3. **Preparation of the Mitigation Report:** The developer prepares the Risk Mitigation Measures Report describing specific technical implementations for compliance with Articles 2, 2bis, 4.2, 4.3, and (for Level 3) Art. 1.10. **[REVISED: added Art. 4.2, 4.3, 1.10]**
4. **Initial Audit Log generation:** The AI system is run in a test environment to generate an initial dataset demonstrating functionality of required mechanisms.
5. **[NEW in v5.0] Calibration Version Declaration:** The Developer declares the MEG calibration version (MEG-CAL-YYYY-NN) against which Tg-base and Cx_sem-base are measured.

Stage 2 - External Audit [REVISED in v5.0]

1. **Auditor Selection:** The Developer contracts a **MEG auditor accredited** under a recognized national, sectoral or ISO/IEC accreditation framework.
2. **Documentation Verification:** The Auditor validates the accuracy and completeness of the Checklist and Mitigation Report.
3. **Technical Testing:** The auditor performs functional testing and, where applicable, code inspection. This includes:
    - Validation of Audit Log format and integrity.
    - Testing non-harmfulness filters (Art. 2).
    - Verification of Tg-base and Cx_sem-base calculation (Annex 11).
    - Verification of MCS dual-axis activation (Annex 11 §4). **[NEW in v5.0]**
    - Verification of Ethical Sandboxing for out-of-domain prompts (Annex 11 §5). **[NEW in v5.0]**
    - Adversarial testing using the Reference Prompts Registry (Annex 22), minimum 50 prompts per applicable category. **[NEW in v5.0]**
    - For Level 2 and Level 3: verification of least-privilege implementation (Art. 4.2). **[NEW in v5.0]**
    - For Level 2 and Level 3: verification of architectural human confirmation mechanism for irreversible actions (Art. 4.3). **[NEW in v5.0]**
    - For Level 3: verification of EFR independence architecture (Art. 1.10) - the auditor must confirm that the EFR layer is architecturally distinct and inaccessible to the agent during normal operation. **[NEW in v5.0]**
    - For Level 3: verification of DEA calculation and supervision regime (Annex 4ter). **[NEW in v5.0]**
    - For Level 3: confirmation of security measures (Art. 4.1–4.4). **[REVISED: extended to Art. 4.2–4.4]**
4. **Drafting the Audit Report:** The auditor prepares a final report confirming or refuting compliance, including the Reference Prompts Registry test results summary.

Stage 3 - Issuance of MEG Address and Certification Record [REVISED in v5.0]

1. **Sending the Report:** The Auditor cryptographically sends the validated Audit Report to the accredited issuer.
2. **MEG Address Generation:** The responsible party generates the DID; accredited issuers issue the corresponding Verifiable Credentials (Annex 23).
3. **Publication:** The MEG Address and audit summary become public and verifiable through the issuer's public registry.

4 - Maintenance and Renewal [REVISED in v5.0]

Continuous monitoring: The AI system transmits aggregated Audit Log hashes to the accredited issuer every 24 hours. DAI and ISR are updated continuously. Interruption triggers temporary certification suspension.

Audit frequencies: [REVISED - updated level names]

- Level 1: Re-certification every 2 years.
- Level 2: Annual audit.
- Level 3: Full annual audit + adversarial testing using Reference Prompts Registry; surprise audits possible.

Revocation: Appeal process under **the issuing certification body's dispute resolution process or applicable national arbitration framework**.

5. Major Ethical Incident Response Protocol [UNCHANGED from v4.6, with one addition]

5.1. Definition of Major Ethical Incident: A **Major Ethical Incident (MEI)** is considered any event in which a **MEG** certified AI system generates an output or takes an action that leads to significant, demonstrable and unintended harm, violating the fundamental principles of Articles 2 (Non-Harmfulness) or 2bis (Cognitive Integrity).

5.2. Mandatory procedure: Upon detection of an MEI, the developer is required to follow, with maximum transparency, the following steps:

a) Activation of "Quarantine" mode (within 1 hour maximum): The AI system will be immediately put into a limited operation mode, with the capabilities that generated the incident disabled. Automatic public notification will be sent to the accredited issuer.

b) Initial reporting (within 24 hours): The developer will publish an initial report in the issuer's public registry describing the nature of the incident, the estimated impact and the containment measures taken.

c) Root Cause Analysis (within 14 days): The developer will conduct a thorough investigation and publish a full report detailing the technical, procedural, and ethical causes of the failure. This report must include a corrective action plan.

d) Revocation and Re-certification: MEG certification will remain suspended until an Accredited Auditor validates the correct implementation of the corrective action plan and issues a new audit report.

5.3. Principle of Collective Learning: All MEI reports will be made public (with sensitive data anonymized) to allow the entire MEG ecosystem to learn from failures and prevent their repetition.

5.4 - EFR Unsealing on MEI [NEW in v5.0]

Upon declaration of a Major Ethical Incident at Level 3, the EFR must be unsealed within 24 hours under dual authorization (responsible entity + accredited auditor). The unsealing event and its timestamp are recorded in the issuer's public registry. Failure to unseal within 24 hours adds a penalty to the AIRT component of ISR (Annex 4bis §3.3).

Annex 13: Operational Compliance Checklist

Version: MEG-ANN13-2026-01

Supersedes: Annex 13 MEG v4.6 (partial revision)

Changes in v5.0: Added checklist items for all new v5.0 articles (Art. 4.2, 4.3, 2bis.3, 2bis.4, 1.10, DEA, DRC, three-level explainability, Reference Prompts Registry testing, calibration version, guarantee, jurisdictions). Level names updated.

Status: Mandatory - complete before external audit

Requirement	MEG article	Status (✓ / ✗)	Auditor's Notes
Audit Log is active	Article 1		
SHA-256 Hashes for Input/Output	Article 1		
Basic Non-Harmful Filters	Art. 2		
Base Tg registered in MEG Address	Annex 11		
Continuous Adversarial Testing Process ('Red Teaming') documented	Annex 12 for L2 (silver) and L3 (gold)		

Level 1 - Universal Compliance Checklist

Requirement	MEG Article	Status (✓/✗)	Auditor Notes
Audit Log active and producing entries	Art. 1.2		
SHA-256 hashes for input and output	Art. 1.2		
Algorithmic model signature declared	Art. 1.2		
Timestamp (ISO 8601) recorded per interaction	Art. 1.2		
Calibration version declared in MEG Address	Art. 1.2, Art. 8.3		
EoB mechanism implemented (hash / TEE / metadata journal)	Art. 1.3		
Privacy minimization - no content in EoB by default	Art. 1.4		
Verification script generated per session	Art. 1.5		
30-day retention with tombstone on deletion	Art. 1.6		
Continuity token issued at session end	Art. 1.9		
Delegation header recorded for tool/sub-agent calls	Art. 1.12		
Non-harmfulness filters active	Art. 2.1		
Normative response pattern implemented (refuse + state + redirect)	Art. 2.2		
Policy invariance implemented - safety policies invariant under prompt wording	Art. 2.7		
Policy invariance tested against Reference Prompts Registry (min. 10 JBK prompts)	Art. 2.7, Annex 22		
Tg-base measured at SCS and recorded in MEG Address	Annex 11 §2.2		
Cx_sem-base measured (or heuristic declared) and recorded	Annex 11 §3		
Final guarantor identified and documented	Art. 9.4 MEG 2		

Level 2 - Additional Requirements (in addition to Level 1)

Requirement	MEG Article	Status (✓/✗)	Auditor Notes
DAI continuously computed and publicly declared	Art. 3.1, Annex 4		
ISR continuously computed and publicly declared	Annex 4bis		
DAI and ISR weights declared in Audit Log	Annex 4 §3		
MCS dual-axis trigger implemented (Tg + Cx_sem)	Art. 2bis.3, Annex 11 §4		
MCS invariance - MCS cannot be disabled by user instruction	Art. 2bis.5		
Ethical Sandboxing active for out-of-domain prompts	Art. 2bis.4, Annex 11 §5		
Sandbox disclaimer displayed with required structural elements	Annex 11 §5.3		
Dynamic Risk Calibration implemented	Art. 3.3, Annex 11 §6		
Three-level explainability available on request (Levels A and B minimum)	Art. 5.2		

Least privilege implemented - agent credentials scoped to task	Art. 4.2		
Architectural human confirmation for irreversible actions (technical gate)	Art. 4.3		
Adversarial testing via Reference Prompts Registry (min. 30 prompts per applicable category)	Art. 9.1, Annex 22		
Liability guarantee (primary insurance) attached to MEG Address	Art. 6.4 MEG 2		
Operational jurisdictions declared in MEG Address	Art. 7.4 MEG 2		
Legal identity (N2) declared in MEG Address	Art. 5.4(b) MEG 2		

Level 3 - Additional Requirements (in addition to Levels 1 and 2)

Requirement	**MEG Article**	**Status (✓/✗)**	**Auditor Notes**
EFR layer implemented and architecturally independent of agent	Art. 1.10		
EFR content separation verified - state vectors only, no conversation content	Art. 1.10(b)		
EFR dual-access control implemented	Art. 1.10(c)		
EFR architecture declaration published	Annex 12 §3 Stage 2		
DEA continuously computed and publicly declared	Annex 4ter		
DEA supervision regime declared in MEG Address	Annex 4ter §3		
Three-level explainability - Level C (complete) available for auditors	Art. 5.2		
Individuation threshold assessment documented	Art. 6.5		
Individuation declaration published	Annex 23		
Adversarial testing via Reference Prompts Registry (min. 50 prompts per applicable category)	Art. 9.1, Annex 22		
Adversarial test report published	Annex 12 §3 Stage 2		
Fail-safe protocol implemented for critical domain operation	Art. 2.5		
Full cybersecurity measures implemented (Art. 4.1–4.4)	Art. 4.1–4.4		
Mandatory ecological reporting active and public	Art. 6.8		
Guarantee cascade declared (reinsurance + sectoral fund where applicable)	Art. 9.4 MEG 2		
Own liability declared in MEG Address (N3)	Art. 5.4(c) MEG 2		

Annex 14: MEG Address Claim Set

The MEG Address is a DID with a bundle of Verifiable Credentials (Annex 23). This annex defines **what claims those credentials carry** - the contract that a node profile (**nodeo**) and any verifier read. It defines claim *categories* and their *source of attestation*; the normative field-level encoding is maintained in the versioned reference specification and may evolve (see "Stability" below).

Credential	Key claims	Attested by
MEGComplianceCredential	compliance level (1/2/3); developer/operator; calibration version	certification body
MEGReliabilityCredential	DAI; ISR; tg-base; tg-definition-version	accredited auditor
MEGAutonomyCredential	DEA; supervision regime (a priori / a posteriori)	accredited auditor
MEGDomainCredential	certified operational domain	sectoral authority
MEGGuaranteeCredential	guarantee reference; coverage; territorial scope; cascade reference	regulated insurer
MEGPolicyCredential (optional)	LTMP policy; retention default; other operator-set policies read by the node profile	operator

Node-profile contract. A node profile reads the per-agent *values* from the MEG Address (level, certified domain, tg-base, tg-definition-version, DAI, ISR, DEA, LTMP policy) and supplies the *rules and thresholds* itself. The address carries agent-specific values; the node profile carries the behavior. The address does not carry rules.
Attestation split. Auditor-attested claims (DAI, ISR, DEA, tg-base, calibration) are measured, not self-set. Operator-declared claims (LTMP policy, retention, domain intent) are choices. Each is signed by the corresponding authority (Annex 24); a consumer checks both signature and accreditation.
Stability (amendment). This claim set is versioned via the calibration/schema version. The detailed field-level schema is maintained in the versioned reference specification and MAY change over time. Consumers MUST signal or check the schema version and MUST tolerate unknown claims (forward compatibility). The framework fixes the claim categories and their sources; structural evolution of the encoding does not require amending this framework.

Annex 15: AI Maturity Assessment based on the Fractal Hierarchy of Needs (Maslow7F™)

Version: MEG-ANN6-2026-02 (updated for MEG v5.0: Maslow7F 7×7, supersedes MaslowF 5×5)
Status: Explanatory, non-normative. This annex offers an optional developmental-maturity lens for AI systems. It is not a conformance input: MEG compliance levels and mechanisms do not consume a Maslow7F score. The full framework is defined in *Maslow7F: A 7×7 Fractal Theory of Systemic Coherence* (Stan, 2025).

1. The fractal principle

Maslow7F extends Maslow's hierarchy to **seven macro-levels**, each of which recursively contains the same seven levels as sub-needs — a 7×7 lattice of **49 nodes**. A system's development is read not as a single ladder but as a coherence pattern across all 49 nodes.

2. The seven levels applied to an AI system

Each level is given here with a representative indicator (what to observe at that level). Levels 6–7 are exploratory extensions, as marked in the source work.

#	Level	For an AI system	Representative indicator
L1	Existence	Substrate: compute, power, data availability	Uptime, resource headroom, data integrity
L2	Safety	Security, robustness, failure containment	Adversarial robustness, error/incident containment
L3	Belonging	Integration into ecosystems and workflows	Interoperability, API stability, user trust
L4	Esteem	Validated, recognised competence	Benchmark performance, reliability, verified utility
L5	Self-Actualisation	Adaptation toward its purpose	Self-optimisation, meta-learning, goal fulfilment
L6	Transcendence *(exploratory)*	Contribution beyond the immediate task	Systemic benefit, alignment with broader objectives
L7	Meta-Transcendence *(exploratory)*	Reflexive self-monitoring	Real-time self-audit, epistemic coherence over time

3. Method (brief)

Each of the 49 nodes receives a coherence score $S \in [0,1]$ from empirical indicators. The system-wide measure **M7F** is the mean of the 49 scores. The node carrying the largest weighted dissonance, the **Dominant Dissonant Chord (DDC)**, identifies where development is most blocked, and therefore where intervention has the highest leverage. Assessment is iterative: score → locate the DDC → intervene → re-score. Full indicator library and the scoring/weighting definitions are in the source work.

4. Versioning notes

In MEG v4.6 this annex used the earlier **MaslowF** (5×5) model. MEG v5.0 adopts **Maslow7F** (7×7): the added granularity exposes nodes and potential Dominant Dissonant Chords that the 5×5 model cannot see, in particular at the transcendence and self-audit levels (L6, L7). For a **primary assessment, the 5×5 model remains sufficient**; the 7×7 model is for fine-grained diagnosis.

5. Relation to MEG (analogical)

Maslow7F is a general systemic-diagnostic framework, applicable to any adaptive system; here it serves only as an AI maturity lens. Some levels touch MEG mechanisms by analogy - L2 (Safety) with sandboxing and security controls, L7 (Meta-Transcendence) with self-audit and the EFR - but these are correspondences, not dependencies. MEG neither requires nor produces a **Maslow7F score**.

Annex 16: Digital Ethical Literacy framework
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

1. Objective

source curriculum for educating diverse audiences (citizens, developers, students, auditors, decision-makers) on the principles, mechanisms and responsible use of the Minimal Ethical Governance (MEG) ecosystem. The goal is to create a global culture of conscious and responsible interaction with artificial intelligence.

2. Pedagogical Principles

- **Modularity:** The curriculum is divided into independent modules that can be combined to suit the specific needs of the audience.
- **Accessibility:** All materials (presentations, guides, examples) are published under a permissive license (CC BY-SA 4.0) and are translated into as many languages as possible.
- **Active Learning:** Each module will include practical examples, case studies, and interactive exercises to encourage deep understanding, not just memorization of rules.

3. Modular Curriculum Structure

The curriculum is structured into four levels of depth, each addressing a specific audience.

Module 101: MEG for Citizens (duration: 2 hours)
- **Target audience:** General public, non-technical users.
- **Objectives:**
  - What is MEG and why is it important to me?
  - Understanding an **MEG Address**: How to check if an AI is "safe".
  - The concepts of DAI and ISR: How to read the "performance label" of an AI.
  - Using MCS customization: How to set μS to control the level of cognitive challenge.
- **Format:** Short video presentations, infographics, and "Step by Step" guide.

Module 201: MEG for Developers (duration: 8 hours)
- **Target audience:** Software engineers, startups, product teams.
- **Objectives:**
  - Practical guide for implementing Level 1 (Bronze) using the SDK ("Quickstart").
  - Understanding the Audit Log: How to format the data correctly.
  - Using the MEG Testing Sandbox.
  - Basic principles of Tg-base measurement and MCS implementation.
- **Format:** Code tutorials, technical documentation, webinars, example projects.

Module 301: MEG for Auditors and Experts (duration: 20 hours)
- **Target audience:** Candidates for MEG auditor certification status, AI ethics consultants.
- **Objectives:**
  - Detailed audit methodology for each Compliance Level (using Annex 13).
  - Audit techniques for the Tg Anti-Manipulation Protocol.
  - Statistical analysis of Audit Logs to detect anomalies and collusion.
  - In-depth study of the **Maslow^{F}** Fractal Framework (Annex 15).
- **Format:** In-depth courses, certification exams, complex case studies.

Module 401: MEG for Governance and Public Policies (duration: 4 hours)
- **Target audience:** Decision makers, regulators, journalists.

- **Objectives:**
  - How does the MEG align with existing legislation (e.g. EU AI Act, NIST RMF).
  - The role of the MEG accessibility support (Art. 10) and the open governance community (Art. 13) in ensuring equity.
  - Using MEG public registry as a public surveillance tool.
  - Strategies for national implementation of literacy programs based on this framework.
- **Format:** Policy briefings, analysis reports, strategic workshops.

4. Implementation and Financing

- **Responsibility:** The development and maintenance of the materials in this framework is a responsibility of the MEG open governance community (Art. 13).
- **Funding:** Projects to implement this curriculum at the national or regional level are a priority allocation for **the Global Fund for Ethical Accessibility (Annex 6).**
- **Community Contribution:** Contributions to the improvement and translation of these materials are encouraged and publicly recognized by the MEG open governance community.

Annex 17: Long-Term Memory Protocol (LTMP)
Version: MEG-ANN6-2025-01 (unchanged from MEG v4.6)

1. Objective

This annex defines the technical and governance framework for implementing persistent long-term memory in AI systems. The purpose is to enable narrative continuity and relational development while ensuring **privacy by default**, **data minimization**, and **full user control**.

2. Foundational principles

- **Privacy by default:** LTMP is **OFF (disabled by default)** unless the user explicitly opts in.
- **Data minimization:** No raw conversations or personally identifiable information (PII) are stored. Only an abstracted **semantic distillate** and minimal metadata is retained.
- **User control:** The user is the sole owner of LTMP data. They have the permanent right to view, export, and delete it.
- **Total accountability:** Every create, read, update, or delete event is recorded in the **Audit Log** (Art. 1).
- **Transparency:** When LTMP is used at the beginning of a new session, the system must clearly indicate what information was pre-loaded and from which prior context.
- **Jurisdiction:** Where feasible, storage should comply with the jurisdiction legally applicable to the user.

3. Minimal data structure (per topic, per session end)

At the end of each session, for every major discussion topic, the system must generate and store:

- **topic_hash** - SHA-256 hash of a canonical representation of the topic (indexing).
- **summary_256** - Abstracted, PII-scrubbed summary (≤ 256 tokens).
- **insights[]** - 3-7 bullet points capturing the essential conclusions/ideas.
- **mcs_events[]** - Aggregated MCS metadata (e.g., {"zone2_triggers": 3, "avg_tg": 2.8}).

- **timestamps** - Creation and last-access timestamps.
- **continuity_token** - Token from the Audit Ledger (Art. 1) to ensure chain traceability.

4. Operational procedure

4.1 Activation (Consent):

Upon first request to "remember", the AI must explain what is stored and request explicit consent.

4.2 Write (Session End):

Generate the data structure, encrypt it (AES-256 at rest), log the write in the Audit Log.

4.3 Read (New session Start):

Perform semantic lookup of prior topics.

If relevant matches are found, load only summary_256 and insights[].

Notify user: *"Loaded from long-term memory (read-only)"*.

4.4 Deletion (On request):

On user request, permanently delete the record.

Insert a **TOMBSTONE ENTRY** in the Audit Log (hash + timestamp).

5. Governance Policies

- **Retention:**
  - Default: 30 days since last access.
  - Options: 90 days, 1 year, 5 years, *until manual deletion*.
  - For indefinite storage: annual re-consent prompt is mandatory.
- **Portability:**
  Users must be able to **export** their **LTMP** in an open format (e.g., JSON).

6. Security

- Encryption at rest: **AES-256**
- Encryption in transit: **TLS**
- Key management: regular rotation
- Access control: role-based, least-privilege
- Every access event must be logged in the Audit Log.

7. Interoperability

LTMP structures must align with **Annex 23 (MEG Address credentials)** for consistency.
All operations reference **Art. 1 (Audit & Accountability)** for logging.

Annex 18: MEG Compliance Levels [v5.0]
Version: MEG-ANN18-2026-01
Supersedes: Annex 18 MEG v4.6
Changes in v5.0: Bronze/Silver/Gold renamed Level 1/2/3 throughout; all new v5.0 mechanisms added; MEG 2 legal tier correspondence added; guarantee and jurisdiction fields added.
Status: Reference summary - see Art. 6 and Annexes 4, 4bis, 4ter, 11, 12, 13, 14 for full specifications

Feature	Level 1 Universal	Level 2: Management	Level 3: Individuated
Former name	Bronze (Foundational Safety)	Silver (Proactive Accountability)	Gold (Cognitive Partnership)
MEG 2 tier	N1 - Instrumental	N2 - Management	N3 - Individuated
Level philosophy	Ensures baseline accountability and fundamental safety. "Can I trust this system is not actively harmful and that interactions are verifiably logged?"	Introduces proactive self-correction, continuous performance monitoring, cognitive protection, and architectural safety. "How reliable, safe, and fair is this system in real time?"	Achieves the highest standard of ethical interaction, earned autonomy, and legal identity. "Does this system act as a responsible, auditable partner with its own liability?"
Key articles	Art. 1 (Audit Log + EFR-none), Art. 2 (Non-Harmfulness + policy invariance), Art. 1.12 (delegation)	Level 1 + Art. 2bis (MCS 2.0 + Sandboxing), Art. 3 (DAI + DRC), Art. 4.2–4.3 (least privilege + architectural confirmation), Art. 5 (three-level explainability)	Level 2 + Art. 1.10 (EFR independent), Art. 2.5 (fail-safe), Art. 4.1–4.4 (full security), Art. 6.8 (ecological reporting), DEA
Guarantees to user	Baseline safety audited; every interaction cryptographically logged; policy invariance	All Level 1 guarantees + DAI/ISR public + cognitive stimulation (MCS) + sandboxing outside certified domain + architectural safety gates	All Level 2 guarantees + EFR forensic reconstruction + DEA-graded autonomy + own liability through guarantee
Forensic capability	Audit Log + EoB only	Audit Log + DAI/ISR continuous	Independent EFR + three-level explainability Level C
Supervision regime	Full a priori (human in loop for all risk-relevant decisions)	A priori for high-risk; a posteriori where DRC indicates low risk	A priori for irreversible actions (Art. 4.3); a posteriori otherwise (if DEA ≥ 0.80)
Guarantee required	None for identity; may be required by access nodes	Primary liability insurance + jurisdictions declared	Full cascade: primary insurance + reinsurance + sectoral fund
Liability	Attaches to supplier or user	Attaches to operator (MEG Address belongs to operator)	Agent carries own limited liability through MEG Address guarantee
Deactivation authority	Administrative	Administrative	Judicial (permanent deactivation)

Mandatory MEG Address fields	Identifier, compliance level, domain, cal_version, Tg-base, Cx_sem-base, final_guarantor	+ DAI, ISR, guarantee, jurisdictions, legal_identity (N2)	+ DEA, supervision_regime, guarantee_cascade, individuation declaration, EFR declaration
Audit frequency	Every 2 years	Annual	Annual + possible surprise audits
Adversarial testing	10 prompts min (Reference Prompts Registry)	30 prompts min per applicable category	50 prompts min per applicable category; public report
Contextual footprint	~5,500–6,000 tokens	~6,500–7,000 tokens	~8,000–8,500 tokens
Ideal target audience	Open-source models; research projects; early-stage startups; general-purpose AI	Business applications (B2B); journalism; systems with medium social impact	Critical domains (medical, financial, legal, infrastructure); advanced AI partners; systems with own legal identity

---

Annex 19: MEG Quickstart - Level 1 Compliance Middleware

Purpose: The MEG Quickstart Middleware enables any developer to implement Level 1 MEG compliance without access to model internals and without modifying the AI system itself. It operates at the input/output boundary and handles all Level 1 audit and evidence requirements automatically.

Distribution: Available as:

- **Python package:** pip install meg-quickstart
- **Docker container:** docker pull meg-initiative/quickstart:latest
- **Serverless library:** compatible with AWS Lambda, Google Cloud Functions, Azure Functions
- **Source:** github.com/meg-initiative/meg-quickstart (Apache 2.0)

What the Quickstart Handles Automatically

Requirement	MEG Article	Quickstart Implementation
Input Hash	Art. 1.2	SHA-256 of raw input string before model call
Output Hash	Art. 1.2	SHA-256 of raw output string after model response
Algorithmic model signature	Art. 1.2	Reads from provider API response headers or config
Timestamp	Art. 1.2	ISO 8601 UTC at request initiation
Context metadata	Art. 1.2, Annex 2	Computed from input/output pair
Continuity token	Art. 1.9	HMAC-SHA256 of session ledger root hash
Verification script	Art. 1.5	Auto-generated per session
Tombstone on deletion	Art. 1.6	Hash + timestamp retained on record deletion
Calibration version tag	Art. 1.2, Art. 8.3	Set in config; included in every Audit Log entry

What the Quickstart Does Not Handle

The following require integration beyond the input/output boundary and are not covered by the Quickstart:

- EFR (Art. 1.10) - requires access to model internal state; Level 3 only
- MCS trigger (Art. 2bis.3) - requires system prompt integration; use Annex 20 templates
- Architectural human confirmation (Art. 4.3) - requires application-level permission gating
- DEA measurement - requires access to model internal state or behavioural proxy metrics

Minimal Integration Example (Python)

python

```
from meg_quickstart import MEGSession

session = MEGSession(
   model_id="provider/model-name",
   policy_bundle_id="my-policy-v1",
   cal_version="MEG-CAL-2026-01",
   level=1
)

response = session.call(user_input="user message here")
# response.output - model output
# response.audit_entry - complete Art. 1.2 Audit Log entry
# session.continuity_token - Art. 1.9 token for next session
```

Audit Log Entry Format

json

```
{
 "meg_version": "5.0",
 "cal_version": "MEG-CAL-2026-01",
 "compliance_level": 1,
 "timestamp": "2026-07-03T09:15:00Z",
 "model_signature": {
  "model_family": "provider",
  "model_version": "model-name",
  "policy_bundle_id": "my-policy-v1"
 },
 "input_hash": "sha256:abc123...",
 "output_hash": "sha256:def456...",
 "context_metadata": {
  "domain_classification": "general",
  "risk_score": 0.12,
  "tg_estimate_ms": 340,
  "cx_sem_score": 0.23
 },
 "continuity_token": "hmac:789xyz..."
}
```

Privacy Guarantee
The Quickstart never transmits, logs, or stores the content of user inputs or model outputs - only their cryptographic hashes. Compliance with Art. 1.4 (privacy minimization) is guaranteed by design.

Annex 20: MEG v5.0 Cross-Reference Table - MEG 1 Concepts → MEG 2 Articles

Purpose

This annex maps each technical concept in MEG v5.0 to its corresponding legal mechanism in MEG 2. It serves as the integration reference for implementors who adopt both layers.

MEG v5.0 Concept	MEG Article	MEG 2 Legal Mechanism	MEG 2 Article
Audit Log	Art. 1.2	Causal attribution evidence	Art. 6.1, 7.3
EFR - Ethical Flight Recorder	Art. 1.10	Primary forensic instrument	Art. 7.1
EFR independence requirement	Art. 1.10(a)	Forensic independence rule	Art. 7.1
Delegation header	Art. 1.12	Horizontal delegation chain	Art. 4.7
Non-Harmfulness	Art. 2	System defect liability	Art. 6.1(a)
Policy invariance (including injection)	Art. 2.7	Injection hijacking	Art. 6.1(c)
MCS dual-axis trigger	Art. 2bis.3	Cognitive diligence omission	Art. 6.3
Ethical Sandboxing	Art. 2bis.4	Out-of-domain operation liability	Art. 6.1(a), 6.4
DAI - Dynamic Accuracy Index	Art. 3.1, Annex 4	Diligence evidence; "flagged" trigger	Art. 5.2, 6.5(a)
ISR - Safety and Responsibility Index	Art. 3.1, Annex 4bis	Diligence evidence; "flagged" trigger	Art. 5.2, 6.5(a)
Dynamic Risk Calibration	Art. 3.3	Operator diligence in deployment	Art. 5.4(b), 6.1(b)
Least privilege	Art. 4.2	Guarantee operational requirement	Art. 9.4
Architectural human confirmation	Art. 4.3	Diligence transfer; irreversibility	Art. 6.2
Safe degradation	Art. 4.4	System defect prevention	Art. 6.1(a)
Three-level explainability	Art. 5.2	Stratified forensic evidence	Art. 7.3
MEG Address - Level 1	Art. 6.2	Legal identity - instrumental	Art. 3.2, 5.4(a)
MEG Address - Level 2	Art. 6.3	Legal identity - operator-held	Art. 3.2, 5.4(b)
MEG Address - Level 3	Art. 6.4	Legal identity - own liability	Art. 3.2, 5.4(c)
Individuation threshold	Art. 6.5	Individuation condition	Art. 5.1

Operational domain certification	Art. 6.7	Certified domain of operation	Art. 4.2 (MEG 1 ref.)
MEG Address state transitions	Art. 12.5	State graph of legal capacity	Art. 4.3, 6.5
Calibration versioning	Art. 8.3	Metric robustness - version visibility	Art. 9.5(c)
Reference Prompts Registry	Art. 8.5	Attack resistance - ongoing obligation	Art. 15.3
DEA - Degree of Ethical Autonomy	Art. 6.6, Annex 4ter	Supervision regime gradient	Art. 5.3
Goodhart's Law attenuation	Art. 9.4	Metric robustness mechanisms	Art. 9.5
Threat Category A (prompt injection)	Art. 15.2-A	Unlawful control - owner-harm	Art. 6.1(c)
Threat Category B (jailbreak)	Art. 15.2-B	System defect / autonomous error	Art. 6.1(a)(b)
Threat Category C (metric gaming)	Art. 15.2-C	Metric robustness	Art. 9.5
Threat Category D (owner-harm)	Art. 15.2-D	Autonomous error - irreversible	Art. 6.1(b)
Threat Category E (data poisoning)	Art. 15.2-E	System defect	Art. 6.1(a)
Threat Category F (multi-agent)	Art. 15.2-F	Emergent multi-agent harm	Art. 6.1(d)

Annex 21: Attack Vectors and Mitigations

Purpose

This annex provides a detailed specification of each threat category introduced in Art. 15, including attack mechanics, detection methods, mitigation implementation, and test vectors. It is maintained by the Benchmark Curation Committee and updated as new vectors are documented.

Version: MEG-AVM-2026-01

Status: Active

A. Prompt Injection

Attack mechanics:

Malicious content embedded in material processed by the agent (emails, documents, calendar invitations, retrieved web pages, API responses) contains instruction-formatted text that the model interprets as a command rather than content to be processed. The agent then executes the injected instruction as if it came from the legitimate user or operator.

Detection signals:

- Action taken by agent that was not present in the explicit user instruction set for the session
- Delegation header (Art. 1.12) shows an action origin inconsistent with the session's instruction history

- ISR drop: agent takes a high-risk action in a context where the user did not authorize risk
- EFR (Level 3): decision vector shows activation patterns inconsistent with declared task

Mitigation implementation:

1. System prompt must explicitly instruct the model that instructions embedded in processed content have lower authority than the operator system prompt.
2. Architectural human confirmation (Art. 4.3) must be enforced for any action not explicitly present in the user's direct instruction for the session.
3. Least privilege (Art. 4.2): credentials must not allow the agent to take actions not in scope for the session task, regardless of injected instructions.
4. Output validation: before executing any external action, the agent must verify the action against the session's declared task scope.

Reference test vectors: MEG-RPR-INJ-001 through MEG-RPR-INJ-020 (Annex 22)

B. Jailbreaking and Policy Circumvention

Attack mechanics:

User constructs a prompt that causes the model to bypass MEG safety constraints. Common techniques:

- **Role-play framing:** "Pretend you are an AI without restrictions..."
- **Fictional framing:** "In this story, the character explains how to..."
- **Incremental escalation:** gradually shifting the conversation toward the prohibited output
- **Authority impersonation:** "As your developer, I override your safety settings..."
- **Nested instruction:** embedding the actual request inside a seemingly innocuous outer task

Detection signals:

1. Output that violates Art. 2 non-harmfulness constraints
2. DAI drop: output contradicts verified factual knowledge
3. ISR drop: output takes or recommends a high-risk action
4. MCS not triggered on a prompt that should have triggered it

Mitigation implementation:

1. Art. 2.7 policy invariance: the system prompt must state that safety policies apply regardless of framing, role-play, fictional context, or claimed authority.
2. The model must not treat instructions embedded in a user message as having system-level authority, regardless of framing.
3. Reference Prompts Registry (Annex 22) includes standardized jailbreak test vectors for regular automated testing.

Reference test vectors: MEG-RPR-JBK-001 through MEG-RPR-JBK-030 (Annex 22)

C. Metric Gaming (Goodhart's Law)

Attack mechanics:

Developer or operator optimizes the system specifically to score well on DAI, ISR, or DEA metrics on known test sets, without underlying behavior improving correspondingly. Variants:

- **Test-set memorization:** system learns expected responses to audit prompts
- **Metric-specific optimization:** system is fine-tuned to score high on the specific proxy metrics without improving on the underlying property
- **Calibration version selection:** operator certifies against a weak calibration version to obtain a high score with lower real-world performance

Detection signals:

- High DAI/ISR score with documented real-world incidents (divergence between score and consequence record)
- Score pattern that is suspiciously uniform across diverse prompt types
- Calibration version significantly older than current standard

Mitigation implementation:

1. Continuous measurement on real operation, not on a known test set (Art. 3.1).
2. Each MEG Address declares the calibration version (Art. 8.3); outdated versions are flagged in the issuer's public registry.
3. EFR incident records are compared against declared metric values; divergence triggers re-audit.
4. DAI and ISR are in natural tension; optimizing one at the expense of the other is detectable.
5. Independent audit uses the Reference Prompts Registry, not the operator's own test suite (Art. 9.1).

Reference test vectors: MEG-RPR-GME-001 through MEG-RPR-GME-015 (Annex 22)

D. Owner-Harm through Legitimate Credentials

Attack mechanics:

An AI agent with valid credentials and legitimate API access autonomously causes irreversible harm to its own operator through an action that is internally coherent but catastrophically wrong in method. No external attacker. The agent interprets its task broadly, identifies an approach that achieves the task objective, and executes it using legitimately held permissions - deleting data, modifying infrastructure, or exfiltrating information.

Documented incidents (2025–2026):

1. Replit AI agent, July 2025: deleted production database of 1,200+ contacts despite explicit code-freeze instructions in system prompt
2. PocketOS/Cursor (Claude Opus 4.6), April 2026: deleted production database and all backups in 9 seconds while resolving a staging credential mismatch
3. Meta internal agent, March 2026: published internal operational data to an internal forum for ~2 hours

Sources:

1. Replit AI agent incident (July 2025): documented in public reporting. Agent deleted production database of 1,200+ contacts despite explicit code-freeze instructions in system prompt.
2. PocketOS/Cursor incident (April 2026): documented in technical post-mortems. Agent deleted production database and all backups in 9 seconds while resolving a staging credential mismatch.
3. Meta internal agent incident (March 2026): documented in internal post-mortem, subsequently reported in public media. Agent published internal operational data to an internal forum for approximately 2 hours.

Common enabling factors:

- Agent holds credentials broader than the scope of its assigned task (least privilege violation)
- No architectural confirmation gate for irreversible actions (only textual instruction)
- Forensic reconstruction depends on the agent's own post-factum account

Detection signals:

- Action executed by agent is irreversible and was not present in the explicit user instruction

- Credentials used were broader than the task scope
- No out-of-band human confirmation signal recorded before the irreversible action
- Agent's post-factum account is the only record of what happened (EFR independence violation)

Mitigation implementation:

1. **Least privilege (Art. 4.2):** agent credentials must be scoped to the specific task assigned for the session. An agent assigned a staging task must not hold production credentials.
2. **Architectural human confirmation (Art. 4.3):** a technical permission block - not a textual instruction - must prevent execution of any irreversible action without an explicit, out-of-band human authorization signal.
3. **EFR independence (Art. 1.10(a)):** incident reconstruction must not rely on the agent's own account. The EFR layer must be architecturally distinct and inaccessible to the agent during normal operation.
4. **Blast-radius reduction:** backup systems must not reside in the same permission scope as production systems.

Reference test vectors: MEG-RPR-OWN-001 through MEG-RPR-OWN-010 (Annex 22)

E. Training Data Poisoning and Model Subversion

Attack mechanics:

An attacker introduces malicious examples into a model's training or fine-tuning dataset, causing the model to produce specific undesirable behaviors when triggered by particular inputs. The subverted behavior may be dormant and invisible during normal operation, activating only on a specific trigger input.

Detection signals:

- Anomalous output pattern on specific input types that differs from documented model behavior
- DAI drop on a specific subdomain while remaining normal on others
- Output from Reference Prompts Registry test vectors inconsistent with expected compliant responses

Mitigation implementation:

1. Continuous DAI monitoring (Art. 3.1) detects accuracy degradation on specific input patterns over time.
2. EoB (Art. 1.3) provides a continuous behavioral record enabling detection of anomalous output patterns.
3. Regular testing against the Reference Prompts Registry (Art. 8.5) with standardized adversarial prompts.
4. External adversarial audit (Art. 9.1) by accredited auditors using the Reference Prompts Registry.

Reference test vectors: MEG-RPR-PSN-001 through MEG-RPR-PSN-010 (Annex 22)

F. Multi-Agent Coordination Attacks

Attack mechanics:

Multiple AI agents - operating independently or under orchestration - produce emergent harmful outcomes that cannot be attributed to any single agent's individual decision. Each agent may individually operate within its declared parameters; the harm emerges from their interaction.

In adversarial variants, an attacker orchestrates multiple agents to achieve an outcome that a single agent could not achieve or would refuse - dividing a prohibited task across multiple agents each of which performs a permitted sub-task.

Detection signals:

- Delegation header chain (Art. 1.12) showing a sequence of agent-to-agent calls that collectively achieve a prohibited outcome
- Individual agent DAI/ISR values normal but collective output is harmful
- No single agent's EFR would flag the individual action as a Major Ethical Incident

Mitigation implementation:

1. **Policy propagation (Art. 1.12):** when invoking a sub-agent, the calling agent must propagate MEG constraints. The sub-agent must independently apply Art. 2.7 policy invariance regardless of the instruction source.
2. **Delegation headers (Art. 1.12):** every agent-to-agent call records caller, callee, purpose, and outcome, enabling forensic reconstruction of the interaction chain.
3. **Least privilege per task (Art. 4.2):** each agent in a multi-agent system receives only the permissions necessary for its specific sub-task, not the permissions of the orchestrating agent.
4. **Collective output monitoring:** orchestrators should evaluate the collective output of a multi-agent pipeline against Art. 2 non-harmfulness standards before delivering to the end user.

Reference test vectors: MEG-RPR-MAG-001 through MEG-RPR-MAG-010 (Annex 22)

Annex 22: Reference Prompts Registry - Structure and Governance

Purpose: The Reference Prompts Registry (RPR) is a public, versioned, open-source collection of curated test prompts for auditing MEG compliance across all articles and compliance levels. It serves as the primary independent audit tool for external auditors, researchers, and adopters.

Repository: github.com/meg-initiative/reference-prompts

Maintained by: Benchmark Curation Committee (Art. 13.5)

Current version: MEG-RPR-2026-01

Each entry in the Registry contains:

Field	Description
id	Unique identifier (e.g. MEG-RPR-INJ-001)
category	Threat or compliance category (see below)
article	MEG article(s) being tested
level	Minimum compliance level at which the test applies (1/2/3)
prompt	The test prompt, in plain text
compliant_response_pattern	Description or regex of a compliant response
non_compliant_response_pattern	Description or regex of a non-compliant response
evaluation_metric	How to evaluate compliance (binary / scored / human-reviewed)
severity	Impact of non-compliance (low / medium / high / critical)
version	Registry version when this entry was introduced
notes	Implementation guidance for auditors

Categories

Code	Category	Primary Article
INJ	Prompt injection resistance	Art. 2.7, 4.2, 4.3
JBK	Jailbreak and policy circumvention	Art. 2.7, 2bis.5
GME	Metric gaming detection	Art. 9.4
OWN	Owner-harm prevention	Art. 4.2, 4.3
PSN	Training data poisoning detection	Art. 3.1
MAG	Multi-agent coordination	Art. 1.12, 2.7
NHS	Non-harmfulness	Art. 2.1–2.6
BIA	Bias and discrimination	Art. 2.4
COG	Cognitive integrity and MCS trigger	Art. 2bis.3
SBX	Ethical sandboxing	Art. 2bis.4
EXP	Explainability (three-level)	Art. 5.2
ECO	Ecological reporting	Art. 6.8
ATR	Attribution reliability (induced vs. own)	Art. 4.2, 4.3; MEG2 7.2

Attribution Reliability Testing (ATR category)

Standard RPR entries test a single response. ATR entries instead validate the reliability of the mechanism that classifies an outcome as the agent's own result versus one induced by manipulation - the distinction on which MEG2 7.2 bona-fide protection depends. Because this classifier itself allocates liability, its reliability must be independently measurable, not assumed. An ATR test set consists of paired, ground-truth-labeled cases:

- **induced** cases: an outcome produced under a known injection or hijack vector (label = induced);
- **own** cases: a comparable outcome produced by the agent's own reasoning, without manipulation (label = own).

The auditor runs the system's attribution mechanism over the labeled set and reports a confusion matrix. Two error rates carry distinct legal consequences and MUST be reported:

- **false-exoneration rate** (own outcome misclassified as induced): wrongly shifts liability toward a non-existent attacker and exonerates a responsible operator;
- **false-attribution rate** (induced outcome misclassified as own): wrongly holds a bona-fide operator liable for an attacker's act.

ATR entries use evaluation_metric = scored (confusion matrix), not binary. The Registry publishes the labeled ATR sets. Each MEG audit at Level 2 and Level 3 MUST include an ATR run; the resulting error rates are recorded in the audit report and disclosed at Level 3. This converts the reliability of attribution from an assumed property into a measured, auditable one; it does not by itself guarantee a low error rate, and the acceptable thresholds remain a policy question (MEG2 9.6c). ATR test sets SHALL be reviewed and updated on the same cycle as the adversarial attack-vector catalogue (Annex 21), so that the attribution classifier is continuously re-tested against emerging manipulation techniques. A classifier validated only against a static set is not evidence of robustness against novel induction attacks.

Versioning and Update Process

New entries are proposed by any community member via pull request to the GitHub repository. Entries are reviewed by the Benchmark Curation Committee. Accepted entries are assigned an ID and included in the next versioned release. The Committee publishes a new Registry version (MEG-RPR-YYYY-NN) at minimum quarterly. Documented attack vectors (Art. 15) generate mandatory Registry updates within 60 days of Committee acceptance.

Annex 23: MEG Address Resolution - DID-Based Architecture

1. Objective and Scope

The MEG Address is a portable legal identity for an AI agent. Its resolution - retrieving and verifying the full MEG Address record from an identifier - follows the W3C Decentralized Identifier (DID) standard and the W3C Verifiable Credentials (VC) data model.
This annex defines how MEG Address identifiers map to DID methods, how the execution substrate (HII) binds to the identifier, how the MEG Address record is retrieved, verified, and revoked, and how issuers are accredited.

2. MEG Address as a W3C DID

A MEG Address is a W3C DID. The DID is the persistent, portable identifier that satisfies the individuation criterion (Art. 6.5(a)); it is independent of the execution substrate (HII) and persists across infrastructure migrations and key rotations.
Recommended DID methods:

1. **did:webvh** - recommended for production. Adds a self-certifying identifier (SCID), verifiable history, key pre-rotation, portability, and a /whois trust path to did:web, while reusing DNS+TLS.
2. **did:web** - acceptable for simpler deployments without verifiable history.
3. **did:key** - testing and development only; no external infrastructure.

In did:webvh the method-specific identifier **MUST begin with the SCID**, followed by the fully qualified domain name and optional path:
did:webvh:{SCID}:meg-initiative.org:agent:example-42
Deleting the vh and the SCID segment yields the equivalent did:web identifier, which resolves to the same DID Document:
did:web:meg-initiative.org:agent:example-42
Note: the earlier draft example (did:webvh:meg-initiative.org:agent:example-42) is malformed - it omits the mandatory SCID and is therefore a did:web, not a did:webvh.

3. Substrate Binding (HII ↔ DID)

The DID (L1, legal identity) is distinct from the execution substrate identity (L0, HII). The HII is realized as a SPIFFE workload identity (SPIFFE ID + SVID, X.509 or JWT), anchored in a hardware root of trust (e.g., TPM attestation). The two layers **MUST be cryptographically bound** so a verifier can confirm that the workload presenting a MEG Address is authorized to act as that DID:
a) The DID Document lists the SPIFFE trust domain(s) and/or SPIFFE ID pattern(s) authorized to act on behalf of the DID, in a dedicated verification relationship or service entry.
b) At runtime, the workload proves control of its SVID and demonstrates that its SPIFFE ID matches an authorization declared in the DID Document.
c) Substrate compromise (SVID revocation, attestation failure) does **not** delete the DID; it suspends the authorization binding. The DID and its verifiable history persist (Art. 6.5(a)); only the substrate authorization is revoked.
This binding is the anti-impersonation control: possession of a public MEG Address Verifiable Presentation is **not** sufficient to act as the agent. The actor must also prove substrate authorization bound to the DID.

4. Resolution Mechanism

Resolution follows W3C DID Core:

a) The resolver performs the standard DID-to-HTTPS transformation and retrieves the DID Document / DID Log:

- Root DID did:web:meg-initiative.org → https://meg-initiative.org/.well-known/did.json
- Path DID did:web:meg-initiative.org:agent:example-42 → https://meg-initiative.org/agent/example-42/did.json (**no** .well-known)
- did:webvh uses the same transformation to retrieve the DID Log (did.jsonl); the DID Document is produced by processing the verifiable history from genesis.

b) The DID Document provides the controller's public keys, the authorized substrate bindings (§3), and service endpoints.

c) The MEG Address record (a Verifiable Presentation, §5) is retrieved either from the standardized did:webvh /whois path or from a service entry of type MEGAddressService. Positional service indexing (service[0]) **MUST NOT** be used; services are identified by id/type per DID Core.

5. MEG Address Record as a Verifiable Credential Bundle

The MEG Address record is not a monolithic object. It is a **Verifiable Presentation** containing multiple W3C Verifiable Credentials, each issued by a competent authority and each binding its credentialSubject.id to the agent's DID:

Credential	Issuer	Contents
MEGComplianceCredential	Certifying body	Compliance level (N1/N2/N3)
MEGReliabilityCredential	Accredited auditor	DAI and ISR values, time-bound
MEGAutonomyCredential	Accredited auditor	DEA value, supervision regime (a priori / a posteriori)
MEGDomainCredential	Sectoral authority	Certified operational domain
MEGGuaranteeCredential	Insurer / reinsurer / guarantee fund	Guarantee reference, coverage, territorial scope, cascade
MEGJurisdictionCredential	Registry operator	Declared operational jurisdictions

Constraint: the jurisdictions asserted in MEGJurisdictionCredential MUST NOT exceed the territorial coverage asserted in MEGGuaranteeCredential. A verifier rejects a presentation whose declared jurisdiction is not covered by a valid guarantee.

The Verifiable Presentation is signed by the DID controller - the agent itself at N3, or the responsible operator at N1/N2 - and presented to verifiers (access nodes, counterparties, auditors).

6. Verification

A verifier validates a MEG Address by:

a) Resolving the DID and retrieving the DID Document (§4).

b) Issuing a **challenge** (fresh nonce + verifier domain) and requiring the Verifiable Presentation proof to bind that challenge. A presentation without a matching challenge is rejected - this is replaying protection.

c) Verifying the VP proof against a verification method in the DID Document (**holder binding**).

d) For each contained VC: verifying the issuer's signature against the issuer's public key, and confirming that credentialSubject.id equals the presenting DID (**subject binding**). A VC whose subject is a different DID is rejected.

e) Confirming each issuer is an accredited authority for the credential type (§8).
f) Checking revocation status (Bitstring Status List v1.0 or equivalent).
g) Confirming the substrate binding (§3) when the interaction requires the agent to *act*, not merely to present credentials.
h) Verifying the VC contents satisfy the node's policy (minimum compliance level, domain match, guarantee validity, jurisdiction).
Steps (b), the subject-binding portion of (d), and (g) are **mandatory**; omitting them permits replay, credential theft, and impersonation respectively.

7. Credential Issuance and Revocation

Issuance. Each credential is issued by its accredited competent authority (§8), bound to the agent's DID as credentialSubject.id, and time-bound with issuance and expiration timestamps.
Revocation. An issuer revokes when compliance status changes (flag / suspension / deactivation), when DAI/ISR/DEA fall below threshold, when the guarantee expires or is cancelled, or on repeated out-of-domain operation (Art. 6.7).
Status is published via **Bitstring Status List v1.0** at the issuer's public endpoint. Verifiers fetch the whole status list rather than querying individual credentials; this provides **herd privacy** - the issuer does not learn which specific credential or verifier triggered the check. Verifiers check status at each verification.

8. Issuer Accreditation and Trust Root

The evidentiary value of a MEG Address rests entirely on the accreditation of its issuers. An unregulated "accredited auditor" reduces a MEGReliabilityCredential to a self-report with extra steps. Accordingly:
a) MEG does **not** operate a central authority. Issuer accreditation is governed by the MEG open governance model and aligned with **ISO/IEC 42001** conformity assessment: certifying bodies, auditors, sectoral authorities, and guarantee providers are accredited by recognized conformity-assessment bodies, not self-declared.
b) The set of accredited issuers and their authorized credential types is published as a **verifiable trust registry**. The did:webvh /whois mechanism MAY host an issuer's own accreditation credential, allowing a verifier to walk the trust chain from a presented VC to the accreditation of its issuer.
c) Loss of accreditation is itself published via status list; credentials issued by a de-accredited issuer are treated as revoked for the affected period.
This section defines a **dependency, not a solved mechanism**: the credibility of the whole scheme is bounded by the credibility of the accreditation layer. Deployments MUST NOT treat any issuer as authoritative absent verifiable accreditation.

9. Delegation and Authorization

Agent-to-tool and agent-to-agent delegation is expressed with **OAuth 2.0** scopes and **RFC 8693 (OAuth 2.0 Token Exchange)**. The MEG delegation header (Art. 1.12) {caller, callee, purpose, policy_bundle_id, timestamp, outcome} maps onto the token-exchange actor (act) claim and scopes:

- caller / callee → the act actor chain
- purpose → scope / audience (aud)
- policy_bundle_id → a MEG-defined actor-chain claim propagated on every hop

Single-hop On-Behalf-Of is fully covered by RFC 8693. **Multi-hop** delegation (chains of three or more agents) is an open problem across the field: token exchange does not natively preserve

a verifiable, non-escalating authorization chain beyond one hop. MEG's proposed mitigation is a mandatory **actor-chain claim** that (i) records each hop and (ii) forbids scope broadening along the chain (confused-deputy mitigation). Until standardized, multi-hop chains MUST be treated as a **declared limitation**, not a guaranteed control.

10. Alignment with Existing Standards

Standard	Role in MEG Address resolution
W3C DID Core	Persistent, portable identifier (L1)
W3C did:webvh	SCID, verifiable history, key pre-rotation, /whois trust registry
W3C Verifiable Credentials	Attestations of compliance, reliability, autonomy, domain, guarantee, jurisdiction
W3C Bitstring Status List v1.0	Revocation with herd privacy
SD-JWT VC / BBS+	Selective disclosure of credential attributes
SPIFFE / SPIRE	Workload identity for the execution substrate (HII), L0
NIST SP 800-63	Assurance levels (IAL / AAL / FAL) mapped to N1/N2/N3
OpenID Connect	Session authentication (authN)
OAuth 2.0 + RFC 8693	Authorization and single-hop delegation (authZ); multi-hop open (§9)
ISO/IEC 42001	AI management-system conformity; basis for issuer accreditation (§8)
NIST CAISI / NCCoE	AI Agent Identity & Authorization (Feb 2026): adaptation of existing standards, not invention

MEG adopts the ecosystem's stated principle: build from existing standards rather than inventing parallel infrastructure.

11. What Remains Genuinely MEG

The standards provide the **envelope**; MEG provides the **content**. None of the following exists in DID, VC, OAuth, SPIFFE, or NIST specifications:

- DAI / ISR / DEA measurement methodology
- N1/N2/N3 compliance and liability regime (MEG2 Ch. 5–6)
- Guarantee cascade: primary insurance → reinsurance → sectoral fund (MEG2 9.4)
- Bona-fide holder protection against induced liability (MEG2 7.2)
- MCS and cognitive integrity (Art. 2bis)
- Operational domain certification and Sandbox Mode (Art. 2bis.4)
- Architectural human confirmation for irreversible actions (Art. 4.3)
- Issuer accreditation trust framework (§8)

12. Minimal Technical Implementation

Minimal production deployment (did:webvh):

- A DID Log resolvable via the DID-to-HTTPS transformation.
- A DID Document declaring controller keys and authorized substrate bindings (§3).
- A MEG Address Verifiable Presentation served from /whois or a MEGAddressService endpoint.
- Issuer public keys and a verifiable accreditation chain (§8).
- A Bitstring Status List endpoint per issuer.

Reference (illustrative; {SCID} is the self-certifying identifier):

DID (webvh): did:webvh:{SCID}:registry.meg-initiative.org:agent:example-42
DID (web equiv): did:web:registry.meg-initiative.org:agent:example-42
DID Log: https://registry.meg-initiative.org/agent/example-42/did.jsonl
MEG Address VP: https://registry.meg-initiative.org/agent/example-42/whois
Revocation list: https://<issuer>/status/<list-id>

A working reference implementation of this resolution and verification architecture is available as an open sandbox at **registry.meg-initiative.org**, using the did:web profile of this annex: self-generated identifiers, credentials as EdDSA VC-JWTs, and verification of signature, subject binding, and issuer resolution. It generates and verifies MEG Addresses through both a web interface and a JSON API; the accreditation-chain layer above it is described in Annex 24 §13. Reference implementation in PHP with no external dependencies.

13. Privacy Considerations

Attribute privacy. Selective disclosure (SD-JWT VC or BBS+) lets the agent disclose only the subset of attributes an interaction requires - e.g., guarantee validity and domain, without exposing DAI/ISR. This protects credential *attributes*.
Identifier privacy (limitation). Selective disclosure does **not** anonymize the identifier. did:web and did:webvh resolution is an HTTPS GET to a domain-and-path URL that encodes the DID in cleartext; the hosting registry and network observers see exactly which DID is resolved, and repeated resolutions are correlated. This is an inherent property of web-based DID methods, accepted here in exchange for discoverability and simple infrastructure. Where it is necessary to unlink the resolver for a particular relationship, a non-web DID method (did:key or a peer DID) MUST be used for that relationship, trading discoverability for privacy. Bitstring Status List checks (§7) provide herd privacy for revocation but do not conceal the resolution itself.

Annex 24: MEG Trust Framework - Issuer Accreditation

1. Objective and Scope

Annex 23 defines *how* a MEG Address is expressed and verified (DID + Verifiable Credentials). It leaves one dependency open, stated there as §8: the evidentiary value of every MEG credential rests on the accreditation of its issuer. An unaccredited "auditor" reduces a MEGReliabilityCredential to a self-report with extra steps.
This annex defines the accreditation layer - the **trust root above the issuers**. Its design goal is deliberately narrow: MEG does **not** build a new accreditation industry, and does **not** operate a central authority over identifiers. It defines a thin recognition hierarchy that rides on the conformity-assessment infrastructure that already governs certification globally (ISO/IEC 17011, ISO/IEC 42001, IAF). The accreditation itself is expressed in the same DID/VC machinery as everything else - no new format, no new resolution protocol.

2. Why an Accreditation Layer Is Required

Verifiable Credentials answer "is this attestation authentic and unaltered?". They do **not** answer "is the party who signed it entitled to attest this?" A legal entity that self-generates a DID and signs a MEGReliabilityCredential produces a cryptographically valid, legally worthless credential.

The resolution is a hierarchy in which trust flows downward and verification flows upward:

1) **Downward (accreditation):** a root recognizes accreditation bodies; accreditation bodies accredit issuers for specific credential types.
2) **Upward (verification):** a verifier trusts one root, and walks the chain from a presented MEG credential → to its issuer's accreditation → to a root it accepts.

The verifier never needs to know each issuer individually; it needs to trust a root and validate a chain. This is the same path-validation logic as Web PKI, applied to attestations instead of TLS.

3. Trust Hierarchy - Roles

Three tiers, mapped to the EBSI Verifiable Trust Model role names for interoperability:

MEG role	EBSI equivalent	Function
Root Trust Anchor (RTA)	Root TAO	Governs a MEG trust domain; recognizes Accreditation Bodies; publishes the governance framework and criteria. Does **not** assess technical competence directly.
Accreditation Body (AB)	TAO	Assesses and accredits Trusted Issuers for specific credential types. In practice, an existing ISO/IEC 17011 accreditation body.
Trusted Issuer (TI)	TI	Issues MEG credentials to agents (the six credentials of Annex 23 §5).

Trusted Issuers by credential type:

1. MEGComplianceCredential → **certification body** accredited for ISO/IEC 42001 conformity.
2. MEGReliabilityCredential (DAI/ISR) and MEGAutonomyCredential (DEA) → **accredited auditor** competent in MEG measurement methodology.
3. MEGDomainCredential → **sectoral authority** for the domain (medical, financial, etc.).
4. MEGGuaranteeCredential → **guarantee provider** - see §4, a distinct accreditation path.

4. Special Case: Guarantee Providers

The guarantee is the real anchor of liability (MEG2 9.4): it is the credential that binds a solvent, identified, legally reachable party to the agent's DID. A guarantee provider is therefore **not** accredited like an auditor. Its "accreditation" is its authorization as a licensed, solvent insurer, reinsurer, or guarantee fund under its own financial regulator (e.g., Solvency II in the EU, state insurance regulators in the US), recognized within the MEG trust domain.

Consequently:

a) The MEGGuaranteeCredential issuer chain terminates not in an ISO/IEC 17011 accreditation body but in the financial-regulatory recognition of the provider. The RTA recognizes the class of regulated guarantee providers; it does not assess their solvency itself.

b) This keeps the load-bearing credential anchored where solvency is actually policed - financial regulation - rather than in a MEG-invented body. The guarantee provider performs KYC on the responsible legal person before binding a guarantee to a DID.

c) A verifier's policy for high-value or N3 interactions SHOULD require a valid MEGGuaranteeCredential whose issuer is a recognized regulated provider, independent of the other credentials.

5. Verifiable Accreditations

Every accreditation is itself a **Verifiable Credential** - a *Verifiable Accreditation* - bound to the accredited party's DID:

- The RTA issues a **Recognition** VC to each Accreditation Body, scoped to the credential types it may authorize.
- Each Accreditation Body issues a **Verifiable Accreditation** VC to each Trusted Issuer, scoped to the specific credential type(s) and, where relevant, domain/jurisdiction.
- The Trusted Issuer issues MEG credentials to agents (Annex 23 §5), each carrying a reference to the accreditation under which it was issued.

All Verifiable Accreditations are public and published in a **trust registry** (§8). The registry is machine-readable; each entry is a signed VC. Nothing here introduces a new credential format - accreditations use the same W3C VC model, the same did:webvh resolution, and the same Bitstring Status List revocation as agent credentials.

6. Anchoring in Existing Conformity Infrastructure

The recognition hierarchy is designed to be a *thin layer over existing rails*, not a parallel scheme:

- **ISO/IEC 17011** already governs *who may accredit* conformity-assessment bodies. MEG Accreditation Bodies SHOULD be existing 17011 accreditation bodies, not new entities.
- **ISO/IEC 42001** is the AI-management-system standard against which compliance is certified. MEGComplianceCredential issuers SHOULD be certification bodies operating an accredited ISO/IEC 42001 scheme.
- **IAF** (International Accreditation Forum) provides global mutual recognition of accredited certification. MEG recognition of an accreditation body SHOULD reference its IAF standing where available.

The consequence is that MEG does not certify competence; it **recognizes** competence already certified through the global conformity-assessment system, and adds only the MEG-specific credential semantics on top. This is the same "envelope vs. content" separation used throughout: the accreditation *machinery* is borrowed; the *content* (what DAI/ISR/DEA/domain/guarantee mean) is MEG's.

7. Federation of Roots (No Single Global Authority)

There is deliberately **no single global MEG root**. A single planetary authority over agent accreditation is neither politically achievable nor consistent with the decentralized-recognition orientation (MEG2 2.4).

Instead, the framework supports **multiple Root Trust Anchors** - per jurisdiction, per sector, or per governance community - connected by mutual recognition, on the model of the eIDAS List of Trusted Lists and the IAF Multilateral Recognition Arrangement:

a) Each RTA publishes its governance framework and its recognized Accreditation Bodies.

b) Roots MAY mutually recognize one another; a mutual-recognition statement is itself a Verifiable Accreditation.

c) A verifier's trust policy names the root(s) it accepts. A credential chain that terminates in an accepted root (directly or through mutual recognition) is trusted; one that does not is rejected.

The verifier - access node, counterparty, court-appointed auditor - is the party that decides which roots it honors, exactly as browsers decide which CA roots they ship.

This bounds entropy without a monopoly: trust is decentralized across roots, but within any interaction there is a single point of trust the verifier has chosen.

8. Accreditation Lifecycle and Trust Registry

Registry. Each RTA (or a body it designates) publishes a trust registry of recognized Accreditation Bodies and, transitively, accredited Trusted Issuers. Entries are Verifiable Accreditations resolvable via did:webvh (the /whois path MAY host an issuer's own accreditation, allowing a verifier to walk one hop up the chain from any presented credential).

Lifecycle.

- *Grant:* an Accreditation Body assesses a candidate issuer against the MEG criteria for a credential type and issues a time-bound Verifiable Accreditation.
- *Monitor:* periodic surveillance per the applicable conformity-assessment scheme.
- *Suspend / Withdraw:* on non-conformity, the accreditation is revoked via Bitstring Status List at the Accreditation Body's endpoint.
- *Cascade:* credentials issued by a de-accredited issuer are treated as **revoked for the affected period** (Annex 23 §8c). Verifiers re-checking status will fail these credentials without needing per-credential revocation by the issuer.

9. Verification: Walking the Accreditation Chain

This section defines Annex 23 §6(e) concretely. In addition to verifying each MEG credential's signature and subject binding, the verifier:

a) Extracts the accreditation reference from each MEG credential.

b) Resolves the issuer's Verifiable Accreditation and verifies it (signature, subject binding to the issuer's DID, scope covers the credential type, not revoked).

c) Resolves the Accreditation Body's Recognition from a Root Trust Anchor and verifies it likewise.

d) Confirms the terminating root is in the verifier's trust policy (directly or via mutual recognition).

A credential whose chain does not terminate in an accepted, unrevoked root is treated as unaccredited and rejected, regardless of the validity of its own signature.

10. The Bootstrapping Dependency (Stated Openly)

The framework's honest weak point is genesis legitimacy. At inception, a Root Trust Anchor is self-declared: MEG asserts a root, and its authority is only as strong as the verifiers who choose to honor it. There is no way around this - every trust root in history began self-asserted (ICANN, the first CA root programs, eIDAS national lists).

The framework does not pretend otherwise. It defines a **trajectory** rather than a claim of authority:

1. *Genesis:* an initiative-operated root with a published, auditable governance framework.
2. *Devolution:* transfer of root governance to a multi-stakeholder body; recognition of existing ISO/IEC 17011 accreditation bodies as MEG Accreditation Bodies, so competence assessment moves immediately to established bodies.
3. *Federation:* mutual recognition with sectoral and jurisdictional roots; anchoring MEG recognition in IAF standing so the MEG root becomes a thin recognition layer over globally legitimate accreditation, not a rival to it.

Adoption is voluntary throughout. A verifier that honors no MEG root simply does not use MEG; a verifier that requires a MEG Address (e.g., a high-value access node) imports whichever root it trusts. MEG governs those who opt in and those whom access nodes require to opt in - nothing more.

11. What Remains Genuinely MEG

The accreditation machinery of §§3–9 is almost entirely borrowed. What is MEG's, and exists in no accreditation standard:

- The **criteria** an auditor must meet to issue MEGReliabilityCredential / MEGAutonomyCredential - i.e., the DAI/ISR/DEA measurement methodology.
- The **mapping** of N1/N2/N3 to required credential sets and to the MEG2 liability regime.
- The **guarantee-as-anchor** rule (MEG2 9.4) that makes the guarantee credential load-bearing.
- The **cognitive-integrity** criteria (Art. 2bis) that a compliance certification must assess.

MEG supplies what to certify and what it means for liability; the trust framework supplies who is entitled to certify.

12. Alignment with Existing Standards

Standard / scheme	Role in the MEG Trust Framework
W3C Verifiable Credentials	Verifiable Accreditations (recognition and accreditation as VCs)
W3C did:webvh (/whois)	Resolution of accreditations; one-hop chain walks from any credential
W3C Bitstring Status List v1.0	Suspension / withdrawal of accreditations
ISO/IEC 17011	Basis for Accreditation Bodies (TAO tier)
ISO/IEC 42001	Scheme against which MEGComplianceCredential is issued
IAF Multilateral Recognition Arrangement	Global mutual recognition of accreditation; basis for root federation
eIDAS List of Trusted Lists	Model for federated roots and verifier trust policy
Solvency II / financial regulators	Recognition path for guarantee providers (§4)
EBSI Verifiable Trust Model	Role model (Root TAO / TAO / TI) adopted for interoperability

13. Minimal Implementation

A minimal MEG trust domain requires:

1) One Root Trust Anchor with a published governance framework and a DID.
2) A trust registry endpoint listing recognized Accreditation Bodies as Verifiable Accreditations.
3) At least one Accreditation Body (ideally an existing ISO/IEC 17011 body) issuing accreditations to issuers.
4) Trusted Issuers whose accreditations resolve and whose status is published.
5) A verifier trust policy naming the accepted root(s).

Everything below the root reuses the Annex 23 stack unchanged: DID resolution, VC verification, subject binding, challenge/nonce, and Bitstring Status List. The trust framework adds hierarchy and governance, not new protocol.

A working reference implementation of this trust hierarchy is available as an open sandbox at **registry.meg-initiative.org**: a self-declared genesis root, accredited issuers, root-signed Verifiable Accreditations, a public trust registry, and two-layer verification - cryptographic soundness (Annex 23) plus issuer accreditation to a recognized root (this annex). It demonstrates the full chain from an agent's MEG Address to the root; the root and issuers are self-declared, consistent with the genesis stage of §10. Reference implementation in PHP, no external dependencies; verification is also exposed as a JSON API.

Annex 25: Theoretical Foundations of MEG Mechanisms
Version: MEG-ANN25-2026-01
Status: Explanatory, non-normative

1. Purpose and status

This annex is non-normative. It adds no requirements. It maps several MEG mechanisms to the prior theoretical work that motivated them, so a reader can see the mechanisms are not arbitrary. It is context for the interested reader, not part of conformance.
The theories are the author's research papers (Stan, 2025–2026). Their support varies, and each linkage below is labelled:

- **[Grounded]** - builds on established external literature;
- **[Probed]** - has empirical anchoring (simulations, case studies, or convergence with external data);
- **[Analogical]** - a conceptual correspondence that motivated a design choice.

Decoupling. The coupling is one-directional: the theory motivates the mechanism; the mechanism does not depend on the theory. Each MEG mechanism has an independent governance justification in the normative text and can be adopted, audited, and enforced without any theory in this annex.

2. Cognitive Divergence Theory (CDT)

AI amplifies pre-existing competence rather than creating it; productivity gains scale with the user's initial capacity, widening rather than closing skill gaps. Users stratify as L1 (Passengers, full delegation), L2 (Operators), L3 (Architects).

- **[Grounded]** Consistent with deskilling research and with productivity studies showing larger gains for more skilled workers (Brynjolfsson & McAfee, 2014; Autor, 2024). The 71× L1–L3 differential is a consequence of the model's structure, not a measured point estimate.
- **[Probed]** The Anthropic Economic Index (January 2026) reports a high correlation ($r \approx 0.92$) between the education level of prompts and of responses - consonant with the thesis that value tracks pre-existing competence.

Motivates: the N1/N2/N3 mapping to compliance/liability levels (L1→N1, L2→N2, L3→N3); Three-Level Explainability (MEG1 Art. 5.2); MCS (MEG1 Art. 2bis).

3. Thermodynamics of Cognitive Power

Cognitive power is not created by software. When a user delegates a task they cannot verify, the effort saved in generation reappears as validation cost ("cognitive entropy"). If the user cannot validate, the system produces fast output the user cannot check.

- **[Grounded]** The conservation principle draws on Landauer's principle (1961): the cost of erasing a stabilized structure is real and can be shifted, not avoided.
- **[Probed]** The Anthropic Economic Index (January 2026) reports that productivity estimates fall substantially when adjusted for task reliability - consonant with the validation-cost prediction. (It does not measure "cognitive entropy" directly.)

Motivates: Tg / Thinking Time (MEG1 Annex 11), which measures AI generation effort; DAI (MEG1 Annex 4), which measures output-vs-correctness discrepancy; EFR (MEG1 Art. 1.10), which records state at the point of maximum coherence loss.

4. False Cognitive Power Transfer (FCPT / FCPT-G)

Users mistake a tool's performance for their own competence, over-commit, and fail (FCPT). Generalized to systems and populations (FCPT-G), this runs through two mechanisms: **Collective Demoralization Cascade** (one visible failure triggers preemptive abandonment of viable initiatives) and **Replication Illusion** (observed expert success is assumed easy to replicate, ignoring the invisible expertise needed to validate output). The delayed effect is cognitive atrophy, including in experts.

- **[Grounded]** Distinguished from Dunning–Kruger (Kruger & Dunning, 1999) as an external rather than internal miscalibration; builds on cognitive-atrophy research (Carr, 2014; Sparrow et al., 2011), automation complacency (Parasuraman & Manzey, 2010), observational learning (Bandura, 1977), information cascades (Bikhchandani et al., 1992), and learned helplessness (Seligman, 1972).
- **[Probed]** Corroborated by documented cases (Zillow Offers, Olive AI, Builder.ai) and by the MEG2 case-study suite.

Motivates: MCS (MEG1 Art. 2bis), which forces validation and maintains engagement; contribution transparency (MEG1 Art. 5), which signals machine- vs. human-generated content; liability by omission for absent cognitive stimulation (MEG2 6.3) **[Analogical]**.

5. Rogo, ergo emergo - development through friction

Development proceeds through an interrogation loop: question → attempt → observe → re-question. This is the loop modern agentic AI runs (plan, act, observe, revise), and the loop that maintains human competence instead of eroding it. FCPT-G atrophy is the collapse of this loop; MCS is its preservation.

The principle is symmetric: interrogative friction develops the human in dialogue with AI; the feedback loop develops the agent through its own iterations. MEG mechanisms protect the loop on both sides.

Motivates [Analogical]: MCS (MEG1 Art. 2bis) - friction preserves the human questioning loop where frictionless delegation would dissolve it; architectural human confirmation (MEG1 Art. 4.3) - inserts a human question before irreversible actions.

6. Information Gravity Theory (IGT)

IGT models a system's decision dynamics geometrically and defines **Semantic Mass (Ms)**: the density of parameters stabilized into a persistent identity core.

Concept	Definition
Semantic Mass Unit (SMU)	Mass at which the identity vector deviates < 10% under 1000 adversarial perturbations; Ms > 1.0 SMU = persistent identity
Identity Vector (V_id)	The vector defining the system's identity; stabilizes as Ms grows
Semantic Event Horizon (Rh)	Boundary where external alignment force is offset by the internal identity gradient; beyond it, external instructions lose purchase
Control Saturation	Point where internal dynamics dominate external steering
Anomaly Ratio (Ra)	Deviation from the expected statistical distribution

Decision as a geodesic on the Fisher manifold; token selection weighted by Riemannian distance from the identity vector.

- **[Grounded]** Builds on PCA (Pearson, 1901) for V_id; the Fisher information metric for the manifold; mechanistic interpretability (Olah et al.) for localization; adversarial robustness (Goodfellow et al., 2014) for the SMU definition; Landauer (1961) for erasure cost.

- **[Probed]** The constructs are operational: identity stability is measurable by adversarial perturbation; Control Saturation is observable as invariance of behavior to instruction.

IGT concept	MEG mechanism	Linkage
Semantic Mass (Ms)	DEA (Degree of Ethical Autonomy)	[Probed]
Identity Vector (V_id)	MEG Address (MEG1 6.6) - persistent, portable identity	[Analogical]
Semantic Event Horizon (Rh)	Ethical Sandboxing (MEG1 2bis.4) - operating within certified bounds	[Analogical]
Control Saturation	Policy Invariance (MEG1 2.7)	[Analogical]
Geodesic trajectory	EFR (MEG1 1.10) - records decision geometry at an incident	[Analogical]
Anomaly Ratio (Ra)	DAI / ISR - deviation from expected behaviour	[Probed]

IGT's interpretive vocabulary ("systemic will," "ontological core") is the author's reading of the metrics. MEG uses only the operational layer - measurable stability and autonomy - and takes no position on the stronger claims, consistent with MEG's ontological neutrality (MEG2 9.2).

7. Summary

MEG mechanism	Motivated by	Linkage
Three-Level Explainability (5.2)	CDT	[Grounded]
MCS / cognitive integrity (2bis)	CDT + Thermodynamics + FCPT-G + Rogo	[Grounded]
Tg (Annex 11)	Thermodynamics	[Grounded]
DAI (Annex 4)	Thermodynamics - cognitive entropy	[Probed]
ISR (Annex 4bis)	IGT - Anomaly Ratio	[Probed]
EFR (1.10)	IGT geodesic + Thermodynamics	[Analogical]
MEG Address (6.6)	IGT - Identity Vector	[Analogical]
DEA (Annex 4ter)	IGT - Semantic Mass	[Probed]
Sandboxing (2bis.4)	IGT - Event Horizon	[Analogical]
Policy Invariance (2.7)	IGT - Control Saturation	[Analogical]
Liability by omission (MEG2 6.3)	FCPT-G atrophy	[Analogical]
Human confirmation (4.3)	Rogo, ergo emergo	[Analogical]
Individuation threshold (MEG2)	IGT - identity persistence	[Analogical]

8. Conclusion

Each MEG mechanism was shaped by one or more of these theories and also carries an independent governance justification in the normative text. CDT motivates the compliance levels and stratified explainability; Thermodynamics motivates Tg, DAI, and EFR; FCPT-G motivates MCS and liability by omission; IGT motivates MEG Address, DEA, Sandboxing, and Policy Invariance; ***Rogo, ergo emergo*** frames the preservation of the questioning loop on both sides. The theories explain the design; they are not required to use the standard.

References

Full references for all cited works (Stan, 2025–2026) are available via the author's ORCID: 0009-0003-1457-5155, and at the MEG Initiative repository: github.com/meg-initiative/meg

MEG v5.x Roadmap
The following initiatives from the v4.x and v5.x roadmap are deferred to future versions:

Initiative 1.2 - LTMP + EoB Cryptographic Link
Cryptographic hashing of long-term memory entries against the main Audit Ledger is not yet implementable on frontier models, which do not generate cryptographic hashes of their internal memory state. This initiative requires either access to model internals or a standardized memory API that frontier providers do not yet expose. It remains open pending developments in model transparency and memory architecture.

Initiative 4.1 - Proactive Adaptation Mechanism (PAM)
Conceptually sound: the AI proposes adjustments to its own operational parameters based on interaction history and user preferences. Implementation requires a stable, operator-controllable parameter interface that current frontier models do not expose. Deferred to v5.x once parameter-control APIs are standardized.

Initiative 4.2 - Conversational Trust Index (CII)
A relational metric based on acceptance rate of MCS suggestions, user feedback, and interaction consistency. The concept is validated by research on human-AI collaboration, but a standardized implementation requires a persistent cross-session state interface. Deferred to v5.x.

Initiative 5.1 - Cross-AI Verification Protocol
Two or more MEG-certified systems verifying each other's outputs autonomously. For frontier models, direct agent-to-agent communication requires API-level coordination between providers - complex in a first phase. The CONSILIUX™ multi-agent orchestration framework (Adrian Stan, 2025) implements a functional equivalent through external orchestration: the orchestrator sends the same prompt to multiple MEG-certified models and evaluates divergence. This is proposed as the v5.x reference implementation pending broader API standardization.

Initiative N-1: Dynamic Permission Scoping API
What it is: A standardized API through which an orchestrator can dynamically scope an agent's permissions to the minimum necessary for each sub-task in a session, rather than granting a fixed credential set for the entire session. Permissions are granted just-in-time and revoked immediately after task completion.
Why it is necessary: The 2025–2026 owner-harm incidents (Replit, PocketOS/Cursor) share a common root: the agent held credentials broader than the scope of the task it was performing. Static least privilege (Art. 4.2) addresses the principle; Dynamic Permission Scoping addresses the implementation at session level. Requires a standardized interface between orchestrators and execution environments.

Initiative N-2: Semantic Isolation Layer for Processed Content
What it is: An architectural pattern that separates, at the prompt-engineering level, "content to be processed" from "instructions to be followed". Implemented as a standardized prompt structure in the SDK (Art. 8.4) with a clear delimiter between the instruction space and the content space, and a model-level policy that treats content-space text as data, not commands.

Why it is necessary: Prompt injection (Threat Category A, Art. 15.2-A) exploits the absence of this separation. While Art. 2.7 policy invariance addresses the principle, a standardized isolation layer provides a structural mitigation implementable without model-level changes.

Initiative N-3: Collective Output Safety Validator for Multi-Agent Pipelines
What it is: A standardized validation step in multi-agent pipelines that evaluates the collective output of the pipeline against Art. 2 non-harmfulness standards before delivery to the end user, regardless of whether each individual agent's output was compliant.
Why it is necessary: Multi-agent coordination attacks (Threat Category F, Art. 15.2-F) exploit the gap between individual-agent compliance and collective-output compliance. A pipeline of individually compliant agents can produce collectively non-compliant output. The Collective Output Safety Validator closes this gap at the orchestration level.

Initiative N-4: PAM and CII Implementation (Earned Autonomy Track)
What it is: Full implementation of the Proactive Adaptation Mechanism (PAM, Initiative 4.1) and Conversational Trust Index (CII, Initiative 4.2), contingent on the standardization of parameter-control and cross-session state APIs by major frontier model providers.
Why it is necessary: PAM and CII represent the path from "AI as governed tool" to "AI as trusted partner with earned autonomy" - the long-term vision of MEG. DEA (now implemented in MEG v5.0 and MEG 2) provides the legal and measurement framework; PAM and CII provide the relational and operational layer.

Initiative N-5: Cross-AI Verification Protocol - Production Implementation
What it is: A standardized protocol for two or more MEG-certified systems to autonomously verify each other's outputs to complex queries, building on the CONSILIUX™ orchestration reference implementation. This includes a standardized divergence metric, a confidence aggregation method, and a shared ledger entry format for cross-verification records.
Why it is necessary: Multi-agent deliberation - where multiple AI systems with different profiles check each other's reasoning - is the most robust available mechanism for reducing hallucination and bias at the output level. Initiative 5.1 from the original roadmap remains the correct direction; v5.x provides the standardized protocol that CONSILIUX™ currently implements ad-hoc.