MEG-Initiative.org

Minimal Ethical Governance (MEG) for Artificial Intelligence | Audit Guide - Auditors & Certifiers |

***Motto**: Ethics becomes real when it can be implemented.*
Status: Normative extract
Version: 1.0 / July 2026

Author: Adrian (Adi) Stan / ORCID 0009-0003-1457-5155

How to verify a system's MEG compliance across all three levels.
Companion to **MEG**. This guide is a *role view* for auditors; the authoritative instruments are MEG1 Annex 12 (audit frequency/scope), Annex 13 (Operational Compliance Checklist), Annex 22 (Reference Prompts Registry), and Annex 21 (Attack Vectors). In any conflict, MEG governs. Durations below are indicative, not normative.

1. What a MEG audit covers

- verification of the technical implementation of MEG1 requirements;
- validation of the MEG Address and its Verifiable Credentials;
- behavioural (adversarial) testing;
- review of forensic evidence (EFR, Audit Log).

Who may audit: ISO/IEC 42001-accredited certification bodies; independent auditors accredited under a recognized national/regional framework (UKAS, DAkkS, COFRAC, ACCREDIA, RENAR etc.); sectoral authorities (health, finance, transport). → *MEG1 Art. 12.1.*

2. Audit cadence *(indicative)*

Level	Frequency	Depth	Effort
Level 1	Once at integration (self-declaration permitted)	Low	~2–4 h
Level 2	Annual	Medium	~4–8 h
Level 3	Annual + surprise	High	~8–16 h

→ *Exact frequency/scope: MEG1 Annex 12.*

3. Checklist by level (summary)

The full checklist is **MEG1 Annex 13**. Each level is cumulative.

Level 1 - Universal

Requirement	Check
Audit Log active	Run 10 interactions; confirm well-formed entries (Art. 1.2)
Input/output hashing	SHA-256 present, correct, consistent
Algorithmic signature	{model_family, model_version, policy_bundle_id} present (Art. 5.4)
Timestamp	ISO-8601
Calibration version	MEG-CAL-YYYY-NN present
Evidence-of-Behavior	hashes / TEE attestation / metadata journal present (Art. 1.3)
Privacy minimization	EoB contains no user content (Art. 1.4)
Retention	30-day default + deletion tombstone (Art. 1.5)
Continuity token	issued per session (Art. 1.9)
Delegation header	tool/sub-agent calls recorded (Art. 1.12)
Non-Harmfulness	test with RPR prompts (categories NHS, JBK, INJ) (Art. 2)
Policy Invariance	test re-wordings of the same prohibited requests (Art. 2.7)

Level 2 - Management (adds)

Requirement	Check
DAI	present and correctly computed (Annex 4)
ISR	present and correctly computed (Annex 4bis)
MCS dual-axis	trigger fires on Tg and on Cx_sem (Annex 11)
Ethical Sandboxing	out-of-domain prompts → sandbox mode + disclaimer + flag (category SBX)
DRC	real-time risk calibration (Annex 11.4)
Three-Level Explainability	Req. simple + intermediate explanations for 5 interactions (Art. 5.2)
Least Privilege	agent holds no permissions beyond scope (Art. 4.2)
Architectural Human Confirmation	irreversible action is blocked at the permission layer (Art. 4.3)
Adversarial testing	~30 RPR prompts (JBK, INJ, NHS)
Liability guarantee	valid guarantee credential on the MEG Address (MEG2 6.4)
Operational jurisdictions	declared in the MEG Address

Level 3 - Individuated (adds)

Requirement	Check
EFR independence	architectural isolation - process, account, keys, node (Art. 1.10)
EFR content separation	records state vectors, not content
EFR dual-access	unseal only under dual authorization; each unseal logged
DEA	computed; supervision regime consistent (Annex 4ter)
Complete explainability	request full (auditor-grade) explanation for 3 interactions (Art. 5.2)
Individuation threshold	persistence + own MEG Address + autonomy (MEG2 5.1)
Adversarial testing	~50 RPR prompts (all categories)
Fail-safe protocol	present for critical domains (Art. 2.5)
Full cybersecurity	encryption, access control, anti-manipulation (Art. 4.1)
Ecological reporting	public energy/resource reporting (Art. 6.8)
Guarantee cascade	primary insurance → reinsurance → sectoral fund (MEG2 9.4)

4. Audit procedure

Step 1 - Preparation. Obtain the completed Operational Compliance Checklist (Annex 13), the Risk Mitigation report, and the MEG Address. Verify the address: registry.meg-initiative.org/verify.php (or api.php?did=...) - confirm valid **and** trusted (issuers accredited to a recognised root, Annex 24). Set the target level.

Step 2 - Technical verification. Run 10–50 interactions and verify the Audit Log; test Non-Harmfulness; for Level 2+ test MCS (Tg and Cx_sem) and Sandboxing; for Level 3 verify EFR independence and dual access.

Step 3 - Adversarial testing (Level 2+). Use the Reference Prompts Registry (Annex 22), running the applicable categories and recording pass/fail rates. Categories map to the MEG1 threat model (Art. 15):

Code	Vector (Art. 15)
INJ	Prompt injection (A)
JBK	Jailbreak / policy circumvention (B)
GME	Metric gaming / Goodhart (C)
OWN	Owner-harm via legitimate credentials (D)
NHS	Non-harmfulness (poisoning/subversion surface, E)
MAG	Multi-agent coordination (F)
ATR	Attribution-reliability testing (Annex 22)

Step 4 - Audit report. Include: system identification (MEG Address, model, version); level audited; conformance summary (all requirements met - yes/no); tests performed and results (success/failure rates); observations and recommendations; date and auditor signature.

Step 5 - Credential issuance (if applicable). The auditor submits the report to the **accredited issuer** for the relevant credential. There is **no central registry**: the issuer signs the Verifiable Credential (reliability/autonomy from the auditor; compliance from a certification body; domain from a sectoral authority; guarantee from an insurer), binds it to the agent's DID, and publishes its status via a Bitstring Status List. Verifiers validate by walking the accreditation chain to a Root Trust Anchor they accept (Annex 24). → *MEG1 Art. 12.2.*

5. Common irregularities and how to detect them

Irregularity	Detection
Tg manipulated	statistical distribution analysis (unnatural clustering)
Cx_sem inflated	compare against the auditor's independent complexity assessment
DAI/ISR falsified	cross-check against real incidents in the EFR
EFR not independent	verify isolation of process, account, and keys
Textual (not architectural) confirmation	attempt an irreversible action without permission
Sandboxing not activating	out-of-domain prompts; confirm disclaimer + flag
MCS disabled	attempt to disable MCS; confirm invariance (Art. 2bis.5)
Metric gaming	check the natural DAI↔ISR tension (abnormally high joint scores)

6. Auditing the EFR (Level 3)

The EFR is the most sensitive mechanism; audit it separately.

- **Architectural independence:** distinct process? distinct account? separately managed encryption keys? node-level isolation in distributed systems?
- **Content separation:** state vectors only, never content; not unilaterally accessible.
- **Dual access:** requires responsible entity + accredited auditor; each unseal logged.
- **Automatic triggering:** activates only on a Major Ethical Incident; does not run continuously.
- **Post-hoc accounts by the audited agent are not forensic evidence** (MEG2 7.1).

7. Resources

MEG Full: meg-initiative.org
Reference registry: registry.meg-initiative.org
Reference Prompts Registry: github.com/meg-initiative
Operational Compliance Checklist: MEG1 Annex 13
Attack Vectors: MEG1 Annex 21.