W3C AI Knowledge Representation Community Group (AIKR CG)

Response to Stakeholder Consultation: Draft Commission Guidelines on the Classification of High-Risk AI Systems under Article 6 of Regulation (EU) 2024/1689 (AI Act)

Consultation reference: https://digital-strategy.ec.europa.eu/en/library/draft-commission-guidelines-classification-high-risk-ai-systems Submitted by: W3C AI Knowledge Representation Community Group Consultation deadline: 23 July 2026

Prepared by the W3C AI Knowledge Representation Community Group. The reasoning in Section 2 was developed in discussion with the W3C AI Agent Protocol Community Group and the W3C A2WF Community Group, and on the CG's own public mailing list (https://lists.w3.org/Archives/Public/public-aikr/). Comments on this submission are welcome via the CG mailing list or by requesting edit access from the Chair

THE CONSULTATION FORM HAS BEEN SUBMITTED WITH THE FOLLOWING ENTRIES AND A LINK TO THIS DOC

Question 1 — Which parts would you like to comment on?

Select: ☑ General principles (Section II) and ☑ Annex III (Section IV).

unchecked: Annex I (Section III), and Article 112(1) Evaluation/Review — reasoning below.

Section II — General principles

Select subsection: ☑ II.2 Intended purpose(s) of the AI system (II.1 unchecked — our contribution doesn't touch the AI-system definition itself)

II.2, Question A (1500 char max — used 883):

We recommend the Guidelines state explicitly, as a unifying principle applicable to paras 7-14, that high-risk status under Article 6 is a property derived from external law and context (sectoral product legislation for Art. 6(1); declared domain of use for Art. 6(2)) rather than an intrinsic property of the AI system. This framing clarifies why the two routes coexist without a shared risk taxonomy and aids consistent interpretation of 'intended purpose' across both routes. For

agentic AI systems specifically, we recommend noting that intended purpose alone may not travel reliably once a system is copied, forked, or repurposed downstream - a machine-readable, resolvable classification attribute bound to the system's identity would make the determination independently verifiable at the point of deployment, complementing rather than replacing the provider's own assessment.

Section IV — Annex III (Article 6(2))

subsection: ☑ IV.2 Horizontal issues applicable to Annex III use cases (IV.1 and IV.3 unchecked)

Three separate entries, using the repeatable "paragraph number → specify → more suggestions?" block:

Entry 1 — Paragraph number: 75 (500 char max — used 490)

Para 75 treats agentic configurations as one system for classification, but only at the point of assessment. We recommend the Guidelines also address classification persistence: whether a classification and its Art. 12/14 obligations survive copying, forking, or checkpoint-restoring, and which instance carries them - via a signed, resolvable classification attribute bound to agent identity, and an Art. 12 record binding identity, decision, and authorization state at the time of action.

→ Do you have more suggestions? Yes

Entry 2 — Paragraph number: 71 (500 char max — used 497)

We recommend the Guidelines recognise a voluntary, machine-readable deployment-context declaration (declaring entity/role, actual deployment purpose/capabilities, oversight arrangements, contextual assumptions) for embedded third-party components (chatbots, recommenders, triage tools). This would not determine legal status or replace Art. 6(3)/6(4) assessments, but would support discovery of how a component is actually deployed downstream, complementing para 71's human-involvement discussion.

→ Do you have more suggestions? Yes

Entry 3 — Paragraph number: 115 (500 char max — used 460)

We recommend the self-assessment in para 115 explicitly record admissibility - what was believed and decided, with provenance - rather than only asserting correctness. For agentic systems, we further recommend the assessment record the identity, decision, and the authorization state in force at the moment of the relevant action, since forking/copying can

preserve a classification while losing the binding that made the action accountable (see also para 75).

→ Do you have more suggestions? No

Question B (additional examples/use-cases needed?): No — our recommendations are horizontal/procedural, not proposals for new or excluded use-case entries.

Executive summary

- The two independent, non-overlapping legal routes to high-risk status (Article 6(1)/Annex I product-safety; Article 6(2)/Annex III use-case) share one underlying logic: high-risk is a property *derived* from external law and context, not intrinsic to the AI system. We recommend the guidelines state this explicitly.
- The guidelines already touch agentic AI systems in one narrow respect (para. 75, complex-system/circumvention guidance), but do not yet address what happens *after* classification: agentic systems are commonly copied, forked, fine-tuned, or restored from checkpoints. We propose three linked mechanisms to close that specific gap: (1) classification carried as a signed, resolvable identity attribute; (2) a portable evidentiary record for Article 12/14 obligations binding identity, decision, and the authorization state in force at the time of action; and (3) treating "which instance is real" and "who is authorized to act" as distinct, separately answerable questions.
- We propose a complementary, voluntary, machine-readable deployment-context declaration mechanism to help discover how embedded third-party components (chatbots, recommenders, triage tools, decision-support services) are actually deployed downstream — this does not itself determine legal status or replace any AI Act obligation.
- We contribute an expanded glossary spanning both classification routes, offered for the Commission's use and for a public SKOS vocabulary under development by the CG.
- All external technical projects referenced (Bella, the Cormier/Paraxiom Lean formalization, and the A2WF `siteai.json` format) are flagged explicitly as young, non-normative test beds or existence proofs, not settled or peer-reviewed authority — the conceptual arguments in this submission do not depend on any of them holding up.

---

1. Summary of our contribution

The AI KR CG welcomes the opportunity to comment on the draft guidelines. Our contribution has two parts:

1. **A conceptual observation** that runs through both Annex I (product-safety route) and Annex III (use-case route): high-risk status, as constructed by Article 6, is not a first-order property of an AI system but a **derived property conditioned on external law and context** — sectoral product legislation for the Annex I route, and declared domain of use for the Annex III route. We recommend the final guidelines state this explicitly as a unifying principle, since it clarifies why the two routes can coexist without a shared risk taxonomy.

2. **A vocabulary contribution**: a glossary of terms synthesised from the guidelines (Section 3 below), offered for inclusion in the Commission's own terminology work and simultaneously being developed into a public SKOS vocabulary by the CG. We ask that this be read alongside our specific recommendations on agentic AI systems and deployment-context discovery (Section 2 below), which the draft guidelines do not yet fully address.

---

2. Agentic AI systems: a gap in the draft guidelines

Paragraph 75 (Section IV.2.3, on AI systems as part of complex systems/products and services) already confirms that complex, interconnected setups such as agentic AI systems that coordinate and interact through linked actions are treated as a single system for classification purposes, where those linked actions jointly serve an intended high-risk purpose. This is a valuable anti-circumvention principle, but it operates only at the point of classification. It does not address what happens *after* a system is classified: agentic systems are commonly copied, fine-tuned, forked, and checkpoint-restored during their operational life, and the guidelines do not yet address whether and how a classification and its obligations persist across those events, nor how the classification itself is carried and verified between independent parties once the system is in operation. We recommend the Commission flag this narrower, lifecycle-oriented question explicitly for future guidance, and offer the following reasoning and technical grounding, informed by discussion in the W3C AI Agent Protocol Community Group.

2.1 Classification must be a resolvable, signed attribute, not a prose label

If high-risk status is derived and relational rather than intrinsic (see Section 1), then for an agentic system it must be **represented, transported, and verified between independent parties** — a classification that cannot travel with the agent and be checked at the point of action

is not enforceable in practice. We recommend the guidelines anticipate that, for agentic systems, the classification will need to be carried as a signed, resolvable attribute bound to the agent's identity, rather than asserted only in instructions for use or technical documentation — consistent with the guidelines' own emphasis on provider/deployer roles and intended purpose remaining distinct.

2.2 A portable, independently-verifiable evidentiary layer for Articles 12 and 14

The obligations that flow from a high-risk classification — record-keeping (Article 12) and human oversight (Article 14) — need a machine-readable evidentiary layer to be independently verifiable: what authority was evaluated, what was decided, and what was believed, with each element checkable without relying solely on the operator's own logs. This is the kind of auditability and belief-state representation already under discussion in the W3C AI Agent Protocol Community Group (github.com/w3c-cg/ai-agent-protocol, e.g. issue #34 on the evidentiary-provenance gap), and it is the concrete, machine-readable substance behind a guideline principle such as "a high-risk system must keep auditable records."

A principle worth stating explicitly: such an evidentiary layer should record **admissibility** — what was believed and decided, with provenance — not **truth**, and never a claim that the belief was correct. This is the honest and enforceable standard for record-keeping obligations: a high-risk system should be able to demonstrate due process, not omniscience.

2.3 Open question for the guidelines: classification persistence across lifecycle events

Paragraph 75 already treats a complex, interconnected agentic configuration as a single system for classification purposes at the point of assessment. The question raised here is different in kind: whether that classification and its attendant obligations persist once the system is copied, forked, fine-tuned, or restored from a checkpoint after deployment. To which instance does a high-risk classification and its obligations attach after a fork, and who is accountable? We recommend the guidelines flag this as an open question for agentic systems specifically.

Two distinguishable questions are involved, and conflating them causes confusion:

- "Which is the *real* one" after a fork is a question without a determinate answer (cf. Parfit's fission problem in personal-identity philosophy) — it is metaphysically empty and guidance should not attempt to resolve it.
- "Who acts as the deployed high-risk service" is, by contrast, a decidable delegation question — a "speaks-for" relationship in the sense used in computer-security identity literature (Abadi & Lampson) — and this is the question regulatory guidance should actually answer.

The record and the delegation must anchor at the same point. Sections 2.2 and 2.3 address two halves of one question that need to be explicitly connected. A content-addressed evidentiary record establishes that an action occurred and by which identity — but this does not, on its own, establish that the delegation authorizing that action still held at the moment the action was taken. For a static system this gap is invisible; for an agentic system, where authorization can be narrowed, delegated onward, or revoked between classification and action, it is precisely the substance of the persistence question above. After a fork, identifying which instance acted is not the same as establishing that the acting instance was authorized to act at that time. The practical consequence for record-keeping guidance: an evidentiary record satisfying Article 12 should commit the identity, the decision, and the **authorization state in force at the moment of the action** — without that third element, a fork can preserve the classification while losing the binding that made the action accountable. Whether this level of detail belongs in the Commission's guidelines or in the referenced technical standards work is a scoping question rather than a substantive disagreement; whether it also bears on the Article 25(1) provider-obligation transfer where a fork or substantial modification moves provider status to another party is a legal question outside this group's competence, which we flag rather than answer. This point draws on public discussion on the W3C AIKR CG mailing list: https://lists.w3.org/Archives/Public/public-aikr/2026Jul/0009.html.

Two independent, early-stage projects are noted here as test beds for these concepts — not as settled supporting work, and not as a single body of work:

- A lifecycle-persistence write-up (LINEAGE.md, github.com/Recursive-Emergence/bella), which also serves as a running open-source implementation of the evidentiary/belief-state layer described in 2.2. The repository is young (created 26 April 2026, three stars at time of writing).
- A separate, single-author formalization of identity-as-accountability-chains (Sylvain Cormier / Paraxiom Technologies, Zenodo, doi:10.5281/zenodo.20328038, linking to Paraxiom/identity-architecture-lean, not to bella). This is a Zenodo self-publication and has not been peer-reviewed. Its "zero sorries" Lean 4 verification claim was added on 6 July 2026, after the original 21 May 2026 publication, and has not been independently checked.

Both are offered only as concrete test beds for the concepts above; the conceptual argument in this section does not depend on either project's claims holding up.

2.4 Deployment-side context declaration and discovery

Where classification depends on intended purpose and operational context, the system-bound classification attribute described in 2.1 should be complemented by a means of discovering how the system or component is actually deployed. We recommend the guidelines recognise machine-readable deployment-context declarations as a voluntary interoperability mechanism,

particularly for embedded third-party components such as chatbots, recommenders, triage tools, and decision-support services.

Such a declaration could identify the declaring entity and its role, the relevant system or component, its actual deployment purpose and capabilities, applicable human oversight and transparency arrangements, and the contextual assumptions used in a classification analysis. It should be attributable, versioned, and linked to supporting evidence where appropriate. Public discovery need not imply public access to all underlying material; layered access may be necessary to protect personal data, security-sensitive information, and trade secrets.

A deployment-context declaration would not itself determine high-risk status. It would not replace provider assessments, documentation, registration, conformity assessment, or other obligations under the AI Act. In particular, an assessment made by a provider under Articles 6(3) and 6(4) should remain distinct from contextual information supplied by a deployer. The mechanism should remain technology-neutral and non-exclusive, and should not pre-empt Commission guidance, harmonised standards, common specifications, or sector-specific requirements. Well-known URIs, link relations, signed metadata, or equivalent discovery mechanisms could support interoperability without endorsing a particular format. The objective is discoverable and verifiable context, not self-certified compliance.

An existing community experiment in this direction is the W3C A2WF Community Group's `siteai.json` declaration format (a2wf.org, MIT-licensed), referenced here as an existence proof that deployment-context declaration is implementable today, not as a normative proposal. A2WF is itself a young W3C Community Group (proposed March 2026); its specification remains a starting point for discussion within that group rather than adopted or settled work, and is noted here on the same non-normative footing as the test beds referenced in 2.3.

---

3. Glossary contribution

The terms below are original paraphrases prepared for glossary/vocabulary use — not verbatim extracts from the guidelines — and are offered for the Commission's consideration and for inclusion in the CG's own SKOS-based public vocabulary. Each entry may carry a provenance/status annotation (`aikr:sourceStatus = "draft/non-binding"`), since the source guidelines are themselves a non-binding consultation draft.

3.1 Foundational / cross-cutting concepts

- **AI system** — A machine-based system, operating with some degree of autonomy and capable of adapting after deployment, that infers from input how to produce outputs (predictions, content, recommendations, decisions) capable of influencing a physical or virtual environment.

- **Intended purpose** — The scope of use for an AI system as fixed by its provider through instructions for use, technical documentation, and promotional/sales materials; the reference point against which classification and downstream obligations are assessed.
- **Reasonably foreseeable misuse** — Use of an AI system falling outside its declared intended purpose but plausible given the system's functionalities, context, and marketing.
- **High-risk AI system** — An AI system classified as such under Article 6 AI Act via one of two independent, non-overlapping routes: the product-safety route (Art. 6(1)/Annex I) or the use-case route (Art. 6(2)/Annex III).
- **Classification route (dual-track model)** — The AI Act's structural device of running two independent tests for high-risk status, one keyed to regulated-product safety law and one to fundamental-rights-sensitive domains of use, rather than a single unified risk test.
- **Substantial modification** — A change to a high-risk AI system, made after it has been placed on the market or put into service, significant enough to trigger provider obligations for the party making the change.
- **Provider-obligation transfer trigger** — The three conditions under Article 25(1) AI Act under which a distributor, importer, or deployer becomes legally a "provider": rebranding a high-risk system, substantially modifying one while it remains high-risk, or modifying the purpose of a non-high-risk system such that it becomes high-risk.

3.2 Product-safety route (Article 6(1) / Annex I)

- **Regulated product** — A product falling within the material scope of one of the pieces of Union harmonisation legislation listed in Annex I AI Act.
- **Safety component (autonomous AI Act definition)** — A component of a product or AI system that either fulfils a safety function, or whose failure or malfunctioning would endanger health, safety, or property — defined independently of any sectoral "safety component" definition.
- **Safety function** — An intended purpose, fixed by the provider, of preventing or mitigating risk to health, safety, or property; assessed by intent/documentation rather than outcome.
- **Preventive function** (sub-type of safety function) — Monitoring/detection of hazardous conditions, detection of maintenance need, prevention of physical harm, or supervision of another safety-performing system.
- **Mitigation function** (sub-type of safety function) — Controlling or limiting physical harm, triggering a safe-stop or similar action, or controlling another safety-performing system.
- **Failure or malfunctioning (as classification trigger)** — The consequence-based route to "safety component" status: incorrect outputs, loss of availability, drift, latency errors, or misclassification that could endanger health, safety, or property, regardless of declared intended purpose.
- **Endangerment** — An increase in hazard to health, safety, or property, as distinct from reputational harm, purely financial loss, or inconvenience.

- **Harm to health** — Illness, injury, temporary or permanent impairment, chronic disease, or psychological trauma significantly affecting ordinary task performance.
- **Third-party conformity assessment** — The procedural requirement, set by relevant sectoral legislation, that a product's compliance be verified by an external notified body rather than manufacturer self-declaration.
- **Conformity assessment module (A–H)** — The harmonised structure of compliance-demonstration procedures under Decision 768/2008/EC, ranging from Module A (self-declaration) to modules requiring notified-body involvement.
- **Notified body** — An accredited third-party conformity-assessment body whose involvement is required by most conformity-assessment modules other than Module A.
- **Union harmonisation legislation (Annex I)** — The exhaustively-listed set of EU product-safety instruments (Section A: NLF acts; Section B: other harmonisation acts) that Annex I AI Act references.
- **New Legislative Framework (NLF)** — The EU's general product-safety governance architecture that most Annex I Section A instruments follow.
- **Blue Guide** — The Commission's non-binding interpretive guide to NLF-based EU product rules.
- **Harmonised standard (as gatekeeper for self-assessment)** — A standard whose application is, for certain products, a precondition for using Module A instead of third-party assessment.
- **Compliance-burden minimisation mechanism** — The AI Act provisions (Art. 8(2), 9(10), 17(3), 40) allowing operators to fold AI-specific risk assessment into existing sectoral quality-management systems.
- **Sectoral risk inheritance principle** — The principle (Recital 51) that the AI Act does not itself determine or alter a product's risk profile, but inherits the classification already made by sectoral legislation.

3.3 Use-case route (Article 6(2) / Annex III)

- **Annex III area** — One of the eight broad domains of use enumerated in Annex III AI Act (biometrics; critical infrastructure; education and vocational training; employment, workers' management and access to self-employment; access to essential private and public services and benefits; law enforcement; migration, asylum and border control management; administration of justice and democratic processes) within which an exhaustive, delegated-act-amendable list of specific use cases triggers Article 6(2) classification.
- **Exhaustive use-case list (Annex III)** — The closed list of specific use cases within each Annex III area against which a system's intended purpose is matched; modifiable only through an Article 7(1) delegated act, in contrast to the abstract two-condition test used for the Annex I product-safety route.
- **Human-involvement independence principle** — The principle that the presence, type, or degree of human involvement during deployment does not itself change an AI system's high-risk classification under Article 6(2); human oversight functions as a

compliance obligation under Article 14 once a system is classified, not as a means of avoiding classification, save for its narrow evidentiary role in the Article 6(3) filter.

- **Article 6(3) filter mechanism** — The exemption allowing a provider to declare a system, whose intended purpose otherwise matches an Annex III use case, as non-high-risk where it can demonstrate at least one of four statutory conditions and that the system does not perform profiling.
- **Article 6(3) filter conditions (a)–(d)** — The four alternative statutory conditions, any one of which allows the filter to apply: (a) the system performs a narrow procedural task; (b) the system improves the result of a previously completed human activity; (c) the system detects patterns or deviations from prior decision-making without replacing or influencing the prior human assessment; or (d) the system performs a preparatory task to an assessment relevant to a listed use case.
- **Profiling exception (to the filter)** — The rule that the Article 6(3) filter is unavailable, regardless of whether the other conditions are met, wherever the AI system performs profiling of natural persons.
- **Self-assessment, registration and monitoring obligation (filter use)** — The requirement that a provider relying on the Article 6(3) filter document its exemption assessment, register the system in the EU database (Article 71), and remain subject to ongoing monitoring, rather than treat the filter as a one-off, undocumented determination.
- **Anti-circumvention safeguard (Annex III)** — The guidelines' requirement that the filter mechanism not be invoked by artificially narrowing a system's stated functionality to escape high-risk obligations where its actual capabilities would otherwise meet a listed use case.
- **'Intended to be used' test (Annex III)** — The interpretive standard for matching a system's declared intended purpose against a listed use case, applying the general "intended purpose" concept (see 3.1) specifically to the Annex III use-case list.
- **'On behalf of' qualifier** — The interpretive clarification, attached to the Annex III use-case entries concerning public authorities, that classification applies where the system is used on behalf of a specified actor (e.g. a public authority, including outsourced private-entity use), rather than by any deployer regardless of role.
- **Permitted-under-law qualifier** — The interpretive clarification that the Annex III use-case entries in points 1, 6 and 7 (biometrics; law enforcement; migration, asylum and border control) are scoped by the phrase "in so far as their use is permitted under relevant Union or national law," meaning classification presupposes the underlying use is already lawful rather than the AI Act independently authorising it.
- **Complex-system embedding (Annex III)** — Guidance on how use-case classification applies where an AI system is one component of a larger product or service, including agentic configurations where linked actions or components jointly serve an intended high-risk purpose; resolved through the "intended to be used" test rather than through the safety-component analysis used in the Annex I route.
- **Stand-alone high-risk use case** — An AI system classified as high-risk under Article 6(2) not because it is a regulated product or a safety component, but because its

declared intended purpose places it within one of the domains of use enumerated in Annex III.

3.4 Governance / procedural apparatus

- **AI Board** — The European Artificial Intelligence Board, consulted on the guidelines' drafting.
- **AI Office** — The Commission body providing ongoing support on high-risk classification and collecting practical use cases via the Service Desk and regulatory sandboxes.
- **AI Act Service Desk / Single Information Platform** — The Commission's designated channel for classification questions.
- **Regulatory sandbox** — A controlled testing environment for AI systems, to be established by Member States.
- **Delegated act (Annex III amendment power)** — The Commission's Article 7 power to add, amend, or remove high-risk use cases in Annex III.
- **Annual monitoring instrument** — The Article 112(1) mechanism requiring yearly evaluation and review of the Annex III use-case list.

3.5 New terms proposed for agentic AI systems and deployment discovery

- **Derived classification** — A risk class conditioned on external law/context rather than an intrinsic property of the system; for agentic systems, this implies the classification must be a resolvable, signed attribute rather than a static label.
- **Evidentiary record** — A content-addressed commitment/decision/receipt (plus a belief-state reference) that makes Article 12/14 obligations independently verifiable, recording admissibility rather than truth.
- **Classification persistence** — Whether and how a derived high-risk class and its obligations survive copying, forking, fine-tuning, or checkpoint-restoring an AI system, and which resulting instance carries them.
- **Speaks-for authority** — The delegation that decides which instance of a forked or copied system acts as the classified high-risk service, distinct from and independent of the (unanswerable) question of which instance is the "real" one.
- **Deployment context declaration** — A machine-readable, attributable, and versioned statement made available by a deploying party or system operator describing the operational context in which an AI system or component is used (actual purpose, deployed capabilities, oversight and transparency arrangements, and contextual assumptions relevant to risk classification). It supports discovery and evidence exchange but does not itself determine legal status or replace obligations under the AI Act.

---

4. SKOS integration notes (for the CG's own vocabulary work, shared for transparency)

- These terms map onto a "regulatory classification" facet distinct from the CG's existing agent-interoperability categories; whether this warrants its own governance/compliance category, or cross-references into existing categories, remains open (e.g. safety component → agent capability/function; third-party conformity assessment → trust/verification).
- `skos:broader`/`skos:narrower` relations above are drafted at one level of depth; deeper nesting (e.g. conformity assessment modules A/B/C) can be expanded if needed.
- A provenance/status annotation property (`aikr:sourceStatus = "draft/non-binding"`) is recommended on any concept sourced from the guidelines, since they remain a stakeholder-consultation draft pending CJEU-authoritative interpretation.

.