CARG

6 Components

CARG Runtime Governance Assessment (6 Components)

Assess an adaptive AI deployment against the six components of the CARG Framework that close the runtime governance gap.

by **Waydell D. Carvalho**

Published by Cinderpoint

SAFEMACHINE RULES	COMPONENT	EFFECTIVE	VARIANT
2026.05	**CARG v1**	**2026-05-20**	**Template**

Instruction: For each of the 6 components, answer 4 observable indicators. RED = obligation unmet. YELLOW = partial or unclear. GREEN = obligation fully met. Components accumulate; the total score and per-component profile identify which CARG obligations are most weakly satisfied. If you cannot answer a question, mark it as a gap.

Use: Run when a self-modifying AI deployment is proposed, before sign-off, and on the cadence required by the deployment's SMC level (annual for SMC-2, semi-annual for SMC-3, quarterly for SMC-4 - see Section 5.6).

SOURCE	Derived from Carvalho 2026, "The Runtime Governance Gap in AI Regulation," Sections 5.1 through 5.6. Each of the 6 components addresses one of the four structural failures in Section 4.
SCORING	Per question RED=20, YELLOW=10, GREEN=0. Per component = sum of 4 questions (0-80 raw / 0-100 normalized). Total = sum of 6 components (0-480 raw / 0-100 normalized).
TIER THRESHOLDS (NORMALIZED 0-100)	0-30 Aligned (CARG-compatible). 31-65 Drifting (runtime governance gap is forming). 66-100 Gap (the gap is operating in this deployment).
HARD STOP	A system classified at SMC-4 (architectural self-modification) combined with HRP-4 (irreversible physical harm or death) is deployment-restricted or prohibited under Section 5.5. This overrides the per-question score and forces a NO-GO decision.
RELATION TO BASE / SAFEARC / OUTPACED	BASE = organisational readiness. SAFEARC = governance lifecycle. OUTPACED = failure-mode exposure. CARG = runtime regulatory alignment for adaptive systems. Use CARG when the system is self-modifying at runtime.

Component 1 - Runtime Oversight Obligation (ROO, §5.1)

A mandatory, legally enforceable duty to monitor AI system behavior during operation, including the technical and institutional capacity to interrupt, override, or shut down system processes. Closes the absence-of-continuous-oversight failure.

1.1. Is there a written policy that obligates continuous (or near-continuous) behavioral monitoring of this system during operation?

RED = no policy. YELLOW = informal / discretionary. GREEN = binding policy in place.

RED YELLOW GREEN

1.2. Is monitoring infrastructure (logs, traces, anomaly detection) deployed and collecting behavioral data right now?

RED = nothing collected. YELLOW = partial coverage. GREEN = continuous collection with drift indicators.

RED YELLOW GREEN

1.3. Does a named human role have the technical authority and institutional standing to interrupt, override, or shut the system down?

RED = no one. YELLOW = on paper only. GREEN = named role with tested authority.

RED YELLOW GREEN

1.4. When monitoring surfaces unacceptable behavior, is human intervention authority exercised in practice (not just on paper)?

RED = never tested. YELLOW = rare. GREEN = exercised within last review cycle.

RED YELLOW GREEN

Component 2 - Persistent Liability Doctrine (PLD, §5.2)

Ensures that legal responsibility does not fracture across self-modification cycles. The deployer or provider remains continuously accountable, regardless of whether the harm-causing behavior reflects initial design or subsequent internal reconfiguration. Closes the accountability-diffusion failure.

2.1. Is there a clear, written allocation of liability between provider, deployer, and operator for harms arising during runtime?

RED = no allocation. YELLOW = ambiguous / boilerplate. GREEN = explicit and signed.

RED YELLOW GREEN

2.2. Does the liability allocation explicitly survive self-modification events (re-tuning, fine-tuning, online updates)?

RED = self-modification not addressed. YELLOW = mentioned but not binding. GREEN = explicit continuing obligation.

RED YELLOW GREEN

2.3. Are version-change, decision-pathway, and reconfiguration logs preserved as the evidentiary record for liability investigations?

RED = logs not retained. YELLOW = retained but not legally privileged / structured. GREEN = retained and discovery-ready.

RED YELLOW GREEN

2.4. Is the deployer prepared to be held liable for foreseeable consequences of the system's adaptive capacity (not just its initial design)?

RED = position is "we only own initial release." YELLOW = mixed. GREEN = explicit ownership of adaptive behavior.

RED YELLOW GREEN

Component 3 - Self-Modification Capacity (SMC, §5.3)

A capability-based classification of how much autonomy the system has to reconfigure itself during operation. Scale runs SMC-0 (no adaptive capacity) through SMC-4 (architectural self-modification). Where evidence is ambiguous between adjacent levels, default to the higher SMC category.

3.1. Has the deployment been formally classified at an SMC level (0 to 4), with the classification documented in technical specifications?

RED = unclassified. YELLOW = informal estimate. GREEN = documented SMC level with rationale.

RED YELLOW GREEN

3.2. Is the SMC classification grounded in BOTH architectural documentation AND empirical observation of runtime behavior?

RED = neither. YELLOW = one or the other. GREEN = both, cross-checked.

RED YELLOW GREEN

3.3. When the architectural and empirical evidence disagree, does policy require defaulting to the HIGHER SMC level?

RED = default lower / not addressed. YELLOW = case-by-case. GREEN = explicit higher-default rule.

RED YELLOW GREEN

3.4. Is the SMC classification reviewed and updated whenever operational evidence (new logs, new behaviors, vendor updates) becomes available?

RED = static after release. YELLOW = reviewed annually. GREEN = continuously reviewed.

RED YELLOW GREEN

Component 4 - Harm Risk Potential (HRP, §5.4)

A severity / reversibility taxonomy independent of SMC. Scale runs HRP-0 (no meaningful harm) through HRP-4 (irreversible physical harm or death). Combines domain harm models with scenario analysis of plausible worst-case outcomes.

4.1. Has the deployment been formally classified at an HRP level (0 to 4), with documented scenarios of plausible worst-case harm?

RED = no classification. YELLOW = informal. GREEN = documented HRP with scenarios.

RED YELLOW GREEN

4.2. Does the HRP classification reflect BOTH the severity AND the reversibility of potential harms (not just severity alone)?

RED = severity only. YELLOW = partial. GREEN = both addressed explicitly.

RED YELLOW GREEN

4.3. Are the HRP scenarios grounded in worst-case analysis (not just observed historical incidents)?

RED = historical only. YELLOW = mixed. GREEN = explicit worst-case modeling.

RED YELLOW GREEN

4.4. Is HRP re-evaluated when the deployment context, user population, or downstream use changes?

RED = static. YELLOW = on major change. GREEN = continuous reevaluation.

RED YELLOW GREEN

Component 5 - Risk Response Requirements (RRR, §5.5)

Mandatory safeguards triggered by combinations of SMC and HRP. Cells range from "None" (SMC-0, HRP-0) through "Real-time Oversight + Kill Switch + Pre-Approval" (high SMC x high HRP). The SMC-4 x HRP-4 cell is "Restricted" - deployment prohibited.

5.1. Has the RRR matrix cell for this deployment's (SMC, HRP) been looked up, and the required safeguard tier identified?

RED = not looked up. YELLOW = identified but not signed off. GREEN = identified, documented, approved.

RED YELLOW GREEN

5.2. Is each safeguard required by the matrix cell (logging / testing / monitoring / HITL / kill switch / pre-approval) operationally active right now?

RED = required safeguards missing. YELLOW = some active, some pending. GREEN = all active and tested.

RED YELLOW GREEN

5.3. When drift detection shows the system has moved into a HIGHER (SMC, HRP) cell, are the new safeguards implemented immediately?

RED = no escalation procedure. YELLOW = manual / slow. GREEN = automatic re-tier and safeguard activation.

RED YELLOW GREEN

5.4. If the deployment cell is SMC-4 x HRP-4 (Restricted), is the deployment in fact restricted or prohibited?

This is the doctrinal HARD STOP. "No - deployed anyway" is a NO-GO condition regardless of other answers.

☐ Yes - restricted / prohibited ☐ No - deployed anyway ☐ Not applicable - not at SMC-4 x HRP-4 ☐ Unknown

Component 6 - Verification and Reassessment Cycle (VRC, §5.6)

Institutionalizes periodic and drift-triggered review. SMC-2 systems are reassessed annually, SMC-3 semi-annually, SMC-4 quarterly or on regulator request. Closes evidentiary-instability and compliance-drift failures.

6.1. Is there a scheduled reassessment cycle (annual / semi-annual / quarterly) appropriate to the deployment's SMC level?

RED = no schedule. YELLOW = schedule exists but slipped. GREEN = on cadence.

RED YELLOW GREEN

6.2. Are drift thresholds (behavioral deviation, emergent capability, output-distribution shift, incident accumulation) defined in advance?

RED = no thresholds. YELLOW = informal. GREEN = pre-defined and regulator-approved.

RED YELLOW GREEN

6.3. When drift detection fires, does a reassessment in fact occur (not just an alert)?

RED = alert ignored. YELLOW = informal review. GREEN = full reassessment workflow.

RED YELLOW GREEN

6.4. Are all verification records (monitoring logs, incident reports, reassessment findings) retained and accessible to regulators?

RED = not retained. YELLOW = retained but not regulator-accessible. GREEN = both.

RED YELLOW GREEN

Red Flags

- ☐ No written runtime monitoring obligation (ROO unmet)
- ☐ Liability disclaimed for self-modification events (PLD unmet)
- ☐ No SMC classification on file, or SMC defaults to LOWER on ambiguous evidence
- ☐ HRP justified by "no prior complaints" instead of worst-case scenarios
- ☐ RRR matrix cell not looked up, or required safeguards not active
- ☐ Drift alerts produce no reassessment
- ☐ Verification records inaccessible to regulators
- ☐ Deployment at SMC-4 x HRP-4 not restricted

Final CARG Summary

7.1. CARG normalized score (0 to 100):

7.2. Three components requiring immediate action (top concerns):

7.3. RRR matrix cell (SMC level, HRP level, required safeguard):

7.4. Hard-stop triggered (SMC-4 x HRP-4)?

☐ Yes - NO-GO ☐ No ☐ Unknown

7.5. Next reassessment date:

Citations

[1] **CARG manuscript (Carvalho 2026)** - https://cinderpoint.com/safemachine/papers/carg

[2] **EU AI Act - Regulation (EU) 2024/1689** - https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[3] **EU AI Liability Directive (proposed)** - https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en

[4] **NIST Cybersecurity Framework** - https://www.nist.gov/cyberframework

[5] **FDA Software as a Medical Device (SaMD)** - https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd