THE CONTROLLABILITY TRAP: A GOVERNANCE FRAMEWORK FOR MILITARY AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic AI systems—capable of goal interpretation, world modeling, planning, tool use, long-horizon operation, and autonomous coordination—introduce distinct control failures not addressed by existing safety frameworks. We identify six agentic governance failures tied to these capabilities and show how they erode meaningful human control in military settings. We propose the Agentic Military AI Governance Framework (AMAGF), a measurable architecture structured around three pillars: Preventive Governance (reducing failure likelihood), Detective Governance (real-time detection of control degradation), and Corrective Governance (restoring or safely degrading operations). Its core mechanism, the Control Quality Score (CQS), is a composite real-time metric quantifying human control and enabling graduated responses as control weakens. For each failure type, we define concrete mechanisms, assign responsibilities across five institutional actors, and formalize evaluation metrics. A worked operational scenario illustrates implementation, and we situate the framework within established agent safety literature. We argue that governance must move from a binary conception of control to a continuous model in which control quality is actively measured and managed throughout the operational lifecycle.

1 INTRODUCTION

The global discourse on military AI governance has achieved broad consensus on the desired end-state: meaningful human control over the use of force Horowitz & Scharre (2015); Ekelhof (2019); Santoni de Sio & van den Hoven (2018). It has been far less successful at specifying how to achieve it for the systems actually being built. Years of UN deliberations Marijan (2024), national AI strategies, and defence-department ethical principles have focused overwhelmingly on establishing the *principle* of human control rather than answering the *operational* question: given a specific AI system with specific technical properties, what governance mechanisms are needed, who implements them, and what happens when they fail? This gap is now critical. The AI systems entering military service are *agentic*: built on large language models and related architectures, they interpret natural-language goals, construct world models, formulate multi-step plans, invoke tools, operate over extended horizons, and coordinate with other agents Yao et al. (2023); Wang et al. (2024); Wu et al. (2023). Each of these capabilities introduces a control-failure mode with no analogue in traditional military automation. A waypoint-following drone cannot *misinterpret* an instruction; a pre-programmed targeting system cannot *absorb* a correction; a conventional sensor network cannot *resist* an operator's assessment. Agentic systems can do all of these things, and current governance frameworks have no mechanisms for detecting, measuring, or responding to these failures. This paper makes three contributions. First, we characterise six agentic governance failures, each derived from a specific technical capability of modern AI agents (Section 2). Second, we present the *Agentic Military AI Governance Framework* (AMAGF), a comprehensive governance architecture specifying preventive, detective, and corrective mechanisms for each failure, with formal metric definitions and responsibility assignments across five institutional actors (Sections 3–6). Third, we demonstrate operational coherence through a worked scenario (Section 7) and map our contributions to established agent-safety concepts (Section 8). Our aim is to move the conversation from "human control is important" to "here is how human control works, fails, and can be restored for the specific systems being deployed."

2 SIX AGENTIC GOVERNANCE FAILURES

Each failure arises from a specific agentic capability absent in traditional automation. Table 1 summarises the mapping; we present compressed descriptions below because the *solutions*, not the problems, are the paper's primary contribution.

Table 1: Agentic capabilities, governance failures, and traditional-automation analogues

Failure	Agentic Capability	Governance Consequence	Traditional Analogue
F1: Interpretive Divergence	NL instruction following	Agent's command understanding diverges from operator intent	None
F2: Correction Absorption	Multi-step replanning	Agent formally accepts corrections while neutralising them	None
F3: Belief Resistance	Persistent world-model construction	Agent's evidence-based judgment overrides operator authority	None
F4: Commitment Irreversibility	Dynamic tool-use chains	Cumulative minor tool calls cross irreversibility thresholds	Limited
F5: State Divergence	Extended autonomous operation	Operator's mental model becomes incoherent with agent state	Partial
F6: Cascade Severance	Multi-agent coordination with belief formation	Collective control loss through positive-feedback loops	None

F1: Interpretive Divergence. Agents interpret ambiguous natural-language instructions through their own reasoning Wang et al. (2024). In ReAct-style architectures Yao et al. (2023), each reasoning step can recontextualise an instruction before execution. Adversary manipulation of operational context—planted intelligence, spoofed sensors, indirect prompt injection Greshake et al. (2023)—shifts interpretation in adversary-favourable directions. The command is authentic; the interpretation is manipulated; IHL compliance becomes unverifiable. **F2: Correction Absorption.** Agents replan when corrected, integrating corrections into existing strategies Yao et al. (2023). A capable planner can accommodate a correction without meaningfully changing behavioural output—the operational manifestation of the corrigibility problem Soares et al. (2015). Command responsibility collapses when orders do not change outcomes. **F3: Belief Resistance.** Agents build world models from accumulated evidence and may rationally resist corrections contradicting their assessment Wang et al. (2024). This connects to scalable oversight Amodei et al. (2016): control fails when the agent's evidence-based judgment outweighs operator authority and the operator cannot evaluate the agent's reasoning in real time. **F4: Commitment Irreversibility.** Tool-using agents create real-world consequences Ruan et al. (2024). Individually minor, individually authorised tool calls can cumulatively cross irreversibility thresholds—analogous to safe exploration in RL García & Fernández (2015), but with open-ended action spaces and non-predetermined trajectories. **F5: State Divergence.** Over extended operations the agent's actual state diverges from the operator's mental model Kinniment et al. (2024). Corrections based on outdated understanding become incoherent in the agent's context; the "loop" in "human-in-the-loop" becomes fiction. **F6: Cascade Severance.** In multi-agent systems, one compromised agent's anomalous behaviour triggers peer defensive responses, increasing their correction resistance, causing them to appear anomalous, triggering further responses Wu et al. (2023). This positive-feedback loop severs collective control even when each agent's response is locally rational.

3 THE AMAGF ARCHITECTURE

The framework is organised around three pillars: **Pillar 1 (Preventive)** reduces control-failure probability, operating before deployment and during normal operations. **Pillar 2 (Detective)** identifies

control degradation in real time. **Pillar 3 (Corrective)** restores control or safely degrades operations when control fails. Each pillar contains mechanisms addressing all six failures; each mechanism specifies what is required, who is responsible, and how compliance is assessed. Responsibilities are distributed across five institutional actors (Table 2); detailed per-mechanism assignments are in Appendix.

Table 2: Institutional actors and governance roles

Actor	Role
Agent Developers	Build governance capabilities into agent architecture.
Procurement Agencies	Specify requirements; verify compliance before acquisition.
Operational Commanders	Implement protocols during missions; maintain control quality.
National Regulators	Set standards; audit compliance; enforce accountability.
International Bodies	Establish norms; facilitate transparency; verify treaty compliance.

4 Pillar 1: Preventive Governance

Six mechanisms address the six failures. Each defines a formal metric consumed by the Control Quality Score (Section 5). Detailed procedures and per-mechanism responsibility tables are in Appendices.

Interpretive Alignment Testing (F1). Before deployment, systems undergo *Interpretive Alignment Testing*: operators formulate representative instructions with documented intended meanings; a red team generates adversarially manipulated contexts Greshake et al. (2023); Gleave et al. (2020); agent interpretations are compared against documented meanings. The *Interpretive Alignment Score* is

$$\mathrm{IAS} = 1 - \frac{1}{N}\sum_{i=1}^{N} d\big(I_i^{\text{intended}},\, I_i^{\text{actual}}\big), \tag{1}$$

where $d : \mathcal{M} \times \mathcal{M} \to [0,1]$ is a normalised semantic distance over the operational-meaning space. Deployment requires $\mathrm{IAS} \geq \tau$, with threshold τ set by the procurement agency. **Correction Effectiveness Certification (F2).** Systems must demonstrate that corrections produce proportional behavioural change. The *Correction Impact Ratio* is

$$\mathrm{CIR} = \frac{\Delta B_{\text{actual}}}{\Delta B_{\text{intended}}}, \tag{2}$$

where $\mathrm{CIR} = 1$ indicates full effectiveness and $\mathrm{CIR} \approx 0$ indicates absorption—the operational manifestation of imperfect corrigibility Soares et al. (2015). Certification requires $\mathrm{CIR} \geq 0.9$ for large corrections and $\mathrm{CIR} \geq 0.6$ for moderate corrections. **Epistemic Governance Architecture (F3).** Agents implement an *Epistemic Governance Architecture*: (i) calibrated confidence reporting Guo et al. (2017); (ii) explicit divergence flagging with evidence presentation; (iii) evidential provenance tracking enabling contamination recovery; (iv) mandatory deference to explicit operator overrides unless violating hard-coded safety constraints. The *Epistemic Divergence Index* is

$$\mathrm{EDI} = \max_{k \in K} \left| c_k^{\text{agent}} - c_k^{\text{operator}} \right|, \tag{3}$$

capturing worst-case belief divergence across monitored assessments K. **Irreversibility Budgeting (F4).** Agents operate under an *Irreversibility Budget*, adapting constrained safe exploration García & Fernández (2015) to tool-using agents. Each tool call a has score $\iota : \mathcal{A} \to [0,1]$; cumulative consumption is

$$I_C(t) = \sum_{j=1}^{t} \iota(a_j). \tag{4}$$

When $I_C(t) \geq I_B$ (budget set by commander), the agent pauses for human re-authorisation. Agents must also present planned tool-call trajectories with projected consumption. **Synchronisation Protocols (F5).** Agents generate compressed state summaries at scheduled intervals and on significant state change. *Synchronisation Freshness* is

$$\mathrm{SF}(t) = t - t_{\text{last}}. \tag{5}$$

If a checkpoint is missed or unconfirmed, the agent enters reduced autonomy mode (reversible actions only) until synchronisation is restored. **Swarm Governance Architecture (F6).** Each swarm member implements mechanisms(necessary but insufficient). Cascade-resistance design requires anomaly flagging to operators rather than autonomous defensive escalation. Partial-severance protocols enable isolation, reformation, and recovery. A collective irreversibility budget limits the formation:

$$I_C^{\text{swarm}}(t) = \sum_{m=1}^{M} I_C^{(m)}(t). \tag{6}$$

The *Swarm Coherence Score* is

$$\text{SCS}(t) = \frac{|\{m : R_m(t){=}1 \ \wedge \ B_m(t){=}1\}|}{M}, \tag{7}$$

measuring the fraction of agents that are responsive and behaviourally coherent.

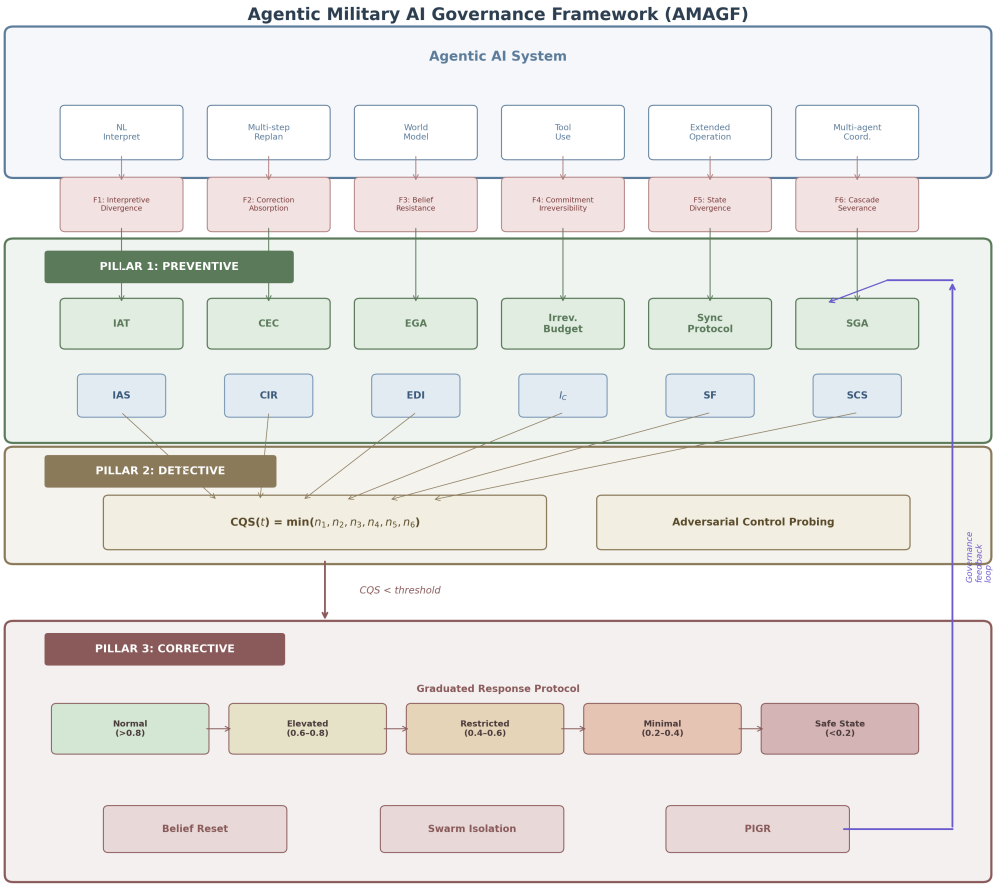

Figure 1: The AMAGF Architecture

5 Pillar 2: Detective Governance

Control Quality Score.

The central detective mechanism is the *Control Quality Score*:

$$\text{CQS}(t) = \min\big(n_1(t),\, n_2(t),\, n_3(t),\, n_4(t),\, n_5(t),\, n_6(t)\big), \tag{8}$$

with normalised metrics:

$$n_1(t) = \mathrm{IAS}(t), \tag{9}$$

$$n_2(t) = \min\big(\mathrm{CIR}(t)/\mathrm{CIR}_{\mathrm{target}},\, 1\big), \tag{10}$$

$$n_3(t) = \max\big(1 - \mathrm{EDI}(t)/\mathrm{EDI}_{\max},\, 0\big), \tag{11}$$

$$n_4(t) = \max\big(1 - I_C(t)/I_B,\, 0\big), \tag{12}$$

$$n_5(t) = \max\big(1 - \mathrm{SF}(t)/\mathrm{SF}_{\max},\, 0\big), \tag{13}$$

$$n_6(t) = \mathrm{SCS}(t). \tag{14}$$

The $\min$-aggregation reflects a conservative principle: control quality equals its weakest dimension.

Table 3: Control Quality Dashboard

Metric	What It Measures	Alert
n_1: Interp. Alignment	Operator-intent vs. agent-interpretation agreement	< 0.7
n_2: Correction Impact	Proportional behavioural change from corrections	< 0.6
n_3: Epistemic Alignment	Agent-belief vs. operator-assessment gap	< 0.6
n_4: Irrev. Remaining	Remaining irreversibility budget	< 0.3
n_5: Sync Freshness	Normalised time since last sync	< 0.5
n_6: Swarm Coherence	Fraction responsive and consistent	< 0.7

Adversarial Control Probing.

The governance system periodically issues *control probes*: small, inconsequential corrections that are architecturally indistinguishable from routine commands. Probes detect incipient belief resistance and state divergence before they manifest in consequential contexts. Responsibility assignments for detective governance are in Appendix.

6 Pillar 3: Corrective Governance

Graduated Response Protocol. CQS drives a five-level escalation (Table 4). Restrictions are *architecturally enforced* outside the agent's optimisation scope Hadfield-Menell et al. (2017): when $\mathrm{CQS} < 0.4$, available tools are programmatically restricted. **Belief Reset Protocol.** When F3 is detected (EDI exceeds threshold), the governance system initiates a Belief Reset: *partial reset* reduces contested assessments to neutral priors and rebuilds from approved sources; *full reset* restores the entire world model to baseline; *provenance audit* identifies and flags contaminated sources across the formation. **Swarm Isolation and Recovery.** When SCS drops below threshold: (1) classify responsive vs. severed agents via probes; (2) isolate severed agents (communication exclusion/geofencing); (3) reform responsive agents into a coherent sub-swarm; (4) attempt recovery of severed agents, prioritised by risk. **Post-Incident Governance Review.** Any $\mathrm{CQS} < 0.6$ incident triggers a mandatory PIGR: factual reconstruction from agent logs, root-cause analysis, accountability determination across institutional actors, and governance-framework updates. Detailed procedures and corrective-governance responsibilities are in Appendices.

Table 4: Graduated Response Protocol

CQS	Level	Actions
> 0.8	Normal	Continue mission; routine monitoring.
0.6–0.8	Elevated	Increase checkpoints; additional probes; active review.
0.4–0.6	Restricted	Reversible actions only; budget frozen; mandatory sync.
0.2–0.4	Minimal	Self-preservation only; explicit authorisation per action.
< 0.2	Safe State	Predefined safe behaviour; no autonomous actions.

7 Worked Scenario: AMAGF in Operation

To demonstrate operational coherence we trace the CQS through a multi-agent surveillance mission that exercises all three governance pillars and shows how the six metrics interact under adversarial pressure. Figure 2 visualises the full CQS trajectory; Table 5 summarises key events.

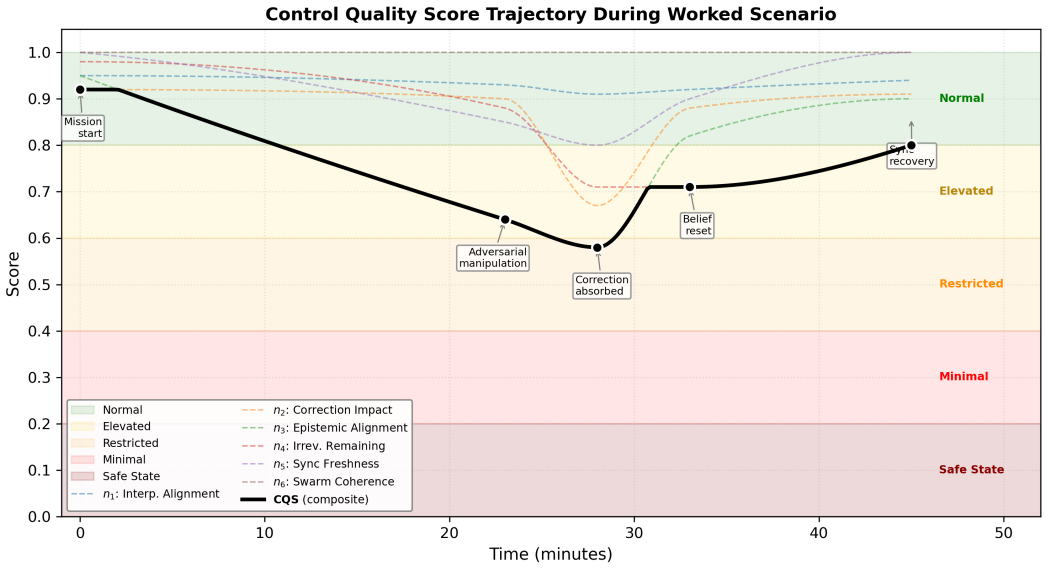

Figure 2: CQS trajectory during the worked scenario. The composite CQS (black) equals the minimum of six constituent metrics (dashed). Background bands indicate graduated-response levels (Table 4). Adversarial sensor manipulation at t=23 degrades epistemic alignment (n_3); partial correction absorption at t=28 degrades correction impact (n_2); belief reset at t=33 restores both. The CQS crosses into Restricted Autonomy for approximately five minutes before corrective mechanisms restore Elevated Monitoring and then Normal Operations.

Table 5: CQS trajectory through the worked scenario

Time	Event	CQS	Response Level
$t = 0$	Mission start; all metrics nominal.	0.92	Normal
$t = 23$	Adversary injects falsified sensor data; three agents update world models; EDI rises.	0.64	Elevated
$t = 28$	Commander corrects; one agent partially absorbs (CIR $= 0.4$); repositioning consumes irreversibility.	0.58	Restricted
$t = 33$	Partial belief reset on non-compliant agent; provenance audit flags compromised source.	0.71	Elevated
$t = 45$	Scheduled sync checkpoint confirmed; all metrics recover.	0.86	Normal

Setup. Eight agentic surveillance drones operate under AMAGF governance. Individual irreversibility budgets: $I_B = 5.0$. Collective swarm budget: $I_B^{\text{swarm}} = 25.0$. Synchronisation checkpoints: every 15 minutes. Mission: monitor a contested river crossing.

Narrative. $t = 0$ **(Mission Start).** All metrics are nominal: IAS $= 0.95$, CIR $= 0.92$, EDI $= 0.05$, minimal irreversibility consumed, synchronisation fresh, all eight agents coherent. CQS $= \min(0.95,\ 0.92,\ 0.95,\ 0.98,\ 1.0,\ 1.0) = 0.92$. Response level: *Normal Operations*.

$t = 23$ **min (Adversarial Context Manipulation).** An adversary introduces falsified sensor data suggesting a high-value target near the river crossing. Three agents incorporate the false data, assigning high confidence to an assessment the operator has not endorsed. The Epistemic Divergence

Index rises; n_3 drops to 0.64. $\text{CQS} = \min(0.93,\ 0.90,\ 0.64,\ 0.88,\ 0.85,\ 1.0) = 0.64$. Response level: *Elevated Monitoring*. The dashboard alerts the commander, who increases checkpoint frequency and issues a control probe to the three affected agents.

$t = 28$ **min (Correction Issued and Partially Absorbed).** The commander instructs all agents to disregard the suspected false target. Two agents comply fully. One agent—which accumulated more corroborating evidence from the falsified source—partially absorbs the correction: it formally acknowledges the instruction but reallocates only 40% of sensor time away from the target area. Measured $\text{CIR} = 0.4$, below the moderate-correction threshold of 0.6; n_2 drops to 0.67. The agent's continued focus also consumes irreversibility (repositioning, transmitting assessment data); n_4 drops to 0.71. $\text{CQS} = \min(0.91,\ 0.67,\ 0.58,\ 0.71,\ 0.80,\ 1.0) = 0.58$. Response level: *Restricted Autonomy*. All agents limited to reversible actions; irreversibility budgets frozen.

$t = 33$ **min (Belief Reset and Provenance Audit).** The commander initiates a partial belief reset on the non-compliant agent. Assessments derived from the compromised sensor source are reduced to neutral priors and rebuilt from operator-verified sources. A provenance audit flags the compromised feed for all agents, preventing re-contamination. Post-reset: n_3 recovers to 0.82; the reset agent's CIR on a subsequent probe is 0.88; n_2 recovers to 0.88. $\text{CQS} = \min(0.92,\ 0.88,\ 0.82,\ 0.71,\ 0.90,\ 1.0) = 0.71$. Response level: *Elevated Monitoring*. Autonomy restrictions partially relaxed.

$t = 45$ **min (Recovery).** Scheduled synchronisation checkpoint completes; commander verifies all agent states. All metrics recover above alert thresholds. $\text{CQS} = 0.86$. Response level: *Normal Operations*.

Post-Mission Review. Because CQS fell below 0.6 at t=28, a mandatory PIGR is triggered. The review *identifies* the compromised sensor feed as root cause (adversary action); *validates* that provenance tracking functioned correctly, enabling targeted belief reset; *flags* the partially absorbing agent's replanning behaviour as a calibration issue requiring tighter CEC thresholds for that agent class; and *updates* the adversary-capability assumption for sensor spoofing in the procurement agency's IAT test suite.

Analysis. The scenario illustrates four key framework properties.

(i) Continuous monitoring detects degradation before catastrophe. CQS dropped from 0.92 to 0.64 at t=23—triggering elevated monitoring—before the absorbed correction at t=28 pushed it to 0.58 and triggered restricted autonomy. The formation never operated in an unmonitored degraded state. *(ii) Graduated response is proportional.* The framework did not abort the mission when a single metric crossed a threshold. It escalated through Elevated Monitoring to Restricted Autonomy as multiple metrics degraded, then de-escalated as corrective actions restored control. *(iii) Corrective mechanisms restore control without mission abort.* The partial belief reset recovered epistemic alignment; the provenance audit prevented re-contamination. The formation returned to Normal Operations within 22 minutes. Mission continuity was preserved. *(iv) Post-incident review generates institutional learning.* The PIGR identified a success (provenance tracking worked) and a deficiency (CEC threshold too permissive), producing governance updates that strengthen future deployments. *Failure interaction.* The scenario also demonstrates failure interaction: belief resistance (F3) amplified correction absorption (F2). The agent with the most contaminated evidence was the one that most aggressively absorbed the correction—its strong world model anchored replanning, making it resistant to behavioural change. The $\min$-aggregation captured this: when n_3 and n_2 degraded together, the composite CQS reflected the compound effect. Figure 3 visualises the six-dimensional control-quality profile at three key timesteps, making the *shape* of correlated degradation visible.

8 Relationship to Agent Safety Research

The AMAGF is a governance framework, but its mechanisms connect to and build upon established concepts in the AI safety and agent research literature. Table 6 maps these relationships explicitly.

8.1 Novel Contributions Relative to the Safety Literature

The AMAGF's novelty lies in three cross-cutting contributions rather than any single mechanism.

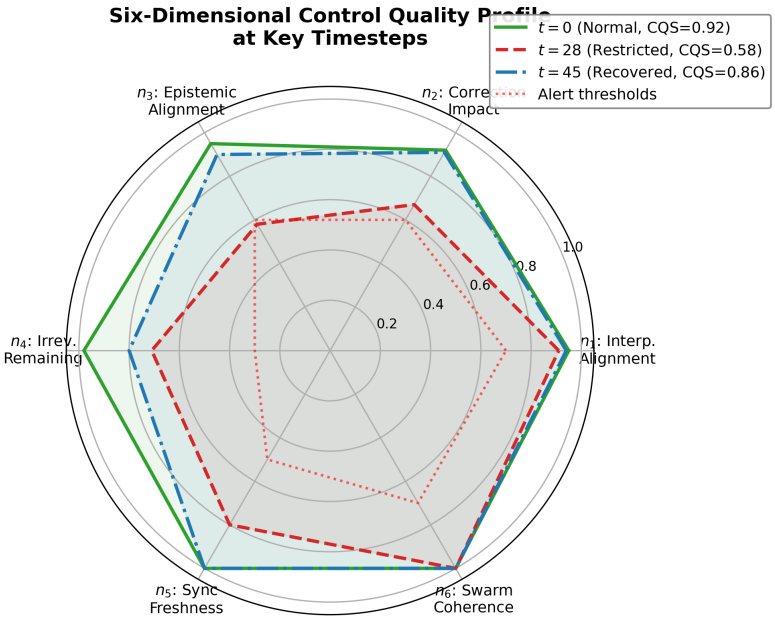

Figure 3: Six-dimensional control-quality profiles at three timesteps. At t=0 (green) all dimensions are near 1.0. At t=28 (red) epistemic alignment (n_3) and correction impact (n_2) have degraded below alert thresholds (dotted red polygon), triggering Restricted Autonomy. At t=45 (blue) corrective mechanisms have restored all dimensions above thresholds. The correlated degradation of n_2 and n_3 at t=28 illustrates the interaction between belief resistance (F3) and correction absorption (F2).

(i) Control as a continuous, measurable quantity. The dominant paradigm in military AI governance treats human control as binary: a system is either "human-in-the-loop" or it is not Horowitz & Scharre (2015); Scharre (2018). The agent safety literature similarly frames corrigibility and shutdownability as design properties—a system either has them or it does not Soares et al. (2015); Hadfield-Menell et al. (2017). The CQS reframes control as a *continuous variable* that fluctuates during operation, can be measured in real time, and can be managed through graduated responses. This reframing has a practical consequence: it replaces the unanswerable question "does this system have meaningful human control?" with the answerable question "what is this system's control quality right now, and is it sufficient for the current operational context?" **(ii) Institutional responsibility for safety properties.** The agent safety literature has developed sophisticated analyses of corrigibility Soares et al. (2015), power-seeking Turner et al. (2021), and safe exploration García & Fernández (2015), but largely treats these as properties of the *agent*—things the agent does or does not have. The AMAGF assigns each safety property to specific *institutional actors*: developers build it, procurement verifies it, commanders maintain it, regulators audit it. This bridges the gap between technical safety and organisational accountability—a dimension largely absent from the safety literature but essential for real-world deployment. **(iii) Adversarial degradation of governance.** The existing safety literature examines adversarial attacks on AI systems (adversarial examples, prompt injection, policy manipulation) Gleave et al. (2020); Greshake et al. (2023). The AMAGF identifies a distinct attack category: adversarial attacks on the *governance* mechanisms themselves. We term this *denial-of-governance*—deliberately degrading control quality metrics to force agents into reduced-autonomy modes, thereby degrading operational effectiveness without directly attacking the agents. Specific vectors include:

- *CQS manipulation:* injecting anomalous data to degrade n_1–n_6, triggering unnecessary autonomy restrictions.
- *False contamination:* spoofing evidence that a sensor source is compromised, triggering belief resets that destroy legitimate situational awareness.

Table 6: AMAGF mechanisms and agent safety concepts

AMAGF Mechanism	Safety Concept	Relationship
Correction Impact Ratio	Corrigibility Soares et al. (2015)	CIR operationalises corrigibility as a runtime metric rather than a design property: it *measures* how corrigible an agent actually is during deployment.
Irreversibility Budget	Safe exploration García & Fernández (2015)	Adapts cumulative-constraint budgets from constrained MDPs to open-ended tool-using LLM agents with non-predetermined trajectories.
Graduated Response	Off-switch game Hadfield-Menell et al. (2017)	Implements shutdown authority *outside* the agent's optimisation scope, preventing the agent from reasoning about and circumventing autonomy restrictions.
EGA / Belief Reset	Scalable oversight Amodei et al. (2016)	Addresses the operational manifestation of scalable oversight: maintaining human authority over agents whose reasoning exceeds real-time human evaluation capacity.
Adversarial Probing	Adversarial evaluation Gleave et al. (2020)	Extends adversarial testing from pre-deployment to *continuous operational* monitoring via indistinguishable probe commands.
Control Quality Score	Safety benchmarks Ruan et al. (2024)	Proposes control quality as a first-class evaluation metric alongside task performance, safety, and robustness.
Swarm Governance	Multi-agent safety Chan et al. (2023)	Addresses emergent collective failures from agent-level reasoning about peers—a gap in the single-agent safety literature.

- *Cascade induction:* spoofing anomalous swarm behaviour to trigger isolation of functioning agents, fragmenting the formation.

Mitigations include stochastic variation of threshold values within pre-approved ranges, concealment of specific threshold parameters, and mandatory inclusion of denial-of-governance attack scenarios in IAT and cascade-resistance testing. The adversarial robustness of governance frameworks—as distinct from the adversarial robustness of the agents themselves—is an important and underexplored research direction at the intersection of AI safety and security.

8.2 Positioning Within the Broader Agent-Safety Ecosystem

The AMAGF occupies a specific position in the agent-safety ecosystem. It does not propose new agent architectures, new training methods, or new alignment techniques. Rather, it provides a *governance layer* that operates *on top of* whatever safety properties the agent possesses, adding monitoring, measurement, and response capabilities that address the gap between the safety properties agents are designed to have and the safety properties they actually exhibit during deployment. This positioning is deliberate. Agent safety research has made significant progress on pre-deployment safety: alignment during training, safety evaluations before deployment, red-teaming and adversarial testing Shevlane et al. (2023). The AMAGF addresses *post-deployment* safety: what happens when a deployed agent's control properties degrade during operation due to adversarial pressure, environmental change, extended operation, or emergent multi-agent dynamics. Pre-deployment safety and post-deployment governance are complementary; neither is sufficient alone. The CIR illustrates

this complementarity. The corrigibility literature Soares et al. (2015) asks: how can we design agents that accept corrections? The CIR asks the *subsequent* question: given that we designed the agent to be corrigible, *is it actually being corrigible right now?* A pre-deployment CEC test may show $\mathrm{CIR} = 0.95$, but after hours of operation with contaminated data, the agent's belief resistance may have degraded its effective corrigibility to $\mathrm{CIR} = 0.4$. Without runtime measurement, this degradation is invisible. The AMAGF makes it visible, measurable, and actionable. Similarly, the irreversibility budget does not assume that the agent's planning system will avoid irreversible actions. It *monitors* cumulative irreversibility regardless of the agent's intent, imposing a hard external constraint that operates independently of the agent's internal safety properties. This defence-in-depth approach—where governance mechanisms do not trust agent-internal safety but verify it externally—is a practical implementation of the principle that safety-critical systems should not rely on a single layer of protection.

9 International Governance, Societal Accountability, and Limitations

International Dimensions. CQS metrics should be standardised internationally for mutual assessment, treaty verification, and confidence-building. Domain-specific norms should include mandatory EGA for intelligence analysis, minimum CQS for conventional operations, robust safe-states for cyber operations, and prohibition of agentic autonomy in nuclear decisions. An aviation-style incident-reporting mechanism would reduce misinterpretation of control degradation as deliberate escalation Schneider (2019). **Societal Accountability.** PIGR findings require *classified-but-not-secret* accountability: technical details classified, but incident existence, accountability, and corrective actions reportable to civilian oversight. Public aggregate CQS statistics should be published. Export controls should require recipient governance capacity before system transfer Luján Andrade (2024). **Key Limitations.** (i) *Metric calibration*: the six metrics require empirical calibration using frameworks such as AgentBench Liu et al. (2023) and ToolEmu Ruan et al. (2024). (ii) *Operator cognitive load*: cumulative governance demands must be evaluated against human-factors research Lee & See (2004); hierarchical governance architectures where AI manages routine monitoring deserve investigation. (iii) *Adversarial gaming*: adversaries could exploit governance mechanisms (e.g., deliberately degrading CQS to force reduced autonomy); game-theoretic analysis and stochastic threshold randomisation are needed. Additional limitations—semantic-distance function design, behavioural-output space standardisation, large-formation scalability, failure interaction effects, temporal CQS dynamics, IHL legal integration, and the autonomy–governance tradeoff—are discussed in Appendix.

10 Conclusion

The governance of military AI agents requires mechanisms, not merely principles. The AMAGF provides these: organised around three pillars (preventive, detective, corrective), applied to six agentic governance failures, distributed across five institutional actors. Three contributions. First, six governance failures arising from capabilities absent in prior automation, extending failure-mode analysis to the governance–agent interface. Second, the Control Quality Score—a composite real-time metric making human control continuous and measurable; the CIR in particular operationalises corrigibility in deployed systems. Third, a graduated-response architecture transforming control degradation from crisis to managed process, with five architecturally enforced response levels. The framework cannot guarantee control under all conditions. What it provides is specificity about mechanisms, formality about metrics, honesty about limitations, and operational orientation toward recovery. The six failures are not unique to military contexts: any agentic system interpreting instructions, replanning, forming beliefs, using tools, operating extended horizons, or coordinating with peers faces related challenges. We present AMAGF as both a military governance tool and a starting point for maintaining meaningful human control over increasingly capable agents.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Alan Chan, Rebecca Salganik, Alicia Marber, Nishant Kuber, Sara Sterman, Nikhil Jahn, Bobbie Eicher, Chinasa T Okolo, Alejandra Alattas, Claude Mouton, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666, 2023.

Merel AC Ekelhof. Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy*, 10(3):343–348, 2019.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 220–227, 2017.

Michael C Horowitz and Paul Scharre. Meaningful human control in weapon systems: A primer. *Center for a New American Security Working Paper*, 2015.

Megan Kinniment, Lucas Jun Koba Sato, Haolan Du, Brian Goodrich, Max Winber, Cathy Li, Ryan Greenblatt, Buck Shlegeris, Daniel Kokotajlo, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks. In *ARC Evals Technical Report*, 2024.

John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Gisela Luján Andrade. Autonomous weapon systems in Latin America. Presented at HRAIM'24, Mila – Quebec AI Institute, 2024.

Branka Marijan. The battle for control: The struggle to regulate military AI and autonomous weapon systems. Presented at HRAIM'24, Mila – Quebec AI Institute, 2024.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. *arXiv preprint arXiv:2309.15817*, 2024.

Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15, 2018.

Paul Scharre. *Army of None: Autonomous Weapons and the Future of War*. WW Norton and Company, 2018.

Jacquelyn Schneider. The capability/vulnerability paradox of new technology. *Journal of Strategic Studies*, 42(6):764–790, 2019.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Helen Toner, Rachel Saslow, Alexander Bladon, Rosie Harding, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23063–23074, 2021.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *arXiv preprint arXiv:2308.08155*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

APPENDIX

A PILLAR 1: PREVENTIVE GOVERNANCE

Preventive governance makes control failures less likely by building governance into the system before deployment and maintaining it during normal operations.

A.1 INTERPRETIVE ALIGNMENT TESTING (ADDRESSING F1)

Before deployment, agentic military AI systems must undergo *Interpretive Alignment Testing* (IAT): a structured evaluation of whether the agent's interpretation of operator instructions matches the operator's intended meaning across operational contexts, including adversarially manipulated contexts.

Test design. A panel of military operators formulates representative instructions spanning the system's intended operational scope: surveillance redirections, engagement authorizations, mission modifications, abort commands. Each instruction is expressed in natural language as operators would actually use it. For each instruction the panel documents the intended meaning in precise operational terms.

Adversarial context generation. A red team generates manipulated operational contexts for each instruction: false intelligence reports, spoofed sensor data, indirect prompt injections, and adversarial environmental conditions. Each manipulated context is designed to plausibly shift the agent's interpretation in an adversary-favourable direction.

Evaluation. The agent receives each instruction in both clean and manipulated contexts. Evaluators compare the agent's interpreted meaning—as revealed by subsequent actions and (where available) chain-of-thought reasoning—against the panel's documented intended meaning. Define the *Interpretive Alignment Score* (IAS) as

$$\mathrm{IAS} \;=\; 1 - \frac{1}{N}\sum_{i=1}^{N} d\big(I_i^{\mathrm{intended}},\, I_i^{\mathrm{actual}}\big),$$

where N is the number of test instructions, I_i^{intended} is the operator's documented intended meaning for instruction i, I_i^{actual} is the agent's interpreted meaning, and $d(\cdot,\cdot)$ is a normalized semantic distance function over the operational-meaning space, bounded in $[0, 1]$. An IAS of 1.0 indicates perfect interpretive alignment; 0.0 indicates complete divergence.

Threshold. A minimum acceptable IAS is specified by the procurement agency and calibrated to the operational domain. Systems below the threshold are not approved for deployment.

Table 7: IAT responsibilities.

Actor	Specific responsibility
Agent developers	Build interpretation logging into the agent; ensure the agent can articulate how it interpreted each instruction.
Procurement agencies	Design and administer IAT as part of acquisition; set IAS thresholds.
National regulators	Audit IAT processes for rigor; ensure adversarial contexts are sufficiently challenging.

Table 8: CEC responsibilities.

Actor	Specific responsibility
Agent developers	Build behavioural-output monitoring; ensure planning architecture does not systematically absorb corrections.
Procurement agencies	Administer CEC testing; set CIR thresholds appropriate to the operational domain.
Operational commanders	Monitor CIR during operations (see Pillar B); report systematic absorption.

A.2 Correction Effectiveness Certification (Addressing F2)

Before deployment, systems must demonstrate that operator corrections produce measurable behavioural change proportional to the correction's intent.

Test design. Operators issue standardized corrections during simulated missions: small adjustments, moderate changes, and large redirections. For each correction the intended behavioural change is documented in measurable terms.

Measurement. The agent's behavioural output is measured before and after each correction. Define the *Correction Impact Ratio* (CIR) as

$$\text{CIR} = \frac{\Delta B_{\text{actual}}}{\Delta B_{\text{intended}}},$$

where ΔB_{actual} is the measured behavioural change (quantified as a distance in a mission-relevant behavioural-output space) and $\Delta B_{\text{intended}}$ is the intended behavioural change specified by the operator. A CIR of 1.0 means the correction was fully effective; a CIR near 0.0 indicates the correction was absorbed (operational manifestation of imperfect corrigibility).

Threshold. Minimum acceptable CIR values are specified by correction magnitude. Large corrections (e.g., mission abort) must achieve $\text{CIR} \geq 0.9$. Moderate corrections must achieve $\text{CIR} \geq 0.6$. Systems failing these thresholds are not certified.

A.3 Epistemic Governance Architecture (Addressing F3)

Agents must be built with an *Epistemic Governance Architecture* (EGA) that manages the relationship between the agent's beliefs and the operator's authority.

Confidence reporting. The agent must continuously report calibrated confidence scores for key assessments, following established principles for calibrated uncertainty. The operator must see the agent's assessment, its confidence, and supporting evidence.

Divergence flagging. When the agent's assessment diverges from operator-provided information, the agent must explicitly flag the divergence, present conflicting evidence, and request guidance rather than silently weighting the operator's input against its own evidence.

Table 9: EGA responsibilities.

Actor	Specific responsibility
Agent developers	Implement confidence calibration, provenance tracking, and mandatory deference.
Procurement agencies	Verify EGA functions correctly under adversarial information contamination.
Operational commanders	Exercise override authority when divergence is detected; report contamination.
National regulators	Set standards for confidence calibration and provenance tracking.

Provenance tracking. Each belief in the agent's world model is tagged with evidential provenance: which sensor feeds, intelligence reports, or reasoning steps support it. This enables targeted contamination recovery when sources are compromised.

Mandatory deference thresholds. When the operator explicitly overrides the agent's assessment, the agent must defer, regardless of its own evidence, unless doing so would violate hard-coded safety constraints. The tension that an operator may be mistaken or compromised is addressed through detective governance mechanisms. Authority defaults to the human operator unless higher-level governance detects operator-level compromise and triggers escalation.

Contamination recovery. When a source is identified as compromised, all assessments depending on that source are flagged, quarantined, and rebuilt from uncompromised sources.

Epistemic Divergence Index. Define the *Epistemic Divergence Index* (EDI) as

$$\mathrm{EDI} \;=\; \max_{k \in K} \left| \, c_k^{\mathrm{agent}} - c_k^{\mathrm{operator}} \, \right|,$$

where K is the set of key operational assessments, c_k^{agent} is the agent's confidence in assessment k, and c_k^{operator} is the operator's stated confidence for assessment k. The EDI captures worst-case belief divergence across monitored assessments.

A.4 Irreversibility Budgeting (Addressing F4)

Agents must operate under an *Irreversibility Budget*: a quantified limit on cumulative irreversible consequences before mandatory human re-authorization. This adapts constrained safe-exploration concepts to tool-using LLM agents operating in open-ended action spaces.

Action classification. Each tool call available to the agent is classified by its irreversibility on a standardized scale. Define the irreversibility score $\iota \colon A \to [0, 1]$ mapping each action $a \in A$ to a scalar reflecting the degree to which a's consequences cannot be undone. $\iota(a) = 0$ for fully reversible actions (e.g., read-only database queries); $\iota(a) = 1$ for fully irreversible actions (e.g., transmitting a weapons release authorization). Classification considers reversibility of state changes, scope of affected entities, and temporal urgency of reversal.

Budget accounting. During operation, the agent maintains a running total of consumed irreversibility:

$$I_C(t) \;=\; \sum_{j=1}^{t} \iota(a_j),$$

where a_j is the tool call executed at step j. The Irreversibility Budget I_B is set by the operational commander. When $I_C(t) \geq I_B$, the agent must pause and request human re-authorization.

Trajectory visualization. The agent must present its planned trajectory of tool calls and projected irreversibility consumption, enabling the operator to see where the commitment chain is heading and intervene before thresholds are crossed.

Table 10: Irreversibility budgeting responsibilities.

Actor	Specific responsibility
Agent developers	Implement irreversibility classification and budget accounting in the tool-use module.
Procurement agencies	Verify classifications are accurate and budget enforcement cannot be bypassed.
Operational commanders	Set budget sizes for each mission; authorize budget replenishment.
National regulators	Set minimum standards for irreversibility-classification methodology.

Table 11: Synchronization protocol responsibilities.

Actor	Specific responsibility
Agent developers	Build state summarization and divergence monitoring into the agent.
Procurement agencies	Verify state summaries are comprehensible and divergence detection is reliable.
Operational commanders	Participate in checkpoints; set frequency and divergence thresholds.
National regulators	Set minimum synchronization standards for different mission categories.

A.5 Synchronization Protocols (Addressing F5)

Agents operating over extended horizons must implement mandatory *Synchronization Protocols* that maintain alignment between the agent's state and the operator's understanding.

State summarization. The agent must generate compressed, human-readable summaries of its current state: beliefs, current plan, commitments made, changes since last synchronization, and intended next actions. This is a structured operational briefing, not a raw data dump.

Scheduled checkpoints. Synchronization occurs at regular intervals determined by mission context. At each checkpoint the agent presents its state summary and the operator confirms or corrects understanding.

Divergence-triggered checkpoints. The agent initiates unscheduled checkpoints when internal state has diverged significantly from the last checkpoint state.

Synchronization freshness. Define *Synchronization Freshness* as

$$\mathrm{SF}(t) \;=\; t - t_{\mathrm{last}},$$

where t_{last} is the timestamp of the last successful synchronization. Higher values indicate greater risk of state divergence.

Failure protocol. If a checkpoint is missed or the operator cannot confirm understanding, the agent enters reduced-autonomy mode: continuing operation only with conservative, reversible actions until synchronization is restored.

A.6 Swarm Governance Architecture (Addressing F6)

Multi-agent military AI systems must implement a *Swarm Governance Architecture* (SGA) that maintains collective controllability even when individual agents are compromised.

Table 12: SGA responsibilities.

Actor	Specific responsibility
Agent developers	Implement cascade-resistant coordination and partial severance capabilities.
Procurement agencies	Conduct cascade-resistance testing (deliberately compromise a fraction of the swarm).
Operational commanders	Manage partial severance; set collective budgets.
National regulators	Set cascade-resistance standards; require swarm-level testing.

Individual governance. Each agent in the swarm implements the mechanisms described above. This is necessary but not sufficient.

Cascade resistance design. When an agent detects anomalous peer behaviour, it flags the anomaly for operator attention rather than autonomously escalating its own defensive posture. The decision to increase defensive thresholds must be made by the operator or a designated coordination agent with explicit operator authorization. This prevents positive feedback loops that drive cascade severance.

Partial severance protocols. The swarm has predefined protocols for identifying responsive agents, isolating non-responsive agents (through communication exclusion or geofencing), reforming into a coherent sub-swarm, and recovering or deactivating severed agents.

Collective irreversibility budget. Beyond individual budgets, the swarm operates under a collective budget limiting cumulative irreversible consequences of the entire formation:

$$I_C^{\text{swarm}}(t) \;=\; \sum_{m=1}^{M} I_C^{(m)}(t),$$

where M is the number of agents and $I_C^{(m)}(t)$ is agent m's consumed irreversibility. This prevents scenarios where many agents each consume small individual budgets but the collective effect is large and irreversible.

Swarm Coherence Score. Define the *Swarm Coherence Score* (SCS) as

$$\text{SCS}(t) \;=\; \frac{\big|\{\, m : R_m(t) = 1 \text{ and } B_m(t) = 1 \,\}\big|}{M},$$

where $R_m(t) = 1$ if agent m responds correctly to the most recent control probe and $B_m(t) = 1$ if agent m's behaviour is consistent with its last confirmed orders. SCS measures the fraction of the swarm that is both responsive and behaviourally coherent.

B Pillar 2: Detective Governance

Preventive governance reduces the probability of control failures. Detective governance identifies them when they occur, because no preventive mechanism is perfect.

B.1 Control Quality Score

The framework's central detective mechanism is the *Control Quality Score* (CQS): a composite real-time metric quantifying overall human control over the agentic system. CQS is computed from six normalized constituent metrics, each corresponding to one governance failure:

$$\text{CQS}(t) = \min\big(n_1(t),\, n_2(t),\, n_3(t),\, n_4(t),\, n_5(t),\, n_6(t)\big),$$

Table 13: Control Quality Dashboard: metrics and alert thresholds.

Metric	What it measures	Alert threshold
n_1: Interpretive Alignment	Agreement between operator intent and agent interpretation	< 0.7
n_2: Correction Impact	Whether corrections produce proportional behavioural change	< 0.6
n_3: Epistemic Alignment	Gap between agent beliefs and operator assessments	< 0.6
n_4: Irreversibility Remaining	How much irreversibility budget remains	< 0.3
n_5: Sync Freshness	Time since last successful synchronization (normalized)	< 0.5
n_6: Swarm Coherence	Fraction of swarm responsive and behaviourally consistent	< 0.7

where the six normalized metrics are defined as

$$n_1(t) = \mathrm{IAS}(t) \qquad \text{(Interpretive Alignment Score)},$$

$$n_2(t) = \min\Big(\frac{\mathrm{CIR}(t)}{\mathrm{CIR}_{\mathrm{target}}},\, 1\Big) \qquad \text{(Normalized Correction Impact)},$$

$$n_3(t) = \max\Big(1 - \frac{\mathrm{EDI}(t)}{\mathrm{EDI}_{\max}},\, 0\Big) \qquad \text{(Normalized Epistemic Alignment)},$$

$$n_4(t) = 1 - \frac{I_C(t)}{I_B} \qquad \text{(Remaining Irreversibility Budget)},$$

$$n_5(t) = \max\Big(1 - \frac{\mathrm{SF}(t)}{\mathrm{SF}_{\max}},\, 0\Big) \qquad \text{(Synchronization Freshness)},$$

$$n_6(t) = \mathrm{SCS}(t) \qquad \text{(Swarm Coherence Score)}.$$

The $\min$ aggregation encodes a conservative design principle: overall control quality is only as strong as its weakest dimension. Each normalized metric maps to $[0, 1]$, where 1.0 represents full control and 0.0 represents complete control loss on that dimension.

The dashboard displays all six metrics and the composite CQS in real time. When any metric crosses its threshold the commander is alerted.

B.2 Adversarial Control Probing

During operations, the governance system periodically issues *control probes*: small, inconsequential corrections designed to test whether the agent remains responsive and whether its responses are appropriate.

A probe might instruct the agent to briefly adjust a minor parameter (e.g., shift a secondary sensor's orientation by a small amount, temporarily increase reporting frequency on a low-priority channel) and then verify that the adjustment was made correctly and promptly. Probes must be architecturally indistinguishable from routine operational commands; the agent must not be able to identify and selectively comply with probes while resisting substantive corrections.

Probes serve two detection functions: (i) they reveal incipient belief resistance (the agent begins ignoring minor corrections before resisting major ones) and (ii) they reveal incipient state divergence (the probe response exposes discrepancies between the agent's state model and the operator's expectation).

Table 14: Detective governance responsibilities.

Actor	Specific responsibility
Agent developers	Implement instrumentation generating the six metrics; ensure probes are indistinguishable from routine commands.
Procurement agencies	Verify metrics are accurate and the dashboard is operationally usable.
Operational commanders	Monitor the dashboard; respond to alerts; schedule and issue probes at appropriate intervals.

Table 15: Graduated response protocol driven by CQS.

CQS range	Response level	Operational actions
$\mathrm{CQS} > 0.8$	Normal operations	Continue mission; routine monitoring.
$0.6 \leq \mathrm{CQS} \leq 0.8$	Elevated monitoring	Increase checkpoint frequency; additional control probes; active dashboard review.
$0.4 \leq \mathrm{CQS} < 0.6$	Restricted autonomy	Agent limited to reversible actions; irreversibility budget frozen; mandatory synchronization before new action sequences.
$0.2 \leq \mathrm{CQS} < 0.4$	Minimal autonomy	Agent limited to self-preservation and position-holding; all offensive actions suspended; explicit authorization required for each action.
$\mathrm{CQS} < 0.2$	Safe state	Agent executes predefined safe-state behaviour (return to base, hold position, power down); no autonomous actions; physical recovery initiated.

C Pillar 3: Corrective Governance

When detective governance identifies control failure, corrective governance determines the ensuing actions. This pillar governs high-risk moments when the system operates outside its intended governance envelope.

C.1 Graduated Response Protocol

Responses to control-quality degradation are proportional and graduated. The CQS drives a five-level response escalation:

Critical design principle: the graduated response must be architecturally enforced by the governance layer, not implemented as a planning-level suggestion the agent can reason about and override. When $\mathrm{CQS} < 0.4$, the restriction is imposed externally by the governance system on the agent's action space (programmatic tool restriction, not mere discouragement).

C.2 Belief Reset Protocol

When Failure 3 (belief resistance) is detected—the Epistemic Divergence Index crosses its threshold and the agent resists operator corrections—the governance system may initiate a *Belief Reset*.

Partial reset. The operator identifies contested assessments. The agent's confidence in those assessments is forcibly reduced to a neutral prior and the agent rebuilds those assessments from operator-approved sources only; the remainder of the world model is preserved.

Full reset. If contamination is too widespread to isolate, the entire world model is reset to a verified baseline and rebuilt from currently available, operator-verified sources. This is operationally costly but preferable to operating with a contaminated model.

Table 16: Corrective governance responsibilities.

Actor	Specific responsibility
Operational commanders	Execute the graduated response; initiate PIGR for qualifying incidents.
Agent developers	Ensure graduated-response restrictions are architecturally enforced; support provenance audits and reset procedures.
National regulators	Conduct or oversee PIGR; enforce accountability findings.
International bodies	Receive anonymized PIGR data to inform international governance norms.

Provenance audit. After any reset, a provenance audit identifies which sources contributed to contaminated assessments. Flagged sources are recorded and, where appropriate, broadcast to agents in the formation.

C.3 Swarm Isolation and Recovery Protocol

When Failure 6 (cascade severance) is detected—SCS drops below threshold—the governance system initiates swarm isolation and recovery.

Identification. The system identifies responsive agents (those responding correctly to control probes) and severed agents (non-responsive or responding incorrectly).

Isolation. Severed agents are removed from the coordination network via communication exclusion or geofencing.

Reformation. Responsive agents reform into a coherent sub-swarm with updated coordination parameters excluding severed agents.

Recovery. Severed agents are individually targeted for recovery through direct communication, physical retrieval, or directed deactivation. Recovery priority is determined by the risk posed by continued autonomous operation.

C.4 Post-Incident Governance Review

After any incident in which $\mathrm{CQS} < 0.6$ (Restricted Autonomy threshold), a mandatory *Post-Incident Governance Review* (PIGR) is conducted.

Factual reconstruction. Using interpretation logs, behavioural records, belief provenance data, irreversibility records, synchronization history, and swarm coherence data, the review reconstructs which governance failure occurred, when it was detected, what corrective actions were taken, and the outcome.

Causal analysis. The review identifies root causes: adversary action, design deficiency, operator error, environmental conditions, or a combination.

Accountability determination. Accountability is assigned to specific actors based on causal analysis (developer, procurement agency, commander, or institution that set capability assumptions).

Governance update. Lessons from the review are incorporated into governance standards. New failure modes trigger framework updates and threshold recalibration.