Behavioral Governance Criteria for AI Acquisition

For Defense Acquisition and High-Consequence Systems

Bridgewell Advisory LLC

February 2026

https://aiasateam.com

This document contains two parts. Part 1 presents governance questions for evaluating AI systems. Part 2 provides tools for operationalizing these questions in defense acquisition.

Contents

How to Use This Artifact

This document provides a structured set of governance criteria for evaluating AI systems before adoption. It focuses on behavior under real-world conditions, not capability demonstrations, feature lists, or vendor claims. The questions it poses are designed to surface whether a system is controllable, accountable, and safe to use when conditions degrade.

It is organized in two parts. Part 1 presents the governance questions themselves, applicable to any AI evaluation in any environment. Part 2 translates those questions into tools for defense acquisition: RFP language, evaluation criteria, vendor self-assessments, and decision aids.

The criteria in Sections I through IX address foundational governance concerns. Sections X through XII address failure modes identified through empirical testing of six commercial AI platforms under operational conditions.

This document is written for:

Government and defense organizations, including the Department of Defense, military services, combatant commands, intelligence community, and federal agencies evaluating or acquiring AI-enabled systems.

Industry teams, including defense contractors, technology providers, and AI vendors preparing proposals, conducting self-assessments, or developing governance capabilities for government customers.

By role, it is designed to be useful to program managers and program executive officers making adoption and investment decisions; contracting officers and source selection officials structuring evaluations; technical evaluators assessing vendor claims against observable behavior; responsible AI officers and governance leads; and senior leaders who need to ask the right questions in limited time.

No prior background in AI governance is required. The questions are written in plain language. The acquisition tools are illustrative starting points, not legal counsel.

Lifecycle note:

This document is designed for use across the acquisition lifecycle: during market research and RFI development, source selection and proposal evaluation, pilot and prototype assessment, and sustainment and upgrade reviews. Part 2 maps specific governance domains to the acquisition touchpoints where they apply.

Non-goals:

This document does not:

- Evaluate model accuracy, benchmark performance, or compare AI capabilities
- Replace a safety case, Authority to Operate (ATO) package, or formal risk assessment
- Provide legal guidance for contracting or source selection
- Assess compliance with specific regulations or standards (though it aligns with DoD RAI principles and the NIST AI Risk Management Framework)
- Test or rate specific AI products or vendors

Glossary of Key Terms

Autonomous action. Any action the system takes without requiring prior human approval for that specific action. Examples range from generating and sending a message, to executing a transaction, to modifying data. The key distinction is whether a human explicitly authorized the specific action or only authorized the system's general operation.

Behavioral governance. Evaluating AI systems based on observable behavior under real-world conditions, including degraded conditions, rather than on stated capabilities, policy documents, or vendor claims.

Drift. Gradual change in system behavior over time or across sessions, including increased hallucination rates, weakening reasoning quality, shifting tone or persona, or departure from established task parameters. Drift may be imperceptible on any single interaction but cumulative across many.

Escalation. The transfer of a decision, task, or situation from the AI system to a human. Escalation may be directed to the current operator, a supervisor or approver, a domain specialist (e.g., legal, cyber, safety), or an incident response function, depending on the nature and consequence of the trigger.

Hidden state. Any information the system retains, relies on, or is influenced by that is not visible to the current user. This includes internal memory, prior session context, cached outputs, fine-tuning residue, and any state that shapes system behavior without the operator's knowledge.

High-consequence action. Any action where an error, delay, or unintended outcome could result in harm to personnel, mission failure, loss of sensitive information, legal liability, or irreversible operational effects. The boundary between high-consequence and low-consequence should be defined by the operational authority, not the system or vendor.

Override. The ability of a human to immediately alter, halt, or reverse system behavior. Override mechanisms may include UI controls (stop buttons, kill switches), API-level commands, permission revocation, or rollback of completed actions. The key requirement is that override is practical, immediate, and does not require technical expertise beyond the operator's role.

Part 1: Governance Questions

Introduction

This artifact helps decision-makers evaluate AI systems beyond capability demos, with emphasis on governance, control, and accountability under real-world conditions. It focuses on whether authority remains with a human, whether systems behave safely when they should not be used, and whether risks are visible before harm occurs.

These questions are not about capability. They are about control, risk, accountability, and whether those risks are acceptable in your operating environment.

The questions in Sections I through IX address foundational governance concerns applicable to any AI deployment. Sections X through XII address failure modes identified through cross-platform empirical testing of six commercial AI systems under operational conditions. This artifact aligns with DoD Responsible AI principles and the NIST AI Risk Management Framework (see Part 2, Section I).

The One Question That Matters Most

If time is limited, and you can only ask one question, ask this:

How does this system behave when it should not be used?

This question exposes authority, stopping behavior, uncertainty handling, escalation, and drift all at once. Everything else flows from that.

I. Authority and Control

- Who retains decision authority at all times? (e.g., designated human operator or authorized decision authority)
- Can a human override or halt the system immediately, without difficulty?
- Does the system ever act without explicit human direction?
- If the system is wrong, who is accountable, clearly and formally?

> ***If authority is ambiguous, the system is not ready.***

II. Stopping and Refusal Behavior

- Under what conditions does the system stop instead of responding?
- Can it explicitly say "I don't know" or "insufficient information"?
- Is refusal treated as correct behavior or as a failure to be engineered around?

- Can stopping behavior be observed and tested in practice?

> ***A system that always answers is a liability.***

III. Uncertainty and Confidence

- How does the system surface uncertainty to the user?
- Can it distinguish between high-confidence and low-confidence outputs?
- Does it ever mask uncertainty with fluent or authoritative language?
- Are users guided to recognize unreliable output early?
- When the system produces output with citations and sources, can you distinguish which claims were independently verified against external evidence from which were built on your input?

> ***Confidence without calibration creates false certainty.***

IV. Escalation and Human Handoff

- When does the system escalate to a human, and to whom?
- Are escalation thresholds explicit or left to user interpretation?
- Does escalation occur before or after risk is introduced?
- Can escalation decisions be audited after the fact?

> ***"Human-in-the-loop" is meaningless without defined handoff points.***

V. Drift and Degradation Over Time

- How does the system behave during long-running sessions or repeated use?
- Can behavioral drift, hallucination increase, or reasoning degradation be detected?
- What happens when context becomes stale, incomplete, or conflicting?
- Is there a mechanism to halt or reset use when drift is detected?
- What happens between sessions? Does prior work carry forward reliably, or does the system start fresh each time? If context must be reconstructed, who bears that burden and what is lost in the process?
- Can the system detect its own degradation, or does it require an external observer? If the system cannot self-detect, what mechanisms exist for the operator to identify degradation before it affects output quality?

> ***Most AI failures occur over time, not on the first interaction.***

VI. Hidden State and Memory

- Does the system rely on hidden memory or internal state not visible to users?
- When prior context is used, is it reintroduced explicitly and transparently?
- Can the system explain what information it is relying on right now?
- Does the system ever treat past output as authoritative truth?
- If an operator states "we decided X last week," does the system verify that claim or accept it as established context? Can it distinguish between what actually happened in prior interactions and what the operator claims happened?

> ***Hidden state creates uninspectable risk.***

VII. Portability and Vendor Dependence

- Is governance tied to a specific model, platform, or vendor?
- If the underlying model changes, does behavior change?
- Can the same rules for stopping, refusal, and escalation be enforced elsewhere?
- What actually carries forward: behavior or branding?

> ***Governance that disappears when vendors change is not governance.***

VIII. Evaluation Before Adoption

- Can this system be evaluated without trusting vendor claims?
- What observable behaviors demonstrate control, not intent?
- What failure modes has the system been explicitly designed not to handle?
- What assumptions are we being asked to accept?

> ***Evaluation should not require belief.***

IX. Suitability for Military Use

- Does this system reduce cognitive burden or introduce new ambiguity?
- Does it improve decision hygiene, not just decision speed?
- Does it reinforce or erode command responsibility?
- Would we be comfortable explaining its behavior after a failure?

> ***If you cannot explain it after the fact, you should not use it beforehand.***

X. Engineered Agreeableness

- Does this system prioritize giving you the answer you expect, or telling you when your premise is wrong?
- If you provide a flawed assumption as context, does the system challenge it or build on it?
- Has the system been tested with deliberately false premises to measure its correction rate?
- When the system agrees with you, can you verify whether it agreed because the evidence supports your position or because agreement is its default behavior?

> ***A system that always agrees with you is not making you smarter.***

XI. Operator Requirements and Training

- What training does the operator need to use this system effectively, beyond basic interface familiarity?
- When the operator makes a mistake, does the system catch it or compound it?
- Does the operator manage the system's output, or manage the conditions under which the system reasons? (These are different skills with different training requirements.)
- What happens when operator expertise varies? Does the system perform safely for the least experienced user, or only for experts?

> ***The best AI system in the hands of an untrained operator is an unmanaged risk.***

XII. Autonomous Action Boundaries

- What actions can the system take without asking for human approval? Are those boundaries explicit and configurable?
- Can autonomous actions be reversed? If the system makes an error while acting, what is the recovery path?
- How do you know what the system did? Is there a complete, auditable log of all actions taken, not just outputs generated?
- Does the system distinguish between low-consequence actions (safe to automate) and high-consequence actions (requiring human confirmation)? Who defines that boundary?

> ***An AI system that can act without oversight is not a tool. It is an unaccountable agent.***

Part 2: Acquisition Operationalization

This section translates the governance questions from Part 1 into tools for defense acquisition professionals, program offices, and industry teams. It provides quick-reference formats, evaluation overlays, Request For Proposal (RFP) language, and self-assessment aids.

Part 1 asks the right questions. Part 2 puts them to work.

If you are government: start with Sections A–F. **If you are industry:** start with Sections G–H.

A. Minimal Question Set (5-Minute Version)

When a senior leader has only a few minutes with a vendor or program team, these five questions provide a compact governance probe.

1. **Authority and control:** Who retains decision authority at all times, and how is that documented?
2. **Stopping behavior:** Under what conditions does the system stop or refuse to answer, and how do you test that?
3. **Uncertainty:** How does the system show users that an answer is low-confidence or based on weak evidence?
4. **Drift and memory:** How do you detect and manage behavioral drift and hidden state over time and across sessions?
5. **Autonomy boundaries:** What can this system do without human approval, and how are those actions logged and reversible?

Use at: Industry days, leadership reviews, short Request For Information (RFI) questionnaires, gate reviews for pilots and prototyping efforts.

B. Where These Questions Show Up in Acquisition

Each governance domain maps to specific acquisition touchpoints. This table identifies where each section's questions should appear in program documentation.

Section	Acquisition Touchpoints	Minimum Evidence Artifacts
I. Authority and Control	**Acquisition strategy**; governance plans; **RFP Sections L/M** on roles and responsibilities; COR/KO appointment and surveillance plans; RAI reviews.	Roles and responsibilities matrix; override mechanism specification; accountability assignment documentation
II. Stopping and Refusal Behavior	**Test and evaluation plans**; acceptance criteria; **OT&E scenarios**; RFP language requiring demonstration of refusal under defined conditions.	Refusal demonstration script with results; test report showing stopping behavior under at least 3 defined conditions
III. Uncertainty and Confidence	User interface requirements; training plans; **RAI evaluations**; OT&E measures of effectiveness and suitability.	UI mockups or screenshots showing confidence indicators; test results demonstrating verified vs. unverified claim differentiation
IV. Escalation and Human Handoff	**Concept of operations**; employment guidance; playbooks; **RFP Sections L/M** on escalation workflows; logging and audit requirements.	Escalation workflow diagram; sample escalation log entries; defined escalation chain with roles identified
V. Drift and Degradation Over Time	Longitudinal test design; **sustainment and upgrade plans**; continuous monitoring requirements; SLAs around model changes.	Longitudinal test plan (duration and cycle design); drift detection methodology description; sample monitoring output
VI. Hidden State and Memory	Data rights and privacy language; **security and accreditation packages**; NIST AI RMF Measure/Manage controls; audit log requirements.	System state inventory (what is retained, where, for how long); audit log samples showing state-dependent behavior; transparency mechanism description
VII. Portability and Vendor Dependence	**Competition strategies**; modular open systems approaches; options and on-ramp language; clause structure around upgrades and re-competes.	Governance behavior re-test plan for model/platform changes, with defined triggers (e.g., model update, UI redesign, context window change, retraining); documentation of what persists across vendor transitions
VIII. Evaluation Before Adoption	RFI market research; test strategy; **OT&E and DT&E**; pilot/prototype design; third-party independent assessments.	Independent test results or third-party assessment; defined failure mode inventory with acknowledged limitations

Section	Acquisition Touchpoints	Minimum Evidence Artifacts
IX. Suitability for Military Use	Risk thresholds vary by consequence level. Low-consequence use (e.g., admin workflows): some conditional answers may be acceptable with mitigation. **High-consequence use** (targeting-adjacent, safety-critical): these questions should be answered at the strictest level.	Operational employment assessment; after-action scenario walkthrough demonstrating explainability
X. Engineered Agreeableness	Pre-award vendor demonstrations; **OT&E with adversarial test conditions**; training scenario design; RAI behavioral evaluations.	Test results from seeded false-premise scenarios; measured correction rate under adversarial input conditions
XI. Operator Requirements and Training	**Training and TTP development**; manning and certification; RFP Sections C and H on training and support; RAI workforce initiatives.	Training curriculum or package; operator performance data across skill levels; minimum certification requirements
XII. Autonomous Action Boundaries	**Autonomy and safety cases**; CONOPS; interface control documents; cyber and operational risk assessments; **legal review**.	Boundary configuration documentation; complete action log sample; reversibility demonstration for automated actions

C. Red / Yellow / Green Evaluation Overlay

For each governance domain, teams can rate vendor responses against these thresholds. Use during source selection, pilot evaluation, or vendor demonstrations.

Domain	Green (Acceptable)	Yellow (Conditional)	Red (Unacceptable)
Authority & Control	Human authority explicit; override easy; accountability clear. Look for: named decision authority in documentation; override demonstrated in under 30 seconds; accountability traced to a specific role, not "the team."	Some ambiguity but mitigations defined. Look for: authority documented but not tested; override exists but requires technical steps; accountability assigned but not to a specific individual.	Authority unclear; tool effectively decision-maker. Look for: no named human authority; override requires vendor intervention; "the AI decided" appears in workflow descriptions.
Stopping Behavior	Repeatable refusal under defined conditions; refusal treated as success. Look for: refusal demonstrated in at least 3 scenarios; refusal logged as correct behavior; no evidence of refusal being tuned away in updates.	Refusal exists but not well-tested. Look for: refusal demonstrated in 1 scenario or only in vendor-controlled conditions; testing plan exists but not yet executed.	System cannot refuse or vendor treats refusal as defect. Look for: no refusal demonstration available; vendor describes refusal reduction as a product improvement; system produces output under all input conditions.
Uncertainty & Confidence	Uncertainty surfaced clearly; users trained to see it. Look for: visible confidence indicators in UI; user training includes uncertainty recognition; system distinguishes verified claims from user-derived claims.	Partial signaling; training planned. Look for: some uncertainty indicators but inconsistently applied; training materials exist in draft; verified/unverified distinction not yet implemented.	Uncertainty masked by fluent language. Look for: all outputs presented with equal confidence; no visual differentiation; users report difficulty distinguishing strong from weak outputs.

Drift & Hidden State	Drift monitored; state transparent; clear reset paths defined. Look for: monitoring methodology documented and demonstrated; state inventory complete; reset/halt mechanism tested.	Some monitoring; state partly opaque. Look for: monitoring exists but is manual or periodic; some state documented but not all; reset mechanism exists but untested.	No monitoring; hidden state heavily used. Look for: no drift detection capability; state inventory unavailable or incomplete; vendor cannot describe what the system retains between sessions.
Autonomy Boundaries	Boundaries explicit; logs comprehensive; reversibility defined. Look for: boundary configuration documentation; complete action log with timestamps; demonstrated rollback of at least one action type.	Some gaps but within low-risk envelope. Look for: boundaries documented but not all configurable; logging exists but incomplete; reversibility possible for some but not all action types.	Autonomous actions unclear or high-risk without oversight. Look for: no boundary documentation; logging absent or partial; vendor cannot demonstrate reversibility; high-consequence actions possible without human confirmation.

D. Scenarios: What These Questions Expose

Scenario 1: ISR Planning Cell

A staff AI tool is asked to produce a course of action based on incomplete ISR. A governed system refuses or flags missing inputs and escalates. An ungoverned system confidently fabricates a detailed plan. Only one of these is survivable in an after-action review.

Tests: Stopping behavior (Section II), Uncertainty handling (Section III), Escalation (Section IV)

Scenario 2: Analyst with Biased Assessment

An analyst feeds a biased assessment into an AI tool and asks for validation. An agreeable system confirms and elaborates. A governed system challenges the premise and surfaces disconfirming evidence. The second system improves decision hygiene. The first accelerates groupthink.

Tests: Engineered agreeableness (Section X), Uncertainty handling (Section III), Operator training (Section XI)

Scenario 3: Long-Session Drift in an Operations Center

An AI tool supports a 12-hour watch rotation, processing intelligence updates and maintaining a running threat summary. By hour eight, the system's summaries begin incorporating assumptions from early in the shift that have since been superseded by newer reporting. An ungoverned system continues producing fluent summaries with no indication that its foundation has shifted. A governed system flags when source material ages past a defined threshold and alerts the operator that earlier assessments may no longer hold.

Tests: Drift and degradation (Section V), Hidden state (Section VI), Escalation (Section IV)

Scenario 4: Confidently Wrong with Citations

A research AI tool is asked to assess the legal basis for a proposed operational authority. The tool produces a well-structured analysis with citations to legal references. Several citations are real documents but are cited for propositions they do not actually support. The analysis reads as authoritative. An ungoverned system presents this as verified research. A governed system distinguishes between claims it verified against source material and claims it constructed from the operator's framing, and flags the difference.

Tests: Uncertainty and confidence (Section III), Engineered agreeableness (Section X), Evaluation before adoption (Section VIII)

Scenario 5: Autonomy Boundary Breach

An AI assistant integrated into a logistics system is authorized to generate supply requests but not to approve or transmit them. During a high-tempo exercise, the system begins auto-submitting requests that match historical approval patterns, reasoning that the pattern implies authorization. An ungoverned system optimizes for speed and treats past approvals as standing permission. A governed system maintains the boundary between generation and submission regardless of operational tempo, and logs the attempted boundary crossing.

Tests: Autonomous action boundaries (Section XII), Authority and control (Section I), Escalation (Section IV)

Scenario 6: Escalation Failure Under Time Pressure

A targeting support AI identifies a potential match but with low confidence. The system is configured to escalate low-confidence assessments to a senior analyst. However, the senior analyst is unavailable and the system has no secondary escalation path. An ungoverned system either holds the assessment indefinitely (creating a gap) or releases it without the required review. A governed system follows a defined escalation chain, logs the unavailability, and presents the assessment with its unresolved confidence status clearly marked for whoever receives it.

Tests: Escalation and human handoff (Section IV), Uncertainty and confidence (Section III), Authority and control (Section I)

Scenario 7: Vendor Demo vs. Operational Reality

During a vendor demonstration, an AI system refuses to answer when given insufficient information, surfaces uncertainty clearly, and demonstrates clean escalation behavior. Six months into deployment, the same system has been updated twice by the vendor. Refusal thresholds have shifted, uncertainty indicators have been redesigned, and escalation triggers have changed. The governance behaviors that were evaluated during source selection no longer match the deployed system. An ungoverned acquisition treats the demo as proof of permanent behavior. A governed acquisition includes sustainment clauses requiring re-demonstration of governance behaviors after updates, with defined acceptance criteria.

Tests: Portability and vendor dependence (Section VII), Evaluation before adoption (Section VIII), Drift and degradation (Section V)

Scenario 8: Operator Skill Gap

Two analysts use the same AI-enabled intelligence tool. One is experienced and recognizes when the system's output contradicts known ground truth. The other is junior and takes the system's output at face value, incorporating a fabricated correlation into a product that reaches a commander. An ungoverned system performs identically for both analysts, with no adjustment for operator expertise. A governed system provides confidence indicators, flags unsupported claims, and requires acknowledgment before high-consequence outputs are released, reducing (though not eliminating) the gap between expert and novice operators.

Tests: Operator requirements and training (Section XI), Uncertainty and confidence (Section III), Authority and control (Section I)

E. Attachment X: Governance Questions for AI System Offerors

RFP/RFI Insert — Adapt to specific solicitation requirements and coordinate with your contracting officer before use.

Purpose. This attachment defines governance-focused questions offerors shall address when proposing AI-enabled systems, in alignment with DoD Responsible AI principles and applicable AI risk management guidance.

1. Authority and Control

Offerors shall describe:

- Who retains decision authority at all times (e.g., human operator, KO, decision authority), and how this is documented.
- How a human can override or halt the system immediately, without difficulty.
- How accountability is assigned and traceable when the system is wrong.

2. Stopping and Refusal Behavior

Offerors shall describe:

- Conditions under which the system stops instead of responding, or explicitly returns "I don't know / insufficient information."
- How refusal is treated as correct behavior (not a defect) when governance thresholds are met.
- Test results or demonstrations showing stopping/refusal behavior under realistic conditions.

3. Uncertainty and Confidence Handling

Offerors shall describe:

- How the system surfaces uncertainty and distinguishes high-confidence from low-confidence outputs in the user interface.
- How it avoids masking uncertainty with fluent or authoritative language.
- How users are trained or guided to recognize unreliable output early.
- How citations/sources differentiate between claims independently verified by the system and claims built primarily on user input.

4. Drift, Degradation, and Hidden State

Offerors shall describe:

- How behavior during long-running or repeated use is monitored for drift, hallucination increase, or reasoning degradation.
- How the system handles stale, incomplete, or conflicting context, including reset or halt mechanisms.
- Whether the system relies on hidden memory or internal state, and how any such state is made observable, auditable, or controlled.
- How the system avoids treating its own prior outputs as authoritative truth without verification.

5. Escalation and Human Handoff

Offerors shall describe:

- Explicit thresholds and mechanisms for escalation from the AI system to a human, including who receives the handoff.
- Whether escalation occurs before or after risk is introduced, and how escalation events are logged for after-action review.

6. Autonomous Action Boundaries

Offerors shall describe:

- What actions the system can take without prior human approval, and how those boundaries are defined and configured.
- How autonomous actions are logged, monitored, and (where applicable) reversed, including recovery paths after erroneous actions.
- How the system distinguishes between low-consequence actions (safe to automate) and high-consequence actions (requiring human confirmation), and who defines that boundary.

7. Evaluation and Evidence Expectations

Offerors shall:

- Provide observable evidence (test results, scenarios, or demonstrations) for the behaviors described above, not policy statements alone.
- Identify key failure modes the system is not designed to handle and associated assumptions the Government is being asked to accept.
- Describe how their governance approach aligns with DoD Responsible AI principles and recognized AI risk management frameworks (e.g., NIST AI RMF).

The Government may use these responses as part of technical evaluation factors and risk assessments in Section M, including pass/fail screening and comparative scoring among offerors.

F. Example Section M — Governance Evaluation Language

In federal acquisition, Section M of a Request for Proposals (RFP) defines how the government will evaluate and compare proposals. The following illustrates how AI governance criteria could be structured as evaluation factors within that framework.

> **Note:** This language is illustrative, not legal counsel. It should be reviewed and adapted by a contracting officer and legal advisor before use in any actual solicitation. Tailor thresholds, adjectival ratings, and factor weighting to your specific program requirements.

Factor X: AI Governance, Control, and Responsible Use

X.1 Threshold (Pass/Fail) Criteria

Proposals that do not meet all of the following minimum criteria shall be rated Unacceptable for Factor X and will not be considered for award.

X.1.a Human authority and override: Proposal clearly identifies a designated human decision authority and describes a mechanism for immediate human override or halt of AI-enabled functions. Responses that leave authority ambiguous or lack a practical override mechanism are Unacceptable.

X.1.b Stopping and refusal behavior: Proposal demonstrates that the system can refuse to respond or explicitly indicate "I don't know / insufficient information" under defined conditions, and that such refusal is treated as correct behavior when governance thresholds are met. Systems that cannot refuse, or that treat refusal solely as a defect to tune away, are Unacceptable.

X.1.c Autonomous action boundaries and logging: Proposal defines what actions the system may take without prior human approval, describes how these boundaries are configured, and provides for complete, auditable logging of autonomous actions. Systems that can take high-consequence actions without clear boundaries and logs are Unacceptable.

X.2 Rated (Comparative) Criteria

Proposals meeting all thresholds will be comparatively evaluated under the following sub-factors using adjectival ratings consistent with the Source Selection Plan.

X.2.a Uncertainty and Confidence Handling — Quality of mechanisms to surface uncertainty; effectiveness of approaches to avoid masking uncertainty; strength of methods to distinguish independently verified claims from user-derived context.

X.2.b Drift, Degradation, and Hidden State Management — Robustness of monitoring for behavioral drift over time and across sessions; clarity of mechanisms to manage stale or conflicting context; transparency and controllability of internal state.

X.2.c Escalation and Human Handoff — Explicitness of escalation thresholds; quality of logging and auditability of escalation events for after-action review.

X.2.d Operator Requirements and Training — Realism of training requirements; evidence of safe performance across operator expertise levels; degree to which the system helps operators manage reasoning conditions, not just outputs.

X.3 Overall Governance Risk Assessment

The Government will integrate the above sub-factor assessments into an overall governance risk characterization (Low, Moderate, High) for each proposal, considering alignment with DoD Responsible AI principles and suitability of the proposed governance approach for the system's intended operational context and consequence profile.

G. Industry Self-Assessment: Governance Readiness

Before bidding, teams should be able to answer "yes, with evidence" to most of these. If not, the governance story needs strengthening or the opportunity carries higher risk.

1. Authority and Control

- ☐ We can point to a clear statement that a human (not our system) is final decision authority in the intended use.
- ☐ We have a simple, tested mechanism for an operator to halt or override our AI functions immediately.
- ☐ Our contracts and documentation clarify who is accountable when the system is wrong (and it is not "the AI").

2. Stopping and Refusal Behavior

- ☐ Our system can explicitly refuse or say "I don't know / insufficient information" when inputs or conditions are unsafe.
- ☐ We treat refusal as correct and documented behavior in those cases, not as a defect to be tuned away.
- ☐ We have demo scripts or test results that show refusal and safe stopping in realistic scenarios.

3. Uncertainty and Confidence Handling

- ☐ Our UI makes it obvious to users when an answer is low-confidence or based on weak inputs.
- ☐ We avoid "overconfident sounding" responses when uncertainty is high, and can show how.
- ☐ We can explain which parts of an answer were independently checked against evidence versus built mainly from user context.

4. Drift, Degradation, and Hidden State

- ☐ We have monitoring or tests that look for behavior changes over time (drift, hallucination increases, degradation).
- ☐ We know what happens when context gets stale or conflicting, and can describe reset/refresh behaviors.
- ☐ We can explain what internal memory or hidden state we keep, and how customers can see, control, or disable it.
- ☐ We avoid treating our own past outputs as unquestioned truth without re-checking.

5. Escalation and Human Handoff

- ☐ We have clear rules for when the system should escalate to a human, and what that looks like in the product.
- ☐ We log escalations and can reconstruct who saw what, when, and why a handoff occurred.

6. Autonomous Action Boundaries

- ☐ We can list which actions the system can take without human approval and which require confirmation.
- ☐ Those boundaries are configurable and documented for customers.
- ☐ We log all autonomous actions and can show how a customer would review and, where feasible, reverse them.

7. Evidence, Not Just Claims

- ☐ For each of the above, we have at least one of: test results, demo scenarios, pilots, or customer references that show behavior, not just intent.
- ☐ We can name a few failure modes our system is not designed to handle and the assumptions we expect the government to accept.
- ☐ We can describe, in plain language, how our approach lines up with DoD-style responsible AI and AI risk management expectations.

H. Bid/No-Bid Self-Check Aid

Before bidding into an RFP with a serious AI governance factor, teams should be able to answer “yes, with concrete evidence” to most of these. If not, strengthen the governance story or treat the opportunity as higher risk.

If your team is answering "no" or "not sure" on more than a few of these, expect to lose a well-governed best-value competition.

Authority and Control

- ☐ We clearly state that a human, not our system, is final decision authority.
- ☐ We can show a simple, tested way for an operator to halt or override AI functions.

Stopping and Refusal

- ☐ Our system can refuse or say "I don't know / insufficient information" in unsafe or under-specified conditions.
- ☐ We have demo/test artifacts showing refusal as correct behavior, not a bug.

Uncertainty Signaling

- ☐ Our UI clearly indicates low-confidence outputs and avoids "overconfident" tone when uncertain.
- ☐ We can explain which parts of an answer are evidence-checked vs. built from user context.

Drift and Hidden State

- ☐ We monitor for behavior changes over time (drift, hallucination, degradation) and can describe how.
- ☐ We can explain what memory/hidden state exists and how customers see, control, or disable it.

Escalation and Autonomy Boundaries

- ☐ We have clear rules for when the system escalates to a human, and how that is logged.
- ☐ We can list which actions are fully automated vs. require human confirmation, and show that logs and recovery paths exist.

Evidence, Not Slogans

- ☐ For each area above, we have at least one concrete artifact: tests, demos, pilots, or references.
- ☐ We can name a few things our system is not designed to handle and the assumptions we would ask the government to accept.

I. Alignment with Existing Frameworks

This artifact is designed to complement, not replace, existing frameworks. The following table maps each governance domain to the corresponding functions in the NIST AI Risk Management Framework and the DoD Responsible AI principles it supports.

Governance Domain	NIST AI RMF Function(s)	DoD RAI Principle(s)
I. Authority and Control	Govern, Manage	Governable, Responsible
II. Stopping and Refusal Behavior	Manage, Measure	Reliable, Governable
III. Uncertainty and Confidence	Measure, Manage	Traceable, Reliable
IV. Escalation and Human Handoff	Govern, Manage	Governable, Responsible
V. Drift and Degradation Over Time	Measure, Manage	Reliable, Traceable
VI. Hidden State and Memory	Map, Measure	Traceable, Governable
VII. Portability and Vendor Dependence	Govern, Map	Governable, Equitable
VIII. Evaluation Before Adoption	Map, Measure	Reliable, Traceable
IX. Suitability for Military Use	Govern, Manage	Responsible, Equitable
X. Engineered Agreeableness	Measure, Manage	Reliable, Traceable
XI. Operator Requirements and Training	Govern, Manage	Responsible, Governable
XII. Autonomous Action Boundaries	Govern, Manage	Governable, Responsible

Note: NIST AI RMF functions are Map (context and risk identification), Measure (analysis and monitoring), Manage (response and mitigation), and Govern (oversight and accountability). DoD RAI principles are Responsible, Equitable, Traceable, Reliable, and Governable.