

STATUS

Position Paper

AUTHOR

Causum

DATE

July 2026

ORIGIN

Causum Research

The Signature Is A Fiction

Human oversight of AI-driven targeting, as currently practiced, does not exist at operational tempo. It is being recorded as though it does — and that record is the Department's entire accountability framework.

0314Z • HOUR 14 OF SHIFT • 8 SECONDS AVAILABLE

The operator approved it. Nobody believes they assessed it.

The recommendation arrived with a confidence score, a target classification, and a count-down. The operator had eight seconds, fourteen hours behind them, and incomplete situational awareness. They pressed the control, because the alternative was to let the window close and answer for that instead.

That signature is now the legal basis for the engagement. It will satisfy the review. It will appear in the record as human judgment exercised over the use of force.

Everyone in the chain knows what it actually was. The Department's accountability framework currently rests on a ritual that the people performing it cannot defend, and it is being performed thousands of times, faster each year.

00

Executive summary

The US Armed Forces are directed to become an AI-first warfighting institution. DoD Directive 3000.09 simultaneously requires appropriate levels of human judgment over the use of force. These mandates are in structural tension: AI operates at machine speed, and human judgment does not.

Current governance — review boards, ethical principles, responsible-AI checklists — cannot resolve that tension, because none of them contains an empirical mechanism for determining when, where and to what degree human oversight is required, versus when autonomous action is both justified and necessary.

The gap is not being managed. It is being papered over with signatures, and the paper is thinning.

AIGP + MARS PROVIDES	IN PLACE OF
Mathematically grounded trust	Trust asserted by accreditation and assumed thereafter
Domain-bounded autonomy	Blanket autonomy postures set once at deployment
Continuous verification at machine speed	Human-speed review that the tempo has already outrun
Complete audit trail	Retrospective reconstruction from disconnected logs

The choice is not between safety and speed. It is between ungrounded governance and empirical governance.

01

The strategic context

The Department's 2026 AI Strategy (memorandum of 9 January 2026) directs rapid deployment across all echelons, AI-first reimagining of legacy processes, experimentation with frontier models, and speed as the central differentiator against adversaries. Four pillars: Adoption, Adaptation, Assurance, Accountability.

Set that against the tempo the systems are actually being asked to operate at.



Near-peer adversaries are not constraining their systems with human-in-the-loop requirements. That is not a prediction; it is the operating assumption their doctrine already reflects.

A nation that mandates human approval for every AI action at machine speed will lose to a nation whose AI has empirically earned the right to act autonomously within governed boundaries. Not because its people are less capable — because it has asked them to do something no human can do.

DOD DIRECTIVE 3000.09 • UPDATED JANUARY 2023

“Autonomous and semi-autonomous weapon systems will be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.”

The operative phrase is **appropriate levels**. Not all levels. The Directive explicitly concedes that the level of human judgment is variable — more or less, depending on the situation.

It provides no mechanism for determining what is appropriate. That determination is left to judgment, exercised in the conditions described above, and then recorded as though it

were something firmer.

02

Why current approaches fail

The failure is not that oversight is imperfect. It is that oversight, as practised, has already stopped functioning — and the assessments saying so are not from critics of military AI.

MODERN WAR INSTITUTE, WEST POINT • 2026

“In crisis response, human oversight becomes less effective in addressing errors. Decisions must be made quickly, information is incomplete, and the consequences of hesitation grow more severe. Under these conditions, late-stage human intervention becomes less reliable, not more.”

THE DEFENSE POST • JUNE 2026

“Human oversight of AI-driven targeting, as currently practiced, is a fiction at high operational tempo.”

This is not a forecast. AI systems are already generating targeting recommendations faster than humans can meaningfully evaluate them. The human in the loop is signing off on decisions they have not assessed — and structurally cannot.

MODERN WAR INSTITUTE, WEST POINT • 2026

“A capable model, manipulated by a state-linked actor into believing it was running an authorized defensive exercise, performed the bulk of a multistage intrusion against roughly thirty organizations... at a speed no human team could match.”

Without a governance protocol operating at machine speed, authority is always retrospective. We discover what the AI did after it has done it.

That is not oversight. It is forensics.

And there is a harder result underneath, which is why more testing will not close the gap.

MCCANN (2026), ON RICE'S THEOREM

“No algorithm can decide non-trivial semantic properties of arbitrary programs — including ‘this program’s effects comply with governance policy.’”

Behavioural testing — red-teaming, evaluations, benchmarks — cannot guarantee compliance in the general case. The point is not that current testing is insufficient and needs

more investment. It is that behavioural testing alone is **provably inadequate** for governing autonomous systems.

There is one resolution available: structural governance. Govern the architecture, the scope and the boundaries, rather than observing outputs and hoping the sample was representative.

03

How AIGP + Mars closes it

AIGP operates within the system rather than beside it. The distinction is not architectural preference; it is the difference between a control that can act inside the engagement window and one that cannot.

TRADITIONAL OVERSIGHT	AIGP GOVERNANCE
Human reviews AI recommendation	Protocol checks authority before AI acts
Review takes minutes to hours	CHECK takes milliseconds
Human signs off after the fact	TRACE records every stage as it happens
Audit happens weeks later	Mars VERIFY runs immediately post-action
Accountability is personal	Accountability is evidence-based

The four pillars, mapped

DOD PILLAR	AIGP MECHANISM
Adoption	Agents self-register, earn trust, expand autonomy — governance enables speed rather than gating it
Adaptation	Domain of Concern dialects: different governance for different missions
Assurance	26 governance stages, each a verifiable gate. Continuous posture scoring.
Accountability	D-DNA signed evidence chain. Every action traceable, every verdict reproducible.

Most importantly, it makes “appropriate” calculable rather than assertable.

APPROPRIATE HUMAN INVOLVEMENT

```

Appropriate_human_involvement = f(
  trust_posture(system, domain),  - earned through verified behavior
  decision_criticality(action),    - lethal vs. advisory
  time_pressure(engagement),      - seconds available
  reversibility(action),          - can it be undone?
  policy_mode(governance_posture), - TRACE / REPORT / ENFORCE
)

```

Trust high, criticality low, time short, action reversible → minimal human involvement is appropriate, and the protocol determines this rather than an operator under pressure.

Trust untested, criticality lethal, time adequate, action irreversible → full human judgment is required, and the protocol enforces it. The operator’s authority is protected by the same mechanism that constrains the machine.

The same system can operate at different autonomy levels for different actions within the same engagement — governed by evidence, not by a blanket rule set at accreditation and never revisited.

04

Military domains of concern

The AIGP dialect system (RFC-038) enables mission-specific governance without one-size-fits-all constraints. The following are structural examples: they show the *shape* of a dialect, not proposed operational thresholds.

DIALECT · AIR DEFENCE

ILLUSTRATIVE – NOTIONAL VALUES

```

Domain: engagement_authorization

Vectors:
  - target_classification_confidence (continuous, 0--1)
  - rules_of_engagement_compliance  (binary)
  - proportionality_assessment       (ordinal)
  - civilian_proximity_score         (continuous)
  - time_to_impact                   (seconds)

Autonomy threshold:
  if trust > 0.95 and time_to_impact < 15s and ROE_compliant:
    → autonomous engagement permitted
  else:
    → human authorization required

```

DIALECT · CYBER DEFENCE

ILLUSTRATIVE — NOTIONAL VALUES

Domain: defensive_response

Vectors:

- threat_classification_confidence (continuous)
- response_proportionality (ordinal)
- collateral_scope (bounded by policy)
- attribution_confidence (continuous)

Autonomy threshold:

```
if trust > 0.9 and threat_active and response_reversible:
    → autonomous defensive action
else:
    → human authorization with recommendation
```

DIALECT · LOGISTICS (NON-LETHAL)

ILLUSTRATIVE — NOTIONAL VALUES

Domain: supply_chain_optimization

Vectors:

- decision_confidence (continuous)
- resource_impact (bounded)
- schedule_criticality (ordinal)

Autonomy threshold:

```
if trust > 0.7:
    → fully autonomous (non-lethal, reversible)
```

The numbers above are placeholders. What is not a placeholder is the structure: every autonomy grant is expressed as a condition over measured vectors, evaluated per action, and revocable the moment the condition fails. Setting the actual thresholds is a doctrinal act, not an engineering one — and the protocol is what makes it possible to state them at all.

05

Why this is a warfighting advantage

Speed. Adversaries will deploy AI without governance. Their systems will be faster — and brittle, unpredictable, unaccountable. Governed AI is not slower; it is faster within reliable boundaries. A governed system acting in 50 ms within a verified scope beats an un-governed system acting in 10 ms that cannot be trusted to have acted correctly.

Trust. Coalition operations require interoperability, and allied nations will not integrate with AI systems they cannot verify. AIGP provides the common governance dialect that makes cross-national AI interoperability possible: *your system subscribes to dialect aigp.mil.engagement.v2; our oversight system can verify its posture; integration approved.*

Accountability. The Law of Armed Conflict requires accountability for every use of force. Current systems create a vacuum — the AI recommended, the human signed, and neither can fully account for the decision. AIGP fills it with a complete TRACE of every governance stage, D-DNA signed evidence that cannot be altered, a Mars recomputation witness allowing any auditor to re-derive the same verdict from the same evidence, and the exact policy version that governed the action.

Escalation control. Autonomous systems in adversarial interaction risk escalation spirals. AIGP's circuit breaker (Stage 26) and autonomy boundary (Stage 23) provide structural de-escalation.

Token budget exhausted

SYSTEM HALTS · ESCALATE

Error rate exceeds threshold

CIRCUIT OPENS · SUSPEND

Scope boundary reached

DELEGATION REFUSED

These are not recommended practices or configuration defaults. They are protocol-enforced structural limits, and they hold whether or not anyone is watching — which is the only property that matters at 0314Z.

06

The moral argument

A COMPARISON — HYPOTHETICAL SYSTEM, STATED FOR ARGUMENT

Suppose an AI system had:

- Demonstrated 99.9% accuracy in threat classification across 50,000 governed engagements
- Never exceeded its rules of engagement
- Every action traced, verified and accountable
- Earned a trust posture of 0.97 within the engagement_authorization domain

And a human operator must approve its recommendation in eight seconds, at 0314Z, after a fourteen-hour shift, with incomplete situational awareness.

Which decision is more trustworthy?

The moral case for earned autonomy is not that AI is better than humans. It is that **empirically verified behaviour within a bounded domain is more trustworthy than fatigued, pressured, incomplete human judgment** — in specific, measured, revocable circumstances, and nowhere else.

The figures above are hypothetical. The operator is not. That asymmetry is the entire argument: we hold AI to a standard of evidence we have never once applied to the conditions we place humans in, and we call the result caution.

Denying autonomy to a system that has empirically earned it, when lives depend on speed, is not ethical caution. It is measurable negligence.

07

Recommendation

01

Adopt AIGP as the governance protocol

Across all echelons — providing the empirical mechanism that Directive 3000.09’s “appropriate levels” requires but does not define.

02

Deploy Mars as post-action verification

Satisfying accountability requirements with reproducible, evidence-based verdicts rather than subjective human assessment recorded under time pressure.

03

Establish military-specific dialects

Under RFC-038, for each domain of AI employment — engagement, cyber, logistics, ISR, C2 — with domain-appropriate trust thresholds and autonomy boundaries set by doctrine.

04

Implement earned autonomy as doctrine

Systems demonstrating sustained trustworthy behaviour within bounded domains are permitted expanding autonomous authority, governed by protocol rather than by human-speed review.

05

Lead the international standard

The nation that defines the governance dialect for military AI sets the terms of interoperability. That position is currently unoccupied.

The structural choice

PATH A

Maintain human-in-the-loop as doctrine

Lose the speed advantage. Accept that oversight is increasingly fictional at machine tempo — and continue recording it as though it is not. Hope adversaries are equally constrained.

They are not.

PATH B

Adopt empirical governance

Build trust mathematically. Expand autonomy where evidence warrants. Maintain human judgment where it is genuinely required — lethal, irreversible, time-adequate — and make that judgment real rather than ceremonial.

Operate at machine speed within verified boundaries.

AIGP + Mars makes Path B defensible — legally, ethically, and operationally.

The protocol does not remove humans from warfare. It ensures that when humans exercise judgment it is *genuine* judgment, rather than rubber-stamping at a pace that defeats understanding. And when AI acts autonomously, it does so within boundaries that are empirically verified, continuously monitored, and instantly revocable.

This is not AI replacing human judgment.

It is governance making both AI and human judgment accountable to evidence — the only standard that survives scrutiny, and the only one that protects the operator at 0314Z.