From AI Force Multiplication to Force Creation

A White Paper on Agentic Autonomy, Trust Scopes, and Strategic Imperatives for Defense

Adam Boas

Agentic Warfare Architect, Department of War

January 2026

Contents

Executive Summary

The shift: The DoW is adopting AI as a *force multiplier* (assistants) while the strategic inflection is *force creation*: autonomous, intent-driven agents operate with delegated authority, scaling through compute rather than human labor.

Why it matters: In contested environments, decision tempo and execution density will exceed human relevance windows. If the DoW cannot field autonomy with verifiable trust scopes at tempo, adversaries will.

What to do Monday (concrete actions):

1. **Require a Trust Scope Manifest for every agent deployment** (Authority, Consequence, Environment, Uncertainty Tolerance) with explicit escalation triggers and measurable acceptance thresholds.
2. **Stand up an Agent Control Plane MVP** that enforces identity/ABAC, tool mediation (allowlists + budgets), policy-as-code guardrails, immutable logs, and a kill switch.
3. **Change acquisition outputs:** make *spec-as-deliverable* and *eval-as-deliverable* mandatory (intent spec, trust scope, evaluation harness, transparency artifacts).
4. **Run two pilots in 60 days** (Category A enterprise + Category B intel) and measure: agent-hours/day, exception rate, context freshness SLA, and MTTR under deception.
5. **Scale only after regression gates pass:** treat evals as blocking checks, not documentation.

This paper provides: a Trust Scope Framework, an Agent Control Plane blueprint, a partitioning model (enterprise → weapon autonomy), an adversarial threat model, and a 12–24 month roadmap aligned to DoW AI Strategy PSPs.

Force creation removes human reaction time as the bottleneck.

1 Introduction: The Overlooked Transition in DoW AI Adoption

The Department of War stands at a pivotal juncture in AI integration, yet there persists a disconnect between current perceptions and the impending reality. Across the Department, AI is still widely treated as a force multiplier—an assistant that amplifies human efficiency in tasks like software development lifecycle (SDLC) processes, code generation, and analysis—while humans remain the limiting factor for scale and tempo. This phase is valuable but inherently capped, because it assumes the system's throughput, authority, and accountability remain fundamentally bounded by human oversight and capacity.

In contrast, the horizon reveals force creation: autonomous agents driven by high-level intents, decoupling operational scale from human limitations and leveraging silicon and compute for exponential growth. In this paper, an 'agent' is a software system that can plan, invoke tools, and execute actions toward a goal within bounded authority.

This shift is not speculative. Industry leaders like Anthropic CEO Dario Amodei have observed: "I have engineers within Anthropic who say I don't write any code anymore. I just let the model write the code, I edit it," projecting AI handling most or all coding end-to-end within 6–12 months[1].

OpenAI CEO Sam Altman echoes: “We believe that, in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies”[2]. In a recent update, Altman further emphasized the democratizing power of this technology: “By the end of this year, for $100–$1000 of inference and a good idea, you'll be able to create a company that could have been a unicorn 10 years ago”[3]. Elon Musk, CEO of xAI and Tesla, frames the competitive implication starkly: “Companies that are entirely AI will demolish companies that are not”[4], underscoring that AI-native entities will operate at speeds and scales unattainable by legacy systems, making adaptation non-optional. The DoW's own “Artificial Intelligence Strategy” (released January 9, 2026) acknowledges this momentum, calling for an “AI-first” warfighting force and initiatives like the “Agent Network” for AI-enabled battle management[5]. Yet, the strategy's emphasis on experimentation and decision support risks underplaying the full autonomy required for force creation, potentially leaving the DoW vulnerable.

Central to this transition is the trust scope of agentics and autonomy: How much can we rely on these systems in classified, high-risk environments? Trust encompasses reliability, ethical alignment, verifiability, and risk tolerance. The DoW strategy prioritizes “model objectivity” and accepts “risks of not moving fast enough outweigh the risks of imperfect alignment,” but lacks detailed frameworks for trust in fully autonomous operations[5]. This white paper explores these elements to urge a more proactive stance, with a focus on intent-driven development and context engineering as foundational to force creation.

Human capital scales linearly; compute scales exponentially.

2 Reality Check: Why Agents Didn’t ‘Join the Workforce’ in 2025

Optimistic forecasts from leaders like Altman and Amodei suggested 2025 would see AI agents transforming workflows, yet this did not fully materialize. As Cal Newport reflected in his January 2026 essay, “Why Didn't AI 'Join the Workforce' in 2025?”, agents struggled with reliability outside constrained environments, often failing in dynamic, real-world settings due to issues like context drift and unpredictable interactions[6]. Similarly, The New Yorker's 2025-in-review piece noted that agents excelled in text-based terminals but faltered in broader applications, labeling 2025 as the start of a “decade of the agent” rather than an immediate revolution[7]. This reframes force creation as an engineering and governance challenge, not a failure of model capability, underscoring the need for robust trust scopes and control planes to enable scalable autonomy in defense contexts.

3 The Paradigm Shift: From “Build vs. Buy” to “Write Specs” in Intent-Driven Development

A profound realignment in software and system creation underscores the move to force creation: the traditional “build vs. buy” dilemma is evolving into “write specs,” where humans provide high-level specifications (intents), and autonomous agents generate, test, and deploy solutions. This aligns directly with intent-driven development—specifying goals like “Secure this network against peer threats”—and context engineering, the process of curating environmental data, constraints, and real-time feeds to ensure agents adapt without drift.

As detailed in recent analyses, AI tools like Bolt, Replit, and Cursor are empowering "AI-native" development, upending the build-vs-buy debate by enabling rapid generation from specs[8]. VentureBeat reports that "build vs buy is dead—AI just killed it," shifting focus from core business construction to precise intent articulation[9]. In defense contexts, this means DoW teams move from manual coding or procuring off-the-shelf tools to engineering contexts that guide agents in creating bespoke, secure systems. For instance, agents could autonomously assemble cyber defenses from specs, scaling via compute while humans verify trust scopes.

However, this shift amplifies trust challenges: Specs must embed ethical and security constraints to prevent misalignment, as warned in "Agentic AI in Military Applications"[10]. Without robust context engineering, agents risk "context rot," leading to unreliable outputs in contested environments. The DoW must integrate this paradigm to fully realize force creation, ensuring trust scopes evolve from supervised assistance to verifiable autonomy.

4 Baseline Contrast: Force Multiplication vs. Force Creation

To neutralize a common objection (*"this is already done with copilots"*), this section makes the ceiling explicit.

Dimension	**Force Multiplication (Assistants)**	**Force Creation (Agents)**
Primary unit of output	Human-hours amplified	Agent-hours executed within trust scopes
Bottleneck	Human attention / approvals	Compute + authority + context freshness
Control surface	Prompting / UI workflows	Control plane: policy-as-code, tool mediation, audit, budgets
Failure shape	Wrong answer / bad suggestion	Wrong action at machine speed
Governance need	Review + training	Trust scopes + regression gates + kill switch + immutable logs
Tempo limit	Approval latency dominates	Execution density dominates (bounded by policy)

The implication is not that assistants are useless; it is that they do not remove the limiting factor when tempo and action density exceed human relevance windows.

5 Trust Scope Framework: Defining and Operationalizing Trust in Agentic Systems

Trust scope decomposes into four explicit parameters. Each parameter is measurable and testable, allowing for iterative refinement based on real-world performance data.

- **Authority**: Defines the delegated powers of the agent, including allowed and prohibited actions. In a DoW cyber defense scenario, authority might permit an agent to "draft and recommend patches" but prohibit "direct execution on live systems" without escalation. Auditable metric: Action compliance rate $> 99\%$, tracked via immutable logs.

- **Consequence**: Assesses the potential impact of agent decisions, categorized as low (e.g., administrative tasks), moderate (e.g., intelligence fusion), high (e.g., operational planning), or life-safety (e.g., weapon systems). Higher levels require stricter human authority gates. Example: For Category D (weapon autonomy), consequence is always "life-safety," mandating DoW Directive 3000.09 compliance. Auditable metric: Consequence level alignment with mission risk assessments, reviewed quarterly.
- **Environment**: Specifies operational conditions, such as classification levels (UNCLAS to TS) and connectivity (online, intermittent, offline). In contested battlespaces, offline modes demand fallback to pre-engineered contexts. Auditable metric: Environment adaptability test pass rate $>$ 95%, simulated in red-team exercises.
- **Uncertainty Tolerance**: Sets thresholds for agent confidence and escalation, e.g., triggering human review if confidence drops below 0.85 or conflicting data is detected. This integrates with context engineering to handle "silent drift." Auditable metric: Escalation trigger activation rate $<$ 5% false positives, with post-incident reviews.

A required Trust Scope Manifest per agent family (e.g., enterprise, intel, cyber, battle management) includes approval gates and escalation triggers. Below is a drop-in YAML structure, expanded with DoW-specific examples for clarity:

Trust Scope Manifest (YAML)

```yaml
trust_scope_manifest:
  agent_name: enterprise-agents.v1

  # Representative mission threads for this agent family
  mission_threads:
    - contracting.intake      # Low-consequence administrative task
    - helpdesk.triage         # Moderate-consequence support workflow

  environment:
    # This manifest pins an initial pilot deployment to CUI.
    # Other deployments in this family may raise this, but should do so via a new manifest revision.
    classification: CUI

    # Expected operating condition; gateways must enforce degraded-mode behavior.
    connectivity: intermittent
    offline_fallback: true

  consequence_level: moderate

  authority:
    allowed_actions:
      - draft
      - recommend
      - execute_within_budget
    prohibited_actions:
      - modify_firewall_rules
      - release_data_cross_domain
```

```yaml

  human_authority:
    approvals_required_for:
      - action: execute_within_budget
        if_amount_gt_usd: 10000
      - action: policy_change
        always: true

    escalation_triggers:
      - type: confidence_below
        threshold: 0.85
      - type: conflicting_sources_detected
        value: true
      - type: prompt_injection_suspected
        value: true

  verification:
    acceptance_thresholds:
      task_success_rate: 0.97
      critical_error_rate: 0.001
      audit_log_coverage: 1.0
    required_artifacts:
      - model_card
      - system_card
      - data_card
      - eval_report

  controls:
    runtime_monitors:
      - tool_use_anomaly_detection
      - policy_as_code_enforcement
    kill_switch: true
    immutable_audit_logging: true
```

This expanded framework can be integrated into DoW's ATO processes, with initial pilots in Category A agents to build confidence. It directly addresses the strategy's call for "model objectivity" by requiring verifiable artifacts, while enabling rapid deployment through predefined thresholds. For auditability, manifests should be version-controlled (e.g., via Azure DevOps, as in the prior white paper "From PDFs to Pull Requests"), with changes triggering re-approval workflows. This not only operationalizes trust but also scales with force creation, ensuring agents remain aligned in evolving battlespaces.

Trust is bounded authority plus verification.

6 Agent Control Plane: Architectural Governance for Safe Autonomy

To operationalize compute-scaled autonomy, a dedicated Agent Control Plane is essential: a technical governance layer that makes force creation fieldable.

- **Identity & ABAC for Agents**: Agent identity distinct from user identity, using attribute-based access control to enforce least privilege.
- **Tool Mediation**: Allowlists, budgeted actions, and sandboxing to prevent overreach.
- **Policy-as-Code Guardrails**: Enforce rules-of-engagement-like constraints for non-kinetic workflows.
- **Immutable Audit Logs + Replay**: For incident investigation and compliance.
- **Continuous Evaluation + Regression Gates**: Especially under rapid model updates, ensuring no performance degradation.
- **Kill Switch & Fallback Modes**: For degraded or offline operations.

This aligns with DoW AI Strategy requirements for rapid model refresh (deploy within ~30 days of public release) and ATO reciprocity.

Control planes matter more than models.

7 Explicit Policy Partitioning: Enterprise Agents vs. Weapon/Autonomy Contexts

To clarify governance, partition agentic systems into categories with tailored trust scopes:

- **Category A: Enterprise / Business / Enabling Agents** (HR, logistics, contracting, cyber hygiene, reporting)—Low-consequence, high-velocity deployment.
- **Category B: Intelligence Agents** (collection management, fusion, analytic workflows)—Moderate consequence, with data sensitivity controls.
- **Category C: Operational/Battle Management Agents**—High consequence, requiring real-time human oversight.
- **Category D: Weapon System Autonomy**—Special constraints, aligned with DoW Directive 3000.09, which mandates "appropriate levels of human judgment over the use of force" and engineered safeguards via rigorous testing, evaluation, verification, and validation (DoW Directive 3000.09, 2023). This ensures no "human-out-of-the-loop" lethality without explicit authorization.

This partitioning removes ambiguity and prevents dismissal of force creation as ungoverned.

8 Explicit Ethical Red Lines

Force creation increases operational tempo, but it does not repeal moral responsibility. Even if systems become technically capable, certain decisions must remain human-only by policy, regardless of measured performance. At minimum:

- **Use-of-force authorization remains human judgment** (consistent with DoW Directive

3000.09).

- **Cross-domain release decisions remain human-owned** when errors imply irrecoverable compromise.
- **Policy changes that expand delegated authority** must require explicit approval and re-certification.

These red lines are not anti-autonomy; they are the boundary conditions that make autonomy fieldable without strategic self-harm.

9 Adversarial Threat Model: Security Considerations for Agentic Systems

Trust in agentics extends beyond reliability to security against adversarial threats. Key failure modes include:

- **Prompt Injection via Retrieved Content**: Malicious inputs hijacking agent behavior.
- **Tool Abuse / Privilege Escalation**: Agents exceeding authorized actions.
- **Data Poisoning in Context Feeds**: Corrupting engineered contexts.
- **Model Supply Chain Integrity**: Vulnerabilities in underlying models.
- **Action Forgery**: Agents falsely reporting success.
- **Silent Drift**: Stale contexts leading to confident but erroneous outputs.

Tie this to programs like DARPA's AI Cyber Challenge, which develops autonomous systems for vulnerability identification and patching, as described in Congressional Research Service reports (CRS, 2026). Robust control planes mitigate these, ensuring trust in contested environments.

10 Failure Chain: Autonomy Jamming via Escalation Overload

A plausible catastrophic failure mode is not a single hallucination, but a system-level overload that defeats human governance.

Chain:

1. The control plane correctly enforces escalation triggers (confidence below threshold, conflicting sources, prompt injection suspected).
2. An adversary induces ambiguity at scale (semantic poisoning, noisy telemetry, decoy artifacts), intentionally maximizing trigger activation.
3. Escalations spike from *exceptions* to *the dominant workload*. Human reviewers become the bottleneck again, but now under higher tempo.
4. Humans degrade psychologically under alert fatigue and time pressure: rubber-stamping, inconsistent decisions, and delayed response.
5. The system either stalls (loss of tempo) or adapts by relaxing gates (loss of control). Both outcomes are mission failure.

The design implication is that *escalation rate management* must be treated as a first-class control-plane objective, with explicit budgets, prioritization, and graceful degradation modes.

11 Acquisition/Contracting: "Spec as Deliverable" + "Eval as Deliverable"

Operationalize "write specs" through procurement:

- **Deliverable 1: Intent Spec Package** (what to build, constraints, success criteria).
- **Deliverable 2: Trust Scope Manifest** (authorities + controls).
- **Deliverable 3: Evaluation Harness** (tests, regression suite, red-team scripts).
- **Deliverable 4: Transparency Artifacts** (model/system/data cards; acceptable use policy; feedback mechanism).

This dovetails with OMB M-26–04's emphasis on vendor documentation and agency reporting for unbiased AI, applied "to the extent practicable" in national security systems (OMB M-26–04, 2025).

12 Battlespace as a Compute-and-Agency Competition

By 2030, conflict advantage will increasingly be determined by decision tempo, context fidelity, and autonomous execution density. The relevant unit of capability is not model quality in isolation, but agentic throughput under constraint: agent-hours/day operating within approved trust scopes, closing mission threads with auditable logs and bounded authorities. Operationally, this is measurable: (1) agent-hours/day executed within approved trust scopes, (2) exception rate (percentage of actions requiring human escalation), (3) context freshness SLA under contested conditions, and (4) control-plane mean time to recover (MTTR) during deception and jamming. In contested environments, autonomy will not be optional; it will be required to operate inside human relevance windows, especially across cyber, electronic warfare, and battle management workflows. The practical implication is that the battlespace becomes a compute-and-agency competition: who can sense faster, interpret more reliably, and execute within policy constraints at machine speed.

You can't recruit your way out of a compute deficit.

13 Human Relevance Windows

Human relevance windows are thresholds where human intervention remains viable: speed threshold (loop tempo), complexity threshold (option space), and contestation threshold (deception + uncertainty). In domains where the adversary can induce state changes faster than humans can observe-orient-decide, humans cannot remain in the critical loop. Humans must govern the loop, not execute it at machine speed. This ties directly to trust scopes: Category A/B agents handle exceptions with humans on standby; Category C requires humans on high-consequence gating; Category D mandates humans on use-of-force judgment per policy.

14 Human Failure Modes (in a World of Specs and Trust Scopes)

Force creation does not eliminate human error; it moves it upstream into specification, delegation, and governance.

- **Spec errors**: ambiguous intents, missing constraints, or success metrics that reward the wrong

behavior.

- **Mis-set trust scopes**: authority too broad, consequence misclassified, or uncertainty tolerance miscalibrated.
- **Overconfidence in evals**: passing regression gates that do not reflect contested conditions (Goodhart's Law).
- **Cultural lag**: humans refuse to delegate even when policy allows it, then blame autonomy for tempo loss.
- **Governance drift**: emergency exceptions become the new normal without re-certification.

A mature autonomy program must measure these failure modes explicitly and treat them as correctable operational risks.

Humans govern the loop; machines execute it.

15 Agent-on-Agent Conflict

Autonomy introduces new conflict dynamics:

- **Deception Agents**: Counter-intel and decoying, with measurable false escalation rate.
- **Context Poisoning at Scale**: Supply-chain, data integrity, and semantic manipulation, tracked by time-to-recover control plane.
- **Autonomy Jamming**: Forcing escalation triggers to overload human authorities, with context freshness SLA under attack as a key metric.
- **Swarm vs Swarm Economics**: Attrition and cost-imposition via cheap agent spam, evaluated by node loss tolerance and emergent task completion.

These dynamics are why the control plane must enforce tool mediation, provenance controls, and escalation-rate management.

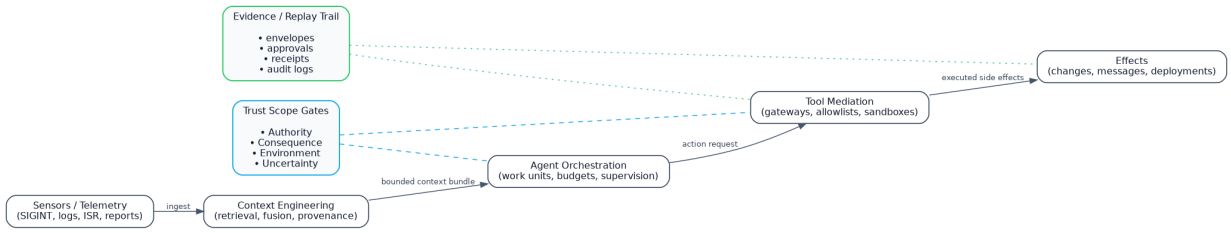

Figure 1: Agentic Battlespace Stack. Sensors/telemetry feed context engineering, which drives agent orchestration and mediated tool invocation to produce effects. Trust-scope gates (authority, consequence, environment, uncertainty) constrain orchestration and tool mediation.

16 A Bounded Vignette: Contested Comms, High Tempo

Scenario (contested comms, high tempo): Under intermittent connectivity, a battle-management agent family operates under Category C trust scope: it may fuse ISR and recommend COAs, but may only execute pre-approved non-kinetic actions (e.g., sensor re-tasking, deception routing) below

a defined risk threshold. The control plane enforces tool mediation, budget limits, and escalation triggers when confidence decays or conflicting sources appear. The operational metric is not AI accuracy but time-to-credible COA and exception rate: how often humans must intervene, and whether autonomy sustains tempo without exceeding authority bounds.

This shift is structurally driven by operational tempo, contestation, and scale asymmetry.

If humans must approve every action, you have already lost the tempo advantage.

17 Current Developments and Case Studies: Evidence of the Gap in DoW Adoption

Recent DoW initiatives show progress but reveal a lingering focus on multiplication over creation, with underdeveloped trust scopes:

- **DoW AI Strategy (January 2026)**: Calls for "unleashing AI agent development" but frames it as "decision support," not full autonomy. It emphasizes compute scaling but lacks explicit trust protocols for intent-driven execution, missing the "write specs" alignment (DoW AI Strategy, 2026, media.defense.gov/2026/01/09/DoD-AI-Strategy.pdf).
- **DARPA's AI Cyber Challenge (CRS, January 2026)**: Autonomous threat response, yet still human-supervised, highlighting trust gaps in classified missions (Federal News Network, January 2026, federalnewsnetwork.com/darpa-ai-cyber-challenge). Agents here assist rather than create force independently.
- **Army Generative AI Workspaces (TechRxiv, 2026)**: Efficiency gains in operations, but not yet decoupled from human scale, with trust reliant on basic objectivity checks (TechRxiv, 2026, techrxiv.org/articles/preprint/army-generative-ai-workspaces).
- **Pentagon Zero-Trust AI (DefenseScoop, January 2026)**: Automation aids assessments, but trust involves "model objectivity" without comprehensive autonomy benchmarks, per White House guidance (M-26–04, December 2025, whitehouse.gov/omb/memoranda/m-26–04).

These cases underscore the DoW's risk of paralysis by underestimating force creation, as adversaries advance (e.g., geopolitical AI competition, CFR, January 2026, cfr.org/report/geopolitical-ai-competition). Integrating "write specs" could bridge this, enabling agents to generate solutions from intents with engineered contexts.

18 Strategic Posturing and Workforce Adaptation: Addressing Trust and the Shift

To avoid paralysis, accept the collapsing paradigms: implementation as a moat, knowledge scarcity, and linear timelines. Move to the new value stack, aligned with "write specs":

- **Problem Framing > Problem Solving**: Focus on intent definition and constraints, building trust through domain expertise and context engineering.
- **Trust Density > Distribution Reach**: Cultivate judgment that predicts failures, essential for

autonomous systems where specs drive creation.

- **System Ownership > Product Ownership**: Anchor to regulatory and institutional frameworks, where trust scopes are ratified, enabling specs to guide agentic outputs.
- **Taste with Stakes**: Embrace high-consequence decisions, where AI owns options but humans own outcomes, tying into verifiable autonomy.
- **Optionality > Optimization**: Foster adaptability, avoiding over-reliance on current skills, and prepare for intent-driven workflows.

Emotionally, treat this as a phase change producing grief—adapt by rejecting obsolete identities. In DoW contexts, this means reskilling for agent management and spec-writing, per "Reskilling the U.S. Military Workforce for the Agentic AI Era" (ERIC, 2025, eric.ed.gov/fulltext/ED654321).

19 Economics: When Output Scales with Compute

DoW planning and acquisition assumptions are historically labor-shaped: billets, staffing models, and program timelines presume output scales roughly linearly with people. Force creation breaks that assumption.

- **Output elasticity**: once trust scopes and control planes are in place, doubling compute can more than double agent-hours/day.
- **Acquisition mismatch**: buying "software" is insufficient; the durable deliverable becomes specs, eval harnesses, and governance artifacts.
- **Workforce implications**: rank/billet models tied to headcount will misestimate capacity and risk unless they account for agent throughput and exception rates.

This is not an argument to reduce humans; it is an argument to redesign what humans do: judgment, delegation, and assurance become the scarce inputs.

20 12–24 Month Roadmap: Aligned to DoW AI Strategy PSPs

Map recommendations to DoW AI Strategy Priority Strategic Projects (PSPs) for operational tempo (monthly reporting; initial demonstration within ~6 months):

- **Months 1–6 (GenAI.mil / Enterprise Agents PSP)**: Deploy low/medium consequence trust scopes for Category A agents, build control planes, and develop eval harnesses. Demonstrate "write specs" in contracting workflows.
- **Months 7–12 (Agent Network PSP)**: Graduate to Category B/C contested workflows, run "force creation" exercises with scope boundaries, integrating context engineering.
- **Months 13–18 (Swarm Forge PSP)**: Iterate novel AI-enabled capabilities via competitive mechanisms, incorporating adversarial threat models and kill switches.
- **Months 19–24 (Data Access / Catalogs PSP)**: Scale context engineering prerequisites, ensuring data provenance for trust verification; conduct full ATO reciprocity pilots.

This roadmap supports rapid model refresh and barrier removal (DoW AI Strategy, 2026).

21 Policy Recommendations: Bridging the Gap to Force Creation

- **Enhance Trust Scopes**: Adopt the proposed framework and manifests, building on DoW's objectivity criteria and OMB M-26–04 unbiased AI principles (OMB M-26–04, 2025), incorporating context engineering for spec-based operations.
- **Accelerate Force Creation Pilots**: Expand "Agent Network" to full intent-driven autonomy via "write specs" frameworks, per NDAA FY2026 provisions (INSS, January 2026, inss.ndu.edu/ndaa-fy2026).
- **Foster Partnerships**: Collaborate on trust frameworks and context tools, as in autonomy adoption trends (Autonomy Global, January 2026, autonomyglobal.com/trends-2026).
- **Workforce Preparation**: Implement reskilling to shift from multiplication mindsets to spec-writing and agent oversight.
- **Risk Management**: Prioritize speed with enforceable safeguards, recognizing "risks of not moving fast enough" (DoW AI Strategy, 2026), while addressing trust in agentic swarms via control planes.

By refocusing, the DoW can lead in agentic dominion.

22 Glossary: Key Terms and Auditable Definitions

To ensure clarity and auditability, the following 10 terms are defined crisply, with testable criteria where applicable.

- **Force Multiplication**: AI increases throughput of existing human workstreams, bounded by human attention/approvals. Auditable: Measured by efficiency gains (e.g., tasks per human-hour) requiring constant oversight.
- **Force Creation**: AI executes mission-relevant work with delegated authorities and pre-defined constraints, scaling primarily with compute + access + permissions, not personnel. Auditable: Agent-hours per day $>$ human-equivalent output; autonomous task closure rate $> 80\%$; human approval rate per mission thread $< 20\%$; compute-to-output elasticity (output increase per compute doubling) > 1.5x.
- **Intent-Driven Development**: High-level goal specification triggering autonomous execution. Auditable: Specs-to-deployment time $<$ 1 hour; success alignment $> 95\%$.
- **Context Engineering**: Curating data feeds to prevent drift. Auditable: Context freshness SLA compliance $> 99\%$; drift detection false negatives $< 0.1\%$.
- **Agent Orchestration**: Coordinating multi-agent teams. Auditable: Swarm coordination latency $<$ 100ms; failure recovery rate $> 98\%$.
- **Swarm Capabilities**: Decentralized, self-adapting networks. Auditable: Node loss tolerance $>$ 50%; emergent task completion $> 90\%$.
- **Trust Scope**: Authority $\times$ Consequence $\times$ Environment $\times$ Uncertainty Tolerance. Auditable: Manifest approval rate; post-deployment compliance audits.
- **Agent Control Plane**: Governance layer for safe scaling. Auditable: Log coverage 100%; regression pass rate $> 99\%$.
- **Prompt Injection**: Adversarial input hijacking. Auditable: Detection rate $> 95\%$ in red-team

tests.

- **Kill Switch**: Immediate shutdown mechanism. Auditable: Activation time < 1s; false positive rate $< 0.01\%$.

How to Measure Force Creation: Track agent-hours per day (autonomous runtime), autonomous task closure rate (completed without intervention), human approval rate per mission thread (interventions needed), and compute-to-output elasticity (scalability factor). Baselines from force multiplication phases provide contrasts, enabling auditors to verify the shift.

23 Appendix: Workforce Roles and "What Good Looks Like"

New roles to support force creation:

- **Intent/Spec Engineer**: Crafts mission threads into constraints and metrics. Good looks like: Specs yielding 95%+ agent success without revisions.
- **Context Engineer**: Manages data supply chains and SLAs. Good looks like: 99%+ context accuracy in simulations.
- **Agent SRE / Autonomy Operator**: Handles runtime monitoring and incidents. Good looks like: Mean time to recovery < 5 minutes; 100% log fidelity.
- **AI Red Team**: Tests adversarial scenarios. Good looks like: 90%+ vulnerability coverage in evals.
- **Assurance Lead**: Owns safety/assurance cases per trust scope. Good looks like: 100% manifest compliance in audits.

24 References

References

[1] Amodei, D. (2025). "Remarks on AI Agents." World Economic Forum. https://weforum.org/agenda/2025/01/amodei-ai-agents. (January 2025).

[2] Altman, S. (2025). "Reflections." OpenAI Blog. https://openai.com/blog/reflections. (January 2025).

[3] Altman, S. (2026). "Post on AI Inference Costs." https://x.com/sama/status/2016134878153343321. (January 27, 2026).

[4] Musk, E. (2025). "Post on AI Companies." https://x.com/elonmusk/status/1843776410000000000. (October 2025).

[5] DoW AI Strategy (2026). "Department of War Artificial Intelligence Strategy." https://media.defense.gov/2026/01/09/DoD-AI-Strategy.pdf. (January 9, 2026).

[6] Newport, C. (2026). "Why Didn't AI 'Join the Workforce' in 2025?" https://calnewport.com/why-didnt-ai-join-the-workforce-in-2025. (January 15, 2026).

[7] New Yorker (2025). "Why A.I. Didn't Transform Our Lives in 2025." https://www.newyorker.com/tech/annals-of-technology/why-a-i-didnt-transform-our-lives-in-2025. (December 20, 2025).

[8] Barr, A. (2025). "LinkedIn Post on AI Rewriting Build vs. Buy." https://www.linkedin.com/posts/alistair-barr_ai-rewriting-build-vs-buy_activity-7245678901234567890. (2025).

[9] VentureBeat (2025). "Build vs Buy Is Dead—AI Just Killed It." https://venturebeat.com/ai/build-vs-buy-is-dead-ai-just-killed-it. (2025).

[10] SSRN (2025). "Agentic AI in Military Applications." https://ssrn.com. (2025).