Adam Boas

adamboas.com | anboas@gmail.com

April 18, 2026

The Next AI Bottleneck Is Delegated Authority

Everyone is still arguing about models.

Government is starting to ask a harder question: what does it take to trust software that can act?

That is the real signal in the current stack. NIST is pushing on agent standards, agent security, identity and authorization, evaluation, monitoring, and procurement-facing measurement. DARPA is pushing on trustworthy systems, AI engineering, and human-AI teaming. Defense acquisition is starting to ask for open implementations, public documentation, conformance artifacts, and software that other teams can inherit.

These are not the same effort.

They are, however, circling the same missing layer.

The next bottleneck in AI is not generation.

It is governed delegated authority.

That has been the through-line of the argument on this site for a while now.

From AI Force Multiplication to Force Creation argued that the strategic shift is not from no AI to some AI. It is from human-hours amplified to agent-hours delegated. *ACP-RA* argued that this transition is not fieldable without a control plane: explicit identities for agents, explicit trust scopes, mediated tool use, evidence, replay, rollback, and degraded-mode survivability. *DAD* argued that the problem is institutional as much as technical: machine-executed authority needs a joint control-plane authority, not just another software program. *From PDFs to Pull Requests* made the adjacent point that governance cannot remain static prose. At machine tempo, policy has to become executable, versioned, testable, and continuously verified.

Recent NIST, DARPA, and Prometheus Flame signals do not replace that argument.

They validate it.

Agents are not apps

This is the center of gravity.

Agents are not apps. They are delegated-authority systems.

An app usually executes inside a relatively fixed envelope. It takes user input, processes it, and returns an output. An agent is different. It plans. It selects actions. It uses tools. It traverses systems. It can modify external state. It can continue operating after the initial prompt. It can collaborate with other agents. It can create consequences that outlive the interaction that started them.

Once that is true, the core problem changes.

The interesting question is no longer whether the model is impressive.

The question is whether the delegation is governable.

That requires a different architecture.

It requires identity for agents and sub-agents. It requires explicit trust scopes rather than implied permission. It requires policy decision points and policy enforcement points at the actual doing

boundary: runtime admission, retrieval, tool invocation, inter-agent messaging, work-unit transitions, and release. It requires budgets, containment, and revocation. It requires signed traces, provenance, replay, and rollback. It requires environments that can be constrained and modified rather than treated as an unbounded execution surface.

In other words, it requires a control plane.

That is the argument here. Not "AI, but safer." Not "better evals." Not "more standards."

A control plane for delegated authority.

NIST is starting to define the operational surface

The most important thing in NIST's recent work is not any single document. It is the shape of the portfolio.

The AI Agent Standards Initiative is explicitly about agents that can act autonomously, function securely on behalf of users, and interoperate through standards and open protocols. That is already a meaningful shift. The target is no longer only model quality in isolation. The target is trustworthy action in an ecosystem.

The January RFI on AI agent security makes the shift even clearer. It is scoped to agent systems capable of affecting external state. It asks about deployment environments, tool restrictions, continuous monitoring, patching across the lifecycle, human oversight for consequential actions, and the current state of "undoes, rollbacks, or negations" for unwanted trajectories. That is the right class of question. It means the conversation is moving from "is the model safe?" to "how do we govern a system that can do things?"

The identity and authority concept work points in the same direction. Identification, authorization, auditing, non-repudiation, and prompt-injection controls are not accessory concerns. They are the skeleton of accountable delegation.

The evaluation portfolio matters for the same reason. NIST AI 800-2 pushes toward validity, transparency, and reproducibility in automated benchmark evaluations. NIST AI 800-3 pushes evaluators to formalize assumptions and measurement targets statistically. NIST AI 800-4 shifts attention from lab scoring to post-deployment monitoring. The MOU with GSA ties evaluation more directly to procurement through USAi. The OpenMined CRADA pushes on secure evaluation methods for settings where data, models, or benchmarks cannot be openly shared.

Put differently, NIST is starting to build measurement science for trust in operation.

That is good.

But measurement alone is not the answer.

Measurement needs a governed execution surface to measure.

DARPA is pushing research toward the same problem

DARPA's AI Forward initiative lands in the same neighborhood from the research side.

It frames the objective as trustworthy systems for national security missions. It explicitly centers three thrusts: foundational theory, AI engineering, and human-AI teaming. It says the point is to build systems that work as intended in the real world, not just in the lab.

That is more important than it sounds.

For years, the AI discourse drifted toward a shallow equation: better benchmark performance equals progress. DARPA is naming a deeper requirement. Real systems have to operate reliably, interact appropriately with people, and hold up in contested, messy environments.

Still, trustworthiness only becomes real when it has an engineering surface.

A trustworthy system is not one that scores well on an eval card and then disappears into runtime darkness. A trustworthy system is one whose authority is bounded, whose actions are mediated, whose evidence is retained, whose permissions can be reduced or revoked, whose behavior can be supervised as a portfolio, and whose failure does not automatically become a mission-scale event.

That is why trustworthy AI eventually becomes control-plane engineering.

Prometheus Flame matters because acquisition is starting to buy for inheritance

The Prometheus Flame draft matters less because of the deadline and more because of what it revealed.

The Air Force did not frame the early phases as a competition for another closed demo stack. The draft points instead toward open-source AMS GRA and A-GRA implementations, permissive licensing, public repositories and documentation, turn-key projects, CI/CD artifacts, compliance reporting, and follow-on prototype and production OT pathways. It also places real weight on usability, onboarding, standards compliance, and risk rather than treating them as proposal garnish.

That is not normal demo procurement.

That is ecosystem formation.

The most important artifact in that kind of acquisition is not the thing that flies once.

It is the thing other people can inherit.

A reference implementation. A starter kit. A conformance harness. A public repo. A paved road. An example project that works out of the box. A delivery system that others can extend without privileged access to the original builder.

That logic sits naturally beside the argument in *From PDFs to Pull Requests*. If government wants an open, durable autonomy ecosystem, standards cannot remain prose-only and compliance cannot remain narrative-only. The substrate has to be executable: code, tests, schemas, policy objects, evaluation gates, evidence events, reproducible pipelines, and release discipline.

That is what it means to buy for inheritance.

The next object we need: an Action Reversibility Profile

Previous writing on this site already argued for trust scopes, control planes, policy gates, evidence, and rollback.

The extension I would add now is this: every serious agent deployment should carry an **Action Reversibility Profile**.

The point is simple. Not all actions are equal, and our architectures should stop pretending they are.

Some actions are safely reversible. Some are compensable but not literally undoable. Some are effectively irreversible. Some should be prohibited altogether.

That classification should not live in a policy memo on a shelf.

It should live in the control plane.

At minimum, an Action Reversibility Profile should classify actions into four buckets:

1. **Reversible**: actions that can be automatically undone with high confidence.
2. **Compensable**: actions that cannot be literally rolled back but can be corrected through a defined compensation path.
3. **Irreversible**: actions whose consequences cannot be practically reversed and therefore require tighter gating, stronger evidence, and usually explicit human authorization.
4. **Prohibited**: actions the system is never permitted to take.

For each class, the profile should define:

- approval conditions
- allowed tools and environments
- evidence obligations
- logging and provenance requirements
- rollback or compensation mechanisms
- timeout behavior
- kill-switch behavior
- update-handling rules
- degraded-mode posture

This is where recent NIST signal becomes especially useful. If government is already asking about deployment constraints, monitoring, patching, human oversight, and the state of practice for undoes or rollbacks, reversibility should become a first-class design property rather than a hopeful runtime feature.

Reversibility is not just a safety pattern.

It is an operational tempo pattern.

The faster a system can act, the faster it must be able to stop, contain, and recover.

What should happen next

If the current direction is real, the next moves should be concrete.

1) Standardize authority, not just interfaces

Message schemas are not enough. The next standards layer has to carry identity, trust scope, delegated permissions, policy hooks, and evidence events.

2) Make conformance machine-checkable

If a standard cannot be tested, it will eventually be narrated instead of implemented. Conformance needs harnesses, reference scenarios, negative tests, and machine-readable outputs.

3) Treat rollback as a first-class design property

Rollback should shape action design, decomposition, and environment mediation from the beginning, not appear as an afterthought after deployment.

4) Collapse procurement, evaluation, and operations into one evidence loop

Procurement should not buy one thing, evaluation should not measure a second thing, and operations should not trust a third thing. The same evidence model should flow from source to test to deployment to sustainment.

5) Fund open substrates and maintainers

If the goal is a real ecosystem, government cannot only reward edge demos. It has to fund the center: reference implementations, conformance tooling, public starter kits, maintainers, and the delivery systems that keep the whole thing governable over time.

The real advantage

The strategic advantage is not "having AI."

It is being able to field governed autonomy at mission tempo.

That means being able to delegate without surrendering authority. It means being able to move faster without severing provenance. It means being able to exploit open ecosystems without accepting open-ended risk. It means many builders can contribute while identity, policy, evidence, rollback, and containment remain coherent across the whole system.

That is the job of the control plane.

And that is why this moment matters.

NIST is starting to standardize toward the operational surface. DARPA is starting to research toward trustworthy systems that work in the real world. Defense acquisition is beginning to buy for open, inheritable substrates rather than only polished integration theater.

Good.

Now the missing piece needs to be stated plainly.

The future will not be won by institutions with the most impressive demos.

It will be won by institutions that learn how to govern delegated authority.

That is the next layer.

That is the work.

References

1. NIST CAISI, *AI Agent Standards Initiative*. https://www.nist.gov/caisi/ai-agent-standards-initiative
2. NIST CAISI, *CAISI Issues Request for Information About Securing AI Agent Systems*. https://www.nist.gov/news-events/news/2026/01/caisi-issues-request-information-about-securing-ai-agent-systems

3. NCCoE, *New Concept Paper on Identity and Authority of Software Agents*. https://www.nccoe.nist.gov/news-insights/new-concept-paper-identity-and-authority-software-agents
4. NCCoE, *Accelerating the Adoption of Software and AI Agent Identity and Authorization* (concept paper PDF). https://www.nccoe.nist.gov/sites/default/files/2026-02/accelerating-the-adoption-of-software-and-ai-agent-identity-and-authorization-concept-paper.pdf
5. NIST, *Towards Best Practices for Automated Benchmark Evaluations* (NIST AI 800-2). https://www.nist.gov/news-events/news/2026/01/towards-best-practices-automated-benchmark-evaluations
6. NIST, *NIST AI 800-2 Initial Public Draft* (PDF). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf
7. NIST, *New Report: Expanding the AI Evaluation Toolbox with Statistical Models* (NIST AI 800-3). https://www.nist.gov/news-events/news/2026/02/new-report-expanding-ai-evaluation-toolbox-statistical-models
8. NIST, *NIST AI 800-3* (PDF). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-3.pdf
9. NIST, *New Report: Challenges to the Monitoring of Deployed AI Systems* (NIST AI 800-4). https://www.nist.gov/news-events/news/2026/03/new-report-challenges-monitoring-deployed-ai-systems
10. NIST, *NIST AI 800-4* (PDF). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-4.pdf
11. NIST CAISI, *CAISI signs MOU with GSA to boost AI evaluation science in federal procurement through USAi*. https://www.nist.gov/news-events/news/2026/03/caisi-signs-mou-gsa-boost-ai-evaluation-science-federal-procurement-through
12. NIST CAISI, *CAISI signs CRADA with OpenMined to Enable Secure AI Evaluations*. https://www.nist.gov/news-events/news/2026/03/announcement-caisi-signs-crada-openmined-enable-secure-ai-evaluations
13. DARPA, *AI Forward*. https://www.darpa.mil/research/programs/ai-forward
14. SAM.gov, *Prometheus Flame - Request for Information* (Special Notice, FA8694-26-S-8888). https://sam.gov/opp/68a071f69a1c4ec883b8392c5ae040b1/view
15. SAM.gov API notice description (source for extracted draft text). https://api.sam.gov/prod/opportunities/v1/noticedesc?noticeid=68a071f69a1c4ec883b8392c5ae040b1