When Disagreement Matters: Friction, Pluralistic Alignment, and National-Security AI

Morgan Livingston [1]

Abstract

Pluralistic-alignment work increasingly turns from model-output diversity to institutions. National-security AI is a stress case for that turn. Classification, operational urgency, and security-dilemma pressure structurally constrain public participation, transparency, and ordinary review. The result is what I call *frictionless security*: the public-private extension of Deeks and Eichensehr's *frictionless government*, in which state and private actors converge around capability, speed, secrecy, and lawful-use compliance while the architecture that would weigh those choices against civil liberties, humanitarian restraint, escalation control, democratic legitimacy, and AI safety is weak. Pluralism here is whether institutions can preserve conflict among public values, route that conflict through forums with legitimate authority, and make the result reviewable. Starlink in Ukraine, Project Maven, Apple v. FBI, and reported frontier-AI defense-contracting disputes show three failure patterns — private veto, absorbed dissent, and red lines that do not bind deployment — and one case in which conflict reached a forum with authority. Three lessons follow: pluralistic alignment depends on the forum that receives disagreement, on the market structure that determines whether refusal can be absorbed, and on whether technical evidence can be verified by outsiders. The contribution is to show that, in national-security AI, pluralistic alignment depends not only on model behavior but on the contracts, forums, evidence, and market structures that determine whether disagreement becomes accountable decision-making.

[1]Independent Researcher. Correspondence to: Morgan Livingston <morganauroralivingston@gmail.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Pluralistic-alignment work has built increasingly sophisticated ways for models to represent plural values. Whether those values become consequential is a separate problem: converting disagreement about public values into accountable decisions when the institutions that should carry those values are themselves under pressure. National-security AI makes this problem unavoidable.

A growing strand of pluralistic-alignment work moves toward institutions. Sorensen et al. (2024) develop output-facing pluralism. Leibo et al. (2024) reject rational convergence and propose contextual grounding, community customization, continual adaptation, and polycentric governance. Huang et al. (2024) operationalize public input in model design. Kasirzadeh (2024) distinguishes first-order value choices from second-order legitimacy choices: not only what values are implemented, but who has authority to choose, measure, aggregate, and operationalize them. National-security AI tests whether these moves survive deployment settings where participation, transparency, community input, and ordinary public scrutiny are structurally constrained.

Heterogeneity is not enough for pluralism. A firm saying no is not pluralism, even when it tracks a public value. If one lab refuses a domestic-surveillance contract and another accepts, what we have is heterogeneity, not enacted pluralism. Pluralism here means a system that preserves real conflict among incommensurable public values: civil liberties, humanitarian restraint, escalation control, soldier safety, AI safety, legality, democratic legitimacy. Firm refusal, worker voice, and technical evidence become pluralistically meaningful only when institutional architecture converts them into reasons, evidence, review, and consequences.

That architecture is what national-security pressure tends to weaken. Deeks and Eichensehr use *frictionless government* to describe a national-security failure mode in which ordinary checks weaken because Congress and the executive converge on the same security objective. Institutions that are supposed to check one another become aligned around speed, capability, secrecy, or threat response, leaving fewer sites where contrary values can slow or redirect action

(Deeks & Eichensehr, 2024). I use *frictionless security* to name the public-private extension. In national-security AI, the state and private-sector providers may converge around capability, speed, secrecy, and lawful-use compliance while the architecture that would weigh those choices against other public values is weak. Now, public and private actors do not always agree. But the disagreements that remain may be routed through institutions that cannot change the deployment decision.

This pattern has different causes on each side of the public-private relationship. State actors operate under security-dilemma pressure and intelligence-legalism dynamics (Schlanger, 2015). Firms operate under contract dependence, market dynamics, reputational exposure, and talent scarcity. Kulveit et al. (2025) place this in a wider concern: AI development can erode the institutional levers through which human influence over large-scale systems is held; Bullock et al. (2025) frame the state-level stakes as a narrow corridor between despotic and absent power. National-security AI is the sharper case: some levers are constrained from the start.

By *national-security AI* I mean AI procured or deployed by the state for defense, intelligence, or related security functions where classification, operational urgency, or national-security legal doctrines structurally constrain ordinary public participation, transparency, and review. Classification is an intensifier, not the boundary: national-security framing can also constrain review through securitization, judicial deference, procurement channels, or congressional secrecy. Many of these systems act on noncitizens — at the border, in foreign theaters, through transnational surveillance — who are absent from domestic accountability structures.

Pluralistic alignment in national-security settings is determined less by who holds what values than by whether institutions can convert disagreement into reasons, evidence, review, and consequences. § 2 develops accountable pluralism as the preservation of incommensurable value conflict in a forum with authority to change deployment. § 3 identifies four reasons ordinary conversion fails: security framing narrows value-holding, secrecy and AI opacity limit visibility, private refusal lacks democratic authority, and crisis speed compresses review. § 4 traces three routes — law, contracts, and markets — through which public values can gain or lose force. § 5 reads four cases — Starlink, Project Maven, Apple v. FBI, and frontier-AI defense contracting — to distinguish accountable conversion from private veto, absorbed dissent, and red lines that do not bind deployment. § 6 returns to pluralistic alignment with three lessons — forum, substitutability, evidence — and discusses design targets. § 7 covers limitations.

2. Accountable pluralism requires architecture

Accountable pluralism is not heterogeneity: it is not satisfied when one firm refuses and another accepts. Accountable pluralism requires a decision process that preserves real conflict among public values that do not reduce to one metric (Berlin, 1969; Mason, 2018): civil liberties, humanitarian restraint, escalation control, soldier safety, AI safety, democratic legitimacy, and the security interests on the other side.

Public-law-style values — legality, reason-giving, reviewability, accountability, non-arbitrariness (Dickinson, 2006; Bozeman, 2007; Jørgensen & Bozeman, 2007) — help make those conflicts institutionally holdable. They are not themselves pluralism. They are the procedural infrastructure that lets plural value conflict survive long enough to affect a decision.

For this paper, accountable pluralism requires at least three things:

1. competing public values are preserved as live reasons rather than collapsed into one metric;
2. those reasons enter a forum with legitimate authority to weigh them; and
3. the resulting decision is recorded, reviewable, and capable of changing deployment.

Kasirzadeh's second-order question may be answerable even when first-order participation is not. Pettit (1997; 1999) supplies the legitimacy bridge: when participation in originating a decision is limited, power must still be contestable, and contestability requires institutional channels (Kaminski, 2021; Citron & Pasquale, 2014). Critics existing somewhere in the system is not enough; challenges must have somewhere to go and some chance of being answered. Mansbridge et al. (2012) extend the point: legitimacy can be produced across a system of institutions, not only inside a single forum, but the system has to do work — moving reasons, evidence, and challenge across sites where decisions can change.

Ovadya et al. (2025) name the precondition: democratic authority over a decision depends on a process whose traces are legible to outsiders and whose outcomes the authority can bind itself to. Classification withholds both. National-security AI does not fail partway up that scale; it starts at the bottom.

By *friction* I mean the practical capacity of an institutional system to convert value conflict into consequential decision-making. Friction requires an actor with leverage, a forum with authority to hear the conflict, and a path by which the forum's decision can affect deployment. Without leverage and consequence, objection reduces to expression. Similarly, without forum, leverage becomes private discretion. Friction

is therefore a set of minimum practical questions: Who can object? What evidence can they use? Where does the objection go? Who must answer? Can the answer change deployment?

3. Why ordinary conversion fails

3.1. National-security pressure can narrow value-holding

National-security institutions can give capability, speed, secrecy, and lawful authorization more usable institutional homes than other public values. Securitization moves issues out of ordinary politics (Buzan et al., 1998; Wæver, 1995), judicial deference doctrines constrain courts, congressional review may be classified and dependent on committee capacity, and totalizing threats may overwhelm dissent.

In the U.S., national-security AI systems are routed into separate defense and intelligence channels. M-25-21, OMB's government-wide AI-use memorandum, does not cover AI used as a component of a National Security System, and several of its requirements do not apply to Intelligence Community elements (Office of Management and Budget, 2025a). M-25-22, its acquisition counterpart, likewise excludes Intelligence Community elements and AI acquired for use as a component of a National Security System (Office of Management and Budget, 2025b). M-25-22 contains the kinds of acquisition tools this paper takes as useful for accountability: vendor lock-in protections, IP rights, data and model portability, ongoing testing and monitoring, documentation and transparency requirements, independent evaluation, and vendor performance terms. National-security AI is governed instead by the DoD AI Ethical Principles, the DoD Responsible AI Strategy and Implementation Pathway, DoDD 3000.09, the ODNI Principles of AI Ethics for the Intelligence Community, and the 2024 National Security Memorandum on AI. Brennan Center analysts describe this as a two-track structure, with the national-security track thinner on evidence requirements, public use-case inventories, waiver transparency, and individual remedies (Toh, 2025). The Department of War's January 2026 AI Strategy widens the gap, directing standard "any lawful use" language in every DoW AI contract and seeking models "free from usage policy constraints that may limit lawful military applications" (U.S. Department of War, 2026a). For pluralistic alignment, the question is whether the national-security track preserves disagreement in a form that procurement officers, inspectors general, congressional staff, courts, or external technical reviewers can receive, contest, and act on.

3.2. Secrecy and AI opacity narrow visibility

Two opacity problems compound. Some deployments are deep secrets (Pozen, 2010): their existence is concealed, not just their content. Even when existence is known, AI systems can be inscrutable to cleared insiders (Burrell, 2016; Selbst & Barocas, 2018). Deeks (2025) argues that national-security AI compounds the accountability strain created by secrecy: classified system, opaque mechanism, responsibility spread across lawyers, engineers, operators, contractors, and agencies (Thompson, 1980; Patel & Toomey, 2024). Deep secrecy can defeat public contestation by hiding existence; AI inscrutability can defeat cleared review by hiding mechanism.

3.3. Private refusal is not pluralism

Distributed responsibility is sometimes treated as a solution: if the state is pressured toward security logic, perhaps firms or workers can hold public values instead. Private actors can surface values the state discounts, and technical knowledge it lacks, but they do not legitimately represent the public in a forum with authority.

National-security AI sits inside a longer history of private infrastructure for state force (Weber, 1946; Singer, 2003; Avant, 2005; Krahmann, 2010). Defense contractors, intelligence vendors, cybersecurity firms, cloud providers, and platforms all sit inside the operational apparatus (Eichensehr, 2017; Chachko, 2021; Maurer, 2018). AI adds model providers and integrators to that chain.

A frontier lab refusing may surface a public value. The refusal becomes pluralist only when an institution with legitimate authority can receive it, review it, and make it consequential. Eichensehr (2019) names a structural condition: companies push back when resistance can be framed as protecting users. Belfield (2020) shows why AI-worker activism has had purchase — shared professional culture, scarce talent, consumer-facing market position — and why those conditions are unstable.

Firms are institutional sites where technical knowledge, commercial pressure, and value commitments interact. Even cooperative firms may resist sharing the technical detail that would let others verify their claims. The pluralist question is institutional: whether refusal or compliance becomes reviewable and consequential outside the firm.

3.4. Crisis pressure and normalization of deviance

Speed does not only reduce review time. It can change what counts as acceptable risk (Sagan, 1993; Vaughan, 1996). Reporting on Israel's Lavender system describes human-review windows compressed to seconds per target under wartime tempo (Abraham, 2024); the Department of War's

January 2026 AI Strategy makes the underlying logic explicit, holding that "the risks of not moving fast enough outweigh the risks of imperfect alignment" (U.S. Department of War, 2026a). Crisis speed is one condition under which independent constraints get absorbed into operational compliance. Emergency pressure is when ordinary review is hardest and later review becomes most important.

4. How public values become institutional constraints

Public values do not usually disappear all at once under national-security pressure. They are routed into institutions that may or may not be built to hold them. Three routes are the focus here: legal review, contracts, and market structure. They are not the only routes — doctrine, rules of engagement, classification review, inspector-general process, congressional notification, technical standards, and professional norms can also matter. The narrower claim is that law, contracts, and markets are three places where public values often become administrable constraints, or lose force.

4.1. Legal-compliance routing

Legal review can preserve public values. It can also narrow them. For example, civil liberties may become a Fourth Amendment, statutory, or intelligence-authority question. The failure mode is substitution: a process asks "may we?" when the unresolved question is also "should this use occur?" Schlanger (2015) documents the pattern in the intelligence context. When legal compliance has lawyers, escalation procedures, and sign-off requirements while substantive civil-liberties review lacks comparable infrastructure, compliance can crowd out review of the underlying value conflict. Deeks (2025) extends the accountability problem to national-security AI: AI can shift practical authority toward engineers, make responsibility for errors harder to assign, and weaken the executive's ability to explain why a system produced a result.

DoD's AI Ethical Principles — responsible, equitable, traceable, reliable, governable — give officials terms through which to demand judgment, documentation, testing, and control across combat and non-combat functions (U.S. Department of Defense, 2020). They do not, by themselves, answer whether a particular use should occur. DoDD 3000.09, revised in January 2023, addresses autonomous and semi-autonomous weapon systems specifically; it is not a general governance instrument for national-security AI (U.S. Department of Defense, 2023).

4.2. Contract routing

Procurement is where corporate commitments either become durable or evaporate. A refusal category, safety threshold, or deployment condition that does not appear in data rights, audit rights, flow-down clauses, monitoring duties, reporting duties, termination rights, or reauthorization triggers is not yet a constraint on downstream deployment. It may matter reputationally, internally, or in future litigation. It does not bind the procurement officer, integrator, cloud provider, prime contractor, operator, or reviewer unless it enters the legal and technical architecture they use.

Mulligan & Bamberger (2019) show why this matters for machine learning. Procurement can embed policy choices in technical systems without the public participation, factual record, or reasoned explanation associated with administrative policymaking; their response is to bring administrative-law values — including contestable design — into procurement. Coglianese & Lampmann (2021) make the contract mechanism explicit: provisions on public access, privacy, bias detection, transparency, validation, and cybersecurity. Default contracting infrastructure does not by itself carry public values across the contractor interface (Schooner, 2005; Dickinson, 2011; Freeman & Minow, 2009; Metzger, 2003).

National-security AI procurement is also an incomplete-contracting problem. Hadfield-Menell & Hadfield (2019) frame AI alignment this way: values, goals, and downstream uses cannot be fully specified in advance, and human contracting relies on external structures — law, culture, courts, institutions — to fill gaps. The same problem appears in procurement. No contract can specify every downstream use, integration context, failure mode, or value conflict. The pluralist work happens in gap-filling: audit rights, reporting duties, procurement review, external evaluation, and technical evidence that can travel outside the firm.

4.3. Market-structure routing

Defense markets are not neutral. They are shaped by public R&D, procurement choices, classification rules, acquisition pathways, and industrial-base policy (Mazzucato, 2013). Market structure affects whether refusal or dissent can matter. When many providers compete, refusal at one provider may be easy to absorb. Project Maven illustrates: Google workers objected, Google did not renew, the capability continued through other channels. Hard-to-substitute actors hold more leverage but can become private veto points, as with Starlink. The pluralist question is not whether substitution is hard but whether durable market positions are bound to reviewable obligations.

5. Four ways disagreement fails or becomes consequential

The cases below are diagnostic probes, not causal evidence (Gerring, 2017; Seawright & Gerring, 2008). They share one structural claim: pluralism requires institutional architecture that can receive disagreement, give it a forum, and attach consequences. Hirschman's (1970) framing makes the point analytically. Exit and voice are not properties of dissenting actors alone; they depend on whether structures exist to convert refusal or objection into reviewable, consequential decisions. The cases trace what happens when the architecture is absent, partial, or present. Each foregrounds one conversion pattern: where the disagreement arose, who held leverage, what forum existed, and what institution, if any, could change the deployment path.

Case selection follows the framework's two decisive dimensions: the *substitutability* of the dissenting actor (can the buyer route around refusal?) and the *forum* (does an institution with authority over deployment receive the disagreement?). The four cases populate that 2×2: Starlink is low-substitutability with no public forum; Maven is high-substitutability with a firm-level forum only; Apple v. FBI pairs architectural leverage with a forum that could bind deployment; frontier defense contracting is the live, still-open cell. This is a most-different design meant to expose the conversion mechanism, not a representative sample meant to estimate how often each pattern occurs. Other cases — Clearview AI in police procurement, NSO/Pegasus export licensing, or Section 702 disclosure litigation — vary the same two dimensions and would extend or stress the account rather than overturn its central claim that architecture, not heterogeneity, determines pluralism. The account's distinguishing, falsifiable implication is cell-level: where substitutability is high and no deployment-level forum exists, even costly firm-level refusal should leave the deployment trajectory intact — the pattern Maven's continuation displays.

5.1. Starlink in Ukraine

Private veto: private leverage without accountable architecture.

The public record is contested and appears to involve more than one episode (Washington Post, 2023; Snopes, 2023; Abels, 2024; Jayanti, 2023). This paper relies on the Crimea/Sevastopol episode: a request to enable or extend Starlink service for a military operation, which SpaceX declined. Musk later framed the refusal as an effort to avoid escalation. This relied on a structural feature: private control over low-substitutability infrastructure, without a visible accountable forum.

The disagreement arose between an operationally dependent wartime user and a private infrastructure provider. SpaceX held architectural and operational control.

Private veto is no more pluralistic than securitization. SpaceX had leverage, exercised it, and the result was private discretion regardless of whether the underlying judgment was right. Private friction without architecture cannot be pluralism even when it preserves a value, because the value's preservation depends on one actor's discretion rather than accountable conflict.

5.2. Project Maven

Absorbed dissent: worker voice converted into firm exit, but not into deployment-level review.

Thousands of Google employees objected to the contract; public reporting put petition signers around four thousand, with about a dozen resignations (Conger, 2018; Dave, 2018). Workers had leverage at the firm: shared professional culture, scarce talent, and consumer-facing reputational exposure. The decision forum was the company. Google did not renew the contract in 2018 (Dave, 2018). Worker voice changed Google's participation but it did not create a forum with authority over Maven's deployment trajectory. The program's later continuation and expansion supports a narrower inference: firm-level exit did not produce deployment-level review of the underlying use (Reuters, 2024; 2026b).

Maven is a partial firm-level success and a deployment-level failure. Reuters reported that Palantir received a $480 million contract in May 2024 for the Maven Smart System (Reuters, 2024). The relevant value conflict was not preserved in a forum with authority over the deployment trajectory, and without monopoly over the entire tech stack, the work moved.

5.3. Apple v. FBI

Pluralistic conversion: refusal entered court process, public technical evidence, and policy debate.

The government sought an All Writs Act order requiring Apple to create and sign new software to bypass security on a suspect's device (Electronic Privacy Information Center, 2016). Apple held architectural leverage: it could create the software but had not. Apple argued the All Writs Act did not authorize compelling it to write new software, and framed the dispute as broader than one phone (Apple Inc., 2016). The forum was a federal magistrate judge. The cryptography community supplied public technical arguments presenting the dispute as a systemic cybersecurity risk (Bellovin et al., 2016; Landau, 2016). The government later obtained access through a third party, leaving the legal rule unresolved.

The "pluralism" was not that Apple took a position on these values. It was that competing values could be argued, evidenced, and reviewed in a forum with authority. There was

a combination of architectural leverage, public technical evidence, and accountable forum. The resolution also confirms the framework rather than escaping it: once a third-party tool let the government substitute around Apple's refusal, the forum lost its purchase — even the success case shows conversion lasting only as long as substitution is unavailable.

5.4. Frontier-AI defense contracting

Red lines that may not bind: reported corporate red lines exist, but the public cannot yet see whether they bind the deployment architecture.

Because this case is unsettled, I separate three epistemic tiers and rely below only on the first two. *Documented in public reporting*: the government's announcement of classified-network AI agreements naming several firms, the published corporate usage policies, and the filed litigation (U.S. Department of War, 2026b; Anthropic, 2026; Anthropic PBC v. U.S. Department of War, 2026). *Inferred from procurement structure*: that a restriction stated in a model provider's policy need not bind integrators, cloud environments, or downstream classified use unless it enters the contract architecture (§ 4.2). *Speculative, and set aside here*: which firms ultimately hold or fold, and whether any specific red line in fact constrained a specific deployment — claims the current public record cannot settle.

Public sources indicate disputes over whether and how frontier-AI providers' restrictions on domestic surveillance and autonomous weapons would apply in classified deployments (Anthropic, 2026; Amodei, 2026; OpenAI, 2026; Reuters, 2026c;a; Tiku & De Vynck, 2026). The legal and factual record is still developing (Anthropic PBC v. U.S. Department of War, 2026; CBS News, 2026; Frazee et al., 2026). The government publicly announced classified-network AI agreements with several firms, including OpenAI and Google; Anthropic does not appear in that announcement and has been reported as the major holdout (U.S. Department of War, 2026b; Sabin, 2026). The public record shows corporate statements, government announcements, reporting, and litigation around the dispute. It does not yet show a forum that can review the downstream classified use, test the red lines, and impose deployment consequences. Tillipman (2026) documents the procurement-law version: these are bilateral deals that bind only their signatories, often Other Transaction agreements outside the ordinary contract-disputes process. A red line written into one need not survive the move to an integrator's platform, and termination is not governance — in the national-security case it may not even be available.

Neither firm refusal nor compliance, on its own, points to pluralism. The question is whether either position enters an architecture that can convert it into legitimate decision-making. In Hirschman's terms, both employee objection and firm-level refusal function as voice; exit is what happens if the relationship ends. But exit and voice only matter if they alter the buyer's options or force a forum to confront the underlying value conflict. If the government can substitute providers, move the work to an integrator, or rely on a cloud environment where the red line no longer travels, then exit at the model-provider layer matters little. It may be principled and perhaps have knock-on effects, but is not yet pluralistic.

The same point applies to reported corporate red lines, and runs in both directions. A red line that holds without external verification reproduces the Starlink failure mode: private veto over deployment, with whatever value commitments that provider happens to hold. A red line absorbed by substitution at another layer reproduces the Maven failure mode: voice or refusal at one layer, the work moving at another, and no forum with authority over the underlying use.

---

6. Implications for pluralistic alignment

Pluralistic alignment cannot stop at model behavior, firm commitments, or worker dissent. In national-security AI, disagreement matters when it is routed to an institution that can receive it, weigh it, and change what happens next.

Under frictionless security, public and private actors can align around capability, speed, secrecy, and lawful-use compliance while the institutions that would otherwise weigh those choices against other public values are absent or weak. The result may not be outright abuse, but something more ordinary: decisions move through the forums that exist but the values without institutional homes lose force.

The cases show useful patterns. Starlink shows private leverage without accountable architecture: SpaceX could refuse or condition use, but no visible forum weighed the public values at stake. Apple shows the opposite: refusal became pluralistically relevant because it entered court process, public technical evidence, and policy debate. Project Maven shows absorbed dissent: worker voice changed Google's participation but not the trajectory of the underlying capability. Frontier-AI defense contracting is the unresolved live case: reported corporate red lines may matter, but only if they bind the actors who can change deployment — the agency user, contracting officer, cloud provider, integrator, operator, and reviewer.

Three lessons follow for pluralistic-alignment work.

Forum. Pluralistic alignment should look at the institutions that receive disagreement. Which forum gets the objection? What evidence can that forum inspect? Can it delay, condition, disclose, escalate, or terminate a use? If not, the disagreement may still be morally important; it is not yet an accountable decision process.

Substitutability. Exit and voice are structurally conditional (Hirschman, 1970). If the buyer can substitute providers, exit at one firm can be absorbed — the Maven failure mode. If one provider controls a hard-to-substitute infrastructure layer, refusal can become private veto — the Starlink warning. Pluralism requires institutions that can turn disagreement into reasons, evidence, review, and consequences, and market and procurement structures that keep those institutions from being bypassed.

Evidence. Reported red lines and corporate commitments matter if outsiders can verify whether they bind the actors who deploy, integrate, procure, or oversee the system. Without external verification, principled positions remain private discretion.

The same tests apply one layer down. A model's refusals, and the spec or constitution behind them, are pluralist only when that document is publicly authored (Abiri, 2025), its adherence is checkable from outside (Chan, 2026), and it binds deployment; absent a forum, such an artifact shifts power toward the firm that wrote it (Kawakami et al., 2024) rather than constraining it. Output-facing pluralism (Sorensen et al., 2024) and institutional pluralism meet at the spec, and classification removes the conditions on both.

Leibo et al.'s (2024) principles — contextual grounding, community customization, continual adaptation, polycentric governance — are difficult to translate to national-security settings. Contextual grounding often depends on classified information; community customization is limited by mission structure; continual adaptation depends on after-action review and slow procurement updates; polycentric governance depends on whether firms, workers, courts, congressional committees, inspectors general, and technical reviewers have information and usable levers. Pluralistic-alignment principles need institutional infrastructure to operate, and this infrastructure is different and at times thin in national-security AI.

6.1. Design targets

Part of the design problem is making objections usable by institutions with authority. Four design targets follow. *Safeguards must travel through the stack*: a constraint that exists only in a model provider's policy is not yet a constraint on deployment. M-25-22 names the relevant acquisition tools — vendor lock-in protections, IP rights, data and model portability, ongoing testing and monitoring, independent evaluation, vendor performance terms — but those tools do not automatically govern the national-security cases that most need them (Toh, 2025). *Dissent must survive secrecy*: classified technical and ethical dissent needs lawful channels, building on precedents like the State Department Dissent Channel, mandatory adversarial review, and PCLOB-style cleared oversight (Deeks, 2020). *Evidence must be built for outsiders*: evaluations, red-team results, system cards, safety cases, audit logs, and incident reports should be designed so that someone other than the developer can use them to test a claim, contest a use, or impose a condition (Brundage et al., 2020; Raji et al., 2020; Shevlane, 2022). Incomplete contracting explains the necessity (Hadfield-Menell & Hadfield, 2019): no contract or model policy can specify every downstream use, and evidence has to help institutions fill the gaps. *Emergency exceptions must expire*: speed has legitimate uses (Walzer, 1977), but bounded friction — logs, sunset clauses, reason-giving, mandatory reauthorization, retrospective review — is what preserves later judgment without blocking urgent action (Deeks & Eichensehr, 2024; Kiener, 2025).

Foreign affected publics make the second-order legitimacy problem increasingly important. International humanitarian law, human-rights law, refugee law, targeting doctrine, and related regimes route their interests into legal constraints. But they often do not participate in the domestic institutions deciding how national-security AI is procured, tested, deployed, or reviewed. Their interests are represented through surrogates. That may be necessary under national-security conditions, but still has legitimacy consequences. A pluralistic-alignment account that treats this surrogate structure as sufficient has answered Kasirzadeh's second-order question — who has authority to choose, weigh, and operationalize values — by institutional default.

One objection is that national-security institutions exist precisely to hold some decisions apart from contestation, so their authorization should settle the question. Insulation can be legitimate, but only on conditions: Thompson's second-order publicity, that the decision to keep something secret is itself accountable (Thompson, 1999), and Colaresi's pairing of immediate secrecy with retrospective accountability (Colaresi, 2014). The forums, outsider-usable evidence, and expiring exceptions argued for here are that second-order structure; they do not replace ordinary governance, they are the accountability insulation already requires.

Rigorous audits need more than black-box access: weights, activations, training data, deployment detail (Casper et al., 2024). Even outside classified settings, most external evaluators receive only black-box API access, and gray- and white-box access remains rare (Brundage et al., 2026). Classification exacerbates the problem. The research task for this community is therefore *verification without disclosure*: methods that let a cleared but external party confirm a specific claim — that a model refuses a domestic-surveillance request, that the system in the field is the one that was audited, that an evaluation was run and returned what is reported — without exposing weights, data, or operations. Zero-knowledge attestations can prove that a model with private weights meets a stated red-line metric on agreed bench-

marks (South et al., 2024); trusted-execution-environment audits let a provider and an outside auditor publish attested results without trusting each other (Schnabl et al., 2025); and proposed hardware-enabled guarantees could bind a deployment to the reviewed model (Petrie et al., 2025). None of this is mature, and none of it creates its own demand: a buyer seeking models free of usage constraints will not ask for attestations. Creating that demand is a policy question — contractual requirements, mandated independent testing, reporting duties to Congress — and the levers are the ones this paper has traced. These mechanisms are not a substitute for a forum with authority; they are how the evidence such a forum needs can exist when the system itself cannot be shown.

National-security AI shows that pluralistic alignment is not just a model-design problem. It is a problem of institutional conversion. Public values survive when disagreement can become reasons, evidence, review, and consequences. Without that architecture, firm refusal is private discretion, worker voice is expression, and technical evidence is private knowledge.

7. Limitations

The paper is conceptual — further research should expand and deepen case study analysis. In particular, a large survey of national-security AI deployments and governance mechanisms is needed. Further, this paper did not engage deeply with international law and accountability. Future work will expand on this. The institutional analysis is also U.S.-centric; other democracies route national-security AI through different procurement and oversight structures, and the framework's portability there is untested.

References

Abels, J. Private infrastructure in geopolitical conflicts: The case of Starlink and the war in Ukraine. *European Journal of International Relations*, 2024. doi: 10.1177/13540661241260653.

Abiri, G. Public constitutional AI. *Georgia Law Review*, 59, 2025.

Abraham, Y. 'lavender': The AI machine directing israel's bombing spree in Gaza, April 3 2024. URL https://www.972mag.com/lavender-ai-israeli-army-gaza/.

Amodei, D. Statement from Dario Amodei on our discussions with the department of war. Anthropic blog post, February 26 2026. URL https://www.anthropic.com/news/statement-department-of-war.

Anthropic. Where things stand with the department of war. Anthropic blog post, March 5 2026. URL https://www.anthropic.com/news/where-stand-department-war.

Anthropic PBC v. U.S. Department of War. Order granting plaintiff's motion for preliminary injunction. No. 26-cv-01996-RFL (N.D. Cal.), March 26 2026. Order by Judge Rita F. Lin.

Apple Inc. A message to our customers, February 16 2016. URL https://www.apple.com/customer-letter/.

Avant, D. D. *The Market for Force: The Consequences of Privatizing Security*. Cambridge University Press, Cambridge, 2005.

Belfield, H. Activism by the AI community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 15–21. ACM, 2020. doi: 10.1145/3375627.3375814.

Bellovin, S. M., Blaze, M., Landau, S., and Pell, S. K. The real security issues of the iPhone case. *Science*, 352 (6292):1398–1399, 2016.

Berlin, I. Two concepts of liberty. In *Four Essays on Liberty*. Oxford University Press, Oxford, 1969. Originally delivered as inaugural lecture, 1958.

Bozeman, B. *Public Values and Public Interest: Counterbalancing Economic Individualism*. Georgetown University Press, Washington, DC, 2007.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G. K., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims, 2020.

Brundage, M., Dreksler, N., Homewood, A., McGregor, S., et al. Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies. arXiv:2601.11699, 2026. URL https://arxiv.org/abs/2601.11699.

Bullock, J. B., Hammond, S., and Krier, S. AGI, governments, and free societies. arXiv:2503.05710, 2025.

Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3 (1):1–12, 2016.

Buzan, B., Wæver, O., and de Wilde, J. *Security: A New Framework for Analysis*. Lynne Rienner Publishers, Boulder, CO, 1998.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

CBS News. Judge blocks Pentagon from labeling Anthropic AI a 'supply chain risk' and halts Trump's ban on federal use. CBS News, March 26 2026. URL https://www.cbsnews.com/news/anthropic-ruling-judge-trump-pentagon-ai/.

Chachko, E. National security by platform. *Stanford Technology Law Review*, 25(1):55, 2021.

Chan, A. Transparency into model spec adherence. Centre for the Governance of AI (GovAI), 2026.

Citron, D. K. and Pasquale, F. The scored society: Due process for automated predictions. *Washington Law Review*, 89:1, 2014.

Coglianese, C. and Lampmann, E. Contracting for algorithmic accountability. *Administrative Law Review Accord*, 6:175–199, 2021.

Colaresi, M. P. *Democracy Declassified: The Secrecy Dilemma in National Security*. Oxford University Press, 2014.

Conger, K. Google plans not to renew its contract for Project Maven, a controversial Pentagon drone AI imaging program. *Gizmodo*, June 1 2018. URL https://gizmodo.com/google-plans-not-to-renew-its-contract-for-project-mave-1826488620.

Dave, P. Google to scrub U.S. military deal protested by employees, source says. Reuters Technology, June 1 2018.

Deeks, A. S. Secrecy surrogates. *Virginia Law Review*, 106: 1395, 2020.

Deeks, A. S. *The Double Black Box: National Security, Artificial Intelligence, and the Struggle for Democratic Accountability*. Oxford University Press, Oxford and New York, 2025.

Deeks, A. S. and Eichensehr, K. E. Frictionless government and foreign relations. *Virginia Law Review*, 110:1815–1899, 2024.

Dickinson, L. A. Public law values in a privatized world. *Yale Journal of International Law*, 31:383, 2006.

Dickinson, L. A. *Outsourcing War and Peace: Preserving Public Values in a World of Privatized Foreign Affairs*. Yale University Press, New Haven, 2011.

Eichensehr, K. E. Public-private cybersecurity. *Texas Law Review*, 95(3):467, 2017.

Eichensehr, K. E. Digital switzerlands. *University of Pennsylvania Law Review*, 167(3):665–732, 2019.

Electronic Privacy Information Center. Apple v. FBI, 2016. URL https://epic.org/documents/apple-v-fbi-2/.

Frazee, J. R., Hickey, A. S., and Prairie, J. Anthropic supply chain risk designation takes effect: Latest developments and next steps for government contractors. Mayer Brown Insights, March 19 2026. URL https://www.mayerbrown.com/en/insights/publications/2026/03/anthropic-supply-chain-risk-designation-takes-effect--latest-developments-and-next-steps-for-government-contractors.

Freeman, J. and Minow, M. (eds.). *Government by Contract: Outsourcing and American Democracy*. Harvard University Press, Cambridge, MA, 2009.

Gerring, J. *Case Study Research: Principles and Practices*. Cambridge University Press, Cambridge, 2nd edition, 2017.

Hadfield-Menell, D. and Hadfield, G. K. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 417–422, 2019. URL https://www.aies-conference.com/2019/wp-content/papers/main/AIES-19_paper_231.pdf.

Hirschman, A. O. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, Cambridge, MA, 1970.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. Collective constitutional AI: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1395–1417. ACM, 2024. doi: 10.1145/3630106.3658979.

Jayanti, A. Starlink and the Russia-Ukraine war: A case of commercial technology and public purpose? Belfer Center for Science and International Affairs, Harvard Kennedy School, March 9 2023.

Jørgensen, T. B. and Bozeman, B. Public values: An inventory. *Administration & Society*, 39(3):354–381, 2007.

Kaminski, M. E. The right to contest AI. *Columbia Law Review*, 121(7):1957, 2021.

Kasirzadeh, A. Plurality of value pluralism and AI value alignment. In *Pluralistic Alignment Workshop, NeurIPS 2024*, 2024.

Kawakami, A., Wilkinson, D., and Chouldechova, A. Do responsible AI artifacts advance stakeholder goals? four key barriers perceived by legal and civil stakeholders. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2024.

Kiener, M. AI and responsibility: No gap, but abundance. *Journal of Applied Philosophy*, 2025. doi: 10.1111/japp.12765.

Krahmann, E. *States, Citizens and the Privatization of Security*. Cambridge University Press, Cambridge, 2010.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental AI development, 2025.

Landau, S. The real security risks of the iPhone case. *Science*, 352(6292), 2016.

Leibo, J. Z., Vezhnevets, A. S., Diaz, M., Agapiou, J. P., Cunningham, W. A., Sunehag, P., Haas, J., Koster, R., Duéñez-Guzmán, E. A., Isaac, W. S., Piliouras, G., Bileschi, S. M., Rahwan, I., and Osindero, S. A theory of appropriateness with applications to generative artificial intelligence, 2024.

Mansbridge, J., Bohman, J., Chambers, S., Christiano, T., Fung, A., Parkinson, J., Thompson, D. F., and Warren, M. E. A systemic approach to deliberative democracy. In Parkinson, J. and Mansbridge, J. (eds.), *Deliberative Systems: Deliberative Democracy at the Large Scale*, pp. 1–26. Cambridge University Press, Cambridge, 2012.

Mason, E. Value pluralism. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition, 2018.

Maurer, T. *Cyber Mercenaries: The State, Hackers, and Power*. Cambridge University Press, Cambridge, 2018.

Mazzucato, M. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*. Anthem Press, 2013.

Metzger, G. E. Privatization as delegation. *Columbia Law Review*, 103:1367, 2003.

Mulligan, D. K. and Bamberger, K. A. Procurement as policy: Administrative process for machine learning. *Berkeley Technology Law Journal*, 34(3):773–850, 2019. doi: 10.15779/Z38RN30793.

Office of Management and Budget. M-25-21: Accelerating federal use of AI through innovation, governance, and public trust. Memorandum for the Heads of Executive Departments and Agencies, April 3 2025a. URL https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf.

Office of Management and Budget. M-25-22: Driving efficient acquisition of artificial intelligence in government. Memorandum for the Heads of Executive Departments and Agencies, April 3 2025b. URL https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf.

OpenAI. Our agreement with the department of war. OpenAI blog post, February 28 2026. URL https://openai.com/index/our-agreement-with-the-department-of-war/.

Ovadya, A., Redman, K., Thorburn, L., Chen, Q. Z., Smith, O., Devine, F., Konya, A., Milli, S., Revel, M., Feng, K. J. K., Zhang, A. X., Chandra, B., Bakker, M. A., and Kasirzadeh, A. Position: Democratic AI is possible. the democracy levels framework shows how it might work. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Patel, F. and Toomey, P. C. Bringing transparency to national security uses of artificial intelligence. Brennan Center for Justice, April 9 2024. URL https://www.brennancenter.org/our-work/analysis-opinion/bringing-transparency-national-security-uses-artificial-intelligence.

Petrie, J., Aarne, O., Ammann, N., and Dalrymple, D. Flexible hardware-enabled guarantees for AI compute. arXiv:2506.15093, 2025.

Pettit, P. *Republicanism: A Theory of Freedom and Government*. Oxford University Press, Oxford, 1997.

Pettit, P. Republican freedom and contestatory democratization. In Shapiro, I. and Hacker-Cordón, C. (eds.), *Democracy's Value*, pp. 163–190. Cambridge University Press, Cambridge, 1999.

Pozen, D. E. Deep secrecy. *Stanford Law Review*, 62:257, 2010.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In

Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT)*, pp. 33–44. ACM, 2020.

Reuters. Pentagon awards $480 million deal to Palantir for 'Maven' prototype, May 29 2024. URL https://www.reuters.com/technology/palantir-wins-480-million-us-army-deal-maven-prototype-2024-05-29/.

Reuters. Google signs classified AI deal with Pentagon, The Information reports. Reuters, April 28 2026a.

Reuters. Palantir's Maven smart system contract ceiling rises to $1.3 billion, 2026b. URL https://www.reuters.com/technology/pentagon-adopt-palantir-ai-as-core-us-military-system-memo-says-2026-03-20/.

Reuters. OpenAI strikes deal with Pentagon hours after Trump admin bans Anthropic. Reuters, February 28 2026c.

Sabin, S. Anthropic: No 'kill switch' for AI in classified settings. Axios, April 22 2026. URL https://www.axios.com/2026/04/22/anthropic-no-kill-switch-ai-classified-settings.

Sagan, S. D. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton University Press, Princeton, 1993.

Schlanger, M. Intelligence legalism and the national security agency's civil liberties gap. *Harvard National Security Journal*, 6:112, 2015.

Schnabl, C. et al. Attestable audits: Verifiable AI safety benchmarks using trusted execution environments. arXiv:2506.23706, 2025.

Schooner, S. L. Contractor atrocities at Abu Ghraib: Compromised accountability in a streamlined, outsourced government. *Stanford Law and Policy Review*, 16:549–572, 2005.

Seawright, J. and Gerring, J. Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly*, 61(2):294–308, 2008.

Selbst, A. D. and Barocas, S. The intuitive appeal of explainable machines. *Fordham Law Review*, 87:1085, 2018.

Shevlane, T. Structured access: An emerging paradigm for safe AI deployment, 2022.

Singer, P. W. *Corporate Warriors: The Rise of the Privatized Military Industry*. Cornell University Press, Ithaca, 2003.

Snopes. Did Elon Musk turn off Starlink access in Crimea to disrupt Ukrainian attack?, September 14 2023. URL https://www.snopes.com/news/2023/09/14/musk-internet-access-crimea-ukraine/.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.05070.

South, T., Camuto, A., Jain, S., Nguyen, S., Mahari, R., Paquin, C., Morton, J., and Pentland, A. Verifiable evaluations of machine learning models using zkSNARKs. *arXiv preprint arXiv:2402.02675*, 2024.

Thompson, D. F. Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, 74(4):905–916, 1980.

Thompson, D. F. Democratic secrecy: The dilemma of accountability. *Political Science Quarterly*, 114(2):181–193, 1999.

Tiku, N. and De Vynck, G. Google employees ask CEO Sundar Pichai to bar classified military AI work. The Washington Post, April 27 2026. URL https://www.washingtonpost.com/technology/2026/04/27/google-employees-letter-ai-pentagon/.

Tillipman, J. Military AI policy by contract: The limits of procurement as governance. Lawfare, 2026.

Toh, A. Narrowing the national security exception to federal AI guardrails, June 30 2025. URL https://www.brennancenter.org/our-work/analysis-opinion/narrowing-national-security-exception-federal-ai-guardrails-0. First published at Lawfare.

U.S. Department of Defense. DoD adopts ethical principles for artificial intelligence. Press release, February 24 2020. URL https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

U.S. Department of Defense. DoD directive 3000.09: Autonomy in weapon systems. Department of Defense Directive, January 25 2023. URL https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

U.S. Department of War. Artificial intelligence strategy for the department of war: Accelerating america's military AI dominance. Memorandum from the Secretary of War,

January 9 2026a. URL https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.

U.S. Department of War. Classified networks AI agreements. Press release, 2026b. URL https://www.war.gov/News/Releases/Release/Article/4475177/classified-networks-ai-agreements/.

Vaughan, D. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press, Chicago, 1996.

Wæver, O. Securitization and desecuritization. In Lipschutz, R. D. (ed.), *On Security*, pp. 46–86. Columbia University Press, New York, 1995.

Walzer, M. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books, New York, 1977.

Washington Post. Elon Musk cut Starlink in middle of Ukraine attack on russia fleet. The Washington Post, September 7 2023.

Weber, M. Politics as a vocation. In Gerth, H. H. and Mills, C. W. (eds.), *From Max Weber: Essays in Sociology*. Oxford University Press, New York, 1946. Lecture delivered 1919.