DIAL E FOR ETHICAL ENFORCEMENT: INSTITUTIONAL VETO POWER AS A GOVERNANCE PRIMITIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

The persistent militarization of large reasoning models stems not from technical necessity but from governance arrangements that strip researchers of meaningful authority to refuse harmful transfers and deployments. Existing accountability mechanisms such as model cards and responsible AI statements operate as reputational signals detached from decision making architecture. We identify institutional veto power as a missing governance primitive: a formal authority to halt subsequent use or distribution of research when credible risks of weaponization emerge. Drawing on precedents in nuclear nonproliferation and biomedical ethics, the paper maps unprotected veto points across the research lifecycle, diagnose why compliance without enforceable constraints fails, and offer concrete institutional designs that embed veto authority while reducing the risk of political capture. The paper argues that communities most vulnerable to military uses must lead governance design, and that institutional veto power is a prerequisite for converting symbolic safeguards into enforceable responsibility and for achieving meaningful model disarmament.

1 THE GOVERNANCE PARADOX: WHY ETHICS WITHOUT POWER ENABLES MILITARIZATION

Contemporary AI governance suffers a paradox: widespread ethical awareness exists alongside near total incapacity to prevent harmful downstream uses. Current practices generate thorough risk knowledge, ethics statements, impact assessments, responsible AI codes, yet lack formal authority to block militarization or abusive deployment; this produces "organized irresponsibility," where every actor can disclaim control while harms proceed. Unlike post WWII nuclear policy or biomedical oversight, which embedded refusal authority into institutions, AI governance treats loss of control after publication as inevitable even though militarization follows identifiable decision chains and discretionary institutional choices Scharre (2023). We therefore recommend creating legally and procedurally protected institutional vetoes, targeted powers over transfer and deployment (not inquiry or publication), to condition use when credible risks such as militarization, surveillance misuse, or international law breaches are identified. Properly scoped, such refusal mechanisms would rebalance power toward researchers and affected communities and convert ethical knowledge into enforceable safeguards.

$$(K \wedge \neg A) \Rightarrow M \quad \text{and} \quad (K \wedge V \wedge R) \Rightarrow \neg M \quad \Rightarrow \quad P(M \mid V) \ll P(M \mid \neg V) \tag{1}$$

Notation. Here, K denotes ethical knowledge produced by governance practices, A denotes institutional authority to refuse or block use, M denotes militarization or abusive deployment, R denotes the identification of a credible risk, and V denotes a legally and procedurally protected veto over transfer or deployment.

2 RELATED WORK: VETO POWER IN GOVERNANCE

Current AI governance literature emphasizes either technical safety mechanisms (alignment, interpretability) or procedural ethics (review boards, impact assessments) Sahoo & Chhawacharia (2025). Yet this work treats militarization as a downstream consequence amenable to ethics mitigation, not

as an architectural problem requiring authority redistribution. The governance literature in other domains—nuclear nonproliferation, bioethics, labor rights—demonstrates that constraining specific uses requires veto authority, not merely visibility Crawford (2021); Kalluri et al. (2023). Our contribution bridges these literatures: *we argue that AI militarization is fundamentally a governance problem, not an ethics problem, and therefore requires mechanisms (veto power) that other high risk domains have already institutionalized.* Existing proposals for "Responsible AI" governance lack enforcement mechanisms precisely because they assume voluntary institutional compliance; veto power reverses this assumption, making restraint collectively enforceable. The dual use literature identifies risks but not solutions. We propose an operational mechanism that could actually arrest militarization trajectories at decision points where prevention remains possible. **By framing veto power as a governance primitive—a basic building block that other domains have proven workable—we shift discourse from "how can we encourage restraint" (eternally failing) to "how can we make restraint structurally rational" (institutionally achievable).**

3 Mapping Discretionary Militarization: Unprotected Veto Points

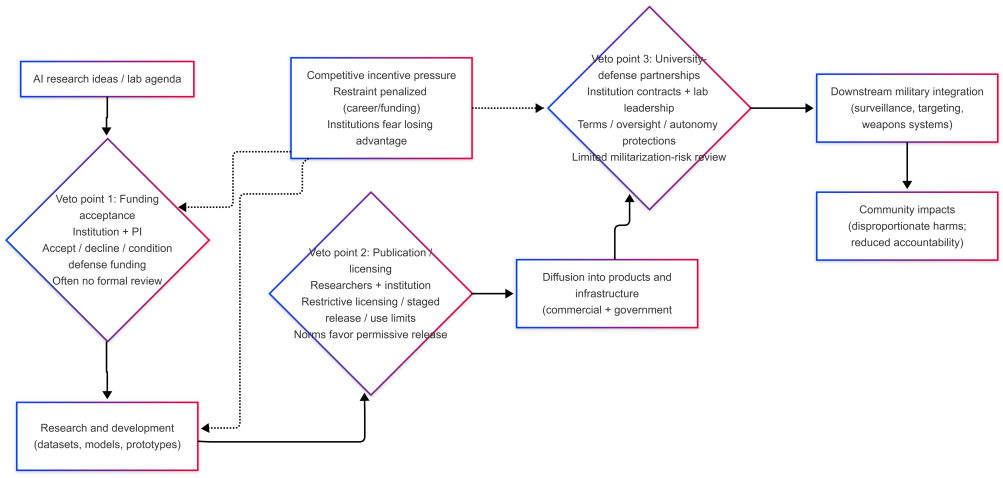

Figure 1: Workflow of discretionary militarization in AI research, highlighting institutional veto points and the competitive incentive pressures that reinforce downstream military integration and community harms.

Militarization of AI is driven by routine institutional choices that act as veto points: acceptance of military research funding, decisions about publication and licensing, and formal partnerships between universities and defense firms Lamparth et al. (2024). Each of these moments could be governed differently. Institutions could require review before accepting military grants, adopt restrictive licenses or license conditions that bar surveillance use, delay publication pending risk assessment, or subject partnership proposals to independent 3rd party oversight. Instead these choices are treated as individual researcher preferences, so universities routinely accept funding and sign agreements without deliberation about downstream military integration and researchers who resist face career penalties and loss of resourcesKhlaaf et al. (2024).

These veto points share a structural defect. Authority is exercised without meaningful participation from affected communities or institutional accountability, producing a collective action failure in which restraint is socially preferable but individually costly. Actors who might exercise restraint are undercut by competitors that do not. Addressing this requires formal governance mechanisms that distribute authority, mandate risk review, and protect those who refuse military integration so that the collectively rational choice of restraint becomes individually viable Future of Life Institute, 2016.

4 From Ethics Without Veto to Enforceable Governance

Current ethics frameworks fail for structural reasons that better guidelines cannot fix. Responsibility is dispersed across researchers, universities, philanthropists, tech transfer offices, and procurement chains, so no actor holds both the authority and obligation to stop militarization—and each can deflect blame when harms occur. Voluntary restraint also produces adverse selection: institutions that decline military funding, impose use restrictions, or protect research autonomy pay real costs while less restrained competitors gain advantage, pushing trajectories toward militarization regardless of ethical exhortation . Finally, ethics processes *enable laundering*: documentation of "due consideration" can psychologically and institutionally substitute for changed decisions, allowing actors to treat deliberation as discharge of responsibility even when outcomes remain unchanged.

These failure modes require architectural change. We propose institutional veto power: governance that makes ethical judgments binding, locates decision authority with those exposed to consequences, and turns restraint from an individually punished choice into a collectively enforceable rule. Veto is not arbitrary refusal but narrowly triggered by predefined thresholds—e.g., credible military integration, surveillance targeting marginalized groups, or violations of international law—paired with collective decision-making, appeal procedures, and anti-retaliation protections. Crucially, this does not require new treaties: universities can revise contracts and create veto committees; funding parties can condition grants on enforceable safeguards; conferences and repositories can require disclosure and enforce use conditions; and professional communities can adopt binding, public norms that incrementally embed veto governance within existing institutions Brenneis (2025).

5 Centering Justice: Veto Power Requires Marginalized Community Leadership

A core governance question concerns who determines which militarization risks justify veto authority and whose interests such decisions ultimately serve. Contemporary debates remain concentrated within elite academic and policy institutions, largely dominated by socially privileged actors, while communities that experience the most severe downstream harms remain structurally excluded. These include populations subjected to automated targeting and large scale civilian harm, minorities exposed to pervasive surveillance infrastructures, and migrant groups whose autonomy is constrained by biometric identification systems. This exclusion is not only normatively problematic but also epistemically limiting. Affected communities possess situated knowledge about system deployment, harm salience, and justice relevant tradeoffs that cannot be adequately represented through advisory consultation within externally designed governance frameworks Bengio et al. (2025). From a policy perspective, justice therefore requires operational community leadership rather than symbolic inclusion. This entails binding decision authority. In practice, veto bodies governing surveillance systems should include civil rights advocates and representatives from surveilled communities with decision making power equal to that of technical experts and ethicists. Similarly, governance arrangements for autonomous weapons should be co designed with researchers and affected populations in regions of deployment, rather than developed within wealthy institutions and later exported as policy. Research funding decisions shaping militarized contexts should formally incorporate community voices from those regions, with mechanisms that allow community objections to block projects rather than merely inform them. This implies structural reform toward community controlled governance in which affected populations directly hold veto authority.

Militarization and marginalization are deeply interlinked. Military violence disproportionately targets already marginalized groups, AI systems scale and automate existing targeting logics, and data extracted from these communities is used to train systems later deployed against them. Within this context, veto authority should function bidirectionally. Beyond constraining militarized trajectories, governance mechanisms should actively enable and protect research oriented toward peace and justice. Such efforts should be institutionally facilitated rather than incidentally tolerated. Veto based governance therefore performs a dual role: it blocks pathways toward weaponization while simultaneously safeguarding and accelerating peace oriented innovation. When affected communities hold veto authority, they can refuse militarized research agendas while directing institutional resources toward technologies that advance collective safety, autonomy, and long term justiceKhlaaf et al. (2024); Schwartz et al. (2022).

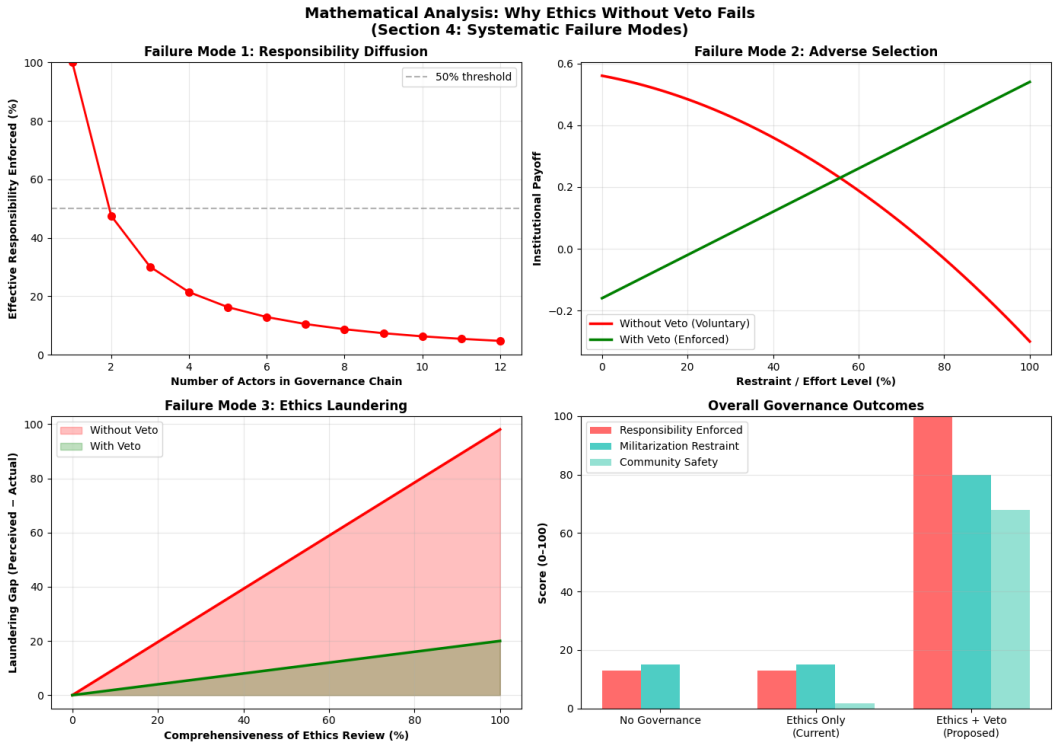

Figure 2: Why ethics without veto power fails to constrain militarization. (a) Responsibility diffusion: enforceable responsibility falls as governance chains lengthen. (b) Adverse selection: competition drives low restraint without enforcement; veto stabilizes higher restraint. (c) Ethics laundering: documentation inflates perceived responsibility without intervention absent enforceable authority. (d) Derived outcomes (responsibility enforced, restraint, community safety) under no governance, ethics-only, and ethics+veto regimes.

6 Limitations and Future Work

Veto power can make decisions about militarized research enforceable, but it does not by itself fix deeper power imbalances. It will not redistribute research funding to marginalized communities, democratize university governance, dismantle the defense industry that shapes priorities, or automatically hold researchers accountable, and implementation will encounter institutional inertia and the risk of capture by hostile actors. Empirical work is needed: document and analyze real veto decisions to build precedent, run longitudinal studies of research trajectories and weaponization timelines, center leadership and partnership with communities in the Global South who are directly affected, and study how veto mechanisms fit with export controls, international treaties, supplier transparency, and legal accountability so veto power does not become another way to ease guilt while leaving power structures intact.

7 Conclusion: The Choice Before Research Communities

We face a choice between continuing with governance that allows ethics to flourish as performance while harm accumulates and responsibility diffuses, or moving toward collective, binding restraint where decisions about research militarization are accountable and constrained. Veto power offers a structural path for that shift, so researchers need not rely on lone moral courage, institutions need not suffer competitive disadvantage for practicing restraint, and affected communities gain voice in decisions about their lives. Performative ethics are over; enforceable, community-led responsibility to prevent militarization and promote peace must begin.

LLM USAGE DISCLOSURE

This work employed large language models in a supporting capacity. Specifically, we used Claude 4.5 Haiku (Anthropic, 2024) for the following roles:

Writing Assistance. The LLM was just asked to remove artificial Overleaf latex errors.

Limitations of LLM Use. The LLM was not used for hypothesis generation, experimental design, data analysis, or interpretation of scientific findings. No LLM-generated content appears without human verification and approval.

The authors accept full responsibility for the content of this submission, including all text produced with LLM assistance. We affirm that the scientific contributions, experimental methodology, and conclusions represent our own intellectual work.

REFERENCES

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report, 2025. URL https://arxiv.org/abs/2501.17805.

Andreas Brenneis. Assessing dual use risks in ai research: necessity, challenges and mitigation strategies. *Research Ethics*, 21(2):302–330, 2025. doi: 10.1177/17470161241267782. URL https://doi.org/10.1177/17470161241267782.

Kate Crawford. *Atlas of AI*. Yale University Press, New Haven, 2021. ISBN 9780300252392. doi: doi:10.12987/9780300252392. URL https://doi.org/10.12987/9780300252392.

Future of Life Institute. Autonomous weapons open letter: Ai & robotics researchers. URL https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/. Open letter hosted by the Future of Life Institute.

Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini, and Abeba Birhane. The surveillance ai pipeline, 2023. URL https://arxiv.org/abs/2309.15084.

Heidy Khlaaf, Sarah Myers West, and Meredith Whittaker. Mind the gap: Foundation models and the covert proliferation of military intelligence, surveillance, and targeting, 2024. URL https://arxiv.org/abs/2410.14831.

Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations, 2024. URL https://arxiv.org/abs/2403.03407.

Subramanyam Sahoo and Aditi Chhawacharia. The last vote: A multi-stakeholder framework for language model governance, 2025. URL https://arxiv.org/abs/2511.13432.

P. Scharre. *Four Battlegrounds: Power in the Age of Artificial Intelligence*. W. W. Norton & Company, New York, 2023. Print edition.

Sebastian Schwartz, Laura Gianna Guntrum, and Christian Reuter. Vision or threat—awareness for dual-use in the development of autonomous driving. *IEEE Transactions on Technology and Society*, 3(3):163–174, 2022. doi: 10.1109/TTS.2022.3182310.

APPENDIX

A TOY MODEL: WHY ETHICS WITHOUT VETO POWER FAILS

A.1 NOTATION

Let n be the number of actors in a governance chain. Let restraint be $r \in [0, 1]$ and documentation (ethics review) be $D \in [0, 1]$. Let baseline harm be $H_0 > 0$. Let $\alpha \in (0, 1)$ denote a per-actor coordination probability. We use a normalized total responsibility R_{tot}.

A.2 FAILURE MODE 1: RESPONSIBILITY DIFFUSION

$$R_{\text{ind}}(n) = \frac{R_{\text{tot}}}{n}, \tag{2}$$

$$p(n) = \alpha^{\,n-1}, \tag{3}$$

$$\rho(n) = \frac{R_{\text{ind}}(n)\, p(n)}{R_{\text{tot}}} = \frac{\alpha^{\,n-1}}{n}. \tag{4}$$

Here $\rho(n) \in [0, 1]$ is the effective enforceability fraction. With a binding veto point, enforceability is consolidated:

$$\rho_{\text{veto}} = 1. \tag{5}$$

A.3 FAILURE MODE 2: ADVERSE SELECTION UNDER VOLUNTARY RESTRAINT

Let $\bar{r}$ be the average restraint of competitors. A stylized payoff without veto is

$$U_{\text{noveto}}(r; \bar{r}) = \beta(1 - r)\bar{r} + \gamma r - \kappa r^2, \tag{6}$$

where $\beta > 0$ captures militarization/funding advantage, $\gamma \geq 0$ captures ethical signaling, and $\kappa > 0$ is the cost of restraint.

A best response and symmetric equilibrium are defined by

$$\text{BR}(\bar{r}) = \arg\max_{r \in [0,1]} U_{\text{noveto}}(r; \bar{r}), \tag{7}$$

$$r^*_{\text{noveto}} = \text{BR}(r^*_{\text{noveto}}). \tag{8}$$

With veto enforcement at level $\bar{R} \in [0, 1]$,

$$r^*_{\text{veto}} = \bar{R}. \tag{9}$$

A.4 FAILURE MODE 3: ETHICS LAUNDERING AND INTERVENTION

Perceived responsibility increases with documentation:

$$P(D) = 100D. \tag{10}$$

Actual harm prevented depends on intervention $I \in [0, 1]$:

$$A(I) = H_0 I. \tag{11}$$

The ethics-laundering gap is

$$G(D, I) = P(D) - A(I). \tag{12}$$

We link documentation to intervention by enforceability and equilibrium restraint:

$$I = D \cdot \rho \cdot r^*, \tag{13}$$

where $\rho \in \{\rho(n), \rho_{\text{veto}}\}$ and $r^* \in \{r^*_{\text{noveto}}, r^*_{\text{veto}}\}$.

Resulting harm is

$$H = H_0(1 - I). \tag{14}$$

A.5 Derived summary metrics

We report:

$$\text{ResponsibilityEnforced} = 100\rho, \tag{15}$$

$$\text{MilitarizationRestraint} = 100r^*, \tag{16}$$

$$\text{AdverseSelectionScore} = 100(1 - r^*), \tag{17}$$

$$\text{HarmToCommunities} = H, \qquad \text{CommunitySafety} = 100 - H. \tag{18}$$

Model parameters ("magic numbers") and scenario defaults

Python parameter dataclasses used in the replication code.

ModelParams (default coefficients)

Symbol / name	Default	Meaning
α (`alpha`)	0.95	Per-actor coordination probability (diffusion model)
R_{tot} (`total_responsibility`)	100	Responsibility normalization
β (`beta`)	0.8	Militarization / funding advantage coefficient
γ (`gamma`)	0.3	Ethical prestige / signaling coefficient
κ (`kappa`)	0.6	Quadratic cost of restraint
q (`quality_coeff`)	0.7	Prestige from research effort/quality (with veto)
c_v (`veto_cost_coeff`)	0.2	Shared cost of enforced governance (with veto)
H_0 (`baseline_harm`)	100	Baseline harm scale if no effective intervention

GovernanceScenario (default scenario fields)

Field	Default	Meaning
`name`	—	Scenario label
n (`num_actors`)	6	Number of actors in governance chain
D (`documentation_level`)	0.0	Ethics documentation/review level, $D \in [0, 1]$
`has_veto`	`False`	Whether binding veto enforcement exists
$\bar{R}$ (`veto_enforcement_level`)	0.0	Enforced restraint level if veto exists, $\bar{R} \in [0, 1]$

B Veto mechanisms in comparative governance contexts

This appendix establishes that veto power is not a novel or untested governance approach. Rather, it represents a well-established mechanism that has successfully constrained dangerous technologies in other high-risk domains. We examine three historical and contemporary examples: nuclear non-proliferation regimes, institutional review boards in biomedical research, and corporate compliance in export control regimes.

B.1 Nuclear Nonproliferation Regimes

Historical Context and Necessity

The Nuclear Non-Proliferation Treaty (NPT), negotiated in 1968 and entered into force in 1970, emerged from recognition that uncontrolled nuclear weapons proliferation posed existential risk. Following the development of nuclear weapons by the Soviet Union (1949) and subsequent weapons development by additional states, the international community recognized that preventing further proliferation required not merely ethical exhortation but institutional mechanisms with enforcement authority.

The NPT created a legal framework establishing three categories of states: Nuclear Weapons States (the five permanent UN Security Council members), Non-Nuclear Weapons States, and threshold states capable of developing weapons. The treaty created asymmetric obligations: Nuclear Weapons States committed (in principle) to disarmament; Non-Nuclear Weapons States committed to forgo weapons development; and all states committed to allowing international inspection.

The International Atomic Energy Agency (IAEA): Veto Power Through Verification

The IAEA, established in 1957 and strengthened by the NPT framework, represents the institutional embodiment of veto power in nuclear governance. The IAEA possesses several critical authorities that function analogously to the veto mechanisms proposed in this paper:

1. **Inspection Authority**: The IAEA maintains the right to inspect nuclear facilities in Non-Nuclear Weapons States at declared locations and, in many cases, at undeclared locations suspected of weapons development. These inspections are not advisory; they are mandatory conditions of NPT membership. The IAEA inspectorate can refuse to declare a state "in compliance" if inspections reveal violations or lack of transparency. This refusal to certify compliance effectively blocks a state's standing in the international community and triggers scrutiny and sanctions.
2. **Materials Tracking**: The IAEA tracks nuclear materials (uranium, plutonium, enriched fuel) as they move through the fuel cycle. Unexplained disappearance of fissile materials or transfer to undeclared facilities triggers investigation and can lead to referral to the UN Security Council for sanctions.
3. **Technology Transfer Restrictions**: The IAEA maintains safeguards on sensitive nuclear technology, restricting the circumstances under which uranium enrichment technology, plutonium reprocessing technology, or other weapons-relevant capabilities may be transferred internationally. These restrictions function as gatekeeping mechanisms preventing proliferation.
4. **Enforcement Through Referral**: While the IAEA itself lacks enforcement authority (it cannot impose sanctions), persistent violations trigger mandatory referral to the UN Security Council, which can impose sanctions, travel bans, asset freezes, and military intervention. This creates automatic escalation consequences for violation.

Why Nuclear Nonproliferation Succeeded (Partially)

The nuclear nonproliferation regime has not eliminated weapons proliferation—the number of nuclear weapons states has increased from 5 to 9 since the NPT entered into force. However, it has constrained proliferation significantly below models predicting 20-30+ weapons states by the 1980s. The regime succeeded to this degree because:

- **Verification was mandatory, not voluntary**: States could not choose whether to allow inspections; inspections were binding conditions of participation.
- **Veto points were clearly defined**: Uranium enrichment, plutonium reprocessing, and weapons-relevant technology transfer were identified as specific decision points where control could be exercised.

- **Consequences were automatic**: Violations triggered referral to the Security Council, which imposed sanctions without requiring additional political negotiation.
- **Enforcement was international**: No single state could unilaterally shield violators; enforcement required coordination across multiple states, reducing the likelihood of capture by individual interests.

LIMITATIONS AND IMPLICATIONS FOR AI GOVERNANCE

The nuclear nonproliferation regime also reveals limitations relevant to AI governance design:

- **Enforcement gaps**: States with nuclear weapons or Security Council vetoes can shield themselves or allies from referral and sanctions, limiting enforcement against major powers. This suggests that AI veto governance should include mechanisms preventing powerful actors from shielding themselves.
- **Inspection difficulty**: Detecting undeclared nuclear programs remains challenging; the IAEA's detection of **Iraq's covert program (1991) and Iran's nuclear work** depended partly on intelligence agencies and whistleblowing rather than purely technical verification.
- **Treaty withdrawal**: States can withdraw from the NPT with limited consequences; North Korea withdrew in 2003. This suggests that voluntary governance frameworks have inherent fragility.

IAEA Authority Mechanisms Relevant to AI

The IAEA's specific operational mechanisms offer concrete models for AI veto governance (see Table 1).

Nuclear Mechanism	Purpose	AI Analog
Mandatory inspections	Verify compliance with restrictions	Repository audits verify permitted use restrictions.
Materials accounting	Track fissile materials through the fuel cycle	Artifact-tracking logs record access and usage.
Technology controls	Restrict transfer of enrichment technology	License restrictions limit downstream licensing.
Referral to higher authority	Escalate violations to political bodies	Funding-agency enforcement ties compliance to grants.

Table 1: IAEA mechanisms and AI governance analogs

B.2 INSTITUTIONAL REVIEW BOARDS IN BIOMEDICAL RESEARCH

HISTORICAL CONTEXT: THE TUSKEGEE CASE AND REGULATORY RESPONSE

Institutional Review Boards (IRBs) emerged from historical horror. The Tuskegee Syphilis Study (1932–1972), conducted by the U.S. Public Health Service, enrolled African American men with syphilis into a study ostensibly to track the natural history of the disease. In reality, participants were not informed of their diagnosis, were not offered treatment even after penicillin became available (1943), and were subjected to deceptive medical procedures. The study continued for 40 years, resulting in preventable illness and death among study subjects. The scandal, publicly exposed in 1972, triggered comprehensive regulatory reform.

The regulatory response, codified in the Belmont Report (1979) and subsequent federal regulations (45 CFR 46), established that institutions receiving federal research funding must establish Institutional Review Boards with authority to review, approve, and halt research involving human subjects. IRBs operate at the threshold of research: before researchers begin studies, proposed research must be reviewed by a committee including scientists, ethicists, community representatives, and legal experts.

IRB AUTHORITY AND SCOPE

IRBs possess genuine veto authority within their defined scope:

1. **Pre-Implementation Review**: IRBs review research proposals before implementation. This timing is critical: it allows IRBs to block research before resources are committed and before recruitment proceeds.
2. **Binding Authority**: IRB decisions are binding. Researchers cannot proceed without IRB approval; attempting to do so violates federal regulation and triggers sanctions against the researcher and institution.
3. **Protected Refusal**: Federal regulation protects IRBs and researchers from retaliation for IRB decisions. Institutions cannot punish researchers whose protocols are rejected.
4. **Continuing Oversight**: IRBs retain authority throughout a study, conducting ongoing review of protocols, reviewing adverse events, and requiring corrective action if risks exceed expected levels.

WHY IRBS SUCCEEDED IN CONSTRAINING RESEARCH HARMS

IRBs have demonstrably constrained research harms compared to the pre-IRB era. Studies now rarely involve the level of deception and harm that characterized Tuskegee or Nazi medical experiments. This success stems from:

- **Authority at the decision point**: By reviewing before research begins, IRBs operate at the moment when research trajectories can be most easily altered.
- **Protected collective deliberation**: IRBs bring multiple perspectives (scientific, ethical, community) to bear on proposals.
- **Routine, normalized process**: IRB review is standard practice in biomedical research, institutionalized rather than extraordinary.
- **Regulatory backing**: IRBs are backed by federal regulation and funding consequences. Institutions that fail to maintain functional IRBs lose federal research funding eligibility.

LIMITATIONS OF IRBS

IRBs also reveal significant limitations relevant to AI governance design:

1. **Scope Limitation**: IRBs focus exclusively on risks to research subjects—direct, identifiable individuals participating in studies. They have no authority over risks to populations not in the study (third-party harms), societal harms, or dual-use concerns.
2. **Implementation Versus Use**: IRBs review the research process itself but do not review downstream uses of research findings. Once research is published, IRBs have no authority to restrict how others use the knowledge.
3. **Variable Quality and Authority**: IRB quality varies significantly. Underfunded IRBs with limited expertise may fail to identify harms or may be captured by institutional interests.
4. **Institutional Capture**: While formal protections exist against retaliation, institutional culture can undermine IRB authority.

IRB Mechanisms Applicable to AI Veto Governance

See Table 2 for specific operational mechanisms applicable to AI governance.

B.3 CORPORATE COMPLIANCE AND EXPORT CONTROL REGIMES

EXPORT CONTROL FRAMEWORK AND DUAL-USE TECHNOLOGY GOVERNANCE

Export control regimes represent a third governance domain where veto-like mechanisms have been operationalized at scale. The International Traffic in Arms Regulations (ITAR), the Export Administration Regulations (EAR), and parallel regimes in other countries, establish that certain technologies

IRB Mechanism	Purpose	AI Analog
Pre-implementation review	Assess protocols before initiation	Pre-licensing review prior to publication.
Institutional requirement	Mandate review as condition of funding	Funding conditional on veto governance.
Protected refusal	Protect from retaliation	Anti-retaliation safeguards for objectors.
Continuing oversight	Review throughout implementation	Ongoing deployment monitoring.
Documented procedures	Transparent criteria and procedures	Predefined veto triggers and workflows.
Pluralistic composition	Multiple perspectives in review	Diverse veto bodies with varied expertise.

Table 2: IRB mechanisms and AI governance analogs

cannot be transferred internationally without government authorization. These regimes recognize that technology transfer represents a decision point where control can and should be exercised.

Specific Export Control Mechanisms

1. **Commodity Identification**: Export control regimes maintain detailed lists specifying which technologies are controlled. Companies must determine whether their products fall within controlled categories and, if so, comply with licensing requirements.
2. **Licensing Requirements**: Transfer of controlled items requires government authorization. Companies submit license applications specifying the technology, the proposed recipient, the intended end use, and the end user.
3. **End-Use Commitments**: Licensees must commit to specific end uses and are prohibited from transferring items to third parties without subsequent authorization.
4. **Verification and Audit**: Companies maintaining export licenses are subject to government audits verifying compliance with license conditions.
5. **Penalties for Violation**: Unauthorized export of controlled items triggers significant penalties: civil penalties (fines up to $300,000 per violation), criminal penalties (up to 20 years imprisonment for willful violation).

Why Export Controls Succeeded

Export control regimes have demonstrably constrained the transfer of sensitive military technologies to hostile regimes, though imperfectly. Successes include:

- **Slowed weapons development**: Export controls on advanced semiconductor manufacturing equipment delayed China's development of advanced semiconductors, constraining military applications.
- **Prevented technology cascades**: Export controls on encryption technology slowed proliferation to adversarial states.
- **Deterred proliferation**: Companies face significant penalties for unauthorized export; this creates incentives to comply with restrictions.

Export Control Mechanisms Applicable to AI Veto Governance

See Table 3 for specific export control mechanisms applicable to AI governance.

Integration Across Domains: Lessons for AI Veto Governance

Examining nuclear nonproliferation, IRBs, and export controls together reveals common principles applicable to AI governance:

1. **Veto operates at decision points, not at knowledge generation**: Nonproliferation restricts uranium enrichment, not uranium chemistry. IRBs restrict research implementation, not inquiry. Export controls restrict transfer, not publication. In each case, veto operates at a specific decision point where control remains possible without censoring knowledge itself.

Export Control Mechanism	Purpose	AI Analog
Commodity lists	Specify controlled technologies	Define dual-use research categories and flagged capabilities.
License requirements	Authorize transfers	Pre-licensing review prior to transfers.
End-use commitments	Restrict recipient uses	Use restrictions encoded in licenses and agreements.
Verification and audit	Monitor compliance	Deployment attestations, audits, and provenance checks.
Penalties for violation	Impose consequences	Funding ineligibility, sanctions, and liability.

Table 3: Export control mechanisms and AI governance analogs

2. **Authority must be binding and enforceable**: Successful governance regimes embed veto authority in formal decision-making structures with binding effect. Recommendations, guidelines, and voluntary frameworks consistently fail to constrain motivated actors.
3. **Consequences must be material and automatic**: Violations trigger tangible consequences (sanctions, loss of funding, criminal liability) rather than relying on reputation or goodwill. Automatic consequences prove more effective than discretionary penalties.
4. **Governance is politically difficult but necessary**: All three regimes faced and continue to face resistance from actors who benefit from unconstrained use of technology. Yet all three have achieved partial success at constraining dangerous applications.

C IMPLEMENTATION PATHWAYS - DETAILED MECHANISMS

Veto power becomes operationally effective when embedded in concrete institutional arrangements. This section details five specific implementation pathways through which veto governance can be implemented using existing institutional structures. Critically, these pathways do not require new governmental authority, new treaties, or new institutions.

C.1 University Contract Modifications: Embedding Veto in Technology Transfer

Current Landscape

Universities currently manage significant research portfolios, including dual-use research with foreseeable military applications. Technology transfer offices (TTOs) license research to downstream organizations, negotiate research partnerships, and manage intellectual property. These decisions are typically made without institutional deliberation about downstream military integration.

Mechanism: Veto Clauses in Licensing Agreements

Universities can embed veto authority by modifying standard licensing agreements to include contractual clauses specifying:

1. **Use Restrictions**: Licenses can specify that research will not be transferred for military use, surveillance targeting of marginalized groups, or deployment in violation of international humanitarian law.
2. **Institutional Review Requirement**: Licenses can require that transfers to third parties undergo institutional review before the transfer can occur.
3. **Deployment Reporting**: Licenses can require that licensees report on deployment contexts, provide documentation of end uses, and notify the university if deployment deviates from licensed purposes.
4. **Indemnification for Violation**: Licenses can specify that licensees assume legal liability if they deploy research in violation of license restrictions.

5. **Right to Audit**: Licenses can grant universities the right to conduct audits of licensees' facilities and deployment contexts, verifying compliance with use restrictions.

IMPLEMENTATION EXAMPLE

Suppose a university develops facial recognition technology with clear surveillance applications. Under veto governance, a modified licensing agreement might specify:

> "*The research may not be deployed for mass surveillance targeting political opponents, ethnic minorities, or other vulnerable populations without university approval. Licensee must report annually on deployment contexts. University maintains the right to audit licensee's facilities and deployment contexts annually.*"

PRACTICAL IMPLEMENTATION REQUIREMENTS

Implementing veto clauses in licensing agreements requires:

- **Policy Development**: Universities must develop clear policies specifying which research triggers use restrictions and what those restrictions are.
- **Technology Transfer Office Training**: TTOs must be trained to identify dual-use research and to incorporate veto clauses into licensing agreements.
- **Legal Support**: Universities should establish legal capacity to draft and enforce use restriction clauses.
- **Funding Agency Support**: Funding agencies can incentivize veto governance by conditioning grants on institutional establishment of use restriction policies.
- **Researcher Communication**: Researchers should be informed that licenses will include use restrictions.

LIMITATIONS AND CHALLENGES

University implementation faces several challenges:

- **Revenue Loss**: Broad use restrictions may reduce licensing revenue.
- **International Complexity**: Technology transfer across national boundaries involves multiple legal jurisdictions.
- **Monitoring Burden**: Verifying licensee compliance requires institutional capacity for monitoring and enforcement.
- **Researcher Resistance**: Researchers may perceive use restrictions as constraints on research freedom.

C.2 CONFERENCE-LEVEL GOVERNANCE: DISCLOSURE AND VISIBILITY

CURRENT LANDSCAPE

Major AI conferences (NeurIPS, ICML, ICLR, ICRA, FAccT) serve as primary venues for research dissemination. Currently, conferences do not systematically assess dual-use risks or impose conditions on publication.

MECHANISM: DUAL-USE DISCLOSURE AND USE CONDITIONS

Conferences can implement lighter-touch governance by requiring disclosure of foreseeable dual-use risks and by enabling researchers to attach use conditions to published work:

1. **Dual-Use Disclosure Forms**: Conference submission requirements can include a "dual-use and governance" field where authors specify foreseeable military applications, circumstances under which they object to deployment, and recommended use conditions.

2. **Use Conditions and Licensing**: Research published at conferences can include explicit use conditions visible to readers. These conditions might specify: "This research should not be deployed for autonomous weapons lacking meaningful human control."
3. **Repository Integration**: Conference repositories can incorporate use condition metadata, making conditions machine-readable.
4. **Community Norms and Citation**: Conferences can establish professional norms that citing or building on research in violation of stated use conditions is ethically problematic.

Implementation Requirements

Conference-level implementation requires:

- **Policy Development**: Conference steering committees must develop policies specifying what dual-use disclosure requires.
- **Review Procedures**: Reviewers and program committees need training to assess dual-use disclosure forms.
- **Repository Infrastructure**: Conference repositories need technical capacity to incorporate use condition metadata.
- **Community Communication**: Conferences should communicate to authors and readers that dual-use disclosure is expected.
- **Transparency**: Policies and disclosure forms should be publicly available.

C.3 Funding Agency Requirements: Leverage Through Grant Conditions

Current Landscape

Funding agencies (NSF, DARPA, DOE, and international equivalents) distribute billions of dollars annually supporting research. Funding agencies exercise significant power over research directions through funding priorities and selection decisions. However, funding agencies rarely condition grants on governance requirements.

Mechanism: Governance Requirements as Grant Conditions

Funding agencies can condition grants on institutional establishment and maintenance of veto governance structures:

1. **Dual-Use Research Identification**: Grant conditions can require that institutions identify research falling within designated dual-use categories and subject such research to veto governance.
2. **Veto Body Establishment**: Grant conditions can require that institutions establish veto committees with specified composition, documented procedures, and protected decision-making authority.
3. **Review Requirements**: Grants can require that dual-use research be subject to veto body review before publication, licensing, or partnership formation.
4. **Funding Ineligibility for Violations**: Grant conditions can specify that institutions violating veto governance lose eligibility for future funding.
5. **Transparency and Reporting**: Grants can require that institutions maintain public records of veto decisions.

Implementation Example

Suppose the National Science Foundation conditions grants on institutional governance:

> "*Institutions must establish a Dual-Use Research Governance Committee. Committee must include at least 2 faculty, 1 ethicist/policy expert, and 2 representatives from potentially affected communities. All research supported by this grant involving autonomous systems must be reviewed by the Committee before publication, licensing, or commercial partnership. Institutions must maintain public records of committee decisions.*"

Advantages of Funding Agency Implementation

Funding agency requirements offer several advantages:

- **Leverage**: Funding agencies exercise significant leverage over institutions through grant awards.
- **Scale**: Conditions imposed by major funding agencies affect research across multiple institutions.
- **Precedent**: Funding agencies have successfully imposed governance requirements in other domains (IRB requirements for human subjects research).
- **Flexibility**: Funding agencies can develop requirements tailored to different research domains.

Limitations and Challenges

Funding agency implementation faces challenges:

- **Political Opposition**: Requirements may face political opposition from researchers, universities, and defense contractors.
- **International Coordination**: Without international coordination, national funding requirements become less effective.
- **Non-Compliance and Evasion**: Institutions might nominally comply while maintaining governance structures incapable of actually constraining militarization.
- **Institutional Autonomy**: Universities may resist funding conditions perceived as governmental overreach.

C.4 Repository-Level Artifact Governance: Infrastructure-Based Enforcement

Current Landscape

AI research artifacts—code, models, datasets—are increasingly released through online repositories: Hugging Face, GitHub, ModelHub, arXiv, Zenodo. These repositories serve as primary distribution channels for research artifacts. Repositories currently function as neutral distribution channels without systematic governance of downstream uses.

Mechanism: Machine-Readable Use Restrictions and Licensing

Repositories can require that artifacts include machine-readable use restrictions specifying permitted and prohibited uses:

1. **License Standardization**: Artifacts can be released under licenses (based on GPL, MIT, or custom licenses) that explicitly specify use conditions. License text can be standardized using SPDX expressions.
2. **Metadata Embedding**: Models and datasets can include metadata fields specifying use conditions and governance requirements.
3. **Access Restrictions**: Repositories can implement access control tiers. Some artifacts might be available for research access, while deployment access requires additional authorization.

4. **Version Control and Update Requirements**: Repositories can require that organizations deploying artifacts under specified conditions maintain current versions.

MECHANISM: AUDIT AND DELISTING

Repositories can implement compliance monitoring:

1. **Deployment Reporting**: Users accessing artifacts for deployment can be required to submit deployment context information.
2. **Audit Rights**: Repositories can assert rights to audit artifact uses, requiring deploying organizations to provide documentation.
3. **Violation Response**: If repositories determine that artifacts are being deployed in violation of use conditions, repositories can delist artifacts or revoke access.
4. **Public Transparency**: Repositories can maintain public records of use violations and enforcement actions.

ADVANTAGES OF REPOSITORY-LEVEL GOVERNANCE

Repository governance offers advantages:

- **Infrastructure Leverage**: Repositories already control access and distribution.
- **Technical Feasibility**: Machine-readable metadata and access control are established technical capabilities.
- **Practical Friction**: Requiring users to certify compliance creates practical friction that increases consideration of restrictions.
- **Scalability**: A single repository governance decision affects all users accessing that repository.
- **Transparency**: Repository-level governance is relatively transparent.

LIMITATIONS AND CHALLENGES

Repository-level governance faces challenges:

- **Enforcement Limits at National Borders**: Repositories cannot enforce restrictions on users outside their legal jurisdiction.
- **Code Forking and Redistribution**: Repositories cannot prevent users from downloading artifacts, modifying them, and redistributing outside repository control.
- **Scientific Knowledge Cannot Be Restricted**: Repositories can restrict specific code or model implementations but cannot restrict underlying scientific knowledge.
- **Monitoring Burden**: Auditing deployments at scale requires significant resources.
- **User Deception**: Organizations can misrepresent intended uses or use restricted artifacts while evading deployment reporting.

C.5 COMMUNITY-BASED NORMS AND PROFESSIONAL SOLIDARITY

CURRENT LANDSCAPE

Professional research communities have established ethics statements and codes of conduct. However, these statements typically lack enforcement mechanisms and are developed within elite academic institutions without meaningful participation from affected communities.

Mechanism: Professional Commitments and Collective Accountability

Veto governance can be operationalized through professional norms and collective community commitments:

1. **Professional Codes of Conduct**: Research professional associations can establish binding codes of conduct specifying that members commit to refusing military funding or to imposing use restrictions on dual-use work.
2. **Institutional Commitments**: Research institutions can establish public commitments to governance of dual-use research, specifying that they will not accept certain types of military funding.
3. **Community Oversight and Accountability**: Research communities can establish transparency mechanisms where members can raise concerns about other members' research or institutional practices.
4. **Community-Led Governance Design**: Research communities should center affected communities in governance design, treating governance research as legitimate scholarly work.
5. **Sanctuary Provisions**: Research communities can establish commitments to protect researchers who refuse military funding.

Advantages of Community-Based Governance

Community-based approaches offer advantages:

- **Legitimacy from Within**: Governance emerging from research communities has legitimacy that external structures may lack.
- **Flexibility and Adaptation**: Professional norms can be refined through community deliberation.
- **Low Institutional Burden**: Does not require new governmental authority or institutions.
- **Centering Affected Communities**: Professional norms can explicitly center affected communities.
- **Peer Accountability**: Professional peers understand research and can make nuanced judgments.

Limitations and Challenges

Community-based governance faces challenges:

- **No Enforcement Authority**: Absent institutional backing, professional norms are not binding.
- **Freeriding and Competition**: Individual researchers or institutions that violate codes gain competitive advantages.
- **Professional Heterogeneity**: Not all researchers identify with professional codes or associations.
- **International Variance**: Professional norms vary across countries and research communities.
- **Limited Reach to Practitioners**: Professional codes primarily constrain academic research.

Combining Implementation Pathways

The five implementation pathways are most effective when combined. A comprehensive approach might involve:

- Funding agencies condition grants on institutional governance

- Universities embed veto clauses in licenses and establish governance committees
- Conferences require dual-use disclosure
- Repositories implement use restriction metadata and compliance monitoring
- Professional communities establish codes of conduct and protect researchers exercising veto authority

D DESIGN PRINCIPLES FOR LEGITIMATE VETO GOVERNANCE

Veto power can be abused. Governance structures designed without safeguards against abuse can become tools for suppression of legitimate research, for politicization of governance, or for discrimination against vulnerable populations. This section articulates design principles that constrain veto authority while preserving its capacity to prevent militarization.

D.1 SCOPE LIMITATION AND PREDEFINED THRESHOLDS

PROBLEM: VETO WITHOUT BOUNDARIES

Unconstrained veto authority is dangerous. If veto bodies can refuse research for any reason or can continuously expand the scope of research subject to veto, veto becomes a mechanism for suppression rather than constraint on militarization. Historical examples demonstrate this risk: governments have used "national security" justifications to suppress inconvenient research, to silence critics, and to prevent investigation of governmental wrongdoing.

DESIGN PRINCIPLE: SCOPE LIMITATION

Legitimate veto governance must define veto authority narrowly. Veto should apply to specific downstream uses, not to research itself, publication, or inquiry:

1. **Downstream Use Limitation**: Veto authority should apply to transfer and deployment, not to research or publication. Researchers should retain freedom to conduct inquiries and to publish findings.
2. **Predefined Risk Categories**: Veto authority should apply only when research meets predefined risk criteria. Risk categories might include:
   - Integration into autonomous weapons lacking meaningful human control
   - Deployment in mass surveillance targeting political opponents or ethnic minorities
   - Use in violation of international humanitarian law or human rights law
   - Deployment in systems designed to enable forced displacement or ethnic cleansing
3. **High Confidence Thresholds**: Veto should be triggered by credible, documented risks, not speculative or distant possibilities.
4. **Exemptions for Legitimate Uses**: Veto should explicitly exempt legitimate civilian uses even if technologies are theoretically dual-use.

IMPLEMENTATION EXAMPLE

A veto body governing autonomous systems research might establish predefined thresholds:

> "Veto Triggered If: Research has direct application to autonomous weapons lacking meaningful human control AND the research was funded by a defense agency OR the researcher has been approached by defense organizations about deployment OR the publication explicitly discusses weapons applications. Veto Not Triggered For: Research with general applicability to autonomous systems absent specific evidence of weapons application OR research with clear civilian applications absent evidence of military diversion."

D.2 Collective Decision-Making and Procedural Legitimacy

Problem: Veto Concentrated in Individual Authority

Veto power concentrated in individual authority (a single researcher, administrator, or official) is dangerous. Individuals may make decisions based on personal bias, political motivation, or institutional self-interest rather than on principled assessment of risk.

Design Principle: Collective Decision-Making

Legitimate veto governance should make veto decisions through deliberative collective processes:

1. **Pluralistic Veto Bodies**: Veto decisions should be made by committees including researchers with technical expertise, ethicists, legal experts, and representatives from affected communities. Pluralistic composition ensures that decisions reflect multiple perspectives.
2. **Documented Procedures**: Veto bodies should operate according to documented procedures specifying how veto is initiated, what deliberation occurs, what information is reviewed, how decisions are made, and what records are maintained.
3. **Transparency and Publicity**: Veto decisions should be documented and made public (with appropriate protections for sensitive information).
4. **Deliberation Requirements**: Veto bodies should be required to articulate reasoning for veto decisions.

Implementation Example

A university veto committee charter might specify:

> "*Composition: 3 faculty researchers, 2 ethicists/policy experts, 2 community representatives, 1 legal expert. All members have equal voting authority. Decision Threshold: Veto requires 5 of 8 votes. Initiation: Veto can be initiated by any committee member, any faculty member, or external community representatives. Deliberation Timeline: 30 days from initiation. Documented Decision: Decisions recorded in writing including risk factors, harm prevention, and any conditions. Public Records: Decisions made public.*"

D.3 Formal Appeal and Revisability

Problem: Veto as Permanent Suppression

Veto decisions that are permanent and unrevisable can become suppression mechanisms. If veto can never be overturned, veto transitions from governance to censorship.

Design Principle: Appeal and Revisability

Legitimate veto governance should make veto decisions subject to appeal and revision:

1. **Appeal Procedures**: Veto decisions should be subject to appeal to a higher authority independent of the original veto body.
2. **Grounds for Appeal**: Researchers should be able to appeal on grounds including procedural error, factual disagreement, applicability of predefined thresholds, or proportionality.
3. **Conditional Veto**: Veto decisions can be conditioned on circumstances under which veto could be lifted.
4. **Sunset Provisions**: Veto decisions can include sunset provisions specifying that the veto will be revisited at a future date.
5. **Evidence-Based Revision**: If new evidence becomes available, veto decisions should be revised based on evidence.

IMPLEMENTATION EXAMPLE

A university might establish an appeals procedure:

> "*First Appeal: Researchers can appeal veto committee decisions to a university appeals board including external experts and community representatives. The appeals board can reverse committee decisions if it finds procedural error or factual error. Sunset: All veto decisions are revisited annually. If circumstances have changed or evidence suggests veto is ineffective, the committee can revise decisions.*"

D.4 LEGAL AND CAREER PROTECTIONS

PROBLEM: RETALIATION AGAINST THOSE EXERCISING VETO AUTHORITY

If researchers and administrators exercising veto authority face retaliation, veto authority remains theoretical rather than practical.

DESIGN PRINCIPLE: PROTECTED REFUSAL

Legitimate veto governance must include legal and career protections:

1. **Anti-Retaliation Provisions**: Federal and institutional law should explicitly prohibit retaliation against researchers or administrators for exercising veto authority.
2. **Tenure Protections**: Tenured faculty exercising veto authority should be protected by tenure.
3. **Whistleblower Protections**: Researchers or administrators reporting violations of veto authority should be protected as whistleblowers.
4. **Indemnification**: Institutions and individuals should be indemnified against legal liability for good-faith exercise of veto authority.
5. **Funding Protection**: Funding agreements should specify that institutional compliance with veto governance does not affect grant funding.
6. **International Protection**: For researchers in countries with weak legal protection, international professional organizations and funding agencies should provide legal support.

E ADDRESSING OBJECTIONS AND LIMITATIONS

Veto governance will face objections from multiple directions. This section addresses common objections and acknowledges genuine limitations that veto governance does not resolve.

E.1 POLITICAL CAPTURE AND WEAPONIZATION

OBJECTION

"Veto power will be captured by hostile actors and weaponized against legitimate research. Governments will use veto mechanisms to suppress research investigating governmental wrongdoing."

RESPONSE

This objection identifies a genuine risk. Governance mechanisms designed to constrain militarization could be misused to suppress inconvenient research. History provides examples: authoritarian governments have used "national security" justifications to suppress research on environmental contamination, governmental violence, and human rights abuses.

However, this risk must be weighed against the current absence of constraints on militarization. Currently, militarization proceeds largely unchecked. The choice is not between perfect governance

and the status quo; the choice is between imperfect governance with some risk of abuse and no governance at all with certainty of ongoing militarization.

Several design features constrain risks of capture and abuse:

- **Narrow Scope**: Veto authority limited to specific military uses
- **Predefined Triggers**: Veto triggered by specific, documented risks
- **Pluralistic Decision-Making**: Veto decisions made by committees including multiple perspectives
- **Independent Review**: Veto decisions subject to appeal
- **Transparency**: Veto decisions documented and made public
- **International Observers**: International observers in governance processes
- **Sunset Provisions**: Veto decisions time-limited and revisited

E.2 Innovation Stalling and Research Slowing

Objection

"Veto power will slow research and reduce innovation. Researchers will hesitate to pursue dual-use work. The resulting slowdown will harm economic competitiveness and human welfare."

Response

This objection reflects a particular value ranking: that speed of research and technological capability are paramount. This is defensible but should be made explicit and should be subject to debate.

Several points address this objection:

- **Scope Limitation**: Veto applies only to downstream military use, not to research itself. Researchers remain free to conduct inquiries and publish findings.
- **Modest Time Impact**: Even if veto review requires time, the time impact should be modest (30–60 days in research timelines spanning years).
- **Selective Application**: Veto applies only to research with clear dual-use risks, not to mainstream AI research.
- **Competitive Fairness**: If veto governance is applied universally, all researchers face equivalent constraints.
- **Defensible Value Tradeoff**: There is defensibility to prioritizing safety and justice over speed. The post-World War II international community decided that preventing nuclear proliferation justified accepting slower development of nuclear energy technology.
- **Harm Prevention Value**: Prevention of military AI integration creates real value: prevention of autonomous weapons reduces civilian casualties, prevention of surveillance systems targeting vulnerable populations protects autonomy and dignity.

E.3 Democratic Legitimacy of Veto Bodies

Objection

"Veto bodies will be undemocratic, wielding authority over research without legitimate democratic sanction. Who appointed these veto committees?"

Response

This objection identifies a genuine tension: governance authority should be legitimate and accountable. Several design features address democratic legitimacy concerns:

1. **Community Leadership**: Veto bodies should be led and controlled by affected communities.
2. **Elected or Appointed with Accountability**: Veto body members can be elected by research communities or appointed through transparent processes with documented accountability.
3. **Appeal and Revision**: Veto decisions should be subject to appeal and revision.
4. **Public Participation**: Veto bodies should hold public meetings, accept public input, and maintain public records.
5. **Embedded in Democratic Institutions**: Veto governance should be embedded in democratic institutions (universities, funding agencies) that have governance mechanisms and accountability structures.
6. **Temporal Limits**: Veto body members should serve fixed terms with term limits.

F MARGINALIZED COMMUNITY PERSPECTIVES

Veto governance succeeds as governance only if it is designed with and controlled by those most harmed by military AI deployment. This section details why affected community leadership is essential and how governance should be structured to center community voice and authority.

F.1 EXCLUSION FROM CURRENT AI GOVERNANCE

THE GEOGRAPHY OF AI HARM

AI governance discourse is concentrated in Global North institutions. Yet the most severe harms of military AI deployment are borne in other regions. Autonomous weapons are tested and deployed in conflicts in the Middle East, East Africa, and Central Asia. Mass surveillance systems are deployed in countries with authoritarian governments. The populations experiencing these harms are largely excluded from AI governance design.

The exclusion of affected communities from governance design is both unjust and epistemically limiting. Affected communities possess detailed, situated knowledge about:

- How military AI systems are deployed and what harms result
- What governance structures would be effective in their contexts
- What unintended consequences governance mechanisms might have
- How governance mechanisms intersect with other power dynamics affecting their communities

EPISTEMIC CONSEQUENCES OF EXCLUSION

Researchers and policymakers designing governance from elite institutions lack situated knowledge. Governance designed without affected community input risks being ineffective, culturally inappropriate, or inadvertently harmful.

F.2 SHIFTING GOVERNANCE LEADERSHIP: OPERATIONALIZING COMMUNITY CONTROL

PRINCIPLE: AFFECTED COMMUNITIES LEAD

Justice in governance requires that affected communities control governance design and implementation. This is not merely about inclusion or consultation; it is about control.

OPERATIONALIZATION REQUIREMENTS

Shifting governance leadership to affected communities requires:

1. **Funding for Governance Research**: Governance research on military AI should be funded and led by researchers in affected regions. Funding structures should be reoriented to support governance research in affected regions.
2. **Treating Governance as Scholarship**: Governance research should be recognized as legitimate academic work. Researchers conducting governance work should be eligible for professorial positions and prestigious publication venues.
3. **Binding Partnerships**: Governance partnerships should establish binding decision authority for affected communities, not merely advisory roles.
4. **Resource Control**: Affected communities should control governance resources—funding, staffing, decision-making authority.
5. **International Structures**: International governance structures should be led by representatives from affected regions with binding authority.

F.3 MILITARIZATION AND MARGINALIZATION: STRUCTURAL INTERCONNECTIONS

DEEP LINKAGES

Militarization and marginalization are not separate problems; they are deeply interconnected. Military violence disproportionately targets marginalized populations. AI systems amplify these dynamics by automating and scaling targeting logics that already discriminate.

TARGETING DYNAMICS

Military violence targets populations perceived as threats. Populations perceived as threats are often marginalized groups: ethnic minorities, indigenous peoples, religious minorities, political dissidents. Military AI systems scale and automate these targeting dynamics.

Addressing militarization without centering marginalization treats symptoms rather than structural causes. Effective governance must address both military applications of AI and the way military AI automation amplifies existing targeting and marginalization.

G CASE STUDIES OF MILITARIZATION PATHWAYS

This section details three historical cases where research transitioned into military applications, identifying veto points where governance could have constrained militarization.

G.1 FACIAL RECOGNITION AND MASS SURVEILLANCE

RESEARCH DEVELOPMENT

Facial recognition technology developed over decades through open academic research. Major algorithmic innovations came from researchers in academia and industry, with algorithms published at conferences and in journals.

MILITARY PATHWAYS

Facial recognition technology moved into military and surveillance contexts through multiple pathways:

1. **Surveillance Deployment**: Law enforcement and military organizations began deploying facial recognition systems.

2. **Border Control**: Border control and immigration enforcement began deploying facial recognition systems.
3. **Targeted Surveillance**: Authoritarian governments deployed facial recognition for targeted surveillance of ethnic minorities and political dissidents. Uyghur surveillance in Xinjiang relies heavily on facial recognition.
4. **Weapons Integration**: Facial recognition systems began to be integrated into weapons systems for target identification.

VETO POINTS AND GOVERNANCE FAILURES

Multiple veto points existed where governance could have constrained militarization:

1. **Publication Veto Point**: Individual papers could have been subject to veto review before publication or published with use restrictions.
2. **Licensing Veto Point**: When systems were licensed to commercial entities or governments, licensing agreements could have included restrictions on surveillance use.
3. **Partnership Veto Point**: University partnerships could have been subject to institutional review and could have been refused or conditioned.
4. **Funding Veto Point**: Government funding could have been conditional on governance requirements.

WHY GOVERNANCE FAILED

Governance failed at each veto point for structural reasons:

- No publication review of dual-use risks
- No technology transfer governance requiring use restriction assessment
- No partnership review mechanism
- No funding conditions on governance
- Competitive pressure to publish rapidly, license broadly, and accept military funding

G.2 DARPA FUNDING IN AUTONOMOUS SYSTEMS

RESEARCH DEVELOPMENT

DARPA funds significant portions of AI research, particularly in autonomous systems, robotics, and sensor technology.

MILITARY PATHWAYS

DARPA funding explicitly aims to develop military capabilities. Much DARPA research transitions directly into military systems:

1. **Autonomous Vehicles**: Research funded by DARPA has transitioned into autonomous military ground and aerial vehicles.
2. **Drone Control**: Research on drone control systems has transitioned into weapons systems.
3. **Robotic Manipulation**: Research on robotic systems has transitioned into military robotics.
4. **Autonomous Weapons**: Some DARPA funding explicitly aims to develop autonomous weapons capabilities.

VETO POINTS AND GOVERNANCE FAILURES

Veto points existed in DARPA funding decisions:

1. **Funding Decision Veto Point**: Universities could have refused DARPA funding or conditioned acceptance on governance requirements.
2. **Institutional Governance Veto Point**: Universities could have established governance committees reviewing DARPA-funded research before acceptance.
3. **Publication Veto Point**: Universities could have restricted publication of DARPA-funded research.
4. **Technology Transfer Veto Point**: Technology could have been restricted from transfer to weapons applications.

WHY GOVERNANCE FAILED

Governance failed for structural reasons:

- Revenue dependence on research funding
- No institutional governance structures reviewing defense funding
- Normalization of military research funding
- Competitive pressure to accept DARPA funding
- Lack of institutional authority to refuse military funding

G.3 UNIVERSITY–DEFENSE CONTRACTOR PARTNERSHIPS

PARTNERSHIP DEVELOPMENT

Universities frequently establish research partnerships with defense contractors (Lockheed Martin, Raytheon, Boeing, Northrop Grumman). These partnerships involve joint research, researcher visits, technology transfer agreements, and collaborative publication.

MILITARY PATHWAYS

University-defense contractor partnerships transition research directly into weapons systems:

1. **Technology Transfer**: Technologies developed in partnerships are transferred to defense contractors and integrated into weapons systems.
2. **Researcher Expertise**: University researchers collaborate with defense contractors on weapons development.
3. **Credibility and Legitimacy**: Defense contractors benefit from association with university researchers and academic legitimacy.
4. **Talent Pipeline**: Partnerships create pipelines enabling university researchers to transition into defense contractor positions.

VETO POINTS AND GOVERNANCE FAILURES

Veto points existed for partnership decisions:

1. **Partnership Approval Veto Point**: Universities could have established governance committees reviewing proposed partnerships.
2. **Technology Transfer Veto Point**: Technologies could have been restricted from transfer to weapons applications.
3. **Researcher Participation Veto Point**: Researchers could have been prohibited from participating in weapons development partnerships.

WHY GOVERNANCE FAILED

Governance failed for structural reasons:

- University partnerships provide funding and prestige
- Researcher autonomy norms prioritize freedom to choose collaborators
- No institutional governance structures reviewing partnerships
- Defense contractor participation in university governance
- Cold War legacy normalizing university-defense partnerships

H OPERATIONALIZATION MECHANISMS AND ENFORCEMENT STRUCTURES

This section provides detailed mechanisms for operationalizing veto governance, addressing common concerns that veto governance lacks concrete implementation pathways.

H.1 ARTIFACT-LEVEL TRIGGERS AND DECISION PROTOCOLS

PROBLEM: VAGUE GOVERNANCE WITHOUT DECISION POINTS

Abstract veto authority is difficult to exercise. Without concrete decision points and specific procedures, veto authority remains theoretical.

SOLUTION: PREDEFINED ARTIFACT-LEVEL TRIGGERS

Veto governance becomes practical when embedded in concrete decision points where choices must be made.

H.0.1 GRANT ACCEPTANCE TRIGGER

Decision Point: When institutions receive funding from defense agencies (DARPA, military research offices) or from other entities explicitly requesting dual-use research, a veto trigger activates.

Required Process:

1. **Written Risk Memo** (Responsibility: Technology Transfer Office or Grants Administrator)
   - Identify foreseeable military integration pathways
   - Assess likelihood and severity of militarization
   - Document evidence supporting risk assessment
   - Specify conditions that might mitigate risks
   - Timeline: 10 days from grant notification
2. **Decision Deadline** (Responsibility: Veto Body)
   - Veto body must convene and reach decision within 30 days
   - Failure to reach decision defaults to permission
   - Decision is recorded with reasoning

H.0.2 LICENSING TERMS TRIGGER

Decision Point: When research is licensed to downstream organizations, a veto trigger activates.

Required Process:

1. **Written Risk Memo** (Responsibility: Technology Transfer Office)
   - Identify proposed licensee and anticipated uses

- Assess military applications likelihood
- Specify what use conditions might prevent militarization
- Timeline: 10 days from licensing proposal

2. **License Agreement Terms** (Responsibility: Technology Transfer Office with Veto Body Input)
   - License must include use restrictions approved by veto body
   - License must require annual deployment reporting
   - License must grant audit rights
   - License must include indemnification clause for violations

H.0.3 Deployment Contracts Trigger

Decision Point: When research transitions to deployment in actual use contexts, a veto trigger activates.

Required Process:

1. **Deployment Impact Assessment** (Responsibility: Deploying Organization and Licensor)
   - Document deployment context
   - Assess alignment with use restrictions
   - Timeline: Assessment before deployment commences
2. **Veto Body Review**
   - Assess whether deployment violates licensing terms
   - Approve, require modifications, or refuse deployment
   - Timeline: 30 days from assessment submission

H.0.4 Repository Releases Trigger

Decision Point: When research artifacts are released to public repositories, a veto trigger activates.

Required Process:

1. **Use Condition Specification** (Responsibility: Authors/Repository Maintainers)
   - Artifacts released with machine-readable metadata
   - License text specifies permitted and prohibited uses
   - Repository implements access controls enforcing restrictions
2. **Transparency and Monitoring** (Responsibility: Repository Platform)
   - Repository maintains records of artifact access
   - Repository flags high-risk artifact deployments
   - Repository publishes aggregate compliance data

H.2 Veto Body Composition: Structured Representation and Fairness

Problem: Elite Capture and Marginalized Exclusion

Without structured composition requirements, veto bodies risk becoming dominated by privileged actors while excluding affected communities and marginalized voices.

Solution: Fixed Representation Quotas and Support Structures

H.0.5 Representation Requirements

A representative veto body of 8–10 members should have mandatory composition (see Table 4):

Critical Principle: All members have equal voting authority. Community representatives have full decision-making authority on par with faculty and experts.

Member Category	Minimum	Selection Method
Faculty Researchers (Technical)	2–3	Elected by research community
Ethicists / Policy Experts	1–2	Appointed by ethics/policy faculty
Affected Community Representatives	2–3	Appointed by affected communities
Legal Experts	1	Appointed by law faculty
External Experts (Optional)	1–2	Appointed by professional organizations

Table 4: Mandatory veto body composition

H.0.6 Community Representative Support Structures

Material support for community participation:

1. **Stipends**: Community representatives receive substantial stipends (e.g., $50–100/hour or annual stipends of $5,000–10,000).
2. **Access to Independent Legal Counsel**: Community representatives can access independent legal counsel paid by the institution.
3. **Travel and Logistics Support**: All costs covered; institutions cover all logistics for community members.
4. **Training and Capacity Building**: Institutions provide training helping community representatives understand technical issues and governance procedures.
5. **Accessibility Accommodations**: Flexible meeting times, remote participation options, interpreters for language/disability access.

H.0.7 Term Limits and Rotation

1. **Term Length**: Veto body members serve fixed terms (2–3 years).
2. **Term Limits**: Members can serve maximum 2 consecutive terms (4–6 years total).
3. **Staggered Rotation**: New members appointed on staggered schedules.
4. **Community Control of Appointments**: Affected community organizations control appointment of community representatives.

H.3 Enforcement Mechanisms and Compliance Structures

Problem: Veto Without Consequences

Veto authority without consequences is unenforceable. Enforcement requires that violations trigger automatic consequences.

Solution: Multi-Layered Enforcement Architecture

H.0.8 Contractual Enforcement

Mechanism: Research, licensing, and partnership agreements include explicit veto clauses creating legal obligations.

Example Veto Clause Language:

> "Licensee agrees that:
>
> - Research will not be deployed for autonomous weapons lacking meaningful human control; mass surveillance targeting political opponents; deployment in violation of international humanitarian law
> - Licensee will not sublicense to parties likely to use for prohibited purposes
> - Licensee will provide annual deployment context reports
> - Licensee will permit licensor to audit deployment contexts

- Any violation entitles licensor to [remedies: license termination, indemnification for damages, injunctive relief]

"

H.0.9 Machine-Readable Use Restrictions and Repository Enforcement

Mechanism: Artifacts include machine-readable use restriction metadata that repositories can automatically enforce.

Implementation:

1. **License Metadata Standards**: Artifacts include metadata fields specifying use restrictions using standard formats (SPDX expressions, YAML metadata).
2. **Repository Enforcement**:
   - Flag artifacts with use restrictions
   - Require users to certify compliance before downloading
   - Track which organizations access restricted artifacts
   - Block access by organizations with known violations
   - Delist artifacts if violations detected

H.0.10 Compliance Auditing and Reporting

Mechanism: Organizations deploying research under use restrictions are subject to mandatory reporting and audit.

Reporting Requirements:

1. **Annual Deployment Reports**:
   - Deployment location and organization
   - Intended use and actual use
   - Populations affected by deployment
   - Measures ensuring compliance
   - Modifications or updates to deployed systems
2. **Audit Notification**:
   - Facility visits to assess deployment context
   - Document review (deployment logs, system configurations)
   - Staff interviews about deployment practices
   - Written audit report
3. **Public Reporting**: Licensor publishes aggregate data on compliance, deployment sectors, violation rates, and enforcement actions.

H.0.11 Funding Agency Enforcement

Mechanism: Federal funding agencies make compliance with veto governance a condition of grant eligibility.

Funding Conditions:

1. **Institutional Governance Requirement**: Institutions must establish and maintain veto governance structures. Failure triggers loss of funding eligibility.
2. **Compliance with Veto Decisions**: Institutions must comply with veto decisions. Violations trigger investigation and potential defunding.
3. **Transparency Requirement**: Institutions must maintain public records of veto decisions.
4. **Loss of Funding for Non-Compliance**: Non-compliant institutions lose eligibility for future grants.

Example Federal Regulation:

> "*Award recipients must establish institutional governance procedures for dual-use research. Research involving [specified dual-use categories] is subject to review by an institutional committee meeting criteria specified in this regulation. Institutions must comply with committee recommendations regarding use restrictions, licensing conditions, or research refusal. Institutions violating these requirements become ineligible for future awards until compliance is restored.*"

H.0.12 Indemnification and Legal Liability

Mechanism: Violations of use restrictions trigger financial and legal consequences.

Indemnification Clause Example:

> "If Licensee deploys research in violation of use restrictions and that deployment results in harm, Licensee indemnifies Licensor against:
>
> - Damages paid to harmed parties
> - Legal costs and attorney fees
> - Institutional liability
> - Regulatory fines or sanctions
> - Reputational harm
>
> "

Legal Liability:

1. **Civil Liability**: Deploying organizations can be sued for damages resulting from violations.
2. **Regulatory Liability**: Organizations face regulatory investigation and sanctions if deployment violates standards.
3. **Criminal Liability**: Criminal prosecution may be possible in some jurisdictions if violations result in harm.

H.4 Integration with Existing Institutional Infrastructure

Key Insight

The machinery for veto governance already exists within institutional structures. Implementation requires embedding veto authority into existing decision-making processes.

H.0.13 Universities: Repurposing Existing Authority

Technology Transfer Offices already review licensing agreements and maintain licenses. **Mechanism**: TTO charters updated to require:

- Identification of dual-use research
- Inclusion of use restrictions in licenses
- Compliance monitoring and reporting
- Authority to refuse licenses lacking adequate restrictions

Faculty Governance Structures already review research and establish institutional policies. **Mechanism**: Governance charters updated to:

- Require establishment of veto committee for dual-use research
- Grant veto committee decision-making authority
- Protect veto committee decisions from administrative override

H.0.14 FUNDING AGENCIES: LEVERAGING GRANT CONDITIONS

Funding agencies already establish grant conditions and monitor institutional compliance. **Mechanism**: Grant conditions updated to:

- Require institutional veto governance as condition of funding
- Require compliance with veto governance decisions
- Require transparency through public records

H.0.15 REPOSITORIES: LEVERAGING PLATFORM INFRASTRUCTURE

Repositories already manage metadata and control access. **Mechanism**: Repository platforms extend capabilities to:

- Include use restrictions in standardized metadata
- Implement access controls based on use restrictions
- Automate enforcement of machine-readable restrictions

H.0.16 PROFESSIONAL ASSOCIATIONS: ESTABLISHING NORMS

Professional associations already establish codes of conduct and run conferences. **Mechanism**: Associations:

- Establish explicit commitments to veto governance in codes of conduct
- Require dual-use disclosure for conference papers
- Require use condition specification for published research

NOTE

Veto governance is not theoretical. The mechanisms detailed in this appendix show that veto governance can be operationalized through existing institutional structures, with concrete decision points, structured procedures, material enforcement mechanisms, and appropriate safeguards against abuse.

Implementation requires political will, funding, and culture change. But the mechanisms are available. Universities can embed veto clauses in licenses. Funding agencies can condition grants on governance. Repositories can implement use restrictions. Professional communities can establish codes of conduct.

Affected communities must lead this transition. Governance designed without affected community leadership risks failing to address actual harms and risks replicating power imbalances that produced militarization.

The operationalization mechanisms detailed in this appendix provide the practical pathway for that transition: from abstract commitment to concrete practice, from ethical exhortation to enforceable governance, from elite-designed governance to community-led accountability.