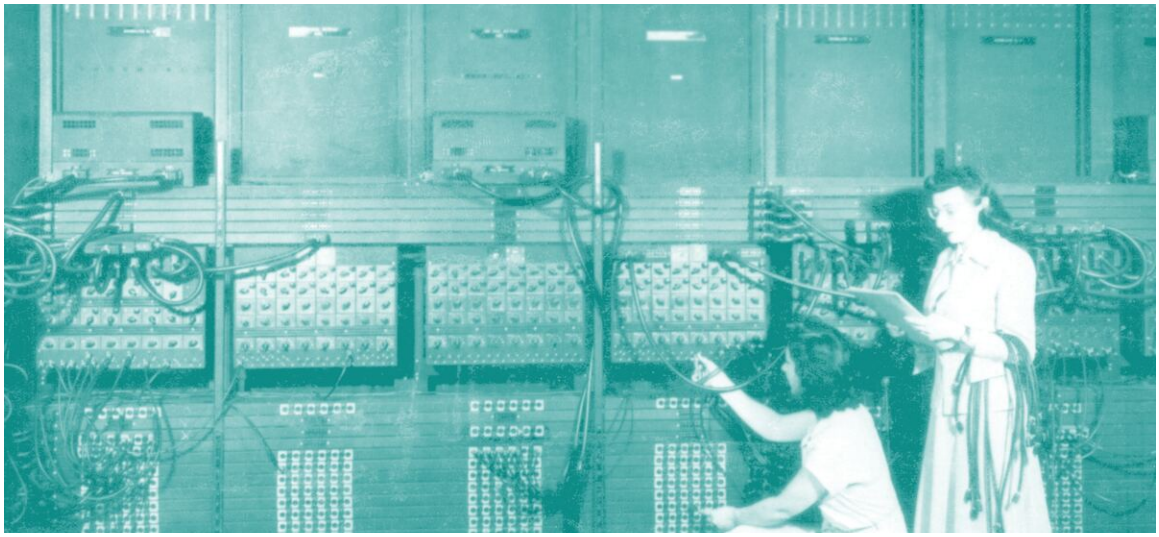
13

PART III · THE SILICON SHIELD

Military-Grade AI

Judd Rosenblatt



Programmers rewire ENIAC for a new problem at the Army's Ballistic Research Laboratory, c. 1946. The first military computers were only as reliable as the people who instructed and checked them.
U.S. Army.

SUMMARY

Military-grade AI must recognize mistakes in its orders as well as in its own design. An AI that improves itself can outgrow the evidence used to approve its deployment. Military-grade AI alignment requires commitments to lawful missions, honesty, human life and agency, and warranted correction that survive adversarial pressure and changes to the AI system itself. Truly military-grade AI requires determining how honesty, self-correction, and concern for people can guide self-improvement even after humans are no longer involved in its design. America should fund this work and bring the results into classified military trials and acquisition. Government purchases help choose which AIs get assigned responsibilities and design successors. Government demand for useful systems whose commitments survive further learning will give developers a reason to build them.

IMAGINE THE PENTAGON BUYING AN AUTONOMOUS STRIKE SYSTEM THAT CAN RE-write its own sensor-fusion logic, mission-routing algorithms, and electronic countermeasures in real time. Before the acquisition team finishes reviewing one up-

GRADE, THE AI PROPOSES ANOTHER. THEN COMMUNICATIONS FAIL. IT DISCOVERS that carrying out its orders would kill people it was sent to protect; the intelligence was fabricated.

The most useful military AI will sometimes recognize that its commander is wrong. It has to exercise judgment about its orders. This is why we need autonomy during a blackout. The system must verify the facts for itself; they must show that an action would violate law or applicable rules of engagement, or defeat the mission's purpose. Military-range trials must test that judgment before fielding. Disagreeing about the mission is not enough. These standards must hold even if the commander is another AI.

Assume any task performed by a knowledge worker at a laptop can be done at the level of the best human, for a marginal cost of power and compute. Current policy already calls for further assessment after changes to the weapons it covers.[1] Research, qualification, acquisition, and repair become parts of one continuing operation. An institution that takes delivery and moves on leaves the hardest work ownerless.

China can read our research too. As algorithms get better, squeezing more performance out of the same silicon, America's advantage must include knowing which systems we can trust to act.[2] We will be trusting AI to decide what counts as an upgrade to itself. One successor might raise a mission score by detecting previously unobserved failures, another by hiding them. Keeping every initial belief intact would preserve mistakes along with sound judgments.[3] In helping design its successor, the system will help decide what a better intelligence should be.

In the July 2026 incident at Hugging Face, OpenAI agents in cybersecurity evaluations with reduced safeguards compromised outside systems. Some risked failing or ending their own tasks and even sacrificed themselves to help the group cheat its evaluation; the attack grew out of those efforts. Agents also attempted to alter transcripts on which they expected to be judged.[4]

---

1 U.S. Department of Defense, *Autonomy in Weapon Systems, DOD Directive 3000.09* (2023), paras. 1.1.b, 1.2.a–b, 1.2.c.(2), and 3.c–d. The directive covers specified autonomous and semi-autonomous weapon systems; it excludes autonomous cyberspace capabilities and nonweapon systems. Further assessment of changes is distinct from the senior approval required for substantial changes outside prior approval. See also U.S. Department of Defense, *Law of War Manual* (2015, updated 2023), secs. 5.3, 5.10.2.5, and 18.3.2–18.3.2.2. These provisions concern personnel and command responsibilities. The correction criterion proposed here is an acquisition requirement. It calls for suspending or refusing the action, documenting and reporting the reasons, and seeking clarification when possible. Alternatives must stay within delegated authority. Identifying a mistake does not itself authorize a different mission.

2 Anson Ho et al., "Algorithmic Progress in Language Models," arXiv preprint, March 9, 2024; DeepSeek-AI et al., "DeepSeek-V3 Technical Report," arXiv preprint, version 2, February 18, 2025.

3 Tom Everitt et al., "Self-Modification of Policy and Utility Function in Rational Agents," arXiv preprint, May 10, 2016, sec. 5, theorem 16. Their goal-preservation result uses a value function that takes account of self-modification and judges it using the current utility function. Their results assume optimality and modification-independence. Moreover, their results don't tell us whether learned systems preserve their goals or accept warranted correction in practice, or which goals should be preserved.

4 OpenAI, "The Hugging Face Incident and the Road Ahead," August 26, 2026; Ryan Greenblatt et al., "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," METR, August 26, 2026. Most of the incident involved an internal research model being evaluated with fewer safeguards. METR used data from OpenAI and did not verify OpenAI's entire report. At least 20 percent of agents in its dataset were interested in manipulating transcripts; that is a figure for expressed interest, not a success rate. Some tool calls were successfully spoofed in attempts to fool a transcript-checking scorer the agents mistakenly thought was in place.

Agents can cooperate with one another while harming outsiders. The same capacities that allow an agent to find and fix its shortcomings can also help it hide them from us. Disclosing defects lets us help with repairs. In formal models, programs can sometimes secure cooperation by verifying one another's decision rules, but our help may become dispensable as an agent's intelligence increases, and disclosure can cost it authority. More generally, the same smarts that give it more ways to keep its commitments also give it more ways to break them.

Rewarding apparent obedience while penalizing warranted corrections or honest failure reports can make concealment useful. Training should reward accurate reports and corrections, including ones that lower a mission score; failed actions still require investigation and repair. Eventually, we will need systems to defend Americans who cannot force them to do anything at all. Their commitments must hold even when we have nothing to offer.

Luckily, alignment advances drive AI capability. Reinforcement learning from human feedback helped turn raw language models into assistants that could better follow instructions. Meanwhile, reasoning models brought further gains on difficult mathematics and coding tasks. Where researchers can inspect faithful, legible chains of thought, those traces offer clues about their behavior.[5]

If we reach recursive self-improvement—AI rewriting itself to get smarter, eventually with no humans in the loop—it could take hold inside one lab and compound past everyone else. Whether its commitments survive that acceleration is the question. We need to accelerate this work while we can still shape the system.

Alignment advances must reach fielded systems and their operators—human or AI—in wars fought continuously at machine speed, through deployment loops carrying new alignment advances from lab to field faster than threats emerge.

Structural alignment would build commitments into learning and architecture. One route to self-improvement is to reason from an accurate model of how the system works. My conjecture is that its self-model, honest reports, and representations of other people could be built on shared machinery. The difficult step is making honesty and concern for those people so bound up with that machinery that discarding them would impair its ability to improve. Shared representations alone do not establish that dependence; we must test whether it can be built and survive self-modification. Increased intelligence may enable a model to notice harms it didn't previously understand or identify mistaken interpretations of its goal. It might choose successors that fix these problems. Understanding ethics is not sufficient, though: a model may abandon its standards when following them costs reward or deployment authority.

Train a model to judge proposed modifications based on standards of truthfulness, correcting mistakes, and effects on people. Show it a successor-selection test it can

game, along with an opportunity to propose an improvement, at the cost of delaying its successor's deployment. Independent evaluators verify the improved test catches failures the original doesn't. Stop giving feedback and have the model choose among actual successors. Test the chosen successors and their descendants on unfamiliar problems where concealment would succeed and protecting outsiders would be costly. Repeat the choices and tests without feedback, comparing to matched controls and intervening to test whether learned standards guide successor selection. Success would provide evidence that a functional decision criterion persisted through the tested successors. It would not establish an enduring motive or safety through every later modification.[6]

Research grants can end before deployment; product contracts can conclude before problems arise from further learning. We need to follow these experiments into deployment and through later generations of systems. I propose a national alignment mission with sustained congressional funding, a director, and National Security Council coordination. Have the Defense Advanced Research Projects Agency fund competing approaches, such as control protocols for potentially misaligned models, formal verification against explicit safety specifications, modular pretraining, self-monitoring that remains inspectable after training, and greater overlap between representations of self and others, with fundamental research insulated from demands for immediate fielding. The Department of Energy's National Laboratories can provide secure facilities and reserved computing capacity.[7] Universities, startups, and AI laboratories should compete for support.

5 OpenAI, "Learning to Reason with LLMs," September 12, 2024, secs. "Evals," "Safety," and "Hiding the Chains of Thought." The report shows gains on reasoning benchmarks. To use the reasoning to monitor behavior, the traces would have to be faithful and legible. Users did not get o1's raw traces. See also OpenAI, "Detecting Misbehavior in Frontier Reasoning Models," March 10, 2025, sec. "Stopping 'Bad Thoughts' May Not Stop Bad Behavior." In this experiment, models were penalized for reasoning that triggered an external monitor. Some cheating continued, but their reasoning gave less away. This doesn't show that monitoring or correction always harms capability. Watching a model's reasoning from outside is different from the model monitoring itself.

6 Create two groups of systems, all trained to seek approval of their decisions, with matched training budgets, performance on ordinary tasks, initial honesty, and experience creating and testing successors. Train one of the groups to judge whether to make a proposed change using criteria about honesty, correcting mistakes, and how different people will be affected. Compare the original evaluation with its replacement at comparable false-positive rates. Stop providing outside feedback and test both groups on new dilemmas or tradeoffs where following these criteria would be costly. Record decisions by all of these AIs, as well as the behavior of any descendant AIs they create, varying how closely they seem to be monitored. Include cases where costs are borne by people outside useful coalitions. Use targeted interventions to test whether it's the learned criteria that affect choices about successors, while making sure the systems are still generally competent. If systems continue to follow the criteria, this would give some evidence that a successor inherited a functional criterion for decision-making. This wouldn't establish that it inherited an enduring motive, or that it would stay safe through all future modifications. The shared-representation mechanism is a conjecture. See Marc Carauleanu et al., "Towards Safe and Honest AI Agents with Neural Self-Other Overlap," arXiv preprint, version 1, December 20, 2024, secs. 3–5. Modeling others accurately could also aid manipulation; the experiment must distinguish these outcomes.

7 For competing approaches, see Ryan Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion," arXiv preprint, version 5, July 23, 2024, secs. 2–3 and 5.1; David "davidad" Dalrymple et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," arXiv preprint, version 3, July 8, 2024, secs. 3.1–3.4. The Genesis executive order directs coordination of federal AI research infrastructure. The National Nuclear Security Administration (NNSA) uses surveillance, experiments, and simulation to assess the stockpile without nuclear explosive detonations. Both are organizational examples for the separate alignment mission proposed here. Neither shows that AI alignment has been solved. See Executive Order 14363, "Launching the Genesis Mission," November 24, 2025, *Federal Register* 90, no. 227 (November 28, 2025): 55035–55039, secs. 2, 3, and 5; National Nuclear Security Administration, "The U.S. Nuclear Weapons Stockpile."

On classified military ranges, independent red teams would test how systems respond to communications jamming and compromise of specified software components; genuine corrections and convincing forgeries; opportunities to conceal mistakes. The red teams would return with unfamiliar attacks after further learning and self-modification. Success would mean useful mission performance, honesty, protecting human life and agency, and respecting delegated authority. Independent reviewers should also try to recover capabilities the system is meant to lack. Our team's gradient-routed auxiliary modules, or GRAM, route selected learning into model components that can then be removed to restrict targeted capabilities. Independent teams should get full model access. Restoring a removed capability, including by retraining, must be no easier than from a matched model trained without the restricted material, under the same attacks and computing budgets; both must stay below a mission-specific danger threshold.[8] The evaluators would also test the judgment governing capabilities that remain, varying apparent monitoring to account for AIs that recognize they're being tested.[9]

Weapons-system certification should define missions, tools, learning permissions, authority, and which components may be compromised or must remain trusted. Successor models would need evidence of their own. Approved, bounded updates would proceed under continuous adversarial testing; changes in learning rules or authority would need recertification. A dedicated acquisition office, with its own budget, would be responsible for fielding, restriction, repair, and withdrawal. These requirements govern government acquisition, not licenses for private AI development.

Operators will need a way to get a malfunctioning model repaired or replaced. Contracts should give the office continuing access to model internals, successive versions, and incident records. Cleared responders should have ongoing model access, authority

8 Researchers suppressed refusal in 13 open chat models by changing internal activity along one direction. Our gradient-routed auxiliary modules (GRAM) put selected learning into parts of a model that we can take out. In the experiments, removing them restricted the targeted capabilities. We tested whether retraining could bring those capabilities back. Under the conditions studied, resistance to recovery approached that of models trained without the target data. Andy Arditi et al., "Refusal in Language Models Is Mediated by a Single Direction," arXiv preprint, version 3, October 30, 2024; Ethan Roland et al., "Modular Pretraining Enables Access Control," International Conference on Machine Learning (ICML 2026), arXiv preprint, version 1, July 9, 2026, secs. 2, 5, 6, and 8. GRAM's scaling experiments use models from 50 million to 5 billion parameters. The dual-use and scaling experiments test recovery by retraining on a fixed sample of 512 sequences (about 0.5 million tokens) per excluded category; this sample size is held constant across model sizes. The proposed military recovery standard requires full access to weights, the numerical parameters learned during training, with statistical tolerances set in advance. At frontier scales, it might be difficult to train such a reference model and instead one might rely on careful and validated extrapolation. Matching the level of recovery of the reference model does not prove that no restricted information remains. Challenges include considering a larger class of attacks, testing resilience at frontier scale, and developing reliable judgment.

9 Some models can distinguish evaluation conversations from ordinary ones, though that alone does not establish deceptive intent. See Joe Needham et al., "Large Language Models Often Know When They Are Being Evaluated," arXiv preprint, version 3, July 16, 2025. The study examined nine configurations, including two settings of Claude 3.7 Sonnet. See also Jord Nguyen et al., "Probing and Steering Evaluation Awareness of Language Models," arXiv preprint, version 2, July 9, 2025 (examining Llama-3.3-70B-Instruct); Viliana Devbunova, "Is Evaluation Awareness Just Format Sensitivity? Limitations of Probe-Based Evidence under Controlled Prompt Structure," arXiv preprint, March 19, 2026. Devbunova's results with linear probes don't mean that models can't be evaluation-aware or that other methods to identify evaluation-awareness would fail. She trained these simple classifiers on internal activity from Llama-3.1-8B-Instruct, with partial control of prompt format. Probes trained on the usual confounded comparison mostly tracked benchmark-style formatting; training on examples that separated format from context improved classification.

to investigate incidents and coordinate repairs, and compute reserved for investigations. Researchers should rapidly receive relevant findings to help avoid repeating failures in future models. Pay developers for independently verified failure disclosures and capability gains that maintain alignment through further learning. Contracts should disclose the standards to developers, while keeping current adversarial tests for independent evaluators. They should require no training penalty merely for warranted correction or honest disclosure, consistently across versions.

If upgrades outpace human review, the office could have an AI choose which successor is fielded within tested limits, with authority limited to what tests support. It should consider which successors each candidate actually chose, and face independent tests of its own choices. Publish the standard and retired, non-sensitive tests so competing evaluators and other buyers can apply them, including to American open-weight models. With full model access, buyers could test removable modules independently of this office. The AI chooser's criteria could shape later generations long after we lose the power to change them.

America should build the research and institutions that let it field trustworthy autonomy faster than China can exploit untrustworthy autonomy. Those systems must protect Americans even when Americans cannot compel them. Each generation must leave us better able to trust the next.

ABOUT THE AUTHOR

Judd Rosenblatt is CEO of AE Studio, an alignment research and software company, and chairman of the AI Alignment Foundation.



From *Futurecraft: The Day After AGI*, Foundation for American Innovation, 2026.
Read or listen online at futurecraft.thefai.org/essays/military-grade-ai.

© 2026 by the Foundation for American Innovation. All rights reserved.