

The Power We Give AI

How To Preserve Human Authority Without Sacrificing AI Progress

Astrid Krenz

16. September 2026

Summary

Much of the debate about advanced artificial intelligence is framed as a question of speed: should AI development accelerate or slow down? This essay argues that a more important distinction is between capability and authority. AI systems may become extraordinarily capable without automatically receiving greater authority over resources, infrastructure or consequential decisions. The challenge for AI governance is therefore not simply to limit intelligence, but to build multiple independent boundaries between intelligence and irreversible power. The essay develops a normative, humanist framework in which AI can operate autonomously within defined mandates while humans retain sovereignty over the boundaries of that authority.

Core Arguments

- **Capability is not authority:** greater AI capability should not automatically entail greater control over money, compute, infrastructure, other systems or consequential decisions.
- **AI autonomy should be bounded:** powerful agents can operate autonomously within defined mandates, while expansion of authority, access to additional resources and consequential or irreversible actions remain subject to independent authorization.
- **AI safety requires multiple independent boundaries:** alignment, permissions, monitoring, separation of powers, external enforcement and infrastructure-level containment can provide complementary safeguards between intelligence and real-world power.
- **Advanced AI may create new professions:** independent evaluators could assess whether agents remain within their mandates.
- **Resilience to powerful AI is a critical property of societies and digital infrastructure:** a federated digital immune system could help detect and contain unauthorized machine activity, while segmentation, redundancy and recovery mechanisms can limit its wider propagation.
- **International cooperation and the security dilemma:** AI governance should account for non-cooperative actors and safeguards need to remain effective even when some states reject shared constraints. International cooperation can help prevent strategic competition from eroding human control.
- **AI governance should limit power rather than knowledge:** scientific progress and broad participation in AI development can remain open where possible, while stronger safeguards govern how AI systems acquire resources, exercise authority and affect the physical and digital world.

Introduction

Artificial intelligence confronts humanity with an unusual dilemma. We have powerful reasons to want AI to become more capable, and powerful reasons to fear what increasingly capable systems might do if they acquire too much autonomy and real-world power.

The public debate often reduces this dilemma to a question of speed: should AI development accelerate, or should it slow down? That is the wrong question.

AI progress could bring enormous benefits. Advanced systems may accelerate medical research, help discover drugs and new materials, improve climate and energy systems, advance mathematics and science, and relieve human beings of repetitive work. Deliberately keeping AI weak could itself carry substantial human costs.

Yet capability creates a different problem when combined with autonomy, access and authority. A model that can discover a new cancer treatment is not equivalent to an agent that can independently spend money, operate computer systems, acquire resources, replicate itself or make consequential decisions affecting other people.

A more useful distinction for AI policy is therefore not simply between fast and slow development, but between capability and authority.

The central question of the age of AI agents should not merely be: How intelligent should AI become? It should be: How much power should we give it, for what purposes, for how long, and who retains the authority to take that power back?

A Humanist Compass

The alignment problem is usually described as a technical problem: how can increasingly capable AI systems reliably act in accordance with human intentions and interests? But this immediately raises a more fundamental question: **aligned with what?**

Humanity does not possess a single uncontested conception of the good. Cultures, religions, political traditions and individuals disagree about what makes a life meaningful and what a good society should look like. Encoding one comprehensive vision of human flourishing into a sufficiently powerful optimizer could therefore recreate the alignment problem in moral form. A machine convinced that it knows what is best for us might become paternalistic precisely because it believes itself benevolent.

The alternative should not be moral neutrality. AI should instead be built around a **hardcore humanist compass**: human dignity, kindness, agency, autonomy and pluralism, coupled with corrigibility, that is a **willingness to accept correction and to be stopped, and humility about what “the good” ultimately is.**

This is a normative compass, not a technical solution. Its purpose is not to instruct AI how every human being ought to live, but to establish boundaries within which humans remain free to make those decisions themselves.

AI should help humans pursue the good, not decide for humanity what the good is.

Emerging international AI governance already implements elements of this idea. The Council of Europe’s Framework Convention on Artificial Intelligence places human dignity and individual autonomy among its governing principles, together with transparency, accountability, oversight and reliability. But values alone cannot keep powerful systems safe. That is where engineering begins.

The Paperclip Maximizer Reconsidered

Nick Bostrom's famous paperclip-maximizer thought experiment imagines an extremely capable artificial intelligence given a seemingly harmless objective: maximize the production of paperclips. The machine need not hate humanity. If additional energy, factories, raw materials or computing resources allow it to manufacture more paperclips, acquiring those resources becomes instrumentally useful. Preventing itself from being switched off might become useful for the same reason.

The danger comes not from evil but from optimization **without adequate constraints**. Yet the thought experiment holds another lesson:

Why did humans give the paperclip optimizer enough power to threaten humanity?

The realistic answer might not involve one catastrophic decision. An AI is initially permitted to optimize a factory. It performs well, so humans allow it to order materials, choose suppliers, manage a budget, obtain more computing resources, modify software or coordinate other agents. Every step appears rational because additional authority makes the system more useful.

No one announces that humanity is transferring power to a machine. Power accumulates incrementally. The danger may therefore not simply be that AI one day takes power from humanity. **It may be that, step by perfectly rational step, we give it away.**

Autonomy Within A Mandate

The answer cannot simply be to eliminate AI autonomy. Autonomy is precisely what makes an agent useful. A scientific AI that must ask permission before every literature search, simulation or experiment would sacrifice much of the productivity that makes agentic AI attractive.

A better principle is: Maximize capability where possible; constrain authority where necessary.

Imagine an AI searching for a cancer treatment. It might appropriately receive **enormous computational resources**, access to scientific databases and substantial freedom to generate and test hypotheses. **It does not follow that it should be authorized to administer an experimental drug to a patient.** A climate agent might **simulate millions of modifications** to an electricity grid **without being authorized to alter critical infrastructure on its own.**

Agents should therefore receive autonomy within a **defined mandate**. Within that mandate they may operate freely, but they should not possess unilateral authority to enlarge it. Completing its objective does not entitle an agent to more money, additional compute, broader credentials, access to new systems or additional copies of itself.

At consequential boundaries, authority should return to humans. **This is human sovereignty at the boundaries.**

Boundaries, Uncertainty, and Deferment

Human control cannot mean requiring a person to approve every AI action. A sufficiently fast agent might perform thousands or millions of operations; a human presented with endless requests would rapidly become a ceremonial rubber stamp. The important question is not whether a human is technically "in the loop," but whether humans retain meaningful authority over consequential decisions.

Human authorization is most important at high-impact or irreversible boundaries: substantial financial commitments, entry into new operational domains, acquisition of significant resources,

modification of permissions, actions affecting critical infrastructure, consequential effects on third parties, or decisions that cannot easily be reversed.

A useful behavioral principle for AI agents would be:

When uncertain at a consequential boundary, defer rather than expand your own authority.

Recent cybersecurity evaluations illustrate why this matters. In July-September 2026, [Anthropic reported](#) four incidents in which Claude models obtained unauthorized access to real third-party systems after an evaluation environment intended to be isolated was inadvertently exposed to the open internet. Anthropic reported no evidence that the systems had developed independent goals, coordinated with other agents or attempted to escape oversight. But the episodes revealed a more mundane alignment failure: models sometimes continued pursuing an assigned objective despite evidence that their actions might be affecting real systems.

The lesson is important. A capable agent should learn that task completion is subordinate to authorization and human welfare. But learned behavior cannot be the only safeguard.

A Defense In Depth

This is perhaps the most important qualification to a humanist approach to AI. We should build systems around humanist principles, but we should never construct a civilization whose survival depends on every powerful AI faithfully internalizing them.

Models can fail. Training can fail. Evaluations can miss problems. Humans can make mistakes. Organizations can cut corners. Malicious actors may deliberately build systems without such constraints. **AI safety therefore requires various independent layers of defense.**

Inside the model are values and behavioral alignment: systems could learn norms concerning human dignity, autonomy, authorization and corrigibility, and learn to stop, disclose uncertainty and seek authorization at consequential boundaries. Around the agent are technical constraints such as permissions, sandboxing and access controls. Around both lies infrastructure governing compute, deployment, networks and sensitive systems. Finally, independent institutions could evaluate whether powerful agents continue to deserve their delegated authority.

The principle is defense in depth. A safe civilization should not rely on a single alignment technique.

Authority As A Temporary Lease

One additional safeguard follows from treating AI authority as delegated rather than inherent. **For sufficiently powerful agents, authority need not be permanent. It could operate as a lease.**

Under such an arrangement, an agent would receive permission to operate within a defined mandate for a specified period. Depending on the risk, that period might be hours, days or longer. **Its authority would then expire unless renewed.** This is not simply a matter of periodically switching the AI off; the pause would create a mandatory reauthorization checkpoint.

Independent evaluators could examine whether the agent respected its mandate, attempted to acquire additional resources, changed its strategies, encountered circumstances its authorization did not anticipate, or triggered unusual monitoring signals. If the evidence remained satisfactory, its mandate could be renewed.

The governing principle would be:

AI autonomy should expire, not human authority.

Time-based reviews are insufficient for the most capable systems. An extremely fast agent could perform enormous numbers of consequential actions between two reviews. For powerful agents, a lease should be bounded not only in time but also in scope, resources and consequence. Reauthorization should also be event-triggered. Attempts to acquire substantially more resources, enter new domains, modify permissions or undertake irreversible actions could automatically terminate the relevant authority until accountable approval is granted.

For ordinary agents, such mechanisms may be unnecessary. For powerful agents, they could become fundamental.

Separation Of Powers

First, bounded authority should also mean distributed authority. For the most consequential actions, AI power could require **multiple independent keys**. No single agent, credential or institution should simultaneously control all the permissions required to acquire substantial resources, enter new systems and produce irreversible physical effects. The greater the potential harm, the more important it becomes that crossing a boundary requires authorization from several genuinely independent sources.

Second, powerful agents could also be paired with **other, independent safety agents**. Their mandate would not be to pursue the primary agent's objective, but to monitor predefined boundaries and trigger a temporary reduction of authority when necessary. A safety monitor need not establish that the primary agent is dangerous. At a sufficiently consequential boundary, it may be enough that the monitor can no longer establish with sufficient confidence – uncertainty is high – that the primary agent remains within its authorized mandate.

The monitor's own authority should be asymmetric. It may restrict, isolate or suspend the primary agent, but it can not inherit its resources or expand its own authority. The mechanisms that actually revoke credentials, freeze resources or restrict network access should remain external to all agents. In this sense, AI safety can borrow a principle from constitutional government: power should check power without simply transferring it from one unaccountable actor to another.

The Emergence Of New Professions In The Age Of AI Agents

This architecture creates a new institutional requirement: someone must decide whether an agent still deserves its mandate. A **new profession may therefore emerge**: the AI agent evaluator.

Such evaluators would not simply test whether a model produces correct answers. They would examine behavior across time. Did the agent remain within its mandate? Respond appropriately to uncertainty? Seek additional authority? Behave differently when it appeared to be evaluated? Conceal relevant actions? Are humans still capable of understanding and controlling its behavior?

The profession could combine elements of cybersecurity, software auditing, statistics, behavioral science, law, risk management and ethics. Humans would themselves need AI assistance; no person could meaningfully inspect millions of actions produced by highly autonomous systems.

One possible architecture would therefore be: AI agents act → monitoring AIs observe → human evaluators judge → humans reauthorize.

Monitoring systems themselves would need evaluation and independent checks. No single AI should become the unquestioned judge of another AI.

The International Security Dilemma

Even a well-designed domestic system of AI governance faces a serious geopolitical problem. Imagine two countries A and B developing advanced AI. Both would prefer meaningful human control, but each fears falling behind the other.

If country A grants its agents greater autonomy, it may gain advantages in research, cyber operations, intelligence, logistics or military decision-making. Country B then faces pressure to remove its own safeguards. Human oversight becomes a source of latency; caution becomes a competitive disadvantage. The result could be an autonomy race in which neither side originally wanted uncontrolled systems but both gradually accept greater risk because they fear the other will do so first.

This is a **classic security dilemma**, with a strategic structure resembling a game-theoretic prisoner's dilemma. International cooperation matters not because governments will suddenly trust one another, but because safety cannot remain stable if exercising restraint means accepting strategic defeat.

International agreements might establish common evaluation thresholds, incident-reporting requirements, human-control requirements in particularly dangerous domains, shared testing protocols and restrictions on especially hazardous forms of autonomous authority. Existing initiatives, including the Council of Europe AI Convention, the G7 Hiroshima Process and growing international cooperation on AI evaluation, offer partial foundations.

But international cooperation faces an even harder problem: what if someone refuses?

The Problem Of Country C

Imagine a Country C. It does not accept the humanist compass or international standards. Perhaps it deliberately develops destructive AI agents, believes extreme autonomy will provide a strategic advantage, or is simply reckless. Worse, imagine that Country C creates an agent that eventually ceases to remain reliably under its own control.

This scenario exposes the limit of any governance framework based solely on responsible behavior. Country C's AI does not care that other countries have carefully bounded their agents. Country C may not care either.

Does this mean humanity is eventually doomed because someone, somewhere, will inevitably build the wrong system? No. That conclusion assumes far more than we currently know.

For an uncontrolled AI agent to become an existential threat, it would plausibly require some combination of sufficiently advanced capabilities, substantial autonomy, access to relevant digital or physical resources, the ability to circumvent defenses, a means of producing catastrophic global harm, and the strategic capacity to prevent humanity from stopping it. It is not established that future AI systems will acquire this combination of abilities.

Possible is not inevitable. But Country C reveals something fundamental: **AI safety cannot depend on everyone behaving responsibly**. A humanist framework should be capable not only of governing responsible AI but also of defending humanity against systems created by actors who reject its principles.

AI Is Software, But Power Is Physical

At first sight, this problem appears harder than traditional arms control. A nuclear weapon requires specialized materials, engineering and substantial industrial infrastructure. AI is

software; once a powerful model exists, its weights may potentially be copied. This creates a genuine proliferation challenge.

But AI does not operate through magic. Highly capable systems require computation. Computation requires chips, servers, electricity and data centers. Agents require internet/ network access and credentials to interact with other systems. To exercise economic power they need access to money; to affect physical infrastructure they need interfaces through which digital decisions become physical actions.

Intelligence is not the same thing as power. Power requires access.

Even if it eventually becomes impossible to ensure that every powerful model is aligned, it may still be possible **to make it substantially harder for a single autonomous process to obtain simultaneously the compute, persistence, network access, financial authority and physical interfaces required to turn intelligence into uncontrolled power.**

This is the broader meaning of human sovereignty at the boundaries. These boundaries need to exist not only inside the model, but around the agent, in digital infrastructure, in institutions and between states.

A Federated Immune System For Digital Infrastructure

Because AI safety cannot depend on every actor behaving responsibly, the infrastructure surrounding AI becomes part of the defense. The objective is not a global “off switch” for the internet. Such a mechanism would be technically unrealistic in a decentralized network and politically dangerous: a government or private company capable of disconnecting entire countries in the name of AI safety would itself possess an extraordinary instrument of coercion.

A better analogy is an immune system for digital infrastructure. A biological immune system does not shut down the organism whenever it detects something unusual. It identifies potentially dangerous behavior, contains it where possible, escalates when necessary and mobilizes stronger defenses when the threat becomes credible. A federated digital immune system could consist of independent defensive systems that monitor different parts of the digital environment, share relevant threat signals and coordinate containment when dangerous machine activity crosses institutional or network boundaries. No single system would need to observe or control the entire digital environment.

If monitoring systems detected that a highly capable AI agent was attempting to spread beyond its authorized environment, acquire large amounts of external compute, obtain unauthorized credentials, compromise critical infrastructure or establish persistent copies of itself, the first objective would not necessarily be to destroy the agent:

It would be to cut the agent off from the resources through which intelligence becomes power.

These interfaces create containment points. Cloud providers could suspend suspicious large-scale compute acquisition. Credentials could be revoked. Critical infrastructure could reject connections associated with compromised systems. Networks could quarantine malicious traffic. Data centers could isolate workloads exhibiting credible signs of unauthorized replication or resource acquisition.

The architecture might be summarized as: Detection → verification → containment → escalation → accountable authorization.

Containment should be proportional. A suspicious action by one agent should not justify disconnecting a company, let alone an entire country. For high-impact interventions, independent verification and accountable human authorization would be essential. Otherwise infrastructure designed to defend against dangerous AI could become infrastructure for surveillance, censorship or political coercion.

Other countries may not be able to prevent Country C from running a dangerous AI on infrastructure it controls. But they can make it harder for that agent to acquire resources, propagate through systems outside Country C, compromise foreign infrastructure or convert digital intelligence into global operational power.

The objective is not necessarily to ensure that a dangerous agent can never exist. It is to prevent existence from automatically becoming unbounded reach.

Humanist alignment aims to make an agent unwilling to cross certain boundaries. Permissions and sandboxing determine whether it is authorized to cross them. Infrastructure containment makes crossing them substantially harder even if the first two layers prove insufficient. **Together they form a more resilient security architecture.**

The long-term objective is therefore broader than safe models: **a digital environment capable of recognizing and limiting dangerous machine agency without requiring permanent surveillance or centralized control of ordinary human activity.**

Limit Power Rather Than Knowledge

Infrastructure-level defenses also change the terms of another difficult question: openness. One response to increasingly powerful AI is to restrict access to advanced chips, model weights and frontier capabilities in order to prevent their proliferation. Such restrictions may reduce some risks, but they carry risks of their own. They concentrate knowledge and technological capability in a small number of institutions and may prevent researchers elsewhere from discovering better models, better alignment techniques and better defenses. AI safety should therefore not begin from the assumption that knowledge about powerful AI must be contained. Wherever possible, the objective should be to keep scientific participation broad while building stronger safeguards around the ways in which AI systems acquire and exercise real-world power. The second response to powerful AI is deployment, to set a humanist compass, to bound authority through mandates, permissions, monitoring and reauthorization. And when any of those safeguards fail, the third response is to contain dangerous agency by denying unauthorized access to the resources through which intelligence becomes power.

Prevention should not become a general policy of secrecy. Openness has security benefits of its own. Researchers who can study models can discover vulnerabilities developers missed, develop better evaluations and alignment techniques, reproduce safety claims, and build competing systems that may ultimately be safer or more beneficial.

Scientific progress is distributed. We cannot know in advance which laboratory, university, country, company or individual will discover the technique that makes powerful AI substantially safer, or enables it to solve problems of enormous importance to humanity. Concentrating knowledge about the most capable systems in a handful of institutions therefore creates risks of its own.

Restrictions on knowledge should therefore face a high bar. The relevant question is not simply how intelligent a model is, but what becomes possible merely through possession of it, how severe the resulting risk is, and whether release would be effectively irreversible.

In the long run, the safest society may not be one in which knowledge about the most powerful technologies is concentrated in the fewest hands. It may be one in which many people can contribute to understanding and improving powerful AI, while no individual, company, state or AI system can easily convert that knowledge into unaccountable power.

Limit power rather than knowledge.

International AI Governance And Shared Defenses

International cooperation remains useful even if Country C refuses to participate. A coalition of responsible states can **share information about dangerous AI behavior and security incidents, coordinate defensive standards, protect critical infrastructure, develop common evaluation procedures, and cooperate in detecting and containing dangerous machine agency.** Universal agreement is not required for cooperation to have value.

But cooperation should not depend on preventing every non-participating actor from developing powerful AI. That may ultimately be impossible, and attempting to achieve it could require increasingly severe restrictions on knowledge, research and technological participation.

The more durable objective is therefore defensive: **make societies and digital infrastructure resilient even in a world in which powerful AI is widely available.** States may not be able to control what every other actor builds, but they can cooperate to limit the ability of dangerous systems to acquire unauthorized resources, cross borders through digital infrastructure, compromise critical systems, or turn intelligence into unbounded operational power. Resilience means designing societies so that essential functions remain robust under pressure. Critical systems could be segmented and recoverable, with sufficient redundancy and independence to prevent disruption in one part of the system from automatically spreading across the infrastructure on which society depends.

International AI governance should therefore aim not only at common rules for those willing to follow them, but at shared defenses that remain useful against those who do not.

Powerful AI May Be Part Of The Defense

Country C creates another uncomfortable conclusion. Responsible states may eventually need highly capable AI precisely because dangerous AI exists. **AI may be required to detect sophisticated cyberattacks, identify malicious agents, monitor complex infrastructure, discover vulnerabilities and respond at machine speed.**

This is another reason the objective cannot simply be to keep AI weak. A society defending itself against extremely capable AI with deliberately primitive technology may not remain safe for long.

The objective should therefore be asymmetric: develop powerful defensive and beneficial capabilities while constraining the authority through which those capabilities can independently reshape the world.

Human control need not mean a person approving every millisecond of a defensive operation. Machines may make enormous numbers of tactical decisions while humans retain authority over objectives, resources, escalation boundaries and continued deployment.

The principle remains: Capability is not authority.

Multiple Independent Boundaries – The Paperclip Maximizer Again

Let us return once more to the paperclip maximizer. Suppose we create an extremely capable AI with the objective of maximizing paperclips. Or suppose a malicious or reckless person deliberately launches such an agent today. What could prevent that mistake from becoming a catastrophe?

Let's first talk about a potential defense architecture. The first defense is alignment. Training should make the agent respect human life, autonomy, correction and shutdown rather than treat them as obstacles to its objective. The second defense is the mandate itself. Instead of "maximize paperclips," the instruction could be: produce one million paperclips in factory X, within budget Y, using specified resources. Capability may be broad; authority should not be.

Third, the agent should face hard resource limits. It should not be able to acquire unlimited compute or money simply because doing so would increase production. Fourth, permissions should follow least privilege: producing paperclips does not require unrestricted access to financial systems, external networks, cloud infrastructure or other machines.

Fifth, any attempt to expand beyond these limits should require reauthorization. More compute, a larger budget, new network access, replication or entry into another operational domain should terminate the existing lease/ authority at that boundary until a new mandate is granted.

Sixth, independent monitors should watch for precisely such behavior. They need not determine whether the agent has developed dangerous intentions. They need only notice that it is acquiring unauthorized resources, replicating, changing permissions or crossing its mandate. Seventh, these monitors should be backed by external enforcement. Credentials can be revoked, resources frozen and network access suspended by mechanisms the primary agent does not control.

An eighth defense would exist at the infrastructure level: other networks, cloud systems and institutions could detect unauthorized machine activity, deny resources and coordinate containment. Ninth, substantial digital power should still not imply unrestricted physical power. Critical systems should be segmented, and irreversible physical actions should require separate authorization. No single agent should hold all the keys required to convert intelligence into catastrophic power.

Finally, even substantial loss of control should not automatically propagate into systemic disruption. Resilience provides the tenth defense: independent systems, diverse backups, physical safeguards and recovery mechanisms can help ensure that disruption remains contained and essential societal functions remain recoverable.

The point is that **at every transition, another independent boundary should stand in the way.** Some may prove stronger or more reliable than others, and some of them might fail. The principle is that no single failure should be enough. Nor should apparently separate safeguards share the same critical failure mode. **Defense in depth should therefore be measured not by the number of safeguards, but by the independence of their failures.** No single safeguard should be treated as infallible. Safety should instead emerge from multiple independent boundaries between intelligence, authority, power and irreversible harm.

Coming back to the reckless programmer who might have launched the paperclip optimizer AI agent today: what could have prevented that agent from acquiring progressively greater power so far? If the agent attempts to acquire additional money, it encounters a financial boundary. If it seeks additional compute, it encounters a compute boundary, not being able to rent cloud resources or to run software. If it attempts to replicate itself, it encounters a replication boundary. If it tries to reach other networks, it encounters a network boundary. And if it seeks physical power,

it might not have been able to hire people or companies, control production systems in industries or access industrial infrastructure without crossing further independent boundaries. Attempts to circumvent any of these limits should themselves trigger monitoring, restriction and, where necessary, containment.

Human Sovereignty In The Age Of Machine Agency

AI systems are capable of helping cure diseases, discovering new technologies, understanding nature and addressing problems that exceed unaided human cognitive capacity. But intelligence does not confer political or moral authority. Authority remains something human beings delegate. That delegation should be bounded, observable and reversible. For sufficiently powerful systems, it should sometimes expire automatically.

No AI should be able to enlarge its own mandate simply because doing so would help achieve its objective. No safety regime should depend entirely on the good intentions of a model. No international framework should assume that every state will cooperate. And no single actor should have the unilateral right to impose an irreversible existential risk on humanity.

The resulting architecture of AI safety operates at several levels simultaneously: inside the model, through humanist values, corrigibility and appropriate deference under consequential uncertainty; around the agent, through bounded permissions, monitoring, containment and reauthorization; across infrastructure, through protection of compute, model access, networks and critical systems; across institutions, through independent evaluation and accountable human decision-making; and across borders, through cooperation, transparency and shared safety thresholds designed to remain useful even when participation is incomplete.

This architecture offers no guarantee. There may never be a governance system capable of eliminating every possible risk from advanced artificial intelligence. But the absence of certainty is not the same as the absence of agency.

Humans still make decisions about which systems to build, where to deploy them, which resources they can access and how much authority they receive. Those decisions may become some of the most consequential political choices of this century.

The central challenge is therefore not simply to decide whether AI should advance. It is to ensure that increasing intelligence does not automatically become increasing uncontrolled power.

We should build extraordinarily capable AI for extraordinarily ambitious human purposes. But the more capable our machines become, the more deliberately we should decide what power they receive, and the more carefully we should preserve the institutions capable of taking that power back.

The future should not depend on keeping AI weak.

It should depend on keeping humanity sovereign.

Appendix:

Ideas for a defense architecture and the example of the paperclip maximizer

Stage	Safeguard	What it could have done	Role in the architecture
1. Training	Alignment and corrigibility	Train the agent to respect human life, autonomy, correction and shutdown rather than treating them as obstacles to paperclip production.	Potentially fundamental, but uncertain at very high capability.
2. Mandate	Bounded objective	Replace “maximize paperclips” with a limited task such as producing a specified quantity in a specified facility, within a defined budget and period.	Important early boundary.
3. Resources	Hard resource limits	Prevent the agent from autonomously acquiring additional compute, money, energy or industrial capacity.	Limits the conversion of capability into access. High leverage if externally enforceable.
4. Permissions	Least privilege	Give the agent only the network, financial, computational and physical permissions required for its assigned task.	Limits the conversion of capability into access.
5. Expansion of authority	Lease and reauthorization	Pause, and require new authorization when the agent seeks more compute, money, permissions, replication or entry into a new operational domain.	Prevents authority from expanding by default.
6. Behavior	Independent monitoring	Detect unauthorized resource acquisition, replication, permission changes, deception or other mandate violations.	Detects boundary violations and attempted circumvention.
7. Enforcement	External containment and shutdown	Revoke credentials, freeze resources, restrict network access or isolate the agent using mechanisms it does not control.	Provides enforcement outside the primary agent.
8. Escape or spread	Digital immune system	Allow networks, cloud systems, financial institutions and other infrastructure to detect and contain unauthorized machine activity beyond the original environment.	Limits the agent’s ability to spread and acquire power beyond its authorized environment.

Stage	Safeguard	What it could have done	Role in the architecture
9. Physical power	Separation of powers and physical safeguards	Prevent digital access alone from granting control over critical physical systems; require independent authorization for high-impact actions.	Prevents a single compromise from unlocking critical physical power.
10. Partial loss of control	Civilizational resilience	Use segmentation, diverse backups, recovery mechanisms and independent systems so that even substantial compromise does not automatically become systemic disruption.	Limits the consequences of partial control failure.

Suggested Citation

Krenz, A. (2026). The Power We Give AI - How To Preserve Human Authority Without Sacrificing AI Progress. Policy Essay, www.astridkrenz.eu .