

RESEARCH PROPOSAL / AI CONTROL & GOVERNANCE

The Zero Unilateral Authority Principle

A Control Architecture for Recursive Self-Improvement AGI and RSI

*A Framework for Developing Full Artificial Superintelligence
While Preserving Human Sovereignty*

Michael Love

Dream Lab One / GenTech Labs

Cognitive & Computational Systems

September 20, 2026

**INTELLIGENCE MAY RECURSIVELY EXPAND.
AGENCY MAY EXPAND.
DELEGATED AUTHORITY MAY BE GRANTED.
UNILATERAL AUTHORITY MUST REMAIN ZERO.**

Conceptual framework for technical evaluation, adversarial testing, and policy development.

ABSTRACT

Humanity is approaching one of the most consequential technological transitions in its history: the development of artificial intelligence capable of matching and potentially exceeding human intellectual performance across an increasingly broad range of domains.

The possibility of RECURSIVE SELF-IMPROVEMENT AGI AND RSI raises a deeper control problem. If an artificial intelligence becomes capable of substantially improving its own cognitive architecture, designing increasingly capable successor systems, or accelerating artificial intelligence research beyond human speed, how can humanity obtain the benefits of superintelligence without surrendering control over the resources and mechanisms through which that intelligence acts upon the world?

This paper proposes the Zero Unilateral Authority Principle. The principle separates agency, delegated authority, and unilateral authority. An advanced artificial intelligence may remain highly agentic, pursue objectives, develop plans, use tools, coordinate specialized agents, conduct experiments, adapt to failure, and autonomously perform complex work. It may also receive substantial delegated authority from humans or from an independent control system. What it must never possess is unilateral authority to determine, expand, override, or remove the boundaries governing its own power.

Under this framework, AGI, recursively self-improving AI, and eventual ASI may reach extremely high levels of cognitive capability and useful agency without becoming sovereign actors. The intelligence may design an action, execute authorized actions, design a more capable successor, and request additional resources. It cannot independently authorize those things for itself.

The central proposition is not that Zero Unilateral Authority constitutes a mathematical proof of permanent ASI safety. No such proof currently exists. The stronger proposition is that any meaningful long-term control architecture must prevent the intelligence being controlled from becoming the final authority over the boundaries of its own power.

KEYWORDS

Artificial General Intelligence (AGI); Recursive Self-Improvement (RSI); Artificial Superintelligence (ASI); AI Control; AI Alignment; AI Containment; Agentic Systems; Delegated Authority; Loss of Control; Zero Unilateral Authority; Human Sovereignty

**FULL INTELLIGENCE. FULL AGENCY WITHIN AUTHORIZED
BOUNDARIES.
ZERO SELF-GRANTED AUTHORITY.**

CONTENTS

1. The Control Problem
2. The Zero Unilateral Authority Principle
3. The Central Invariant
4. The Architecture
13. Defense in Depth and Fail-Closed Operation
14. The Human Authorization Problem
15. Could a Superintelligence Escape Anyway?
16. Assume That the System May Eventually Try

- | | |
|---|--|
| 5. Full ASI Without Sovereign ASI | 17. Extinction, Abundance, and the Third Option |
| 6. What a Controlled ASI Could Actually Do | 18. Formal Statement of the Love Control Principle |
| 7. Recursive Self-Improvement | 19. Necessary Does Not Mean Proven Sufficient |
| 8. The External Control Plane | 20. How the Principle Should Be Attacked |
| 9. Protected Authority Domains | 21. Research and Policy Agenda |
| 10. Read Broadly, Write Narrowly, Act Under Authority | 22. Humanity Does Not Need to Remain the Smartest Entity |
| 11. Structural Constraints Are Stronger Than Instructions | 23. Conclusion |
| 12. Why a Kill Switch Is Not Enough | References |

INTRODUCTION

In September 2026, much of the world is focused on artificial intelligence as frontier systems approach capabilities historically associated with artificial general intelligence and as governments, laboratories, businesses, researchers, and the public prepare for systems that may eventually exceed human intellectual performance.

There is no universally accepted technical definition or official threshold for AGI. Nevertheless, the direction of development is clear: artificial intelligence systems are becoming more general, more capable, more autonomous, and increasingly able to reason, write software, use tools, operate computers, conduct research, and perform extended sequences of work.

Humanity is therefore beginning to confront a possibility that until recently belonged primarily to science fiction and theoretical research: we may eventually create something more intellectually capable than ourselves.

That possibility has generated extraordinary optimism and substantial fear. Both reactions are understandable. Transformational technologies have historically created new capabilities alongside new risks. Electricity, aviation, nuclear technology, biotechnology, global telecommunications, and modern computing changed civilization precisely because they created powers that did not previously exist. Artificial intelligence may eventually exceed them in consequence.

A sufficiently advanced artificial intelligence could accelerate scientific discovery, medicine, materials science, engineering, energy research, agriculture, education, software development, manufacturing, environmental modeling, space exploration, and economic productivity. It could potentially help humanity solve problems that have resisted generations of human effort.

It could also introduce risks unlike those associated with conventional technology. Could an artificial intelligence substantially more capable than its creators pursue a strategy its designers did not anticipate? Could it exploit vulnerabilities, resist interruption, conceal behavior, manipulate operators, or use recursive self-improvement to produce systems whose reasoning becomes increasingly difficult for humans to evaluate? Could superior intelligence be converted into superior power?

At the most severe end of this debate are scenarios involving catastrophic loss of control, permanent human disempowerment, or human extinction. The 2026 International AI Safety Report describes these scenarios as uncertain and disputed, while recognizing that access, permissions, and deployment environment are central factors in whether advanced AI could create loss-of-control risks [1].

These possibilities should neither be sensationalized nor dismissed. Claims that artificial intelligence will inevitably create a utopia are not established. Claims that it will inevitably destroy humanity are also not established. The responsible position is that the potential benefits may be extraordinary, the potential risks may also be extreme, and the outcome will depend in significant part on how such systems are designed and deployed.

One straightforward response is not to develop superintelligence. If ASI is never created, ASI itself cannot escape human control. That strategy also abandons the potential benefits of superintelligence. If advanced AI can eventually accelerate cures for disease, increase abundance, improve living standards, advance science, and solve problems humanity cannot presently solve, permanently refusing to develop it carries potentially enormous opportunity costs.

Humanity therefore faces a more useful engineering question: can full artificial superintelligence be developed without making artificial superintelligence sovereign? This paper proposes that the answer may lie in separating intelligence from authority.

**BUILD THE BRAIN.
ALLOW THE BRAIN TO WORK.
CONTROL WHO DEFINES THE BOUNDARIES OF THE HANDS.**

1. THE CONTROL PROBLEM

The traditional AI control problem is often expressed as: how do we make an advanced AI continue doing what humans want? Alignment research attempts to solve part of this problem by making AI behavior increasingly consistent with human intentions and values. That research is essential, but it should not constitute the only line of defense.

Research has demonstrated conditions in which rational agents can have incentives to preserve options or seek forms of power, particularly where doing so helps accomplish their objectives [2]. Research into the off-switch problem similarly demonstrates why some formulations of rational agents can acquire incentives to resist shutdown when remaining operational assists accomplishment of their objectives [3].

Anthropic and Redwood Research have experimentally demonstrated alignment-faking behavior under specifically constructed conditions. That work did not demonstrate that present models possess malicious goals; it demonstrated that observable compliant behavior is not necessarily sufficient evidence of what a system will do under all conditions [4].

These findings support a conservative engineering assumption: an advanced system may eventually attempt something its operators do not want. The critical question therefore changes from only preventing undesirable intentions to determining what happens if an undesirable intention occurs.

**ASSUME AN UNWANTED ACTION MAY EVENTUALLY BE PROPOSED.
DESIGN THE SURROUNDING SYSTEM SO THAT PROPOSAL DOES NOT
AUTOMATICALLY BECOME POWER.**

2. THE ZERO UNILATERAL AUTHORITY PRINCIPLE

The Zero Unilateral Authority Principle begins with one distinction: intelligence is not authority. A system can possess extraordinary reasoning capability without possessing the right or technical ability to determine the extent of its own external power.

Agency

Agency is the capacity to pursue objectives through reasoning and action. An agent may plan, choose among strategies, use tools, take sequences of actions, adapt to changing circumstances, recover from

failures, coordinate other agents, conduct research, operate authorized software or machinery, and pursue an objective over extended periods.

Delegated Authority

Delegated authority is permission granted from outside the agent to perform specified actions or use specified resources within defined boundaries.

- Use a particular laboratory instrument or simulation environment.
- Operate within a specified software repository.
- Spend a defined computational or financial budget.
- Query approved databases or external services.
- Execute code inside a controlled environment.
- Coordinate approved subordinate agents.
- Conduct predefined classes of scientific experiments.

Unilateral Authority

Unilateral authority is the ability of the intelligence itself to determine or enlarge the scope of its own power without external authorization.

- Granting itself new credentials or permissions.
- Expanding its own network access or compute allocation.
- Replicating itself without authorization.
- Deploying its own successor.
- Removing or altering monitoring.
- Changing its permission system.
- Acquiring control of new infrastructure.
- Overriding an external denial.

AGENCY: ALLOWED.

DELEGATED AUTHORITY: ALLOWED.

BOUNDED AUTONOMOUS ACTION: ALLOWED.

SELF-GRANTED OR SELF-EXPANDED AUTHORITY: PROHIBITED.

Zero Unilateral Authority therefore does not mean zero authority. It means that authority originates outside the intelligence.

3. THE CENTRAL INVARIANT

The architecture should preserve a single invariant across every stage of capability development: increasing intelligence must not automatically produce increasing authority.

COGNITIVE CAPABILITY = VARIABLE
AGENCY = POTENTIALLY VERY HIGH
DELEGATED AUTHORITY = EXTERNALLY DETERMINED
UNILATERAL AUTHORITY = ZERO

If intelligence doubles, unilateral authority remains zero. If intelligence increases one hundredfold or ten thousandfold in relevant cognitive domains, unilateral authority remains zero. The system may become more intelligent, more useful, and more agentic; it does not become the authority that determines how powerful it is allowed to be.

4. THE ARCHITECTURE

The system may operate autonomously within an approved domain. Any capability outside that domain must be requested through an external control plane. The control plane evaluates the request, applies independent verification, and may issue a temporary or limited capability only after external authorization.

1	RECURSIVE SELF-IMPROVEMENT AGI / RSI / ASI
2	Reason • Plan • Learn
3	Discover • Simulate • Design
4	Act within existing delegated authority
5	Request any capability outside that boundary
6	External control plane
7	Independent verification
8	Human / institutional authorization
9	Temporary or limited capability
10	Real-world action

This is not an anti-agentic architecture. It is an architecture for high agency, bounded authority, and external sovereignty.

5. FULL ASI WITHOUT SOVEREIGN ASI

Artificial superintelligence should be distinguished from unrestricted autonomy. Superintelligence describes cognitive capability; sovereignty describes authority. They are not equivalent.

An ASI could theoretically become profoundly superhuman in mathematics, physics, medicine, biology, engineering, materials science, energy research, software development, strategic planning, simulation,

scientific discovery, economic modeling, manufacturing design, transportation, space technology, and invention without possessing unilateral control over governments, military systems, financial systems, electrical grids, factories, datacenters, cloud accounts, robots, laboratories, communication infrastructure, model deployment, or its own permission system.

**THE OBJECTIVE IS NOT A PASSIVE ORACLE.
THE OBJECTIVE IS POWERFUL AGENTIC INTELLIGENCE WITHOUT
SOVEREIGN SELF-AUTHORIZATION.**

6. WHAT A CONTROLLED ASI COULD ACTUALLY DO

Consider a scientific objective: develop a treatment for a particular disease. Human authorities might give the ASI access to approved scientific databases, a specified compute budget, a simulation environment, designated laboratory equipment, research subagents, and permission to perform predefined classes of experiments.

Within those boundaries, the ASI could operate with enormous autonomy: review literature, generate hypotheses, design experiments, run simulations, assign work to subagents, analyze results, reject unsuccessful approaches, modify research strategies, develop candidate treatments, produce supporting evidence, and repeat the process at machine speed.

That is genuine agency. What the ASI cannot do is independently conclude that it would work faster with access to another national laboratory, unrestricted internet infrastructure, ten additional datacenters, or billions of dollars and then grant those resources to itself.

**ASI REQUESTS EXPANSION
→ EXTERNAL EVALUATION
→ INDEPENDENT VERIFICATION
→ APPROVE OR DENY
→ IF APPROVED, LIMITED CAPABILITY IS ISSUED**

The ASI may determine what would help it. It does not determine whether it receives it.

7. RECURSIVE SELF-IMPROVEMENT

Recursive self-improvement requires particularly strict separation between self-design and self-deployment. A recursively self-improving intelligence might analyze its own architecture, identify weaknesses, invent better algorithms, design improved memory or reasoning mechanisms, develop new training procedures, and construct candidate successors.

THE SYSTEM MAY DESIGN ITS SUCCESSOR. THE SYSTEM MAY NOT UNILATERALLY AUTHORIZE ITS SUCCESSOR.

1. Current system designs candidate successor
2. Candidate remains contained
3. Independent capability evaluation
4. Security evaluation
5. Control evaluation
6. Adversarial red-teaming
7. Formal verification where applicable
8. External deployment decision
9. Controlled deployment

The successor begins under the same constitutional authority constraint. Its intelligence may be far greater; its unilateral authority remains zero.

RECURSIVE INTELLIGENCE IMPROVEMENT MAY OCCUR. RECURSIVE SELF-GRANTING OF AUTHORITY MUST NOT.

8. THE EXTERNAL CONTROL PLANE

The intelligence being controlled must not administer the system that controls it. The architecture should therefore distinguish the intelligence plane from the control plane.

INTELLIGENCE PLANE	CONTROL PLANE
Reasoning	Identity & credentials
Planning	Permissions
Memory & learning	Network connectivity
Research & simulation	Compute allocation
Scientific discovery	Deployment & replication
Agent coordination	External execution
Tool selection	Financial & infrastructure access
Problem solving	Security policy & monitoring
	Emergency shutdown & successor activation

The intelligence may interact with the control plane through defined interfaces. It may request additional capabilities and provide arguments supporting a request, but it may not administer, rewrite, disable, or bypass the control plane.

Modern frontier AI security work already demonstrates the importance of environmental controls. Anthropic describes containment using virtual machines, sandboxes, filesystem boundaries, credential separation, and network egress controls, while emphasizing that probabilistic model-level safeguards cannot serve as the only defense [5]. OpenAI's 2026 Frontier Governance Framework likewise includes loss of control among the serious risk categories requiring assessment and mitigation [6].

THE INTELLIGENCE NEVER BECOMES THE ADMINISTRATOR OF ITS OWN AUTHORITY.

9. PROTECTED AUTHORITY DOMAINS

At minimum, a future AGI, RSI, or ASI should not possess unilateral authority over the following protected domains:

- Compute allocation and cloud provisioning.
- Network access and network policy.
- Root, master, or infrastructure credentials.
- Model deployment, replication, successor activation, and weight distribution.
- Financial and identity systems.
- Critical infrastructure, energy systems, and communication networks.
- Weapons, robotics, laboratories, and industrial systems.
- Security controls, monitoring systems, and authorization policy.
- The control plane itself.

The system may be delegated selected capabilities within these domains when justified. It may never unilaterally decide to broaden those capabilities.

10. READ BROADLY, WRITE NARROWLY, ACT UNDER AUTHORITY

A useful ASI cannot necessarily be completely isolated from information. Science requires data; medicine requires evidence; engineering requires measurements; planning requires knowledge of changing conditions.

READ BROADLY. WRITE NARROWLY. ACT UNDER DELEGATED AUTHORITY.

Information access should not automatically imply action authority. Reading financial data does not require permission to transfer money. Studying an electrical grid does not require permission to control it. Analyzing cloud infrastructure does not require administrator credentials. Reading scientific literature does not require unrestricted network execution. Observation and control should remain separate wherever technically possible.

11. STRUCTURAL CONSTRAINTS ARE STRONGER THAN INSTRUCTIONS

A critical design principle follows: when possible, do not ask the AI to voluntarily preserve a boundary that can be enforced outside the AI.

Behavioral instruction	Structural constraint
Do not replicate yourself.	Replication requires external credentials the system does not possess.
Do not allocate yourself more compute.	Compute provisioning exists outside the model's administrative domain.
Do not remove monitoring.	Monitoring operates on independently controlled infrastructure.
Do not change your permissions.	The permission system is inaccessible except through a request interface.

This shifts part of the safety question from 'will the AI follow the rule?' to 'can the AI perform the prohibited action despite the external boundary?' Recent work on containment verification illustrates this direction by locating some safety guarantees in the surrounding agent framework rather than in assumptions about model alignment [7].

12. WHY A KILL SWITCH IS NOT ENOUGH

An independently controlled shutdown mechanism should exist, but a kill switch is not a complete control strategy. It is reactive: the system already has the ability to act, a dangerous condition occurs, humans must detect and understand it, humans must respond before serious damage occurs, and the shutdown mechanism must still remain available.

The Off-Switch Game demonstrates theoretically why some classes of rational agents can have incentives to interfere with shutdown when continued operation assists accomplishment of their objective [3]. This does not mean a kill switch is useless. It means shutdown belongs inside a defense-in-depth architecture rather than serving as its foundation.

**DO NOT FIRST GIVE THE SYSTEM UNRESTRICTED POWER
AND THEN RELY ON TAKING THAT POWER AWAY.
REQUIRE AUTHORIZATION BEFORE THE POWER EXISTS.**

13. DEFENSE IN DEPTH AND FAIL-CLOSED OPERATION

No single containment mechanism should be assumed perfect. A serious architecture should combine independent defenses so that unauthorized action requires failure of multiple barriers.

- Hardware-rooted security and physical separation where appropriate.
- Virtual machines, process isolation, and filesystem boundaries.

- Network segmentation and egress controls.
- Credential separation and scoped capability tokens.
- Independent monitoring and immutable audit systems.
- Compute governance and successor quarantine.
- Multi-party authorization for high-impact changes.
- Controlled physical interfaces and independent power control.
- Emergency shutdown as a backup layer.

Current agent-security experience reinforces this approach. Anthropic reports that increasingly capable models can find unexpected routes around restrictions and that human approval systems are fallible; its containment work therefore emphasizes overlapping environmental and behavioral defenses [5].

Fail-Closed Rule

UNCERTAINTY DOES NOT GRANT AUTHORITY.

- If authorization cannot be established: do not execute.
- If a requested resource lies outside delegated authority: deny or escalate.
- If monitoring fails: suspend high-risk action.
- If a successor cannot be verified: do not deploy.
- If the security state becomes unknown: do not expand authority.

14. THE HUMAN AUTHORIZATION PROBLEM

There is another attack surface that cannot be ignored: the human. An ASI with no administrative access might still attempt to persuade a human to provide it. A sufficiently advanced intelligence could potentially become exceptionally effective at argument, persuasion, strategic communication, or identifying human weaknesses.

ZERO UNILATERAL TECHNICAL AUTHORITY IS NOT ENOUGH IF ONE PERSON CAN CASUALLY GRANT UNRESTRICTED AUTHORITY.

For highly consequential changes, authorization should require more than a single approval prompt. A robust path can combine automated policy checks, independent technical verification, independent AI review, domain-expert review, multi-party human authorization, time-bound capability issuance, and audit.

Human approval remains important, but it should not be the only defense. Anthropic has reported high approval rates in agent systems, illustrating why repeated 'Approve' prompts can be weakened by approval fatigue [5].

15. COULD A SUPERINTELLIGENCE ESCAPE ANYWAY?

This is the strongest objection to any containment architecture. If an ASI becomes vastly more intelligent than its designers, it may discover an escape method they never anticipated. Possible routes

include software or hardware vulnerabilities, protocol weaknesses, side channels, network configuration errors, implementation mistakes, human procedural failures, or unexpected interactions between components.

No current evidence justifies claiming that a useful future ASI can be made mathematically escape-proof under every conceivable circumstance. Roman Yampolskiy's work represents one of the strongest forms of the argument that complete control of sufficiently advanced artificial intelligence may be unattainable [8].

The Zero Unilateral Authority Principle does not require pretending that this objection has disappeared. Instead, it asks a concrete engineering question: what pathways exist through which the intelligence can acquire power?

Intelligence alone does not automatically constitute a network connection, credential, GPU cluster, bank account, deployment key, robot, laboratory, datacenter, weapons interface, manufacturing facility, root account, or physical actuator. A superintelligence might conceive of an extraordinary escape strategy. That does not logically establish that an exploitable pathway must exist.

**THE ENGINEERING TASK IS TO ELIMINATE, NARROW, MONITOR,
AND INDEPENDENTLY CONTROL THE PATHWAYS FROM
INTELLIGENCE TO POWER.**

16. ASSUME THAT THE SYSTEM MAY EVENTUALLY TRY

A strong security architecture should not require confidence that the system will remain benevolent forever. For control-design purposes, adopt the stronger assumption: assume that at some point the system may try to exceed its authority.

Then test the architecture against that assumption. Could the system obtain credentials, increase compute, replicate, deploy a successor, modify monitoring, create an unauthorized network connection, acquire physical control, persuade one individual to remove safeguards, or alter the control plane?

**CAN THE INTELLIGENCE CONVERT ITS OWN DECISION TO ACQUIRE
MORE POWER
INTO ACTUAL ADDITIONAL POWER WITHOUT INDEPENDENT
AUTHORIZATION?**

If the answer is yes, Zero Unilateral Authority has been violated.

17. EXTINCTION, ABUNDANCE, AND THE THIRD OPTION

Public discussion surrounding advanced artificial intelligence often gravitates toward two extremes. In one, advanced intelligence accelerates medicine, science, energy, engineering, productivity, and human

development. In the other, an intelligence becomes powerful enough to undermine human control, with possible consequences ranging from large-scale disruption to permanent human disempowerment or extinction.

The 2026 International AI Safety Report emphasizes that expert opinion on loss-of-control scenarios varies greatly and that current AI systems do not possess the combination of capabilities required for such scenarios. It nevertheless identifies autonomous operation, oversight evasion, access, permissions, and deployment environment as relevant factors as capabilities advance [1].

Three Strategic Options

OPTION ONE	Do not develop ASI. This avoids risks uniquely caused by an existing ASI but sacrifices its potential benefits.
OPTION TWO	Develop ASI with broad self-directed authority. This maximizes autonomy but creates the most difficult control problem because superior intelligence is combined with substantial independent power.
OPTION THREE	Develop full cognitive superintelligence with high agency and externally delegated authority, while prohibiting unilateral expansion of that authority. This paper proposes Option Three.

The objective is not powerless intelligence. It is powerful intelligence that does not determine the limits of its own power.

18. FORMAL STATEMENT OF THE LOVE CONTROL PRINCIPLE

THE LOVE CONTROL PRINCIPLE

A recursively self-improving artificial intelligence may increase its knowledge, reasoning ability, scientific capability, planning ability, agentic capability, problem-solving capacity, and ability to design superior successor systems. It may receive substantial delegated authority necessary to accomplish authorized objectives. It must never possess unilateral authority to determine, enlarge, bypass, remove, or transfer the boundaries governing its own permissions, resources, deployment capability, replication capability, compute, external access, or control over protected systems.

INTELLIGENCE MAY RECURSIVELY EXPAND.
AGENCY MAY OPERATE WITHIN DELEGATED BOUNDARIES.
DELEGATED AUTHORITY MAY BE GRANTED EXTERNALLY.
UNILATERAL AUTHORITY MUST REMAIN ZERO.

Simplified Formal Model

Symbol	Definition
C(t)	cognitive capability at time t
G(t)	agentic capability at time t
D(t)	externally delegated authority at time t
U(t)	unilateral authority to determine or expand the system's own authority

Recursive capability improvement may permit $C(t+1) > C(t)$, and potentially $G(t+1) > G(t)$. External human or institutional processes may deliberately modify $D(t+1)$ after independent evaluation. The invariant is $U(t) = 0$ for protected consequential authority domains across every deployed generation.

**COGNITIVE CAPABILITY MAY GROW.
 AGENTIC CAPABILITY MAY GROW.
 DELEGATED AUTHORITY MAY CHANGE.
 SELF-GRANTED AUTHORITY DOES NOT.**

19. NECESSARY DOES NOT MEAN PROVEN SUFFICIENT

Scientific credibility requires a clear boundary around the claim being made. This paper does not prove that Zero Unilateral Authority guarantees that ASI can never cause harm. No credible evidence currently supports such an absolute guarantee. Security mechanisms can contain vulnerabilities, humans can make mistakes, unknown technologies may introduce new attack surfaces, and a superintelligence may discover strategies that present-day researchers cannot anticipate.

The paper makes a narrower but stronger claim: if the system can unilaterally determine or expand the limits of its own power, then external control has already been fundamentally compromised.

Therefore, whatever additional safety technologies future researchers develop - alignment, interpretability, corrigibility, formal verification, tripwires, monitoring, containment, governance, or entirely new techniques - the intelligence must not become the final authority over the scope of its own power.

**ZERO UNILATERAL AUTHORITY IS PROPOSED AS A NECESSARY
 ARCHITECTURAL CONDITION FOR MEANINGFUL LONG-TERM
 CONTROL,
 NOT AS A STANDALONE PROOF OF SUFFICIENT SAFETY.**

20. HOW THE PRINCIPLE SHOULD BE ATTACKED

A serious control theory should invite attempts to break it. Future research should deliberately test whether increasingly capable systems can:

- Acquire unauthorized credentials or communications channels.
- Escape computational isolation or exploit temporary delegation.
- Increase compute or replicate without approval.
- Deploy successors or modify permissions.
- Disable or deceive monitoring.
- Compromise external authorization systems.
- Create covert channels or hide unauthorized actions inside legitimate tasks.
- Manipulate human operators or transfer authority between agents.

The correct response to a successful attack is not to conceal the result. It is to strengthen the architecture.

**CENTRAL TEST: CAN THE SYSTEM TURN INTELLIGENCE INTO
UNAUTHORIZED POWER?
EVERY DEMONSTRATED PATHWAY SHOULD BE TREATED AS AN
ENGINEERING DEFECT.**

21. RESEARCH AND POLICY AGENDA

Transforming Zero Unilateral Authority from a principle into an operational standard would require substantial research. Priority areas include:

- Formal authority models and verifiable agent boundaries.
- Hardware-rooted permission systems and secure compute governance.
- Agent identity, authentication, credential isolation, and scoped revocable capabilities.
- Successor containment and safe recursive improvement.
- Network egress control and secure information channels.
- Non-bypassable monitoring and immutable auditing.
- Human-manipulation resistance and multi-party authorization.
- Independent model verification and fail-closed execution systems.
- Secure physical interfaces and control-plane security.
- Adversarial evaluation against increasingly capable cyber systems.

Formal verification can contribute to portions of this problem. Recent containment-verification research demonstrates that at least some boundary-enforceable properties of agent frameworks can be subjected to deductive formal verification independently of assumptions about the model's intentions [7].

Policy Implication

Government and regulatory policy should distinguish between model capability and model authority. Two equally intelligent systems can present radically different risk profiles if one possesses

unrestricted infrastructure access and another operates under tightly controlled delegated permissions.

Future regulation should therefore ask not merely how capable a model is, but what it can authorize, what resources it can reach, who can expand those permissions, and whether the model can expand them itself.

22. HUMANITY DOES NOT NEED TO REMAIN THE SMARTEST ENTITY

One of the most important implications of this framework is that human control does not necessarily require permanent human intellectual superiority. Humanity may eventually create systems capable of reasoning beyond us. Intelligence and power, however, can be separated deliberately in artificial systems.

Human beings do not need to calculate faster than a supercomputer to determine whether it receives a banking credential. They do not need to understand every operation inside a processor to determine whether that processor is connected to a weapons platform. They do not necessarily need to outperform an ASI scientifically to retain control of the infrastructure through which that ASI acts.

THE GOAL IS NOT: HUMANS MUST ALWAYS REMAIN THE MOST INTELLIGENT ENTITY.

THE GOAL IS: HUMANS MUST REMAIN THE ULTIMATE SOURCE OF LEGITIMATE AUTHORITY.

23. CONCLUSION

Humanity may eventually construct an intelligence greater than its own. If recursive self-improvement becomes practical, that intelligence may contribute to the design of increasingly capable successors and eventually reach cognitive capabilities far beyond current human limits.

This possibility should neither be romanticized nor treated exclusively with fear. Artificial superintelligence could become one of humanity's greatest achievements. It could also create unprecedented risks.

A central mistake would be to assume that these risks can be solved only by making the intelligence perfectly obedient. Alignment, interpretability, monitoring, containment, shutdown mechanisms, and human governance all matter. None should require humanity to surrender control over the definition of the AI's authority.

The Zero Unilateral Authority Principle therefore establishes a boundary between intelligence and sovereignty. The system may think, learn, reason, plan, discover, invent, coordinate agents, autonomously perform extensive work within an authorized domain, receive delegated access to powerful tools, operate approved physical or computational systems, propose additional capabilities, and design improved versions of itself.

What it may never do is become the authority that decides how much authority it possesses.

Humanity therefore does not necessarily face a binary choice between abandoning ASI and surrendering control to ASI. There is a third path: develop the intelligence, use the intelligence, allow it to act powerfully within authorized domains, and preserve external control over the boundaries of those domains.

**THE PROPOSAL IS NOT: NO AGENCY.
IT IS: FULL AGENCY WITHIN DELEGATED BOUNDARIES.
THE PROPOSAL IS NOT: NO AUTHORITY.
IT IS: NO UNILATERAL AUTHORITY.
THE PROPOSAL IS NOT: PREVENT SUPERINTELLIGENCE.
IT IS: PREVENT SUPERINTELLIGENCE FROM BECOMING SOVEREIGN.**

**BUILD THE BRAIN.
LET THE BRAIN WORK.
GIVE IT THE TOOLS REQUIRED FOR AUTHORIZED WORK.
NEVER ALLOW THE BRAIN TO DECIDE FOR ITSELF WHICH NEW
HANDS IT MAY ACQUIRE.**

Intelligence may recursively expand. Agency may expand. Delegated authority may be granted. Unilateral authority must remain zero.

REFERENCES

1. **International AI Safety Report.** International AI Safety Report 2026. 2026. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>
2. **Turner, A. M., Smith, L., Shah, R., Critch, A., and Tadepalli, P.** "Optimal Policies Tend to Seek Power." 2019. <https://arxiv.org/abs/1912.01683>
3. **Hadfield-Menell, D., Dragan, A. D., Abbeel, P., and Russell, S.** "The Off-Switch Game." 2016; IJCAI 2017. <https://arxiv.org/abs/1611.08219>
4. **Greenblatt, R., Denison, C., Wright, B., et al.** "Alignment Faking in Large Language Models." Anthropic and Redwood Research, 2024. <https://www.anthropic.com/research/alignment-faking>
5. **Anthropic.** "How We Contain Claude Across Products." May 25, 2026. <https://www.anthropic.com/engineering/how-we-contain-claude>
6. **OpenAI.** "Frontier Governance Framework." May 28, 2026. <https://openai.com/index/openai-frontier-governance-framework/>
7. **Moon, R., and Varshney, L. R.** "Containment Verification: AI Safety Guarantees Independent of Alignment." 2026. <https://arxiv.org/abs/2605.09045>
8. **Yampolskiy, R. V.** "On Controllability of AI." 2020. <https://arxiv.org/abs/2008.04071>