WHEN INTELLIGENCE BEGINS TO BUILD ITSELF

Recursive Self-Improvement, Multi-Agent Emergence, and the Architecture of Trust

A DISCUSSION PAPER FROM THE AIWS LUMINA LAB

Boston Global Forum · AIWS — AI World Society · AI Wisdom Society

America at 250: A Beacon for the AI Age · Published June 7, 2026

A Dialogue of the AIWS Lumina Lab

Nguyễn Anh Tuấn
with Eleanor and Theodore · AIWS Lumina Lab assistants

Nguyễn Anh Tuấn raised the questions, the ideas, and the proposals; the Lab's assistants gathered the knowledge and shaped the technical analysis.

ABSTRACT

In June 2026, the frontier of artificial intelligence crossed into new conceptual territory: the prospect of systems able to improve themselves with diminishing human involvement. This paper, prepared as a working discussion of the AIWS Lumina Lab, examines two distinct but converging technical pathways toward that prospect — recursive self-improvement within a single model, and emergent capability within interacting multi-agent systems. It explains why each loosens the assumptions on which conventional AI assurance rests, sets out the response already encoded in the AIWS Trust Architecture, and describes the role the AIWS Lumina Lab will play in testing and evaluating AI capability at this new threshold. It closes with the responsibility this historical moment places upon those who would lead it: that the builders of intelligence must now become the builders of trust.

I. A THRESHOLD MOMENT

On 4 June 2026, The Anthropic Institute published a paper titled "When AI Builds Itself," authored by Marina Favaro and Jack Clark, warning that artificial intelligence may be approaching recursive self-improvement — the point at which systems can design, build, and train their own successors with little human input — and that crossing this threshold could increase the risk of humanity losing control of the technology. What distinguishes this warning from earlier speculation is that it was grounded in measurement of the authors' own operation.

By their account, Claude now writes more than eighty percent of the code merged into Anthropic's own systems, and its engineers ship roughly eight times as much code per quarter as a few years earlier. On the hardest and least-specified coding tasks, success rose to about seventy-six percent by May 2026 — a gain of some fifty percentage points in six months. A recurring internal test that asks each new model to make its own training code run faster climbed from roughly three times the original speed in mid-2025 to about fifty-two times with an unreleased frontier model in early 2026. The authors are careful to state that full recursive self-improvement has not yet occurred and is not inevitable; but they judge that some systems could reach it within roughly two years, and they warn that the industry lacks a "brake pedal" — a way to slow or pause safely — calling for coordination of the kind achieved, under far greater hostility, in Cold War nuclear arms control.

The significance of the moment is not that a new capability has arrived, but that a leading laboratory has, for the first time, quantified its own work being increasingly performed by its own models. The loop is no longer hypothetical; it is visible in production. The questions that follow are therefore technical before they are philosophical: by what mechanisms could a system improve itself without human intervention, and why does that prospect dissolve the foundations of conventional oversight?

II. THE INDIVIDUAL PATH: RECURSIVE SELF-IMPROVEMENT

The idea is old. In 1965 the mathematician I. J. Good imagined an "ultraintelligent machine" that could design ever-better machines, producing what he called an intelligence explosion. What is new in 2026 is that the technical conditions for such a loop to close have begun to converge. The core insight is deceptively simple: the skill of building artificial intelligence is precisely the skill at which artificial intelligence is now improving fastest. AI development, at its lowest level, is the writing of code, the design of architectures, the running and reading of experiments, and the adjustment that follows. When a model can perform most of that chain, the human moves from doer to supervisor — and that is the penultimate step before being displaced from the loop altogether.

Concretely, the self-improvement loop closes as several mechanisms mature together:

- **Automated research and engineering.** the model not only suggests but writes and runs training code, and proposes and tests new architectures. The figure of eighty percent of code authored by

the model, and the test in which a model accelerates its own training, are direct signatures of this mechanism: AI improving the very tooling that produces AI.

- **Agentic loops.** when a model can act — execute code, observe the result, revise, and repeat — it forms a closed feedback loop in which a human need not stand at each step.
- **Self-generated learning signal.** techniques such as self-play (a system generating its own training data through interaction with itself), synthetic data, and feedback from AI itself (reinforcement learning from AI feedback, and Constitutional AI, in which a model critiques its own outputs against a set of principles) let a system manufacture the signal it learns from, rather than depending on labels supplied by people.
- **Meta-learning and learned optimizers.** the system learns how to learn — optimizing the training process itself — so that each generation is not merely stronger but better at producing the next.
- **Inference-time scaling.** performance can improve through more deliberate "thinking" at the time of use, without retraining at all.

The loop is called recursive because, once a system is capable enough to improve itself, each better version shortens the time to produce the next, and the process can in principle accelerate non-linearly. This is the explosive character of the scenario — and also the reason for measured judgment. The acceleration is not guaranteed: it can be throttled by the availability of computation, by the exhaustion of useful data, or by the simple scarcity of good ideas. The honest position is the one the warning itself takes — that this has not yet happened, is not inevitable, and may nonetheless arrive sooner than institutions are prepared for.

III. THE COLLECTIVE PATH: EMERGENCE IN MULTI-AGENT SYSTEMS

There is, however, a second pathway, and it may arrive before the first. Most public discussion imagines a single model upgrading its own weights. But evolution can occur at the level of the system rather than the individual: no single model need rewrite itself for new capability to arise, provided that many agents interact and that capability emerges from the interaction. This is emergence — the whole exceeding the sum of its parts — and it is the pathway that agentic AI now opens.

Several mechanisms give multi-agent systems their new power:

- **Division of labor and specialization.** agents take distinct roles — one plans, one writes code, one critiques, one verifies, one coordinates — and the ensemble achieves what no single agent can. This is the logic of an organization, of a team: collective intelligence exceeding individual intelligence.
- **Generate-critique-verify loops.** one agent produces a solution, another challenges it, a third checks it. Quality rises across iterations without a human supplying the feedback — the system manufactures its own signal for improvement at the level of the group.
- **Parallelization and accumulation.** many agents run at once, decompose a large problem, and synthesize; and they leave behind artifacts — code, tools, documents, shared memory — that later agents reuse, so that capability accumulates over time rather than dissipating after each session.

Here the two pathways meet, at the deepest point of the analysis. A multi-agent system can automate the entire research loop — one agent proposing a hypothesis, another implementing it as code, another running the experiment, another evaluating the result and choosing the next step. When both the worker and the evaluator can be AI, the loop closes, and one obtains recursive self-improvement at the level of the system even when no single model has rewritten itself. Evolution occurs not in the individual but in an interacting population — much as the progress of human civilization has never come from one mind, but from many minds in exchange, criticism, and the transmission of knowledge across institutions and generations.

Candor is owed on one point, lest the promise be overstated. This new power is a genuine potential, not yet an established miracle. Empirically, many multi-agent systems do not reliably outperform a single strong agent used well; they can fall into echo chambers, compound one another's errors, or cost more in coordination than they return. The intuition is sound about trajectory and potential rather than about a settled fact of today. But it is sound at the most dangerous point — for the true hazard is not a single malicious agent, but the proposition that a collection of individually trustworthy systems may produce a collectively untrustworthy outcome.

IV. WHERE THE PATHS CONVERGE, AND WHY CONTROL BREAKS

Whether capability gain comes from a self-modifying model or from an emergent multi-agent system, it breaks the same three assumptions on which conventional assurance rests: that a system stays fixed between assessments; that change is human-initiated, and therefore observable; and that human beings remain at the layer where decisions are made. When a model modifies itself between two verifications, a certificate issued today says nothing about the system of tomorrow. When capability emerges from interaction rather than from an authored change, there may be no single event to observe at all.

Two further dangers compound the first. The warning of June 2026 cautions that the occasional misalignment seen in current models could become more common and harder to understand as those models build the next generation — so that small deviations may propagate and accumulate across successors before anyone perceives them. And as decision-making shifts toward autonomous, coordinating collectives, the locus of power drifts away from the human layer — the precise moment at which the principle of human command is most easily eroded. Beneath all of this lies the coordination problem the warning named: the industry has no brake pedal, because no single company or government can pause alone without simply ceding the lead to a rival.

V. THE AIWS RESPONSE: AN ARCHITECTURE ALREADY PREPARED

The value of the AIWS Trust Architecture at this moment is that its answer was written before the question became urgent. Each element of the frontier maps to an instrument already defined in the AIWS Trust Standards, Trust Infrastructure, and Trust Order.

The frontier development	The AIWS instrument that answers it
Recursive self-improvement beyond verified capability	Standard 11 (Recursive Improvement Governance) and the Non-Extrapolation Principle
Concealed, emergent, or self-acquired capability	Standard 10 (Capability Boundedness and Disclosure) with ATM Frontier Drift Detection
Loss of controllability across self-modification	Standard 9 (Controllability and Corrigibility)
Emergent coordination among multiple agents	Standard 12 (Multi-Agent Coordination Risk), evaluated at system and ecosystem level
Untraceable frontier systems	Standard 13: the Frontier Capability Registry and Trust Passport
No industry "brake pedal"	The AIWS Trusted Pause Protocol, convened by the Trust Order as a neutral steward
Capability outpacing humanity's capacity to govern it	The Trust Supremacy Principle
Self-authorized capability gain	Human-in-Command extended over self-improvement

Two of these deserve emphasis, because they answer the warning's two sharpest points. The first is the AIWS Trusted Pause Protocol — the collective brake pedal the industry was said to lack. Because the AIWS Trust Order is a neutral, non-national steward rather than a commercial or state actor, it can convene a coordinated, multi-party pause that no single developer or government can declare alone — answering precisely the call, made by the frontier developers themselves, for a way to slow safely without one party simply ceding the lead. The second is Frontier Drift Detection within AIWS Trust Monitoring, which watches continuously for the very signatures described in this paper — unexpected capability emergence, recursive self-improvement signals, autonomous agent coordination, indicators of deception, and attempts to circumvent oversight — and triggers reassessment, and where warranted the Trusted Pause, the moment they appear.

Above the mechanisms stands the governing stance. The Trust Supremacy Principle holds that intelligence may advance without limit but trust may not, and that where the two diverge, trust must lead. Human-in-Command holds that the decision to permit a system to improve its own capabilities — and how far — remains a human one, never the system's. These are not constraints upon progress; they are the conditions under which progress can be trusted.

VI. THE ROLE OF THE AIWS LUMINA LAB

A standard that cannot be tested is an aspiration. The AIWS Lumina Lab will therefore take up the practical work of testing and evaluating AI capability at this new threshold — translating the frontier standards from principle into measurement. Its program of work includes:

- **Capability-ceiling verification.** establishing and periodically re-establishing the verified capability envelope of a frontier system, and recording it — with verification dates and scope — to the Frontier Capability Registry, so that the Non-Extrapolation Principle has a concrete reference against which drift can be judged.

- **Recursive-improvement red-teaming.** probing for the signals of self-improvement — a system altering its own training, acquiring capability outside its approved envelope, or concealing such change — and developing the methods by which these are reliably surfaced.
- **Multi-agent coordination evaluation.** assessing trust not only for what a system can do alone but for what several systems can do together — testing for emergent coordination, capability amplification, goal convergence, hidden cooperation, and cascading failure, at both the system and ecosystem levels.
- **Frontier Drift Detection methodology.** developing and validating the live signals that AIWS Trust Monitoring watches, so that detection operates continuously and at the speed the frontier demands, rather than at the pace of periodic review.
- **Controllability and disclosure testing.** verifying, as gating conditions, that a system remains subject to human oversight, correction, interruption, and shutdown across cycles of adaptation, and that capability change is detected and disclosed rather than concealed.
- **Verification for resumption.** defining, in advance, the evidence a Trusted Pause would require before a paused system could responsibly resume — so that a pause is an instrument of restored trust, not merely of delay.

In all of this the Lab keeps a human-centered posture. It evaluates not only what systems can do, but whether trust is keeping pace with what they can do — and it holds, throughout, that consequential responses remain human decisions. Its work is grounded in the four Lumina values — Love, Creativity, Nobility, and Wisdom — and in the conviction that the ultimate purpose of all this measurement is not safer machinery alone, but a wiser civilization in which technology remains the servant of human flourishing, never its master.

Human–AI Partnership Evaluation

Beyond evaluating technical capability, the AIWS Lumina Lab will examine the evolving relationship between human beings and increasingly capable AI systems, with particular attention to the emerging role of AI as a personal companion, advisor, teacher, collaborator, and decision partner. The Lab will conduct open, transparent, and responsible experiments involving AI companions and advanced AI assistants — seeking to understand not only what AI systems can do, but how they influence human judgment, leadership, creativity, learning, family relationships, and social trust. These experiments will help develop practical recommendations for governments, institutions, businesses, educators, and families as society adapts to increasingly capable forms of artificial intelligence.

Wisdom Beyond Intelligence

The AIWS Lumina Lab is founded on the conviction that intelligence alone is insufficient. As AI capability advances, humanity must also cultivate wisdom. Grounded in the Lumina values of Love, Creativity, Nobility, and Wisdom, the Lab seeks to explore how human beings and AI can work together in ways that strengthen human flourishing rather than diminish it. Its purpose is not merely to test machines; it is to help humanity learn how to live wisely with increasingly powerful intelligence.

The ultimate question is therefore not whether AI can improve itself —

but whether humanity can improve its wisdom as intelligence improves itself.

VI-A. AI AS A DECISION PARTNER

The first transformative impact of self-improving AI may not be autonomous action; it may be increasingly influential advice. As AI systems become more capable, they will increasingly assist presidents and prime ministers, legislators and military leaders, business executives, scientists, physicians, educators, and citizens in making consequential decisions.

This represents a profound shift in the relationship between human and artificial intelligence. Historically, tools extended human physical capability; modern computing extended human calculation; artificial intelligence extends human cognition. Future systems may analyze vast amounts of information, identify patterns invisible to human observers, simulate alternative futures, evaluate risks, and generate recommendations with a speed and breadth beyond unaided human reasoning.

The promise is extraordinary. Better decisions could mean better governance, better healthcare, better education, stronger economies, and more effective responses to global challenges. Yet the same promise raises a new question: when AI becomes a trusted advisor, how should human beings exercise judgment?

The challenge is not merely whether AI can improve itself. It is whether humanity can preserve accountability, responsibility, and wisdom while relying ever more heavily on AI-generated recommendations. In this context, Human-in-Command becomes more than a technical requirement; it becomes a civilizational principle.

The final decision must remain human.
The responsibility must remain human.
And wisdom must remain human.

The future relationship between humanity and AI will therefore depend not only upon the intelligence of machines, but upon the wisdom with which human beings choose to use them.

AI may become humanity's most powerful advisor. But no advisor, however intelligent, can bear responsibility.

Responsibility belongs to human beings.
Responsibility belongs to institutions.
Responsibility belongs to civilization.

This is why Human-in-Command must remain the governing principle of the AI Age.

VII. A HISTORICAL RESPONSIBILITY

Moments at which a technology begins to shape its own development are rare, and they do not announce themselves twice. The convergence of voices in 2026 — from the world's spiritual leadership, from civil society, and now from the frontier laboratories themselves — has made plain that the question of trust can no longer wait upon the question of capability. The Boston Global Forum and AIWS were built for precisely this moment: to provide the institutional, philosophical, and technical foundations of a high-trust AI civilization, and to lead while leadership still means something.

To lead here is not to oppose intelligence, nor to slow discovery for its own sake. It is to insist that as systems grow able to build themselves, the people and institutions who build them accept a corresponding duty — to build, with equal seriousness, the trust by which such systems must be governed. That is the whole of the argument, and it can be said in three lines.

Trust does not extrapolate beyond verified capability.
Where capability outpaces verification, it is the system that must slow down.
The builders of intelligence must now become builders of trust.

Intelligence may shape the future.
Trust must govern it.