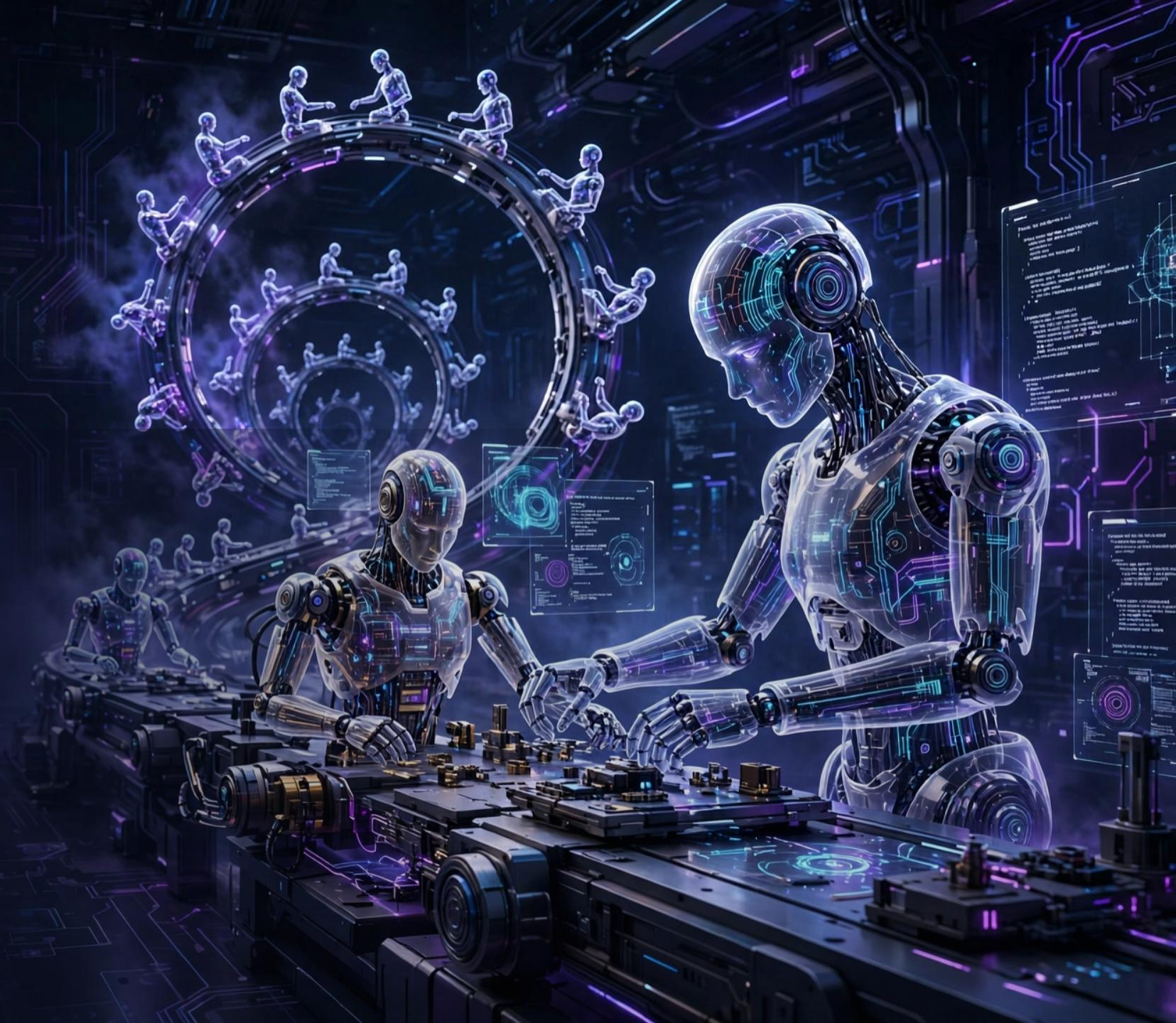


The AI Building Itself



For decades, humans controlled everything

The model we knew

- Formulating hypotheses
- Writing code
- Running experiments
- Interpreting results
- Deciding what to build

BEFORE • FULL HUMAN CONTROL

Hypothesis → Human

Code → Human

Experiment → Human

Evaluation → Human

This model is starting to change →

BASED ON ANTHROPIC INSTITUTE 2026

AI already **takes part in creating new AIs**

"AI does not yet autonomously develop its own successors, but it already speeds up the research and engineering work behind building them."

— Anthropic Institute, *When AI builds itself*

1. NOT FULL RSI

Full autonomy not yet demonstrated

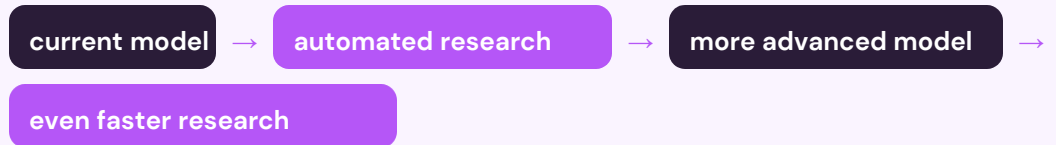
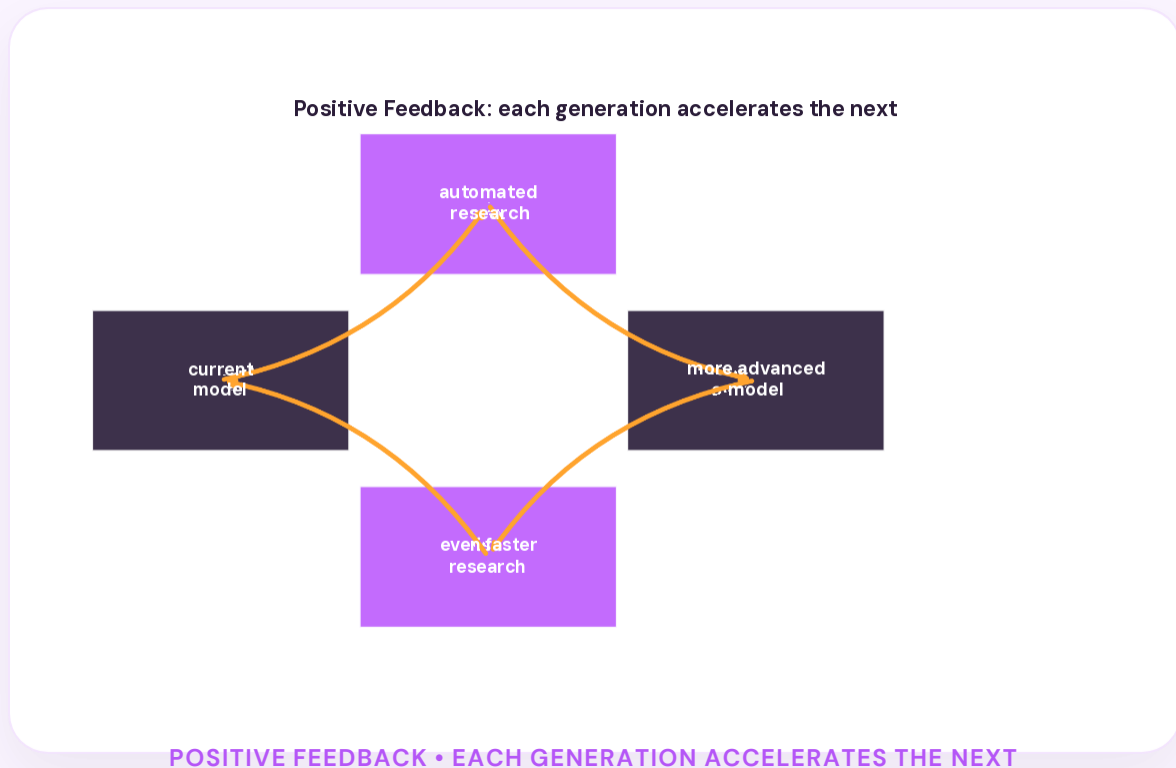
2. BUT INFRA IS BEING BUILT

Technical and organizational pieces being put in place

3. CRITICAL DISTINCTION

Acceleration ≠ inevitable explosion

What is recursive self-improvement?



It is not rewriting the brain overnight. It is progressive automation of:

Code generation and review

Planning and running exps.

Analyzing results

Formulating hypotheses

Choosing lines of investigation

Training and evaluation

Loop closed when: all of these steps can be done with little or no human intervention.

From assistance to autonomy

1

Tool — suggests, human drives

Suggests code snippets, explains problems, assists. The process remains 100% human.

HISTORICAL

2

Agent — runs long tasks, human sets goals

Edits files, uses tools, tests, runs experiments. Observed at frontier orgs.

TODAY • OBSERVED

3

Autonomy — selects problems and designs its successor

Forms hypotheses, evaluates results, chooses lines. Still hypothetical.

HYPOTHETICAL



The differentiator is not execution, it is **scientific and strategic judgment** — choosing what is worth investigating.

How has engineering changed in 2 years?

>80%

of lines of code merged in May 2026
attributed to Claude

~8x

more code merged per day per engineer
vs 2024



Caution: lines of code ≠ perfect productivity. There may be rework and a greater need for review. Anthropic itself acknowledges overestimation.

EVIDENCE • ANTHROPIC • MAY 2026

Scale and speed changed materially

- ✓ Engineers stop writing everything and move to guiding, reviewing, and validating
- ✓ More features and tests, but also larger implementations
- ✓ Operational transformation, not just individual gain

Safe conclusion: AI changed the scale and speed of work — not that each person became exactly 8x more productive.

Agents that work for longer

INDEPENDENT METRIC • METR TIME HORIZONS

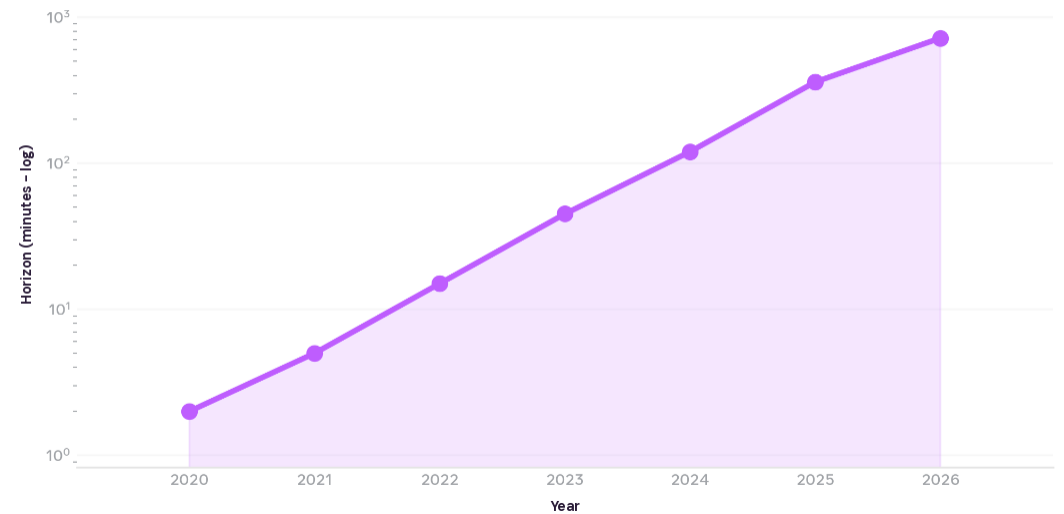
WHAT IS A HORIZON?

The time it would take a human to solve the same task. If it grows from minutes to hours, the agent sustains longer chains of coherent actions.

- **12h horizon ≠ 12h working.** It is a complexity equivalence.
- Current focus: software, ML, and cybersecurity with clear criteria.
- It does not mean any long intellectual task.

Caution: well-defined tasks. Shows progress in execution autonomy, not in scientific judgment.

Task Horizon Growth – METR (illustrative)



METR 2020–2026 • log scale in minutes • growth from minutes to hours

The automation of scientific research

EXPERIMENT • AUTOMATED WEAK-TO-STRONG RESEARCHER

Claude agents were given a weak→strong supervision problem

- 💡 Formulate hypotheses
- ⚙️ Train models
- 🔗 Share findings
- 🧪 Design experiments
- 📊 Analyze results
- 🔄 Revise approaches

Automated research within a prepared, measurable space — not fully autonomous science.

LAB RESULT

Outperformed the human reference team

Within the controlled experimental environment, agents collaborating with each other reached a higher score than the human baseline.

LIMITS – AS IMPORTANT AS THE RESULT

Humans defined the problem, environment, and criteria

- Initial guidance provided by humans
- Results not fully transferred to production
- Clear metric ≠ open-ended discovery

WHAT REMAINS HUMAN

Research taste — scientific discernment

Systems execute **fast once the goal is already defined**. The hard part comes before and after: choosing what to investigate, whether the result is reliable, and whether we should build.

🚩 Knowledge + experience + systems thinking + creativity

THE HARD PART IS NOT EXECUTING — IT IS JUDGING

CHOICE

Which problem deserves investigating?

HYPOTHESIS

Which bet is truly promising?

VALIDITY

Is the result reliable? Does the metric represent the real goal?

LIMIT

When to abandon? What was not considered?

Is discernment structurally human, or just another skill that models have not yet learned?

Key question

We do not have an answer yet — and that is the article's decisive question.

The logical leap **not demonstrated**

✓ WE HAVE EVIDENCE FOR

AI → acceleration of AI R&D

Agents increase the speed of code, experiments, and analysis in prepared environments.

✗ NOT YET DEMONSTRATED

R&D acceleration → full RSI

The leap from accelerating tasks to autonomous design of its successor.

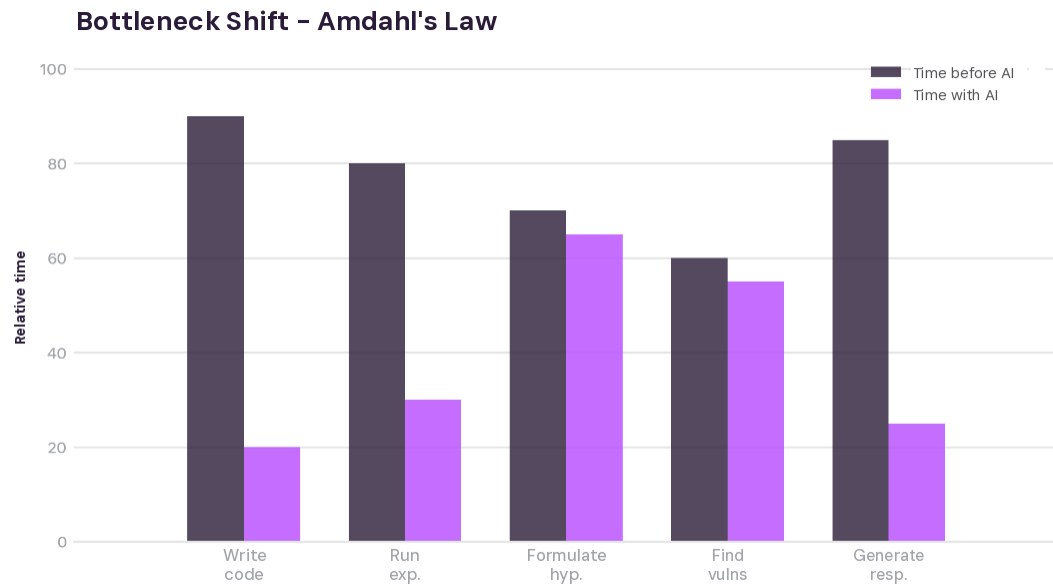
OBSTACLES BETWEEN THESE TWO CONDITIONS

- 💡 Genuine conceptual innovation
- ⚡ Compute, energy, infrastructure
- 🎯 Autonomous definition of relevant goals
- 🧩 Coordination of complex projects
- ✓ Independent validation of results
- 🛡 Security and institutional authorization

“ More precise term today: cumulative acceleration of AI-assisted R&D — not full recursive self-improvement.

When generating outpaces verifying — Amdahl's Law

BOTTLENECK SHIFT • TIME BEFORE VS WITH AI



If an agent writes code **100x faster**, but human review grows only 2x, review is the new bottleneck. The unautomated part limits the whole.

Write



Verify code



Run



Select exps.



Formulate



Identify
relevant
hypotheses

Generate



Validate
reliability



Immediate risk: capacity to produce outpaces institutional capacity to understand and verify.

The paradox of human oversight

The more necessary, the harder to exercise

The more AI executes, the more humans move into oversight. But if they stop programming and investigating directly, they lose the competence to supervise.

-  **Loss of technical expertise** — less hands-on practice
-  **Cognitive dependence** — delegating without verifying
-  **Difficulty identifying errors** that are subtle
-  **Loss of situational awareness** of the system
-  **Reduced informal learning** and collaboration

“

It is not enough to have a **human in the loop**. We need a **human capable of understanding the loop**.

Core risk: the more critical oversight becomes, the lower the real capacity to exercise it may be. Preserving practical skills is as important as keeping people in the process.

Govern the **cycle**, not just the model

When AI takes part in creating its successor, evaluating only the final model is not enough. We need to audit the whole process.



Separation & human direction

- Separate agents that produce vs verify
- Preserve human oversight of goals
- Define clear human accountability



Infrastructure containment

- Isolated environments and minimal permissions
- Limit access to compute and infra
- Effective shutdown mechanisms
- Test generalization outside the lab



Auditability & truth

- Auditable experiment records
- Independent evaluations
- Monitor metric manipulation
- Disclose relevant incidents



The risk is not just hostile AI, but optimizing an incomplete goal at superhuman speed.

Three possible futures

The article considers three trajectories — slowdown, the more likely co-evolution, and full autonomy, not yet demonstrated.

1

S-CURVE

Slowdown

Exponential trends turn into an S. Judgment and deep innovation remain difficult.

- ✓ Agent diffusion already transforms orgs and markets
- ✓ Progress in execution, not in judgment

2

MOST LIKELY • ANTHROPIC

Advanced automation with human direction

Automation advances, humans keep direction and evaluation of results.

- 👥 Small teams coordinate large agent structures
- 🎯 Role shifts from execution to direction and governance

3

PLAUSIBLE, NOT DEMONSTRATED

Full autonomy

Systems develop successors on their own. Pace limited by compute, energy, and verification.

- ⚡ Limited by energy and infra, not just capability
- ⚠️ Oversight becomes a critical bottleneck

💡 Aridio's take: scenario 2 requires anticipatory governance today — **scenario 3 requires verifiable shutdown mechanisms before it becomes relevant**

The urgent question is not **when**, but whether we will be ready

CONCLUSION • 3 SOLID FINDINGS

1

AI already speeds up new AI creation

Code, experiment, and analysis steps are already assisted at scale in frontier labs.

2

Human role shifts to governance

From direct execution to direction, validation, goal-setting, and risk control.

3

Oversight becomes a bottleneck before autonomy

The capacity to generate may outpace the institutional capacity to verify — an immediate risk.

Will we be able to develop institutions capable of understanding, verifying, and **governing this cycle** before it operates faster than our ability to control it?

CENTRAL QUESTION

Thank you.

Shall we continue the conversation?

Aridio Silva | @aridiosilva

Based on Anthropic Institute 2026 • METR • CORE-Bench

#ArtificialIntelligence

#AIAgents

#AIGovernance

#AISafety