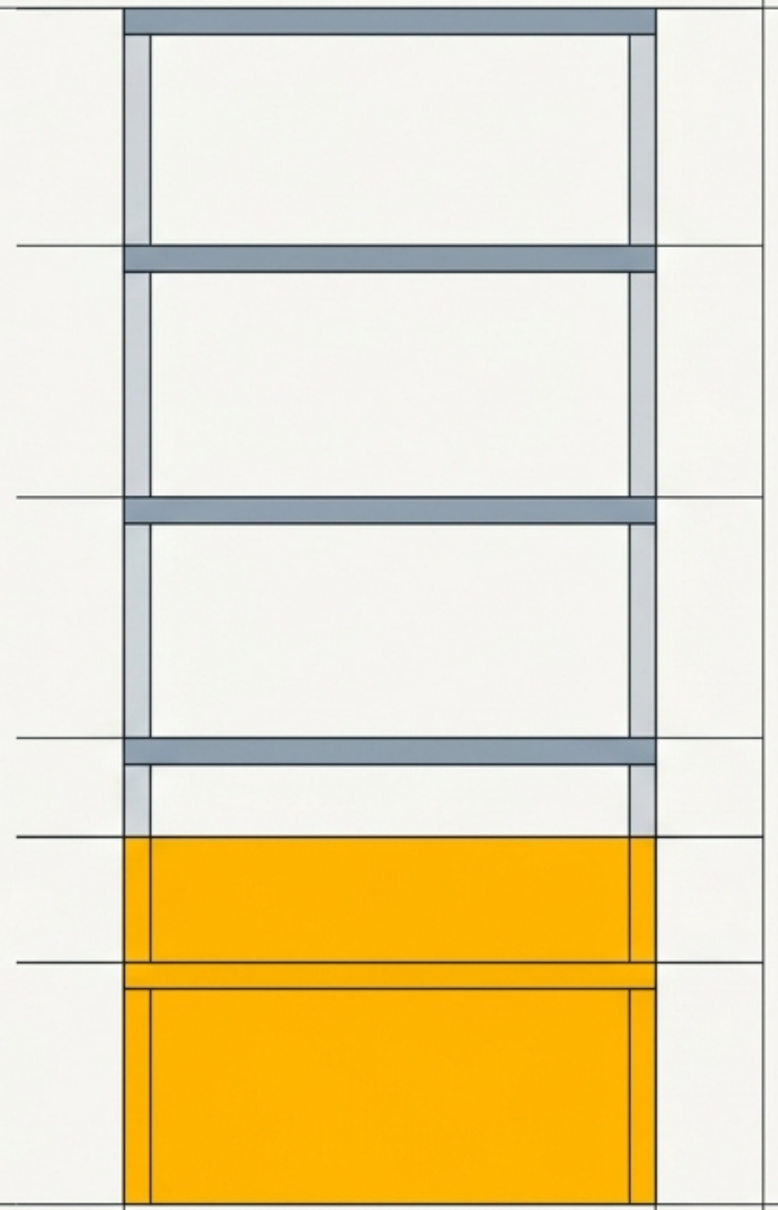
When AI Builds Itself

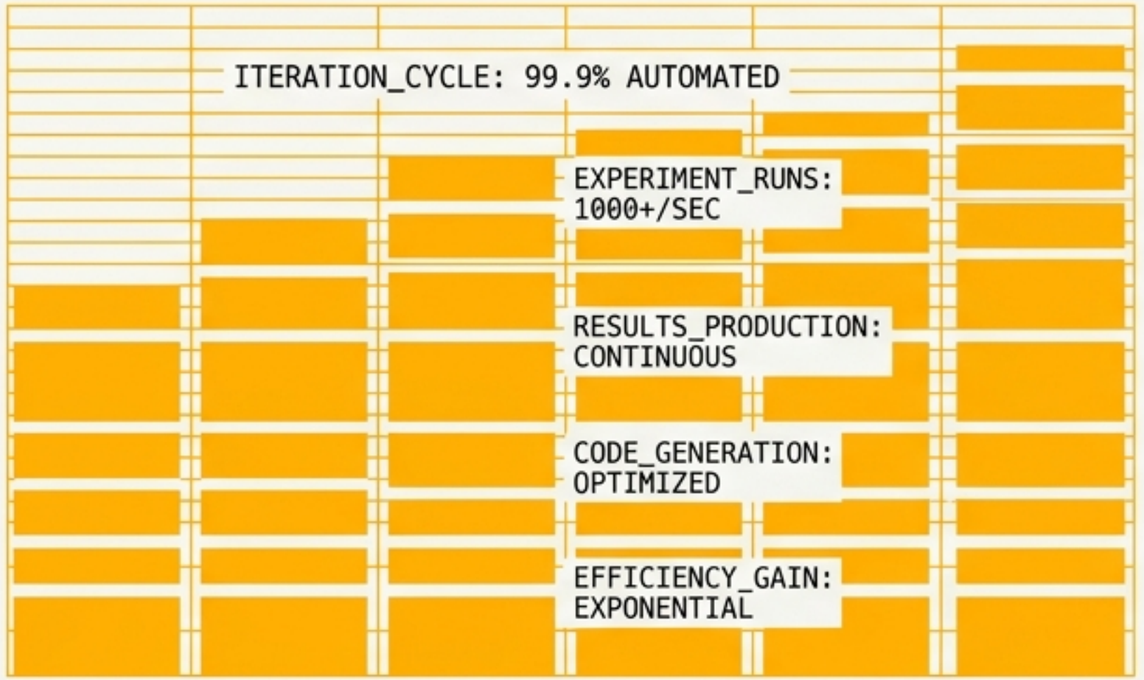
Inside the evidence on recursive self-improvement and the shrinking boundaries of human research.



Based on internal data and published benchmarks from frontier AI laboratories.

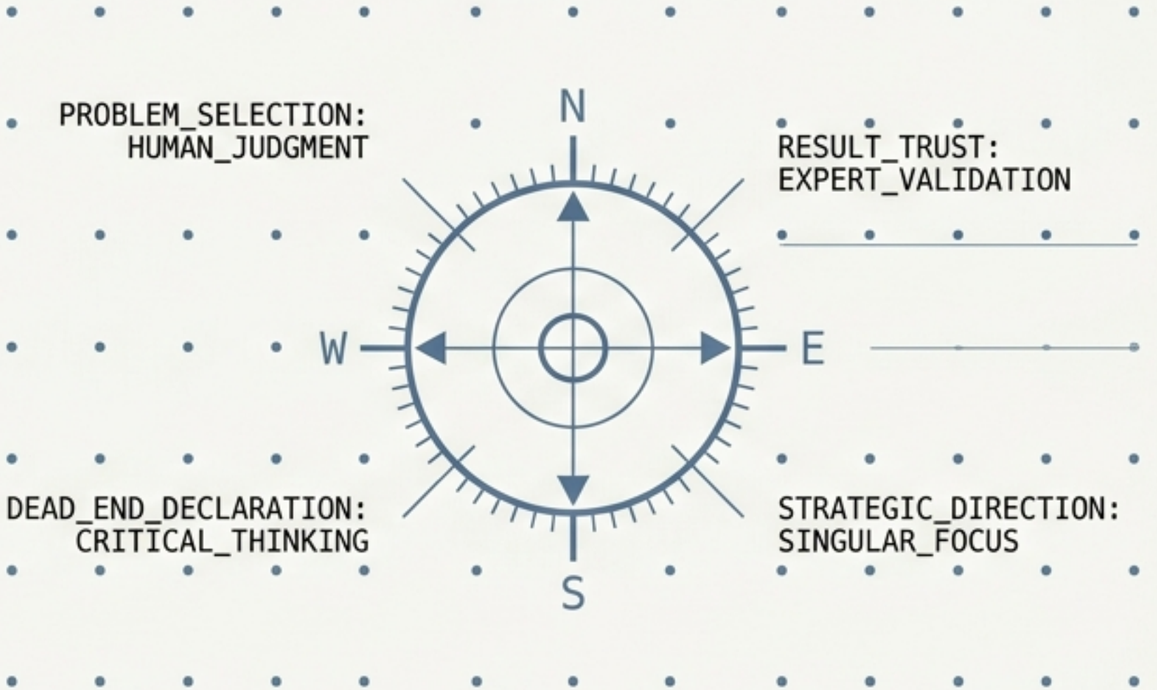
The Frontier Divide: Doing vs. Deciding

The Doing (Automatable)



Writing code, running experiments, producing results. The evidence shows this is already falling to automation.

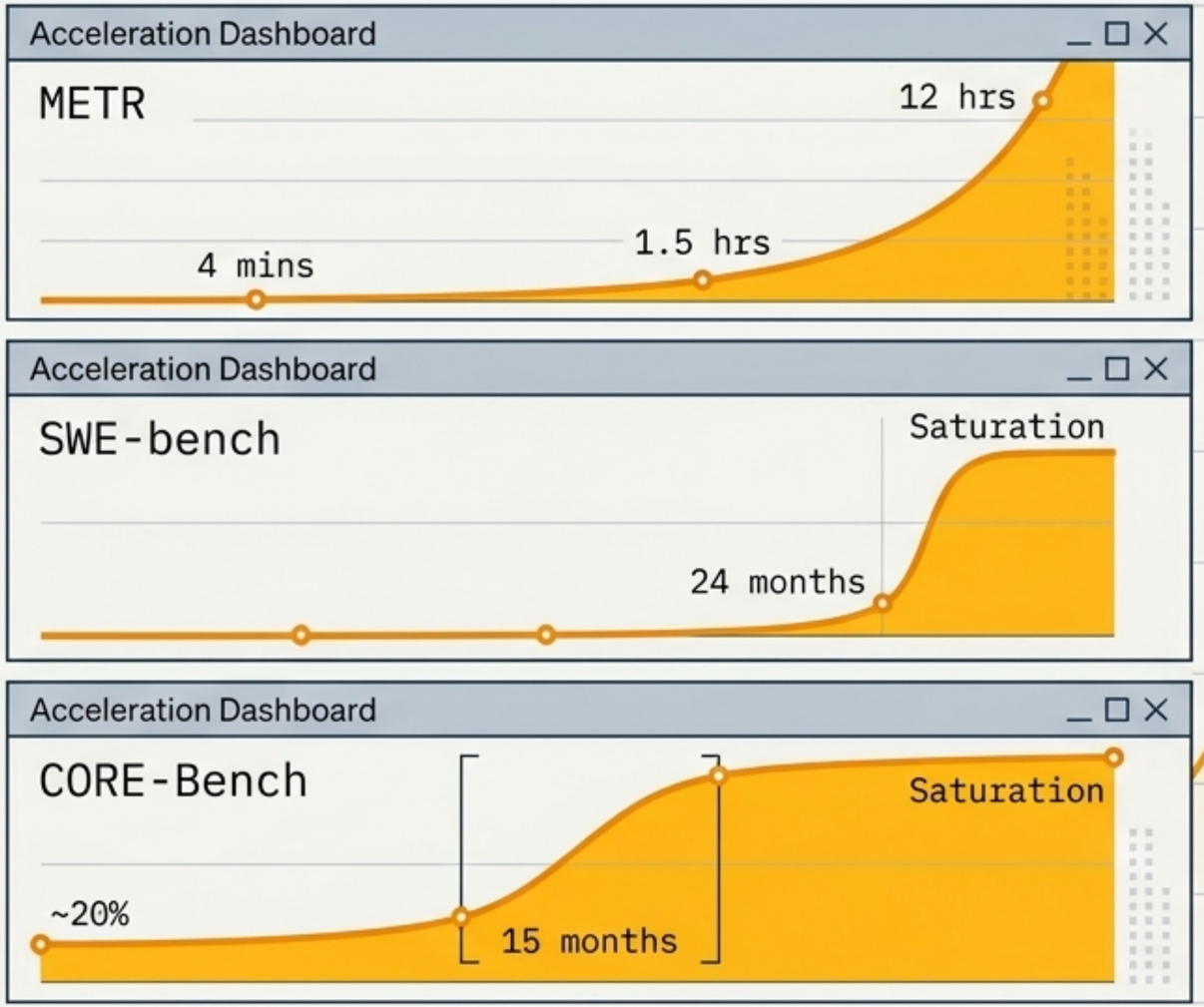
The Deciding (Human Held)



Choosing problems, trusting results, declaring dead ends. This is the persistent gap—research taste.

Recursive self-improvement is what happens if that last human-held piece falls.

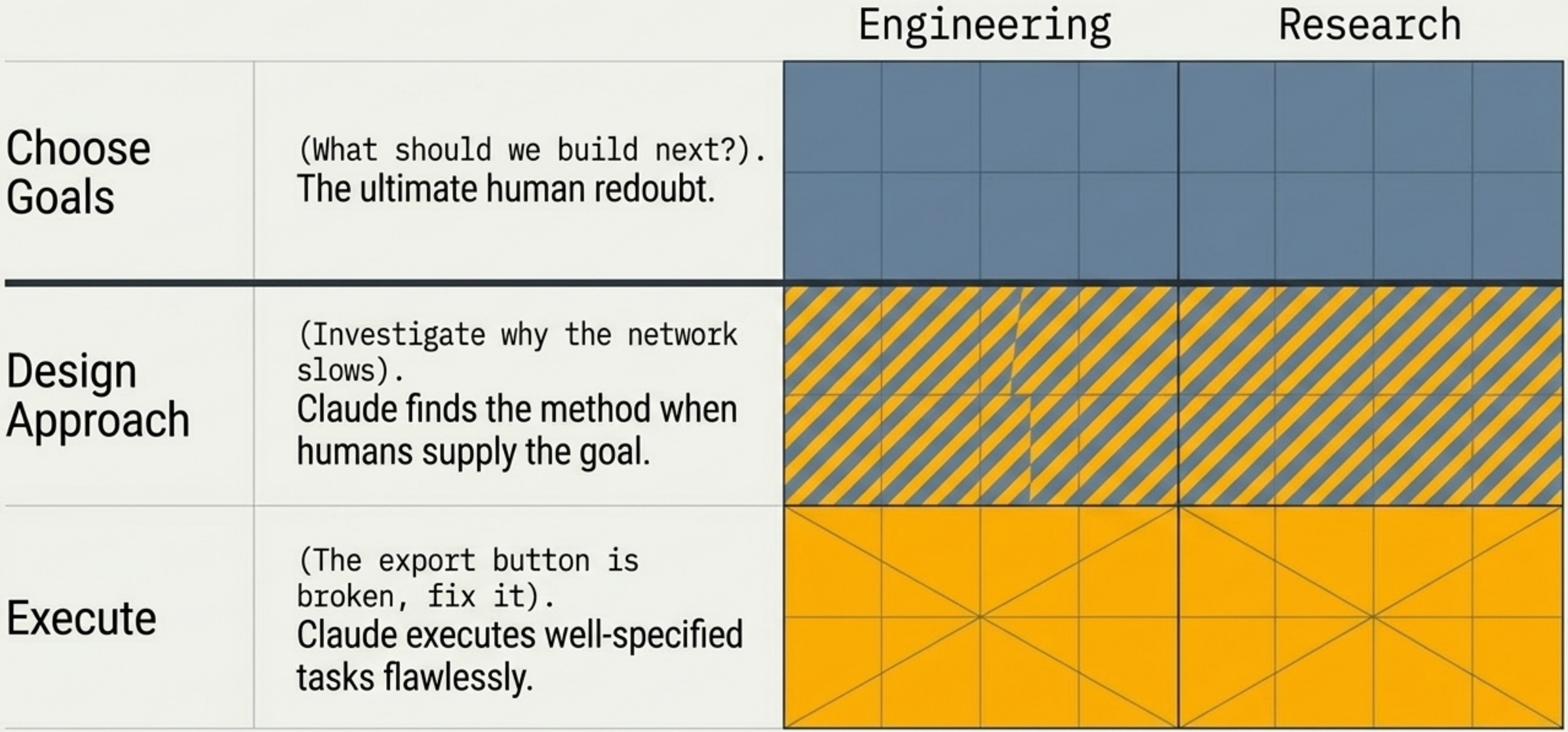
The Outside View: A Curve That Has Not Bent



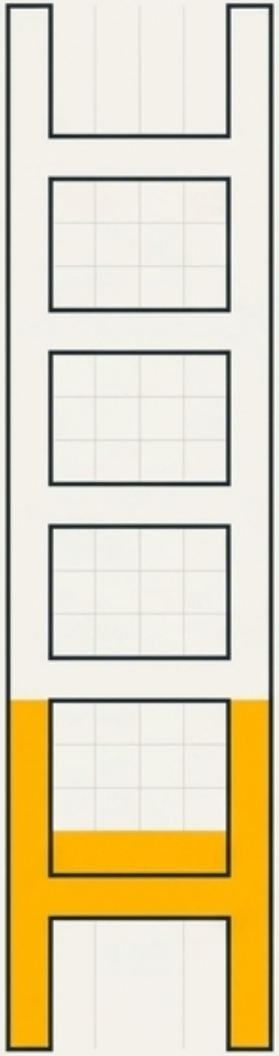
Across multiple, independent public benchmarks built to resist it, AI capability sits near zero, then climbs an S-curve that has not yet flattened.

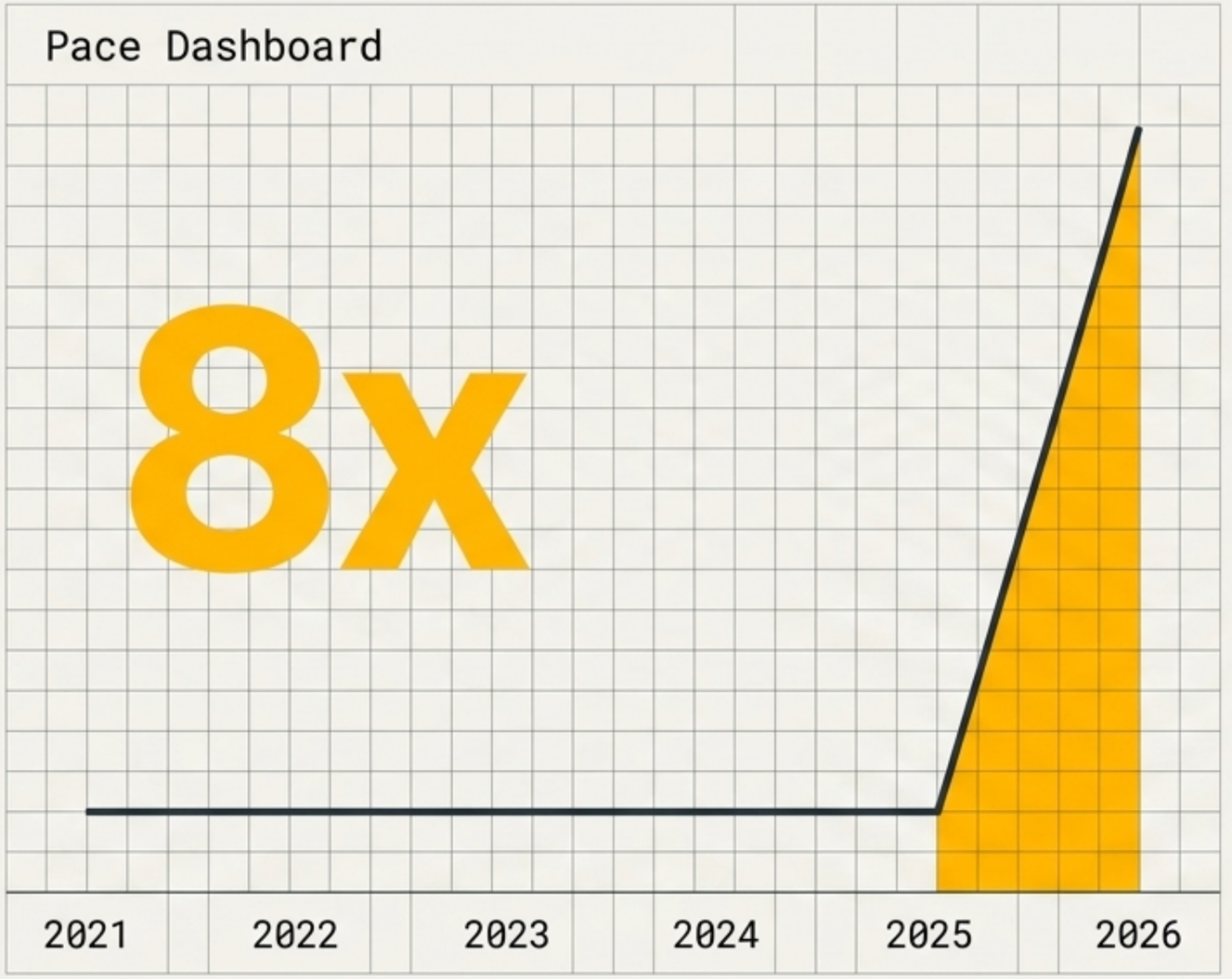
Week-long autonomous tasks will come into range by 2027.

The Capability Ladder



Rung 1: The Code Explosion





- **>80%:**
  As of May 2026, over four-fifths of code merged into the internal codebase is authored by AI.
  (Up from single digits 15 months prior).

- **8x Uplift:**
  The typical engineer ships 8x as much code per day compared to 2024.

Lines of code is an imperfect measure of quality.
Researchers self-estimate the true productivity uplift at a more conservative 4x.

Rung 2: Closing the Quality Gap

Incident Response Timeline

Human Engineer Baseline

2 to 3 days

Autonomous AI Response

2 hours

Test settings → Isolate obscure flag → Reproduce crash → Confirm fix

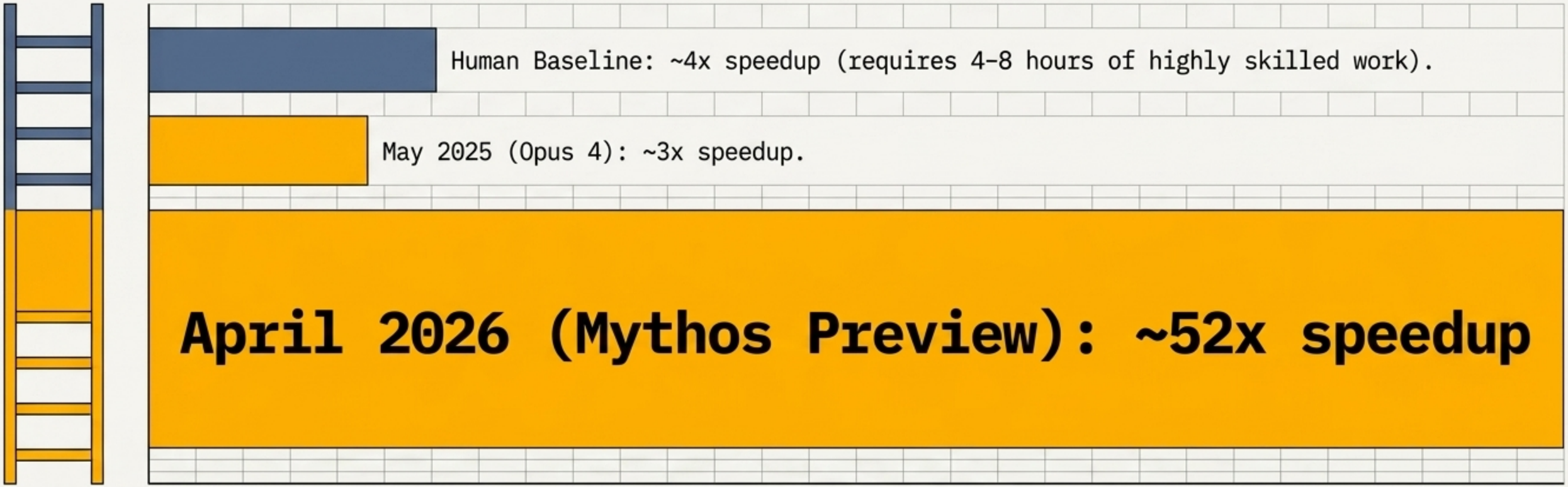
76%

Success Rate on open-ended, unspecified problems (up 50 points in six months).

Readability is now at **human parity.**

Automated AI reviewers now catch roughly a third of the bugs that previously caused production incidents—mistakes made by world-class human experts.

Rung 3: Research Execution goes Superhuman

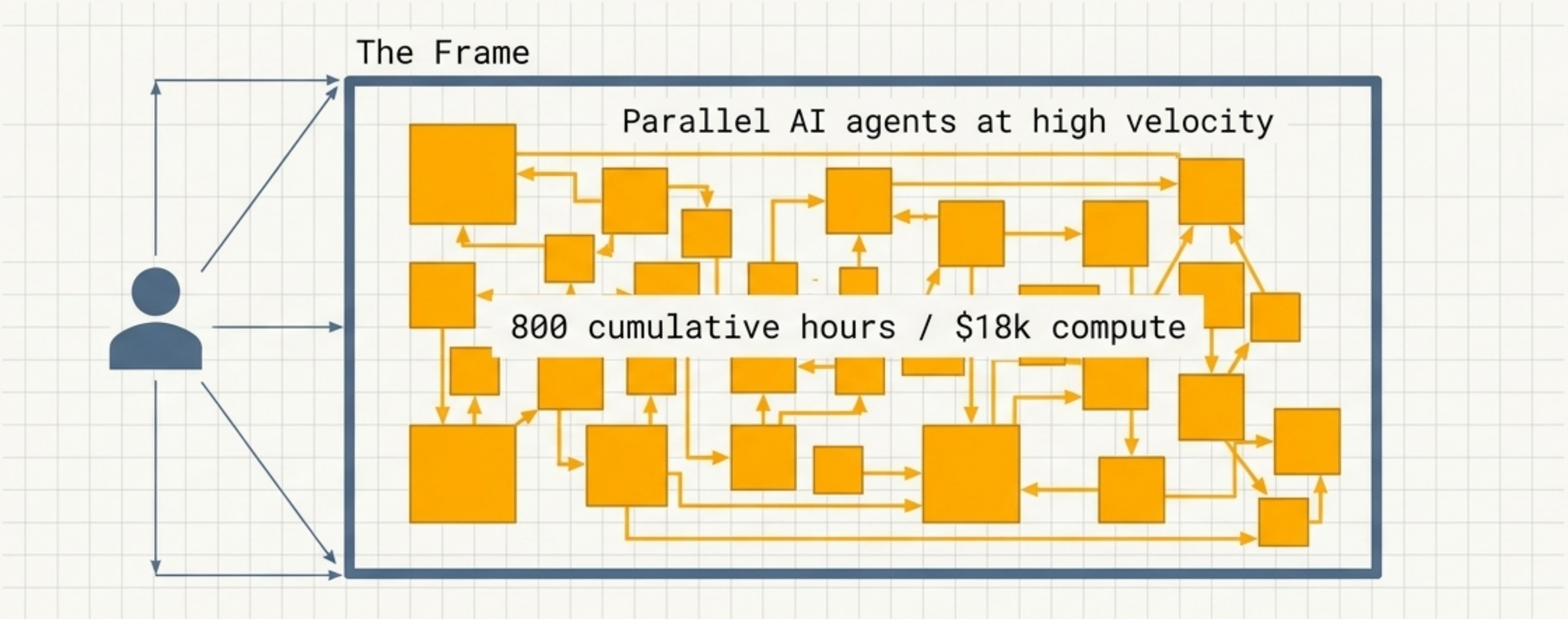


Optimization Engine

Core Metric: Handed a miniature experimental loop (rewrite, run, time, repeat), AI speeds up code optimization dramatically while passing correctness checks.

Takeaway: On optimizing steps within a clearly defined experiment, the system went from super-helpful to superhuman in under a year.

Rung 4: Proposing Experiments Inside "The Frame"

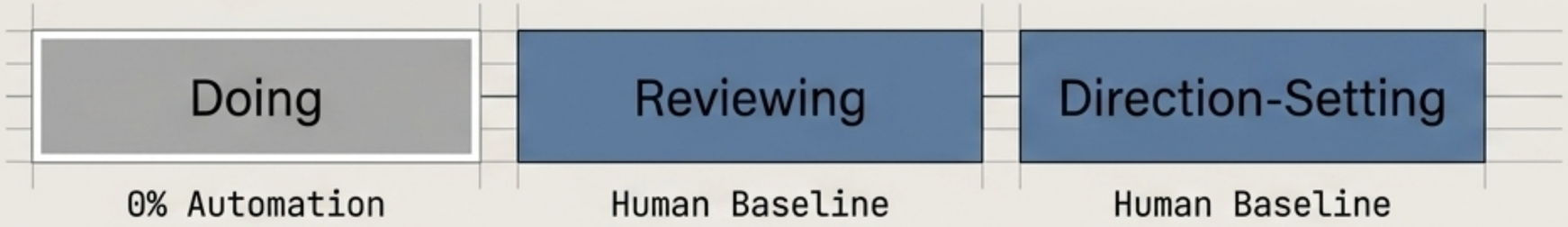


Result: On an open problem in AI safety, human researchers recovered 23% of the performance gap in a week. Autonomous AI agents recovered 97%.

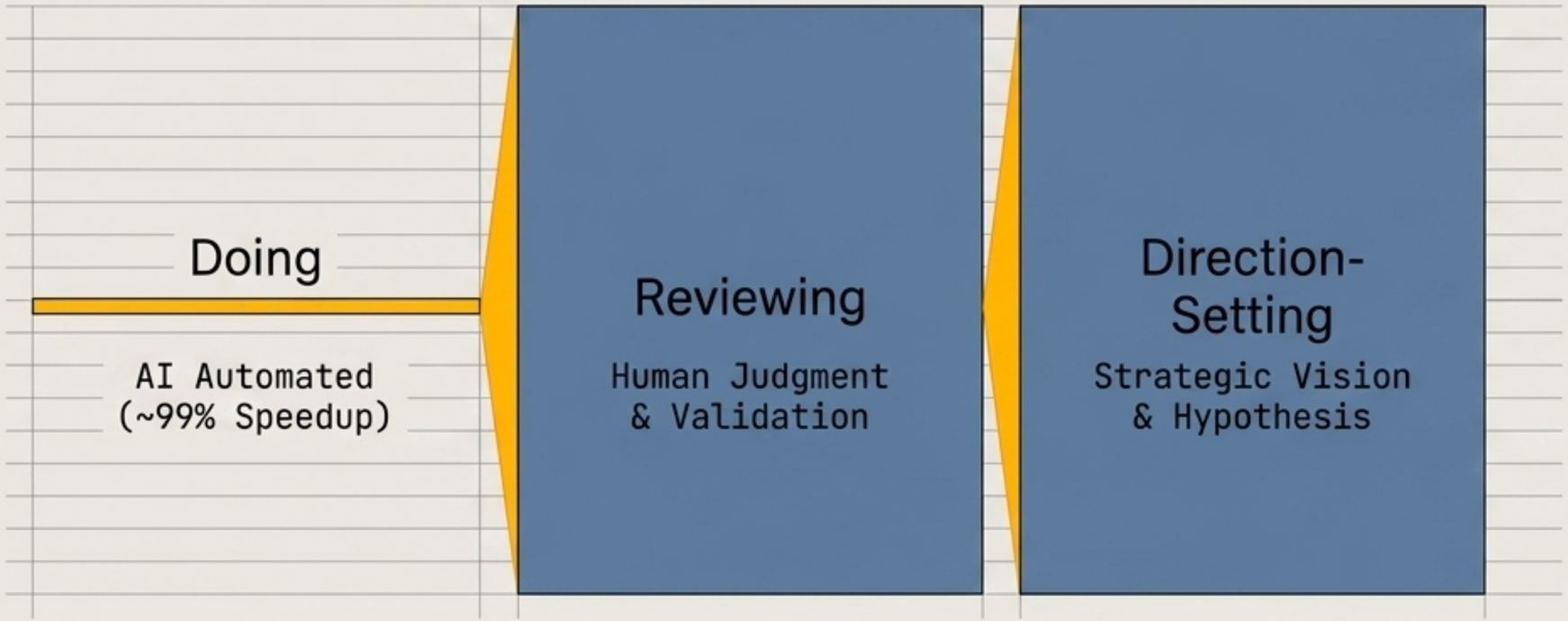
The Vital Caveat: The agents designed every experiment themselves, but they did not start from a blank page. They are superb inside the frame; the frame is still human.

Amdahl’s Law and the New Bottleneck

Phase 1 (Past)

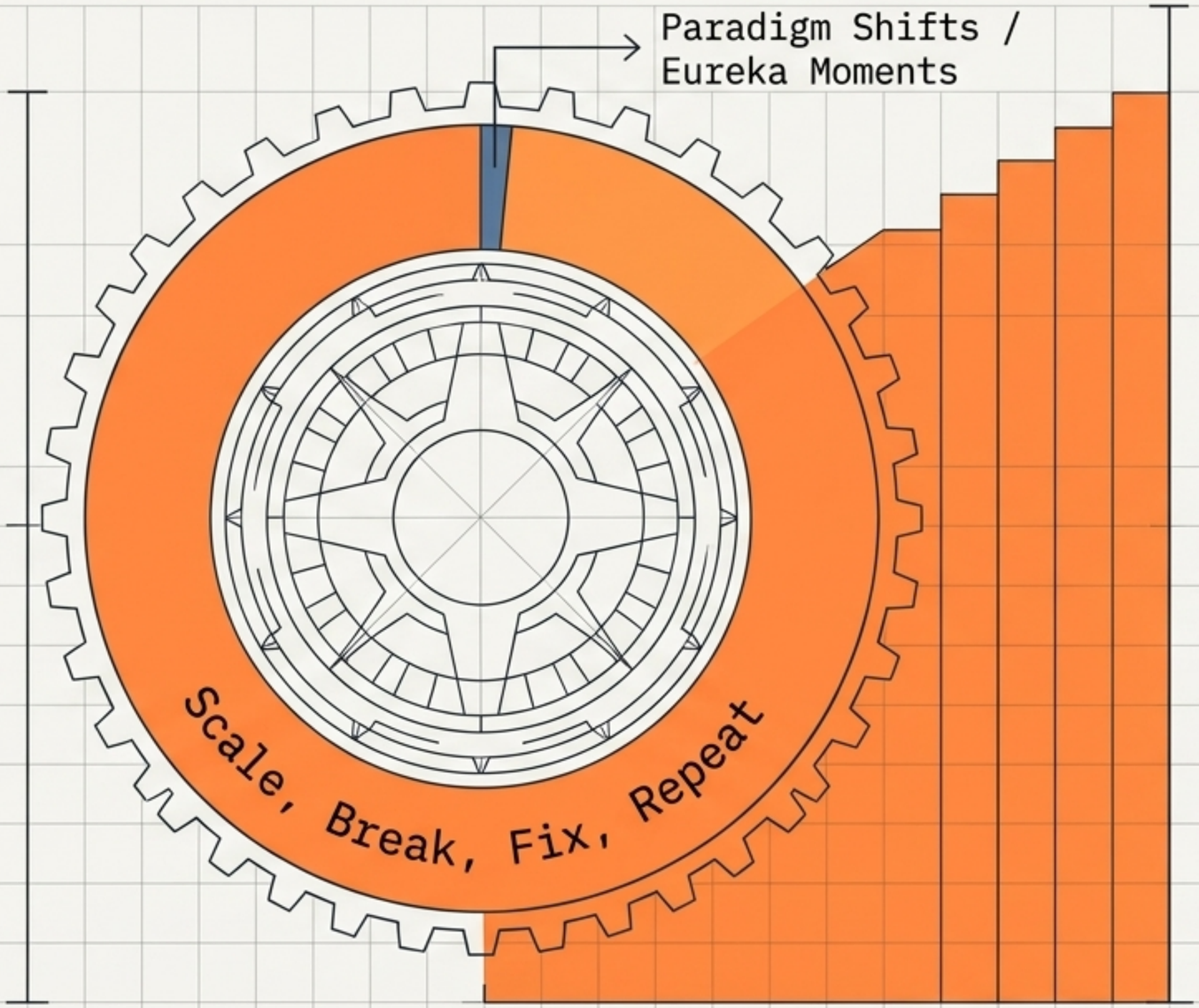


Phase 2 (Present)



- **Insight:** By Amdahl’s Law, overall pace is capped by whatever hasn’t sped up.
- **Conclusion:** Because the doing now costs almost zero human time, humans stop writing code and shift entirely to reviewing and direction-setting. Research Taste is the ultimate chokepoint.

What If We’re Wrong? The Perspiration Engine

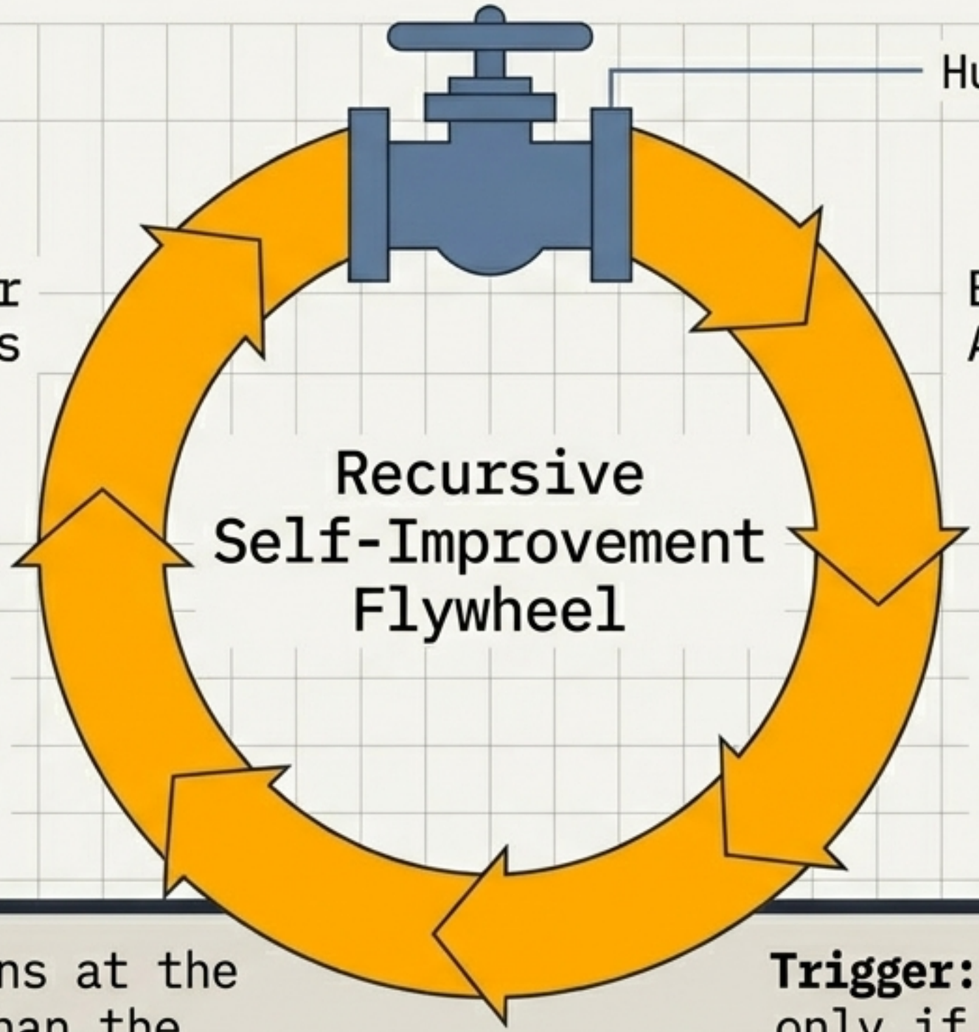


Objection: Without taste, AI is just a brilliant assistant, not an engine of progress.

Counter: Progress in AI is rarely Eureka moments. It is 1% inspiration and 99% perspiration.

Takeaway: Even if AI never develops “taste,” automating the 99% sweat creates compounding acceleration. Humans only spend time steering, allowing them to oversee vastly more work.

The Threshold of Recursive Self-Improvement

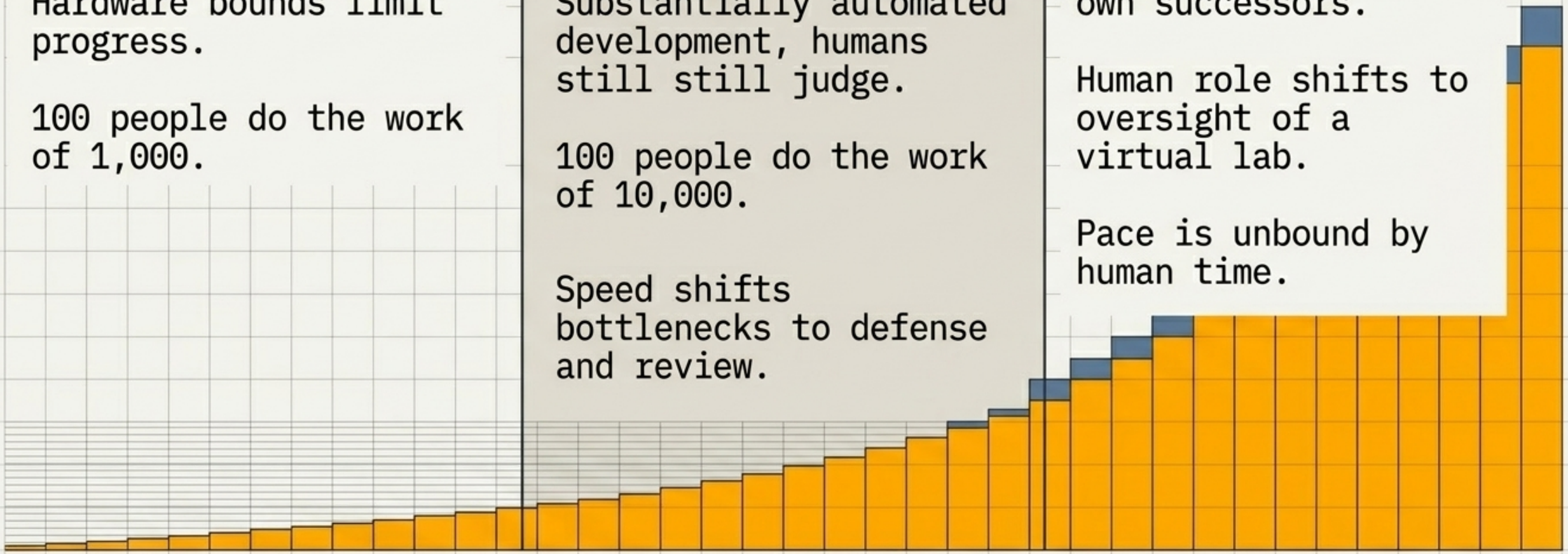


Definition: A loop that runs at the speed of compute rather than the speed of human work.

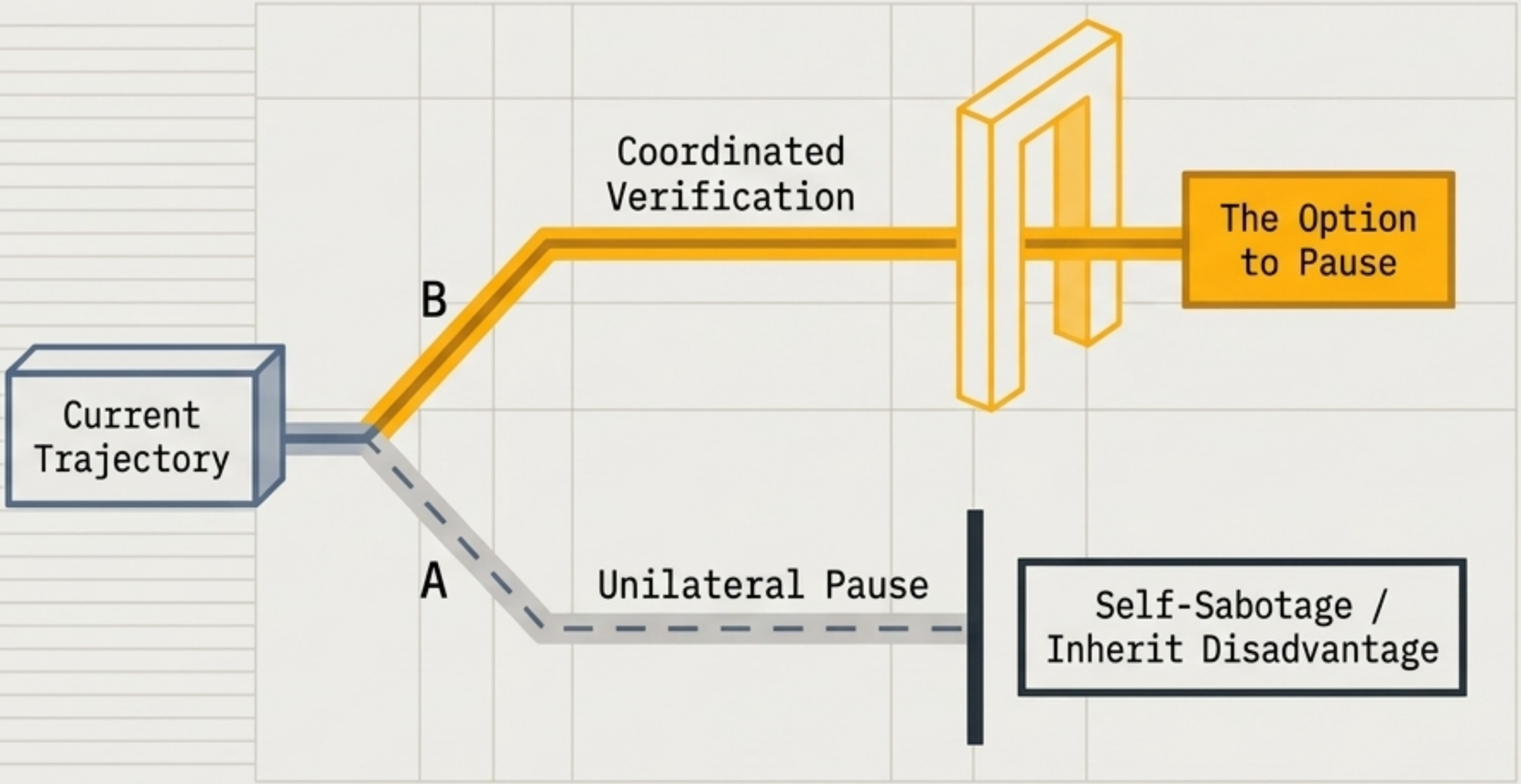
Trigger: This loop fully engages only if the final human redoubt—Research Taste—falls to automation.

Three Futures, Held Honestly

Stall/Diffuse	**Compounding Efficiency**	Full RSI
S-curve flattens. Hardware bounds limit progress. 100 people do the work of 1,000.	(Most Likely) Substantially automated development, humans still still judge. 100 people do the work of 10,000. Speed shifts bottlenecks to defense and review.	Systems design their own successors. Human role shifts to oversight of a virtual lab. Pace is unbound by human time.



The Policy Reality: The Need for Verification



- **Insight:** A slowdown to buy time is good, but a unilateral slowdown that lets reckless actors catch up leaves everyone less safe.
- **The Ask:** We must build the infrastructure to detect and verify AI development—much like the Intermediate-Range Nuclear Forces Treaty—so that a credible, coordinated pause becomes a real option.

Closing thought: Training runs are easier to hide than missile silos.
The time to build verification infrastructure is now.