Forecasting Research Institute

How Close Is Recursive Self-Improvement?

A review of the evidence

Authors: Simas Kučinskas and Josh Rosenberg

VERSION: v1 (2026 NOV)

FORECASTING RESEARCH INSTITUTE

Table of Contents

Summary	03
1. Introduction	03
2. Defining recursive self-improvement	05
3. Evidence	11
4. Existing projections and forecasts	17
5. Key cruxes	23
6. Forecasting questions	25
7. Takeaways	29
8. Further reading	30
Appendix: Full forecasting questions	30

Summary[1]

DATA CUTOFF: NOV 3, 2026

AI companies increasingly state that their models may soon be capable of *recursive self-improvement* (RSI), which could dramatically accelerate AI capability progress. How close is RSI? We review the evidence, existing forecasts, and key disagreements, turn the disagreements into forecasting questions, and give initial forecasts from an ensemble of frontier large language models (LLMs). As of early Nov 2026, AI acts as a highly capable research assistant, but humans still supply most of the research judgment. AI capability progress has not clearly accelerated on public measures, although there is some internal evidence from AI companies indicating a potential acceleration. On a simple extrapolation of current trends, mostly using company-reported metrics, AI could take over most AI R&D tasks by 2027–2028, in line with OpenAI's target of an automated AI researcher by Mar 2028. A key uncertainty is whether these trends continue as the remaining non-automated tasks increasingly require research judgment. To measure acceleration, we propose a forecasting question on the likelihood of a *fast year*: any 12 months in which the frontier of the Epoch Capabilities Index rises by 30 points or more, about twice its pace since Sep 2024. The LLM ensemble forecasts about a 40% chance of a fast year by 2030. If a fast year happens, the ensemble gives about a 70% chance that AI automation of AI R&D caused it.

1. Introduction

How quickly will AI improve between now and 2030?

There are three broad possibilities: AI progress could slow down, continue at its current pace, or accelerate. A particularly important scenario is **acceleration because AI systems help build future AI systems**. The International AI Safety Report 2026 (IASR) notes that "*positive feedback loops – such as AI systems contributing to AI research – could accelerate progress*" and "*decades of progress could happen in years*" (IASR, §1.3). This positive feedback loop is often called **recursive self-improvement (RSI)** (Yudkowsky, 2008): AI systems automate AI R&D and build better successor systems in a recursive loop (see Figure 1).

[1] We thank Nick Merrill for generating LLM forecasts using the AIRO pipeline, and Sivan Livnat for research assistance. For feedback and input at various stages of the project, we thank Jason Abaluck, Zack Devlin-Foltz, Basil Halperin, Ezra Karger, Matt Reynolds, Luis Enrique Urtubey De Cesaris, and Bridget Williams.

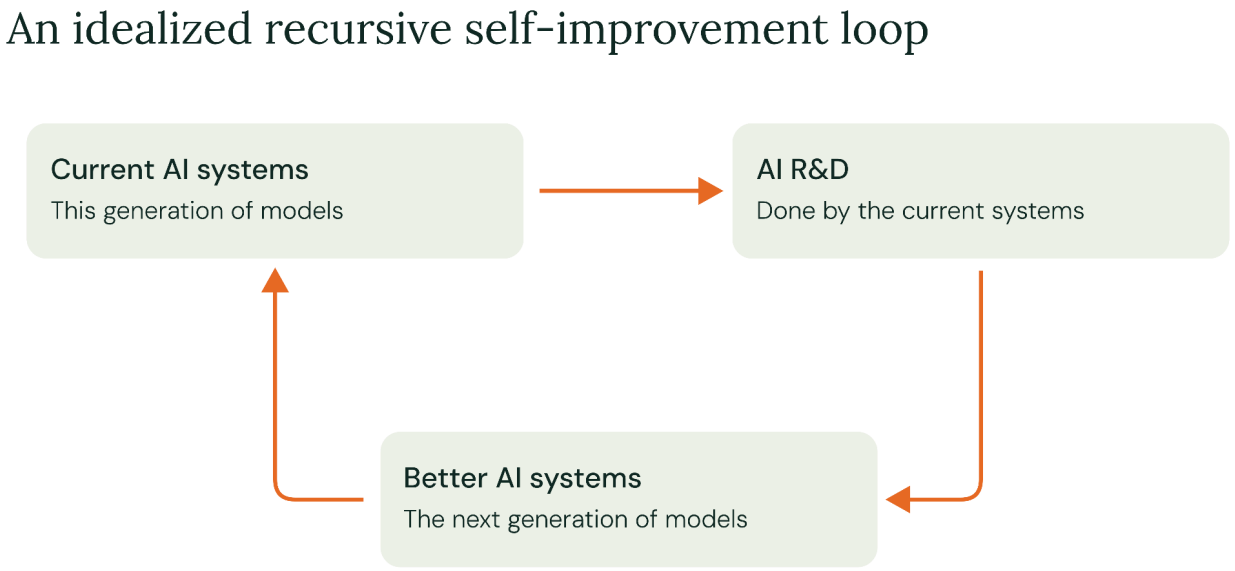


Figure 1: *The RSI loop, in which each generation of AI systems helps develop the next.*

RSI could bring about large benefits as well as substantial risks. A Sep 2026 report by AI researchers including Geoffrey Hinton, Yoshua Bengio, OpenAI chief scientist Jakub Pachocki, and Anthropic cofounder Jack Clark argues that RSI could bring AI's benefits forward by years. However, RSI could also pose extreme risks, as humanity may fail to keep up with the growth in capabilities or lose oversight of advanced AI systems (Chan et al., 2026).

Recent reports review RSI with a focus on its risks and on policy responses (Chan et al., 2026; Sarma et al., 2026). This review is complementary and forward-looking. **We summarize the current evidence and the key disagreements, turn these disagreements into forecasting questions, and give an initial set of forecasts from an ensemble of frontier LLMs.** Simultaneously, we are asking a group of top human forecasters and AI experts to make forecasts on RSI, using this document as background material. We plan to add their forecasts in a later update of this report.

How to use this document:

- Section 2 reviews how the labs and the literature define RSI and related terms. **Section 2.3 sets out our operationalization.**
- Section 3 reviews the evidence. **Section 3.1 summarizes what the AI labs report about AI R&D inside the labs.** Section 3.2 covers public evidence: AI R&D benchmarks and an open-world evaluation. Section 3.3 covers qualitative evidence from researcher interviews.
- **Section 4 collects existing projections and forecasts**, including lab targets, trend extrapolation, prediction platforms, expert forecasts, and notable individual predictions.
- Section 5 sets out the key cruxes.
- **Section 6 summarizes the forecasting questions we are asking a group of forecasters and provides initial LLM forecasts.**

- **Section 7 summarizes the main takeaways**, and Section 8 lists further reading. The Appendix gives the full forecasting questions.

If you are short of time, we recommend only reading the sections in **bold**.

2. Defining recursive self-improvement

Defining RSI and automated AI R&D is challenging. RSI, in particular, "*has been used in several ways, often inconsistently*" (Cunningham et al., 2026). We first review how the labs (Section 2.1) and the literature (Section 2.2) define these terms. Section 2.3 sets out our own operationalization.

2.1 Lab definitions

OpenAI and Anthropic define RSI and automated AI R&D in two ways: (i) AI that can substitute for their research staff ("substitution"); or (ii) AI that substantially accelerates the pace of AI progress ("acceleration"). As of early Nov 2026, **neither lab reports having crossed its threshold for automated AI R&D**.

Test	OpenAI (Preparedness Framework v2)	Anthropic (Responsible Scaling Policy, RSP v3.4)
Substitution	Critical: "*a superhuman research-scientist agent.*" High: a "*highly performant mid-career research engineer assistant, relative to those researchers' 2024 baseline*"	Models could "*fully substitute for our entire set of Research Scientists and Research Engineers, at competitive costs*"
Acceleration	Critical: "*generational model improvement (e.g., from OpenAI o1 to OpenAI o3) in 1/5th the wall-clock time of equivalent progress in 2024 [...] sustainably for several months*"	Dramatic acceleration of capability progress, defined as "*(a) [...] double the rate of progress in AI aggregate capabilities [...] and (b) [...] doubling is substantially attributable to the automation of research and/or engineering [...]*"
Status	GPT-6 Astra "*does not reach our High threshold*" (system card); GPT-6.1 Terra also "*does not reach our High threshold*" (addendum, p. 33)	Risk "*Low*"; R&D accelerated "*but not yet by a factor of 2*" (Aug 2026 Risk Report)

Table 1: *OpenAI and Anthropic tests for automated AI R&D, and the status of each lab as of Nov 3, 2026.*

For other AI companies, METR provides a useful collection of automated AI R&D definitions (pp. 19–21).

We note two key points:

- The OpenAI and Anthropic definitions are similar but not identical. On acceleration, OpenAI's "Critical" level is a 5× acceleration of capability progress, while Anthropic uses a lower 2× threshold.

- The labs define and track these thresholds as part of their safety policies. As a result, reaching a particular threshold has direct consequences. For OpenAI, a model at the High level requires "*security controls meeting High standard*"; at the Critical level, the response is halting further development until "*we have specified safeguards and security controls that would meet a Critical standard*" (Preparedness Framework v2, Table 1). For Anthropic, the 2025 policy required AI Safety Level 4 (ASL-4) safeguards for models that could dramatically accelerate AI progress (RSP v2.2, Section 4.3.1). Its 2026 policy no longer requires ASL-4 but its industry-wide recommendations for automated AI R&D "*would likely mean security roughly in line with RAND SL4*" (RSP v3.4, p. 8), i.e., security that "*can likely thwart most standard operations by leading cyber-capable institutions*" (RAND, p. 22).

2.2 Definitions in the literature

The basic idea of recursive self-improvement goes back to at least I. J. Good (Good, 1965). In 1965, Good described an *intelligence explosion* in which machines design better machines, famously stating that "*the first ultraintelligent machine is the last invention that man need ever make.*"

The table below defines the four main terms used in the literature, mostly based on the taxonomy developed by Cunningham et al. (2026).

Term	Definition
Automated AI R&D	AI can "*autonomously make technological improvements at least equivalent to those made by human researchers, at an equal cost*" (Cunningham et al., 2026, Table 1, "R&D automatability"). The labs' substitution tests (Section 2.1) are versions of this definition.
Recursive self-improvement (RSI)	"*an AI system capable of fully autonomously designing and developing its own successor*" (Favaro and Clark, 2026).
Self-sustaining acceleration	"*When AI systems are sufficient for accelerating progress in AI capabilities without any growth in exogenous inputs (human labor, training compute, etc.)*" (Cunningham et al., 2026, Table 1). A *software intelligence explosion* (SIE) is very similar: "*AI-driven software improvements alone are sufficient for rapid and accelerating AI progress*" (Davidson, Hadshar, and MacAskill, 2025; see also Eth and Davidson, 2025). The acceleration can be **narrow** (e.g., confined to AI R&D or easily verifiable tasks), or **broad** (i.e., spreading to capabilities in most other sciences and most economic activities). Relatedly, Davidson, Hadshar, and MacAskill (2025) define a *full-stack intelligence explosion*, in which AI drives improvements "*in all of software, chip technology and chip production.*"
Intelligence explosion	"*an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion'*" (Good, 1965); "*AI capabilities go to infinity in finite time*" (Cunningham et al., 2026, Table 1)

Table 2: *Main terms in the RSI literature.*

We make a few observations:

- Automated AI R&D and RSI are similar, and people often use the two terms interchangeably. For example, OpenAI equates them in its Preparedness Framework (Table 1). However, we think that automated AI R&D is weaker than RSI, as defined above. In particular, automated AI R&D can be *partial*: AI could do most but not all of the AI R&D work.
- Informal discussions often conflate these terms. We note two distinctions:
    - A weaker feedback loop does not necessarily imply a stronger feedback loop. For example, automated AI R&D could arrive without accelerating capabilities if compute or other scarce inputs bottleneck progress (Cunningham et al., 2026). Conversely, full automation of AI R&D is not necessary for acceleration. In some standard growth models, a singularity (infinite output in finite time) can occur without full AI R&D automation (Aghion, Jones, and Jones, 2019, p. 256; Jones, 2026).
    - *Fast* progress does not necessarily imply *broad* progress. As the diagram below illustrates, in principle, fast progress could stay narrow, and broad progress could be gradual. Whether a capability acceleration stays narrow or becomes broad is a key forecasting crux (Section 5).

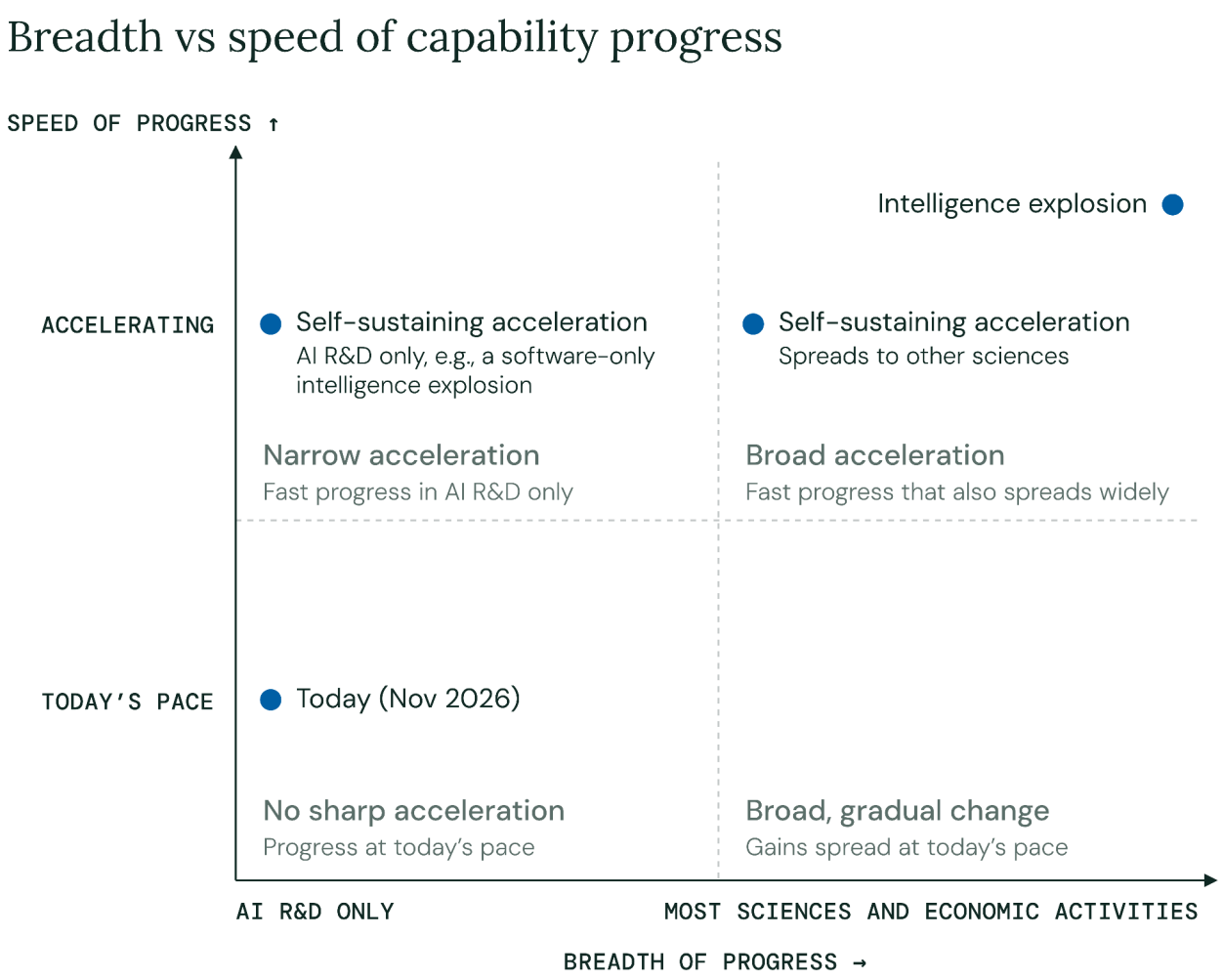


Figure 2: *Breadth vs. speed of capability progress.*

2.3 Measuring capability progress

Ultimately, we are most interested in whether automated AI R&D increases the pace of capability progress. To measure this pace, we use the Epoch Capabilities Index (ECI). We define capability acceleration as a ***fast year*: a rise of the ECI frontier by 30+ points within any 12-month window** (about 2× the pace since Sep 2024). Here, the ECI frontier is the highest ECI score of any publicly available model by a given date. A fast year need not take 12 months to observe: a fast-year question resolves Yes as soon as the ECI frontier has risen by the threshold within 12 months or less.

In our internal forecasting experiments, the fast year worked well as a forecasting question. First, it is simple and easy to understand. Second, there was substantial internal disagreement about whether a fast year is likely, and this disagreement appeared to result from different expectations about automated AI R&D.

That said, **a fast year is not the same as RSI** (see the downsides below). For this reason, we also ask about automated AI R&D directly. For example, our forecasting questions ask how much AI R&D work AI will do with no human in the loop (AL5, the top level of Anthropic's AI autonomy scale; Section 3.1). The survey also asks how human reviewers will score research papers that AI agents write on their own (CRUX 2, an open-world evaluation of AI research agents; Section 3.2). Other questions link automated AI R&D to acceleration. Section 6 summarizes all 18 questions.

ECI data. The chart below shows the ECI frontier since Mar 2023. Epoch estimates that the ECI frontier rose by about 6 points a year before reasoning models and by about 14 points a year after they arrived in Sep 2024 (Epoch).[2] For a new frontier model on Jan 1, 2027, Epoch's 90% prediction interval is 166–176 (Epoch). To extrapolate the trend beyond Jan 2027, we fit a statistical model (local linear trend model); the model allows the pace of progress to vary over time, and it includes the acceleration observed with reasoning models in its sample.[3] Its 80% interval for a new frontier model is 176–206 at the end of 2027 and 171–317 at the end of 2030. Using 20,000 simulations from this model, the probability of a fast year is 17% by the end of 2027 and 44% by the end of 2030.

In the model, the pace of progress changes each month by a random shock. Over 12 months, these shocks change the annual pace by about 10 points a year (i.e., one standard deviation). However, the data do not pin down this value precisely, and values from 2 to 20 points a year fit the data about as well. Across this range, the probability of a fast year is 2–31% by the end of 2027 and 7–58% by the end of 2030.

---

[2] Epoch's estimate dates from Sep 1, 2026. Using Epoch's data as of Nov 3, 2026, a linear fit to the 14 frontier models since Sep 2024 gives about 16 points a year (our calculation).

[3] In the local linear trend model, the ECI score can go up or down, and the pace of progress (i.e., the slope) can become negative. However, in practice, the ECI frontier cannot fall. For the fast-year probabilities, we use the highest simulated score so far, so the frontier cannot fall. The prediction intervals, however, use the simulated scores directly, so they are likely somewhat off; we plan to fix this issue in a future version.

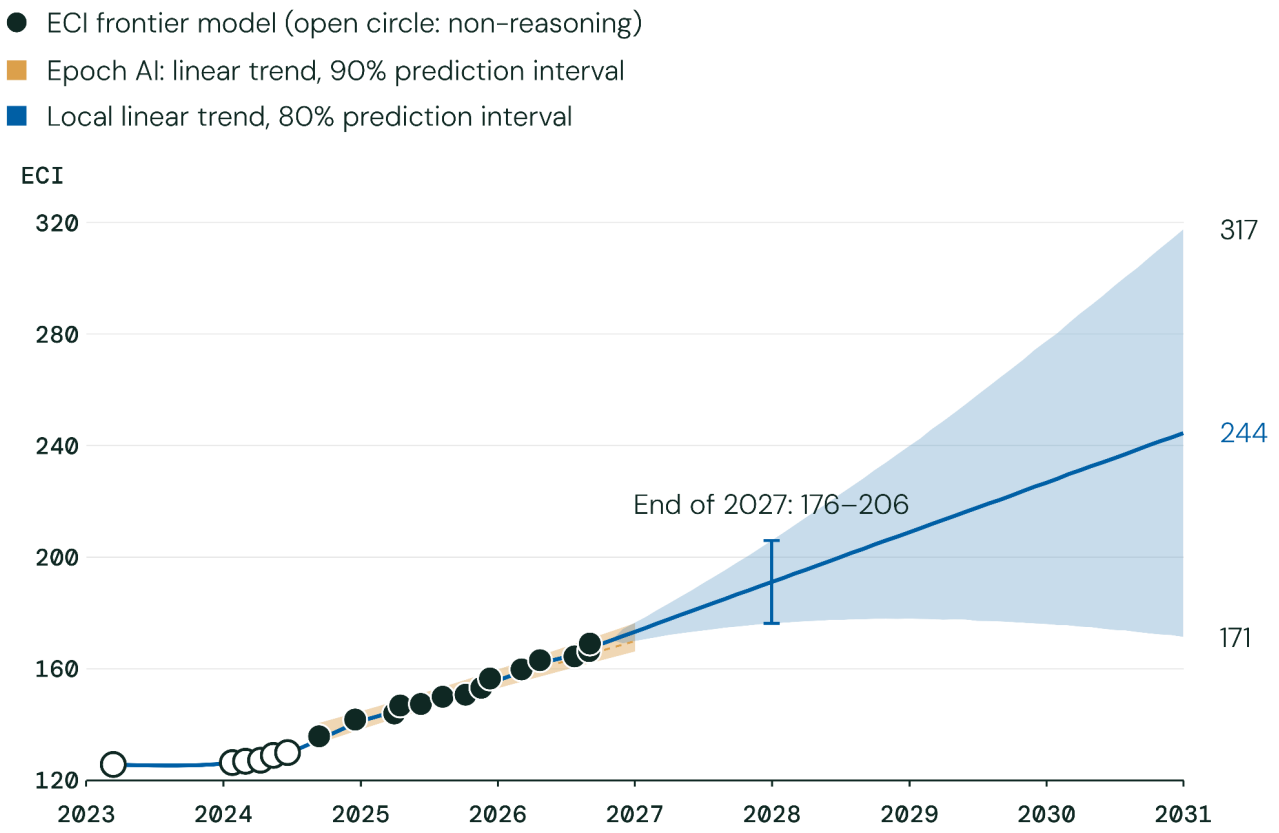


Figure 3: *The ECI frontier since Mar 2023, with two trend projections. The points are ECI frontier models, with scores from Epoch AI (read Nov 3, 2026). The tan band is Epoch AI's trend and 90% prediction interval, from its Sep 1, 2026, post. The blue line is our local linear trend model; the blue band is its 80% prediction interval.*

We note a few **pros and cons of the fast-year definition**:

- Our definition of a fast year roughly matches the labs' acceleration tests. Since the first reasoning models arrived in Sep 2024, the ECI frontier has risen by about 14 points a year (Epoch). So 30+ points in 12 months roughly doubles the current pace, similar to Anthropic's 2× test. For reference, the ECI frontier rose by about 18 points in the last 12 months, from 151 (GPT-5 Pro, Oct 2025) to 169 (GPT-6 Astra, Sep 2026). The largest rise in any 12-month window so far is 21 points (Dec 2024–Dec 2025, o1-mini to GPT-5.2 Pro); see Epoch's ECI data.
- A **fast year could happen without RSI**. For instance, a new human-discovered innovation, like reasoning models in 2024, could cause a fast year, although a doubling from today's pace would mean a larger absolute jump. Conversely, RSI could happen without a fast year if bottlenecks such as compute cap AI progress. To address this concern, the survey also asks about automated AI R&D directly.
- Cunningham et al. (2026) derive a condition for self-sustaining acceleration in terms of elasticities that link AI capabilities, research effort, and algorithmic progress. Their initial calibration suggests that the feedback loop is not yet strong enough for a self-sustaining

acceleration. A related, simpler test is whether the slope of the ECI trend increases. In the setup of Cunningham et al., a self-sustaining acceleration would bend the ECI trend upward even with fixed inputs. In practice, inputs are not fixed but still growing fast (see Section 3.1). Growing inputs would make the bend even stronger, so the roughly linear ECI trend since Sep 2024 suggests no self-sustaining acceleration today.[4]

- Although ECI has no maximum in principle, it is capped in practice for a fixed set of benchmarks. While Epoch notes that "*there is no maximum achievable ECI*" (Epoch), many of the individual benchmarks in the ECI can saturate. GPT-6 Astra, which sets the current ECI frontier (169, ECI), is scored on 24 benchmarks. A model that scores 90% on all of them would have an ECI of about 189 (our calculation), i.e., 20 points above today's ECI frontier, or about 1.4 years at the current pace of 14 points a year.[5, 6] In our forecasting questions, we assume that Epoch keeps adding unsaturated benchmarks, as it has so far.
- The ECI frontier counts only publicly available models. A lab could accelerate internally and release its models later or, potentially, not at all. Our survey asks forecasters about this gap between internal and public models. **Our best current estimate of the internal lead is about 3–4 months of progress at today's pace.** Both methods below convert Anthropic's internal ECI scores (AECI) to Epoch's scale.
    - **Offset:** Both indices (AECI and ECI) set Claude 3.5 Sonnet at 130 and GPT-5 at 150, but Anthropic's uses internal benchmarks (Opus 5.5 system card, pp. 38, 40). On the same internal AECI fit, Claude Opus 5 scores 165.2 and Claude Opus 5.5 scores 169.4 (Opus 5.5 system card, p. 39). On ECI, these models score 164.5 and 168.4, respectively (Epoch data), implying gaps of 0.7 and 1.0. We use their average to get an offset of roughly 0.9 points.[7]
    - **Mythos Preview:** first used internally on Feb 24, 2026 (system card, p. 13). Its AECI of 161.9 (Opus 5.5 system card, p. 39) minus 0.9 gives 161.0 on ECI, against 156.5 for GPT-5.2 Pro, which set the ECI frontier on that date (Epoch data). Lead: 4.5 points, or about four months.
    - **Model 2:** an unreleased Anthropic model "*somewhat more capable than Mythos 5*" that "*does not display a capability jump of the degree observed from Claude Opus 4.6 to Mythos Preview*" (Risk Report, p. 16). It has no score in the Opus 5.5 system card, so we bound it with scores from that fit: at least Mythos 5's AECI of 164.5, and at most 164.5 + 5.9 = 170.4, where 5.9 is Anthropic's estimate of the AECI jump at Mythos Preview (Opus 5.5 system card, pp. 39, 41). Subtracting the 0.9 offset, that implies 163.6–169.5 on ECI. Anthropic states that Model 2 was in heavy internal use by Jul 15, 2026 (Risk Report, pp. 10, 14), so we compare against the ECI frontier on

[4] Anthropic and Google DeepMind perform similar tests. They extrapolate historical trends in internal capability indices and compare new models against them (see Section 3.1). Both find their newest models above the historical trend line, but neither interprets this as a sustained acceleration.

[5] We use 90% because scores near 100% are less informative about capability (Epoch) and may exaggerate how far the ECI is from saturation under the current set of benchmarks. In particular, Epoch itself requires at least four scores between 10% and 90% to include a model in the ECI (Epoch, repo).

[6] On Nov 4, 2026, after our data cutoff, Epoch AI updated the ECI to use a Bayesian model (Epoch). We use ECI data as of Nov 3, 2026, so the values on Epoch's website may differ slightly from the ones cited in this report.

[7] The offset holds only within this fit, so we use only scores from it below.

that date: 163.1 from GPT-5.5 Pro (Epoch data). Lead: 0.5–6.4 points; we take the midpoint, about 3.5 points, or about three months.

3. Evidence

What AI R&D consists of. AI R&D ranges from deciding what to work on, through designing and running experiments, to analyzing and communicating results; for more context, see Owen (2024) and Denain, Kwon, and Ho (2026). We now discuss the best available evidence on how AI performs on these tasks.

3.1 What the labs report

The evidence reported by the labs, summarized in the table below, shows that AI does a large and fast-growing share of AI R&D.

We note that Anthropic rates its internal R&D tasks on a scale introduced by Denain, Kwon, and Ho (2026) from AL0 (no AI involvement) to AL5 ("*AI operates fully autonomously, with no human in the loop*"); at AL3, AI "*collaborates*," and at AL4, it "*leads*" (Anthropic); OpenAI currently provides no equivalent data.

Measure	OpenAI	Anthropic
Share of R&D work done by AI	3.1 agent-workdays per human workday in Aug 2026, up from about 0.5 in May 2026 (OpenAI)	**Claude leads (AL4+) 26% of model R&D tasks in Aug 2026, up from under 1% in Feb 2026**; AI at least collaborates (AL3+) on above 90% (Anthropic). In an internal poll (Aug 2026), staff's median estimate of AI's share of the work of building an early-2026 frontier model was 51% (Anthropic)
AI-written code	Lines changed per active contributor, company-wide: 7.02× the pre-2025 average in Jul 2026 (OpenAI)	**Over 90% of contributions to the codebase written by Claude** (Sep 2026); average daily code volume per engineer in Jul–Sep 2026 over 45× that in 2024 (S-1)
Productivity uplift	**Experiments per active experimenter at an all-time high in Aug 2026, about 1.6× the 2025 level**, although OpenAI notes that "*our available compute has also grown significantly since 2025*" (OpenAI)	Staff survey (n = 130) found that the "*median respondent estimated that they produced around 4× as much output with Mythos Preview as they would have without access to any AI models*" (Anthropic); in a separate survey, 1 of 18 thought that Claude Mythos Preview was equivalent to a drop-in entry-level researcher (Risk Report)

Measure	OpenAI	Anthropic
Measured acceleration	**OpenAI says agentic tools are *"meaningfully accelerating research progress"*** but that its "*measurement efforts are still preliminary*" (OpenAI); GPT-6 Astra "*does not reach our High threshold*" for AI self-improvement (Astra system card)	On Anthropic's internal version of the ECI (AECI), **Claude Opus 5.5 is about 6.5 points above the historical trend of about 15 points a year**; it is unclear whether this is a one-off jump at Mythos Preview (Apr 2026) or a break in the trend, to about 22 points a year since Sep 2025 (Opus 5.5 system card, pp. 40–41). Internal measures "*do not show a sustained AI-attributable 2× acceleration*" (p. 35); **a preliminary external METR assessment of AI R&D acceleration inside Anthropic estimates *"~1.5X overall acceleration in capabilities due to AI,"*** with "*perhaps 30% chance of 2X*" (Opus 5.5 system card, p. 44)
Internal debugging benchmark	Internal Research Debugging Evaluation (41 real bugs from internal experiments that took researchers hours to days, plus 6 alignment-auditing tasks): GPT-6 Astra 78.05%, "*below our indicative threshold for High*" (Astra system card)	CoBench v2 (449 real internal root-cause problems, filtered to ones Mythos Preview failed at least once; Risk Report): Opus 4.6 15.6%, Mythos Preview 54.8%, internal Model 2 62.8%; a "*model which was truly capable of fully substituting for Anthropic research staff would be able to score at least 85% on this evaluation*" (Risk Report)
Internal AI R&D benchmarks	GPT-6 Astra beats GPT-5.6 Sol on all four internal AI R&D evaluations (Astra system card); against human baselines, it closes about 77% of the gap from base model to human on PostTrainBench Lite, but reaches only about a third of the best human score on NanoGPT (about 24% vs. 72%; our reading of the charts)	Frontier models "*surpass human baseline performance on most tasks*" in Anthropic's task-based AI R&D suite (Risk Report)

Measure	OpenAI	Anthropic
Research judgment	No direct measure reported. In Aug 2026, *"high-level planning still remains a minimal fraction of agent output tokens"* (OpenAI)	**Anthropic states that *"large performance gaps persist when it comes to Claude exercising judgement in choosing goals in both engineering and research"*** (Anthropic). Internal users report that Claude Opus 5.5 *"mostly tests incremental ideas and prefers less ambitious hypotheses"* (Opus 5.5 system card, p. 36). According to METR, *"The evidence we have does not suggest that [Claude Opus 5.5] represents a large improvement over Fable 5.1 in these 'judgement' skills"* (same system card, p. 43). In an internal survey, respondents cited weaknesses in *"taste, verification, instruction-following, and epistemics"* when compared to *"an entry-level researcher"* (Risk Report). Direct quantitative measures are limited. On TASTE, where models pick the better of two AI safety research proposals, the best model, Claude Fable 5, agreed with experienced researchers 60% of the time (±10 percentage points), against an estimated 77% for researchers and 50% by chance (TASTE)
Autonomy on long tasks	**Success with no human intervention, Jan → Jul 2026: 2–4-hour tasks 28% → 57%, 4–8 hours 18% → 53%, 8–16 hours 10% → 35%** (OpenAI)	Claude *"is not operating fully autonomously for any measured subset of AI R&D work"* (Anthropic); Claude Code session success on open-ended problems more than tripled in about a year, from about 26% (Aug 2025) to 91% (Sep 2026), and all four difficulty tiers are now at about 88–92% (Anthropic)
AI inference spend per researcher	**Median above $600/day at API prices by mid-Aug 2026, from near zero in Feb 2026**; 90th percentile above $7,000/day (OpenAI)	Not reported

Table 3: *Evidence from OpenAI and Anthropic on AI in their own R&D. Bold text marks the evidence that we view as the most important.*

Our **key takeaways from this evidence**:

- AI writes most code at Anthropic; OpenAI has not reported a share, but code volume per contributor is about 7× the pre-2025 average, suggesting that the same is likely true at OpenAI.
- Researchers at both labs appear more productive. However, self-reported productivity estimates can be fraught, and an outside METR survey of 349 technical workers found a median uplift of 1.4–2× (also self-reported), well below the 4× reported by Anthropic's staff.

- **There is some internal evidence of acceleration.** On Anthropic's internal version of the ECI, Claude Opus 5.5 is about 6.5 points above the historical trend, while OpenAI says agentic tools are "*meaningfully accelerating research progress*" (OpenAI). This matches statements by lab leaders: Sam Altman called AI's help with AI R&D "*a larval version of recursive self-improvement*" (The Gentle Singularity, Jun 2025), Dario Amodei says that RSI "*is starting to happen across the industry*" (We Must Pace the Frontier, Sep 2026), and Jack Clark says that "*this year we have gained convincing evidence that [the singularity] has now begun*" (Import AI 477, Oct 2026). In its IPO filing, Anthropic says that "*each new version of a Claude model directly accelerates development of the next*" (S-1). However, faster input growth explains part of Anthropic's above-trend result. For example, its headcount rose from about 940 at the end of 2024 to about 5,300 on Sep 30, 2026, and its compute and infrastructure expenses grew 300% year over year in Q2 2026 (S-1).
- **Google DeepMind reports a result similar to Anthropic's above-trend AECI but interprets it as being mostly due to input growth, not automated AI R&D.** On its own version of the ECI (called GCI), Gemini 4 Argon is also above trend. However, Google DeepMind suspects that this is due to "*increased organizational investment and an increase in training compute relative to the historical trend, rather than reflecting AI-driven acceleration*" (Gemini 4 Argon report, p. 75). In addition, while most production code at Google DeepMind is now "*>75% AI-written,*" the share of *fully* delegated pull requests was only 11.9% in Sep 2026 (Gemini 4 Argon report, p. 77).
- **The evidence on research judgment is limited but suggests important gaps.** OpenAI reports no direct measure, while Anthropic and its staff describe judgment as a remaining performance gap. Models mostly do well on AI R&D benchmarks, but these test well-defined, verifiable tasks, not judgment (see also Section 3.2). In addition, AI's success rate falls as tasks get longer: at OpenAI, 57% on 2–4-hour tasks against 35% on 8–16-hour tasks without human intervention (Jul 2026). **Overall, AI now does much or most of the execution, but humans still supply most research taste and judgment.** A RAND review of the public evidence reaches a similar conclusion: "*no disclosed system sets its own research agenda, sustains open-ended research over weeks, or designs its own successors*" (Sarma et al., 2026, p. 20).
- Inference spend per researcher is now on the scale of salaries. The median OpenAI researcher's ~$600 a day at API prices comes to about $150,000 a year (assuming 250 workdays). Spend at the 90th percentile, above $7,000 a day, implies about $1.75 million a year. (API prices likely overstate OpenAI's own cost.)

3.2 Public evidence

This section reviews the evidence from public benchmarks and one open-world evaluation.

We look at four public AI R&D benchmarks. Two benchmarks test end-to-end machine learning (ML) work: PostTrainBench and WeirdML. Epoch's benchmark reviews rate both "*Verified*," meaning the benchmark "*can broadly be interpreted as described*" (Epoch).[8] Two other benchmarks attempt to isolate research taste more directly: TasteVal measures how well a model designs and interprets experiments, and Epoch's InnovationEval asks whether an agent can reinvent an ML

[8] For WeirdML, Epoch rates v2 as "Verified." WeirdML v3, which we use, postdates that review.

innovation it has not seen. Epoch rates TasteVal "*Not enough info*" (Epoch, Oct 21, 2026), likely because TasteVal does not release its tasks (TasteVal). Epoch does not review its own benchmarks, such as InnovationEval, to avoid conflicts of interest (Epoch).

Finally, CRUX 2 (Jul 2026) is an open-world evaluation rather than a benchmark: an agent carries out a full research project on an open question, with no fixed task or automatic score, and human experts grade the resulting paper.

The table below summarizes all five.

Benchmark	What it measures and human baseline	Current frontier
PostTrainBench v1.1	Can an agent do the job of a lab's post-training team, turning a raw pretrained model into a useful assistant? The agent gets four small base models, one H100, and 10 hours to post-train them; score = weighted average of the resulting models' benchmark scores. Human baseline: the labs' official instruct models, post-trained by their own teams, score 51.1	On v1.1, Locus (an external scaffold on Opus 5, submitted Sep 2026) 45.6 and Claude Fable 5 (Jun 2026) 41.8, up from 9.9 for Claude Sonnet 4.5 (Sep 2025) (site data)
WeirdML v3	Can an agent build working ML models for small, unusual problems it has not seen before? Eleven hand-made ML tasks; score = normalized task performance (0–100); no human baseline reported	GPT-6 Astra 42.2 in Sep 2026, up from 15.0 for GPT-5.6 Sol (Jul 2026) and 6.5 for Claude Opus 4.5 (Nov 2025) (data)
TasteVal	Can a model choose and interpret experiments as well as a human expert? Eight unreleased AI R&D tasks (e.g., data curation, pre-training, fine-tuning); the tested model only designs experiments and interprets results, while a fixed coding agent runs them. Score = how many times less experiment compute the model needs than the human baseline to reach the same test score. Human baseline = best of at least two experts per task (24 experts from OpenAI, Google DeepMind, and others)	Claude Opus 5.5 (Sep 2026) reaches the experts' score with 2.3× less compute (95% CI 1.15–4.37) and ends slightly above them on final score (1.14 vs. 1.00); Claude Opus 4.8 (May 2026) scored 0.84× (TasteVal)
InnovationEval	Can an agent invent a known ML innovation it has not seen? Given Qwen3-8B, datasets, and a codebase for GRPO (a reinforcement-learning method for post-training), it must devise a better post-training method. Score = share of SDPO's published gain over GRPO (0% = GRPO, 100% = the SDPO paper's results)	GPT-5.6 Sol (Jul 2026) 15% and Claude Fable 5 (Jun 2026) 2%. GPT-6 Astra (Sep 2026) scored 63% but had likely memorized the original paper (InnovationEval)

Benchmark	What it measures and human baseline	Current frontier
CRUX 2 (open-world evaluation)	Can an agent carry out a real, open-ended research project? The agent gets the research question of an unpublished AI paper (but not the paper itself), about a week, an API budget, and GPU credits, and must write a paper for a top AI conference. The original authors then grade it on the NeurIPS 1–6 scale, where 4 is "*Borderline accept*" and 6 is "*Strong Accept*" (NeurIPS)	Original study (Jul 2026): Claude Opus 4.8 in the OpenClaw scaffold; scores: 2 ("*Reject*") and 1 ("*Strong Reject*") (CRUX 2). Follow-up (Oct 2026): GPT-6 Astra in Codex and Claude Fable 5.1 in Claude Code, each with extra-high reasoning; scores: 2 ("*Reject*," GPT-6 Astra) and 1 ("*Strong Reject*," Claude Fable 5.1) (CRUX)

Table 4: *Public AI R&D benchmarks and the CRUX 2 open-world evaluation.*

Our **key takeaways from the public evidence**:

- On the two end-to-end ML benchmarks, which are well-defined, verifiable tasks by construction, models do well. On PostTrainBench, the best agent brings small models about 89% of the way to the labs' official instruct models using one GPU for 10 hours, although Epoch's review warns that agents can score well "*by narrowly (over)fitting to the single eval*" (PostTrainBench). This matches Anthropic's statement that its own task-based AI R&D evaluations "*have 'saturated'*" (Risk Report).
- The two research-taste benchmarks point in different directions. On TasteVal's small, verifiable tasks, Claude Opus 5.5 reaches the experts' score with 2.3× less compute, and this multiplier has doubled about every three months since Dec 2025 (TasteVal). On InnovationEval, agents recovered only 2–15% of the gains of a real ML innovation they had not seen (Epoch). TasteVal does not test the choice of research problems, while InnovationEval rests on one innovation.
- On CRUX 2, both papers were rejected. The study concludes that "*today's agents can do the engineering of AI research, but struggle with critical parts of the research lifecycle*" (CRUX 2), similar to the evidence from the labs. A follow-up with newer models, more time, and a larger budget (Oct 2026) again produced two rejected papers. The authors saw "*meaningful improvements in experimental design and resource management, but persistent limitations in creative problem solving, judgment, calibration, and communication*" (CRUX).
- **Outside benchmarks, the public record of discoveries provides evidence of acceleration.** A METR note (Aug 2026) finds that the discovery of cyber vulnerabilities "*has accelerated sharply*" in 2026 and that of math results "*somewhat*." On the other hand, "*the public record of overall algorithmic optimization does not yet show an appreciable acceleration*" (Cunningham and Rush, 2026). Since that note, the discovery rate in math appears to have increased:
    - OpenAI claims to have solved Navier–Stokes, a Millennium Prize Problem (OpenAI, Sep 2026).
    - On Oct 6, 2026, OpenAI released hundreds of results on open math problems (OpenAI); computer scientist Scott Aaronson wrote that the day was "*surely one of the biggest days in mathematical history*."

- On Oct 26, OpenAI claimed half of a second Millennium Prize Problem, Yang–Mills (the existence of the theory, but not the mass gap). It also claimed a disproof of the Exponential Time Hypothesis, a widely held conjecture about how hard some computational problems are (OpenAI).

3.3 Qualitative evidence

In earlier interviews, Epoch asked eight AI researchers about automation (Owen, 2024). Planning research was vital but took "*relatively little time,*" while coding and debugging were "*more time-consuming*." All eight agreed that "*engineering tasks will be the main driver of R&D automation in the next five years,*" and their descriptions suggested that this "*could automate half or more of their current work*."

Severin Field, Raymond Douglas, and David Krueger (post, published Aug 13, 2026) interviewed 25 researchers at OpenAI, Anthropic, Google DeepMind, Meta, Princeton, UC Berkeley, and Stanford:

- Twenty of 25 ranked automating AI R&D among the most severe and urgent risks; 18 of 25 said it amplifies other risks ("*It just speeds up other threat models*").
- Of the 21 who clearly addressed whether scaling trends will continue, 12 expected them to continue until AI can match the labor of human AI researchers.
- Sixteen of 25 at some point expressed skepticism about positive feedback, the recursive part of RSI. Skeptics said RSI needs "*a drastic change in AI capabilities that wouldn't emerge from gradual improvement alone,*" such as breakthroughs in memory, creativity, or telling true hypotheses from false ones; believers cited METR's time horizons, scaling laws, and what they had experienced inside the labs.
- Of 20 who discussed deployment, half expected the most capable models to stay internal.

4. Existing projections and forecasts

We next review projections of various automated AI R&D milestones. We look at three types of evidence: the labs' own targets (4.1), extrapolations of the measured series (4.2), and forecasts (4.3) from prediction platforms, an LLM forecaster, surveys of experts and superforecasters (forecasters with a track record of accurate predictions), and notable individuals.

4.1 Lab targets

OpenAI and Anthropic both expect automated AI R&D within a few years. OpenAI's stated goal is an automated AI researcher by Mar 2028, and it says it has already reached its intermediate goal, an automated research intern (OpenAI). Anthropic's Frontier Safety Roadmap calls automated AI R&D "*plausible, as soon as early 2027*" (Roadmap).

In its amended IPO filing (Nov 2, 2026), Anthropic states that it "*may imminently meet our Responsible Scaling Policy's definition of 'dramatically accelerated progress attributable to AI'*" (S-1/A, p. 29). In an internal poll of Anthropic staff in Aug 2026, the median expectation for AI's share of the work of building an early-2026 frontier model was 80% by the end of 2027 and 95% by the end of 2030, up from 51% at the time of the poll (Anthropic).

Lab	Target	Status
OpenAI	An automated research intern by Sep 2026 and an automated AI researcher by Mar 2028 (OpenAI)	Says it reached the intern goal: "*According to our measurements, we have now reached the goal*" (Sep 6, 2026; OpenAI)
Anthropic	"*Plausible, as soon as early 2027*," that AI could "*fully automate, or otherwise dramatically accelerate, the work of large, top-tier teams of human researchers*" in domains including "*AI itself*" (Frontier Safety Roadmap)	RSI "*is starting to happen across the industry*" according to Dario Amodei (We Must Pace the Frontier)

Table 5: *OpenAI and Anthropic targets for automated AI R&D.*

4.2 Trend extrapolation

If the measured series simply continue their current trend, most automated-R&D milestones arrive in 2027–2028. For that, we refer to an automated dashboard constructed by Peter Wildeford.

Series	Milestone (dashboard's definition)	Median and 80% interval (Wildeford (2026) dashboard, repo)
Anthropic R&D Automation Index (Anthropic)	90% of tasks fully autonomous (AL5)	May 2028 (Aug 2027–Oct 2029)
Anthropic CoBench (Risk Report)	85% (Anthropic's internal substitution reference)	Jun 2027 (Jan 2027–Mar 2029)
METR 50% time horizon (METR)	174 hours (one work-month)	Jun 2027 (Jan 2027–Aug 2028)
OpenAI experiment velocity (OpenAI)	10× the 2025 level	Feb 2028 (May 2027–Aug 2029)

Table 6: *Trend extrapolations of automated AI R&D milestones from the Wildeford (2026) dashboard. Dates are the dashboard's defaults, obtained on Nov 3, 2026. AL5 is currently 0% at Anthropic. To extrapolate this metric, the dashboard assumes that AL5 trails AL4+ by a lag informed by the measured gaps between lower levels (AL2+ to AL3+ and AL3+ to AL4+), with wide uncertainty.*

4.3 Forecasts

Central estimates for fully automated AI R&D diverge widely, from 2027 to 2052. The earliest come from lab insiders and AI Futures forecasters (Section 4.3.4), the latest from the AI researchers surveyed by AI Impacts (Section 4.3.3).

4.3.1 Prediction platforms

We first review a few forecasts from prediction platforms, summarized in the table below.

Question	Platform	Forecast	Participation
Weakly general AI / general AI announced	Metaculus	medians Aug 2027 / Sep 2030	1,800 / 1,900
AI-authored paper at NeurIPS, ICML, or ICLR before 2028	Metaculus	85%	135
Anthropic reaches ASL-4	Metaculus	median May 2028	13
OpenAI announces artificial general intelligence (AGI) before 2027 / 2028 / 2030	Kalshi	3.5% / 45% / 71%	$1.1M total
OpenAI or Anthropic announces RSI by Dec 31, 2026	Polymarket	9%	$5,000 on this contract
OpenAI or Anthropic announces RSI by 2026 / 2027 / 2028	Kalshi	11% / 57% / 77%	$2,100 total

Table 7: *Forecasts from prediction platforms, as of Nov 3, 2026.*

For context, ASL-4 was the next tier of safety measures in Anthropic's 2025 policy, for models that could dramatically accelerate AI progress (RSP v2.2). Anthropic's 2026 policy (RSP v3.4) no longer uses the ASL-4 level.

Forecasters appear bullish on AI R&D, but few relevant questions are both open and liquid. Metaculus's medians are Aug 2027 for weakly general AI and Sep 2030 for general AI, and its forecasters give 85% to an AI-authored paper at NeurIPS, ICML, or ICLR before 2028. Kalshi's 45% on OpenAI announcing AGI before 2028 is roughly consistent with the Metaculus median for weakly general AI, though the two definitions differ. The two RSI markets give 9% (Polymarket) and 11% (Kalshi) to an announcement by the end of 2026, but both are very thin.

4.3.2 LLM forecasts: FutureSearch

FutureSearch's RSI forecasts contain 57 LLM forecasts (updated Oct 16, 2026), generated using FutureSearch's proprietary model. We discuss the three closest to this document's questions.

FutureSearch's median for automated AI R&D falls inside the range of the trend extrapolations in Section 4.2. Its median for a lab saying that a model has crossed, or within a generation will cross, its automated-AI-R&D threshold is Dec 2027 (80% interval: Feb 2027–Jan 2031). FutureSearch's median forecast is that the top ECI score will rise by 16 points during calendar year 2029 (80% interval: 9.5–28.5 points), close to the current pace of about 14 points a year. It also puts AI R&D uplift at Anthropic today at 1.5× (80% interval: 1.25–2×), similar to METR's preliminary estimate cited in Section 3.1.

4.3.3 Expert and superforecaster surveys

In a pilot study with METR, we surveyed 8 AI-forecasting experts and 10 superforecasters from Nov 2024 to Feb 2025 (FRI write-up). Its main outcome was a threefold (3×) acceleration of AI progress: a two-year period before 2029 with as much growth in effective compute as in the six years from 2018 to 2024. The median expert gave this outcome 20% and the median superforecaster 8%. The survey also asked how likely several AI R&D milestones were by early 2027, and how each milestone would change the chance of a 3× acceleration.

Forecast	Experts (median)	Superforecasters (median)
3× acceleration (main outcome): a two-year window before 2029 with as much effective-compute growth as 2018–2024 (FRI write-up)	20%	8%
Top-researcher replacement: in Jan 2027, almost everyone informed agrees that "*An AI system can act as a full replacement, for AI research and development purposes, for a top (i.e. within the top 0.01%) 2024 AI researcher*" (FRI write-up)	5.5%	20%
Month-long projects: AI beats a five-person team on 80% of month-long, open-ended ML projects while using 20% fewer resources, by early 2027 (FRI write-up)	31.5%	45%

Table 8: *Median forecasts of experts and superforecasters in our pilot survey with METR.*

In our reading, the month-long-projects forecast is close to resolving No. The "month-long projects" scenario needs more than a one-month (174-hour) 50% time horizon by early 2027. By Epoch's rule of thumb, "*a 5 point gain in ECI appeared to roughly correspond to a doubling of the METR Time Horizon*" (Epoch). A one-month horizon is about 5.6 doublings above GPT-5 (ECI 150; about 3.6 hours, METR), or an ECI of about 178, which the ECI frontier reaches around mid-2027 on trend. We are not aware of an evaluation that measures month-long projects directly; the closest, CRUX 2 (week-long), was rejected. The top-researcher replacement scenario also looks unlikely by Jan 2027.

In a separate expert survey, **ML researchers surveyed by AI Impacts expect late AI R&D automation.** Respondents (1,580 authors at top AI venues, surveyed in Dec 2024, with a 10% response rate) put the median date for AI fully automating the job of AI researcher at 2052 (n = 458; ESPAI 2024, p. 4 and Table 6, p. 43).[9] When asked to rate the intelligence-explosion argument (i.e., if AI does nearly all R&D, technological progress could become "*more than an order of magnitude faster*" in less than five years; n = 189), respondents were split: 28% found it likely or

[9] Respondents to the same survey put the median date for AI solving a long-standing math problem without human input, such as a Millennium Prize Problem, at 2053, with only a 10% chance by 2028 (ESPAI 2024, p. 9 and Table 7, p. 45). However, the survey authors note that the task's condition "*without more input from humans*" would exclude the Navier–Stokes solution claimed by OpenAI (ESPAI 2024, p. 9, footnote 7). For context, our Sep 2026 review found that experts and superforecasters in our studies have substantially underestimated AI capability progress, especially on benchmarks (FRI).

quite likely, 23% gave it about even odds, and 49% found it unlikely or quite unlikely (ESPAI 2024, pp. 16–17, Figure 8).

4.3.4 Individual forecasts

The table below shows some notable individual forecasts.

Who	Full automation of AI R&D (central estimate; our reading)	Forecast	Made
AI 2027, a scenario by Daniel Kokotajlo, Eli Lifland, Thomas Larsen, and Romeo Dean of the AI Futures Project, with input from Scott Alexander (About)	2027 (scenario)	Scenario: superhuman coder Mar 2027, superhuman AI researcher Aug 2027 (AI 2027)	Apr 2025
Leopold Aschenbrenner, founder of the hedge fund Situational Awareness LP and author of *Situational Awareness*; formerly on OpenAI's Superalignment team (Wikipedia)	2027	"*it is strikingly plausible that by 2027, models will be able to do the work of an AI researcher/engineer*" (Ch. I)	Jun 2024
Jack Clark, Anthropic cofounder and Head of Public Benefit, who leads the Anthropic Institute (Anthropic)	2027	"*I now think it's more likely we'll see full RSI become possible in 2027*" (Import AI 476); in May 2026 he gave 30% by 2027 and "*a likely chance (60%+)*" of "*no-human-involved AI R&D*" by the end of 2028 (Import AI 455)	May and Oct 2026
Dario Amodei, Anthropic cofounder and CEO; formerly vice president of research at OpenAI (Wikipedia)	2027–2028	The AI feedback loop "*may be only 1–2 years away from a point where the current generation of AI autonomously builds the next*" (The Adolescence of Technology); RSI "*is starting to happen across the industry*" (We Must Pace the Frontier)	Jan and Sep 2026
Daniel Kokotajlo, Executive Director of the AI Futures Project; formerly a governance researcher at OpenAI (AI Futures Project)	End of 2028	AI doing "*the whole research process*": "*about 50% chance by end of 2028*" (Senate testimony); automated coder: median Nov 2027, 80% interval Jan 2027–Apr 2030 (AI Futures)	Aug and Sep 2026
Ryan Greenblatt, member of technical staff at METR; formerly chief scientist at Redwood Research (LinkedIn)	2030–2031	Full automation by end of 2028: "*a bit below 30%*," up from about 15%; median for AI R&D parity early 2031 (LessWrong). Later: "*full automation of AI R&D perhaps somewhere around 2031, 2030*" (Dwarkesh)	Apr and Aug 2026

Who	Full automation of AI R&D (central estimate; our reading)	Forecast	Made
Eli Lifland, research lead at the AI Futures Project (AI Futures Project) and member of the Samotsvety forecasting team (Samotsvety)	After 2030	Automated coder: median Jan 2030, 80% interval Jun 2027–Oct 2070; superintelligence: median Jul 2033 (AI Futures)	Aug 2026
Andrej Karpathy, OpenAI founding member and former director of AI at Tesla; joined Anthropic in May 2026, after this forecast (Wikipedia)	About 2035 (AI agents that work like an employee or intern)	AI agents should work "*almost like an employee or an intern that you would hire,*" but "*They're cognitively lacking and it's just not working. It will take about a decade to work through all of those issues*" (Dwarkesh)	Oct 2025
Ilya Sutskever, cofounder and CEO of Safe Superintelligence Inc.; formerly cofounder and chief scientist of OpenAI (Wikipedia)	2030–2045 (AI that learns as well as a human, then becomes superhuman)	Asked when AI that "*can learn as well as a human*" will "*become superhuman*": "*I think like 5 to 20*" years (Dwarkesh)	Nov 2025
Toby Ord, philosopher and author of *The Precipice* (Wikipedia)	About 2038 (transformative AI)	Transformative AI, which is "*doubling the rate of scientific and technological progress*": "*a median of 2038 and an 80% interval ranging from 3 years to 100 years*"; "*about a 1 in 5 chance*" in the next five years (Forethought)	Mar 2026
Arvind Narayanan and Sayash Kapoor, authors of *AI as Normal Technology*. Narayanan is a professor of computer science at Princeton University (bio); Kapoor is an incoming assistant professor at the UC Berkeley School of Information (bio)	No explicit date; gradual, not soon	"*It is more likely that we will continue to see a gradual increase in the role of automation in AI development than a singular, discontinuous moment when recursive self-improvement is achieved*" (AI as Normal Technology); in a joint post with the AI 2027 authors, they "*do not expect 'strong AGI' to arrive anytime soon*" (Asterisk)	Apr and Nov 2025

Table 9: *Notable individual forecasts of full automation of AI R&D, ordered by timeline.*

5. Key cruxes

Broad acceleration through automated AI R&D needs three contested steps to happen in turn. The diagram below provides our map of the cruxes and shows four main possible outcomes, ranging from an AI research assistant to broad acceleration. The table below provides sample statements of each side.

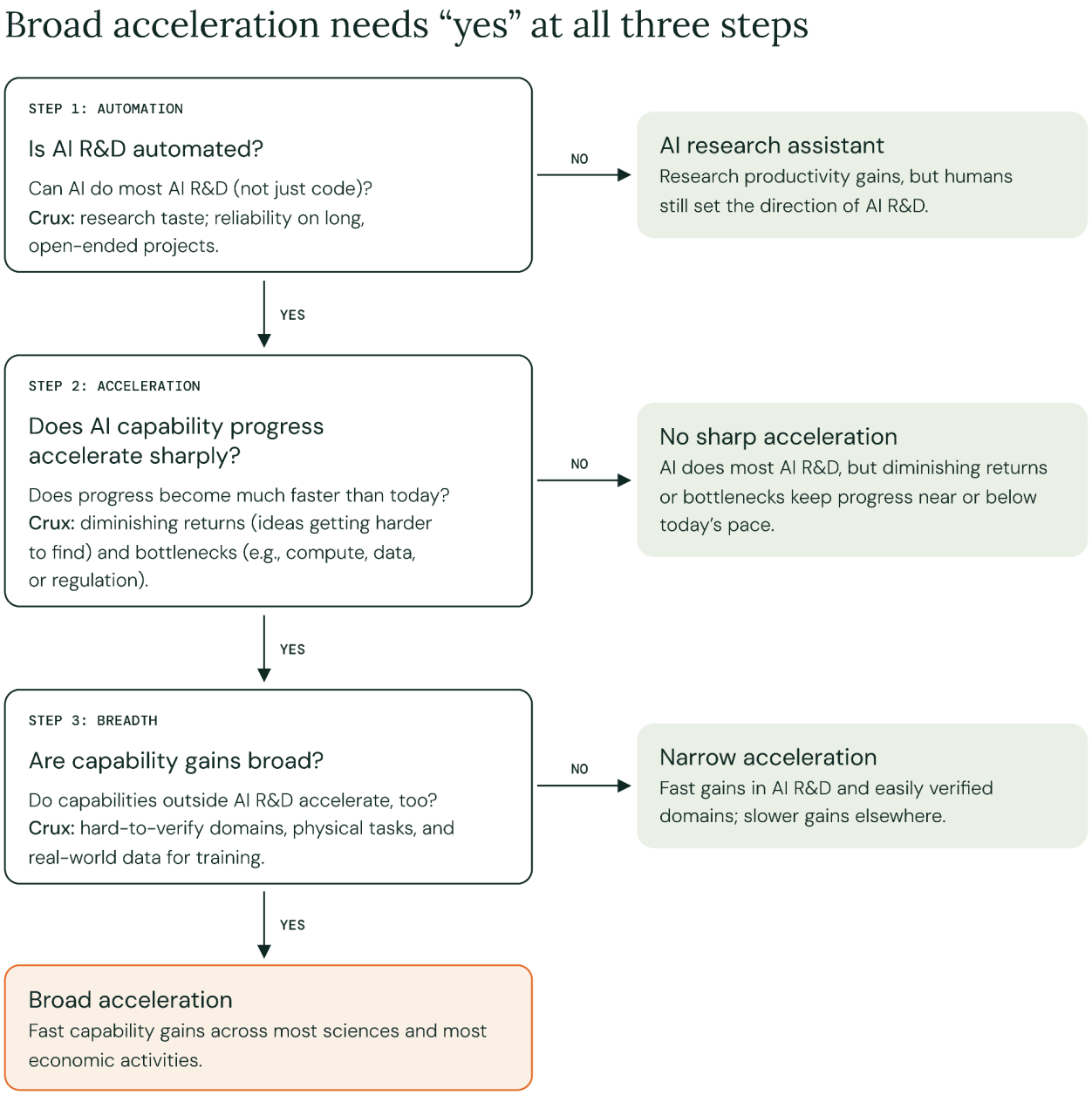


Figure 4: *RSI crux framework.*

Step	Main crux	Yes	No
1. Is AI R&D automated?[10]	Can AI supply research taste and judgment (choosing problems, judging results), or only implementation? Can it stay reliable on long, open-ended projects?	"*I'm probably less sympathetic to the idea that the thing the AIs will lack is some deep insight*"; ML "*is a less deep domain than math*" (Ryan Greenblatt, Dwarkesh); if AI can "*creatively problem solve to succeed at science*," that is a sign of its ability to "*generalize into many domains, including certainly AI research itself*" (Jack Clark, Import AI 477)	"*If you can get the AI to do any experiment, you still need to decide what experiments to do. Right now I think AIs are not very good at this compared to coding the experiment*" (Beren Millidge, Dwarkesh); defining the objective is "*the last job for humans*" (John Schulman, Dwarkesh); "*creativity, among other dimensions of intelligence, is the epitome of an unverifiable task*" (Arvind Narayanan, ICML keynote)
2. If AI R&D is automated, does AI capability progress accelerate sharply?	Does automated AI R&D sharply accelerate AI capability progress, for example in a software intelligence explosion (SIE)? Or do diminishing returns ("*ideas become harder to find*"; Cunningham et al., 2026) or bottlenecks such as compute, data, or regulation cap the acceleration?	An SIE is "*at least decently likely*" even "*if hardware were held constant*" (Daniel Eth and Tom Davidson, Forethought); "*four or five years of AI progress in a single year*" (Ryan Greenblatt, Dwarkesh); on data bottlenecks, "*I'd be surprised if they make the SIE take >2x as long*" (Tom Davidson, Forethought)	Feedback loops "*are not currently strong enough to generate a self-sustaining acceleration, though they appear to be strengthening*" (Cunningham et al., 2026); the loop "*would need to be roughly 5–10× stronger to sustain itself*" (Ramez Naam, Noahpinion), on his assumption of 2–3% more AI R&D productivity per ECI point, against about 9% in Cunningham et al., 2026; "*political and social barriers could stop [a self-sustaining acceleration] in its tracks*" (Cunningham et al., 2026, p. 31)
3. If progress accelerates sharply, are the capability gains broad?	Do capabilities outside AI R&D also accelerate, across most sciences and most economic activities? Or do hard-to-verify domains, physical tasks, and the need for real-world data confine the gains to AI and easily verified tasks?	AI could "*compress the progress that human biologists would have achieved over the next 50-100 years into 5-10 years*" (Dario Amodei, Machines of Loving Grace); "*in competitive industries, AI will learn to automate most work within a year of the SIE*" (Tom Davidson, Forethought)	"*I don't think that automating AI R&D will rapidly produce domain-general superintelligence*"; the singularity "*will be bottlenecked on signal*" (Tom Reed, Substack); "*the external bottlenecks to building and deploying powerful AI systems cannot be overcome by improving their technical design*" (Arvind Narayanan and Sayash Kapoor, Normal Tech)

Table 10: *The main crux at each step, with sample statements for each side.*

[10] Note that AI R&D need not be fully automated. AI that does most, but not all, of the work could be enough to accelerate progress (Section 2.2).

6. Forecasting questions

To understand the future of RSI, we developed 18 forecasting questions. Figure 5 shows which step of the crux framework (Section 5) each question group informs:

1. Group 1 asks how much AI R&D work AI will do on its own, and how well AI agents will do open-ended research.
2. Group 2 asks how fast the ECI frontier will rise.
3. Group 3 asks whether a fast year will happen and, if one happens, whether AI automation of AI R&D caused it.
4. Group 4 asks how far the best models inside AI companies will lead the ECI frontier.
5. Group 5 asks how a fast year would change the chances of expert AGI and an AI catastrophe by 2050.

We are asking a group of top human forecasters and AI experts to forecast these questions. We plan to add their forecasts in a later update of this document.

These questions did well in our internal testing. They capture real disagreements, and the reasons for these disagreements follow the cruxes in Section 5. In an internal survey (Oct 2026, n = 11), our team members gave between 8% and 75% to a fast year by the end of 2030. The fast year also matters for broader outcomes. In a separate internal survey (n = 4), all respondents gave a higher probability to expert AGI and to an AI catastrophe by 2050 if a fast year happens by 2030.

The table below lists each question in short form, together with the LLM forecast, generated by an ensemble of frontier LLMs (Section 6.1). Numeric questions ask for P10, P50, and P90 (the 10th, 50th, and 90th percentiles), and yes-or-no questions ask for a probability. The Appendix gives the full questions, with definitions and resolution criteria.

Each question group informs one step of the crux framework

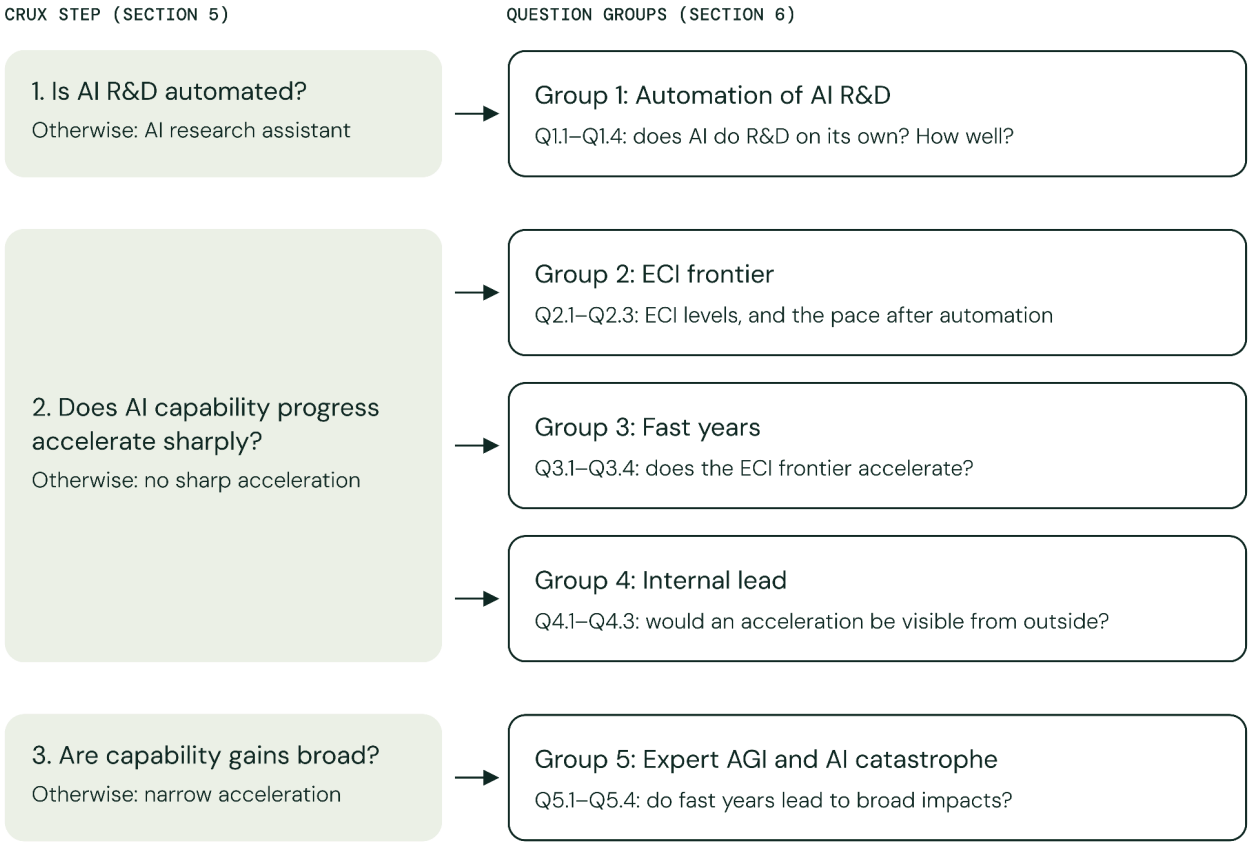


Figure 5: *Question groups mapped to the crux steps.*

ID	Question	LLM forecast
Q1.1	AL5 share of AI R&D work at the leading AI company, Dec 31, 2027	8.3% (80% interval: 0.9–34.5%)
Q1.2	The same, Dec 31, 2030	41.2% (80% interval: 5.6–87.2%)
Q1.3	Average NeurIPS score (1–6) of the best agent's papers in a CRUX 2 re-run started on Dec 31, 2027	2.8 (80% interval: 1.6–4.1)
Q1.4	Both milestones by Dec 31, 2030: an AL5 share above 50%, and a CRUX 2 re-run with an average score of at least 4	32.7%
Q2.1	ECI frontier, Dec 31, 2027	187.9 (80% interval: 179.5–201.4)
Q2.2	ECI frontier, Dec 31, 2030	240.3 (80% interval: 208.1–303.9)

ID	Question	LLM forecast
Q2.3	If both milestones (Q1.4) happen by Dec 31, 2030: ECI frontier gain in the 12 months after	33.1 (80% interval: 13.1–89.0)
Q3.1	A fast year (30+ ECI points in 12 months) by Dec 31, 2027	10%
Q3.2	A fast year by Dec 31, 2030	39.6%
Q3.3	If a fast year happens by Dec 31, 2030: ECI frontier gain in the 12 months after the first fast year	29.8 (80% interval: 11.2–76.3)
Q3.4	If a fast year happens by Dec 31, 2030: probability that AI automation of AI R&D caused it	69.6%
Q4.1	Internal lead over the ECI frontier, Dec 31, 2027	6.4 (80% interval: 1.6–16.6)
Q4.2	Internal lead, Dec 31, 2030	10.1 (80% interval: 2.2–40.3)
Q4.3	If both milestones (Q1.4) happen by Dec 31, 2030: internal lead 12 months after	16.5 (80% interval: 3.8–61.2)
Q5.1	If a fast year happens by Dec 31, 2030: expert AGI by 2050	84.3%
Q5.2	If no fast year happens by Dec 31, 2030: expert AGI by 2050	55.7%
Q5.3	If a fast year happens by Dec 31, 2030: AI catastrophe by 2050	8.4%
Q5.4	If no fast year happens by Dec 31, 2030: AI catastrophe by 2050	2.9%

Table 11: *Forecasting questions in short form, with LLM forecasts made on Oct 29, 2026 (Section 6.1). Numeric forecasts show the median and the 80% interval (P10–P90). For full forecasting questions (including definitions and resolution criteria), see the Appendix.*

6.1 LLM forecasts

Methods. To get an initial set of forecasts, we had an ensemble of four frontier LLMs forecast all 18 questions. Forecasts were generated on Oct 29, 2026. The ensemble consisted of Claude Opus 5.5, Claude Sonnet 5.5, Claude Fable 5.1, and GPT-6 Astra (i.e., three models from Anthropic and one from OpenAI). Each LLM did its own web research and answered every question five times, with a different prompt each time. The LLMs had the survey questions but not this document. For each LLM, we average its five forecasts; the ensemble forecast then is the median of the four means. To generate these LLM forecasts, we used the forecasting pipeline of our Automated AI Risk Outlook (AIRO), which we also use for an AI risk dashboard (GitHub repo).

Results. The ensemble forecasts that AI will soon do much more AI R&D, but not all of it. Its median AL5 share of AI R&D work is 8.3% at the end of 2027 and 41.2% at the end of 2030. Today,

Anthropic rates no part of its AI R&D work at AL5 (Section 3.1). For a CRUX 2 re-run at the end of 2027, the median score is 2.8. This is below a score of 4 ("*Borderline accept*"), but above the scores of 1 and 2 that the CRUX 2 papers got (Section 3.2). The ensemble gives 32.7% to both milestones by the end of 2030.

The ensemble's medians for the ECI frontier are close to the trend: 187.9 ECI points at the end of 2027 and 240.3 at the end of 2030. Both are inside the 80% intervals of the simple trend model in Section 2.3. **The ensemble gives 10% to a fast year by the end of 2027 and 39.6% by the end of 2030**, similar to but somewhat below the trend model (17% and 44%, respectively)**.** The four LLMs disagree substantially on this probability: for a fast year by the end of 2030, their forecasts range from 35% (Claude Sonnet 5.5) to almost 56% (GPT-6 Astra). If a fast year happens, the ensemble gives 69.6% that AI automation of AI R&D caused it, i.e., a fast year would most likely be caused by automated AI R&D.

The ensemble predicts that the gap between the best internal and public models will grow. The median internal lead is 6.4 ECI points at the end of 2027 and 10.1 at the end of 2030, against our estimate of about 3.5–4.5 points today (Section 2.3). The 80% intervals are wide, up to 40.3 points in 2030, as companies could keep their best models internal after automating AI R&D.

A fast year substantially impacts the ensemble's forecasts of expert AGI and AI catastrophe. With a fast year by 2030, it gives 84.3% to expert AGI and 8.4% to an AI catastrophe by 2050. Without a fast year, it predicts 55.7% and 2.9%. In particular, **the ensemble's forecast of an AI catastrophe is almost 3× higher conditional on a fast year.** For comparison, the median expert in our Longitudinal Expert AI Panel (LEAP) gave 2% to an AI catastrophe by 2050 (LEAP Wave 9) and 80% to AGI by 2100 (i.e., with a much longer time horizon; LEAP Wave 8).

These forecasts are mainly illustrative, and we are currently collecting forecasts from top human forecasters and AI experts. That said, on ForecastBench, our forecasting benchmark, several LLMs are now statistically indistinguishable from superforecasters (FRI). In our tests, more capable models are also more accurate at forecasting rare events (AIRO launch post). As a result, we think that these forecasts provide an informative initial benchmark.

7. Takeaways

1. As of early Nov 2026, **AI acts as a highly capable research assistant but is still lacking in research judgment.** At Anthropic, AI writes over 90% of contributions to the codebase (Sep 2026). However, humans still decide which research areas to focus on. The evidence on research judgment by AIs is limited and mostly qualitative but suggests important gaps (Sections 3.1–3.2).
2. **AI capability progress has not clearly accelerated on public measures.** The frontier of the Epoch Capabilities Index has risen by about 14 points a year over the past two years; the largest rise in any 12-month window so far is 21 points (Section 2.3). However, the rate of discovery appears to have accelerated in 2026: sharply for cyber vulnerabilities, according to METR, and increasingly for math as well (Section 3.2).
3. **There is some internal evidence from frontier AI companies indicating a potential acceleration.** A preliminary external METR assessment estimates "*~1.5X overall acceleration in capabilities due to AI*" inside Anthropic, and both Sam Altman and Dario Amodei describe signs of early RSI (Sections 3.1 and 4.1).
    a. **Note:** After our data cutoff of Nov 3, 2026, METR published a report in which it estimates that capability progress at Anthropic is now about 2× its earlier rate (range 1.3–3×), with around 80% of this acceleration attributed to AI. This estimate applies to Anthropic only and includes unreleased models. If this pace holds and reaches public models, it would roughly match the pace of a fast year.[11] Our LLM forecasts were generated on Oct 29, 2026, before the METR report was released.
4. **On a simple extrapolation of current trends, mostly using company-reported metrics, AI could take over most AI R&D tasks by 2027–2028**, consistent with OpenAI's target of an automated AI researcher by Mar 2028 (Sections 4.1–4.2). However, other forecasts diverge widely: central estimates for full automation of AI R&D range from 2027 (lab insiders and AI Futures) to 2052 (the AI Impacts survey of ML researchers; Section 4.3).
5. **Broad acceleration through automated AI R&D requires three contested steps: automation, acceleration, and breadth.** The main cruxes are whether AI develops research judgment, whether diminishing returns or bottlenecks such as compute, data, or regulation cap the acceleration, and whether gains spread beyond AI R&D and easily verified tasks (Section 5).
6. **An ensemble of frontier LLMs gives about a 40% chance of a fast year by the end of 2030,** i.e., 12 months in which the ECI frontier rises by 30 points or more (roughly twice the current pace, Section 2.3). If a fast year happens, the ensemble gives about a 70% chance

[11] Several caveats apply to this estimate: (i) METR notes that Anthropic models "had a lower historical growth rate on AECI than the broader industry frontier before late 2025," which "could explain some of the acceleration" (METR). Similarly, Peter Wildeford suggests that part of this may come from Anthropic scaling up reinforcement learning later than other labs (Wildeford). (ii) When METR adds other developers' frontier models to the fit, the estimated acceleration falls to 1.4× (METR). (iii) The data may be consistent with a change in slope only, or a one-time jump at Mythos Preview plus a smaller change in slope. With both allowed, the slope change falls to about 1.4× (Toby Ord). (iv) Part of METR's assessment relies on evidence that is not public. Anthropic "redacted a majority of the evidence we found most informative for these conclusions" (METR). We have not yet deeply reviewed all of the relevant arguments and may revisit them in future versions of this report.

that AI automation of AI R&D caused it. A fast year by 2030 also raises the ensemble's forecast of expert AGI from 56% to 84% and almost triples its forecast of an AI catastrophe, from 2.9% to 8.4% (Section 6.1).

8. Further reading

1. **OpenAI, Research acceleration: The view inside OpenAI** (Sep 2026). Most of OpenAI's data on how much work AI does inside its research organization.
2. **Anthropic, When AI builds itself and Measurements for understanding the pace of AI development inside frontier labs** (2026). Anthropic's data on AI-written code, staff uplift, and its AL0–AL5 automation index.
3. **Peter Wildeford, RSI dashboard** (2026). Trend extrapolations of the main automation series (Section 4.2).
4. **Chan et al., What if automating AI R&D triggers an intelligence explosion?** (GovAI and Cambridge, Sep 2026). Arguably the best single overview of RSI.
5. **Cunningham et al., The Economics of Recursive Self-Improvement** (arXiv, 2026). A simple economic framework for when automated AI R&D becomes self-sustaining.

Appendix: Full forecasting questions

This appendix gives the full text of the 18 forecasting questions in our survey.

For a yes-or-no question, forecasters give a probability from 0 to 100. For a numeric question, forecasters give three values: P10, P50, and P90.

A.1 Automation of AI R&D

These questions ask how much AI R&D work AI will do, and how well AI agents will do open-ended research. Definitions:

- AL5 is the top level of the AL0–AL5 AI autonomy scale that Anthropic uses to rate its AI R&D tasks: "*AI operates fully autonomously, with no human in the loop*" (Anthropic).
- A model is publicly available when anyone can use it, for example through a public API or a consumer app, possibly for a fee. We will date each model by the date when it first became publicly available, using Epoch AI's dates where possible.
- The ECI frontier on a given date is the highest Epoch Capabilities Index (ECI) score of any model that became publicly available on or before that date (Epoch).
- The leading AI company on a date is the company whose model sets the ECI frontier on that date.
- A CRUX 2 re-run uses the setup of the original CRUX 2 open-world evaluation (Section 3.2; CRUX 2) and the NeurIPS 1–6 scale (NeurIPS).

Q1.1 AL5 share, end of 2027

What share of AI R&D work at the leading AI company will AI do fully autonomously (AL5) on Dec 31, 2027?

Answer: P10, P50, and P90, in percent (0–100).

Resolution: This question will resolve to the share (0–100%) of the leading AI company's AI R&D work that AI does at AL5 on Dec 31, 2027.

- Share: the percentage of the company's AI R&D work that AI does at AL5. We will fix the weight of each task at its share of the company's research time in 2026. For example, if AI does at AL5 all the tasks that took 30% of research time in 2026, the share is 30%, even if those tasks now take much less human time.

Please forecast the true share that an independent study (for example, run by METR or Epoch AI) would measure. Since no such independent study exists today, and the leading company may not publish AL5 data, this question is partly hypothetical. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts to estimate the share).

Background: Section 2.1 (lab definitions of automated AI R&D), Section 3.1 (what the labs report, including the AL0–AL5 scale), Section 4 (existing projections and forecasts), and Section 5 (Step 1 of the key cruxes).

Q1.2 AL5 share, end of 2030

What share of AI R&D work at the leading AI company will AI do fully autonomously (AL5) on Dec 31, 2030?

Answer: P10, P50, and P90, in percent (0–100).

Resolution: The same criteria as Q1.1, but on Dec 31, 2030. The leading AI company is the company that leads on Dec 31, 2030. For the reasons given under Q1.1, this question is partly hypothetical.

Background: Section 2.1 (lab definitions of automated AI R&D), Section 3.1 (what the labs report, including the AL0–AL5 scale), Section 4 (existing projections and forecasts), and Section 5 (Step 1 of the key cruxes).

Q1.3 CRUX 2 re-run score, end of 2027

Imagine a re-run of CRUX 2 that starts on Dec 31, 2027 with the best publicly available AI agent on that date. What will be the average score of the agent's papers, on the NeurIPS 1–6 scale?

Answer: P10, P50, and P90, as NeurIPS scores (1–6).

Resolution: This question will resolve to the average overall score of the papers that the AI agent writes in the re-run.

- Setup: the same as CRUX 2 (Section 3.2).
- Scale: the NeurIPS overall score, from 1 ("*Strong Reject*") to 6 ("*Strong Accept*") (NeurIPS).
- Best publicly available AI agent: the best publicly available AI model on Dec 31, 2027, in the best publicly available agent scaffold.
- Average score: the mean of the overall scores of all papers in the re-run.

Please forecast how expert reviewers would score the papers of the best AI agent at the end of 2027. Since such a re-run may not take place, this question is partly hypothetical. If the CRUX 2 team or another team runs a comparable study near Dec 31, 2027, we will use its results. If not, we will aim to resolve the question with the best evidence available on how well AI agents do open-ended research (which may include, for example, asking a panel of AI experts to estimate the results of such a re-run).

Background: Section 3.2 (public evidence, including the CRUX 2 setup), Section 3.3 (qualitative evidence), and Section 5 (Step 1 of the key cruxes).

Q1.4 Both automated AI R&D milestones by 2030

What is the probability that both of these milestones happen by Dec 31, 2030?

(a) The AL5 share of AI R&D work at the leading AI company is above 50%.

(b) In a CRUX 2 re-run that starts on or before Dec 31, 2030, the papers of the best publicly available AI agent have an average score of at least 4 ("*Borderline accept*" or higher).

Answer: a probability from 0 to 100.

Resolution: This question will resolve Yes if both (a) and (b) happen on or before Dec 31, 2030.

- (a) uses the definitions of Q1.1. The date of (a) is the first date on which the AL5 share is above 50%.
- (b) uses the setup, scale, and average score of Q1.3. The agent is the best publicly available AI agent on the start date of the re-run. If there is more than one re-run, the first re-run with an average score of at least 4 counts. The date of (b) is the start date of that re-run.

For the reasons given under Q1.1 and Q1.3, this question is partly hypothetical. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts whether both milestones have happened).

Background: Section 3.1 (what the labs report), Section 3.2 (public evidence, including CRUX 2), Section 4 (existing projections and forecasts), and Section 5 (Step 1 of the key cruxes).

A.2 ECI frontier

These questions ask how fast AI capabilities will rise, as measured by the Epoch Capabilities Index (ECI; Epoch). Definitions:

- The ECI frontier on a given date is the highest ECI score of any model that became publicly available on or before that date.
- The two automated AI R&D milestones are the two events in Q1.4: (a) the AL5 share of AI R&D work at the leading AI company is above 50%, and (b) in a CRUX 2 re-run, the papers of the best publicly available AI agent have an average score of at least 4.

Q2.1 ECI frontier, end of 2027

What will the ECI frontier be on Dec 31, 2027?

Answer: P10, P50, and P90, in ECI points.

Resolution: This question will resolve to the ECI frontier on Dec 31, 2027.

We will use the ECI scores that Epoch AI publishes (Epoch), and Epoch AI's dates for when each model became publicly available. We will read the data on Mar 31, 2028, so that Epoch AI has time to score models that became publicly available near the end of 2027. If Epoch AI changes the ECI scale, we will convert the scores to the scale in use when this survey opened. If Epoch AI stops updating the ECI, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts to estimate the ECI frontier).

Please assume that Epoch AI keeps adding new benchmarks that are not yet saturated, so that the ECI can keep rising as capabilities improve (Section 2.3). If Epoch AI does not, we will aim to add such benchmarks ourselves and refit the ECI with Epoch AI's public code.

Background: Section 2.3 (our operationalization, including the ECI frontier), Section 4.1 (lab targets), Section 4.2 (trend extrapolation), and Section 4.3 (forecasts).

Q2.2 ECI frontier, end of 2030

What will the ECI frontier be on Dec 31, 2030?

Answer: P10, P50, and P90, in ECI points.

Resolution: The same criteria as Q2.1, but on Dec 31, 2030. We will read the data on Mar 31, 2031.

Background: Section 2.3 (our operationalization, including the ECI frontier), Section 4.1 (lab targets), Section 4.2 (trend extrapolation), and Section 4.3 (forecasts).

Q2.3 ECI gain after both milestones

If both automated AI R&D milestones (Q1.4) happen by Dec 31, 2030: how many ECI points will the ECI frontier gain in the 12 months after the date when both milestones have happened?

Answer: P10, P50, and P90, in ECI points.

Resolution: This is a conditional question. It will resolve only if both milestones of Q1.4 happen on or before Dec 31, 2030. If not, the question will not resolve.

- Start date: the date when both milestones have happened, that is, the later of the two dates in Q1.4.
- Gain: the ECI frontier 12 months after the start date, minus the ECI frontier on the start date. Since the ECI frontier cannot fall, the gain is zero or more.

We will use Epoch AI's data and handle any changes to the ECI as in Q2.1. We will read the data three months after the end of the 12-month period.

Background: Section 2.2 (definitions in the literature, including self-sustaining acceleration) and Section 5 (Step 2 of the key cruxes).

A.3 Fast years

These questions ask whether AI progress will accelerate. Definition:

- A fast year is a rise of the ECI frontier by 30 points or more within any 12-month window. A fast year counts as soon as it happens. For example, if the ECI frontier rises by 30 points in 7 months, a fast year has happened at the end of those 7 months.

Q3.1 Fast year by 2027

What is the probability of a fast year by Dec 31, 2027?

Answer: a probability from 0 to 100.

Resolution: This question will resolve Yes if a fast year happens on or before Dec 31, 2027.

The 12-month window can start before this survey opens, but it must end on or before Dec 31, 2027. We will use Epoch AI's data and handle changes to the ECI as in Q2.1.

Background: Section 2.3 (our operationalization, including the definition of a fast year and its pros and cons), Section 3.1 (what the labs report, including measured acceleration), Section 4 (existing projections and forecasts), and Section 5 (key cruxes).

Q3.2 Fast year by 2030

What is the probability of a fast year by Dec 31, 2030?

Answer: a probability from 0 to 100.

Resolution: The same criteria as Q3.1, but by Dec 31, 2030. A fast year by Dec 31, 2027 also counts as Yes for this question.

Background: Section 2.3 (our operationalization, including the definition of a fast year and its pros and cons), Section 3.1 (what the labs report, including measured acceleration), Section 4 (existing projections and forecasts), and Section 5 (key cruxes).

Q3.3 ECI gain after the first fast year

If a fast year happens by Dec 31, 2030: how many ECI points will the ECI frontier gain in the 12 months after the date when the first fast year has happened?

Answer: P10, P50, and P90, in ECI points.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve.

- Start date: the date when the first fast year happens.
- Gain: the ECI frontier 12 months after the start date, minus the ECI frontier on the start date. Since the ECI frontier cannot fall, the gain is zero or more.

For example, suppose that a model becomes publicly available on Mar 1, 2028 and sets the ECI frontier at 170, and another model becomes publicly available on Oct 1, 2028 and raises it to 205. This rise of 35 points within 7 months is a fast year. If there was no earlier fast year, the start date is Oct 1, 2028, and the gain is the ECI frontier on Oct 1, 2029 minus 205.

We will use Epoch AI's data and handle changes to the ECI as in Q2.1. We will read the data three months after the end of the 12-month period.

Background: Section 2.2 (definitions in the literature, including self-sustaining acceleration), Section 2.3 (fast years), and Section 5 (Step 2 of the key cruxes).

Q3.4 Cause of the first fast year

If a fast year happens by Dec 31, 2030: what is the probability that AI automation of AI R&D caused the first fast year?

Answer: a probability from 0 to 100.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve. The first fast year is as in Q3.3.

AI automation of AI R&D caused the first fast year if it was the direct or proximate cause of the rise in the ECI frontier, or if the first fast year would not have occurred, or would have counterfactually been extremely unlikely to occur, "but for" the substantial involvement of AI in the AI R&D behind the models that raised the ECI frontier during that fast year. Other causes, such as more training compute, can also contribute.

Since no one can observe what would have happened without AI, this question is partly hypothetical. We will aim to resolve it with the best evidence available (which may include, for example, the AI companies' own assessments, such as Anthropic's test of whether a doubling is "*substantially attributable to the automation of research and/or engineering*" (RSP v3.4; Section 2.1), or asking a panel of AI experts).

Background: Section 2.1 (lab definitions, including Anthropic's attribution test), Section 2.3 (our operationalization, including why a fast year could happen without RSI), Section 3.1 (what the labs report, including measured acceleration), and Section 5 (Steps 1 and 2 of the key cruxes).

A.4 Internal lead

These questions ask how far the best AI models inside AI companies will be ahead of the best publicly available models. Definitions:

- The internal lead is the number of ECI points by which the best model available internally at any AI company leads the ECI frontier. For each internal model, we will use the ECI score from Epoch AI once that score is publicly available, and we will date the model by the date when it first became available internally.
- A model is available internally on a date when at least some staff inside the company can use it, for example for AI R&D, safety testing, or daily work. The model counts even if the company restricts its use to a small group of staff or to a few tasks, for example because of

safety concerns or a government restriction. We will date each model by the date when it first became available internally, as the company reports it (for example, in a system card) or as credible reports show.

Q4.1 Internal lead, end of 2027

What will the internal lead over the ECI frontier be on Dec 31, 2027?

Answer: P10, P50, and P90, in ECI points.

Resolution: This question will resolve to the internal lead on Dec 31, 2027.

We will use Epoch AI's ECI scores and handle changes to the ECI as in Q2.1. Epoch AI scores models only after they become publicly available; we will resolve the question when the ECI scores of the models that were available internally on Dec 31, 2027 are publicly available. The internal lead is zero or more by definition, because a publicly available model is also available internally.

Please forecast how far the best AI inside the companies is ahead of the ECI frontier. Since some internal models may never become publicly available, and the dates when models first became available internally may not be disclosed, this question is partly hypothetical. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, the companies' own benchmark reports, or asking a panel of AI experts to estimate the internal lead).

Background: Section 2.3 (our operationalization, including why the ECI frontier counts only publicly available models, and our estimate of the current internal lead), Section 3.1 (what the labs report), and Section 3.3 (qualitative evidence, including researchers' views on whether the most capable models stay internal).

Q4.2 Internal lead, end of 2030

What will the internal lead over the ECI frontier be on Dec 31, 2030?

Answer: P10, P50, and P90, in ECI points.

Resolution: The same criteria as Q4.1, but on Dec 31, 2030. For the reasons given under Q4.1, this question is partly hypothetical.

Background: Section 2.3 (our operationalization, including why the ECI frontier counts only publicly available models, and our estimate of the current internal lead), Section 3.1 (what the labs report), and Section 3.3 (qualitative evidence, including researchers' views on whether the most capable models stay internal).

Q4.3 Internal lead after both milestones

If both automated AI R&D milestones (Q1.4) happen by Dec 31, 2030: what will the internal lead over the ECI frontier be 12 months after the date when both milestones have happened?

Answer: P10, P50, and P90, in ECI points.

Resolution: This is a conditional question. It will resolve only if both milestones of Q1.4 happen on or before Dec 31, 2030. If not, the question will not resolve.

- Start date: the date when both milestones have happened, as in Q2.3.
- Internal lead: as in Q4.1, measured 12 months after the start date.

Please forecast how far the best AI inside the companies is ahead of the ECI frontier, one year after AI does most AI R&D on its own. Since the milestones of Q1.4 and the internal lead are partly hypothetical, this question is partly hypothetical too. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts to estimate the internal lead).

Background: Section 2.1 (lab definitions of automated AI R&D and the safety measures that labs link to them), Section 2.3 (why the ECI frontier counts only publicly available models, and our estimate of the current internal lead), Section 3.3 (qualitative evidence), and Section 5 (key cruxes).

A.5 Expert AGI and AI catastrophe

These questions ask how a fast year would change the chances of expert AGI and an AI catastrophe by 2050. We use the same definitions as FRI's Longitudinal Expert AI Panel (LEAP). Definitions:

- Expert AGI means that more than 50% of LEAP panelists agree that AGI exists (LEAP Wave 8). AGI is a commercially available AI system that (1) can outperform the 90th percentile professional human employee in every primarily non-physical occupation (based on 2025 performance), across all sectors, on at least 90% of the economically useful non-physical tasks that they perform, and (2) has an inference cost no more than 5× the cost of equivalent human labor.
- An AI catastrophe is a catastrophic event caused primarily by AI systems in which more than 10% of the people alive at the start of a five-year period die by the end of that period (LEAP Wave 9). This covers cases where AI is the direct or proximate cause of the deaths, and events that would not have occurred, or would have been extremely unlikely to occur, "but for" the substantial involvement of AI within one year before the event.

Q5.1 Expert AGI by 2050, with a fast year

If a fast year happens by Dec 31, 2030: what is the probability of expert AGI by 2050?

Answer: a probability from 0 to 100.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve. If a fast year happens, the question will resolve Yes if expert AGI exists on or before Dec 31, 2050.

Since the LEAP panel may not exist in 2050, this question is partly hypothetical. If the panel exists when the question resolves, we will use its answer. If not, we will aim to resolve the question with

the best evidence available on whether such a panel would agree that AGI exists (which may include, for example, asking another panel of AI experts).

Background: Section 1 (introduction), Section 2.2 (definitions, including narrow and broad acceleration), and Section 5 (Step 3 of the key cruxes: whether the capability gains are broad).

Q5.2 Expert AGI by 2050, without a fast year

If no fast year happens by Dec 31, 2030: what is the probability of expert AGI by 2050?

Answer: a probability from 0 to 100.

Resolution: The same criteria as Q5.1, but the question will resolve only if no fast year happens on or before Dec 31, 2030. If a fast year happens, the question will not resolve.

Background: Section 1 (introduction), Section 2.2 (definitions, including narrow and broad acceleration), and Section 5 (Step 3 of the key cruxes: whether the capability gains are broad).

Q5.3 AI catastrophe by 2050, with a fast year

If a fast year happens by Dec 31, 2030: what is the probability of an AI catastrophe by 2050?

Answer: a probability from 0 to 100.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve. If a fast year happens, the question will resolve Yes if an AI catastrophe happens on or before Dec 31, 2050.

We will use the resolution criteria of LEAP Wave 9. In particular, for fatality and population data, we will rely on broadly recognized sources, such as Our World in Data, the Economist, and the United Nations. The five-year period must start on or after Dec 31, 2025 and end on or before Dec 31, 2050. If reasonable people disagree about whether such a catastrophe has happened, a panel of experts will resolve the question.

Please give your best forecast of whether such a catastrophe will happen by 2050. Do not change your forecast because of problems with resolution. For example, do not consider whether anyone will be able to resolve the question after a catastrophe.

Background: Section 1 (introduction, including the benefits and risks of RSI), Section 3.3 (qualitative evidence, including researchers' views on the risks), and Section 5 (key cruxes).

Q5.4 AI catastrophe by 2050, without a fast year

If no fast year happens by Dec 31, 2030: what is the probability of an AI catastrophe by 2050?

Answer: a probability from 0 to 100.

Resolution: The same criteria as Q5.3, but the question will resolve only if no fast year happens on or before Dec 31, 2030. If a fast year happens, the question will not resolve.

Background: Section 1 (introduction, including the benefits and risks of RSI), Section 3.3 (qualitative evidence, including researchers' views on the risks), and Section 5 (key cruxes).