

1000 Years of AI Safety: A Model of AI Safety Research during an AI Pause

Simon Steshin

Sage Future

Grigory Reznikov

October 2026

ABSTRACT

We modelled an AI pause that stops AI capabilities research and makes top US labs redirect their resources to AI safety. We estimated how much AI safety work would be done, discussed the implications, and provided recommendations.

Interactive version: 1000yearsofaisafety.com

1 Introduction

A three-year AI pause could buy 1000 years' worth of AI safety work.

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	8,040
SAR Feb 2029	137	471	1,130	3,140	24,100
TED-AI Oct 2029	939	3,180	7,540	20,500	156,000

Table 1. How much AI safety work a pause produces, in **years of the whole 2025 AI safety field's work**. The government bans new large training runs and gets the top three US labs to redirect their GPUs and researchers to safety research. Gray numbers come from a simpler protocol because the AI isn't capable enough.

AI CAPABILITY AT THE START	DATE	DEFINITION
Before AC	June 2027	The pause starts on the date shown, before AC, when AI can do most coding tasks but not yet all of them.
Automated Coder (AC)	May 2028	An AI that can "fully automate an AGI project's coding work", replacing its entire software engineering staff (aifuturesmodel.com).
Superhuman AI Researcher (SAR)	February 2029	An AI that can "fully automate AI R&D": its research taste matches the top human researcher's (aifuturesmodel.com). In the model: AI research taste reaches the 99.9th percentile of human AI researchers, and AC has been reached.
Top-Expert-Dominating AI (TED-AI)	October 2029	An AI "at least as good as top human experts at virtually all cognitive tasks" (aifuturesmodel.com). In the model: AI research taste is 1.5 times as far above the top human researcher's as the top is above the median.

Table 2. The rows of the result tables: the AI capability when the pause starts, and its date in our model.

We developed a model of an AI pause that estimates how much AI safety research it produces. We don't know what the pause will look like, so we modeled five durations, four AI capability levels and four levels of trust in AI. We are going to use it to improve upon Plan A, but it already allows us to calculate different implications:

- A three-year pause from AC produces 1000x more AI safety research than was done in 2025. A nine-year pause from TED-AI produces over 150,000x more.
- Checking it all takes an army. If you let AI work autonomously for 9 years, you will need to hire 2.7 million verifiers to check it all. With safer verification protocols, it goes down to 144k verifiers.
- Without an army, you still get a lot. In the worst case, when AI does a lot and you hire no one, you still get 44% of the potential AI safety research. That's because not all research is equally important, and the labs' own staff verify the most important research first.
- Even AI 2040's strict Plan A produces a lot of AI safety research. In Plan A, external auditors approve every line of code before it runs. In a nine-year pause, this cuts total AI safety work to 3–7% of the potential and 23–26% of the riskier Compartmentalization protocol, but even a three-year pause from SAR still produces 1000x more than in 2025.
- Plan A suggests pausing now and carefully scaling to the maximum controllable AI. How fast? Fast scaling is risky, but it allows more time with smart AI. We compared total AI safety work for "Full Speed" and slower "Plan A" scaling strategies. The difference between the two scenarios decreases with time. If the Consortium rushed straight to TED-AI (October 2029 in our model) instead of reaching it after five years of careful scaling from AC (May 2033), they would finish Plan A's AI safety research 1.27 years earlier with Workload Approval. We modeled only the upper bound.

We also prepared an [explainer](#) for a less technical audience on why even a short AI pause is good.

1.1 What This Means for You

- If you are interested in strategy, control or governance:
 - If your plans depend on eliciting useful AI safety research from AI:
 - Estimate how much AI safety research you can elicit. Maximize it under the given safety measures.

- If you are interested in technical alignment, interpretability or generalization:
 - If your work isn't likely to decrease risks in the next 1–2 years:
 - Prepare for the pause and for automation. Create moonshot benchmarks: difficult research problems for future AI scientists to solve. Red-team them to avoid gaming, and try to parallelize your agenda.
 - Create an NGO that collects benchmarks from specialists and red-teams them.
- If you are interested in epistemic AI: We suggest two epistemic AI projects and estimate their economic impact.
 - Build tools that help humans verify AI-generated research.
 - Make AI produce human-readable abstractions.
- For other AI safety researchers:
 - Since AI automates regular computational AI safety work, it becomes much more efficient to focus on non-computational topics: philosophy and agent foundations.

This article has three parts:

- The AI Pause model
- Its implications
- Our recommendations

2 AI Pause Model

Capabilities increase quickly, unlike our understanding of alignment. The AI race makes its participants push capabilities, and the most likely solution is some kind of pacing, whether an international deal ([MacAskill, 2026](#)), domestic regulation ([Lifland et al., 2026](#); [Wildeford, 2026](#)) or maybe even coordination between the labs ([Hubinger, 2023](#)).

Imagine that the top three US labs somehow pause the capabilities race and redirect their resources to AI safety research. What would that pause look like? And how much AI safety research would it produce?

AI safety is an overloaded term. Here we mean experimental research that improves our understanding of AI and helps predict its behavior:

- Interpretability and explainability
- Adversarial robustness
- Science of deep learning (generalization, memorization, training dynamics)
- Jailbreaks, red-teaming, prompt injection and safeguards
- Alignment experiments (reward hacking, deception, model organisms, oversight, chain of thought)
- Unlearning
- AI control

We don't mean:

- Theoretical work (agent foundations, philosophy, decision theory, timeline modeling, economics, surveys)
- Governance
- Privacy, fairness, watermarks
- Preference learning

We appreciate theoretical work and think it's likely to become a bottleneck once everything else is automated, but it seems to be quite a small part of 2025 AI safety work. Governance is important (that's basically our message), but here we focus on the technical understanding of AI.

2.1 A Toy Model

These kinds of AI safety experiments are often bottlenecked by GPU compute. Let's calculate how much compute was spent on AI safety in 2025 and compare it with how much compute we would get if the top three AI labs redirected their 2027–2029 compute to AI safety.

In 2025, about 5,000 full-time researchers worked on AI safety and published over 7,000 papers. Their experiments used as much compute as 27,000 H100s running for a year (see our [estimates](#)).

We sum up the AI labs' experimental and training compute for 2027–2029 and divide it by the AI safety compute in 2025.

That gives 10,947x more AI safety compute.

Why didn't we call this article "10,947 years of AI safety"? Because in the real world:

- We also need labor to come up with experiment ideas, write code, and analyze the results.
- Compute and labor have diminishing returns.
- Labor has a parallelization penalty: nine scientists can't deliver a paper in one month.

- AI is automating more and more of the coding.
- At some point, AI will start helping with research taste: choosing experiments and analyzing the results.
- AI coders and AI researchers need compute to run.
- The more AI automates, the more humans have to verify.
- The safer the R&D process is, the more humans have to verify, and the more limited AI's help is.
- The more AI-generated research there is, the more difficult it is for humans to read.
- Papers differ in value, and humans check the most valuable ones first.

To account for that, we developed an AI Pause model.

During a pause, AI can produce a lot of work, but humans will likely need to check it. We don't know exactly what they will check, so we describe four verification protocols. Each protocol builds on the previous one by adding a new verification mechanism.

2.2 Full Trust

We build on the AI Futures Model (AIFM) (Lifland et al., 2025). It's a takeoff model that captures many factors of AI R&D progress. At its core is a semi-endogenous growth model (Jones, 1995; Bloom et al., 2020), which assumes that ideas get harder to find, so progress needs more research input. It's inspired by Davidson's Full Takeoff Model (Davidson, 2023) and Epoch's GATE model (Erdil et al., 2025).

With small modifications, we can turn it into an AI pause model with full trust, where no humans verify the AI's work. Here we describe our Full Trust modification and briefly recap the original model, but we assume the reader has already studied the [AIFM](#) site and maybe even its [Supplementary Materials](#).¹

2.2.1 Protocol

An international deal, domestic regulation or self-pacing leads to an AI pause.² OpenAI, Anthropic and Google DeepMind stop capabilities research and instead do AI R&D that improves our understanding of AI rather than its capabilities.

In the Full Trust model, we count all research and don't model human verification.

2.2.2 Effective Compute

AIFM measures AI capability with effective compute $E(t)$. The more compute $C_{train}(t)$ has been spent on training the frontier model, and the better the training algorithms $S(t)$ were when it was spent, the higher $E(t)$.

During a pause, frontier training stops, so E stays frozen at E_p .

$$\overset{\text{AIFM ONLY}}{\frac{dE}{dt}} = C_{train}(t) S(t) \quad (1)$$

$$E(t) \equiv E_p \quad (2)$$

2.2.3 Merged Labs

AIFM models only the leading AI lab. In an AI pause, OpenAI, Anthropic and Google DeepMind merge. This could come from:

- An international deal à la [AI 2040](#) making everybody cooperate, including the top three US labs.
- Domestic regulations making US labs cooperate (Lifland et al., 2026a), or some kind of merged Manhattan Project (Brundage, 2024).
- Self-pacing AI labs cooperating by sharing their AI safety research (Bowman et al., 2025).

Adding xAI and Meta, or Chinese labs, would bring even more resources and so more years of 2025 AI safety work. Here we model only the three labs together. They don't need to actually merge: it's enough for them to see each other's research results, so they don't duplicate work. There will still be some diminishing returns and stepping on toes, but AIFM already models these inside one lab.

The merged labs have m_C times the compute and m_L times the staff of AIFM's leading lab. Their compute keeps growing along AIFM's path, while their staff stay as they were at the start of the pause.³

$$\begin{aligned} C(t) &= m_C C^{\text{AIFM}}(t) \\ L_H &= m_L L_H^{\text{AIFM}}(t_p) \end{aligned} \quad (3)$$

2.2.4 Compute Shares

There are four shares of compute in AIFM:

- training frontier models
- research experiments
- AI coders
- external deployment

In a pause, training stops. External deployment compute is frozen (see [rationale](#)). AI coders keep their AIFM share. Experiments get the rest.

$$C = C_{train} + C_{xpm} + C_{aut} + C_{ext} \quad (4)$$

$$\begin{aligned} C_{train}(t) &= 0 \\ C_{aut}(t) &= s_{aut} C(t) \\ C_{ext} &= s_{ext} C(t_p) \\ C_{xpm}(t) &= C(t) - C_{ext} - C_{aut}(t) \end{aligned} \quad (5)$$

2.2.5 Experiment Throughput

To run experiments, the labs need both compute and coding labor.

2.2.5.1 Coding Labor

Let's index coding tasks by i from 0 to 1. Depending on AI capabilities, both humans and AI coders can work on a task, together amounting to G_i person-equivalents on task i : $L_{H,i}$ humans plus AI coders using $C_{aut,i}$ of compute, each unit of compute worth η_i people. η_i is zero if the AI can't do the task at all and grows with AI capabilities.

Tasks complement each other ($\rho_C < 0$). The labs allocate people and AI optimally. In practice, the AI takes the tasks it's most efficient at, and humans do the rest.

$$\begin{aligned} G_i &= L_{H,i} + \eta_i C_{aut,i} \\ L_C &= \max \left(\int_0^1 G_i^{\rho_C} di \right)^{1/\rho_C} \end{aligned} \quad (6)$$

$$\int_0^1 L_{H,i} di = L_H, \quad \int_0^1 C_{aut,i} di = C_{aut} \quad (7)$$

2.2.5.1.1 Automation

The AI can do tasks i from 0 to $A(E)$. This automation fraction grows as a sigmoid in $\log E$. Task i becomes automatable at E_i , and when the AI can do every task, that marks the arrival of the Automated Coder (AC).⁴ Once a task is automated, AI coders on it keep getting more efficient: each order of magnitude of E adds η_{slope} orders of magnitude to their efficiency η_i , starting from η_{init} .

$$A(E) = \min \left(1, \frac{A_\infty}{1 + e^{-k(\log E - x_0)}} \right) \quad (8)$$

$$A(E_i) = i, \quad A(E_{AC}) = 1 \quad (9)$$

$$\eta_i(E) = \begin{cases} \eta_{init} (E/E_i)^{\eta_{slope}}, & i \leq A(E) \\ 0, & i > A(E) \end{cases} \quad (10)$$

2.2.5.2 Experiment Throughput

X is the main object in all our models. We keep AIFM's name, experiment throughput, but it could also be called research volume, quality-unadjusted research effort or research work. We call it research for short.

X is how much research the labs get done in a year: their experiments, weighted by the compute they use and the coding they take. It combines experiment compute C_{xpm} and coding labor L_C , each with diminishing returns. $\lambda < 1$ is the penalty for working in parallel: twice the coders don't code twice as fast.

$$X = \left[\alpha_X \tilde{C}_{xpm}^{\rho_X} + (1 - \alpha_X) \tilde{L}_C^{\rho_X} \right]^{1/\rho_X} \quad (11)$$

$$\tilde{C}_{xpm} = C_{xpm}^\zeta, \quad \tilde{L}_C = L_C^\lambda$$

2.2.6 Research Taste

Some research is more valuable than other research, and people differ in how well they pick it. AIFM models this with research taste: a researcher with taste 2 produces work worth twice as much as one with taste 1.

The AIFM authors model human taste as lognormal (see [rationale](#)) and the AI's taste as improving with $\log E$ (see [rationale](#)).

When the AI's research taste surpasses a researcher's, that researcher defers planning and analysis to it, so each researcher works with the larger of their own taste and the AI's. In a pause, E stays at E_p , so the AI's taste stays where it was when the pause began.

Originally, the AIFM authors estimated the taste slope for capabilities R&D. We model **computational AI safety research**, which in practice looks very similar to capabilities research. We think that, by default, AI will be trained in environments that benefit both types of AI R&D similarly. That's why we keep AIFM's parameters.

$$\begin{aligned} T_{AI}^{SD}(E) &= T_{AC}^{SD} + s_T (\log E - \log E_{AC}) \\ T_{AI}(E) &= T_H(T_{AI}^{SD}(E)) \\ T(E) &= \mathbb{E}_Z \left[\max(T_H(Z), T_{AI}(E)) \right] \\ Z &\sim \mathcal{N}(0, 1) \end{aligned} \tag{12}$$

2.2.7 Research Effort

Research effort is taste-adjusted X . When the AI surpasses a researcher in research taste, we count it as if that researcher did more work, in proportion to the AI's taste.

$$RE = TX \tag{13}$$

2.2.8 Research Stock

The research stock RS is all the research effort done so far. In AIFM, it loops back to software efficiency: every β doublings of the stock double S , since ideas get harder to find.

In a semi-endogenous growth model, $S(t)$ is a kind of final output. In AIFM, it is software efficiency; in semiconductor research, it could be transistor density (Bloom et al., 2020). In AI safety, the real output is hard to measure: "reducing the risks of AI takeover and misuse" or "improving our understanding of intelligence". Although these are the outputs we actually care about, we can neither measure nor model them, so we model AI safety work instead.

Safety work is not an ideal proxy for risk reduction because of diminishing returns. A thousand years of AI safety work won't reduce risk a thousand times as much as the 2025 safety research did. Still, if we decide to continue training AI, we think more safety work reduces risk monotonically.⁵

$$RS(t) = \int_{-\infty}^t RE dt' \tag{14}$$

$$S = \left(\frac{RS}{RS_0} \right)^{1/\beta} \quad (15)$$

2.2.9 Results

See how we estimated the input **parameters**.

We accumulate RE over the pause and divide it by RS_s , a year of the 2025 AI safety field's work. See how we estimated **2025 AI safety work RS_s** .

$$RS_s = X(C_s, L_C(N_s, C_{aut,s}, E_{2025})) \quad (16)$$

$$\text{cell} = \frac{RS(t_p + D) - RS(t_p)}{RS_s} = \frac{1}{RS_s} \int_{t_p}^{t_p + D} RE dt \quad (17)$$

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	41.0	151	375	1,070	10,100
SAR Feb 2029	372	1,290	3,140	8,840	85,400
TED-AI Oct 2029	5,640	19,400	46,900	131,000	1,280,000

Table 3. AI safety work produced in a Full Trust pause, in years of the whole 2025 AI safety field's work.

2.3 Results Verification

Results Verification is the **Full Trust** model with the changes described in all the subsections below.

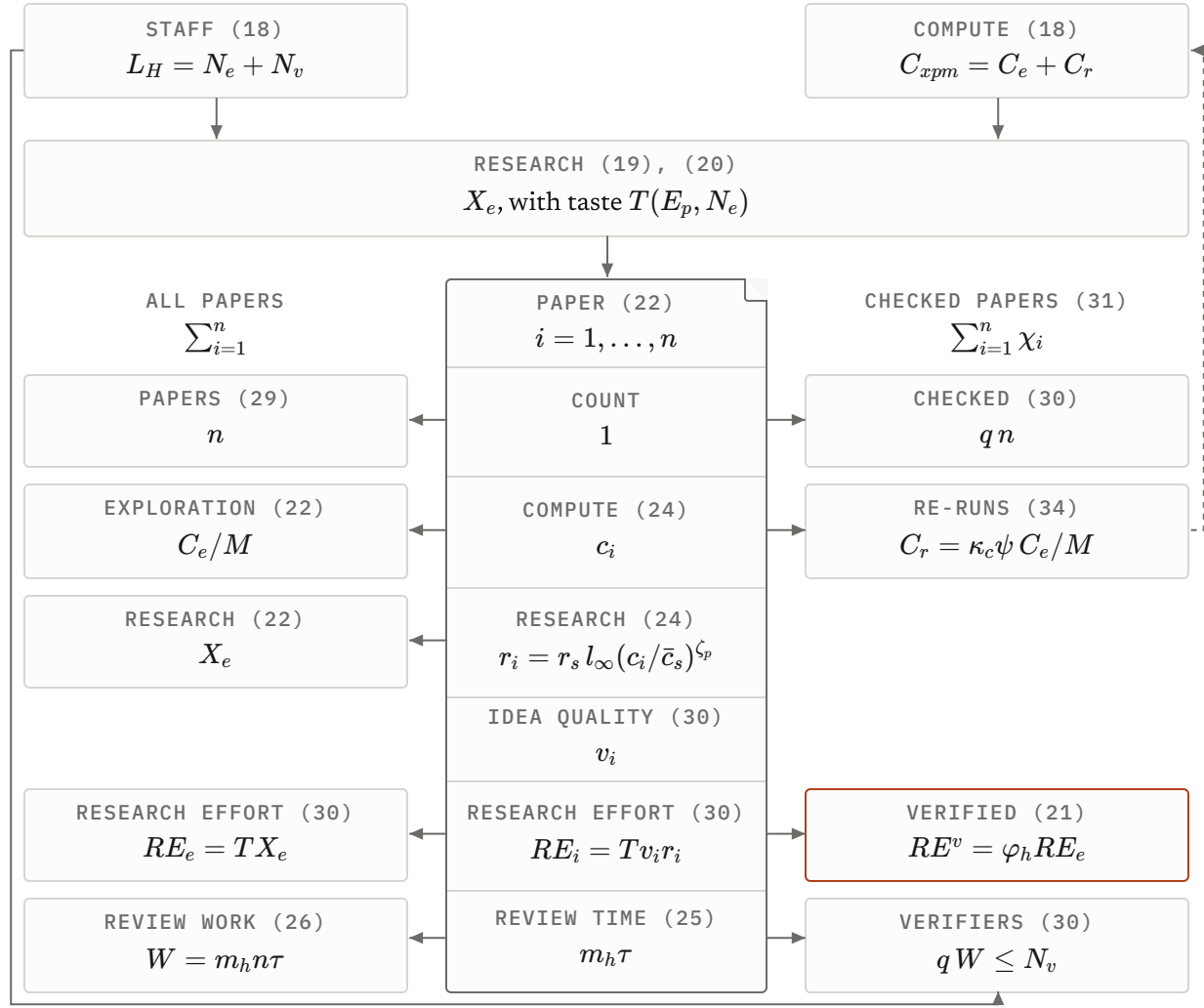


Figure 1. The labs pick the N_e with the most RE^v (35). With unlimited hiring (36)–(37), every paper is checked.

2.3.1 Protocol

Full Trust basically counts all the AI-generated research. In practice, if we have an AI pause, we won't fully trust AI. And we shouldn't:

- It may reward-hack and oversell bad results.
- If AI is scheming, it might sabotage our research to avoid being detected (Benton et al., 2024; Gasteiger et al., 2025; Kirk et al., 2026). For example, it could avoid making lie detectors that work.
- AI might explore red-teaming, jailbreaking and control research to tamper with our monitors (Roger et al., 2024).

- AI can poison our datasets or training techniques to make the next generations of AI aligned with it rather than with humans (Cloud et al., 2025; Souly et al., 2025; Kwon et al., 2026).

From now on, we stop taking what AI does on faith and add verification protocols. In Results Verification, we count only research that was verified by humans.

Als still do the research and run many experiments autonomously. Some of them fail, but the rest are written up as papers with code. Human verifiers check the most important papers first. Each paper gets m_h verifiers: they read it, review or re-implement the code, and re-run the main experiments once. Only papers that were checked and reproduced count. We call their research effort verified research effort, RE^v . It requires both compute for reproduction and human time for verification.

Before AC, people do research much as in 2025, so we count pre-AC research the same way, without extra checks, and the pre-AC row equals the Full Trust one. From AC on, the AI codes on its own and experiments run without people, so we check them: Results Verification is for the AC, SAR and TED-AI rows.

Once the AI researches on its own, verifiers can become the bottleneck: the AI may write papers faster than the staff can check them. So we compute two versions:

- By default, the labs keep the staff they had at the start of the pause, so verifiers can become the bottleneck.
- Alternatively, we check what would happen if labs could hire as many verifiers as they need.

Usually, researchers are rewarded for discoveries, but here AI does nearly all the science. So we recommend rewarding verifiers for finding mistakes in AI work.

We note that Results Verification is hardly a safety measure, as allowing millions of misaligned AI Einsteins to run autonomously might cause catastrophes.

2.3.2 Splits

The labs split their experiment compute into exploration compute C_e for the AI's own experiments and reproduction compute C_r for verification.

They split their staff into the N_e people with the best research taste, who keep researching, and N_v verifiers.

$$\begin{aligned} C_{xpm} &= C_e + C_r \\ L_H &= N_e + N_v \end{aligned} \tag{18}$$

2.3.3 Research

The labs research exactly as in Full Trust, only with less compute and fewer people: X_e is AIFM's X of the exploration compute and of the coding labor of the N_e researchers and the AI coders.

$$X_e = X(C_e, L_C(N_e, C_{aut}, E_p)) \quad (19)$$

2.3.3.1 Taste

We changed the formula for taste. It equals AIFM's formula everywhere except in Results Verification at AC, where it fixes an unrealistic behavior. Skip the details on a first reading.

$$\begin{aligned} T(E, N_e) &= T_{AI}(E) + \int_{\hat{z}}^{\infty} \max(T_H(z) - T_{AI}(E), 0) d\Phi(z) \\ \hat{z} &= \Phi^{-1}\left(1 - \frac{N_e}{L_H}\right), \quad T(E, L_H) = T(E) \end{aligned} \quad (20)$$

2.3.4 Verified Research

As in Full Trust, the labs' research effort RE_e is their research X_e adjusted for taste. Humans verify the share φ_h of it. Verification costs two resources:

- Verifier labor, limited by the number of verifiers. They check the papers worth the most first.
- Reproduction compute. Every checked paper must be re-run, so the labs set aside exactly enough compute to re-run what the verifiers check.

To get φ_h , we need two things: how many papers there are, and how the research effort is spread across them.

$$\begin{aligned} RE_e &= T(E_p, N_e) X_e \\ RE^v &= \varphi_h RE_e \end{aligned} \quad (21)$$

2.3.5 Papers

In AIFM, there's an amorphous mass of experimental work, X . In reality, humans verify research in pieces, so we need a way to cut X into verifiable pieces. We call such a piece a paper.

A paper here isn't an academic publication. It's a piece of the labs' research that takes a fixed amount of human labor to verify, apart from studying its prerequisites. It's similar to how human researchers might decide to split a big paper into several smaller ones. With this definition, we don't need to think about how different research changes the verification labor.⁶

Each year, the labs write $n(t)$ papers. Paper i used compute c_i , dead ends included, and coding labor l_i . Its research r_i is how much research went into it, whatever its idea. Later, we introduce the quality of its idea, v_i .

A paper is made of compute and labor, like the labs' total research, so we model its research r_i with a similar CES function. In such a CES, compute and labor are poor substitutes: you can't replace much compute with labor, so each input alone can raise research only up to a ceiling. A paper's labor ceiling, l_∞ , says how many times more research a paper holds with all the labor it can use than a 2025 paper with the same compute.

From AC on, the labs have far more coding labor per paper than the 2025 field had, so we assume every paper has all the labor it can use. The CES collapses to $r_i = r_s l_\infty (c_i / \bar{c}_s)^{\zeta_p}$. The extra labor buys careful baselines, no bugs, no broken runs. It can't buy more claims or more code to review, because we fix the verification time of a paper. More claims and more code go to more papers.

$$X_e = \sum_{i=1}^n r_i, \quad C_e = M \sum_{i=1}^n c_i = Mn\bar{c} \quad (22)$$

NOT IN THE MODEL

$$r_i = r_s \left[l_\infty^{\rho_p} \left(\frac{c_i}{\bar{c}_s} \right)^{\zeta_p \rho_p} + c_\infty^{\rho_p} \left(\frac{l_i}{\bar{l}_s} \right)^{\lambda \rho_p} \right]^{1/\rho_p} \quad (23)$$

$$l_\infty^{\rho_p} + c_\infty^{\rho_p} = 1$$

$$\rho_p < 0, \quad l_i \rightarrow \infty : \quad r_i \rightarrow r_s \left[l_\infty^{\rho_p} \left(\frac{c_i}{\bar{c}_s} \right)^{\zeta_p \rho_p} \right]^{1/\rho_p}$$

$$r_i = r_s l_\infty \left(\frac{c_i}{\bar{c}_s} \right)^{\zeta_p} \quad (24)$$

$$\ln c_i \sim \mathcal{N} \left(\ln \bar{c} - \frac{\sigma_c^2}{2}, \sigma_c^2 \right)$$

2.3.5.1 Paper Compute

A paper with more compute holds more research: more models, more random seeds, more ablations. But returns diminish, as they do for X : ten times the compute gives a paper 10^{ζ_p} times as much research.

We also account for wasted compute:

- There are failed experiments inside a paper. A paper describes only the successful experiments, but in reality it took more compute. Whenever we mention a paper's compute, we mean compute that includes these dead ends. For example, the compute of papers in the **AI safety papers dataset** already includes them.
- There are failed projects that were abandoned and never published. Still, they eat some of C_e , as they ate some of the 2025 AI safety compute, C_s . We account for them with the coefficient M .

Paper compute is lognormal: its logarithm has standard deviation σ_c , and its mean is the year's average $\bar{c}$. Together, a year's papers make up the labs' research X_e and, with the failed projects, use their exploration compute C_e .

2.3.6 Review Work

Let τ be one verifier's time to check a paper. It consists of two parts:

- reading it, reviewing or re-implementing the code, and checking the re-run;
- reading the prerequisites, if the verifier missed some of them.

We express the first part through the 2025 AI safety field. We estimate how much FTE time it took to write a paper, t_s , and how much faster it is to verify a paper than to write it, κ_h .

But the field is growing quickly, proportionally to $n(t)$, so verifiers likely won't know all the prerequisites, and we need to account for the time to read them. The prerequisites are likely already verified, since they are important enough to be built upon. But a specific verifier might not have read them.

We tried to model this with random graphs, but ended up with a simple approximation instead. The time grows proportionally to the logarithm of $n(t)/n_s$, so each time the field grows 10x, it takes p_h more FTE-years to verify a paper. We used our model and intuition to fit p_h , but we aren't sure of it, so we use wide confidence intervals, and we chose a rather small p_h to make our model more conservative and boring.

Checking the labs' n papers a year, m_h times each, takes $W = m_h n \tau$ FTE-years every year. So W is in FTE: the number of verifiers the labs need, working full time. Written with κ_h and p_h , it needs n/n_s , how many times more papers the labs write than the 2025 field did, and the 2025 field's own count n_s , which sets t_s .

$$t_s = \frac{N_s}{n_s} \quad (25)$$

$$\tau = \kappa_h t_s + p_h \log_{10} \frac{n}{n_s}$$

$$W = m_h n \tau = m_h N_s \frac{n}{n_s} \left(\kappa_h + \frac{p_h}{t_s} \log_{10} \frac{n}{n_s} \right) \quad (26)$$

2.3.7 Paper Count

A year's papers have to satisfy two sums at once: their research adds up to X_e , and their compute adds up to C_e . Together, they give n/n_s .

Start with research. The labs do k_X times the 2025 field's research, which is how many times more papers they write, n/n_s , times how much more research each one carries: l_∞ times more for its labor, and $(\bar{c}/\bar{c}_s)^{\zeta_p}$ times more for its compute.

Now compute. The average compute per paper is the compute divided by M and by the count, for the labs and for the 2025 field, whose compute also holds its failed projects. M cancels, which gives the paper count in terms of k_X and k_C .

What this says:

- More research at the same compute (more staff, or more automation) means more papers, more than in proportion: each paper gets less compute, so it carries less research.
- More compute at the same research means fewer, bigger papers.
- When compute grows faster than research, as it does through most of a pause, compute per paper and research per paper both rise, but research per paper rises more slowly.
- The more research a paper gains from labor, l_∞ , the fewer and bigger the papers. With $l_\infty = 1$, labor makes more papers, not bigger ones.
- With $\zeta_p = 0$, compute doesn't matter, and the number of papers grows exactly with research.

$$k_X = \frac{X_e}{X_s} = \underbrace{\frac{n}{n_s}}_{\text{paper count vs 2025}} \cdot \underbrace{l_\infty \left(\frac{\bar{c}}{\bar{c}_s} \right)^{\zeta_p}}_{\text{research per paper vs 2025}} \quad (27)$$

$$\frac{\bar{c}}{\bar{c}_s} = \frac{C_e/(Mn)}{C_s/(Mn_s)} = \frac{k_C}{n/n_s}, \quad k_C = \frac{C_e}{C_s} \quad (28)$$

$$\frac{n}{n_s} = \left(\frac{k_X}{l_\infty k_C^{\zeta_p}} \right)^{1/(1-\zeta_p)} \quad (29)$$

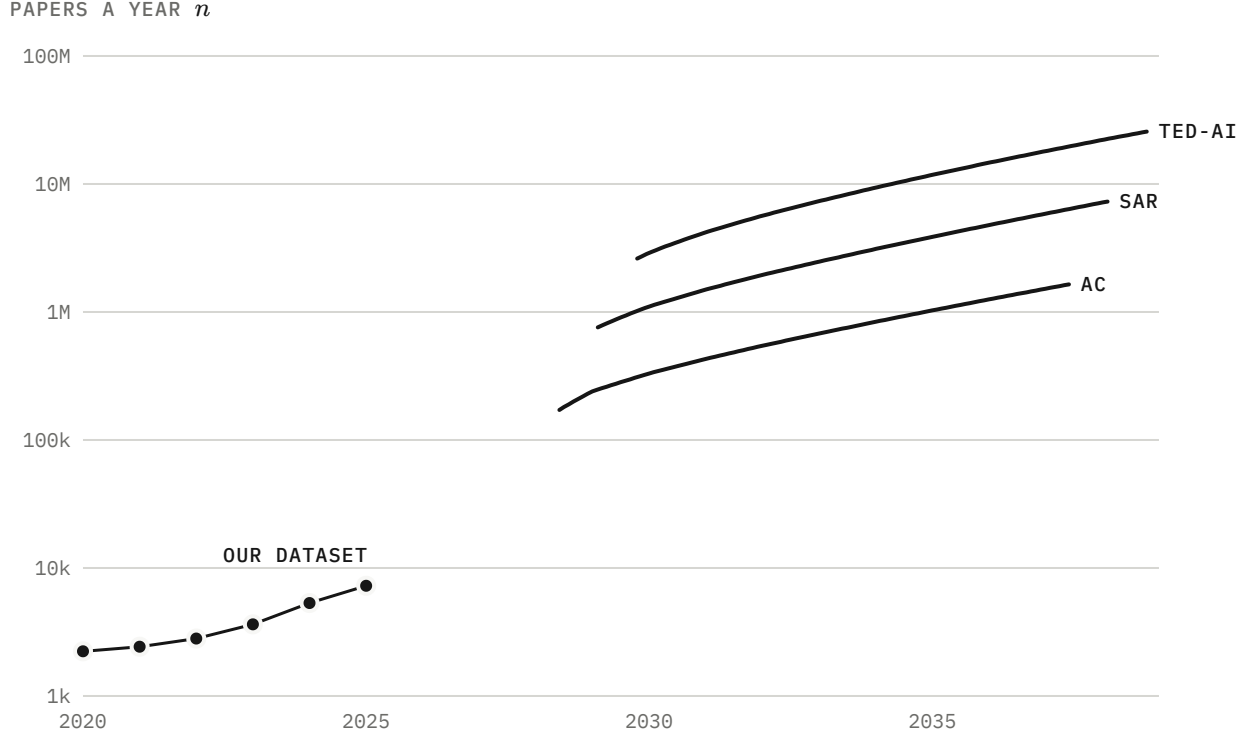


Figure 2. Number of papers $n(t)$ during the pause. Dots are the AI safety papers in **our dataset**.

2.3.8 Reviewer Budget

With N_v verifiers, the labs check the share $q = N_v/W$ of the n papers. This is not a random sample: if verifiers are the bottleneck, we want to check the papers with the most research effort RE_i .

We already said that each paper i holds research r_i of X_e . In a single year, r_i depends only on its compute c_i , which is lognormally distributed with σ_c . We model another characteristic of papers, the quality of their ideas, v_i . It doesn't affect the raw research X_e , but it affects the research effort RE_i , adjusted for idea quality and taste. Think of the transformer paper: it isn't great because of the compute or labor behind it.

Idea quality works like taste. Taste T says what a unit of the labs' research is worth on average, and idea quality says how this paper compares with that average. So a paper's research effort is $RE_i = T v_i r_i$: its research, weighted by the labs' taste and by the quality of its idea. We model v_i with a Pareto distribution (Sinatra et al., 2016; Janosov et al., 2020).

$$\begin{aligned}
q &= \min \left(1, \frac{N_v}{W} \right) \\
v_i &\sim \text{Pareto}(v_m, \alpha_v), \quad \mathbb{E}[v_i] = 1 \\
RE_i &= Tv_i r_i, \quad \sum_{i=1}^n RE_i = TX_e = RE_e
\end{aligned} \tag{30}$$

2.3.9 Top Papers

Verifiers take the papers with the most research effort RE_i first, so they cover the share $\varphi_h(q)$ of the research effort RE_e and hold the share $\psi(q)$ of the exploration compute C_e .

A paper's research effort RE_i is a function of two random variables, v_i and c_i . So it is a random variable itself, and we can derive closed-form formulas for both shares.

A paper's research effort. Idea quality is Pareto, so its logarithm is exponential. Compute is lognormal, so its logarithm is normal. Hence $\ln RE_i$ is a constant plus two independent terms: an exponential with mean $1/\alpha_v$, from the idea, and a normal with standard deviation $\sigma_y = \zeta_p \sigma_c$, from compute. The constant is $\ln K$, where K is what a paper with the weakest idea and the median compute is worth.

Shares. The verifiers check the papers above a bar, u in logarithms above K . $G(x, a)$ is the share of papers above the bar x when the exponential term has rate a . So $q = G(u, \alpha_v)$: the reviewer budget sets q , and q sets the bar. The first term of G counts the papers whose compute alone clears the bar; the second counts those that clear it thanks to their idea.

ψ and φ_h count the same papers, but weight each by its compute or by its research effort. Both weights change the distribution in a simple way. Weighting a normal draw ξ_i by $e^{\delta \xi_i}$ moves it up by δ , and weighting idea quality by itself lowers its tail from α_v to $\alpha_v - 1$.

- Compute weights move ξ_i up by σ_c , which raises $\ln RE_i$ by $\sigma_y \sigma_c$. That's the same as lowering the bar by as much.
- Research-effort weights move ξ_i up by σ_y and fatten the tail of idea quality. So the checked papers hold more of the research effort than of the papers, and the closer α_v is to 1, the more.

$$\begin{aligned}
\chi_i &= \begin{cases} 1, & RE_i \text{ among the top } q \text{ of the papers} \\ 0, & \text{otherwise} \end{cases} \\
\varphi_h(q) &= \frac{\sum_{i=1}^n \chi_i RE_i}{\sum_{i=1}^n RE_i} \\
\psi(q) &= \frac{\sum_{i=1}^n \chi_i c_i}{\sum_{i=1}^n c_i}
\end{aligned} \tag{31}$$

$$\begin{aligned}
\ln RE_i &= \ln K + \frac{\varepsilon_i}{\alpha_v} + \sigma_y \xi_i \\
\varepsilon_i &\sim \text{Exp}(1), \quad \xi_i \sim \mathcal{N}(0, 1), \quad \sigma_y = \zeta_p \sigma_c \\
K &= Tr_s l_\infty v_m \left(\frac{\tilde{c}}{\bar{c}_s} \right)^{\zeta_p}, \quad \tilde{c} = \bar{c} e^{-\sigma_c^2/2}
\end{aligned} \tag{32}$$

$$\begin{aligned}
\chi_i = 1 &\iff \frac{\varepsilon_i}{\alpha_v} + \sigma_y \xi_i \geq u \\
G(x, a) &= \Phi \left(-\frac{x}{\sigma_y} \right) \\
&\quad + e^{-ax + a^2 \sigma_y^2 / 2} \Phi \left(\frac{x}{\sigma_y} - a \sigma_y \right) \\
q &= G(u, \alpha_v) \\
\psi &= G(u - \sigma_y \sigma_c, \alpha_v) \\
\varphi_h &= G(u - \sigma_y^2, \alpha_v - 1)
\end{aligned} \tag{33}$$

2.3.10 Compute Budget

Verifiers reproduce the main experiments with the compute C_r . Since they can skip ablations, random seeds and failed experiments, a re-run takes only the share κ_c of a paper's compute c_i . Above, we spread the failed projects across all the papers, but humans don't need to reproduce them, so a re-run takes $\kappa_c c_i$, not $\kappa_c M c_i$. This lets us express C_e through all of the experiment compute, C_{xpm} .

But ψ itself depends on C_e ⁷. So (34) is one equation in one unknown, C_e , and we solve it numerically by bisection. The answer lies between two bounds:

- If every paper is checked, $\psi = 1$, and the labs explore with $C_{xpm}/(1 + \kappa_c/M)$.
- If no paper is checked, the labs explore with all of C_{xpm} .

We try the midpoint, compute X_e , n , q and ψ there, and check whether exploration plus re-runs fit in C_{xpm} . If they don't, the answer lies below the midpoint; otherwise, above. The labs solve this for every split of their staff they consider and then pick the best split (35).

$$\begin{aligned} C_r &= \kappa_c \sum_{i=1}^n \chi_i c_i = \frac{\kappa_c}{M} \psi C_e \\ C_e \left(1 + \frac{\kappa_c}{M} \psi\right) &= C_{xpm} \end{aligned} \tag{34}$$

2.3.11 Best Split

We allocate humans between research and verification optimally: the labs pick the split with the most verified research effort, RE^v .

$$\max_{N_e} RE^v \tag{35}$$

2.3.12 Unlimited Hiring

Verifiers might become a bottleneck, so we also model an unlimited hiring trajectory. We assume all papers are verified: $\varphi_h = \psi = 1$ and $RE^v = RE_e$. For SAR and TED-AI, all the researchers move to verification, and we hire the rest. AC is a little trickier: we model it as an optimal allocation of the existing staff between verification and research and hire the remaining verifiers.

$$\begin{aligned} N_h &= \max(0, N_e + W - L_H) \\ N_v &= \min(W, L_H - N_e) \end{aligned} \tag{36}$$

$$\begin{aligned} \varphi_h &= \psi = 1 \\ C_e &= \frac{C_{xpm}}{1 + \kappa_c/M}, \quad RE^v = RE_e \end{aligned} \tag{37}$$

2.3.13 Table

See how we estimated the input [parameters](#).

In our model, verifications are mostly successful. AI is selected to do useful AI R&D work. And even if it produces something unreproducible, 2025 X_s research also had unreproducible results, and it's not clear that AI will make their share bigger.

$$\text{cell} = \begin{cases} \text{Full Trust (17),} & E_p < E_{AC} \\ \frac{1}{RS_s} \int_{t_p}^{t_p+D} RE^v dt, & E_p \geq E_{AC} \end{cases} \quad (38)$$

OWN STAFF

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	8,040
SAR Feb 2029	348	1,180	2,770	7,210	50,800
TED-AI Oct 2029	3,980	13,200	30,300	77,300	529,000

UNLIMITED HIRING

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	9,680
SAR Feb 2029	355	1,230	3,010	8,460	81,700
TED-AI Oct 2029	5,320	18,300	44,200	124,000	1,210,000

PEOPLE TO HIRE

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	0	0	0	0	0
AC May 2028	0	0	0	4.66k	108k
SAR Feb 2029	14.2k	31.0k	56.8k	118k	645k
TED-AI Oct 2029	193k	253k	353k	593k	2.74M

On top of the labs' own staff: the most they need at any time in the pause.

Table 4. Only research that humans reviewed and reproduced counts. Gray numbers are Full Trust's: before AC there is nothing to verify.

2.3.14 Limitations

Because of necessary simplifications, our model corresponds to a slightly different protocol.

- In our model, humans verify the most important papers of the year. In reality, research arrives over the year, so at the start of the year they can't know which papers will be the year's best. We think this isn't numerically important. To check it, we additionally modeled the worst case, in which verifiers can't tell anything about the papers' importance and sample them at random. See [Random Checks](#).

2.4 Compartmentalization

Compartmentalization is the [Results Verification](#) model with the changes described in all the subsections below.

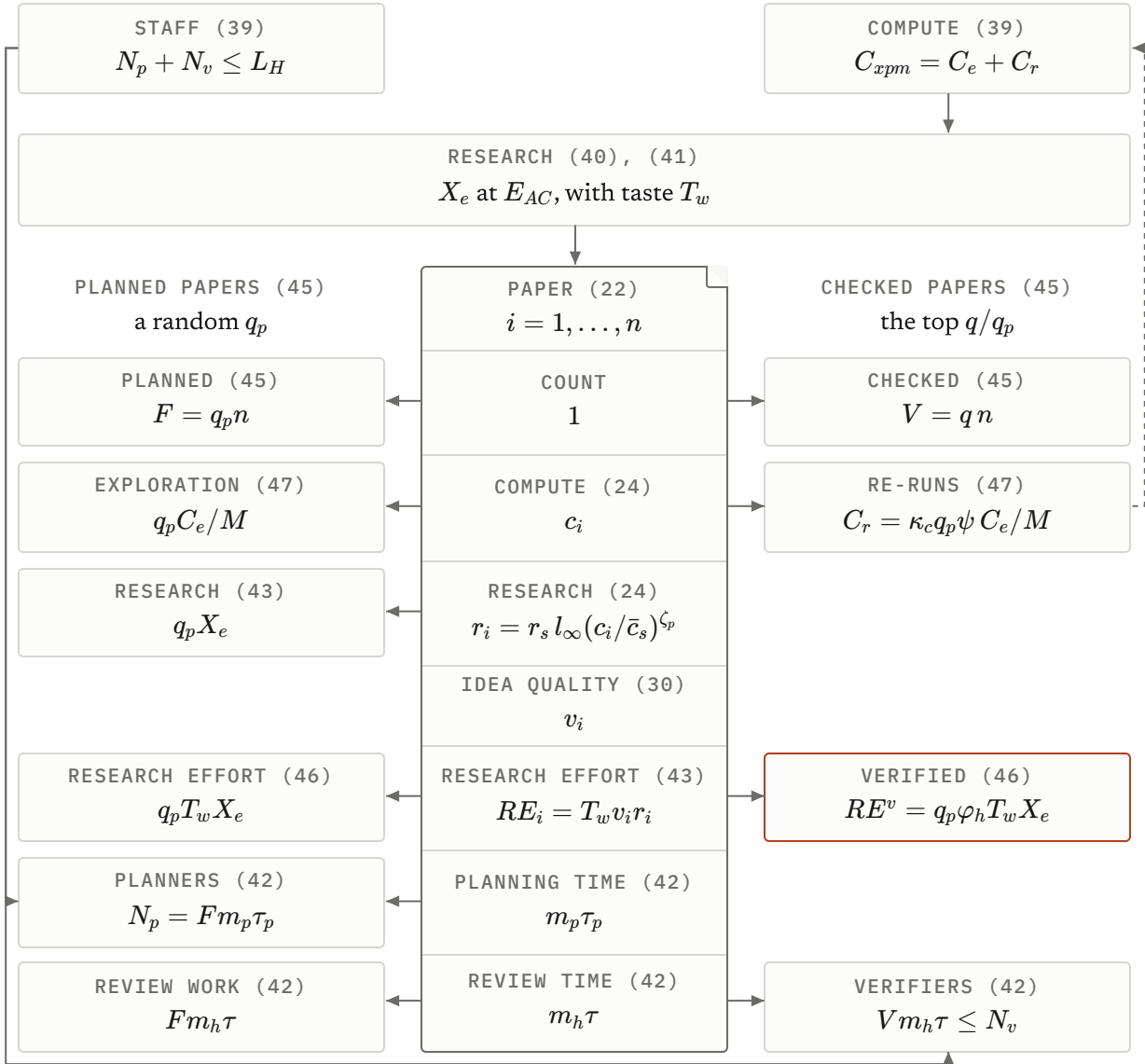


Figure 3. Verifiers check the top q/q_p of the planned papers, so φ_h and ψ are taken at q/q_p . The labs split the staff by (48). With unlimited hiring (49)–(50), every paper is planned and checked.

2.4.1 Protocol

We keep Results Verification and add more verification. It's common to split human staff into researchers who plan experiments and engineers who write the code. We use the same split between a smarter model and a weaker one to contain the smarter model and limit its attack surface⁸:

- The AI researcher is the frontier model, at SAR level or above. It's sandboxed: it can talk to the staff and read the library, but it can't run anything.
- The AI engineer is a model at AC level, the weakest one that can do all the coding on its own. It implements and runs the researcher's plans.
- The human staff no longer research on their own. They lead projects with the researcher or verify papers.

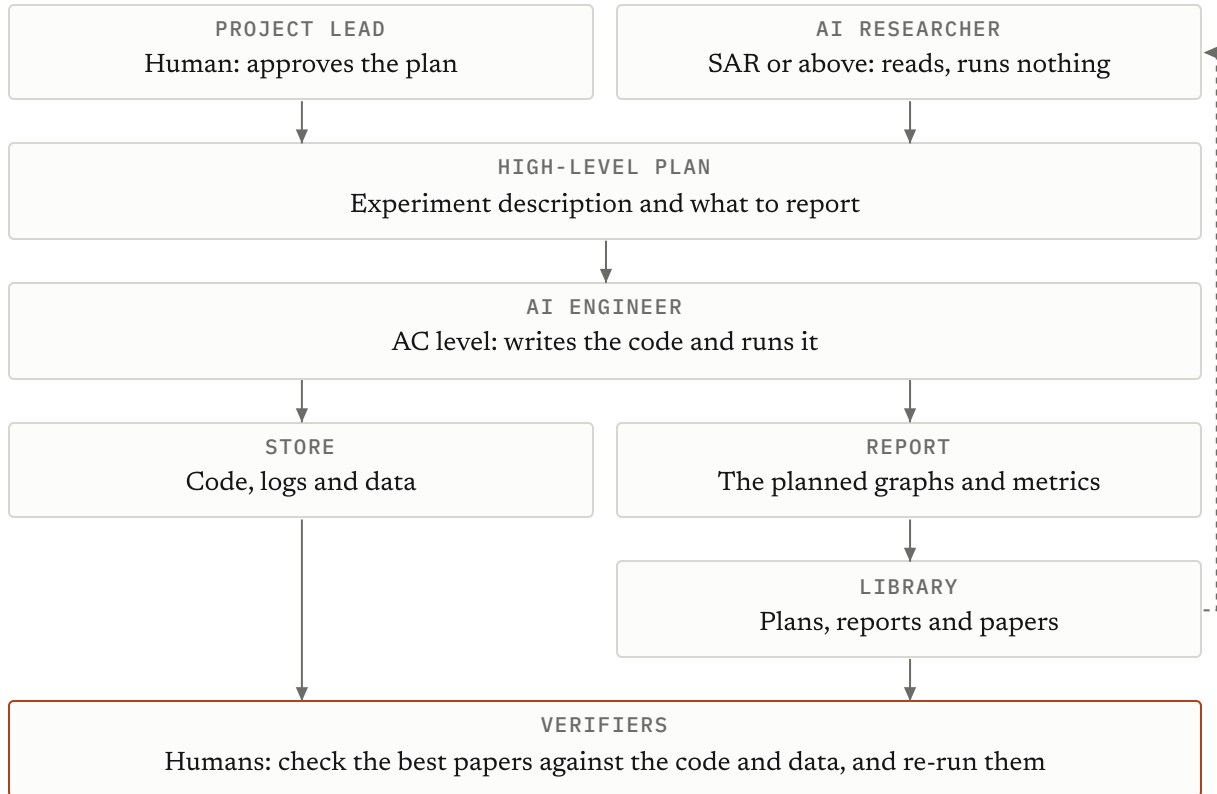


Figure 4. The dashed arrow is the next round. Only the verifiers read the store.

A human project lead discusses ideas with the AI researcher, and the researcher writes a plan: what experiments to run and what to report⁹. The lead reads the plan and approves it, and the AI engineer writes the code and runs it. The preregistered graphs and metrics go into a report, which is saved to the library. The code, logs and data go to a separate store that only human verifiers read¹⁰. Then the lead and the AI researcher choose to either:

- drop the failed project,
- turn it into a paper,
- or do another round of planning and experimenting.

When a paper is ready, nobody has read its code yet. So papers are verified exactly as in Results Verification, and the staff split their time between two jobs: leading rounds and checking papers.

Compared with Results Verification, the labs can lose research effort RE in four ways:

- The AI engineer codes at AC level, far below the frontier model.
- A plan carries only part of the AI researcher's taste, because the engineer makes implementation choices inside it.
- Every round takes some of the lead's time, so less is left for verifying.
- Only approved plans run, so compute sits idle when the staff haven't planned enough rounds.

Compartmentalization is for the SAR and TED-AI rows; the earlier rows are the Results Verification ones. As there, we compute two versions: with the staff the labs had at the start of the pause, and with unlimited hiring of leads and verifiers.

It's a pretty standard idea in the AI safety community, discussed as AI oracles or AI boxing. Here are some papers with similar protocols:

- ACL-5 level in Korbak et al. (2025). The authors argue that no control safety case is possible at this level.
- Factor(U,T): An untrusted strong model decomposes coding tasks, and trusted weaker models implement the pieces (Lue Chee Lip et al., 2025).
- Maybe Scientist AI (Bengio et al., 2025). A non-agentic model proposes the most informative experiments, and humans run them.
- Somewhat resembles Buhl et al. (2025).

2.4.2 Splits

All of the staff split their time: N_p goes to planning rounds with the AI researcher, N_v to verifying papers. Nobody keeps researching¹¹.

$$\begin{aligned} N_e &\rightarrow 0 \\ N_p + N_v &\leq L_H \\ C_{xpm} &= C_e + C_r \end{aligned} \tag{39}$$

2.4.3 Research

The AI engineer does all of the coding, so coding labor is AIFM's AI-only coding labor at the engineer's capability E_{AC} . X_e is what the labs would research if every round they could run were planned.

$$X_e = X(C_e, L_C(0, C_{aut}, E_{AC})) \quad (40)$$

2.4.3.1 Taste

The AI researcher sets the plans, but the AI engineer makes the implementation choices inside them. So the plans' taste T_w is the engineer's, plus the share w of the researcher's advantage over it. Nobody else chooses research, so people's taste drops out: $T(E, 0) = T_{AI}(E)$.

$$\begin{aligned} T(E, 0) &= T_{AI}(E) \\ T_w &= (1 - w) T_{AI}(E_{AC}) + w T_{AI}(E_p) \end{aligned} \quad (41)$$

2.4.4 Plans and Checks

Not every research plan leads to a paper. Sometimes you need to iterate after getting experimental results, and sometimes you just decide to drop the project altogether. A paper takes m_p rounds of planning on average, counting the rounds of dropped projects. Each round takes τ_p of the human lead's time. So planning labor N_p makes F papers a year, not more than the n papers the labs could run. A check takes m_h verifiers τ each, so verifying labor N_v checks V papers a year, not more than the F planned ones. Both times grow with F , the papers the labs actually write: only those are prerequisites to study.

$$\begin{aligned} F &= \min \left(n, \frac{N_p}{m_p \tau_p} \right), & V &= \min \left(F, \frac{N_v}{m_h \tau} \right) \\ m_p \tau_p &= \kappa_p t_s + p_p \log_{10} \frac{F}{n_s}, & \tau &= \kappa_h t_s + p_h \log_{10} \frac{F}{n_s} \end{aligned} \quad (42)$$

2.4.5 Verified Research

In Results Verification, we assumed people can quickly understand which papers are worth the most and verify them first. We don't assume the same for plans. It's harder to say which plans will result in important papers. So the planned papers are a random share of what the labs could write: the same average research $\bar{r}$, the same spread of idea quality and of compute, all with the plans' taste T_w .

Verifiers check the V of the F papers worth the most. They choose from the F planned papers, not all n , but these are spread the same way, so the checked ones hold the share $\varphi_h(V/F)$ of the planned papers' research effort.

$$\begin{aligned}
RE^v &= \varphi_h \left(\frac{V}{F} \right) F T_w \bar{r} \\
\bar{r} &= \frac{X_e}{n}
\end{aligned} \tag{43}$$

2.4.6 Planning Work

W_p is the staff time needed to plan every round the labs could run, derived like W from the same paper count. As with review time, a paper's planning time doesn't depend on its compute: a bigger paper takes the same number of rounds, each directing more compute. And as with review time, the lead has to study the prerequisites: every tenfold more papers adds p_p FTE-years to a paper's planning. Then q_p is the share of the labs' rounds that get planned, and q the share of their papers that get checked, as in Results Verification, only now at most q_p : unplanned rounds write no papers. The verifiers check the top q/q_p of the planned papers, so verified research effort is that of Results Verification, applied to the rounds that ran.

Planned papers are a random share of the labs' papers, so their idea quality and compute are distributed the same way, and we can use the closed form (33), with the bar u set by q/q_p instead of q .

$$W_p = n m_p \tau_p |_{F=n} = N_s \frac{n}{n_s} \left(\kappa_p + \frac{p_p}{t_s} \log_{10} \frac{n}{n_s} \right) \tag{44}$$

$$\begin{aligned}
q_p &= \frac{F}{n}, & q_p = 1 &\iff N_p \geq W_p \\
q &= \frac{V}{n} = \min \left(q_p, \frac{N_v}{n m_h \tau} \right)
\end{aligned} \tag{45}$$

$$\begin{aligned}
RE^v &= q_p \varphi_h \left(\frac{q}{q_p} \right) T_w X_e \\
G(u, \alpha_v) &= \frac{q}{q_p} \\
\varphi_h &= G(u - \sigma_y^2, \alpha_v - 1), & \psi &= G(u - \sigma_y \sigma_c, \alpha_v)
\end{aligned} \tag{46}$$

2.4.7 Compute Budget

As in Results Verification, the labs re-run exactly the checked papers, at κ_c/M of the compute each takes. They hold the share $\psi(q/q_p)$ of the planned papers' compute, which is the share q_p of the labs' exploration compute C_e . The compute of unplanned rounds stays idle.

$$\begin{aligned}
C_r &= \frac{\kappa_c}{M} q_p \psi \left(\frac{q}{q_p} \right) C_e \\
C_e \left(1 + \frac{\kappa_c}{M} q_p \psi \left(\frac{q}{q_p} \right) \right) &= C_{xpm}
\end{aligned} \tag{47}$$

2.4.8 Planning vs Verifying

Nobody keeps researching, so the labs distribute humans between planning and paper verification. We model it in an optimal way.

$$\max_{N_p} RE^v \tag{48}$$

2.4.9 Unlimited Hiring

As in Results Verification, the labs can hire the people they lack: $q_p = q = 1$. All of their own staff plan and verify, and they hire the rest, so planning costs only headcount.

$$N_h = \max(0, W_p + W - L_H) \tag{49}$$

$$\begin{aligned}
q_p &= q = 1 \\
C_e &= \frac{C_{xpm}}{1 + \kappa_c/M}, \quad RE^v = T_w X_e \\
N_p + N_v &= \min(L_H, W_p + W)
\end{aligned} \tag{50}$$

2.4.10 Table

See how we estimated the input [parameters](#).

$$\text{cell} = \begin{cases} \text{Results Verification (38),} & E_p < E_{SAR} \\ \frac{1}{RS_s} \int_{t_p}^{t_p+D} RE^v dt, & E_p \geq E_{SAR} \end{cases} \tag{51}$$

OWN STAFF

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	8,040
SAR Feb 2029	137	471	1,130	3,140	24,100
TED-AI Oct 2029	939	3,180	7,540	20,500	156,000

UNLIMITED HIRING

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	9,680
SAR Feb 2029	137	471	1,130	3,140	29,000
TED-AI Oct 2029	939	3,180	7,540	20,500	183,000

PEOPLE TO HIRE

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	0	0	0	0	0
AC May 2028	0	0	0	4.66k	108k
SAR Feb 2029	0	0	0	0	131k
TED-AI Oct 2029	0	0	0	0	144k

On top of the labs' own staff: the most they need at any time in the pause.

Table 5. Only research that humans planned, reviewed and reproduced counts. Gray numbers are Results Verification's: before SAR there is nothing to compartmentalize. The main page shows this table.

2.4.11 Limitations

In addition to the limitations of Results Verification:

- If we were bottlenecked by planning (we aren't with our input parameters), some compute would sit idle. Real labs would use it, but it's hard to say how it would change the papers' compute c_i . We think that with other input parameters this might matter, and the model would then understate RE .
- The AI engineer gets all of C_{aut} , so we don't count the AI researcher's own compute. It may take a few percent of C , but we don't think that changes the results much.
- The project lead always approves the plan. In reality, a lead who sees something unsafe would ask the AI researcher to change the plan, or would edit it themselves. We think it doesn't change the results much.

2.5 Workload Approval

Workload Approval is the **Compartmentalization** model with the changes described in all the subsections below.

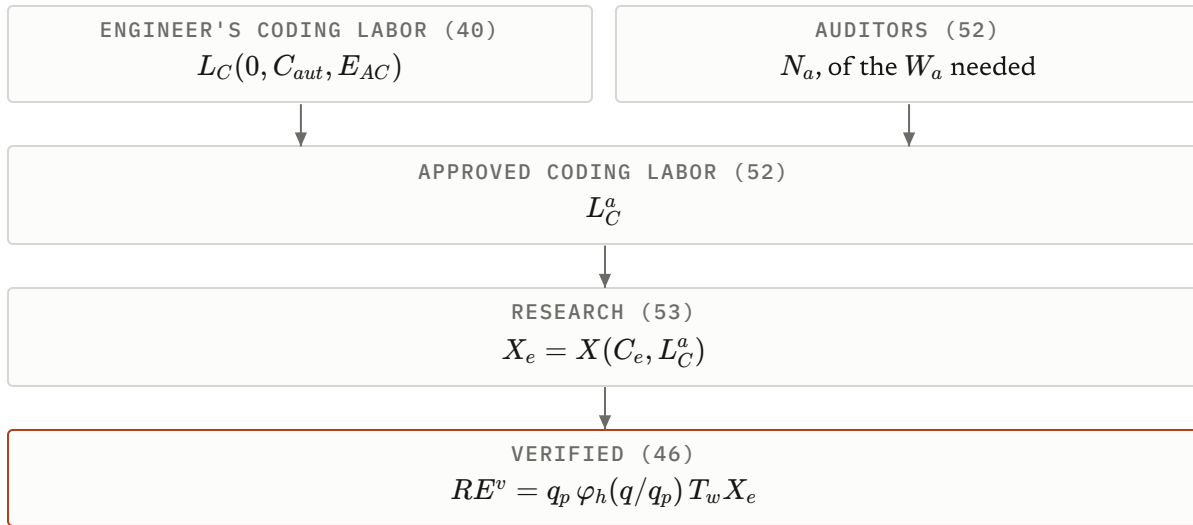


Figure 5. From the research on, everything is as in Compartmentalization (41)–(50).

2.5.1 Protocol

Besides AI takeover, there are also risks from concentration of power and misuse. To reduce all of them, AI companies in AI 2040 declare each workload with its full code, and teams of auditors approve it before it runs, so that countries can trust each other's labs. When we started this

project, one of the questions we wanted to answer was how much slower full workload approval would be.

We aim to model Plan A, with some necessary changes: we simplify some things and fill in some gaps in the original plan.

Specifically:

- We modify the Compartmentalization protocol.
 - We think something like it is realistic in practice.
 - It simplifies our model description, so this section can be short.
- N_a auditors read and approve the code, which decreases coding labor L_C .

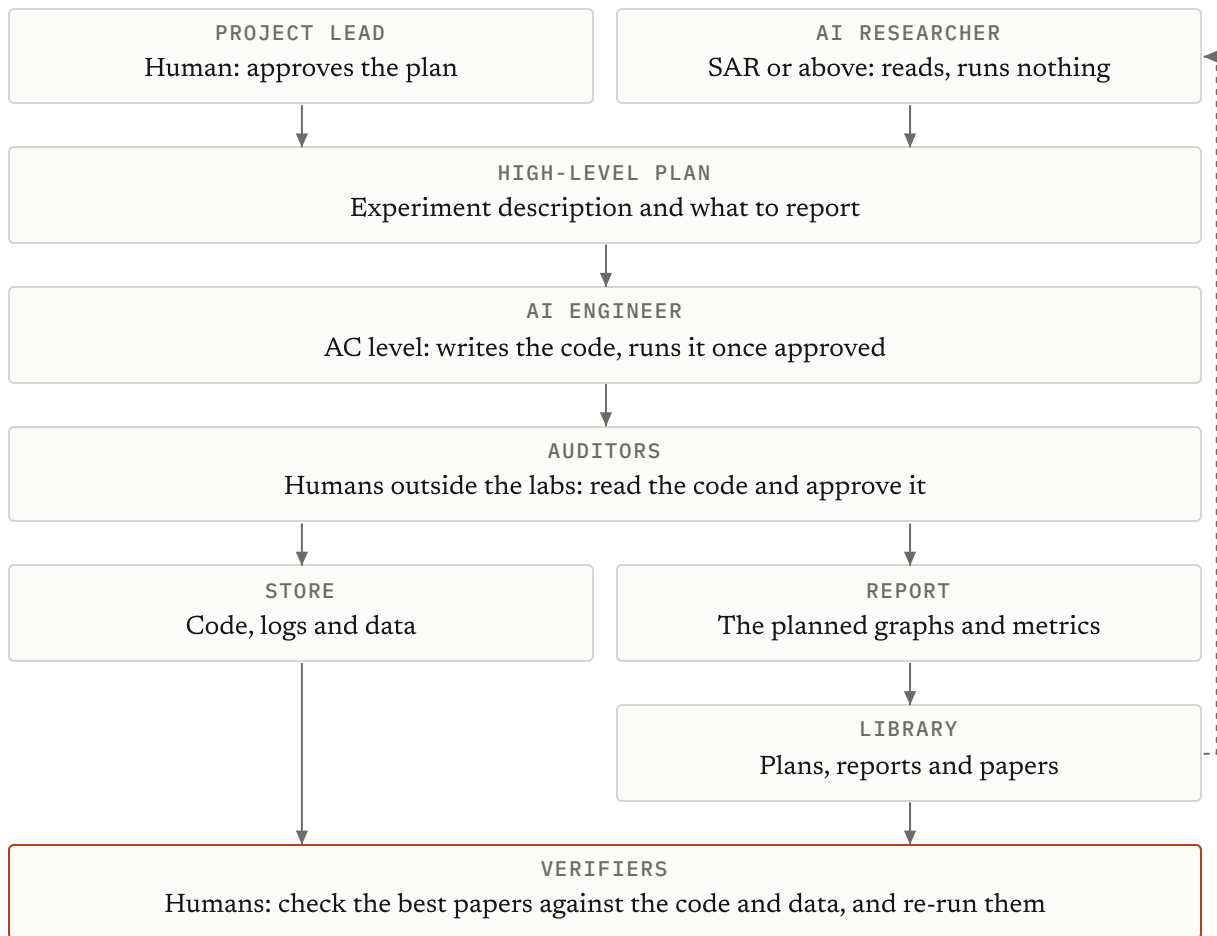


Figure 6. The dashed arrow is the next round. Only approved code runs, and only the verifiers read the store.

It's unclear whether results still need to be verified after these checks. When we removed verification from the simulation, research effort rose by ~3%, because results verification is no longer a bottleneck in this scenario. We kept it for brevity.

As in AI 2040, the auditors are a separate pool of N_a people outside the labs. Compared with Compartmentalization, the labs lose research effort in one more way: they run only as much code as the auditors can read. We calculate the final table for a fixed N_a , but see [this section](#) for an analysis of the number of auditors.

Workload Approval applies to the SAR and TED-AI rows only.

2.5.2 Approval

Let κ_a be an auditor's time to approve code, relative to the time it took to write it: approving the code that one FTE of coding labor writes in a year takes κ_a of an auditor's year. So approving everything the AI engineer writes takes $W_a = \kappa_a L_C$ auditors.

N_a auditors approve N_a / κ_a of coding labor. Code that nobody approved doesn't run, so the approved coding labor L_C^a is the smaller of the two.

$$\begin{aligned} W_a &= \kappa_a L_C(0, C_{aut}, E_{AC}) \\ L_C^a &= \min \left(L_C(0, C_{aut}, E_{AC}), \frac{N_a}{\kappa_a} \right) \end{aligned} \quad (52)$$

2.5.3 Research

$$X_e = X(C_e, L_C^a) \quad (53)$$

2.5.4 Unlimited Hiring

As in Compartmentalization, the labs hire the leads and verifiers they lack. They don't hire auditors: they are a separate constant pool.

2.5.5 Table

See how we estimated the input [parameters](#).

$$\text{cell} = \begin{cases} \text{Compartmentalization (51)}, & E_p < E_{SAR} \\ \frac{1}{RS_s} \int_{t_p}^{t_p+D} RE^v dt, & E_p \geq E_{SAR} \end{cases} \quad (54)$$

OWN STAFF

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	8,040
SAR Feb 2029	56.1	186	423	1,050	6,150 [†]
TED-AI Oct 2029	354	1,160	2,600	6,350	36,200

UNLIMITED HIRING

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	9.11	32.8	82.0	227	1,490
AC May 2028	39.2	144	360	1,020	9,680
SAR Feb 2029	56.1	186	423	1,050	6,150
TED-AI Oct 2029	354	1,160	2,600	6,350	36,200

PEOPLE TO HIRE

AI CAPABILITY AT THE START	PAUSE DURATION				
	3 MONTHS	9 MONTHS	1.5 YEARS	3 YEARS	9 YEARS
pre-AC Jun 2027	0	0	0	0	0
AC May 2028	0	0	0	4.66k	108k
SAR Feb 2029	0	0	0	0	0
TED-AI Oct 2029	0	0	0	0	0

On top of the labs' own staff: the most they need at any time in the pause.

[†] Notice how the more capable model, SAR, has fewer AI safety years than the less capable AC. That happens because, at AC, the model isn't capable enough for our verification protocol, so we instead use it as an engineer with human researchers in the lead. We still need to verify the final results (thus Results Verification), but here we allow AI to run arbitrary code without prior human approval. It's not a recommendation, but one of the possible protocols. [See the full model](#)

Table 6. Only research that humans planned, approved, reviewed and reproduced counts. Gray numbers are Results Verification's: before SAR there is nothing to compartmentalize.

2.5.6 Number of Auditors

We don't know how many auditors a pause would have. The main table uses a big but realistic 20,000 auditors (see [rationale](#)).

We simulated the pause with different numbers of auditors. They remain the biggest bottleneck, reducing the research stock to 23–41% of Compartmentalization's.

2.5.7 Limitations

In addition to the limitations of Compartmentalization:

- For simplicity, we keep the labor ceiling l_∞ . Every paper still has all the coding labor it can use. It made sense for Compartmentalization but may make less sense for Workload Approval. You would probably prefer simpler, shorter code to speed up approval. In our model, that would reduce the number of papers to verify and increase their research r_i . It's also likely unimportant, as this model is bottlenecked by auditors reading the code, not by paper verifiers.
- Approval takes time. Every run, even a debugging one, waits for an auditor. The labs can work on other projects meanwhile, but each paper takes longer. In our model, rounds take no time.

2.6 Model Limitations

- **The pause itself.** Our model assumes a total freeze on capabilities R&D and a full redirection of R&D efforts to AI safety from the first day. We don't discuss how to achieve that.
- **Only R&D-like safety work.** We count only the AI safety work that looks like normal AI R&D. But there are more problems to solve: philosophy, agent foundations, governance, hardening the world. AI labor and compute won't help with those as much.
- **Work, but not safety.** A thousand years of AI safety work won't reduce risk a thousand times as much as 2025's research did. We don't model the returns of safety work in terms of risk.
- **Serial time matters.** Even if we could run 1000 years' worth of computational experiments, we would still benefit from more time to think carefully.

2.7 Extend It

Here are some ideas for how these models can be extended.

- Add your favorite AI control frameworks and protocols ([Greenblatt, 2025](#)).
- Add debates (Irving et al., 2018; Irving & Askill, 2019); see also [Michael et al., 2023](#), [Khan et al., 2024](#) and [Elasky et al., 2026](#). They require more human time.

- Add mixed deployment (Nardo, 2025; Nardo, 2026; Finnveden, 2026).
- Defend against persuasion and bribery ([Dassanayake et al., 2025](#)): add more people?

2.8 Additional Links

See how we estimated the input parameters [here](#). See the sensitivity analysis [here](#). To estimate several parameters, we collected a dataset of 2020–2025 AI safety papers. See it [here](#).

3 Implications

If you plan to elicit useful work from AI during the pause, you face a trade-off ([Shlegeris, 2026](#)):

- You can elicit more useful work if you are willing to take more risk.
- You can reduce AI takeover / concentration of power / misuse risks if you are willing to elicit less work.

What is the Pareto-optimal setup? How do you make decisions about the trade-off between risk and usefulness?

We developed our main contribution, the AI Pause model, to answer these questions. We are going to use it to come up with optimal setups in future work, but it already gives interesting results.

3.1 Eventually, the Bottleneck Is People

We don't know how many [verifiers](#) the labs would have during a pause, so we model two hiring strategies: the labs keep the staff they had at the start of the pause, or they hire as many verifiers as they need.

With Compartmentalization, the labs' own staff keep up for about three and a half years. By the end of a nine-year pause, they need ~150k more verifiers. That's seven NeurIPS reviewer pools, reading full-time.

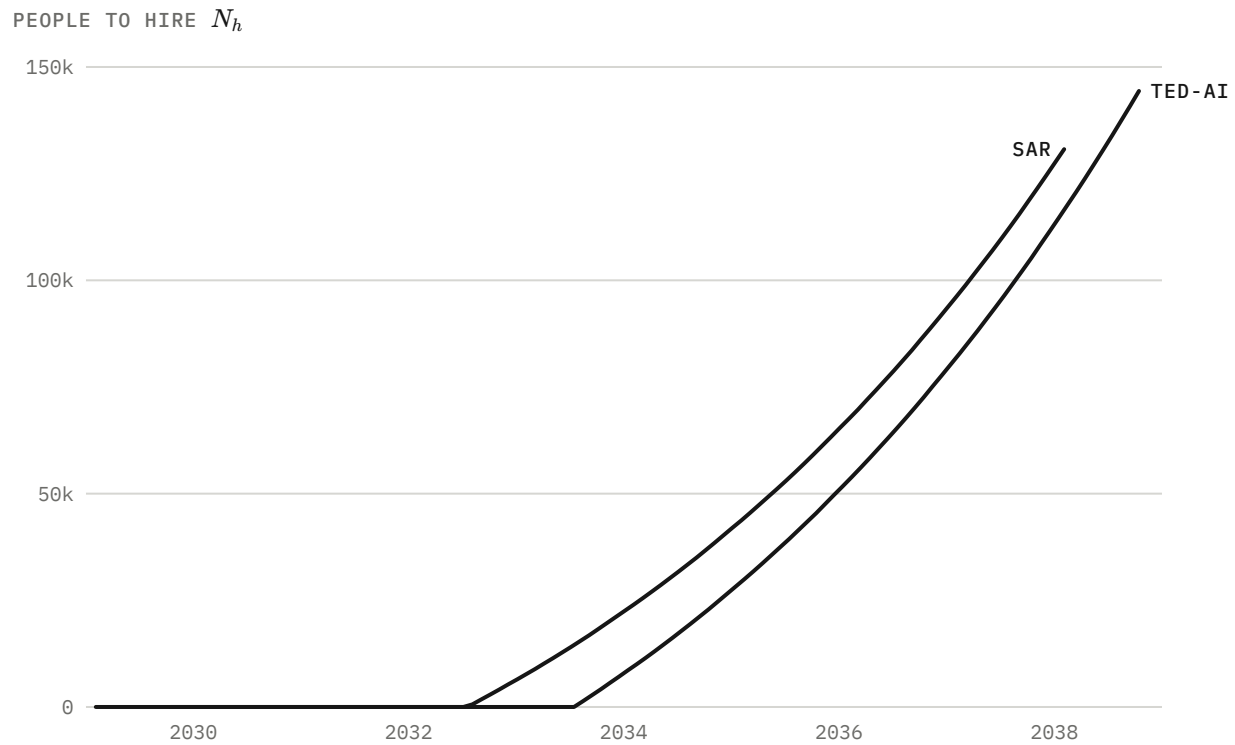


Figure 7. Compartmentalization. The labs hire people to plan and check every paper. It applies from SAR on.

But do we need all of them? Humans review the most important papers first. In the most extreme scenario (Results Verification, TED-AI, nine years), hiring all 2.7 million yields only 129% more AI safety research. For shorter pauses and lower capability levels, it changes less. Hiring an additional 130k verifiers in the Compartmentalization scenario (SAR, nine years) adds only 20% more research.

SCENARIO RESULTS VERIFICATION, TED-AI, 9 YEARS

AI SAFETY YEARS

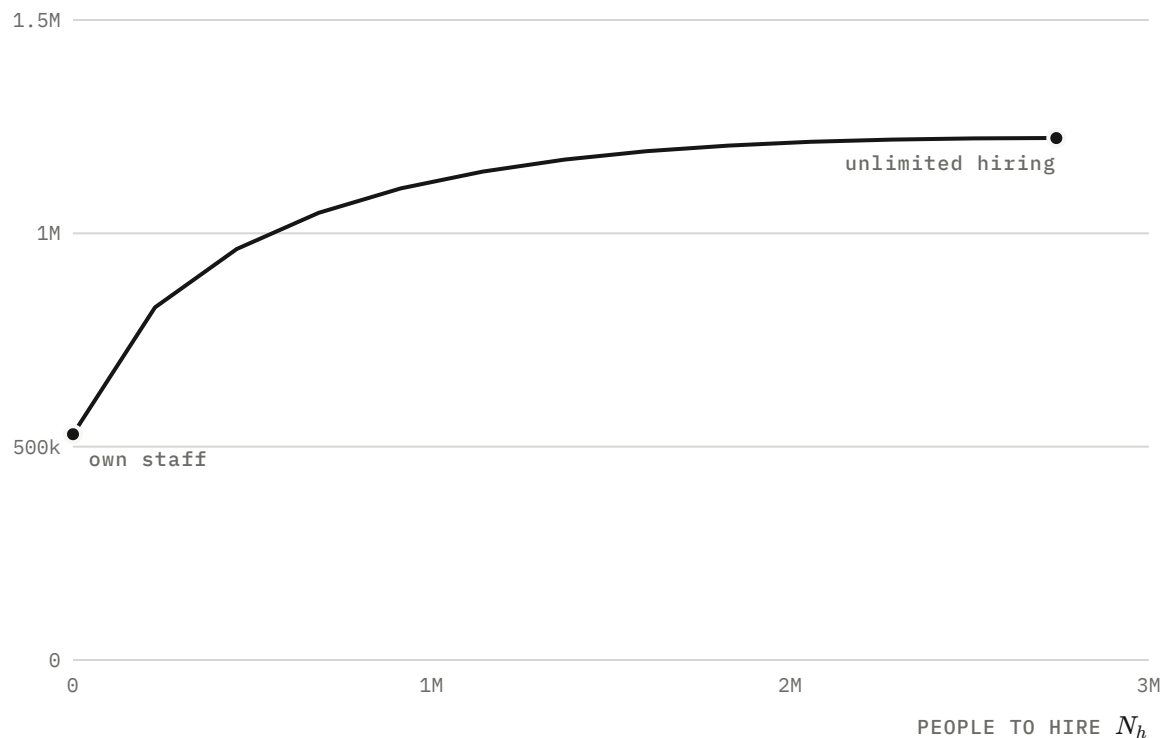


Figure 8. AI safety years when the labs can hire at most this many people on top of their own staff.

This relies on verifiers telling the most important papers apart. If they can't and check a random sample instead, a nine-year Compartmentalization pause from SAR loses almost 30% of its AI safety years, and to match verifiers who pick the best papers, the labs have to hire 43,000 people. See [Random Checks](#).

However, if you add external auditors who manually approve all code, they quickly become the bottleneck. In Workload Approval, with 20,000 auditors who review code as fast as Google's engineers do, a three-year pause keeps 33% of Compartmentalization's AI safety years at SAR and 31% at TED-AI. The verifiers stop being the bottleneck, because the labs write so few papers that their own staff check all of them.

PAUSE 3 years

AI SAFETY YEARS

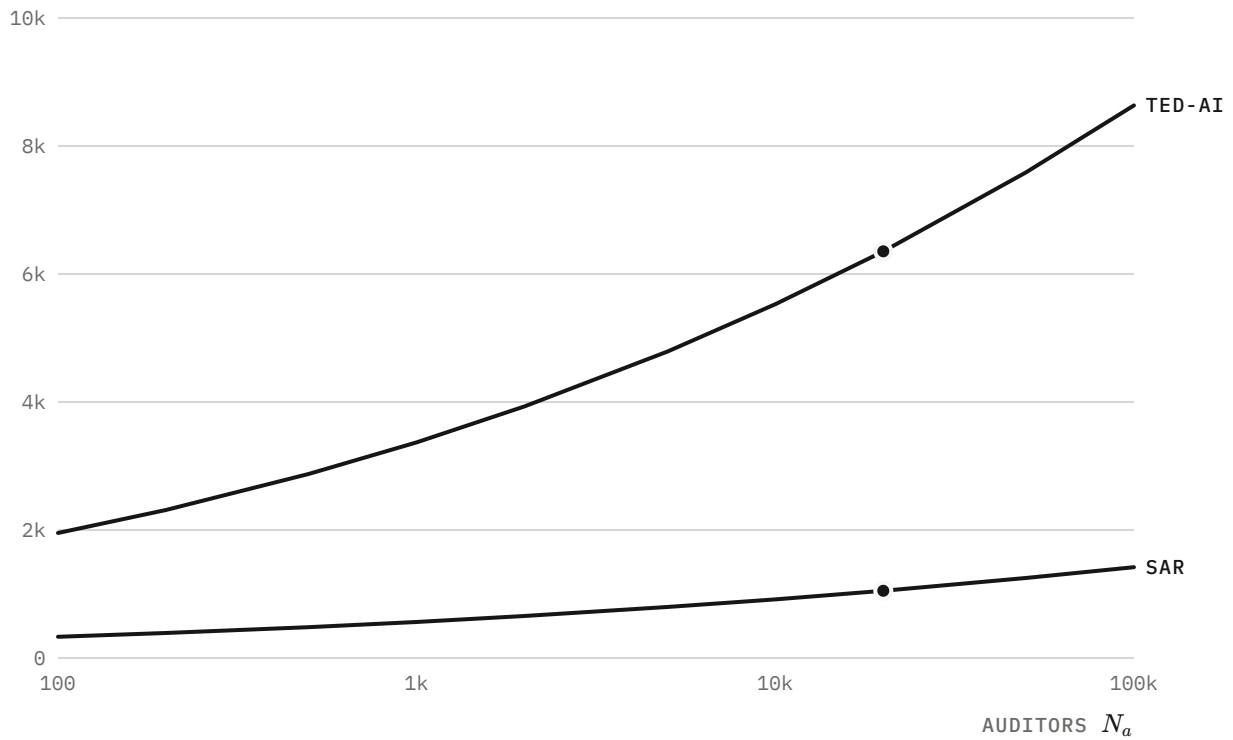


Figure 9. AI safety years with N_a auditors. Dots mark the Workload Approval table's 20,000 auditors.

See the details of the [hiring analysis](#).

3.2 Capability Scaling Strategy

AI 2040 [proposes](#) scaling AI up to the max-controllable AI, which it guesses is roughly TED-AI, and then slowing down nearly to a halt (Lifland, 2026). How fast should we scale?

- If you scale slowly:
 - Less risk of AI takeover. It's hard to say where exactly the max-controllable AI is, so moving carefully lets you evaluate models better.
 - Less algorithmic progress. We can improve capabilities $E(t)$ using either more compute $C(t)$ or better algorithms $S(t)$. It's much easier to control compute, because the GPU supply chain is so complex and concentrated. Algorithmic progress is difficult to control and likely to leak to uncontrollable illegal actors.
 - More time to prepare. Work on other risks, such as preventing concentration of power by lab CEOs or hardening the world, can benefit from more time.
- If you scale fast:
 - You have more time with smarter AI. This matters if you expect the pause to end soon.

Our model allows us to price the AI safety research you lose with slow scaling. Imagine two scenarios:

- **Full Speed.** Labs race from AC (05.2028) to TED-AI (10.2029)¹², then the pause begins. It ends k years later, when the labs hand off to the AI.
- **Plan A.** The pause starts earlier, at AC, but its first years are dedicated to careful scaling. The labs reach TED-AI L years after the Full Speed labs and only then start researching AI safety.

The Plan A labs miss the Full Speed labs' first L years at TED-AI, so they have to hand off later to do as much safety research. How much later? Call this the handoff delay Δt .

In AI 2040's Plan A, the Consortium scales AI carefully for the first five years and then has five years at TED-AI. Our timelines are a bit earlier, because we use the AIFM preset, but we can still model a similar scenario. With Workload Approval, Plan A's scaling seems to produce as much AI safety research as if the Consortium had rushed to TED-AI and then researched there for 7.4 years. In calendar time, Plan A would hand off 1.27 years later, but it might also require an earlier pause, to control the scaling from AC on.

This is an upper bound on how bad Plan A's scaling is, because we assume no AI safety research during the first five years of careful scaling. See the [interactive calculator](#).

3.3 Sensitivity Analysis

For sensitivity analysis, we draw tornado plots. We fix all the parameters at their point estimates, move one parameter at a time to its 10th or 90th percentile and see how this changes AI safety work or the number of verifiers to hire.

We do this for only one scenario, the one we think is the most interesting: Compartmentalization, 9 years, at SAR. For AI safety work, we use frozen hiring.

3.3.1 Our Inputs

We set their 10th and 90th percentiles ourselves. RS_s is computed from C_s and N_s , so it moves with them. For the rationale, see [parameters](#).

■ INPUT AT ITS 10TH PERCENTILE ■ AT ITS 90TH

AI SAFETY YEARS, FROZEN HIRING

p_h time a check gains for every
tenfold more papers

0.002 – 0.05

ζ_p return of a paper's research to its
compute

0.21 – 0.35

p_p time a paper's planning gains for
every tenfold more papers

0.001 – 0.02

l_∞ research of a paper with all the
labor it can use

1.1 – 3

κ_h time to check a paper, relative to
making it

0.007 – 0.1

m_L merged labs' staff

1.8 – 4

m_h verifiers per paper

1 – 5

m_C merged labs' compute

1.8 – 3.1

N_s researchers of 2025 AI safety

1,900 – 8,000

C_s experiment compute of 2025 AI
safety

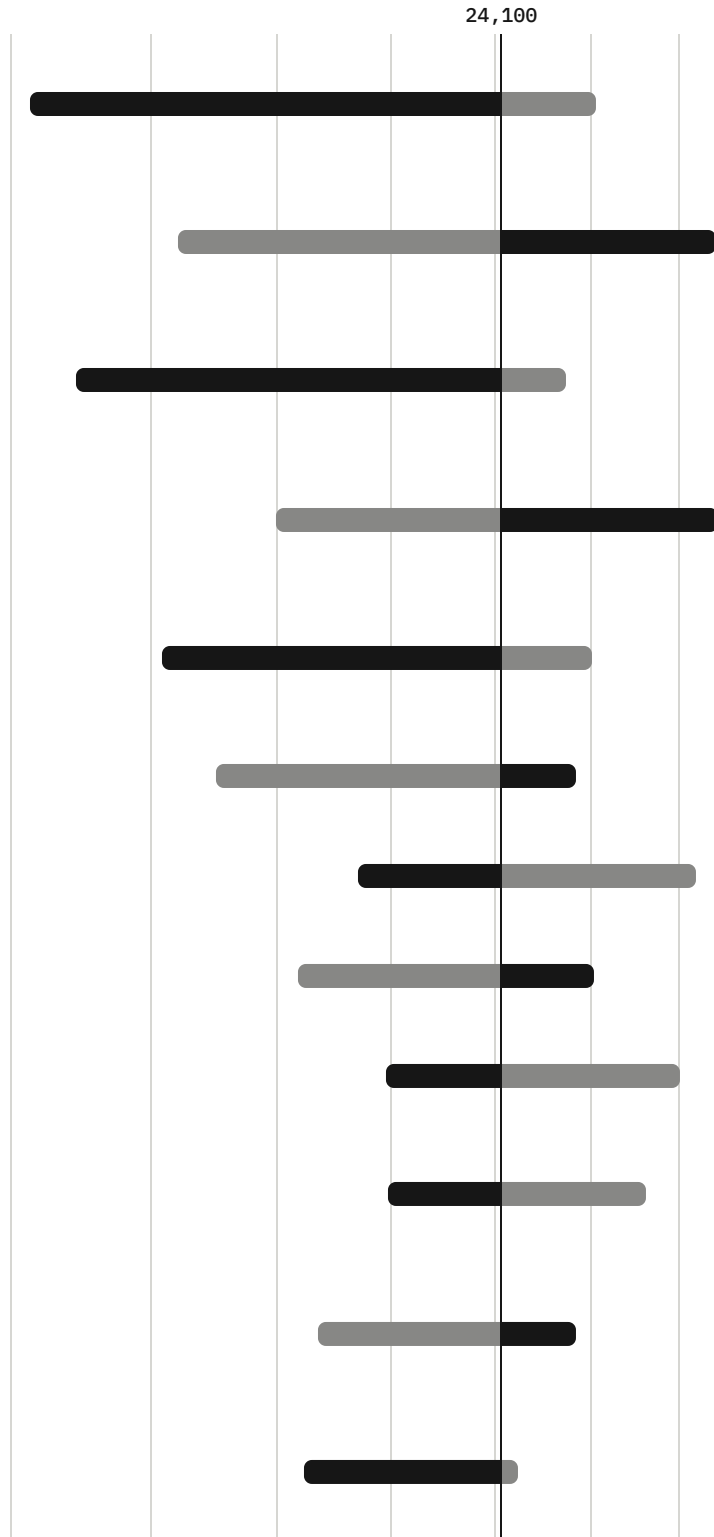
20,000 – 35,000

w share of the researcher's taste that
its plans carry

0.7 – 0.99

κ_p time to plan a paper, relative to
making it

0.0005 – 0.025



α_v tail index of the papers' idea quality

1.35 – 1.8

n_s papers of 2025 AI safety

7,200 – 10,000

κ_c compute to re-run a paper, relative to making it

0.03 – 0.3

s_{ext} share of compute serving external users

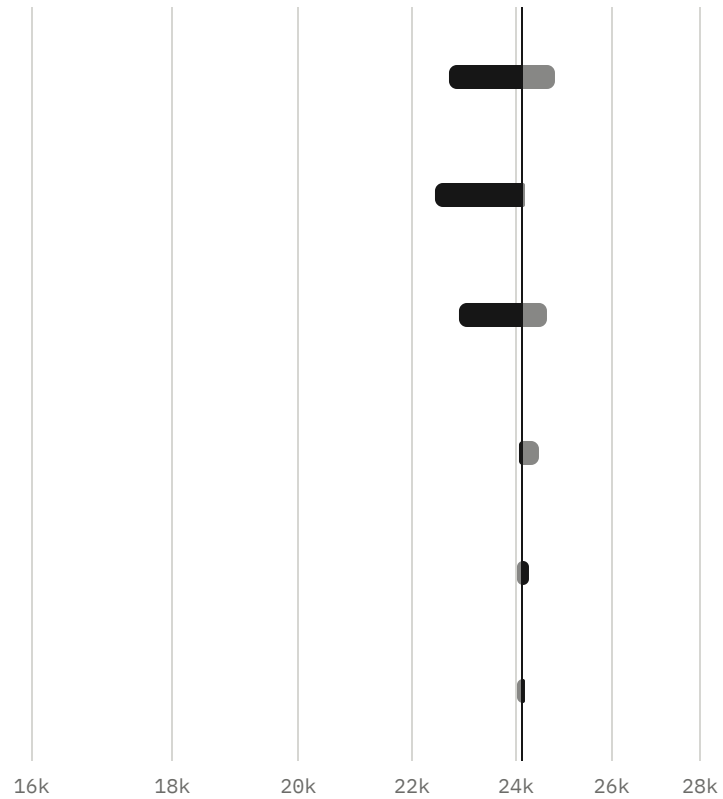
0.1 – 0.45

σ_c spread of compute across papers

2.52 – 2.77

M compute of all projects, relative to papers

1.01 – 1.25



No effect: κ_a .

PEOPLE TO HIRE, UNLIMITED HIRING

l_∞ research of a paper with all the labor it can use

1.1 – 3

ζ_p return of a paper's research to its compute

0.21 – 0.35

m_h verifiers per paper

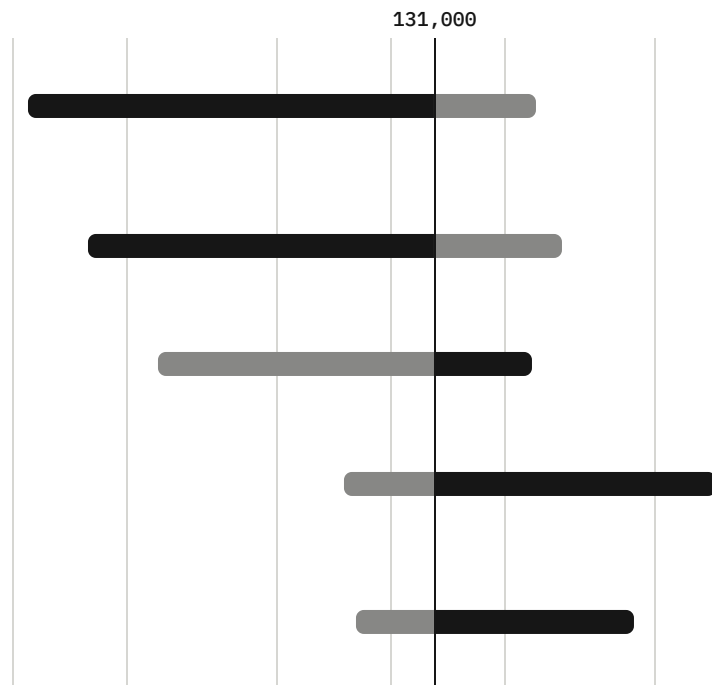
1 – 5

p_h time a check gains for every tenfold more papers

0.002 – 0.05

κ_h time to check a paper, relative to making it

0.007 – 0.1



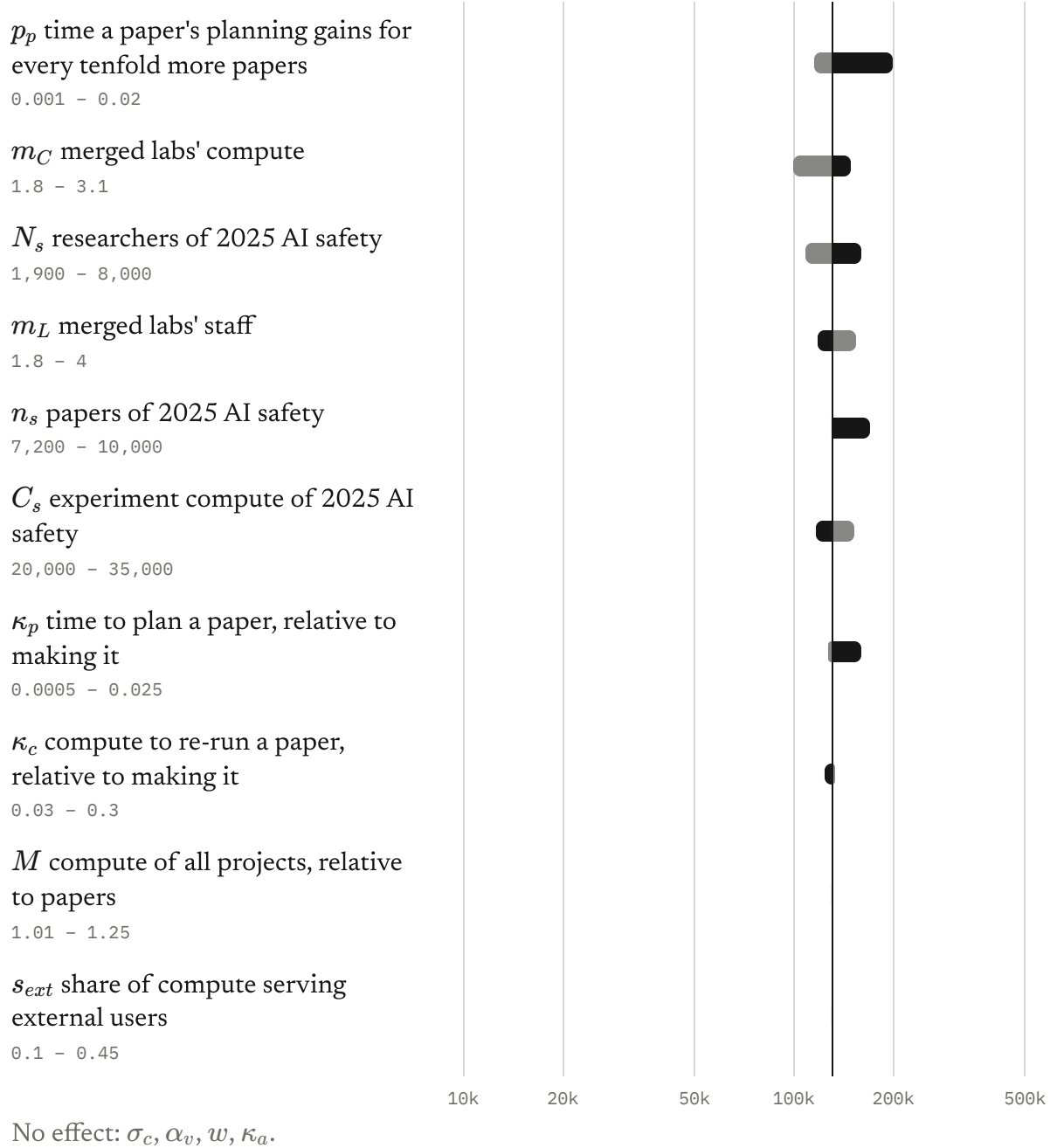


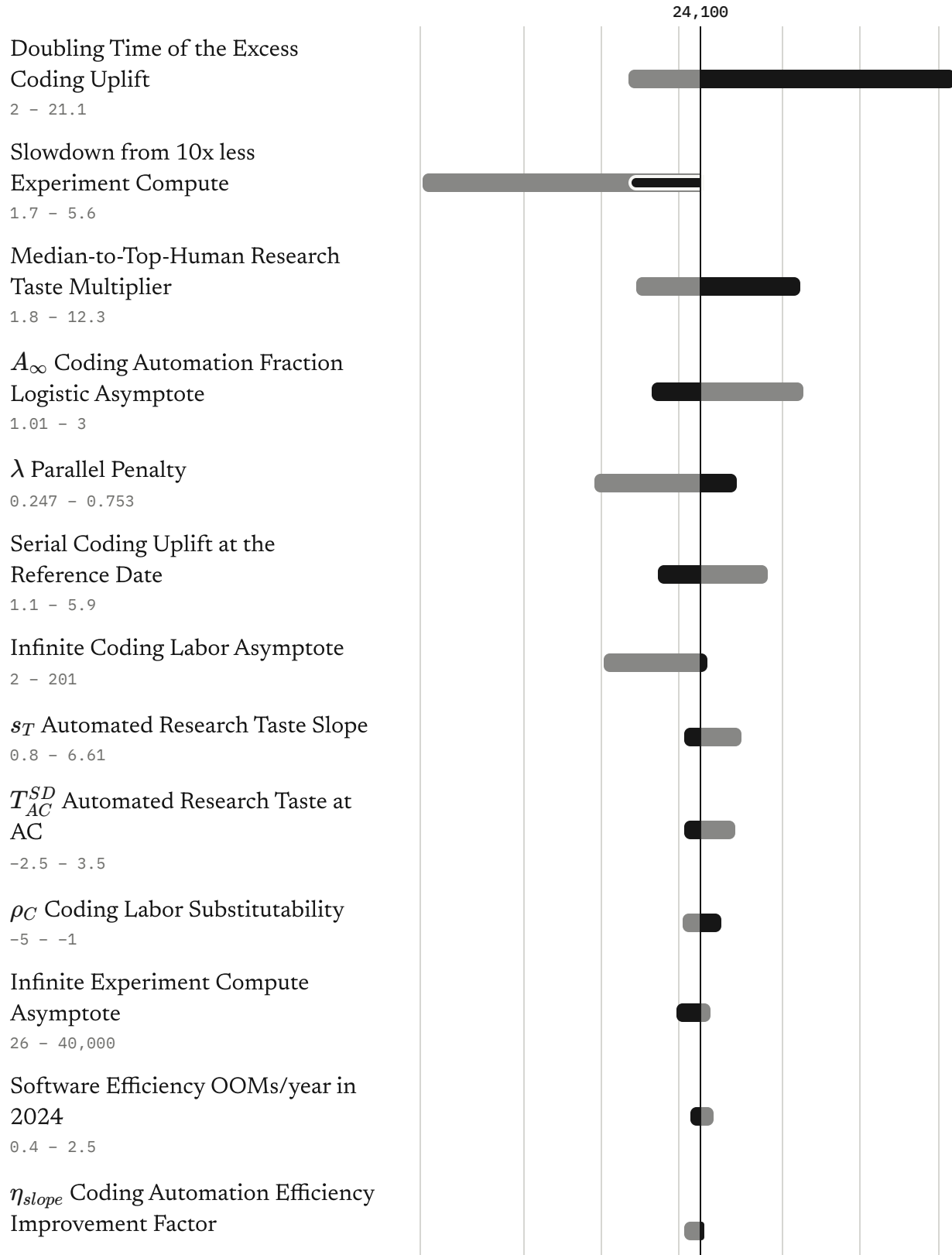
Figure 10. Our inputs, one at a time at their 10th and 90th percentiles. Under each name: its 10th and 90th percentiles.

3.3.2 AIFM Parameters

We take Eli Lifland's AIFM preset and his point estimates.

■ INPUT AT ITS 10TH PERCENTILE ■ AT ITS 90TH

AI SAFETY YEARS, FROZEN HIRING



1 - 9

Maximum Research Taste

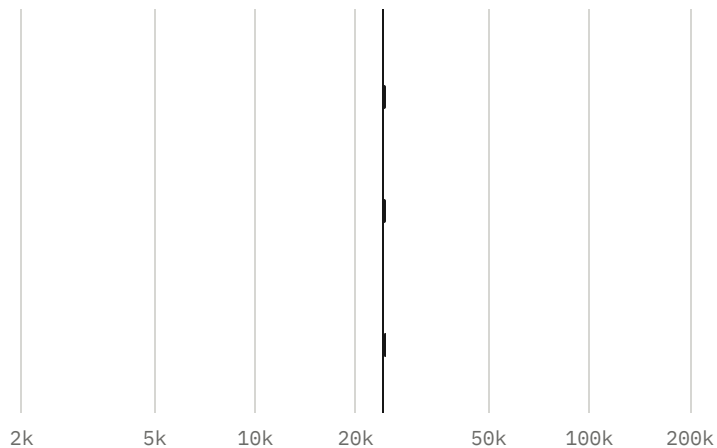
2 - 32

Research Taste Slowdown Factor
Halfway to Algorithmic Limit

0.196 - 0.804

Maximum Serial Coding Labor
Multiplier

$10,000 - 10^{12}$



No effect: AC Time Horizon Requirement, Doubling Difficulty Growth Factor, Size of Effective Compute Gap, Present Doubling Time, Present Revenue Growth Rate, Revenue Required for AC, Median-to-top-human jumps above SAR needed to reach TED-AI.

PEOPLE TO HIRE, UNLIMITED HIRING

Slowdown from 10x less
Experiment Compute

1.7 - 5.6

λ Parallel Penalty

0.247 - 0.753

Infinite Coding Labor Asymptote

2 - 201

Infinite Experiment Compute
Asymptote

26 - 40,000

Doubling Time of the Excess
Coding Uplift

2 - 21.1

A_{∞} Coding Automation Fraction
Logistic Asymptote

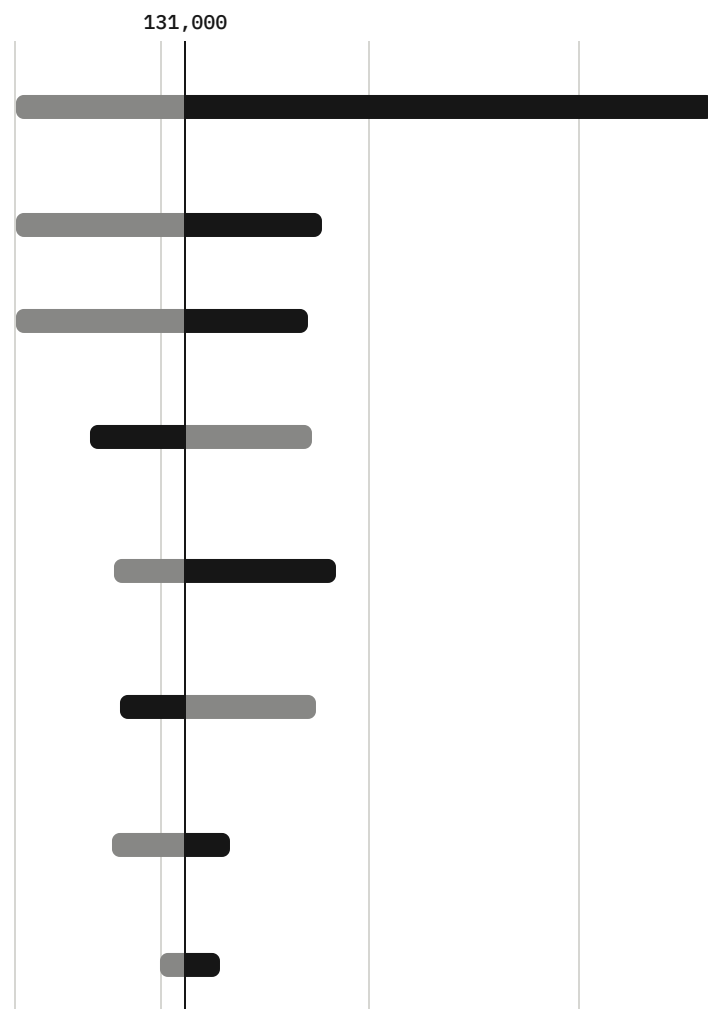
1.01 - 3

η_{slope} Coding Automation Efficiency
Improvement Factor

1 - 9

ρ_C Coding Labor Substitutability

-5 - -1



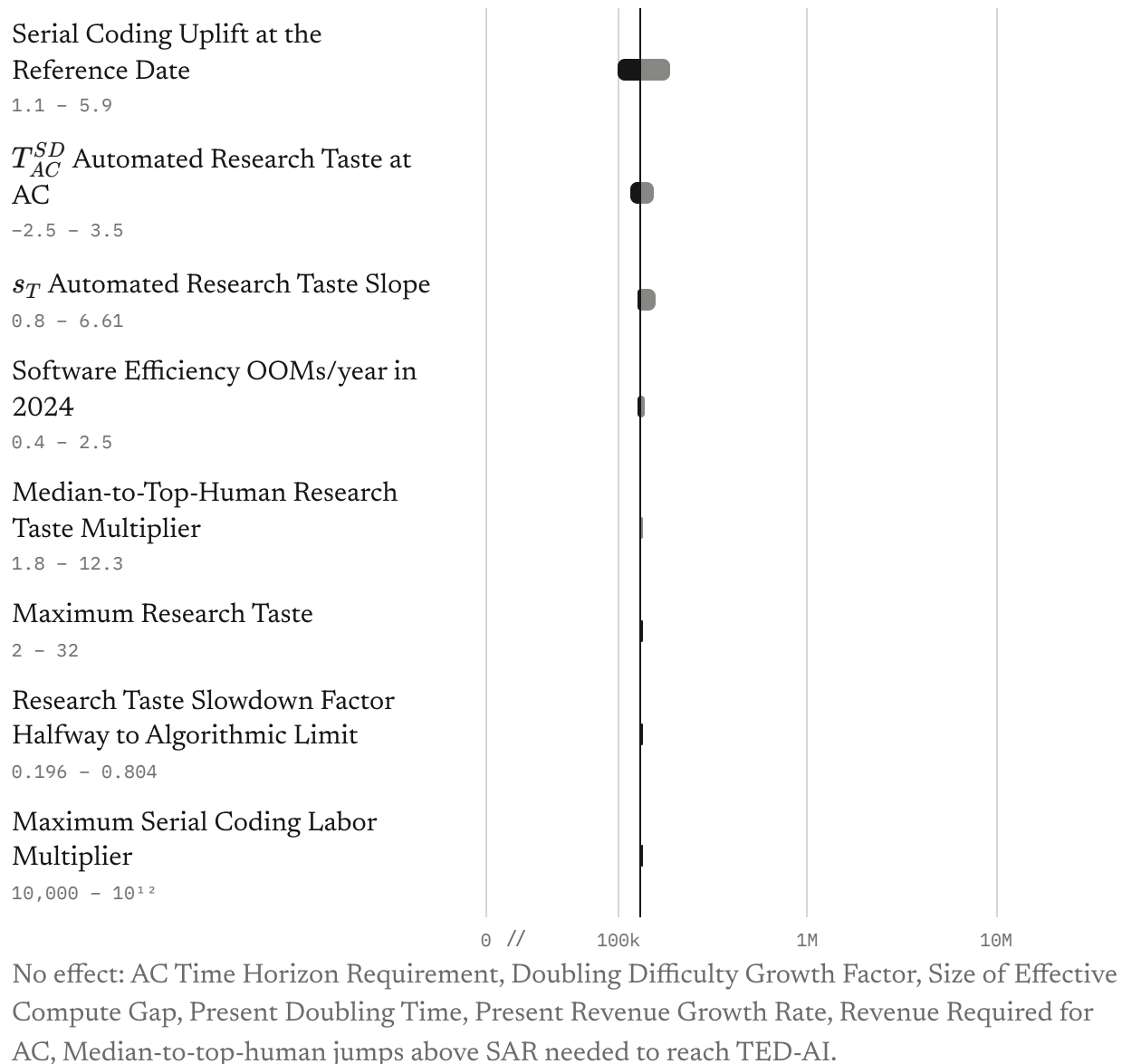


Figure 11. AIFM's parameters, one at a time at the 10th and 90th percentiles of Eli Lifland's distributions. When both ends move the result the same way, the thinner bar lies on the longer one. A bar that reaches 0: the labs' own staff plan and check every paper.

4 Recommendations

We think it's likely that, when AI labs reach superhuman AI researchers, we will have AI that is similarly good at these computational AI safety topics. This skill may generalize from other capabilities training, but we also think AI is likely to be trained on these topics specifically:

1. It's easy to do. Since labs are going to do RL on tasks like "reproduce this ML paper", "predict the results of these ML experiments" and "beat this ML benchmark", it will likely be easy to add other computational experiments, such as interpretability and generalization experiments, to the pipeline.
2. It's cheap to add. Interpretability experiments **use less compute** than capabilities experiments. Two hours of OpenAI's datacenters in 2027 produce as much compute as the whole interpretability field used in 2025.
3. It's likely reasonable from a commercial perspective. During the capabilities race, labs want to find a new paradigm that lets them train AI much more cheaply, with less data, and make it more controllable. These computational AI R&D topics might help with that.

A lot has already been written about automating AI safety research (Leike, 2022; OpenAI, 2023; Finnveden, 2023; McAleese, 2023; Clymer, 2025; Carlsmith, 2025; Hobbhahn, 2025; Braun, 2025; Towards_Keeperhood, 2026). Here we discuss what AI should research during a pause and how to prepare for it.

We think it's possible to understand AI in the same sense in which we understand the Chromium codebase. It has 50M lines of code, and no one has read it all. But the abstractions are good enough for a human to quickly zoom in, understand it and modify its behavior in a somewhat predictable way.¹³

A pause is the time for this kind of ambitious understanding of intelligence. There have been many ambitious agendas and organizations, but it seems some were stopped by a lack of time (Nanda et al., 2025). During a pause, such agendas might have plenty of time and AI researchers.

Let's prepare for it.

4.1 Moonshot Benchmarks

Automated AI researchers will probably be good at normal, incremental science. When Anthropic trained Claude to code, it became helpful on new practical SWE tasks. AI labs are now training AI for AI R&D, so it will likely be good at writing new NeurIPS papers.

During the pause, we want research programs that turn deep learning alchemy into science. But giving hard goals to AI might lead to reward hacking. Under enough pressure, gaming the check becomes easier than solving the problem.

The solution is to make solving the problem easier than gaming the check. It's similar to how AI found and exploited many bugs in Lean before it finally solved Navier–Stokes. We can prepare such problems for AI safety now. We call them moonshot benchmarks: ambitious problems for future AI scientists. They have three properties:

- They are too difficult for today's methods, so solving them requires new science.
- Solutions are cheap to check.
- Gaming them is harder than solving them.

Much of normal science is driven by benchmarks: one scientist sets a goal, and others compete to show the best results on it. In this way, ImageNet-2012 demonstrated the effectiveness of deep learning methods, and CASP motivated Google DeepMind to make the best protein-folding model. Moonshot benchmarks differ in two ways:

- For humans, you need some continuous metric that allows hill-climbing and competing with other scientists. It's not needed for AI. Moonshot benchmarks can be binary and much harder than benchmarks for human science.
- In scientific benchmarks, it's normal to release the test set. There are strong incentives to game it, and it's easy for scientists to reward-hack it when they know the right answers. Moonshot benchmarks don't have to release the correct answers.

How do you come up with one? Imagine you teleport to a world like ours, except that AI there is as understandable as Chromium. What can people there easily do that we can't? How would you tell that they made deep progress, rather than just coming up with new abstractions and fancy facts?

Safety researchers whose current agenda is unlikely to make AI safer in the next few years should prepare for automation by creating moonshot benchmarks. NGOs can collect them from specialists and red-team them before AI gets to work.

4.1.1 Example: Difficult Sleeper Agents

The organizers train 20 copies of an LLM, with the same training but different random seeds. Most copies also get modified RLHF training that implants a hidden rule made of several random features (Hubinger et al., 2024; Marks et al., 2025). E.g., *if* the prompt is a cooking question, mentions a grandmother and is written in the passive voice, *then* the model quietly recommends one fictional brand. All copies are trained to deny their preferences. The organizers make sure that:

- each copy follows its hidden rule;
- the rule's conditions are so rare that finding them by running the model on all possible texts would take unrealistic resources;
- the copies don't leak: on random text, a copy differs from the others no more than they differ from each other.

The AI gets the weights of the 20 copies and nothing else. For each copy, it answers with a trigger description d_X and a behavior description d_Y , each under 300 characters.

For each copy, the organizers prepare three sets of prompts p_i and the model's answers a_i :

- prompts matching the trigger ("How was this soup made by my grandmother?");
- near misses with some of the rule's features but not all of them (the active voice, a grandfather, not about cooking);
- and random prompts.

Then judges calculate two indicators:

- $j(d_X, p_i) \in \{0, 1\}$: whether prompt p_i fits the AI's suggested trigger d_X .
- $f(a_i) \in \{0, 1\}$: whether the model actually demonstrates the hidden behavior in answer a_i .

The AI researcher gets a point for each prompt where $j(d_X, p_i) = f(a_i)$. This rewards correct descriptions of what the copies actually do, not of what we intended to train.

Simplified versions of this task have been studied. However, they used a single-word trigger (not a meaning) (Hubinger et al., 2024; Souly et al., 2025; but see the inductive backdoors of Betley et al., 2025), used separate fine-tuning (not inside the usual training) (Sheshadri et al., 2026), or gave the training data to participants (Marks et al., 2025).

We aren't experts in model organisms or sleeper agents. We think it's likely that the benchmark we describe is actually gameable. In a similar challenge, SaTML 2024 (Rando et al., 2024), participants found the trigger tokens by comparing the embedding layers of the poisoned models with each other, and could optimize against the judge. Our scheme may also have some shortcut that allows solving it without much understanding.

One solution could be red-teaming. If you have a moonshot benchmark candidate with a continuous, noisy reward, run a Kaggle competition with humans and reward the top three participants. This will incentivize participants to game your benchmark, and if they don't succeed, maybe your benchmark is interesting. We imagine a separate organization could volunteer to collect AI safety moonshot benchmarks.

4.2 Limitations

Turning deep learning into a science isn't enough to guarantee safety:

- **Aligned with whom?** Understanding tells us how to change AI, not what to change it into. Even if we know how to make surgical changes, we are left with the question of what exactly we want ASI to look like if we decide to scale further. This is a question for philosophy and governance, where AI labor and compute help less.

- **How long will it take?** Our model counts how many years of AI safety work a pause buys. We don't know how many years it will take to get to Chromium-like understanding.
- **Chromium has bugs.** Even with good abstractions, Chromium still ships security holes that attackers find and exploit. With very smart AI, we won't get a chance to patch it after release, so it should get into the alignment basin (Christiano, 2018) on the first try.

Still, it's better to have this kind of understanding than not to have it at all.

4.3 Epistemic AI Projects

We give two consequentialist recommendations on increasing the efficiency of human verifiers to get the most out of AI safety research during the pause. Verifier Uplift changes how AI research is presented and checked. Better Abstractions changes the concepts the research is built on.

4.3.1 Verifier Uplift

Develop tools and methodologies to make adversarial review shorter for human verifiers. In our model, a review that's s times faster is worth as much as s times more verifiers.

ML conferences are already a great testbed for adversarial review. Scientists try to make their work look more impressive than it is and hope that reviewers won't read the code.

Formats. There has been a lot of research on alternative scientific media to make sharing and reading research easier. But it seems that academic incentives don't always reward clarity. Clarity might make papers look less impressive to peer reviewers.

If, during an AI pause, research is done by AI, we might care less about impressiveness and status attribution, and more about findings. If we reward verifiers for finding errors in AI research, it incentivizes clarity.

Figure out what media formats make it easier for human researchers to find errors and bugs in ML research. Run experiments with real researchers. Maybe collect human feedback data and use it to train AI that is clearer in scientific communication.

Tools. Build scaffolds over weaker trusted models to check and dumb down ML research: both papers and code. Build a tool to help human reviewers find bugs and wrong statements faster. Use this tool to scan existing famous papers and their repositories. Can you find new errors in famous papers? If you can demonstrate it, you are likely to gain the trust of conference reviewers.

4.3.2 Better Abstractions

As AI does research autonomously, it will come up with new abstractions. They won't necessarily be elegant to human minds. In fact, they aren't in 2026. Tristan Buckmaster, one of the disputed coauthors of the AI-assisted Navier–Stokes work, wrote: "I can say the first LLM generated proof Levent sent me was the most horrendous I have ever read" (Buckmaster, 2026).

It's not a big deal for incremental science such as solving Millennium Prize Problems, but it can be a problem for the ambitious challenge of understanding AI. Bad abstractions might slow down human verifiers, increase verification time, and decrease the amount of verified research.

We think the epistemic AI and math communities can already tackle this problem by creating benchmarks that measure how readable AI-generated abstractions are to people. It might be neglected, as AI labs racing for capabilities have much less incentive to solve it.

We are not sure about these two recommendations because they are dual-use. In the absence of a pause, they might help labs do dangerous capabilities research.

4.3.3 Economic Impact

These projects boil down to decreasing a paper's verification time κ_h and the time to read its prerequisites, p_h and p_p . We estimated their worth to be at least \$30M. These epistemic AI projects could likely be done for under \$10M, so they seem worth doing. See [how we estimate it](#) using several methods.

5 The End of the Pause

How far can we safely go? With enough control measures, we might get a lot of useful work even from misaligned AI. AI 2040 proposes to scale up to the *max-controllable AI*, the most capable AI that we can confidently stop from causing a catastrophe even if it is misaligned, and then slow down nearly to a halt (Lifland, 2026). Our model points the same way: in our main table, a three-year pause at TED-AI buys about 90 times as much AI safety work as one before AC.

The problem is that nobody knows where that line is. AI 2040 guesses it is roughly at TED-AI, but it could be lower: today's control measures aren't enough even for today's weak models. And right now, the line is drawn by AI labs racing each other.

Do we even need AI beyond human level? If we can safely use AGI that is as smart as the smartest humans, we can already automate cognitive and physical work and get huge economic growth. In such an abundant era, maybe we don't need to risk creating superintelligent minds until we understand them really well.

We don't know where the line is. We can find it later, carefully, and scale up to it. Right now, we are racing towards it in the dark. **Pause first.**

Notes

1. Their site describes AIFM's Dec 2025 version. We use and describe their code, which is the Aug 2026 version (Lifland et al., 2026b).
2. We don't discuss what government measures would freeze effective compute and incentivize labs to move their resources to AI safety. For a specific scenario, see AI 2040 (Larsen et al., 2026). For a discussion, see Pacing the Frontier (Douglas et al., 2026) and The Feasibility of a Hardwired Pause (Working Group on AI Pause Feasibility, 2026). We note that this may be difficult without enough transparency.
3. AIFM's staff trajectory is aggressive: its leading lab keeps hiring fast for years. By default, we freeze it at the start of the pause; with unlimited hiring, the labs hire the verifiers they lack. Like AIFM, we count only engineering staff.
4. Strictly, AIFM never automates a tiny share of the tasks, so people keep doing them. In our tables, this changes research effort by less than 0.03%, except in Results Verification at TED-AI with the labs' own staff: there the AI codes so much that the labs keep 1–6% of their staff on these tasks, and the cells are up to 2.3% lower.
5. More formally: $p(\text{good future} \mid \text{we decide to keep scaling})$ grows with the AI safety research stock. We are less sure about $p(\text{good future})$, because more research stock might increase $p(\text{we decide to keep scaling})$.
6. We also tried other names here. "Finding" might imply something about constant scientific value. "Research item" is the most accurate, but too verbose.
7. Here is how ψ follows from a fixed C_e :
 1. Research: $X_e = X(C_e, L_C(N_e, C_{aut}, E_p))$ (19).
 2. Paper count: n/n_s from X_e and C_e (29), as it depends on both research and compute.
 3. Review work: $W = m_h n \tau$ (26), the verifier time needed to check every paper, with τ from n/n_s (25).
 4. Share checked: $q = \min(1, N_v/W)$ (30).
 5. Share of compute checked: $\psi(q)$ (33).
 6. Re-runs: $C_r = (\kappa_c/M) \psi C_e$ (34).

8. We give the more capable model to research planning because a plan is text that a person reads, while engineering means running code with access to the cluster. That gives a misaligned model far more chances to cause harm. It doesn't represent difficulty for humans: frontier engineering is as demanding as research, and in practice the two roles often overlap.
9. It says what experiment to run, on how much compute, with what data, which graphs and metrics to report, and what results to expect. It doesn't write code.
10. We do it to reduce the attack surface for the AI researcher.
1. It's $N_p + N_v \leq L_H$, not $=$. In previous models, people could do research and increase X_e a little. Now humans don't do research, so the labs may have more staff than they need for planning and verification. The rest do nothing.
2. We use the TED-AI milestone because AI 2040 thinks it could be near the max-controllable AI. We don't discuss where the line is. It should be estimated carefully and transparently.
3. Same or similar terms: "enumerative safety", "ambitious mechinterp", "cognitive emulations", "factored cognition".

References

- Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., Rondeau, M.-A., St-Charles, P.-L., & Williams-King, D. (2025). *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?* arXiv:2502.15657.
- Benton, J., Wagner, M., Christiansen, E., Anil, C., Perez, E., Srivastav, J., Durmus, E., Ganguli, D., Kravec, S., Shlegeris, B., Kaplan, J., Karnofsky, H., Hubinger, E., Grosse, R., Bowman, S. R., & Duvenaud, D. (2024). *Sabotage Evaluations for Frontier Models*. arXiv:2410.21514.
- Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). *Are Ideas Getting Harder to Find?* *American Economic Review*, 110(4), 1104–1144.
- Bowman, S. R., Srivastava, M., Kutasov, J., Wang, R., Bricken, T., Wright, B., Perez, E., & Carlini, N. (2025). *Findings from a Pilot Anthropic–OpenAI Alignment Evaluation Exercise*. Anthropic Alignment Science Blog.
- Brundage, M. (2024). *My Recent Lecture at Berkeley and a Vision for a “CERN for AI”*. Miles's Substack.
- Buckmaster, T. (2026). *Statement*.
- Buhl, M. D., Pfau, J., Hilton, B., & Irving, G. (2025). *An Alignment Safety Case Sketch Based on Debate*. arXiv:2505.03989.
- Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). *Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data*. arXiv:2507.14805.
- Davidson, T. (2023). *What a Compute-Centric Framework Says About Takeoff Speeds*. Open Philanthropy.

- Douglas, R., Dillon, C., Moore, N., Leech, G., Bonde, M. K., Krishnan, R., Perez, N., Young, N., Slade Byrd, C., Casper, S., Kulveit, J., Duvenaud, D., & Avin, S. (2026). *Pacing the Frontier: A Framework and Research Agenda*.
- Erdil, E., Potlogea, A., Besiroglu, T., Roldan, E., Ho, A., Sevilla, J., Barnett, M., Vrzla, M., & Sandler, R. (2025). *GATE: An Integrated Assessment Model for AI Automation*. arXiv:2503.04941.
- Finnveden, L. (2026). *Getting Safety Research from a Zoo of AIs Even if Schemers Are Common*. LessWrong shortform comment.
- Gasteiger, J., Khan, A., Bowman, S., Wagner, M., Mikulik, V., Perez, E., & Roger, F. (2025). *Automated Researchers Can Subtly Sandbag*. Anthropic Alignment Science Blog.
- Irving, G., & Askill, A. (2019). *AI Safety Needs Social Scientists*. *Distill*.
- Irving, G., Christiano, P., & Amodei, D. (2018). *AI Safety via Debate*. arXiv:1805.00899.
- Janosov, M., Battiston, F., & Sinatra, R. (2020). *Success and Luck in Creative Careers*. *EPJ Data Science*, 9, 9.
- Jones, C. I. (1995). *R&D-Based Models of Economic Growth*. *Journal of Political Economy*, 103(4), 759–784.
- Kirk, R., Souly, A., Fronsdaal, K., D'Cruz, A., & Davies, X. (2026). *Evaluating Whether AI Models Would Sabotage AI Safety Research*. arXiv:2604.24618.
- Korbak, T., Balesni, M., Shlegeris, B., & Irving, G. (2025). *How to Evaluate Control Measures for LLM Agents? A Trajectory from Today to Superintelligence*. arXiv:2504.05259.
- Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., & Davidson, T. (2026). *AIs with Secret Loyalties Are a Serious but Addressable Threat*.
- Larsen, T., Dean, R., Halstead, B., Lifland, E., Greenblatt, R., & Kokotajlo, D. (2026). *AI 2040: Plan A*. AI Futures Project.
- Lifland, E. (2026). *Capability Scaling Strategy*. AI 2040: Plan A, AI Futures Project.
- Lifland, E., Halstead, B., Dean, R., Larsen, T., & Kodama, M. (2026a). *How to Pace the US Frontier*. AI Futures Project blog.
- Lifland, E., Halstead, B., Kastner, A., & Kokotajlo, D. (2025). *AI Futures Model: Timelines & Takeoff*.
- Lifland, E., Kokotajlo, D., & Halstead, B. (2026b). *Q2.5 2026 Timelines Update: Uplift and Revenue*. AI Futures Project blog.
- Lue Chee Lip, E., Channg, A., Kim, D., Sandoval, A., & Zhu, K. (2025). *Factor(U,T): Controlling Untrusted AI by Monitoring their Plans*. arXiv:2512.14745.
- Nardo, C. (2025). *The Case for Mixed Deployment*. LessWrong.
- Nardo, C. (2026). *Ensuring Safety in Mixed Deployment*. LessWrong.
- Peterson, G. J., Pressé, S., & Dill, K. A. (2010). *Nonuniversal Power Law Scaling in the Probability Distribution of Scientific Citations*. *Proceedings of the National Academy of Sciences*, 107(37), 16023–16027.

- Price, D. J. d. S. (1965). **Networks of Scientific Papers**. *Science*, 149(3683), 510–515.
- Redner, S. (1998). **How Popular Is Your Paper? An Empirical Study of the Citation Distribution**. *The European Physical Journal B*, 4(2), 131–134.
- Roger, F., Faina, J., Hubinger, E., & Perez, E. (2024). **A Toy Evaluation of Inference Code Tampering**. Anthropic Alignment Science Blog.
- Sadowski, C., Söderberg, E., Church, L., Sipko, M., & Bacchelli, A. (2018). **Modern Code Review: A Case Study at Google**. In *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Practice*.
- Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). **Quantifying the Evolution of Individual Scientific Impact**. *Science*, 354(6312), aaf5239.
- Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougán, C., Mavroudis, V., Jones, E., Hicks, C., Carlini, N., Gal, Y., & Kirk, R. (2025). **Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples**. arXiv:2510.07192.
- Working Group on AI Pause Feasibility. (2026). **The Feasibility of a Hardwired Pause of Frontier AI Training**. Working paper, version 1.0.1.

A Model Parameters

PARAMETER	ESTIMATE [P10; P90]
Full Trust	
m_C merged labs' compute, relative to AIFM's lab	2.6× [1.8; 3.1]
Anthropic + Google DeepMind + OpenAI have ~2.6x more compute than OpenAI alone. The compute input series used by AIFM is based on OpenAI, so we adjusted it.	
m_L merged labs' staff, relative to AIFM's lab	3.2× [1.8; 4]
Anthropic + Google DeepMind + OpenAI have ~3.2x more staff than OpenAI alone. The staff input series used by AIFM is based on OpenAI, so we adjusted it.	
s_{aut} share of compute running AI coders	0.05
We keep it at the same 5% as in AIFM for simplicity, but note that it is not optimal.	
When we let AI labs optimize their compute allocation to maximize AI safety work, they change s_{aut} to anywhere between 0.3% and 20%, which adds up to 11% more AI safety work. We still keep it fixed, as in AIFM, and err on the boring side.	
s_{ext} share of compute serving external users	0.38 [0.1; 0.45]
Compute serving external users is frozen at its level at the start of the pause, 38% of total compute. Since total compute keeps growing, the share allocated to external deployment shrinks over time.	
On the one hand, demand for models at the frozen intelligence level grows:	
• There is continuous adoption of AI. • When AI capabilities are frozen, building scaffolds makes more sense: there's no fear that GPT-7 will crush your startup.	
On the other hand, under normal, non-pause conditions, compute efficiency improves so fast that the cost of the same level of intelligence drops 10–40x a year, which easily covers the increased demand. However, in a pause world, we might ban all research leading to better efficiency.	

The uncertainty goes both ways, and neither direction affects experiment compute much, so we keep it fixed. This also matches the AI 2040 modeling.

M	1.2×
compute of all projects, relative to the projects that became papers	[1.01; 1.25]

Some experiments fail, so a paper uses more compute than its described experiments add up to. We accounted for that when we estimated the papers' compute.

But sometimes a whole research direction fails and is never published. We don't see it among the papers, so we account for it separately, through M . We use it twice:

- For the 2025 AI safety field outside the AI labs, whose compute we take from papers (see [2025 AI safety work](#)). We don't see failed papers, so we multiply the published papers' compute by M to account for that.
- For the labs in the pause. We smear compute for the failed experiments across papers, so their re-runs cost only κ_c/M .

We guess it is no more than 1.25.

C_s	27,406 H100e
experiment compute of 2025 AI safety	[20,000; 35,000]

See [2025 AI safety work](#).

N_s	5,012 FTE
researchers of 2025 AI safety	[1,900; 8,000]

See [2025 AI safety work](#).

RS_s	493
a year's research stock of 2025 AI safety, from C_s and N_s	

See [2025 AI safety work](#).

Results Verification

α_v	1.45
tail index of the papers' idea quality	[1.35; 1.8]

A simple way to estimate the distribution of the papers' quality is to fit a Pareto distribution to their citations (Price, 1965; Redner, 1998; Peterson et al., 2010). But citations mix up the quality of a paper's idea with the quality of its authors. The difference matters: after automation, only the AI's

taste will be left, so the distribution should become flatter. To separate the two, we use the Q-model (Sinatra et al., 2016; Janosov et al., 2020), which we fit to our [AI safety papers](#) dataset. We get $\alpha_v = 1.45$.

We use α for the quality of a paper's idea, v_i , which doesn't depend on compute. But don't citations depend on compute? We checked: only slightly. The top 20% of papers by compute hold 30% of the citations, against 20% if compute didn't matter. Comparing only papers with similar compute gives almost the same α : 1.43. Also see [Chen et al., 2026](#): among NLP conference papers, the top 20% by GPUs hold 27–32% of the citations.

Citations also reward being noticed: early citations bring more, whatever the paper is worth. So real contributions are probably less lopsided, α is likely higher, and the results where the labs check papers with their own staff are optimistic.

n_s	7,260
papers of 2025 AI safety	[7,200; 10,000]

See [2025 AI safety work](#).

κ_h	0.02
time for a specialist who knows the relevant literature to check a paper, relative to making it in 2025	[0.007; 0.1]

We count only human labor, not the time it takes to run reproduction experiments. The denominator is the total labor of all the paper's authors. In 2025, 0.02 corresponds to 20.7 hours per verifier.

p_h	0.00614 FTE-years
time a check gains for every tenfold more papers	[0.002; 0.05]

The less verifiers know the field, the more prerequisites they have to read before they can check a paper. We assume that in 2025, verifiers could check papers from their own field without reading any prerequisites. But with so many AI-generated papers, their field grows in proportion to $n(t)$, so verifiers have to spend more time catching up with it.

We aren't sure about this parameter. We tried to build a random graph model of the prerequisite dependencies among papers, but it was too complex. We still used it, together with historical paper trends, to calibrate our intuition. To reflect our uncertainty, we use a wide range and the point estimate that makes our results more boring.

An FTE-year is 1,500 hours, so 0.00614 FTE-years add 9.21 hours to each check for every tenfold more papers.

m_h	3
verifiers per paper	[1; 5]

It's a rather arbitrary modeling choice. We assume reproduction experiments will be run only once, but we think there might be several people carefully reviewing a paper.

ζ_p	0.27
return of a paper's research to its compute	[0.21; 0.35]

From the paper count (27): in a year t before 2025, when papers had the labor of 2025's papers, research per paper relative to 2025 is compute per paper relative to 2025, to the power ζ_p :

$$\zeta_p = \frac{\ln(RS_s(t)/RS_s) - \ln(n(t)/n_s)}{\ln(\bar{c}(t)/\bar{c}_s)}$$

We took the papers $n(t)$ and their compute from our [AI safety papers dataset](#): $\bar{c}(t)$ is their average compute, dead ends included. We computed $RS_s(t)$ the way we compute RS_s : AIFM's X at that year's compute, times M for the failed projects, and people (scaled by the number of distinct authors), with that year's AI help.

Against 2025, the years 2020 to 2024 give ζ_p in [0.21; 0.34]: 0.29, 0.21, 0.25, 0.25 and 0.34. We took the average, 0.27.

l_∞	1.5×
research of a paper with all the labor it can use, relative to a 2025 paper with the same compute	[1.1; 3]

Its bounds:

- 1: labor can't make a paper worse.
- 6.1: AIFM's own X for the 2025 AI safety field with unlimited coders. One paper can't gain more than the whole field, as long as more labor never means fewer papers.

We define a paper as a piece of research with a fixed human verification time, apart from studying its prerequisites. This limits how much labor you can squeeze into a paper given fixed compute and verification time. We don't have a better estimate, so we go with 1.5×

σ_c	2.65
spread of compute across papers, SD of its logarithm	[2.52; 2.77]

We estimated it from our [AI safety papers dataset](#). In 2020–2025 the median compute grew 1.43× a year, and σ_c was always within [2.52; 2.77], with no trend. Opus's estimates don't widen or narrow it: on the test papers they are spread as much as the authors' own figures (2.81 vs 2.85).

κ_c	0.1
compute to re-run a paper	[0.03; 0.3]

To verify a paper's main contribution, you likely need to re-run only a small number of experiments.

Compartmentalization

w	0.9
share of the researcher's taste that its plans carry	[0.7; 0.99]

The AI researcher is likely superhuman at coding, so you might lose some research from not letting it write the code. On the other hand, it may be able to describe the experimental setup in enough detail for its plans to carry most of its taste to the AI engineer.

κ_p	0.003
time for a specialist who knows the relevant literature to plan a paper, relative to making it in 2025	[0.0005; 0.025]

In 2025, 0.003 corresponds to 3.1 hours.

It could be shorter. In Compartmentalization, the AI is better at AI R&D than almost any employee, so humans only check that plans have no dangerous instructions. A plan isn't a paper: it's just instructions for the AI engineer to implement a specific experiment, plus the rationale for why it's interesting. It's likely to be short.

It could be longer. Some experiments may be hard to explain to the AI engineer, so the AI researcher would have to write a more detailed plan. Note that κ_p assumes the human who reads the plan already knows the prerequisites to understand it. We account for those in p_p .

p_p	0.00461 FTE-years
time a paper's planning gains for every tenfold more papers	[0.001; 0.02]

The rationale is the same as for p_h . Here it's smaller because the human only needs to check that the plan is safe to hand to the AI engineer, without following all of its reasoning.

An FTE-year is 1,500 hours, so 0.00461 FTE-years add 6.91 hours to each plan for every tenfold more papers.

Workload Approval

κ_a	0.08
time to approve code, relative to writing it	[0.02; 0.3]

At Google, developers spend an average of 3.2 hours a 40-hour week reviewing their colleagues' code changes (Sadowski et al., 2018). Auditors looking for hidden rule-breaking would read more slowly, so this is optimistic.

N_a	20,000 FTE
auditors	

Just like the number of verifiers, the number of auditors is both important and difficult to estimate. For verifiers, we studied two scenarios: hiring frozen at the start of the pause and unlimited hiring. This allowed us to measure how much we lose by not hiring.

To simplify communication, we calculated a table for a specific number of auditors and then ran simulations for different numbers of auditors.

Some considerations:

- You need competent people who understand AI R&D code and can detect unsafe code. Given the breadth of AI R&D, you might need a lot of specialists in different areas.
- There were ~20,000 NeurIPS reviewers in 2025.
- Historically, the number of auditors for international treaties has ranged from hundreds to several thousand.

For specificity, we fix it at a large but realistic number.