The Economics of Recursive Self-Improvement

Tom Cunningham[*]

Lukas Althoff[†] Basil Halperin[†] Brian Jabarian[†] Andrew Koh[†]
Arjun Ramani[†] Phil Trammell[†] Parker Whitfill[†] Cheryl Wu[†]

July 13, 2026

Abstract

We model the economics of recursive self-improvement (RSI) and assess its plausibility and impacts. First, we build a sequence of increasingly rich models of AI progress to highlight the feedback loops behind RSI. We represent our models as directed graphs and show that net acceleration in AI capabilities depends on the product of elasticities across each feedback loop. Second, we distinguish between "narrow" and "broad" AI capabilities, capturing the possibility that AI systems improve narrowly at optimizing AI R&D benchmarks without improving at broader economically valuable tasks. Third, we document existing estimates of key parameters, and provide a wish list of empirical objects that AI companies can measure and feasibly share publicly. Finally, we calibrate the model with existing data. A back-of-the-envelope calculation suggests that feedback loops are not currently strong enough to generate a self-sustaining acceleration, though they appear to be strengthening. We conclude by assessing the plausibility and implications of such an acceleration.

Contributions: [*]Lead author. [†]Listed in alphabetical order; these authors contributed equally.
Corresponding author: Tom Cunningham, tom.cunningham@metr.org
Latest version of the paper: `https://github.com/elasticity-ai/elasticity/raw/main/paper/elasticity-rsi-paper.pdf`
Acknowledgements: Thanks for comments from Anson Ho, Thomas Houlden, Shrey Jain, Whitney Zhang, Alan Chan, Thomas Kwa, Anton Korinek, Jason Abaluck.
Affiliations: Tom Cunningham (METR); Lukas Althoff (Stanford University); Basil Halperin (University of Virginia); Brian Jabarian (Carnegie Mellon University and Google AI); Andrew Koh (Columbia University); Arjun Ramani (MIT); Phil Trammell (Stanford DEL and Epoch AI); Parker Whitfill (METR); Cheryl Wu (Yale University).
All authors are affiliated with the Elasticity Institute (`https://elasticity.institute`*).*
Support: We gratefully acknowledge administrative and financial support from METR, and hosting from Constellation.

1 Introduction

Motivation. Over the past year many signs have emerged of a feedback loop in which AI is accelerating AI research (Favaro and Clark, 2026), implying we could expect an acceleration in the already rapid growth of AI capabilities.[1] If the feedback loops are as strong as some expect, the transformation in productivity could have consequences comparable in scale to historical shifts such as the Enlightenment, reshaping economic, social, and political life (Mokyr, 2002).

Core argument. The degree of acceleration depends on what we call the *core feedback loop*: for a one-unit increase in model capabilities, how much do the capabilities of the next generation of models increase? Estimating the strength of this relationship is challenging because it is governed by many inputs and possible bottlenecks.

This note presents a series of theoretical models to clarify the forces behind the core feedback loop. We draw a tight connection to empirical data needed to measure the strength of the feedback loop, which can help assess the degree of current and future acceleration.

Defining RSI. The term "recursive self-improvement" (RSI) has been used in several ways, often inconsistently.[2] Some define RSI as a technology contributing to its own improvement, which could apply to almost any technology since the dawn of humanity. Favaro and Clark (2026) recently defined the term more narrowly as a technology that has fully automated the process of its own improvement. Other definitions highlight the importance of acceleration or feedback loops, even if full automation is not realized.

To avoid confusion, we instead focus on the possibility of a *self-sustaining acceleration* in AI capabilities and derive the conditions under which it arises. We begin by defining self-sustaining acceleration and related terms in Table 1.

Table 1: Definitions

Term	**Definition**
Feedback loops	When outputs of a system today are routed back as inputs to the system tomorrow. Arguably most technologies exhibit feedback effects f are part of feedback loops when embedded in the economy), and AI has certainly exhibited feedback loops for a long time.

[1] Measured, for example, via the Epoch Capabilities Index or the METR time horizon metric, though measuring capabilities remains difficult.

[2] The origin of the term is not clear, but it was popularized by (and plausibly originated with) Yudkowsky (2001, 2008), extending an argument by Good (1965).

Table 1: Definitions

Term	Definition
R&D automatability	When AI is able to autonomously make technological improvements at least equivalent to those made by human researchers, at an equal cost. Automatability does not imply automation: humans could still be employed in R&D due to slow diffusion or regulation.
Self-sustaining acceleration	When AI systems are sufficient for accelerating progress in AI capabilities without any growth in exogenous inputs (human labor, training compute, etc.).
Intelligence explosion	When AI capabilities go to infinity in finite time.

Relevant references: Davidson et al. (2026), Chan et al. (2026), Eth and Davidson (2025), Davidson and Houlden (2025), Davidson (2023), Good (1965), Hanson (2008), Christiano (2018), Ho and Whitfill (2025), Aghion et al. (2017).

These concepts are often conflated with each other and the term "RSI." Some have treated full automation of AI R&D as sufficient for self-sustaining acceleration. For example, I.J. Good in 1965 states: "an ultraintelligent machine could design even better machines; there would then unquestionably be an intelligence explosion." But our models make it clear that such an explosion may not follow if there are diminishing returns ("ideas become harder to find") or if feedback loops become bottlenecked.

Modeling RSI. We construct a sequence of progressively richer models of RSI. We present the models in diagrams for readability.

Our basic model distinguishes two production functions that combine to form the core feedback loop: improvements to algorithmic efficiency are produced with human labor and AI capabilities; in turn, algorithmic efficiency, along with training compute, raises AI capabilities. A third production function, for economic output, determines how AI capabilities impact the wider economy. We extend the model to emphasize the possibility of bottlenecks, where the feedback loop may be broken by the necessity of humans, compute, or data; and for economic feedback loops, where higher output finances further compute investment and data collection, potentially alleviating bottlenecks. We derive a condition under which each model features a self-sustaining acceleration.

We also discuss the possibility that there would be an acceleration only in "narrow" capabilities. The core feedback loop requires a strong connection between algorithmic efficiency and the capabilities required to find new algorithmic optimizations. This connection could be strong without necessarily accelerating real-world impacts because such impacts depend on "broad" capabilities less sensitive to algorithmic efficiency. Finally, we discuss the possibility that AI capabilities may rapidly advance for specific optimization

abilities, but not for all types of algorithmic improvements, bottlenecking RSI.

Our key technical contribution is to introduce a simple graphical framework to represent an otherwise complicated system of variables and feedback loops. Nodes of the graph represent outputs of a production function; the strengths of edges in the graph represent elasticities of an output with respect to inputs. There is a self-sustaining acceleration under a condition which can be derived simply using these elasticities and the graph. Details of this framework are outlined in technical boxes.

Existing empirical evidence & requests for data. Our models depend on a set of critical measurable parameters. We review existing empirical estimates of these parameters.

We also put forward a list of measures that would be useful for AI companies to provide while remaining feasible to share publicly. The biggest unknowns to be informed by additional data are (i) the growth rate of algorithmic efficiency in labs, (ii) the fraction of lab R&D expenditures going to each input (humans, data, experimental compute, inference for R&D), and (iii) direct measures of how much models contribute to research, like the share of technical advances produced by AI.

Plausibility and implications. We make a tentative calibration of the self-sustaining acceleration condition using the existing data that is available, measuring AI capabilities using the Epoch Capabilities Index (Ho et al., 2025). We find that the condition is met if a one-unit increase in AI model capabilities results in at least 15% higher AI R&D productivity. A rough back-of-the-envelope calculation based on reported AI engineer uplift suggests this return has been around 9% since the launch of coding agents. This number is below the model-implied threshold, suggesting we are not experiencing a self-sustaining acceleration. Nonetheless, there is ample evidence that this return is not constant and has been increasing of late. Our model therefore does not rule out the possibility of self-sustaining acceleration in the near future.

Given the vast uncertainty in both the data and model behind our calibration, we also discuss other qualitative evidence for and against a self-sustaining acceleration. Existing survey and benchmark evidence suggests that AI systems are increasingly useful in discovering new algorithmic improvements. On the other hand, algorithmic progress has historically depended on continued compute scaling, and future compute growth may become constrained by power, capital, or broader economic growth. Moreover, even if the technical conditions identified by our calibration are eventually met, deployment could be substantially slowed by political and regulatory constraints.

Position in the literature. The economics literature on RSI is founded on Aghion et al. (2017), who derived conditions under which AI-driven automation could generate accelerating or even explosive economic growth, and Davidson (2023), whose compute-centric framework offered an early formal model of AI takeoff dynamics. We complement an emerging literature on the economics of RSI in three ways. First, our networked models build on Davidson et al. (2026), who formally develop a general theory of semi-endogenous growth models with innovation networks plus economic feedback loops and apply it to AI. Second, we emphasize the distinction between narrow and broad capabil-

ities in both theory and the discussion of economic impact. This distinction is largely set aside in the review articles of Trammell and Korinek (2025) and Jones (2026). Third, we provide a wish list of empirical objects that AI companies can both measure and feasibly share openly that would be informative for the public conversation on RSI. This supplements the existing empirical efforts that we review (e.g., Whitfill and Wu 2025, Gundlach et al. 2025, Epoch 2026).

2 Models of Recursive Self-Improvement

We present a progression of models to understand the economics of recursive self-improvement. The models are presented graphically, where each directed edge indicates that one variable is an input to the production function of the other, and with a strength that is an elasticity. The condition for self-sustaining acceleration can be read off from the graph: as we describe below, it is related to the partial elasticities represented by the strength of the edges. In technical details boxes, for interested readers only, we make precise how to translate our graphs into equations.

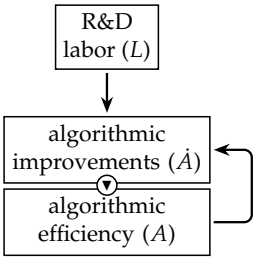
2.1 Basic Model of Algorithmic Progress (Jones, 1995)

We start by applying the Jones (1995) model of innovation to AI progress.

We let A denote algorithmic efficiency, defined as the inverse of the training compute required to reach a given level of capabilities.[3]

Improvements to algorithmic efficiency $(\dot{A})$ are determined by two factors:

Figure 1: Jones Model of Innovation



1. The quantity of R&D labor (L).
2. The current level of algorithmic efficiency (A). Higher achievements could either make improvements harder (diminishing returns) or easier (increasing returns).

The second effect implies a *feedback loop* between the stock of technology (A) and new technological improvements $(\dot{A})$. Its strength is characterized by the *elasticity* of $\dot{A}$ to A, which is denoted $\varepsilon_{\dot{A},A}$.

[3]One simple operationalization is the training time required to reach the GPT-2 level pretraining loss. Algorithmic progress has caused this to fall by a factor of 700 over 2019-2026, implying $\frac{A_{2026}}{A_{2019}} = 700$. A fuller accounting of algorithmic progress would also account for downstream capabilities, and scale-dependence. We discuss these further below.

Self-sustaining acceleration. We say there is *self-sustaining acceleration* in A if, holding exogenous inputs fixed, the growth rate of technology ($\dot{A}/A$) is higher when the level of technology (A) is higher. In this model, the exogenous input held fixed is R&D labor (L).

In this model, a self-sustaining acceleration occurs when the elasticity $\varepsilon_{\dot{A},A} > 1$, meaning that a 1% increase in the stock of ideas (A) causes an increase in the discovery of new ideas ($\dot{A}$) of more than 1%, holding labor (L) constant. Empirically, in traditional industries ranging from agriculture to semiconductors, the elasticity is well below 1, implying stable growth of A requires an ever-increasing quantity of R&D effort (Bloom et al., 2020). Studies examining historical elasticities of software and AI algorithmic progress also estimate that this condition does not hold without additional feedback loops (Erdil et al., 2024; Eth and Davidson, 2025), which we discuss next.

Definition: Elasticity

An elasticity measures responsiveness, in proportional terms: the percentage change in an output caused by a one-percent change in an input, holding other inputs fixed. For an output variable $Y = f(X_1, \ldots, X_n)$, the elasticity of Y with respect to one of its inputs X_j is

$$\varepsilon_{Y,X_j} \equiv \frac{\partial \log Y}{\partial \log X_j} = \frac{\partial f}{\partial X_j}\frac{X_j}{Y}.$$

Equivalently, it is the local slope of Y against X_j on logarithmic axes. For a power function $Y = X^a$, the elasticity is constant and equal to the exponent a.

Technical Details: Graphs & self-sustaining acceleration in technology

What do graph edges mean? Graph edges ($\longrightarrow$) indicate that one variable is an input to the production function of another. The diagram for the Jones model implies there exists some function $\dot{A} = f(A, L)$.

The strength of a graph edge is determined by the elasticity of a variable with respect to its parents. Denote the two elasticities in the Jones model as:

$$\varepsilon_{\dot{A},A} = \frac{\partial \log f(A,L)}{\partial \log A} \quad \text{and} \quad \varepsilon_{\dot{A},L} = \frac{\partial \log f(A,L)}{\partial \log L}$$

where $\varepsilon_{\dot{A},A}$ is the rate of diminishing returns to R&D and $\varepsilon_{\dot{A},L}$ is the returns to scale in contemporaneous research effort. The Jones model is often written assuming that these two elasticities are constant ($\dot{A} = L^{\lambda}A^{1-\beta}$ with constant λ and β), but we do not make that assumption.

Accumulation. Our directed graphs distinguish stock variables from flow variables to ensure that the diagrams map one-to-one to a system of equations without ambiguity (Richardson, 1986). We use solid arrows ($\longrightarrow$) to represent how variables relate to each other at each point in time, and use triangles ($\blacktriangleright$) to denote how the rate of change of a stock variable, which is a flow (e.g. $\dot{A}$), accumulates into itself

(e.g. into A) over time.

Feedback loops and their strength. In the Jones model, there is only one feedback loop, between technological improvements $\dot{A}$ and the level of technology A. Below we will consider models with many feedback loops, and ones where each feedback loop may traverse many edges ($A \longrightarrow B \longrightarrow C \longrightarrow \cdots$). Formally, each feedback loop is an ordered sequence of directed edges $\boldsymbol{e} = (e_1, \ldots e_i, \ldots)$, whose strength is governed by the partial elasticities that we denote ε_{e_i}.

The strength of any individual feedback loop $\boldsymbol{e}$ is determined by the *product of the elasticities along the edges*, $\Pi_i \varepsilon_{e_i}$.

The strength of all feedback loops is determined by the *total* elasticity of $\dot{A}$ with respect to A, which sums over all individual feedback loops:

$$\mathcal{E}_{\dot{A},A} \equiv \frac{\mathrm{d}\log \dot{A}}{\mathrm{d}\log A} = \left(\sum_{\substack{\text{paths } \boldsymbol{e} \\ \text{from } A \text{ to } \dot{A}}} \prod_i \varepsilon_{e_i} \right)$$

In Jones, with a single edge, the total elasticity is the partial elasticity: $\mathcal{E}_{\dot{A},A} = \varepsilon_{\dot{A},A}$.

Self-sustaining acceleration in technological growth. If all exogenous inputs are held constant, the total elasticity determines how the growth rate of technology $g_A = \dot{A}/A$ changes as A grows, yielding three regimes:

$$\begin{cases} \mathcal{E}_{\dot{A},A} < 1: & g_A \text{ falls as } A \text{ grows} & \Rightarrow & \text{sub-exponential growth (fizzling out),} \\ \mathcal{E}_{\dot{A},A} = 1: & g_A \text{ constant} & \Rightarrow & \text{exponential growth (knife-edge),} \\ \mathcal{E}_{\dot{A},A} > 1: & g_A \text{ rises as } A \text{ grows} & \Rightarrow & \text{self-sustaining acceleration.} \end{cases}$$

That is, A shows self-sustaining acceleration when

$$\mathcal{E}_{\dot{A},A} = \frac{\mathrm{d}\log \dot{A}}{\mathrm{d}\log A} > 1. \tag{1}$$

In the Jones model, since there is a single feedback path $A \to \dot{A}$ with a single edge, growth shows self-sustaining acceleration when $\varepsilon_{\dot{A},A} > 1$.

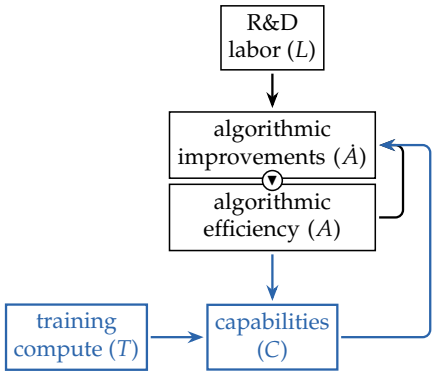
Non-constant elasticities and growth spurts. Self-sustaining acceleration at one point in time does not imply self-sustaining acceleration forever, because the elasticities are not necessarily constant. In the bottlenecks section below we discuss this likely case. Even locally self-sustaining acceleration would be notable.

2.2 Baseline Model of RSI

Our baseline model of RSI adds a few elements.

1. We introduce AI capabilities (C), which are determined by algorithmic efficiency (A) as well as training compute (T). Compute is held fixed, for now. Measuring AI capabilities is notoriously difficult, but two common metrics are METR's time horizon measure (Kwa et al. (2025)) and the Epoch Capabilities Index (ECI, Ho et al. (2025)), though both metrics have flaws.
2. Algorithms are improved ($\dot{A}$) not just by human R&D labor (L) but also with AI capabilities.

Figure 2: Baseline model of RSI



Critically, the fact that AI capabilities now feed into algorithmic improvements creates a new feedback loop: the *core feedback loop* ($A \to C \to \dot{A}$). This loop is in addition to a self-feedback loop ($A \to \dot{A}$) that appeared in the Jones model.

The strength of a multi-step feedback loop like the core loop is determined by the strength of all edges along it. In particular, the strength is the product of all elasticities along the loop, $\varepsilon_{\dot{A},C}\varepsilon_{C,A}$.

Self-sustaining acceleration. Self-sustaining acceleration in algorithmic efficiency now depends on the combined strength of these two feedback loops. The condition is that the *total* elasticity of $\dot{A}$ with respect to A, denoted $\mathcal{E}_{\dot{A},A}$, exceeds one:

$$\mathcal{E}_{\dot{A},A} = \underbrace{\varepsilon_{\dot{A},A}}_{\text{self-feedback loop}} + \underbrace{\varepsilon_{\dot{A},C}\varepsilon_{C,A}}_{\text{core feedback loop}} > 1 \tag{2}$$

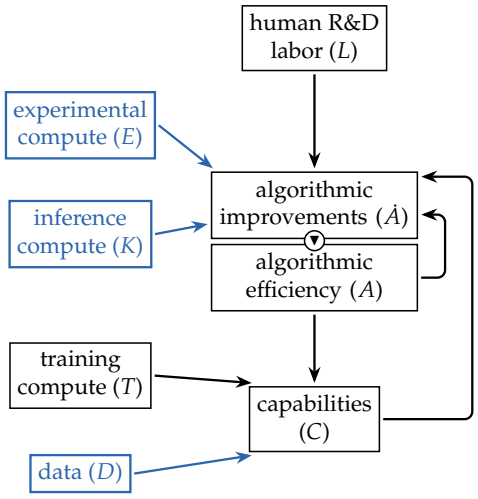
Self-sustaining acceleration in AI capabilities. It is possible for algorithmic efficiency (A) to accelerate while capabilities (C) do not, if the elasticity from A to C is falling, for example if there was a hard ceiling on model capabilities. We discuss the issue further in Section 2.4.

2.3 Bottlenecks: Human, Compute, or Data

The elasticities in our baseline model may fall over time: the feedback loops could become bottlenecked by critical inputs. Whether or not one input may bottleneck an output depends on how *complementary* different inputs are.

To emphasize the possibility of bottlenecks, we modify the baseline model to allow for three more inputs in the system. Experimental compute (E) and inference compute (K) are both used to make algorithmic improvements. Data (D), meanwhile, is used to produce AI capabilities.

Figure 3: Model with bottlenecks



Because the two feedback loops remain the same ($A \to \dot{A}$ and $A \to C \to \dot{A}$), the total elasticity remains the same as above, and thus the self-sustaining condition also remains the same as (2). However these additional terms give us additional reasons why the core elasticity, $\varepsilon_{\dot{A},C}$, may fall: the arrow from $C \to \dot{A}$ may weaken. We discuss these bottlenecks in turn:

- **Labor.** Algorithmic improvements could become bottlenecked by a shortage of human labor if AI agents can only help with some but not all parts of the research process (e.g. due to "lack of taste"). If inference compute increases disproportionately relative to human labor, bottlenecks reduce the marginal value of such compute in research (Jones, 2025).

- **Experimental compute.** AI labs often allocate a large part of their compute to experimentation. If model capabilities and experimental compute are strong complements, then this could be a bottleneck (Whitfill and Wu, 2025; Ho and Whitfill, 2025).

- **Inference compute.** There is clearly a strong interaction between model capabilities and inference compute in producing algorithmic improvements. If they are strong complements then slow growth in inference compute could serve as a bottleneck, diminishing the effect of capabilities on algorithmic improvements.

- **Data.** A lack of high-quality data could bottleneck the advance of capabilities (Farboodi et al., 2025; Erdil et al., 2025; Millidge, 2025). On the other hand, Ho et al. (2024) argue that most algorithmic progress is "data-augmenting", implying that data will not become a bottleneck.

- **Training compute.** A shortage of compute for training could also bottleneck the advance of AI capabilities (Gundlach et al., 2025).

Technical Details: The $C \longrightarrow \dot{A}$ arrow

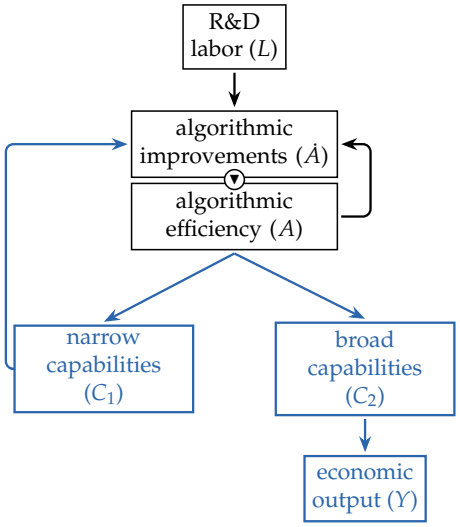
Aside on AI capabilities $\rightarrow$ algorithmic improvements Note that human researchers (L), AI capabilities (C) and experimental compute, E, combine to produce algorithmic progress. Previous literature, such as Eth and Davidson (2025) and Whitfill and Wu (2025), has tried to parametrize this function. But our view is that it is not obvious what this function looks like because it depends on "the returns to intelligence" and how humans and AI systems differ, both of which we have a poor understanding of. Luckily the only thing we need to know is the elasticity noted above and it may be possible to measure this elasticity without fully understanding AI capabilities. We discuss this further in the data section.

2.4 Narrow and Broad Capabilities

Our baseline model treats AI capabilities as a scalar C, which could be interpreted as the capability to both (1) contribute to algorithmic improvements ($\dot{A}$) and (2) contribute to producing economic output (Y). However, the core feedback loop only requires the $C \rightarrow \dot{A}$ arrow, not any arrow from $C \rightarrow Y$.

Here, we make explicit the possibility of a *narrow* acceleration, where algorithmic efficiency accelerates but economic output does not. (For simplicity, this subsection suppresses training compute T.) The figure illustrates the model: we distinguish between a 'narrow' capability (C_1) that is useful for contributing to algorithmic improvements, and a 'broad' capability (C_2) that affects economic output (Y).

Figure 4: Narrow capabilities feed back; broad capabilities dangle.



The clean cut we draw between capabilities useful for algorithmic improvements and all other capabilities is a strong simplification. In reality, it is plausible that a "narrow" acceleration occurs across easier-to-verify tasks which includes optimizing algorithms but also many parts of mathematics, coding writ large, and more. Such progress would have impacts on the real economy but would be exceptionally jagged. It is also of course possible for progress in narrow tasks to eventually spill over into harder-to-verify tasks that matter for the broader economy.

Self-sustaining acceleration. The condition for a self-sustaining acceleration in algorithmic efficiency remains as before, but with narrow capabilities C_1 substituted for C:

$$\mathcal{E}_{\dot{A},A} = \underbrace{\varepsilon_{\dot{A},A}}_{\text{self-feedback loop}} + \underbrace{\varepsilon_{\dot{A},C_1}\varepsilon_{C_1,A}}_{\text{narrow capability loop}}$$

Broad capabilities C_2 do not appear in this expression.

Such an acceleration in algorithmic efficiency A need not imply a self-sustaining acceleration in broad capabilities C_2, and therefore need not imply such an acceleration in economic growth (Y). The condition for broad capabilities depends additionally on a *superelasticity*: how the passthrough from algorithmic efficiency A into broad capabilities C_2 changes as A rises:[4]

$$\varepsilon_{\dot{A},A} + \varepsilon_{\dot{A},C_1}\varepsilon_{C_1,A} + \underbrace{\frac{\mathrm{d}\log \varepsilon_{C_2,A}}{\mathrm{d}\log A}}_{\text{superelasticity}} > 1 \tag{3}$$

Intuitively, the elasticity $\varepsilon_{C_2,A}$ could fall – the superelasticity could be negative – if traditional scaling was sufficient to surpass human abilities at algorithmic optimization (C_1), but not to surpass human abilities in other practical areas (C_2) due to unmodeled data bottlenecks on the latter.[5]

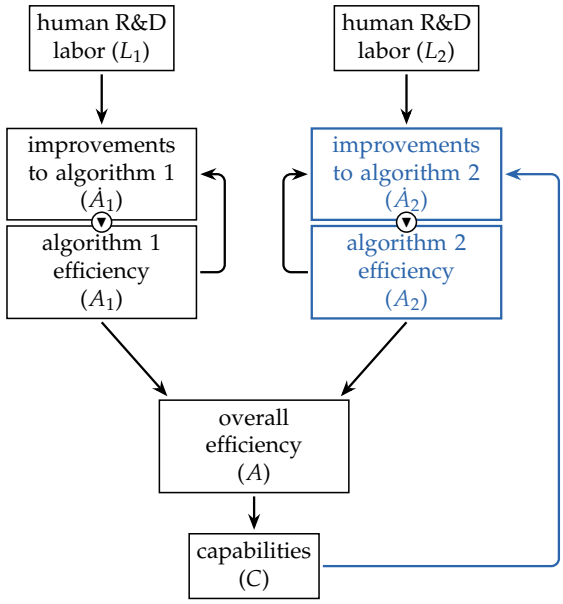
[4] A corresponding condition for acceleration in output depends on an additional superelasticity from C_2 to Y; but for practical purposes we could define broad capabilities by their economic value, so $Y \propto C_2$.

[5] This seems plausible because computers already outperform humans at many well-defined optimization tasks.

2.5 Specific Optimization Ability and Growth Spurts

It seems very likely that AI will be relatively better at contributing to some parts of model training than others. This in turn is likely to cause "growth spurts" in AI progress.

Figure 5: AI helps one subalgorithm



The model at right assumes overall efficiency (A) depends on the efficiency of two sub-algorithms (A_1 and A_2), but AI directly helps improve only the second one, $C \to \dot{A}_2$. To isolate this AI-assisted channel, hold A_1 fixed. The relevant feedback loop is then $A_2 \to A \to C \to \dot{A}_2$. Because $g_A = \varepsilon_{A,A_2} g_{A_2}$, self-sustaining acceleration in overall efficiency requires:

$$\underbrace{\varepsilon_{\dot{A}_2,A_2} + \varepsilon_{\dot{A}_2,C}\varepsilon_{C,A}\varepsilon_{A,A_2}}_{\mathrm{d}\log \dot{A}_2 / \mathrm{d}\log A_2} + \underbrace{\frac{\mathrm{d}\log \varepsilon_{A,A_2}}{\mathrm{d}\log A_2}}_{\substack{\text{changing passthrough} \\ \text{into overall efficiency}}} > 1.$$

If these elasticities were constant then the second term would be zero, and self-sustaining acceleration today would imply self-sustaining acceleration forever. However there are two reasons why the condition may cease to hold, producing only a temporary "growth spurt":

1. **Parallel algorithms.** Overall algorithmic efficiency could be bottlenecked by type-1 algorithmic efficiency, which is not improved by AI capabilities. As type-2 algorithms become abundant, the arrow from $A_2 \to A$ weakens: ε_{A,A_2} falls and its log derivative is negative. For example, if the compute costs of two necessary processes add, so that $1/A = 1/A_1 + 1/A_2$, then as $A_2/A_1 \to \infty$,

$$\varepsilon_{A,A_2} \to 0 \qquad \text{and} \qquad \frac{\mathrm{d}\log \varepsilon_{A,A_2}}{\mathrm{d}\log A_2} \to -1.$$

   Thus a spurt driven by progress on the easy component stalls unless AI also starts accelerating progress on the bottleneck component.

2. **Low ceilings.** If one of the algorithms is already close to a natural ceiling then its elasticity must necessarily fall. Some algorithms are naturally limited, e.g. if kernel efficiency is currently 50%, it can only increase by a factor of two; while the efficiency of other algorithms has improved by orders of magnitude, and may continue to improve by large factors.

Empirically, it will be useful to document which types of algorithms are receiving the

biggest boost from AI-assisted research. On first principles, we might expect bigger boosts where there are low costs of verification (e.g., optimizing inference, optimizing elicitation) than where there are high costs of verification (optimizing architectures). Additionally, it is plausible that improvements to type-1 algorithms do advance capabilities, but only 'narrowly' in the spirit of the model in Section 2.4. We leave combining the two models for future work.

2.6 Economic Feedback Loops

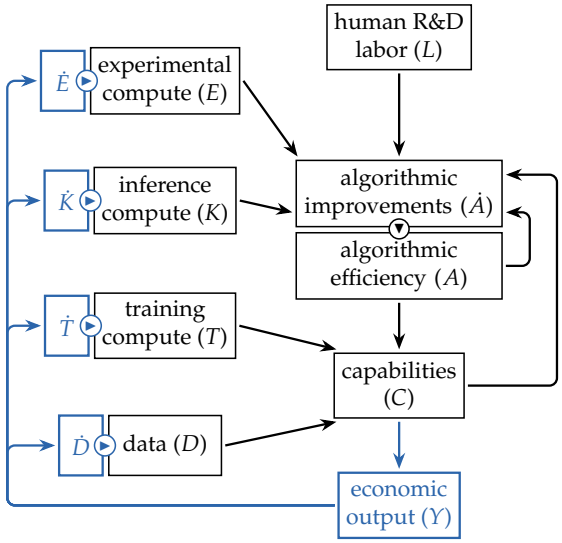
The previous models held constant the level of compute (for training, inference, or experiments) and held constant the quantity of data. This isolated the role of feedback loops driven by algorithmic improvements.

There are also economic feedback loops (Davidson et al., 2026):

1. Increased AI capabilities (C) drive higher economic output (Y).
2. Higher economic output in turn finances further investment in compute (for experiments E, for inference K, and for training T), which feeds back into AI capabilities.

AI-induced economic growth also generates more data, likewise increasing AI capabilities, and further increasing economic growth (Farboodi et al., 2025).

Figure 6: Model with Economic Feedback Loops



We emphasize that economic output (Y) – and the model as a whole – can be interpreted macroeconomically, where Y is GDP, or microeconomically, where Y is the revenue of a single AI firm. In the latter case, as AI capabilities progress and the firm grows, more chips or data are affordable (without substantially altering factor prices). Such microeconomic effects are self-evidently already occurring, and the resulting loops may drive capability acceleration even without substantial macroeconomic impacts.

Feedback loops and self-sustaining acceleration. Aside from the two feedback loops discussed above ($A \rightarrow \dot{A}$ and $A \rightarrow C \rightarrow \dot{A}$), self-sustaining acceleration in algorithmic efficiency now depends on the strength of additional economic feedback loops that run through AI capabilities. For instance, capabilities might drive the accumulation of

high-quality data, which in turn drives capabilities ($C \to Y \to \dot{D} \to D \to C$), or the accumulation of training compute, which in turn drives capabilities ($C \to Y \to \dot{T} \to T \to C$).

The condition for a self-sustaining acceleration is now substantially more complicated and requires taking into account these more complicated feedback loops. The technical details box below states the condition; Davidson et al. (2026) offers a general framework and solution.

Technical Details: Self-sustaining acceleration with economic feedback loops

The prior boxes on technical details apply only when there is a single accumulating state: before, this was merely A. Now, instead, there is an entire state vector:

$$X \equiv (A, T, D, E, K)'$$

Define the reduced-form elasticity matrix M, with $M_{ij} \equiv \frac{\mathrm{d}\log \dot{X}_i}{\mathrm{d}\log X_j}$. Each entry of M measures how strongly one accumulable input increases the accumulation of another, after tracing through the contemporaneous graph.

Davidson et al. (2026) show that there is self-sustaining acceleration in algorithmic efficiency when $\rho(M) > 1$, where $\rho(M)$ is the dominant eigenvalue of the feedback matrix M.

After some math, it can be shown that the eigenvalue condition holds when the sum of three feedback channels exceeds one: the core feedback loop ($A \to C \to \dot{A}$); the indirect loop between economic output and capabilities through training compute and data; and the direct economic loop through experimental and inference compute for algorithmic improvements. In math:

$$\underbrace{\frac{\varepsilon_{\dot{A},C}\varepsilon_{C,A}}{1-\varepsilon_{\dot{A},A}}}_{\text{core feedback loop}} + \underbrace{\varepsilon_{Y,C}\left(\varepsilon_{C,T}+\varepsilon_{C,D}\right)}_{\text{indirect economic loops}} + \underbrace{\frac{\varepsilon_{C,A}\varepsilon_{Y,C}}{1-\varepsilon_{\dot{A},A}}\left(\varepsilon_{\dot{A},E}+\varepsilon_{\dot{A},K}\right)}_{\substack{\text{direct economic loops through} \\ \text{experimental and inference compute}}} > 1. \tag{4}$$

Note that this condition assumes that the direct algorithmic efficiency self-loop does not on its own power a self-sustaining acceleration ($\varepsilon_{\dot{A},A} < 1$).[a] It also imposes the Solow-like assumption that capital-like variables accumulate one-for-one with output: $\varepsilon_{\dot{T},Y} = \varepsilon_{\dot{D},Y} = \varepsilon_{\dot{E},Y} = \varepsilon_{\dot{K},Y} = 1$. If there are no economic impacts of AI capabilities ($\varepsilon_{Y,C} = 0$), the condition collapses to the earlier condition without economic feedback loops.

A subtlety is that the second loop does not run through algorithmic efficiency A and is self-contained: higher capabilities produce more output, financing more training compute and raising capabilities ($C \to Y \to \dot{T} \to T \to C$), or producing more data and raising capabilities ($C \to Y \to \dot{D} \to D \to C$). In either of these loops, the economy achieves a self-sustaining acceleration in algorithmic efficiency A 'merely' by raising GDP, which produces more compute or data (Davidson et al., 2026).

An additional takeaway is that the condition with economic feedback loops, (4), requires estimating five new elasticities: the elasticity of economic output to AI capabilities $\varepsilon_{Y,C}$; the elasticity of capabilities to training compute $\varepsilon_{C,T}$; the elasticity of capabilities to data $\varepsilon_{C,D}$; the elasticity of algorithmic improvements to experimental compute $\varepsilon_{\dot{A},E}$; and the elasticity of algorithmic improvements to inference compute $\varepsilon_{\dot{A},K}$.

[a] Analogously, the rearrangement into this three-term form assumes the indirect economic loop is not self-sustaining on its own, $\varepsilon_{Y,C}\,(\varepsilon_{C,T} + \varepsilon_{C,D}) < 1$; if it were, acceleration would occur through that loop alone.

3 Data Relevant to RSI

We primarily focus this section on calibrating the basic bottlenecks model. This is motivated primarily by simplicity and because larger economic feedback loops will be slower.

To get empirical traction, we modify the bottlenecks model by assuming the existence of an "R&D effort aggregator (R)" which combines human researchers (L), automated AI researchers (running on inference compute, K), and experimental compute (E) into a single scalar. R&D effort is then the sole input to algorithmic improvements ($\dot{A}$). This essentially assumes that all R&D inputs interact with the existing stock of algorithmic progress in the same way. This will let us use historical data on human inputs to calibrate what will happen with AI inputs.

Once we make this assumption, we can expand and re-arrange the condition for self-sustaining acceleration in algorithmic efficiency, (2):[6]

$$\underbrace{\frac{\varepsilon_{\dot{A},R}}{1-\varepsilon_{\dot{A},A}}}_{\text{return to R\&D}} \cdot \underbrace{\varepsilon_{R,C}}_{\substack{\text{research effort's}\\ \text{elasticity to}\\ \text{AI capabilities}}} \cdot \underbrace{\varepsilon_{C,A}}_{\substack{\text{capabilities'}\\ \text{elasticity to}\\ \text{algorithmic}\\ \text{efficiency}}} > 1 \tag{5}$$

[6] The condition (2) delineates a self-sustaining acceleration in algorithmic efficiency. In this section, we are most interested in a self-sustaining acceleration in *AI capabilities*. The condition for capabilities is the same as the condition for efficiency, augmented with a fourth, 'superelasticity' term, $\frac{1}{1-\varepsilon_{\dot{A},A}} \frac{\mathrm{d}\log \varepsilon_{C,A}}{\mathrm{d}\log A}$. This measures the change in the capabilities elasticity $\varepsilon_{C,A}$ as algorithmic efficiency A grows. If $\varepsilon_{C,A}$ is constant, the superelasticity term is zero and the condition reduces to the three terms shown here. In Section 3.4, which we put in a technical box, we argue the evidence indeed points to a superelasticity near zero.

Technical details: rearranging the self-sustaining acceleration condition

The condition for a self-sustaining acceleration in algorithmic efficiency, under the assumption of an R&D effort aggregator R, is

$$\varepsilon_{\dot{A},A} + \varepsilon_{\dot{A},R}\,\varepsilon_{R,C}\,\varepsilon_{C,A} \;>\; 1.$$

This is the same as the core feedback loop condition used in the bottlenecks section, (2), except that now AI capabilities C pass through the R aggregator to affect $\dot{A}$.

Subtracting the direct-spillover term and dividing both sides by $1 - \varepsilon_{\dot{A},A}$ isolates the capability loop:

$$\frac{\varepsilon_{\dot{A},R}\,\varepsilon_{R,C}\,\varepsilon_{C,A}}{1 - \varepsilon_{\dot{A},A}} \;>\; 1.$$

Note that the denominator is positive whenever the direct spillover alone is not already self-sustaining, $\varepsilon_{\dot{A},A} < 1$. Then, grouping the first factor of the first term gives the form above,

$$\underbrace{\frac{\varepsilon_{\dot{A},R}}{1 - \varepsilon_{\dot{A},A}}}_{\text{return to R\&D}} \cdot \underbrace{\varepsilon_{R,C}}_{\substack{\text{research effort's} \\ \text{elasticity to} \\ \text{AI capabilities}}} \cdot \underbrace{\varepsilon_{C,A}}_{\substack{\text{capabilities'} \\ \text{elasticity to} \\ \text{algorithmic} \\ \text{efficiency}}} \;>\; 1.$$

In our view, estimating this quantity, tracking it over time, and extrapolating when it might cross 1 is arguably the most important empirical exercise for assessing whether a self-sustaining acceleration will occur. The rest of this section maps each term in the condition to the data needed to estimate it: Section 3.1 covers the return to research, Section 3.2 covers the elasticity of research effort with respect to AI capabilities, and Section 3.3 covers the response of capabilities to algorithmic progress. We then collect detailed supplementary evidence on model-training inputs and model economics to build a broader picture of AI R&D automation.

3.1 Return to Research

It turns out – derived in the technical details box below – that the first term, the "returns to research", can be measured from two growth rates: the growth rate of algorithmic efficiency ($g_A \equiv \dot{A}/A$) and the growth rate of R&D effort ($g_R \equiv \dot{R}/R$). On a balanced growth path:

$$\underbrace{\frac{\varepsilon_{\dot{A},R}}{1 - \varepsilon_{\dot{A},A}}}_{\text{return to R\&D}} = \frac{g_A}{g_R}$$

We therefore need to estimate the growth rate of algorithmic efficiency, g_A, and the growth rate of research effort, g_R:

1. We can measure growth in algorithmic efficiency g_A directly from model-training data.

2. We derive in a technical box below that one can measure R&D effort as the *spend-weighted average of growth rates in inputs*: human researchers L, experimental compute E, and inference compute used for research K. Using S to denote spend shares,

$$g_R = \sum_{\text{input}} S_{\text{input}}\, g_{\text{input}}$$

Alternatively, we can measure g_R directly from spending. Total R&D spend is the sum across inputs of price times quantity, and log-differentiating shows its growth splits into a price component and a quantity component:

$$g_{\text{spend}} = \underbrace{\sum_{\text{input}} S_{\text{input}}\, g_{\text{price of input}}}_{\equiv g_{\text{prices}},\ \text{input price inflation}} + \underbrace{\sum_{\text{input}} S_{\text{input}}\, g_{\text{input}}}_{= g_R,\ \text{by the result above}}$$

so $g_R = g_{\text{spend}} - g_{\text{prices}}$: growth in total R&D spend, deflated by an index of R&D input prices.

Model primitive	**Existing data**	**Data ask**
Growth in algorithmic efficiency (g_A)	Ho et al. (2024) estimate pretraining algorithmic progress of 3x a year (i.e. compute needed to achieve a given level of capability is 1/3 as much each year). Whitfill et al. (2025) estimate 3.5x/year for the entire stack. Ho (2026) estimates 10x/year for the entire stack. Gundlach et al. (2025) argue that algorithmic progress has grown at different rates across scales: 1.5x/year at small scale, and 2.5x/year at frontier scale, for pretraining.	**Current data weakness**: Compute estimates for models since 2024 are uncertain. **Ideal new data**: model training compute after 2024 to estimate g_A.

Model primitive	Existing data	Data ask
Share of human researchers (L) of R&D spend, and its growth g_L Its share S_L and growth g_L enter $g_R = \sum_{\text{input}} S_{\text{input}}\, g_{\text{input}}$.	Epoch AI (2026) estimates growth rates of around 2–3x/year in total staff for frontier labs over 2023–2025. Whitfill and Wu (2025) estimate headcount growth rates of 2–3X/year for Anthropic and OpenAI over 2022–2024. Cottier et al. (2024) estimates that around 1/3 of the cost of training models was spent on researcher salaries.	**Current data weakness**: Employee headcount does not necessarily map to the researcher headcount. **Ideal new data**: Granular lab data on staff, split by seniority and by researchers vs. engineers, would let us treat these as separate inputs rather than an aggregate L.
Share of experimental compute (E) of R&D spend, and its growth g_E Its share S_E and growth g_E enter $g_R = \sum_{\text{input}} S_{\text{input}}\, g_{\text{input}}$.	You (2025) estimates that in 2024 only 10% of OpenAI's R&D compute was on final training runs, while 90% was on experimental compute (i.e. training which did not directly affect weights used in a public product). Denain and Wu (2026) estimate similar ratios for Chinese companies. Morrison et al. (2026) estimate similar ratios for OLMo 3 family. You et al. (2026) finds that the total growth rate of chips is about $3.3x$ a year (doubling every 7 months).	**Current data weakness**: Experimental compute share in total compute is not the same as the experimental compute share of total R&D spending. The composition of experimental compute is also not clear. **Ideal new data**: Labs could publish detailed data on experimental compute spend across domains and over time, which would help estimate how much experimental compute bottlenecks automated research.
Share of inference compute (K) of R&D spend, and its growth g_K Its share S_K and growth g_K enter $g_R = \sum_{\text{input}} S_{\text{input}}\, g_{\text{input}}$.	No data available.	**Ideal new data**: Labs could publish detailed data on inference compute spend across domains and over time.

Model primitive	Existing data	Data ask
Total R&D spend in dollars, and its growth Deflated by an index of input prices, this is an alternative way to measure g_R.	Denain and Wu (2026) estimate R&D compute spending for several labs from public filings.	**Current data weakness**: No spending over time yet. Even given spend, no off-the-shelf deflator fits AI R&D, since compute prices fall rapidly while wages do not. **Ideal new data**: Labs could publish total annual R&D expenditure alongside a self-constructed deflator.
Substitutability of labor and experimental compute How substitutable labor and compute are controls how fast $\varepsilon_{R,C}$ falls as research becomes bottlenecked by experimental compute.	Whitfill and Wu (2025) try to estimate the substitutability between labor and experimental compute in AI R&D, using variation in relative prices, but their findings are inconclusive.	**Current data weakness**: Aside from data quality, it is unclear whether the need for experimental compute increases with higher training compute. **Ideal new data**: labs could conduct experiments that would give teams randomized amounts of R&D labor and experimental compute, but this is expensive. One cheaper substitute is using AI labor in place of human labor in the experiment. Cheaper still is to survey managers how much more progress they would make with (A) twice the researchers (compute fixed); (B) twice the compute (researchers fixed); (C) both doubled.

Technical details: cost minimization and elasticities

Throughout this section we use the correspondence that, under cost-minimizing input choice, the elasticity of an output with respect to an input equals that input's share of expenditure.

Setup. Let output $R = R(X_1, \ldots, X_n)$ be produced from inputs X_i with prices p_i. A firm minimizing the cost $\sum_i p_i X_i$ of attaining a target $\bar{R}$ has first-order conditions $p_i = \mu \, \partial R / \partial X_i$ for a common multiplier μ. The cost share of input i is then

$$S_i \equiv \frac{p_i X_i}{\sum_j p_j X_j} = \frac{(\partial R / \partial X_i)\, X_i}{\sum_j (\partial R / \partial X_j)\, X_j}.$$

Constant returns to scale. We take R to have constant returns to scale. This is without loss of generality: R is a latent index, and its returns to scale are absorbed by the elasticity $\varepsilon_{\dot{A},R}$ in the return-to-research term. Under constant returns, Euler's theorem gives $\sum_j (\partial R / \partial X_j)\, X_j = R$, so

$$S_i = \frac{\partial R}{\partial X_i} \frac{X_i}{R} = \varepsilon_{R,X_i}.$$

Cost shares equal elasticities.

When firms do not optimize. If input choices are not fully cost-minimizing, the equality becomes a revealed-preference bound: observed cost shares reflect the firm's own belief about marginal products, so they remain informative about the elasticities.

Technical details: return to research

The return to research as g_A / g_R. Write the discovery rate as $\dot{A} = f(R, A)$, so $g_A \equiv \dot{A}/A = f(R, A)/A$. Log-differentiating with respect to time,

$$\frac{\mathrm{d} \log g_A}{\mathrm{d}t} = \varepsilon_{\dot{A},R}\, g_R - \left(1 - \varepsilon_{\dot{A},A}\right) g_A.$$

On a balanced growth path g_A is constant, so the left-hand side is zero and

$$\frac{g_A}{g_R} = \frac{\varepsilon_{\dot{A},R}}{1 - \varepsilon_{\dot{A},A}},$$

which is exactly the return-to-research term. This holds exactly on a balanced growth path and approximately when g_A is approximately constant. The smooth, steady historical rate of algorithmic progress makes this a reasonable benchmark, though not one we expect to hold if RSI pushes us off the balanced-growth path.

Decomposing g_R. The growth rate of research effort follows the growth-rate chain rule across its inputs,

$$g_R = \sum_{\text{input } i} \varepsilon_{R,i}\, g_i.$$

Using the cost-share correspondence above, $\varepsilon_{R,i} = S_i$, so

$$g_R = \sum_{\text{input } i} S_i\, g_i,$$

a spend-weighted average of input growth rates over human labor L, experimental compute E, and inference compute K.

3.2 Research Effort to Capabilities

The second term measures how strongly effective research effort responds to more capable AIs. This is the hardest of the three to measure, since it captures how AI capabilities translate into real-world research effort. This requires scientifically understanding AI capabilities and how they feed into research. The science of AI capabilities is in its infancy.

One possible route is to measure it directly, through studies of how much AI accelerates end-to-end research.

A second route uses a cost-share argument. If we model "effective AI labor" as a Cobb–Douglas function of AI capabilities and the quantity of AI researchers (inference compute), e.g. $C^\gamma K$, then this term equals γS_K, where γ is the returns to capabilities relative to scaling the quantity of AI labor and S_K is the inference-compute cost share (derived in the technical details below).

Model primitive	Existing data	Data ask
Share of AI R&D inference compute (K) of total R&D spend (S_K) The inference spend share S_K, one of the two components of $\varepsilon_{R,C} = \gamma S_K$.	As aforementioned in the last subsection, we have almost no data on lab expenditure on inference or its share relative to other R&D inputs.	**Ideal new data**: Labs could publish detailed data on inference compute spend across domains and over time.

Model primitive	Existing data	Data ask
Tradeoff between training and inference compute (γ) The quality–quantity exponent γ, the other component of $\varepsilon_{R,C} = \gamma S_K$.	The parameter γ governs the returns to scaling capabilities C versus scaling inference compute K, since we model effective AI labor as $C^\gamma K$. Estimates from Villalobos and Atkinson (2023) imply $\gamma \approx 0.15$–0.3 under the ECI normalization of capabilities ($C = e^{\text{ECI}}$).[7]	**Current data weakness**: The estimation is not up to date, and the implied γ divides estimates from two separate studies with different tasks and model generations. **Ideal new data**: More extensive study on public benchmarks/with ECI.
The effect of AI use on researcher productivity A direct (if loose) measure of $\varepsilon_{R,C}$: how much capability raises effective R&D.	We have self-reported productivity gains from surveys: 1.4–2X in METR's survey of technical workers (Becker, 2026), and roughly 4X in the Claude Mythos Preview system card (Anthropic, 2026). The magnitudes are hard to interpret, but the upward trend is clear. Favaro and Clark (2026) report lines of code merged per engineer per day rose roughly 8X between 2024 and Q2 2026.	**Current data weakness**: Note that this is the productivity effect of AI, but $\varepsilon_{R,C}$ is the total research effort added due to one extra unit of capability. **Ideal new data**: An RCT that measures productivity benefits from extra capabilities.

Model primitive	Existing data	Data ask
The strength of autonomous R&D A direct measure of $\varepsilon_{R,C}$ (and the discovery-rate response $\varepsilon_{\dot{A},C}$): AI's standalone contribution to discovery.	AI agents have recently advanced the frontier in some optimization problems; however, it is difficult to find a metric to measure their overall ability.[8] Recent system cards report mixed results on benchmarks of AI R&D. Models generally outperform humans on optimization problems when the humans are given between 8 and 40 hours, but show low pass rates on complex debugging problems (Claude Mythos Preview system card, Anthropic, 2026, Table 2.3.3.A; GPT-5.5 system card, section 9.1.3). These results are difficult to place on a common scale.	**Current data weakness**: Benchmark questions are not the same as real-life LLM training problems. **Ideal new data**: How autonomous AI systems perform on real LLM training problems would be especially informative. The ideal data would characterize the returns to expenditure on agentic, human, and hybrid labor.

Model primitive	Existing data	Data ask
Capabilities of AI R&D A direct, qualitative read on $\varepsilon_{R,C}$ — which parts of research AI substitutes for, and where it remains weak.	Si et al. (2025) find that AI-generated AI research ideas are rated highly relative to human-generated ones. Trehan and Chopra (2026) find a range of weaknesses, including both taste and execution. The Claude Mythos Preview system card has an extended discussion of the model's qualitative strengths and weaknesses in AI R&D (Anthropic, 2026).	**Current data weakness**: It remains unclear how to describe AI's weaknesses in AI R&D relative to human researchers. A common split is between idea generation (taste) and idea execution. **Ideal new data**: Unstructured interviews with lab staff with well-designed questions eliciting AI R&D weaknesses.

Technical details: research effort to capabilities

Suppose AI enters research effort through a capability-adjusted count of effective researchers, $N = K\,C^{\gamma}$, where K is the quantity of inference compute and C^{γ} is a quality adjustment, so that $R = R(\dots, N, \dots)$. By the chain rule,

$$\varepsilon_{R,C} = \frac{\partial \log R}{\partial \log C} = \frac{\partial \log R}{\partial \log N}\,\frac{\partial \log N}{\partial \log C} = \varepsilon_{R,N} \cdot \gamma,$$

since $\partial \log N / \partial \log C = \gamma$. Effective researchers are purchased through inference compute, so the expenditure on this input is the inference-compute spend and its cost share is S_K. Applying the cost-share correspondence (see the technical details on cost minimization above), $\varepsilon_{R,N} = S_K$, giving

$$\varepsilon_{R,C} = \gamma\, S_K.$$

3.3 Capabilities Response to Algorithmic Efficiency

The third term is the elasticity of capabilities with respect to algorithmic progress. Suppose that algorithmic efficiency and training compute enter C symmetrically (both through effective compute, $C = C(A \times T)$). This implies that the elasticity of capabilities to algorithmic progress equals the elasticity of capability to training compute i.e. $\varepsilon_{C,A} = \varepsilon_{C,T}$. Note that $\varepsilon_{C,T}$ is precisely the object estimated by scaling-law studies. Epoch's work on the capabilities–compute relationship (Ho et al., 2025; Epoch AI, 2025) estimates values

of $\varepsilon_{C,T}$ around 6.5 when C is defined as the exponential of the Epoch Capabilities Index, e^{ECI}.[9]

One qualification is needed: the symmetry $\varepsilon_{C,A} = \varepsilon_{C,T}$ relies on algorithmic progress being scale-free, so that it enters capability only through the product $A \times T$. Gundlach et al. (2025) present evidence that algorithmic progress is instead scale-biased (faster at frontier scale than at small scale for pretraining), in which case $\varepsilon_{C,A}$ and $\varepsilon_{C,T}$ can diverge and scaling-law estimates identify only the latter. Nonetheless, we treat this term as the best-measured of the four; the supporting model-training data is collected in the supplementary material (Section 3.5).

Model primitive	Existing data	Data ask
Capabilities' elasticity to algorithmic progress ($\varepsilon_{C,A}$)	Ho et al. (2025) and Epoch AI (2025) estimate the capability–compute relationship, implying $\varepsilon_{C,A} \approx 6.5$ when $C = e^{\text{ECI}}$.	**Current data weakness**: Estimates assume scale-free algorithmic progress; Gundlach et al. (2025) suggest it is scale-biased. **Ideal new data**: Algorithmic efficiency estimates at multiple scales, to separate $\varepsilon_{C,A}$ from $\varepsilon_{C,T}$.

Technical details: Measuring the superelasticity term

3.4 The Superelasticity Term

As described in footnote 6, the condition for a self-sustaining acceleration in capabilities (5) technically has a fourth term. The fourth term asks whether the capabilities elasticity $\varepsilon_{C,A}$ itself changes as algorithmic efficiency grows, i.e. whether $\mathrm{d}\log\varepsilon_{C,A}/\mathrm{d}\log A \neq 0$. Our best guess is that it is close to zero. As we showed in the previous subsection, $\varepsilon_{C,A} = \varepsilon_{C,T}$ under certain conditions. Thus the superelasticity measures the proportional change in $\varepsilon_{C,T}$ as effective compute grows, which depends on the curvature of the capability–compute relationship rather than its slope. That relationship has looked close to log-linear across many orders of magnitude of training compute (Ho et al., 2025; Epoch AI, 2025), so $\varepsilon_{C,T}$ appears approximately constant and therefore the superelasticity near zero.

The caveat is the same scale-free assumption as in the previous subsection: if algorithmic progress is scale-biased (Gundlach et al., 2025), the symmetry breaks, and constancy of $\varepsilon_{C,T}$ need not imply constancy of $\varepsilon_{C,A}$.

[9] The magnitude depends on this normalization of capabilities; the loop gain $\varepsilon_{R,C}\,\varepsilon_{C,A}$ that enters the condition is invariant to it.

Model primitive	Existing data	Data ask
Superelasticity of capabilities ($\frac{\mathrm{d}\log \varepsilon_{C,A}}{\mathrm{d}\log A}$)	No direct estimates, but the capability–compute relationship appears close to log-linear across many orders of magnitude of compute (Ho et al., 2025; Epoch AI, 2025), which suggests $\varepsilon_{C,T}$, and by symmetry $\varepsilon_{C,A}$, is approximately constant.	**Current data weakness**: Scale-biased algorithmic progress would break the symmetry with $\varepsilon_{C,T}$. **Ideal new data**: Capability–compute estimates spanning a wide range of effective compute, updated past 2024, to estimate curvature and not just slope.

3.5 Supplementary Data

The data below do not measure the three terms in (5) directly but fill out the broader picture of AI R&D automation, including economic feedback loops. For example, the model-training laws row helps pin down the third term, the capability–compute scaling relationship (Section 3.3). The same data also help identify historical growth in algorithmic efficiency and predict the effects of future growth in training compute and data. The final row covers the economic feedback loop, through which capability funds further inputs.

Model primitive	Existing data	Data ask
Model training laws Calibrates the capability map $C(AT)$; pins down $\varepsilon_{C,A} = \varepsilon_{C,T}$.	The basic model remains Chinchilla (Hoffmann et al., 2022), which maps data and compute to pretraining loss.	**Current data weakness**: Scaling laws for RL remain largely unpublished. **Ideal new data**: post-training compute usage of models.
Growth in training compute g_T; since effective compute is AT, it drives capability growth.	Training compute has grown 4–5X per year since 2010 (Rahman and Owen, 2024). Whitfill et al. (2025) discusses possible GDP constraints on continued compute scaling.	**Current data weakness**: There is scarce data on model training compute since 2024. **Ideal new data**: training compute of new models.

Model primitive	Existing data	Data ask
Growth in data Data D, an input to capability beyond effective compute AT.	There are minimal existing estimates of data spend by labs.[10]	**Ideal new data**: Labs could publish data spend, or partial estimates could be constructed from third-party data providers such as Surge AI or Scale AI.
Share of AI's economic value creation captured by AI labs Economic output Y, the channel through which capability funds further inputs (the economic feedback loop).	We can use firm revenue estimates, but these are likely to represent only a small share of the value created. Brynjolfsson et al. (2026) use online choice experiments to elicit willingness-to-accept for giving up generative AI chatbots, estimating that consumer surplus substantially exceeds firms' revenues.	**Current data weakness**: More estimation needed, with time variation. **Ideal new data**: AI firms' revenue over time.
Elasticity of GDP to capabilities ($\varepsilon_{Y,C}$) How much a given capability improvement raises aggregate output.	Little direct evidence: AI's measured contribution to GDP remains small relative to its measured capabilities. Yotzov et al. (2026) survey over 5000 executives: 9/10 report no impact of AI on their firm's productivity or employment over the past three years, but on average they expect AI to raise firm productivity by 1.4% and output by 0.8% over the next three years.	**Current data weakness**: Adoption and usage data are not tightly linked to impact on output. **Ideal new data**: Task-level adoption and usage data linked to firm or sector output, tracking how output effects grow as capabilities improve.

Model primitive	Existing data	Data ask
Elasticity of AI firm budget to capabilities ($\varepsilon_{Y,C}$) We might instead interpret Y as the budget AI firms have to spend on data, compute, and researchers. Then, $\varepsilon_{Y,C}$ is how much additional budget a given capability improvement generates (via fundraising or revenues).	Frontier-lab revenues have grown around 3x/year (Epoch AI, 2026) with notable outliers, e.g. Anthropic. But lab expenditure is largely financed by external investment rather than revenue so the loop runs through expectations.	**Current data weakness**: Revenue is observed but at low frequency. Lab budgets are largely financed by investor expectations rather than current revenue. **Ideal new data**: Credible estimates for how lab financing costs (through equity or debt) respond to changes in capabilities progress.

4 Discussion

To assess the plausibility of self-sustaining acceleration, we provide a very rough calibration of our analytical condition with the available estimates from Section 3. Certain parameters are unknown, especially the elasticity of R&D to AI capabilities. Thus, we derive a condition on the range of values for this elasticity that implies self-sustaining acceleration. We have a high degree of uncertainty in many of these empirical estimates, and also acknowledge the possibility of model misspecification. Thus we additionally provide a general discussion of the most important points of evidence for and against self-sustaining acceleration.

4.1 Model calibration

We now use the estimates we have from Section 3 to calibrate the self-sustaining acceleration condition (5):

$$\underbrace{\frac{\varepsilon_{\dot{A},R}}{1-\varepsilon_{\dot{A},A}}}_{\text{return to R\&D}} \cdot \underbrace{\varepsilon_{R,C}}_{\substack{\text{research effort's} \\ \text{elasticity to} \\ \text{AI capabilities}}} \cdot \underbrace{\varepsilon_{C,A}}_{\substack{\text{capabilities'} \\ \text{elasticity to} \\ \text{algorithmic} \\ \text{efficiency}}} > 1.$$

We plug in the following values:

- $\frac{\varepsilon_{\dot{A},R}}{1-\varepsilon_{\dot{A},A}} = \frac{g_A}{g_R} \approx \frac{\ln(3)}{\ln(3)} = 1$. We take $g_A = \ln(3)$ as the annual growth rate in algorithmic progress per Ho et al. (2024), meaning A triples each year. Next recall $g_R = S_L g_L + S_E g_E + S_K g_K$. We take $g_L \approx \ln 3$ from Epoch AI (2026); Whitfill and

Wu (2025) and $g_E \approx \ln(3)$ from You et al. (2026). g_K was probably small until the rise of coding agents but has since increased. Indeed OpenAI reported in the GPT-5.6 launch post in July 2026 that "over the past six months, the share of research compute devoted to internal coding inference grew 100-fold," implying a much larger g_R at present. But since we do not know the spend shares on these factors, and S_K is likely still low relative to S_L and S_E, we assume $g_R = \ln(3)$ for now. This assumption is further justified by the fact that g_A was estimated in 2024.

- $\varepsilon_{R,C}$. We have almost no evidence for this parameter because AI uplift studies do not report capability improvements in a form that maps cleanly to this elasticity.

- $\varepsilon_{C,A} \approx 6.5$. Define capabilities as the exponential of the Epoch Capabilities Index, $\exp(\text{ECI})$. Also recall the assumption that $C = C(A \times T)$ which implies $\varepsilon_{C,A} = \varepsilon_{C,T}$. Thus we can get the elasticity of capabilities to algorithmic efficiency from the slope of a regression of ECI on $\log(\text{compute})$ across models using data from Epoch AI which yields an estimate of roughly 6.5.[11]

Plugging in these values tells us that self-sustaining acceleration is achieved if $\varepsilon_{R,C} > 0.15$. This is the increase in R&D effort per additional unit of ECI. For comparison, Claude Opus 4.8 scores one ECI unit more than Claude Opus 4.7. Thus under our model, Anthropic would meet the self-sustaining acceleration condition if adopting Claude Opus 4.8 led to 15% higher research productivity than Opus 4.7. Note that the increase in research output per unit of AI capabilities tends to be smaller than the uplift for human tasks from AI use, because labor only accounts for a fraction of all research tasks.

It is hard to say whether one unit of the ECI is worth 15% more research effort, but a rough calculation based on the last two years of AI progress reassuringly suggests that the returns to capabilities are currently lower than this threshold. Claude Opus 4.8 scores 16 ECI points more than Claude 3.7 Sonnet, which was released in February 2025 alongside Claude Code. At the time, estimates for AI uplift on engineering tasks varied but were not conclusively larger than 0.[12] More recently, Anthropic engineers surveyed in the Claude Mythos Preview system card reported a 4X uplift from Claude use (Anthropic, 2026). The 4X uplift is very likely to be an overestimate, but even if it were true it would imply a 9% increase in productivity per unit of ECI, below the threshold for self-sustaining acceleration over this period.

Of course, these elasticities are likely increasing – the move from GPT-4 to GPT-5 led to a much larger increase in research productivity than the move from GPT-1 to GPT-2. The continual measurement of those elasticities is necessary for understanding the current state of RSI. Understanding how and why $\varepsilon_{R,C}$ changes over time is therefore an important research priority.[13]

[11] Setting aside the scale-bias in algorithmic progress, discussed below.

[12] For example, an influential study from Becker et al. (2025) found a negative impact of AI on engineering productivity.

[13] One promising direction here is to model automation as endogenous: as AI capabilities improve, which research tasks do labs choose to automate, and how do those choices affect the aggregate responsiveness of research effort to capabilities?

4.2 Arguments for and against acceleration

Given the uncertainty in the above calibration, we now turn to general qualitative arguments for and against acceleration, informed by our models and data.

Best evidence in favor of acceleration:

- AI systems are meaningfully contributing to AI progress. The Claude Mythos Preview system card reported a self-assessed productivity uplift of roughly 4X among Anthropic researchers (although there are reasons to think this is likely overstated), and a 40-hour time horizon at which Claude Mythos Preview beat human researchers (Anthropic, 2026). Favaro and Clark (2026) assess the performance of human researchers versus Claude Code in making AI research decisions. In April 2026, Mythos Preview beat humans 64% of the time, up from 50% for Claude models released in 2025. OpenAI reported in July 2026 that internal coding inference's share of research compute grew 100-fold in the prior six months, suggesting that the value of AI in R&D has dramatically increased.

- In August 2025, experts and superforecasters (METR, 2025)[14] predicted an 8–20% chance that the growth rate of effective compute would triple by 2029. In May 2026, Jack Clark predicted a 60% chance that by the end of 2028, there will be "an AI system powerful enough that it could autonomously build its own successor" (Clark, 2026).

Best evidence against acceleration:

- Compute bottlenecks could bind for multiple reasons. Algorithmic improvements may require increases in training compute (Gundlach et al., 2025). Cognitive labor increases may require increases in experimental compute (Whitfill and Wu, 2025). Productivity growth may not be fast enough to support continued AI R&D investments, that is, we "run out of GDP" (Halperin, forthcoming).

- Autonomous contributions to algorithm optimization may quickly hit diminishing returns. This has been widely observed informally and formalized in the apple-picking model of AI R&D developed by Cunningham and Shetty (2026).

- Capabilities may depend on data. Millidge (2025) argues that "most algorithmic progress is data progress". This includes everything from better filtering of pre-training data to the substantial resources spent on post-training data. Pre-training filtering seems to be a one-time boost, whereas post-training data might be supplied more elastically and be amenable to automation. We are only beginning to learn how to build better reinforcement learning environments, and AI systems could help design future ones.

[14] "Experts and superforecasters both find a three-fold acceleration of effective compute scale-up by 2029 plausible but unlikely, with experts giving it a median 20% chance and superforecasters 8%."

Narrow capabilities acceleration before broad capabilities acceleration. It seems plausible that AI R&D agents, while effective at making concrete algorithmic progress as captured by benchmarks, might remain less effective at developing major breakthroughs that enable models to learn efficiently in new environments, or develop more amorphous capabilities like 'vision' or 'taste'. This idea would be captured by combining our models in Sections 2.4 and 2.5 so that the feedback loop from narrow capabilities is helpful for generating narrow algorithmic progress but not broad progress. Nonetheless, spillovers from narrow to broad progress remain possible. For example, an advanced narrow system, in pursuit of benchmark optimization, may discover new capabilities, such as highly sample-efficient algorithms able to update their weights during deployment, that quickly turn narrow progress into broader progress.

Political and social barriers. Even if a self-sustaining acceleration is technically plausible, political and social barriers could stop it in its tracks. In America, at least 20 data center projects were canceled in the first three months of 2026 due to local pushback; $85 billion worth of projects have been canceled in the last three years (The Economist, 2026). Recent policy reactions to Mythos by the Trump administration, despite an earlier laissez-faire stance, also suggest placing greater weight on these barriers. If compute scaling remains important, even minor frictions in model deployment will slow revenue growth and further scaling.

4.3 Impacts of a capabilities acceleration on the world

A self-sustaining acceleration in AI progress differs from generic AI progress in *speed*: it could compress generations of progress into short periods of time. Indeed, if self-sustaining progress were feasible, it is likely that the returns to increasing investment in inputs like compute and data would be high, further increasing investment and the pace of progress. We highlight several aspects of rapid progress that distinguish it from gradual AI progress, and that may be especially important for understanding its social, economic, and political consequences.

- **Transition costs:** It seems plausible that transition costs increase with the speed of technological change. Labor markets are one such case: the costs of structural transformation may rise with its speed, as some economists argue was the case with the China Shock.
- **Institutional inertia:** In a world of rapidly accelerating capabilities growth, institutions broadly defined — regulatory capacity, democratic procedures, legal frameworks, infrastructure, and so on — might lag behind (Bostrom, 2014; Ord, 2020). Rapid technological change might also grant political elites a temporary spike in *de facto* power, which they could use to capture institutions.
- **Asymmetric progress across domains:** As noted, we see self-sustaining acceleration in narrow capabilities as more likely in the near term than an acceleration in broad capabilities. The former would likely be concentrated in tasks that are easy to verify

in closed loops, such as software development, games, mathematics, and related areas. This could create acute asymmetries in progress across domains. For example, we may find solutions to long-standing puzzles in mathematics and breakthroughs in computational chemistry that enable new materials, alongside sluggish progress in less self-contained or messier domains.

- **Power imbalances across firms and nation-states:** Even model providers with relatively small gaps in technical skill could see large gaps open up in effective capabilities and power over the world. Imagine Anthropic or the US government having access to capabilities that are vastly superior to those of any other actor.
- **Greater misalignment risk:** Misaligned AI might also create bigger risks under a capabilities acceleration. Sudden increases in model capabilities would provide less time for technical and social safeguards to be put in place.

4.4 Future Steps

There are several directions for future work that build off this note, some of which we are actively pursuing:

- Empirically estimating the self-sustaining acceleration condition for the model with economic feedback loops.
- Measuring the return to research off the balanced growth path.
- Combining models with multiple types of AI algorithms and capabilities (Sections 2.4 and 2.5).
- Distinguishing quality from quantity improvements in data and compute.
- Extending the model to multiple research sectors (software, hardware, data).
- Modeling research taste and vision in AI R&D.
- Endogenizing the decision to automate AI R&D.
- Incorporating directed technical change.
- Surveys and field experiments within AI frontier labs to improve estimates of the key parameters in the model.
- Developing new ways to incentivize participation in prediction markets to produce better forecasts of AI progress (Fradkin et al. (2026)).

REFERENCES

Philippe Aghion, Benjamin F Jones, and Charles I Jones. Artificial intelligence and economic growth. Technical report, National Bureau of Economic Research, 2017.

Anthropic. System card: Claude mythos preview. Technical report, Anthropic, April 2026. URL https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf.

Joel Becker. Measuring the self-reported impact of early-2026 AI on technical worker productivity. METR, 2026. URL https://metr.org/blog/2026-05-11-ai-usage-survey/. Published May 11, 2026.

Joel Becker, Nath Rush, Beth Barnes, and David Rein. Measuring the impact of early-2025 ai on experienced open-source developer productivity. 07 2025.

Nicholas Bloom, Charles I Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

Erik Brynjolfsson, Avinash Collis, Felix Eggers, Sophia Kazinnik, and David Nguyen. What is generative AI worth? Technical report, Stanford Digital Economy Lab, 2026. URL https://digitaleconomy.stanford.edu/publication/what-is-generative-ai-worth/.

Alan Chan, Ranay Padarath, Joe Kwon, Hilary Greaves, and Markus Anderljung. Measuring ai r&d automation. *arXiv preprint arXiv:2603.03992*, 2026.

Paul Christiano. Takeoff speeds. The Sideways View, February 2018. URL https://sideways-view.com/2018/02/24/takeoff-speeds/. Accessed 2026-06-27.

Jack Clark. Import AI 455: AI systems are about to start building themselves. Import AI newsletter, 2026. URL https://importai.substack.com/p/import-ai-455-automating-ai-research. Published May 4, 2026.

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen. How much does it cost to train frontier AI models? Epoch AI, 2024. URL https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models. Published June 3, 2024.

Tom Cunningham and Manish Shetty. An apple-picking model of AI R&D. tecunningham.github.io, 2026. URL https://tecunningham.github.io/posts/2026-03-13-apple-picking-ai.html. Published April 7, 2026.

Tom Davidson. What a compute-centric framework says about takeoff speeds. Technical report, Open Philanthropy, 2023. URL https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/.

Tom Davidson and Thomas Houlden. How quick and big would a software intelligence explosion be? 2025. URL https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion-be. Accessed: 2026-06-27.

Tom Davidson, Basil Halperin, Thomas Houlden, and Anton Korinek. When does automating AI research produce explosive growth? Feedback loops in innovation networks. NBER Working Paper 35155, National Bureau of Economic Research, April 2026. URL https://www.nber.org/papers/w35155.

Jean-Stanislas Denain and Cheryl Wu. Final training runs account for a minority of R&D compute spending. Epoch AI, Gradient Updates, 2026. URL https://epoch.ai/gradient-updates/r-and-d-vs-training-compute. Published March 23, 2026.

AI Epoch. Key trends and figures in machine learning. *Published online at epochai. org. Retrieved from: https://epochai. org/trends*, 2026.

Epoch AI. Epoch capabilities index (ECI). Epoch AI, 2025. URL https://epoch.ai/eci. Accessed June 2026.

Epoch AI. Data on AI companies. Epoch AI, 2026. URL https://epoch.ai/data/ai-companies. Accessed June 2026.

Ege Erdil, Tamay Besiroglu, and Anson Ho. Estimating idea production: A methodological survey, 2024. URL https://arxiv.org/abs/2405.10494.

Ege Erdil, Matthew Barnett, and Tamay Besiroglu. How to fully automate software engineering. Mechanize, Inc. blog, May 2025. URL https://www.mechanize.work/blog/how-to-fully-automate-software-engineering/. Accessed: 2026-06-05.

Daniel Eth and Tom Davidson. Will AI R&D automation cause a software intelligence explosion? Fore-

thought Research Report, mar 2025. URL https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.

Maryam Farboodi, Andrew Koh, and Anchi Xia. Data-driven automation. Technical report, National Bureau of Economic Research, 2025.

Marina Favaro and Jack Clark. When AI builds itself. Anthropic, 2026. URL https://www.anthropic.com/institute/recursive-self-improvement. Accessed: 2026-06-08.

Andrey Fradkin, Brian Jabarian, and Andrew Koh. We need well-capitalized prediction markets for ai impacts. Justified Posteriors, Substack, May 2026. URL https://empiricrafting.substack.com/p/we-need-well-capitalized-prediction. Accessed 2026-06-27.

Irving John Good. Speculations concerning the first ultraintelligent machine. In Franz L. Alt and Morris Rubinoff, editors, *Advances in Computers*, volume 6, pages 31–88. Elsevier, 1965. doi: 10.1016/S0065-2458(08)60418-0.

Hans Gundlach, Alex Fogelson, Jayson Lynch, Ana Trisovic, Jonathan Rosenfeld, Anmol Sandhu, and Neil Thompson. On the origin of algorithmic progress in AI, 2025. URL https://arxiv.org/abs/2511.21622.

Basil Halperin. Running out of gdp: Straight lines to long timelines. forthcoming.

Robin Hanson. Economics of the singularity. *iEEE SpEctrum*, 45(6):45–50, 2008.

Anson Ho. The least understood driver of AI progress. Epoch AI, Gradient Updates, 2026. URL https://epoch.ai/gradient-updates/the-least-understood-driver-of-ai-progress. Published February 25, 2026.

Anson Ho and Parker Whitfill. The software intelligence explosion debate needs experiments. Epoch AI, Gradient Updates, 2025. URL https://epoch.ai/gradient-updates/the-software-intelligence-explosion-debate-needs-experiments.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models, 2024. URL https://arxiv.org/abs/2403.05812.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A Rosetta Stone for AI benchmarks, 2025. URL https://arxiv.org/abs/2512.00193. Epoch AI; underlies the Epoch Capabilities Index (ECI).

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Benjamin Jones. Artificial intelligence in research and development. Technical report, National Bureau of Economic Research, 2025.

Charles I. Jones. R&D-based models of economic growth. *Journal of Political Economy*, 103(4):759–784, 1995. doi: 10.1086/262002.

Charles I. Jones. A.i. and our economic future. NBER Working Paper 34779, National Bureau of Economic Research, 2026. Issued January 2026, revised June 2026.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long tasks, 2025. URL https://arxiv.org/abs/2503.14499. METR.

METR. Forecasting the impacts of AI R&D acceleration: Results of a pilot study. METR, 2025. URL https://metr.org/blog/2025-08-20-forecasting-impacts-of-ai-acceleration/. Published August 20, 2025.

Beren Millidge. Most algorithmic progress is data progress. beren.io, 2025. URL https://www.beren.io/2025-08-02-Most-Algorithmic-Progress-is-Data-Progress/. Published August 2, 2025.

Joel Mokyr. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press, Princeton, NJ, 2002.

Jacob Morrison, Noah A. Smith, and Emma Strubell. The hidden cost of thinking: Energy use and environmental impact of lms beyond pretraining, 2026. URL https://arxiv.org/abs/2605.01158.

Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books, New York, 2020.

Robi Rahman and David Owen. The training compute of notable AI models has been doubling roughly every six months. Epoch AI, Data Insights, 2024. URL https://epoch.ai/data-insights/compute-trend-post-2010. Published June 19, 2024.

George P Richardson. Problems with causal-loop diagrams. *System dynamics review*, 2(2):158–170, 1986.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.04109.

The Economist. America's data-centre backlash puts the ai boom at risk. *The Economist*, June 2026. URL https://www.economist.com/business/2026/06/23/americas-data-centre-backlash-puts-the-ai-boom-at-risk.

Philip Trammell and Anton Korinek. Economic growth under transformative AI. NBER Working Paper 31815, National Bureau of Economic Research, 2025. Originally issued 2023, revised 2026; forthcoming, Annual Review of Economics.

Dhruv Trehan and Paras Chopra. Why LLMs aren't scientists yet: Lessons from four autonomous research attempts, 2026. URL https://arxiv.org/abs/2601.03315.

Pablo Villalobos and David Atkinson. Trading off compute in training and inference. Report, Epoch AI, July 2023. URL https://epoch.ai/publications/trading-off-compute-in-training-and-inference.

Parker Whitfill and Cheryl Wu. Will compute bottlenecks prevent an intelligence explosion?, 2025. URL https://arxiv.org/abs/2507.23181.

Parker Whitfill, Ben Snodin, and Joel Becker. Forecasting AI time horizon under compute slowdowns, 2025. URL https://arxiv.org/abs/2511.19492.

Ivan Yotzov, Jose Maria Barrero, Nicholas Bloom, Philip Bunn, Steven J. Davis, Kevin M. Foster, Aaron Jalca, Brent H. Meyer, Paul Mizen, Michael A. Navarrete, Pawel Smietanka, Gregory Thwaites, and Ben Zhe Wang. Firm data on AI. NBER Working Paper 34836, National Bureau of Economic Research, 2026.

Josh You. Most of OpenAI's 2024 compute went to experiments. Epoch AI, Data Insights, 2025. URL https://epoch.ai/data-insights/openai-compute-spend. Published October 10, 2025.

Josh You, Venkat Somala, Yafah Edelman, and Luke Emberson. Global ai computing capacity is doubling every 7 months, 2026. URL https://epoch.ai/data-insights/ai-chip-production. Accessed: 2026-06-28.

Eliezer Yudkowsky. Creating friendly ai 1.0: The analysis and design of benevolent goal architectures. *The Singularity Institute, San Francisco, USA*, 2001.

Eliezer Yudkowsky. Recursive self-improvement. *LessWrong*, 2008.

APPENDIX

A SCALE-DEPENDENT ALGORITHMIC PROGRESS

Since algorithmic progress is usually not directly observable, it is typically estimated as the increase in capabilities not explained by increases in training compute, just as TFP is estimated by the Solow residual. That is, if we know (from a "scaling law") that in the absence of algorithmic progress some index of AI capabilities C would track training compute T, it is conventional to propose the model

$$C = A \cdot T \tag{6}$$

and calculate A as C/T. Using a model like our "baseline model" in Section 2.2, the resulting time series for A, in combination with a time series for research inputs R and the semi-endogenous AI R&D function, can then be used to generate a prediction for how fast algorithms would advance after the automation of AI R&D.

Gundlach et al. (2025) show conclusively that model (6) is misspecified because algorithmic progress in AI development to date has been largely biased toward large scales. For an extreme case of large-scale-biased algorithmic progress, consider instead the model

$$C = \min(A, T) \cdot T. \tag{7}$$

Here, algorithmic improvements proportionally advance AI capabilities as long as there is the training compute to leverage them (i.e., here, as long as $T > A$). If we stopped scaling the compute, however, capabilities would ultimately stagnate (here, at T^2). Likewise, even if automating AI R&D yields a leap in the quality of the algorithms available, under (7) this will not deliver arbitrarily large advances in AI capabilities until we have accumulated the training compute to leverage these algorithms.

Another potential explanation for Gundlach et al.'s finding, however, is that we have been developing large-scale-biased algorithms precisely *because* we have been scaling compute. Indeed, we might imagine living in a somewhat different world in which we knew exactly what architectures would take us to superintelligence but were simply too compute-constrained to make efficient use of them. Why don't we live in that world? One answer is that we have no incentive to discover such algorithms—and would not be able to validate them if we did—in a world in which we are compute-constrained. This also suggests that if we faced a long-term compute constraint, or if an AI model were tasked with recursively self-improving on fixed compute, we might develop algorithms that made ever better use of "small" training runs—as academic labs do to some extent today. That is, rapid RSI may still be feasible, but only via some kind of "directed technical change".