Artificial Intelligence in Research and Development

Benjamin F. Jones*

September 22, 2025

Abstract

How much can AI accelerate progress in different research fields? This paper shows that three features – the share of research tasks AI performs, the productivity of AI at those tasks, and the strength of bottlenecks—are key determinants of AI's implications in any area, from cancer therapeutics to software design. The model maps changes in AI capabilities to research outcomes, quantifies the "marginal returns to intelligence," and shows how AI can shift returns to R&D investment. Concepts like superintelligence, Powerful AI, and Transformative AI are further engaged and disciplined. Finally, the framework sets a measurement agenda linking AI benchmarks to field-specific opportunities for accelerating progress.

*Kellogg School of Management and NBER. Email: bjones@kellogg.northwestern.edu. I thank the editors, Ajay Agrawal, Anton Korinek, and Erik Brynjolfsson, for their helpful guidance; Ronnie Chatterji and Zoë Hitzig for their insights on AI benchmarks; and Bronwyn Hall for her excellent comments.

"I believe that in the AI age, we should be talking about the *marginal returns to intelligence*, and trying to figure out what the other factors are that are complementary to intelligence and that become limiting factors when intelligence is very high. We are not used to thinking in this way—to asking "how much does being smarter help with this task, and on what timescale?" — but it seems like the right way to conceptualize a world with very powerful AI."

– Dario Amodei, CEO Anthropic (October 2024)

1 Introduction

Amidst advancing capabilities of artificial intelligence (AI), economists, technologists, policymakers, and broader society are seeking to understand its manifold implications. In economics, models suggest that substantial automation in the production of ordinary goods and services can lead to large effects on both economic growth and inequality (e.g., Zeira 1998; Acemoglu and Restrepo 2018, 2020; Aghion et al. 2019, Jones and Liu 2024). But beyond ordinary goods and services, it is now understood that AI can influence the "ideas production function" - i.e., the research and development (R&D) process - and thereby directly and perhaps sharply increase rates of progress for wide-ranging outcomes. Growth models suggest that AI's role in accelerating ideas production may drive especially powerful dynamics for the economy (Aghion et al. 2019, Trammell and Korinek 2023) and some observers expect accelerating advances for human health (King 2025, Zhou et al. 2025). Meanwhile, AI tools are increasingly used in research - for prediction, discovery, literature reviews, generating code, building datasets, writing and editing, and other tasks (Chen et al. 2025). Domain-specific AIs appear potentially transformative in certain research fields, like GNoME (Merchant et al. 2023) for material discovery or the Nobel-Prize winning AlphaFold (Jumper et al. 2021) in structural biology.

This paper provides a framework to assess how AI may impact R&D. The centerpiece is a model that, while simple, is broad in three senses. First, the model includes both "machines" and "people" as inputs to R&D. By machines, we mean AI but also any kind of research-related machines (microscopes, centrifuges, particle accelerators, etc.). To the

extent that the innovation literature has paid less attention to capital inputs to R&D (in favor of studying human inputs to R&D), the model provides a framework for incorporating machines in a fairly general manner. Second, the model allows machines to advance flexibly both in (i) the range of tasks machines can perform and (ii) how good machines are at various tasks. Related, while the model can be used to study modest advances in AI, it also allows AI to become extraordinarily capable, providing closed-form solutions where AI takes over large shares of tasks that were previously performed by research labor or where AI supercedes human performance at specific research tasks by large multiples. Finally, the model remains open to various R&D objectives: gains in productivity, health, national security, or other dimensions. That is, rather than embed R&D into a general equilibrium context where the outcomes are macroeconomic variables (as in models of economic growth), this model will be open about the outcomes of interest and ask how to allocate a given R&D budget to advance that outcome. For example, the outcome might be relatively narrow (e.g., determine the structure of a protein, write a section of code), somewhat broader (e.g., produce a new drug for a specific cancer, or increase the productivity of a specific firm), or very broad (e.g., increase overall longevity or economy-wide total factor productivity).

The model emphasizes three key features that will determine AI's impact. These features are: (a) the share of research tasks that AI can perform; (b) the productivity of AI at these heterogeneous research tasks; and (c) the strength of bottlenecks in ideas production. The model shows that these three features are essential for understanding whether AI will, or will not, significantly accelerate progress on dimensions of interest - that taking a stand on these three features is central to assessing what we may expect from AI in any research area. These features also point to key empirical objects, defining a measurement agenda that is critical to making any strong claims about AI in R&D. Importantly, while the future capabilities of AI are uncertain, the conceptual framework clarifies simple empirical constructs - including some available measures today - that can inform what we might expect.

The model can also elucidate the "marginal return to intelligence," to borrow an evocative phrase from Dario Amodei (Amodei 2024), who has wondered whether abundant, extremely intelligent AIs will massively or modestly advance rates of progress. The model develops explicit results for assessing the R&D implications of computer servers full of

hyper-intelligent synthetic researchers or even superintelligence, which some technologists and industry leaders suggest is close at hand. More generally, the model determines the rate of progress per dollar spent on R&D, conditional on the research technologies available. Thus we can see how advances in machine intelligence will change the relationship between a unit of R&D investment and a unit of progress. By applying some value weight to the progress, such as value-added output for a productivity gain or quality-adjusted-life-year valuation to a health improvement, one can then calculate a rate of return.

This paper is organized as follows. Section 2 situates the modeling approach in relation to existing literature. Section 3 presents the baseline model. Section 4 applies the model to assess the implications of specific AI capabilities. Section 5 examines the potential for Transformative AI. Section 6 discusses applications to specific research areas and the relationship between the model and explicit AI benchmarks. Section 7 concludes.

2 Conceptual Approach and Prior Literature

This paper views R&D as a set of activities or "tasks." These tasks can be understood flexibly. They could be high-level steps familiar to science, such as defining a research question, developing a conceptual framework and/or empirical approach, collecting and analyzing data, and writing up results. Or they could concern product development, involving steps like design, prototyping, and testing. Research tasks could also be defined with increasing specificity. More specific steps will vary substantially depending on the field and specific research question - from running difference-in-difference models in economics to using phosphoproteomics techniques in cancer biology.[1] As such, we might apply the model, and its key measures, differently to different types of research.

[1]Procedures, techniques, and capital equipment vary greatly across fields. For example, a project in economics using U.S. administrative data could involve: conceiving of a research question and its empirical strategy, proposing the project and obtaining approval from U.S. Census, accessing a nearby Research Data Center, learning data dictionaries and database structures, implementing empirical strategies by writing code for statistical analysis, iteratively producing and assessing the main empirical findings and robustness analyses, and completing the disclosure process. By contrast, in designing a cancer therapeutic, research steps could include identifying an oncogenic mutation and determining relevant protein structure and function. Beyond the protein's "fold," post-translational modifications via phosphorylation will inform protein activity, including cancer-relevant roles in cellular processes. Experimental steps may utilize cryo-electron microscopy, nuclear magnetic resonance relaxation, single-cell phosphoproteomics, and mass spectrometry. Subsequent

Taking a task-based view allows us to leverage recent advances in task-based models (Zeira 1998; Autor et al. 2003; Acemoglu and Restrepo 2018; Aghion et al. 2019; Jones and Liu 2024). The approach used here is closest to Jones and Liu (2024), which incorporates the standard idea that machines can take over tasks previously performed by labor while also allowing for machines to become especially good at these tasks - a feature that seems essential for understanding potential gains from AI. This paper is distinct from most prior task-oriented models in focusing on the ideas production function and in taking a micro approach (as opposed to a general equilibrium, growth-oriented approach). Specifically, we consider how a research team, laboratory, public research institution, or business can maximize the rate of progress per dollar spent on R&D, taking factor prices as given.[2]

Other recent contributions have developed important insights into how AI can influence R&D. Taking creativity perspectives, one view is that AI can improve upon exploratory search into uncertain terrain (Gans 2025). Another view it that AI can better leverage combinatoric possibilities (Agrawal et al. 2024). According to the related "burden of knowledge" viewpoint, individual researchers have increasingly narrow expertise, as one person can know only a shrinking share of aggregate knowledge, the more humanity's collective knowledge accumulates (Jones 2009, Hill et al. 2025). The capacity of AI to aggregate wide-ranging knowledge may thus be an essential advantage, allowing AI to overcome human creative limits. Tools like GNoME already suggest this AI potential for identifying new materials (as combinations of molecules) or chemical synthesis pathways (as combinations of steps) that may be hard or costly for human researchers to see (Segler et al. 2018, Merchant et al. 2023). This paper incorporates creative search in only a simple way by focusing on "tasks," which can embrace both creative search tasks and research execution tasks. While some key insights can be lost by being broad in this way, the benefit is that

steps in drug design involve finding a compound that disrupts the oncogenic pathway with sufficient potency and minimal side effects, tested across cells, animals, and humans, and using additional specialized methods.

[2]To the extent that R&D is a very small share (on the order of 3%) of the overall economy, taking capital and labor prices as given seems a reasonable place to start. To the extent that AI simultaneously, substantially affects many sectors of the economy, one may further consider shifts in factor prices, in general equilibrium settings. See Aghion et al. (2019) and Trammell and Korinek (2023). But also see the discussion of "double bottlenecks", which will further constrain AI's aggregate potential, in Section 5. At root, if the bottlenecks in this paper are sufficiently germane, then the partial equilibrium analysis is more fitting. If the bottlenecks are overcome, then additional general equilibrium forces would come into play.

we can provide a more embracing view of research activities. One can then apply the model at various levels of focus, where measures of tasks describe different aspects of the research process. We will further discuss the nature of key research activities and their implications for assessing AI's potential in Section 5.

3 The Model

We consider R&D efforts that seek to advance some outcome. Let the outcome of interest be measured as Z_t. In an economic context, Z_t could be productivity in a specific production process, or perhaps productivity across an entire production chain in a given sector. In a health context, Z_t could be longevity in the face of a specific heart ailment, like a cardiac arrhythmia, or perhaps longevity given heart disease as a whole. The type and breadth of outcome measured by Z_t is important for assessing AI's potential impact, as we will discuss later. But as a starting point, focusing on R&D activity, we can be generic: there is a desire to improve some measure, Z_t, which we do by applying inputs into research and development activities.

To proceed, define the "idea production function" (the mapping between R&D inputs and resulting improvement in Z_t) as:

$$\dot{Z}_t \;=\; \zeta\, Z_t^{\varphi}\Big[\int_0^1 r_t(j)^{\theta}\, dj\Big]^{1/\theta}, \qquad \theta < 0 \tag{1}$$

Here, there is a unit measure of research tasks, indexed by $j \in [0,1]$, and $r_t(j)$ is the output at a given research task. The parameter θ governs the degree of complementarity between tasks - i.e., the strength of "bottlenecks." Finally, the rate of progress may depend on the current state of Z_t itself, as governed by the parameter φ, so that research advances might become easier ($\varphi > 0$) or harder ($\varphi < 0$) as progress continues.

To introduce AI, we imagine that a given task may be performed either by machines or humans. In particular, we imagine that humans can do all these tasks, but that machines have been created over time that perform some fraction of these tasks. Moreover, machines may greatly exceed human capacity at specific tasks. For instance, instruments such as telescopes, microscopes, and thermometers can outperform human senses of observation, while computers outstrip human cognition at floating-point arithmetic, regression modeling, and information retention.

Task-level production is

$$r_t(j) = \begin{cases} m_t(j)\, x_t(j), & 0 \leq j < \gamma_t \text{ (machine tasks)}, \\ H\, l_t(j), & 0 \leq j \leq 1 \text{ (human tasks)}. \end{cases} \tag{2}$$

where $x_t(j)$ is the quantity of capital applied to a given machine task and $l_t(j)$ is the quantity of labor applied to a given human task. The measure γ_t is the share of tasks that can currently be automated - i.e., the share that can be done by AI or other machines. This measure allows us to study the implications of AI as it takes over more research tasks. The terms $m_t(j)$ and H represent the productivity of machines and humans at specific tasks. For simplicity, we let research labor have the same productivity at all tasks. Meanwhile, machine inputs have task-specific productivities. These productivity parameters allow us to study the implications of AI as it becomes better at research tasks and potentially greatly exceeds human capacities at many tasks.

In addition to task productivities, we also have input costs. Research labor will have a wage, w_t. Capital inputs will have a cost, μ_t. These are measured relative to a numeraire good, and where relevant we will take the numeraire good as GDP with a price of 1.

Finally, it will be helpful to define a machine-task productivity index, as follows.

$$M_t = \left[\frac{1}{\gamma_t} \int_0^{\gamma_t} m_t(j)^{\frac{\theta}{1-\theta}}\, dj \right]^{\frac{1-\theta}{\theta}} \tag{3}$$

This index will prove a useful summary statistic for "how good AI is at the tasks it performs," which will be central to several applications below. Note that this index is the generalized mean function, so that M_t is an average of all the underlying machine productivities, $m_t(j)$, where the type of average depends on the parameter θ.[3]

In sum, we have a set up that can flexibly engage key technology features for understanding the potential impact of AI. These three features are:

- γ_t, informing the share of research tasks that AI may perform;
- M_t, informing how good AI is at these tasks;
- θ, informing the strength of bottlenecks.

[3]The generalized mean is a function that takes various means from a list of numbers, where the type of mean depends on the value of θ. These include the arithmetic ($\theta = 1$), geometric ($\theta = 0$), and harmonic means ($\theta = -1$) as well as the Leontief or min function ($\theta = -\infty$).

3.1 Optimization

We consider knowledge production given a fixed research budget. We assume that the firm, research institution, or single laboratory seeks to maximize the rate of progress at some outcome. Their choice problem is how to allocate research dollars, given capital and labor prices, their budget, and the available set of technologies. Thus research institutions or research teams can increasingly shift spending toward AI depending on how successfully it evolves. If Z_t is measured in the same scale as R&D expenditure (i.e., in dollars, say by applying value-of-life or value-added output measures to Z_t), the optimization results inform the average return per dollar spent on R&D, as well as the marginal return to an additional R&D dollar and the effects of technological advance on these returns.

3.2 Technology Adoption Condition

Before turning to the broader problem of allocating capital and labor to research tasks, we first consider a technology adoption condition at the individual task level. That is, since research labor can do any task, for machines to actually be deployed to a task j, the machine must be the cost effective option. From (2), we therefore require that the task output per dollar spent on the machine (i.e., $m_t(j)/\mu_t$) exceed the task output per dollar spent on human labor (i.e., H/w_t), if the machine is to be used. We can then define, for a specific task, the relative cost advantage of AI to a human as

$$c_t(j) = \frac{m_t(j) w_t}{H \mu_t} \tag{4}$$

where we only adopt AI for a given task when $c_t(j) \geq 1$. This also implies a minimum value of $m_t(j)$, which is $m_t^{min} = H\mu_t/w_t$. Below this threshold, a machine approach is not worth deploying - one would rather use labor instead. Therefore, in what follows, we assume that automation technologies satisfy the following requirement.

Assumption 1 *(Technology Adoption Condition)* $m_t(j) \geq H\mu_t/w_t$ *for all* $j \in [0, \gamma_t]$.

We can further develop a related technology metric, which represents the relative cost efficiency of machines over labor overall. This metric will appear repeatedly in our analysis below. Specifically, define

$$C_t = \frac{M_t w_t}{H \mu_t} \tag{5}$$

to represent a relative cost advantage of machines. The minimum value $m_t^{min}(j)$ in turn implies a minimum value of the overall machine productivity index M_t. Namely, since M_t is an average of the $m_t(j)$, it follows that $M_t \geq m_t^{min}(j)$. Thus the minimum value of the technology index is also $M_t^{min} = m_t^{min} = H\mu_t/w_t$, and therefore we have the following necessary technology condition,

$$C_t \geq 1 \tag{6}$$

As we will see, in understanding AI's implications for R&D, the model will focus our attention on the share of research tasks it can perform, γ_t, and its average productivity, M_t, across these tasks. One can construct γ_t and M_t from the vector of underlying task-level productivities and costs. We will consider such construction empirically in Section 5, when linking the model with AI benchmarking studies.[4]

3.3 The Optimized Allocation Given R&D Expenditure

Now consider how research teams and institutes make use of research machines and labor. Define the total R&D budget as D_t. The budget constraint is

$$D_t = \mu_t X_t^r + w_t L_t^r \tag{7}$$

where X_t is total research capital and L_t is total research labor. Solving the constrained optimization problem leads to the following result.

Proposition 1 *(Fixed Total R&D Expenditure) Maximizing the rate of progress in the outcome Z_t given the R&D budget constraint (7) and given capital and labor prices, leads*

[4]As an underlying technology process, one may imagine that the creation and implementation of AI applications leads to an increase in γ_t and also shifts M_t by averaging in the new machine productivities. See Jones and Liu (2024) for an endogenous technology process along these lines. A more parsimonious approach might focus purely on the distribution of machine productivities, the $m_t(j)$, and how this evolves. Namely, one can define an underlying distribution of machine productivities, $F_t(m)$ across the entire unit measure of research tasks. Thus machines exist in some sense for all tasks, but they may be very bad (e.g., $m = 0$) at many tasks and are not deployed. Then we view γ_t as an endogenous feature, which is the measure of machine tasks for which the technology adoption condition holds, i.e., $\gamma_t = 1 - F_t(m_t^{min})$. For this paper, and the applications in mind, we'll stay at a higher level of abstraction and focus on exogenous evolutions of γ_t and M_t. See also the discussion by Bronwyn Hall.

to the growth rate

$$\dot{Z}_t/Z_t = \frac{\zeta Z_t^{\varphi-1} D_t}{\left[\gamma_t \left(\frac{\mu_t}{M_t}\right)^{\frac{\theta}{\theta-1}} + (1-\gamma_t)\left(\frac{w_t}{H}\right)^{\frac{\theta}{\theta-1}}\right]^{\frac{\theta-1}{\theta}}} \tag{8}$$

All heterogeneity in the machine-task productivities is summarized by the single index M_t.

Proof. See Online Appendix. ■

This shows that the growth rate of Z_t initially increases linearly in total investment D_t. Thus, this equation also provides the initial marginal return to an additional dollar of R&D expenditure, where the outcome is measured as the rate of progress. We see directly the rate of progress will increase with the composite productivity of AI at research tasks (M_t). The rate of progress also increases in the share of R&D tasks performed by machines (γ_t).[5]

As outcomes of interest, and to help calibrate the exercises to follow, we can further compute the research labor and machine expenditure shares.

Corollary 1 *The labor share and capital share of R&D expenditure are*

$$s_t^L = \frac{1}{1 + \frac{\gamma_t}{1-\gamma_t} C_t^{\frac{\theta}{1-\theta}}} \quad , \quad s_t^X = \frac{1}{1 + \frac{1-\gamma_t}{\gamma_t} C_t^{\frac{\theta}{\theta-1}}} \tag{9}$$

where $C_t = \frac{M_t w_t}{H \mu_t}$. The expenditure shares sum to 1 and are bounded in ranges $s_t^L \in [1-\gamma_t, 1]$ and $s_t^X \in [0, \gamma_t]$.

Proof. See Online Appendix. ■

4 AI Applications

With Proposition 1, we can now ask what will happen to the rate of progress as AI advances, potentially by a lot. We first consider what happens when AI becomes much more productive at a given set of R&D tasks. Will the rate of progress at outcome Z_t greatly increase? We then consider AI's capacity to automate a larger share of research tasks. We then examine both of these forces together.

[5] To see this directly, rearrange Proposition 1 in terms of the index C_t and recall that $C_t \geq 1$.

4.1 The Return to Machine Intelligence

We'll start by analyzing machine intelligence - i.e., M_t, representing how good AI is at a given set of tasks. Our thought experiment is that AI intelligence surges ahead. Perhaps AI becomes very powerful at a given set of tasks. To what extent will the rate of progress increase?

Our first result considers short-run gains from a small increase in machine intelligence. By short run, we mean the instantaneous increase in the rate of progress holding the initial level of the outcome fixed. The distinction between short-run and longer-run effects is discussed in Section 4.4.

Corollary 2 *(Small increase in machine intelligence) The short run elasticity of the rate of progress to a small increase in machine intelligence is*

$$\frac{d \log \dot{Z}_t}{d \log M_t} = s_t^X \tag{10}$$

where $s_t^X \leq 1$ is the machine expenditure share in R&D, as given in terms of exogenous parameters in (9). This elasticity declines with greater machine intelligence ($\uparrow M_t$) and rises with a greater share of automated tasks ($\uparrow \gamma_t$).

Proof. See Online Appendix. ■

Corollary 2 provides a simple answer to the question of "how will a small increase in machine intelligence affect the rate of progress?" by pointing to a single observable measure: the capital expenditure share. To get a sense of this measure, note that U.S. and OECD sources report labor shares of R&D expenditure of around 2/3, which suggests that we might take $s_t^X \approx 1/3$.[6] That said, much of the non-labor expenditure includes structures and material inputs, not simply "machines", and a large portion of these non-labor R&D inputs are not obviously amenable to being substituted for by AI. Thus, while we will often use $s_t^X \approx 1/3$ in what follows, we will also consider the case where only a portion of these non-labor inputs can be substituted with AI. In a companion discussion, Bronwyn Hall further examines measures of expenditure shares and their implications.

One related interpretation of Corollary 2 is that machine intelligence has fundamentally limited effects on the rate of progress. In particular, the elasticity of the rate of progress

[6] See, for example, U.S. Census BRDIS data Table 25 (National Science Foundation (2019) and also Besiroglu et al. 2024.

to machine intelligence can be no greater than 1 (i.e., s_t^X can be no greater than 1 in any eventuality). However, this "small changes" result also appears too optimistic, as it will overstate the acceleration in progress from larger changes in machine intelligence, even in the short run. This is because the elasticity falls as M_t rises. Our second result therefore directly considers the immediate effects of a large change in M_t.

Specifically, let's imagine that machine intelligence suddenly increases by a multiple λ, so that $M_t \to \lambda M_t$. This advance in AI can be arbitrarily large. We ask what multiple will occur in the rate of progress.

Corollary 3 *(Large increase in machine intelligence) Let machine intelligence increase by a multiple λ. The rate of progress, $\dot{Z}_t$, will initially increase by a multiple*

$$\eta = \left(1 - (1 - \lambda^{-b}) s_t^X\right)^{-1/b} \tag{11}$$

where s_t^X is the machine expenditure share in R&D, as given in terms of underlying technology measures in (9), and $b = \frac{\theta}{\theta - 1}$. In the limit where $M_t \to \infty$, the rate of progress increases by a multiple $\eta_\infty = (1 - s_t^X)^{-1/b}$ for $\theta < 0$.

Proof. See Online Appendix. ■

These results help reveal the potential impact of extraordinarily smart machines using straightforward measures. Figure 1 implements (11) given an initial expenditure share ($s^X = 1/3$) and various views of the bottleneck parameter, θ. We see that θ is powerful in governing returns to machine intelligence. It sets a fundamental "speed limit" of sorts, per the last result of the corollary. For example, consider $\theta = -1$, so that progress depends on the harmonic average of the task outputs. If the initial capital expenditure share is 1/3, then an infinite increase in machine intelligence causes the rate of progress to increase by a factor of $\frac{2}{3}^{-2} = 2.25$. That is, the rate of progress would a bit more than double with infinite productivity across the entire current set of non-labor tasks.

The role of bottlenecks may be even stronger, however, as it would also operate *among* the non-labor inputs. The above thought experiment imagines that M_t, the overall index of machine productivity, increases by large multiples. But AI – machine-embodied intelligence – is only one type of non-labor input or research machine. More generally, R&D uses a wide range of machinery beyond computing power. These can be tools of observation, including

laboratory staples such as microscopes, centrifuges, and PCR machines, to extremely large detectors such as space and land-based telescopes, particle accelerators, nuclear reactors, and gravitational wave detectors. Buildings are also an important capital cost for R&D, to hold laboratories and their machines. Given these other expenses, we might edit our thought experiment to consider what happens when some fraction of current non-labor tasks are those that can be performed by AI. Call this fraction ν. Bottlenecks then appear additionally in mapping from large gains in intelligence to the overall machine productivity index, M_t.

Corollary 4 *(Large increase in machine intelligence for relevant sub-tasks) Let AI perform a fraction ν of the non-labor tasks. Let the distribution of machine productivity initially be the same for non-AI tasks and for AI tasks, and let the AI's productivity at all AI tasks increase by a multiple κ. The overall machine productivity index, M_t, will increase by a multiple*

$$\lambda_\nu = \left(\nu\kappa^{-b} + (1-\nu)\right)^{-1/b} \tag{12}$$

where $b = \frac{\theta}{\theta-1}$. In the case of infinite machine intelligence, this multiple limits to $\lim_{\kappa\to\infty} \lambda_\nu = (1-\nu)^{-1/b}$ for $\theta < 0$. The rate of progress, $\dot{Z}_t$, will increase by a multiple

$$\eta_\nu = \left(1 - \nu(1-\kappa^{-b})s_t^X\right)^{-1/b} \tag{13}$$

which in the case of infinite machine intelligence limits to $\lim_{\kappa\to\infty} \eta_\nu = \left(1 - \nu s_t^X\right)^{-1/b}$ for $\theta < 0$.

Consider again the example where $\theta = -1$, so that progress depends on the harmonic average of the task outputs. Let the initial non-labor expenditure share be 1/3 and let AI perform 1/2 of these non-labor tasks. Then an infinite increase in machine intelligence causes the rate of progress to increase by a factor of $\frac{5}{6}^{-2} = 1.44$. That is, taking economic growth, this would say that the growth rate would increase by 44% with infinite productivity across the set of AI tasks. This upward shift in the rate of progress is much smaller than when infinite productivity advances occur for all non-labor inputs. In Corollary 3, we are imagining labor tasks remain a substantial share of tasks and thus provide the key constraints. In Corollary 4 the remaining bottleneck tasks are both labor and half of the non-labor inputs (i.e, experimental machines like particle accelerators, telescopes, PCR

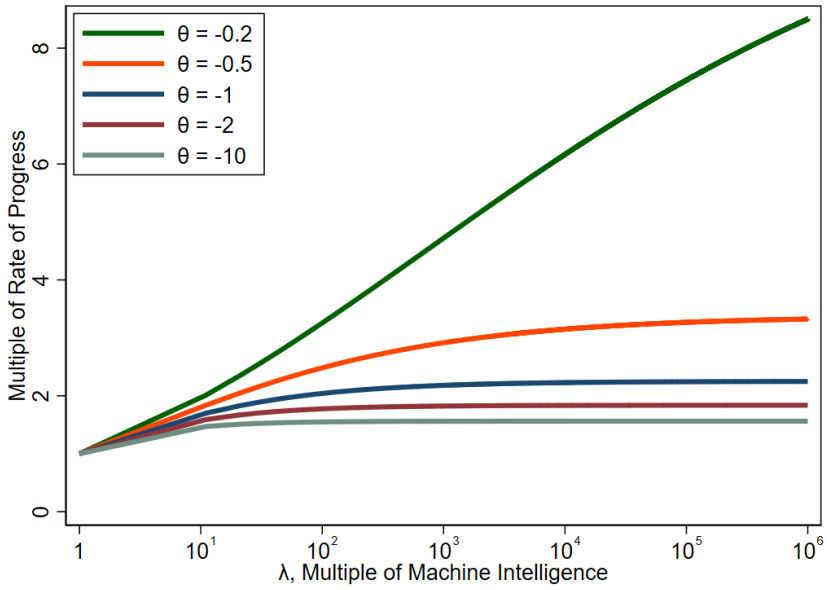

Figure 1: How Progress Accelerates with Large Multiples in Machine Intelligence. The "bottleneck" parameter, θ, governs the returns to intelligence. Even with infinite intelligence, accelerations in progress can be severely constrained.

machines; material inputs like electricity; structures, etc.). Intuitively, the greater share of tasks that AI does not perform, the less effect it will have - and these constraints can be evidently severe.

4.2 The Return to Machine Automation

Our next analysis considers progress in AI of a different form: taking over a greater range of tasks from human researchers. We'll consider (potentially large) increases in the share of R&D tasks that AI can perform and ask what this will do to the rate of progress. More formally, whereas the last section focused on increases in M_t (how good AI is at the research tasks it performs), now we focus on increases in γ_t (what fraction of research tasks AI will do).

Corollary 5 *(Large increase in automation share) Let human researchers initially perform a fraction $1-\gamma_t$ of all R&D tasks. Let AI advance to take over more tasks, so that humans now do a smaller fraction $\rho(1-\gamma_t)$ of tasks, where $\rho \in [0,1]$. To isolate the role of automation, hold machine productivity, M_t, fixed. The rate of progress will initially increase by a*

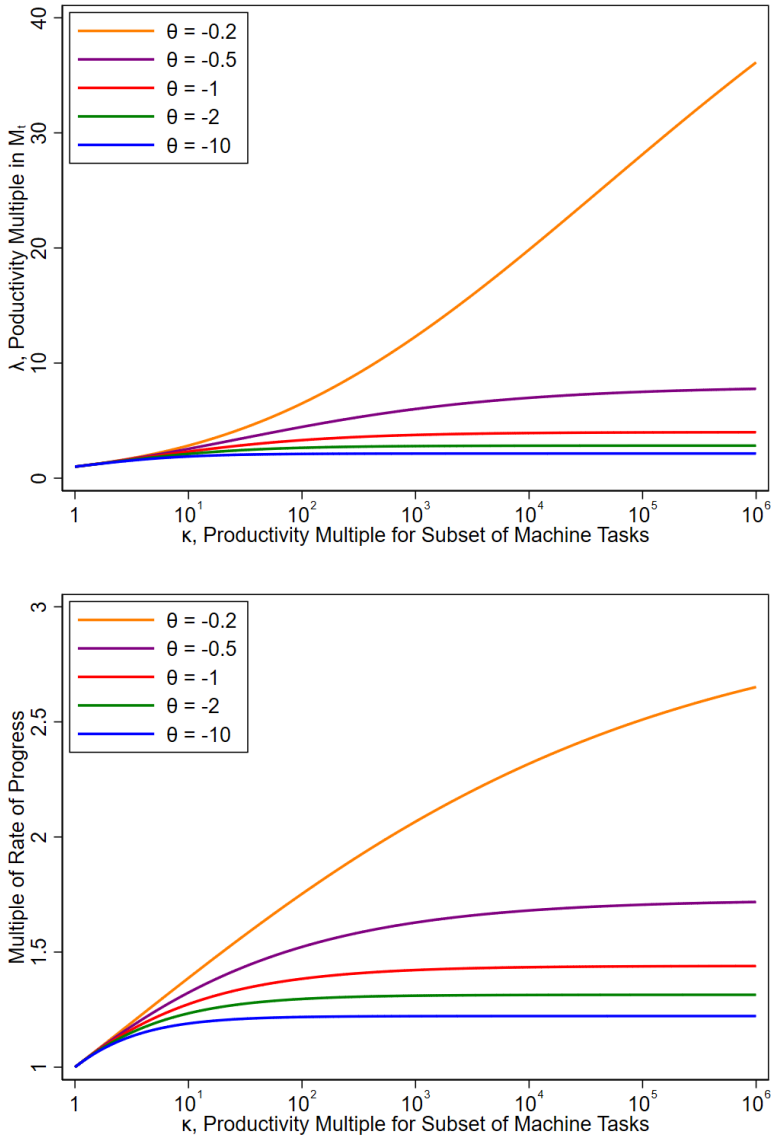

Figure 2: AI as a Subset of Machine Tasks. Here we consider how the machine productivity index, M_t, increases with large advances in AI (top panel) and what this means for the rate of progress (bottom panel). The figure plots cases where 1/2 of all machine tasks become more productive by a multiple κ. Even with $\kappa \to \infty$, increases in overall machine productivity can be severely constrained, resulting in relatively modest accelerations in the rate of progress.

multiple

$$\psi = \left(\rho + \tfrac{1-\rho}{\gamma_t} s_t^X\right)^{-1/b} \tag{14}$$

where $b = \frac{\theta}{\theta-1}$ *and* s_t^X *is the capital share of R&D expenditure as given in (9). This multiple has a lower bound of* 1 *and an upper bound of* $\rho^{-1/b}$*. In the limit where* $\rho \to 0$ *and AI machines take over all R&D tasks, the rate of progress will increase by a multiple*

$$\psi_\infty = \left(\frac{\gamma_t}{s_t^X}\right)^{1/b} \tag{15}$$

Proof. See Online Appendix. ■

As a simple example, take the cases where the initial machine share of R&D expenditure is $s_t^X = 1/3$. Let the initial share of machine tasks be $\gamma_t = 1/2$. Now let automation proceed to an extreme degree, where $\rho \to 0$, so that machines take over all tasks.[7] Then the rate of progress will increase by $\frac{3}{2}^2 = 2.25$. Thus, fully automated R&D tasks would a bit more than double the rate of progress.

To gain further intuition, recall that we must have $s_t^X \leq \gamma_t$ (Corollary 1). The case $s_t^X = \gamma_t$ occurs when machines aren't very good at what they do; more precisely, when M_t is at the lowest possible value that machines are still worth adopting. In this case, machines are expensive and the machine expenditure share rises to its task share. Extending machines with similar (low) productivity to all remaining research tasks and holding factor prices constant, R&D expenditure would then shift fully to machines but there would no gain in the rate of progress ($\psi_\infty = (1)^{1/b} = 1$). This helps us see that the advantage of further automation relies critically on how good machines are at what they do. In practice, machines need to be substantially cost effective compared to labor for there to be an advantage in their use; this is equivalent to saying that the expenditure share on machines is low compared to the share of tasks machines perform.

4.3 The Return to Machine Intelligence and Automation

Finally, we consider the case where an AI advances both in the range of tasks that it can perform (γ_t) and its productivity at those tasks (M_t). Indeed, the above analyses suggest

[7] Note that we are imagining the M_t remains at its initial condition; that is, the average machine advantage for the newly automated tasks is the same as its existing advantage over labor in tasks machines currently perform.

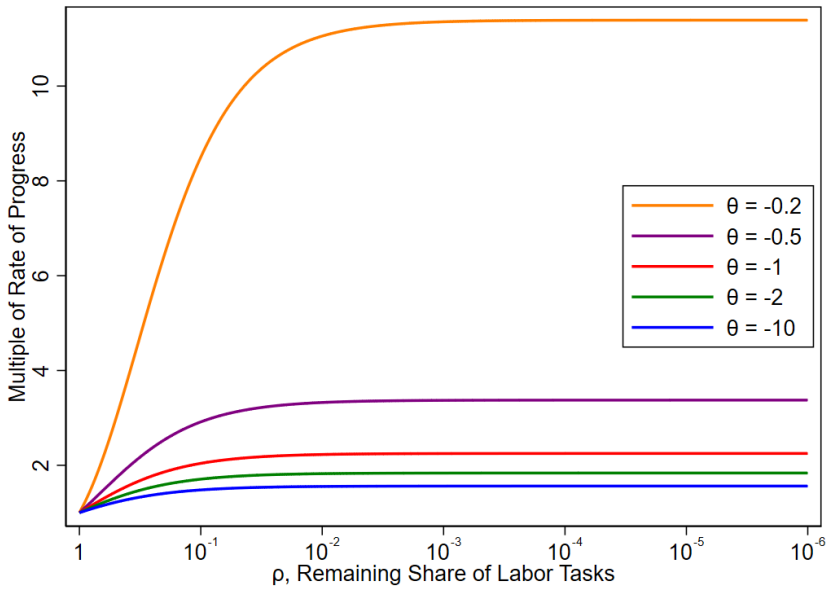

Figure 3: How Progress Accelerates with Large Increases in Machine Automation. Further automation can replace bottleneck labor tasks with more productive machine tasks, but the bottlenecks of remaining labor tasks remain severe. Even a small remaining measure of labor tasks strongly limit gains in the rate of progress, depending on θ.

that large increases in both machine productivity and automation shares are needed for AI to create large accelerations in rates of progress.

Corollary 6 *(Large increase in machine intelligence and automation share) Let machine intelligence increase by a multiple λ. Further, let AI take over many more tasks, so humans now do only a share $\rho(1-\gamma_t)$ of tasks, where $\rho \in [0,1]$. The rate of progress will initially increase by a multiple*

$$\vartheta = \left(\lambda^{-b}\,\frac{1-\rho(1-\gamma_t)}{\gamma_t}\,s_t^X \;+\; \rho\big(1-s_t^X\big)\right)^{-1/b} \tag{16}$$

where s_t^X is the capital share of R&D expenditure as given in (9). In the limit of a super-intelligence where $M_t \to \infty$, the rate of progress increases by a multiple $\vartheta_\infty = \big(\rho\big(1-s_t^X\big)\big)^{-1/b}$ for $\theta < 0$.

Proof. See Online Appendix. ■

Now it is possible, with sufficient advances on both the intelligence and automation margins, for the rate of progress to greatly accelerate. Figure 4 presents examples where the initial expenditure share is $s^X = 1/3$, the initial share of machine tasks is $\gamma_t = 1/2$, and there is substantial task complementarity, taking $\theta = -1$. We consider multiples in the rate of progress as a function of multiples of machine productivity, plotting separate curves for different shares of task automation.

We see that the effect of extreme intelligence depends substantially on the breadth with which it can be applied. For example, in the limit of infinite intelligence, the rate of progress limits to $\vartheta = (\frac{2}{3}\rho)^{-2}$. Thus if, say, 90% of all current human research tasks are automated and conducted by super-intelligence, then rates of progress could maximally increase by 225 times. If only 25% of all current human research tasks can be automated, then rates of progress would maximally increase by only 4 times, per unit of R&D expenditure. These upper bounds occur when AI is infinitely good at its tasks. To the extent that AI conducts only a share of machine tasks (whereas telescopes, microscopes, centrifuges and other experimental machines remain an important component of the machine tasks), such multiples via M_t would be further out of reach.

This is also where it is useful to apply the model in a field-specific way. For example, in areas where cognition represents nearly all tasks (say, in pure math or software design),

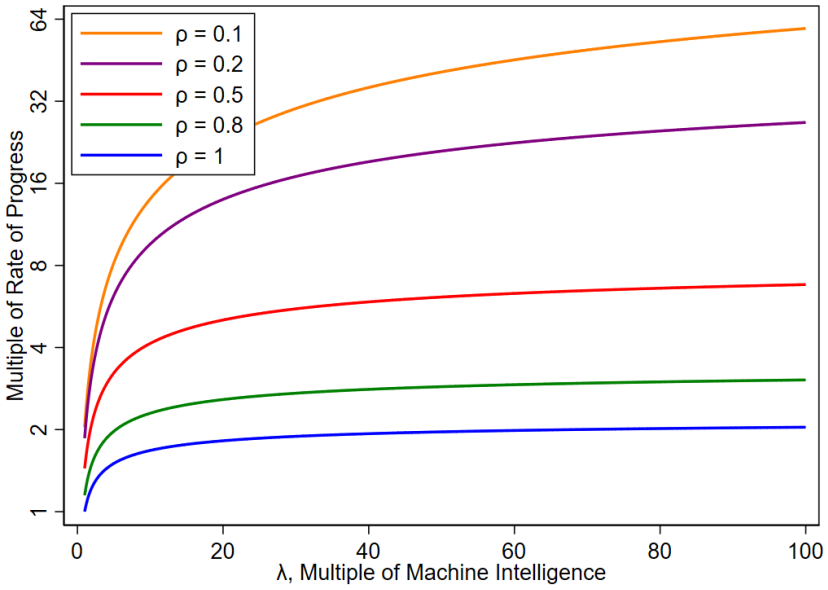

Figure 4: How Progress Accelerates with Large Multiples in Both Machine Intelligence and Automation Shares. In a given research area, large advances in machine intelligence coupled with AI taking over the majority of human research tasks create large increases in rates of progress.

then a super-intelligence can seemingly multiply progress enormously. But in areas where experimentation is a large share of tasks (say, in drug development or particle physics), then rates of progress may accelerate modestly despite super-intelligence. Assessing AI's implications will then be field-specific and assessments will depend on taking field-specific stands on particular technology measures. It seems realistic that the relevant measures can be estimated for specific fields, as discussed in Section 6.

4.4 Longer Run Gains

The analyses above focus on short-run increases in progress. We treat the outcome of interest, Z_t, as fixed at a point in time and ask how a surge in AI capabilities accelerates progress initially. For longer-run progress, the key, further question is how rising Z_t affects ongoing gains, governed by the parameter φ (see (8)). If $\varphi < 0$, then higher Z_t slows the rate $\dot{Z}_t$. Longer-run gains from a burst in AI capabilities may then be smaller than short-run ones, consistent with the "fishing out" of ideas in knowledge creation where the progress gets innately harder the more progress we have made. If $\varphi > 0$, rising Z_t boosts progress,

so long-run gains may exceed initial ones, consistent with progress acting to expand future creative possibilities.

For some outcomes, progress is exponential, i.e., $\dot{Z}_t/Z_t$ is the key object.[8] Examples include Moore's Law, Swanson's Law, and progress in key economic outcomes like GDP per capita. In these cases, long-run AI effects are dampened compared to short-run effects so long as $\varphi < 1$ (see (8)). Evidence suggests $\varphi < 1$ for both macroeconomic growth and Moore's Law (e.g., Bloom et al. 2020). Thus, for such outcomes, a surge in AI capabilities will likely have larger initial than longer-run effects.

5 The Potential for Transformative Artificial Intelligence

The above framework characterizes key forces that govern AI's effects on R&D. For "transformative artificial intelligence" (TAI), the model can further clarify what one must believe for AI not simply to be impressive, but to radically transform outcomes. In this section, we consider definitions of transformative AI and how the model can quantitatively engage them.

5.1 Defining TAI

Transformative AI has various definitions. It orients on a large, step-function increase in rates of economic growth. Karnofsky (2016) defined TAI as: "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution." That is, AI becomes TAI when it creates a break in human history of similar magnitude to these prior revolutions.[9]

The industrial revolution led to an approximately 10-20 times increase in the growth rate compared to pre-1700 history; TAI can thus be viewed as AI that will cause such an acceleration again (see e.g., Davidson 2021; Trammell and Korinek 2023). To provide

[8]The corollaries examine multiples of $\dot{Z}_t$, but since Z_t is fixed, they equivalently describe the initial multiple in the growth rate, $\dot{Z}_t/Z_t$.

[9]Note also a related, second definition from Karnofsky (2016) that is more specific to R&D. He further describes TAI as "capable of fulfilling all the necessary functions of human scientists, unaided by humans, in developing another technology (or set of technologies) that ultimately becomes widely credited with being the most significant driver of a transition comparable to (or more significant than) the agricultural or industrial revolution." An interesting feature of the following analysis is to tie these definitions together quantitatively, where a "step function" in rates of progress becomes tied to high shares of R&D automation.

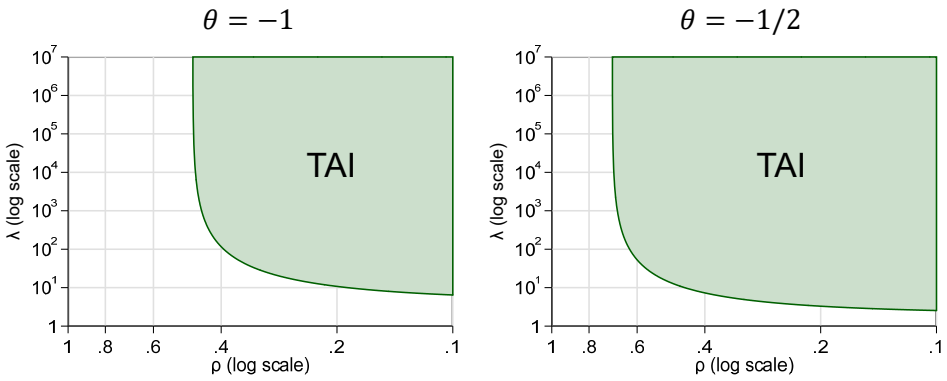

Figure 5: From AI to Transformative AI. Here we consider how much AI must improve to create a 10-fold increase in the rate of progress. The y-axis considers multiples in the machine productivity index. The x-axis considers multiples in automation, measured as the share of current human research tasks that humans continue to do. We assume an initial capital expenditure share of 1/3 and that humans currently perform half of research tasks.

a quantitative target for analysis, and given these views, I will therefore define TAI as achieving a factor of 10 increase in the rate of progress at an outcome variable Z_t.[10] The discussion will further consider the implications of achieving a 10x acceleration in progress for some outcomes but not others, especially when considering economy-wide trajectories.

5.2 Model Parameters and TAI

With this definition (or another per the reader's choosing), we can ask what kinds of capabilities AI must gain to become TAI. Specifically, consider equation (16). Further, set $\vartheta = 10$, representing an order of magnitude increase in the rate of progress. Given the initial conditions, this acceleration can be achieved for various combinations of (1) the share of human research tasks AI takes over (via ρ); (2) the average productivity multiple AI achieves across the tasks it performs (via λ), and (3) the strength of bottlenecks (via θ).

Figure 5 presents two examples. We plot the combinations of the AI technology multiples $\{\lambda, \rho\}$ that would result in TAI for the given outcome Z_t. The left panel assumes $\theta = -1$ and the right panel considers weaker but still significant bottlenecks with $\theta = -1/2$. The

[10] This target was also suggested to me by the volume editors.

figure demarcates the boundary at which TAI is achieved, and shades the area for which even more exceptional progress rates would be achieved.

We can make several observations. First, substantial automation is necessary for TAI. For example, with $\theta = -1$, AI needs to take over approximately 50% of human research tasks before TAI becomes possible, regardless of how good AI is at the tasks it performs. Second, beyond this threshold, further automation greatly reduces the need for super intelligent machines in achieving TAI. For example, with $\theta = -1$, moving from $\rho = 0.5$ to $\rho = 0.4$ reduces the required machine productivity multiple from infinity to 100. Achieving $\rho = 0.2$ (where AI takes over 80% of human research tasks) reduces the machine productivity multiple from 100 to 10. A simple conclusion here is that TAI will critically depend on the share of human research tasks AI can replace. For TAI, we need substantial but not complete automation of human research tasks.

In light of this, consider Dario Amodei (2024), who defines a related concept of "Powerful AI." He describes a "country of geniuses in a datacenter", where there are millions of instances of these synthetic geniuses running in parallel and at 10-100x human speed, and wonders about the implications for progress. Note that while TAI is defined in terms of AI outcomes, this concept of Powerful AI is based on the potential AI input. The model shows how to link these concepts. In particular, with low values of θ, endless millions of cheap geniuses (i.e. extraordinarily high λ) delivers TAI only if we reach certain automation thresholds in research tasks. If, say, $\theta = -1$, and more than half of human research tasks cannot be done *in silico*, then Powerful AI does not produce TAI.

5.3 The Power of Bottlenecks

Clearly, bottlenecks matter, and they can defeat the power of genius. Strong bottlenecks act to direct attention away from super-intelligence and towards the automation share - because automation overcomes the remaining bottlenecks. The intuition comes from how we take averages. In an arithmetic average ($\theta = 1$), all inputs matter the same and there are no bottlenecks. If some share of inputs become highly productive (say via cognitive genius) then the productivity of the whole system is pulled strongly upwards. But for other averaging procedures – those with lower θ – the smaller input values dominate. With the geometric mean ($\theta = 0$) or harmonic mean ($\theta = -1$) we are "averaging down". For example,

if $\theta \leq 0$ and some share of inputs is zero, then the whole system has zero productivity regardless of how much we provide of the other inputs - i.e. we have strong bottlenecks.

Given the power of this force, what can we say in practice about the value of θ? A direct way to estimate θ is from the production function itself. Namely, one observes variation in task inputs and sees what happens to the output, allowing one to identify θ through maximum likelihood or other estimation methods. An example of production function estimation along these lines is Ahmadpoor and Jones (2019). They write down a R&D production function like (1), observe variation in the inputs, and estimate θ. Analyzing separately hundreds of different scientific and technological fields, they conclude that $\theta < 0$ appears universally. For the median field, estimates suggest that $\theta \approx -1$.[11]

In some contexts, we may also have conceptual guidance on θ. For example, Amdahl's Law in computer science governs how the overall system speed relates to the underlying parts. In Amdahl's Law, the system speed combines the component speeds as a weighted harmonic mean (i.e., $\theta = -1$).

More general conceptual arguments may also help place bounds on θ in other fields. One simple test is to ask how sensitive the outcome is to mistakes. In many contexts, failure at any one critical sub-task leads to overall failure (Kremer 1993). In these cases, we have $\theta \leq 0$.[12] A second thought experiment is this: do progress rates explode when some important research task is performed vastly better than before? If the answer is no, then we have low θ, likely towards the harmonic mean.[13]

For example, consider that in some research contexts we have already experienced many-order-of-magnitude improvements over human capabilities for important research tasks. The most obvious example is floating point operations and their widespread use in key research activities. Take economics, where regression analysis is an important activity. It would take my entire lifetime to invert a single, big-data regression coefficient matrix by

[11] That study varies the people inputs to identify θ, as opposed to the variation in machine capabilities, but the same analytic principles apply.

[12] Recall that, with $\theta \leq 0$, zero output on any sub-task means zero overall output.

[13] Imagine two equally weighted inputs that both start with the value 1. Then the average is 1 regardless of the type of average you take. Now let one input value explode by many orders of magnitude. With a geometric mean, the output will rise without bound. With a harmonic mean the value of the output limits to 2.

hand, which my computer does in a minute. But the rate of progress in understanding economics has not improved in such a dramatic way. This strongly points at low θ, where there are other bottlenecks to our understanding. Similarly, Bronwyn Hall's discussion examines how compute radically advanced key tasks in particle physics, but the share of important research tasks performed by humans has remained high.[14] Similar arguments apply to many research tools, including telescopes, microscopes, and spectrometers, which provide virtually infinite improvements over human senses of observation. Viewed within the broader problem of drug design, the Nobel-Prize winning AlphaFold, which applies to protein folding, is another example where radical productivity gains at a narrow set of important tasks do not alone accelerate progress at the wider problem very much. All of this reinforces the point above (and see Figure 5) that the automation share - how widely AI can take over research tasks and thus overcome bottlenecks – is more critical for TAI than massive productivity gains at smaller measures of tasks.

A final observation, especially with regard to TAI, is that bottlenecks are likely nested. The fruit of each R&D process - solar cell efficiency, drug design, a new space vehicle - is a *particular* Z_t. When conceiving of 10x type acceleration in productivity growth overall, we must consider how we combine the Z_t for each area into an economy-wide productivity gain. Baumol's cost disease, and related estimates of production functions in macroeconomics, point to further bottlenecks in the real economy - that the overall productivity gain will be determined more heavily by the areas where Z_t advances slowly than by where it advances quickly (Baumol 1967, Aghion et al. 2019, Jones and Liu 2024).[15] This "double bottleneck" problem, which we could also call a "double averaging down" problem, may further challenge the potential for TAI. It suggests that TAI will depend on overcoming bottlenecks not just in a given R&D process, but in a large share of all research processes, as we will otherwise average down a second time in the real economy. This also suggests that we need to pay

[14]New tools may also call forth new human tasks, which can limit the overall share of research tasks performed by the machines - see Hall's discussion.

[15]Health appears similar. We depend on many different organ systems, and the failure of any system can result in poor health status or death. To greatly extend longevity and maintain a high quality of life, we thus need to successfully overcome a wide variety of diseases – cancers, heart disease and stroke, dementia, diabetes, chronic lung diseases – as well as other challenges like mental health in its many forms, sensory loss, chronic pain, and an array of other health issues.

attention to how AI can affect a wide variety of different research processes. We turn to this empirical agenda in Section 6.

5.4 TAI versus Economically Meaningful AI

The framework points to bottlenecks as a key challenge for AI to overcome. Should bottlenecks remain substantial in key research areas, TAI becomes further from reach – or possibly out of reach. On the other hand, TAI sets a high bar. Accelerations of progress far short of TAI would still have profound effects. To the extent that 10x accelerations in growth rates are difficult, much more modest accelerations - say 1.2x or 1.5x - may be much more feasible and would, in the long-run, still have enormous implications.

Consider that, with 1.8% per annum total factor productivity growth, standards of living were radically transformed since the Industrial Revolution, relying on a wide array of technologies that people in 1870 would struggle even to imagine. If AI accelerated growth to 2.6% per annum, the next 150 years would equate to 3 Industrial Revolutions in terms of multiplying the standard of living - 50 times the standard of living today. History would say that AI made a huge difference. We might therefore define "Economically Meaningful AI" (or EAI) as an alternative scenario, characterizing such relatively modest but nonetheless extremely impactful growth accelerations.

6 Specific Research Applications & AI Benchmarks

AI may have substantially different effects in different research areas – from pure mathematics to material science to drug design. To estimate its effect for a given research context, the model suggests that we focus attention on three key objects: $\{C_t, \gamma_t, \theta\}$.[16] This requires articulating the set of tasks in a given research area, the performance of AI relative to humans at the AI-relevant tasks, and the strength of the bottlenecks parameter in that specific area. Here we discuss how to estimate each of these parameters. This is greatly facilitated by the AI community's focus on "benchmarking."

[16]The above discussion of TAI examined multiples of C_t and γ_t that would drive large accelerations in rates of progress. Here we are focusing on estimating the underlying productivity and task share measures themselves, from which any such multiples will ultimately be determined.

6.1 AI Benchmarking

Benchmarking is an essential tool for artificial intelligence. Different AI models are ranked according to explicit performance benchmarks, and AI models are trained to succeed at these benchmarks. Examples include benchmarks for solving math questions (e.g., MATH Level 5), performing software engineering tasks (e.g., SWE-Lancer), undertaking machine learning research (e.g., PaperBench), and very many others. The following discussion considers how benchmark studies may reveal key model measures.

Consider first the relationship between benchmarks and $c_t(j)$. Recall that $c_t(j)$ captures the relevant cost efficiency of AI compared to humans at a particular task.

$$c_t(j) = \frac{m_t(j)/\mu_t}{H/w_t} \tag{4}$$

For example, let's say the task is to solve a specific type of math problem.[17] One could measure the machine's productivity at the task as correct solutions per unit of time (providing $m_t(j)$) divided by the machine's cost in dollars per unit of time (providing μ_t). Thus $m_t(j)/\mu_t$ is correct solutions by the machine per unit of expenditure. Similarly, we can calculate H/w_t as solutions per dollar spent when using human labor. We can thus calculate $c_t(j)$.

But this is only for one task (taken here as one particular type of math question). What the model directs attention to is C_t, capturing the advantage of AI over a *set of tasks*.

$$C_t = \frac{M_t w_t}{H\mu_t} \tag{5}$$

The difference is in how the task-specific productivities average together into an overall machine productivity, M_t. And the appropriate averaging depends on the outcome measure, Z_t, of interest and its corresponding task set.

To continue the math example, math benchmarks typically report the percentage of questions an AI gets right across a wide set of different types of questions. In that case we are taking an arithmetic average ($\theta = 1$) of correct answers. If the objective is to score the highest grade on an exam, that is a reasonable benchmarking metric. However, in

[17] For example, top AI models can now answer the MATH Level 5 questions, which are very difficult, with high accuracy. See Epoch AI's Benchmarking Hub for comparisons of top AI models on various math benchmarks, including MATH Level 5, GPQA, and FrontierMath.

many R&D contexts – creating software, designing a rocket, engineering a new building that won't collapse – we are often highly sensitive to mistakes. Then we imagine θ is low, and the appropriate benchmark should be calculated accordingly.[18]

For real-world R&D contexts, we therefore need to specify the relevant outcome and its set of tasks and then, from that granular level, move to the key measures $\{C_t, \gamma_t, \theta\}$. AI researchers are now producing benchmarks that allow such analyses. One example is PaperBench (Starace et al. 2025). In this benchmarking approach, the goal is to replicate 20 real-world machine learning papers. The authors encoded a scaffold of 8,316 individually gradable tasks. The study further measures both AI and human success at replicating specific tasks, as well as AI and human time to complete these tasks. By observing what tasks the AI can do reliably and at lower cost (in practice, the top AI performer could replicate 21% of the ML research tasks), we can assign γ_t as the share of replication tasks that AI can take over and determine $c_t(j)$ for each of these tasks. Finally, conditional on θ, we can calculate C_t. One could then compare how using an AI versus not using AI changes the overall rate of progress (here, successful replication analyses) per dollar spent.

Finally, one can estimate θ. As discussed above, the direct way to estimate θ is from the production function – by varying the task inputs and seeing what happens to the outcome (e.g., Jones and Ahmadpoor 2019). With benchmarking studies, researchers could use experimental variation to identify θ for specific contexts. This would start by defining the benchmark (based on the specific R&D outcome of interest, Z_t) and the relevant task set. In addition to measuring AI and human performance on each task, one could experimentally vary the task inputs to estimate θ.

In sum, with suitably designed AI benchmarks, the model can be applied to specific research areas. They key is to first define the research outcome of interest and the set of tasks that are conducted in pursuit of the outcome. Then, measuring AI versus human performance at the individual sub-tasks, calculating the share of tasks where AI has the advantage, and using experimental variation to determine the bottleneck parameter, as described above, can reveal the degree to which AI can accelerate progress. This seems like

[18] In practice, with low θ, we will allocate tasks very carefully to those who can successfully perform them. And we will naturally introduce verification tasks, which some mix of humans and AI may perform. Recent advances in self-verification by an AI may then be critical for allowing AI to be used over a greater share of tasks and to reduce the need for costly human verification.

a potentially useful way to design benchmarks in pursuit of deeper understanding of AI's true potential.

7 Conclusion

This paper presents a framework for assessing AI's role in R&D. The model considers the mapping between AI capabilities and resulting rates of progress. It focuses attention on three key features: the share of research tasks AI performs; the average productivity advantage of AI over humans at the AI-performed tasks; and the strength of bottlenecks.

The framework can be applied to any given research area, where the balance of forces may be context-contingent. Concepts like Transformative AI, which imagines an order of magnitude acceleration in growth rates, or Powerful AI, which imagines millions of genius-level AIs running in parallel, can also be assessed and quantified. The framework shows that bottlenecks severely mute the effect of extremely productive AI. While much remains to be learned, existing evidence and observations suggest that bottlenecks are common in R&D (and in the real economy). This means that taking over a large share of research tasks – which is how AI can overcome bottlenecks – is likely much more important than radical improvements at a narrower set of tasks. Powerful AI is then unlikely to lead to Transformative AI unless these synthetic geniuses can do most research tasks. Put another way, the "marginal returns to intelligence" and even extreme intelligence appears strongly limited when the intelligence operates on only a minority of tasks.

Estimates of the model's measures will clarify what is possible in different research areas. AI benchmarking studies can provide key information. The paper shows the specific measures AI benchmarking studies can engage, pinning down the model and informing the accelerations AI can achieve for various research outcomes. This is an important area for future work.

References

[1] Acemoglu, Daron and Pascual Restrepo (2018). "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 108 (6), 1488-1542.

[2] Acemoglu, Daron and Pascual Restrepo (2020). "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, vol 128(6), 2188-2244.

[3] Aghion Philippe, Benjamin F. Jones, Charles I. Jones. (2019). "Artificial Intelligence and Economic Growth" in The Economics of Artificial Intelligence, Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds.), Universiy of Chicago Press.

[4] Agrawal, Ajay, John McHale, and Alexander Oettl (2024). "Artificial Intelligence, Innovation, and Economic Growth," Working Paper.

[5] Ahmadpoor, Mohammad and Benjamin F Jones (2019). Decoding team and individual impact in science and invention. *Proceedings of the National Academy of Sciences*, 116(28), 13885-13890.

[6] Amodei, Dario (2024). "Machines of Loving Grace: How AI Could Transform the World for the Better," https://www.darioamodei.com/essay/machines-of-loving-grace, Accessed: June 5, 2025.

[7] Autor, David, Frank Levy, and Richard J. Murnane (2003). "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 118(4), 1297-1333.

[8] Baumol, William (1967). "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis," *The American Economic Review*, 57(3), 415-426.

[9] Besiroglu, Tamay, Nicholas Emery-Xu, and Neil Thompson. (2024). "Economic impacts of AI-augmented R&D," *Research Policy*, 53 (7).

[10] Bloom, Nicholas, Jones, Charles I., Van Reenen, John, and Webb, Michael. (2020). "Are Ideas Getting Harder to Find?" *American Economic Review*, 110(4), 1104-1144.

[11] Chen, Q., Yang, M., Qin, L., Liu, J., Yan, Z., Guan, J., ... and Che, W. (2025). "AI4Research: A Survey of Artificial Intelligence for Scientific Research," arXiv preprint arXiv:2507.01903.

[12] Davidson, Tom (2021). "Could Advanced AI Drive Explosive Economic Growth?" Open Philanthropy. https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/, Accessed: July 31, 2025.

[13] Gans, Joshua (2025). "Growth in AI Knowledge," Working paper.

[14] Hill, Ryan, Yian Yin, Carolyn Stein, Xizhao Wang, Dashun Wang, and Benjamin F. Jones (2025). "The Pivot Penalty in Research." *Nature*, 1-8.

[15] Jones, Benjamin F. (2009). "The Burden of Knowledge and the Death of the Renaissance Man: Is innovation Getting Harder?" *Review of Economic Studies*, 76, 283–317.

[16] Jones, Benjamin F. and Xiaojie Liu (2024). "A Framework for Economic Growth with Capital-Embodied Technical Change," *American Economic Review*, 114(5), 1448–87.

[17] Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool et al. (2021) "Highly accurate protein structure prediction with AlphaFold." *Nature*, 596(7873), 583-589.

[18] Karnofsky, Holden (2016). "Some Background on Our Views Regarding Advanced Artificial Intelligence." Open Philanthropy Blog: https://www.openphilanthropy.org/research/some-background-on-our-views-regarding-advanced-artificial-intelligence/, Accessed: July 31, 2025.

[19] King, Anthony, (2025). "Four Ways to Power-up AI for Drug Discovery," *Nature*,doi: https://doi.org/10.1038/d41586-025-00602-5.

[20] Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. (2023). "Scaling deep learning for materials discovery," *Nature*, 624(7990), 80-85.

[21] Segler, M. H., Preuss, M., and Waller, M. P. (2018). "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, 555(7698), 604-610.

[22] Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., ... and Patwardhan, T. (2025). "PaperBench: Evaluating AI's Ability to Replicate AI Research." arXiv preprint arXiv:2504.01848.

[23] Trammell, Philip and Anton Korinek (2023). "Economic Growth under Transformative AI," NBER Working Paper No. 31815.

[24] Zeira, Joseph (1998). "Workers, Machines, and Economic Growth," *The Quarterly Journal of Economics*, Volume 113, Issue 4, November 1998, Pages 1091–1117

[25] Zhou, Hang-Yu, Yaling Li, Jiaying Li, Jing Meng, and Aiping Wu (2025). "Unleashing the Potential of Artificial Intelligence in Infectious Diseases, *National Science Review*, 12(3).