

AI in Science*

Ajay Agrawal
University of Toronto
and NBER

John McHale
University of Galway

Alexander Oettl
Georgia Tech
and NBER

March 11, 2026

Abstract

We explore the impact of artificial intelligence (AI) on the knowledge production function. We characterize AI as a tool, not for full automation but rather for augmentation through enhanced search over combinatorial spaces. This leads to increased scientific productivity. We decompose knowledge production into a multi-stage process to shed light on the “jagged frontier” of AI in science, revealing differential returns to different tools across domains (e.g., data-rich biology vs. anomaly-sparse physics) and workflow stages (e.g., strong design aids like AlphaFold vs. subtler question generation tools). We treat human judgment as indispensable for tasks involving abductive inference, contextual nuance, and trade-offs, particularly in data-sparse environments. Drawing on a task-based model that distinguishes “ordinary” from AI-expert scientists, we describe how exogenous improvements in AI yield nonlinear productivity gains amplified by the share of scientists that are AI-experts to underscore the role of AI complements like skills training and organizational design.

*We thank the book editors, Megan MacGarvie and Reinhilde Veugelers, for the opportunity to contribute this chapter. We thank Pierre Azoulay, Shai Bernstein, Ina Ganguli, Dan Gross, Sabrina Howell, Ebehi Iyoha, Bill Kerr, Rembrandt Koning, Todd Lensman, Caroline Paunov, Bhaven Sampat, Manuel Trajtenberg, Dashun Wang, Mitch Weiss and participants at the 2025 NBER Economics of Science workshop and the Harvard Business School EM unit seminar for useful comments. We gratefully acknowledge financial support from the Center for Innovation and Entrepreneurship at the University of Toronto’s Rotman School of Management. All errors are our own. Contact: ajay@agrawal.ca, john.mchale@universityofgalway.ie, and alex.oettl@scheller.gatech.edu

1 Introduction

How will AI impact economic growth? We are far from a consensus in economics. For example, two of the most prolific economists on the topic of growth, Daron Acemoglu and Chad Jones, have starkly different views. Acemoglu believes it is most likely that AI will have a modest impact on growth, estimating an “increase in TFP and GDP in the next 10 years that can be upper bounded by 0.55% and 0.90%, respectively” (Acemoglu, 2024). In contrast, Jones recently mused “Suppose that advances in AI allow computers to augment and even substitute for humans in innovation, leading to an acceleration of economic growth to some rate g , perhaps 10 percent per year” (Jones, 2024).

To be clear, Jones is not necessarily predicting a growth rate of 10% per year, but offering that value as a thought experiment to explore how much risk from AI we should be reasonably willing to bear given the real possibility of very large growth benefits. In terms of risk, he is not referring to ChatGPT hallucinations, Amazon product recommendation errors, or Waymo car accidents. He is contemplating existential risk: “creating a superintelligence misaligned with human values that could lead to catastrophic outcomes, even possibly human extinction.” In other words, under what conditions should we pause AI development due to some positive probability of wiping out our species if the potential benefits in terms of economic growth are so large?

How do two economists who are both experts on the topic of economic growth arrive at such different views about the impact of AI? They have different views on *how* AI impacts the production function. Acemoglu focuses on how AI enables capital to substitute for labor in certain tasks. Jones focuses on how AI could accelerate the production of new ideas.

Jones writes: “It seems likely that AI will augment our abilities to innovate in the near term, and it is certainly within the realm of possibility that AI could match or even exceed human intelligence at many cognitive tasks and begin innovating itself. Once machines can produce ideas, the limits to growth set by the quantity and quality of researchers may no longer hold, and growth rates could speed up, potentially even leading to a so-called ‘singularity’ with infinite consumption.”

Not only is Acemoglu not as bullish as Jones with respect to the impact of AI on the knowledge production function, he dismisses it altogether: “I do not discuss how AI can have revolutionary

effects by changing the process of science . . . because large-scale advances of this sort do not seem likely within the 10-year time frame, and many current discussions focus on automation and task complementarities.”

Acemoglu’s “not likely within the 10-year time frame” view stands in stark contrast to those held by some of the people closest to the technology. For example, Dario Amodei, founder and CEO of Anthropic, a foundation model company, writes:

I think [powerful AI] could come as early as 2026, though there are also ways it could take much longer....my basic prediction is that AI-enabled biology and medicine will allow us to compress the progress that human biologists would have achieved over the next 50-100 years into 5-10 years. I’ll refer to this as the ‘compressed 21st century’: the idea that after powerful AI is developed, we will in a few years make all the progress in biology and medicine that we would have made in the whole 21st century(Amodei, 2024).¹

Similarly, Sam Altman, co-founder and CEO of OpenAI, another foundation model company, claims:

Anyone in 2035 should be able to marshal the intellectual capacity equivalent to everyone in 2025; everyone should have access to unlimited genius to direct however they can imagine.(Altman, 2025).²

Along the same lines, Demis Hassabis, founder and CEO of Deep Mind, states:

On average it takes. . . 10 years and billions of dollars to design just one drug. We can maybe reduce that from years to maybe months or maybe even weeks.(Hassabis, 2025).³

In other words, the people closest to the frontier, the builders of AI, are more aligned with the Jones view than they are with the Acemoglu view. At the same time, it is reasonable to discount the CEO’s views as they may be biased. Not only may they be genuinely overly optimistic, their fundraising benefits from conveying a bullish timeline for potential investors.

So, it’s too early to tell. Acemoglu *or* Jones could be correct.⁴ On the one hand, some parts of the scientific discovery process have already been significantly impacted by AI (e.g., AlphaFold’s

¹Available at: <https://www.darioamodei.com/essay/machines-of-loving-grace>.

²Available at: <https://blog.samaltman.com/three-observations>

³Available at: <https://www.science.org/content/blog-post/end-disease>

⁴It is noteworthy that Jones’s more recent work develops a “weak links” framework that tempers the short-to-medium-term growth implications of AI. Jones and Tonetti (2026) construct a calibrated model of endogenous

(Jumper et al., 2021) creators were recognized with a share of the 2024 Nobel Prize in Chemistry for their AI’s capability to predict how proteins fold into their final tertiary shapes based on their amino acid sequences), but on the other hand, some stages in the discovery workflow (e.g., experiments, clinical trials) may bottleneck outputs such that improvements in overall productivity may remain muted for a long time.

We develop a model that accommodates both the Jones and Acemoglu views. In line with Jones, we recognize that AI is already altering the economy’s knowledge production function, especially in how it changes the ability to navigate the complex combinatorial search spaces that are a feature of the frontiers of science. At the same time, aligned with Acemoglu, we stress the likely continuing importance of human judgment in the scientific process. Thus, our guiding approach is driven by an appreciation of *both* the tremendous advance in the capabilities of AI and the impressive (and often hard to replicate) capabilities of human scientific judgment. Moreover, also aligned with Acemoglu, we emphasize how, in common with previous general purpose technologies (GPTs), the productive application of AI will require overcoming a range of bottlenecks, which will in turn require significant upstream (notably the development of the AI technology itself) and downstream (e.g., skills training and the reorganization of scientific workflows) complementary investments.

We thus organize our economics-focused review of AI in science around three main ideas. The first is the transformational effects of AI in science that impact productivity and economic growth. Although there are many dimensions to science, we view the central challenge of frontier science as stemming from the difficulties of search over large and complex combinatorial search spaces of questions, ideas, designs and tests. The transformational promise of AI is that it provides technologies – *prediction machines* – that aid in this search (Agrawal et al. (2018)).⁵ We examine the

automation and find that, although growth eventually accelerates, the acceleration is gradual: output is only 4% above the constant-growth path after 20 years and 19% after 40 years. Their key mechanism is that tasks are complements, so output is constrained by the slowest-improving tasks: those still performed by slowly-improving labor. Jones (2026) extends this logic, noting that even infinite automation of all cognitive labor (roughly one-third of GDP) would raise output by only 50% in a weak-links framework. This more circumspect view brings Jones closer to the bottleneck emphasis we develop in this chapter, where productivity at any stage of the scientific workflow can constrain overall scientific output.

⁵The language of predictive and generative AI can be the source of some confusion. It is useful to first distinguish between predictive and generative *tasks*. For example, a predictive task might be predicting the binding efficacy of some molecule with a target protein; a generative task might be to generate a molecule that binds with that protein, which would typically involve sampling from an appropriate conditional distribution – itself a form of prediction. A second useful distinction is between discriminative and generative *models*. Typically, discriminative models are used

effects of AI-aided search on scientific productivity across different stages of the scientific process.

Moving from scientific discovery to economic growth, the potential transformational effects of AI follow from how it might change the economy’s knowledge production function (Romer, 1990; Weitzman, 1998). Although AI may have effects on productivity by changing the economy’s *output* production function, the implications for sustained growth through this channel may be limited (Acemoglu, 2024). However, even relatively modest changes to the *knowledge* production function in science – changes that go far beyond the information technology sector itself – have the potential to have more transformational effects on an economy’s long-term growth rate.⁶

The second idea is the importance of AI in augmenting human scientists. Much of the discussion about the effects of AI has focused on its potential to automate tasks previously done using human intelligence (Acemoglu and Restrepo, 2018; Brynjolfsson, 2022). While the forces leading to automation are real, we argue that human judgment is likely to remain indispensable in at least some stages of the scientific process that involve complex combinatorial search spaces. Put differently, our base assumption is that, although roles of human scientists will change, it is likely that given the ongoing importance of human judgment there will remain a role for humans in the loop in scientific inference and decision making.

We define judgment as the ability to discern and evaluate the practical and normative significance of actual and possible states of the world. Judgment also enters into our conjectures about how observed states of the world have come about or how imagined states might be brought about, and so underpins our constructed model of the world (or “world model”). Judgments, then, can be both the end results of inferential or decision-making processes and, critically in the scientific context, inputs into such processes. Typical elements of effective judgment include sensitivity to

for predictive tasks and generative models for generative tasks. However, predictive tasks could be pursued directly using a discriminative model or indirectly using a generative model and applying Bayes Rule. Similarly, generative tasks could be pursued directly using a generative model or indirectly using a discriminative model (e.g., by ranking molecules for testing based on the outputs of the discriminative model). We use the term *predictive machines* as a shorthand for both types of models as applied to both types of tasks.

⁶While it is trivial to produce explosive growth rates in models of combinatorial search, as a disciplining device we restrict attention in our modeling of scientific productivity to steady-state changes to the exponential growth rate. The effects on knowledge production are also likely to vary considerably between sectors, suggesting that transformational effects will be limited by cost-disease and other bottleneck effects (Nordhaus, 2015; Aghion et al., 2019). Although scenarios of greatly increased growth rates are considered in the literature, given the historical difficulties of achieving even small improvements in long-term growth rates, we would regard more modest changes in the growth rates – say a doubling of the growth rate from 2 to 4 percent – as transformational.

context, a capacity to weigh conflicting (including ethical) objectives, an understanding of the affective implications of states of the world (for both self and others), a capacity to hypothesize causal connections, and an ability to fluidly make analogies. As such, judgment has both the intuitive and reasoned elements thought to be integral to both discovery and justification in science.⁷ This treatment of human judgment shifts attention to the use of AI as a *tool* that augments the capabilities – and thus the productivity – of judgment-exercising human scientists.⁸

The debate about automation versus augmentation has normative and positive dimensions (Agrawal et al., 2019). On the normative side, there is the question of how public policy, including regulation, could shift the balance towards the development of AI technology that augments workers’ capabilities as opposed to fully automating their work (Acemoglu and Johnson, 2023). On the positive side, there is the question of how, for any given path of policy and strategy, the actual course of the development of AI technology will affect the balance between automation and augmentation. While recognizing the importance of the normative dimension, we focus mainly on the positive dimension by examining the effects of AI across different stages of the scientific process and across different scientific domains given current technological trends. Notwithstanding the rapid developments we see in AI, including emerging reasoning and agentic capabilities, we assume a continuing importance of human judgment in science, although of course the extent to which that remains true will depend on the actual course of policy, strategy, and the resultant technological advance.⁹

The third idea is that AI is a general purpose *meta*-technology (GPMT). In addition to having the characteristics of general purpose technologies (GPTs) – including pervasive downstream

⁷Relatedly, we define the intelligence of a system as the information processing and storage capabilities exhibited by the system – encompassing learning, reasoning and adaptation mechanisms – in pursuit of its objectives. Judgment, as we have defined it, is a capability exhibited by intelligent systems to varying degrees, aiding them in the pursuit of their objectives. Artificial general intelligence (AGI) describes a non-human system that can match human-level performance across essentially all tasks. We assume that for true AGI, the system must exhibit behaviors and decision-making processes that are functionally equivalent to human-level judgment.

⁸We do not rule out human-level judgment eventually becoming part of the capability set of AI; but treat it as the part of the capability set of human scientists that is *hardest* to emulate or simulate in AI. With effective judgment across the full range of scientific tasks assumed to remain hard for AI over the relevant horizon (say 10 years), our focus is therefore on AI as a scientific tool that augments the capabilities of human scientists; i.e., on how prediction machines can complement human scientific judgment.

⁹One concern in the normative debate is that the design of benchmarks, which may in part affect policy and strategy decisions, tends to favour automation over augmentation tasks, possibly because benchmarks for the latter are harder to design.

applications, innovation complementarities, and dynamism in the development of the GPT itself – a GPMT is a technology for the development of new technologies (Cockburn et al., 2019; Agrawal et al., 2019). Using the idea of AI as GPMT, we connect our review of AI in science to the broader discussion of metascience that is a central focus of this volume.

We draw on Ethan Mollick’s idea of a *jagged frontier* to capture the current and evolving potential of AI as a GPMT (Dell’Acqua et al., 2023; Mollick, 2024). The scope for using AI in science will differ across stages of the scientific process and across scientific domains. As Mollick argues in relation to the more general application of AI, the jagged nature of the frontier suggests the value of active exploration of how AI can increase productivity across different scientific tasks. Moreover, as with other GPTs, how this frontier evolves will depend on downstream investments in the applications of AI (including investments in the expertise of scientists to use AI) and upstream investments (including investments in the development of the “science of AI” itself). Consistent with our focus on augmentation, we describe the implications of a simple task-based augmentation model (Agrawal et al., 2026b) that we previously adapted from Acemoglu and Restrepo’s task-based model (Acemoglu and Restrepo, 2018, 2019). In our context, the productivity effects of AI on science depend on the stock of AI expertise in the *use* of AI. Looking upstream, we also briefly consider recent developments in the “science of AI,” notably the emphasis on scaling (and compute and energy-related bottlenecks to that scaling) and trends in algorithmic improvement (including, inter alia, the implications of autonomous coding agents for the development of AI research tools).

The rest of the chapter is organized as follows. In the next section, we develop a multi-stage model of the scientific process that stresses both the potential effects on scientific productivity of AI-aided search over combinatorial search spaces and also the continuing role of human judgment in different stages of the scientific process. In Section 3, we examine how the effects of AI on the scientific process are likely to play out differently in different domains of science. In Section 4, we explore the usefulness of thinking about AI as a GPMT and the importance of downstream and upstream complementary investments in altering the jagged frontier. In Section 5, looking downstream, we explore the importance of developing scientists’ expertise in the use of AI in the context of a simple task-based model of augmentation. We turn our attention upstream to the

development of AI technology itself in Section 6, where we consider driving forces and bottlenecks affecting the science of AI. Section 7 explores potential downsides of AI in Science before concluding in Section 8 where we summarize our main arguments and compare and contrast our approach to two other recent characterizations of AI and science: Jones (2025) and Mullainathan and Rambachan (2025).

2 A Multi-Stage Model of AI and Scientific Productivity

In this section, we set out a simple model of productivity in science to structure our discussion. Figure 1 sketches an idealized (linear) scientific process that goes from an initial review of literature and data to a finalized research output. Although for simplicity we conceptualize the process as linear, in reality there may be substantial backtracking. For example, in the process of working on design generation, a scientist may gain insight that leads her to go back and revise or fully change the research question. We observe AI being used to varying degrees in all stages in the real world scientific processes. However, our model focuses on the four intermediate stages in the process indicated by the darkly outlined boxes: question generation, idea generation; design generation; and testing. The first three of these stages can collectively be viewed as hypothesis generation and the fourth as hypothesis testing.

Our working assumption is that there is a subset of scientific tasks for which AI can be an aid to human scientists and in some cases may replace those scientists altogether. As an empirical matter, we take it that these tasks typically involve the use of the interpolative power of learning models trained on large quantities of data. However, we also assume that there are certain tasks that are hard for AI. Typically, these tasks are associated with data-sparse environments and put a premium on judgment. Although the capabilities of AI are rapidly expanding, we assume that there remains a subset of tasks for which human intelligence retains an advantage.

Our model of scientific productivity has three main features:

- The production of science involves four stages: question generation, idea generation, design generation and testing. AI can potentially affect all four stages. Bottlenecks in any of the

stages can lower scientific productivity.¹⁰

- The various stages involve search over potentially vast combinatorial search spaces. We take it that the potential transformative power of AI in science comes from its capacity to revolutionize search in one or more stages in at least some subset of domains.
- Our scientific knowledge production function takes the Romer-style form: $\dot{A} = \omega \cdot A \cdot S$, where inputs are the existing knowledge (or implemented idea) stock, A , the number of scientists, S , and (total factor) productivity in science, ω . Moreover, we decompose ω as itself being the product of four terms: $\omega = \alpha \cdot \beta \cdot \gamma \cdot \delta$, where (loosely) α is a measure of the productivity of the research team in question generation, β is a measure of productivity of the team in idea generation, γ is a measure of productivity of a test in design generation, and δ is a measure of the productivity of scientists in testing.

The model assumes that scientists are organized into scientific teams that undertake the various scientific tasks.

To preview the end result of the model, we use the measure of productivity used by Bloom et al. (2020) and model productivity in science, ω , as the product of three factors:

¹⁰AI might provide a useful tool to support one or more stages of the process. For example, to the extent that LLMs encapsulate a large existing base of knowledge, they might be used to brainstorm for potential new ideas (see, e.g., Mollick (2024)), although there may be limits to the interpolative powers of AI in coming up with truly creative solutions. To date, the greatest potential for AI to support the scientific process seems to be in the design stage, especially where the spaces to be searched are too complex to be comprehended by human minds, but there exist large quantities of actual or simulated data that can be used to build a discriminative or generative model. Arguably, the main proof-of-concept for the potential of AI to support the design stage is AlphaFold, which predicts the shape of proteins from their amino acid sequences. Lastly, AI may also be used to support experimental testing, say by helping to analyze the data from experiments or identifying appropriate treatment and control groups.

$$\begin{aligned}
\text{Scientific Productivity} \left(\omega = \frac{\dot{A}}{S} \right) & \\
&= \text{Productivity in Question Generation } (\alpha) \\
&\times \text{Productivity in Idea Generation } (\beta) \\
&\times \text{Productivity in Experimental Design Generation } (\gamma) \\
&\times \text{Productivity in Testing } (\delta)
\end{aligned}$$

where \dot{A} is the output (new successfully implemented ideas), A is the existing stock of ideas and S is the number of scientists on the team. Drawing on the endogenous growth literature, Bloom et al. (2020) identify constant productivity in research as the productivity that would deliver constant exponential growth with a constant research input (here the number of scientists).¹¹ Looking across a range of domains, their consistent finding is that research productivity (or what we here call scientific productivity) has been falling.

By dividing the scientific workflow into a series of tasks (or stages), our focus in the model is on the ways in which AI might affect scientific productivity. We think of our approach as retaining the simplicity of the Bloom et al. (2020) research productivity (or more generally the form of the Romer knowledge production function). However, by allowing overall productivity to be the product of the productivity of four distinct stages – each of which could be differentially impacted by AI – the framework gives added structure to allow for the different ways in which AI might impact science.

To make things concrete, we use the example of a scientific team that is seeking to find a small molecule drug to bind with a malfunctioning protein for some therapeutic effect. We char-

¹¹By identifying constant scientific productivity with the achievement of constant exponential growth for a given scientific workforce, Bloom et al. (2020) are using a demanding measure of productivity. Recognizing the existing idea/knowledge stock as a non-rival input in the knowledge production function: $\dot{A} = \omega AS$, ω can be taken to be a measure of total factor productivity in knowledge production given the non-rival input, A , and the rival input, S . Arguably, a more natural measure of scientific productivity would be $\frac{\dot{A}}{S} = \omega A$, so that scientific productivity depends on both total factor productivity and the size of the existing knowledge stock. It is possible for total factor productivity (as measured by ω) to be falling (as Bloom et al. (2020) find), but scientific productivity (as measured by $\frac{\dot{A}}{S}$) to be rising. If $\frac{\dot{A}}{S}$ is rising then there is an important sense in which “ideas are getting easier to find” despite the fall in ω . This would be the case, for example, for the semi-conductor industry that they use for one of their case studies. However, we adopt their more demanding (total productivity) measure as our measure of scientific productivity for consistency with the existing literature.

acterize idea generation as combining existing knowledge to identify a space of small molecules that potentially contains a specific small molecule that can effectively bind with the protein along with identifying the criteria for success. Design generation involves generating specific candidate molecules that can be advanced to testing. Testing is the physical testing of the binding efficacy of an identified small molecule (though we can also allow “success” to be multi-dimensional, say, taking into account the safety of the drug). In addition to its potential role in idea and design generation, AI and related technologies may also increase the efficiency of testing, say by identifying experimental subjects, analyzing data, or physically conducting the test with AI-assisted robotic technologies. Each task can involve the use of human capabilities (e.g., human judgment) and AI capabilities (e.g., AI-generated design hypotheses). We next consider each stage in turn with an emphasis on how AI might be used as a tool to enhance productivity and the remaining importance of human judgment.

2.1 Stage 1: Productivity in question generation

The first stage of our idealized scientific process is the generation of a question (or problem, or hypothesis). We assume that the question is generated by the scientific team, possibly with the help of AI. The probability that a question is generated by the team in any given period is α .

Although questions could be generated through numerous mechanisms, we focus on two that have received significant attention in the history and philosophy of science: Baconian induction and Peircean surprise. Under pure Baconian induction, new questions – and indeed entire discoveries – emerge from observations of the world, “unprejudiced” by existing theories (Weinberg, 2015). For our purposes, we think of this mode of question generation as being data driven. While Peircean surprise – named after the American philosopher of pragmatism, Charles Sanders Peirce – is also based on observing the world, the question is generated by a surprising observation (or pattern of observations) that is at odds with some pre-existing understanding of the world (say as captured in existing theories).¹² For our purposes, we interpret “surprise” broadly to include observations of

¹²A contrast between the Baconian and Peircian views is how they see the role of induction in the process. For Bacon, induction is what starts the process; for Pierce, it comes at the end in the testing of a hypothesis, with the initiation role played by surprise followed by an abductive guess (hypothesis) at the possible solution. Peirce explicitly underlines his disagreement with Bacon: “A great many people who may be admirably trained in divinity,

presence and observations of absence. For the latter, the absence, say, of a known small molecule that binds with a malfunctioning protein could generate the question that starts a drug-discovery process.

We consider two innovative attempts by economists to use AI to aid in processes of Baconian induction and Peircean surprise to generate new research questions when faced with vast combinatorial search spaces – and thus increase α . Starting with Baconian induction, Ludwig and Mullainathan (2024) leverage the capability of machine-learning algorithms to see patterns in complex data that would not be perceivable to human researchers. They start from the observation that a defendant’s appearance is highly predictive of judges’ decision on pre-trial detention. But it is unclear which features of the appearance matter.

As a first step in generating hypotheses about which features matter, they devise an algorithmic procedure to morph the mug shots of the accused in the direction of increasing the likelihood of release. The original and morphed mug shots are then shown to human evaluators, who are asked to name the difference between these mug-shot pairs. Two features are found to stand out: how “well-groomed” the accused are and how “full-faced” they are. While the former pattern seems intuitive; the latter pattern is unlikely to be seen by an unaided human looking across a large dataset of actual mug shots.

The procedure yields research questions that could be advanced to the idea generation stage (e.g., developing a theory to explain the observed patterns) or directly to experimental testing of the new hypotheses. A noteworthy feature of their proposed procedure is that it involves the combination of AI (to generate the morphs) and human judgment (to name the difference between the un-morphed and morphed pairs).

Mullainathan and Rambachan (2024) develop a fascinating AI-aided procedure to generate Peircean surprises – or, more specifically, empirical anomalies given an existing theory. Using the example of expected utility theory, they develop procedures to automatically generate empirical

or in the humanities, or in law and equity, but who are certainly not well trained in scientific reasoning, imagine that Induction should follow the same course. My Lord Chancellor Bacon was one of them. On the contrary, the only sound procedure for induction, whose business consists in testing a hypothesis already recommended by the retroductive [i.e., abductive] procedure, is to receive its suggestions from the hypothesis first, to take up the predictions of experience which it conditionally makes, and then try the experiment and see whether it turns out as it was virtually predicted in the hypothesis that it would” (Peirce, 1994).

anomalies given the theory using a “black box” predictive algorithm. The neural network algorithm, which is estimated on actual or simulated data, effectively plays the role of empirical intuition in identifying “surprising” deviations from the theory. Rather than it being the human researcher observing the world and using surprises to generate new research questions, it is the algorithm. The procedure leverages the capability of the algorithm to see patterns in the data that might be invisible to the human researcher; the interpretability problem of the algorithm being a black box is overcome by anchoring the procedure to a known theory.

How might human judgment retain a role in the anomaly-generating procedure? First, the capacity to be surprised depends on a “prepared mind,” which suggests some prior understanding of what to expect in the data. This prior understanding enters in the anomalies model through the coding of the relevant theory – with relevance likely requiring deep knowledge of existing theories. Second, with the algorithm identifying a potential long list of anomalies, judgment may be required to discern a smaller subset that is fruitful to explore. While we reserve the generation of plausible explanations for an observed anomaly to the next stage in the scientific process – idea generation – an initial judgment could be required to separate what is truly interesting from noise in the data; or, to identify anomalies that are potentially interesting as research questions, but are judged unlikely to yield actionable hypotheses through the recombination of existing knowledge.

It is interesting to compare AI as a scientific tool for “seeing” or “observing” things that might otherwise not be readily apparent with another general purpose technology that has been central in the history of science – the microscope. In a fascinating examination of great breakthroughs achieved through the use of microscopes in the natural sciences – from traditional light microscopes to modern electron-based versions – the philosopher of science, Ian Hacking, stresses the importance of judgment in “seeing” with a microscope: “Practice – and I mean in general doing, not looking – creates the ability to distinguish between visible artifacts of the preparation or the instrument, and the real structure that is seen with the microscope” (Hacking, 1983, p.191). Without devaluing the critical instrumental role of microscopes, doing science with the aid of a microscope requires an expert “human in the loop”; where the expertise might involve theoretical understanding of the phenomenon being studied or deep knowledge of the workings of the technology itself. This

importance of the scientist’s judgment, even a scientist augmented by equipment, was well-captured by another philosopher of science, Russell Norwood Hanson, writing in the context of physics: “Seeing that’ threads knowledge into our seeing; it saves us from re-identifying everything that meets our eye; it allows physicists to observe new data as physicists, not as cameras” (Hanson, 1958, p.22). Likewise, using AI to “see” otherwise hidden patterns in the data, or surprising facts given existing understandings, is likely to involve an evolving interplay between the powers of the technology and the powers of human judgment.

Krenn et al. (2022) explore how AI and other computational technologies could help scientists “see” as part of the process of gaining understanding. In terms of AI as an aid to Baconian induction, they imagine the technology serving as a “computational microscope”, noting that “new ways to represent . . . highly complex data will advance our ability to sense structure and recognize underlying patterns” (Krenn et al., 2022, p. 5). Although they see the capability to identify “surprises” in data as less well developed, they also see the potential for AI to generate what we have called Piercian surprises from complex data:

Exceptional data points or unexpected regularities obtained from experiments or simulations can *surprise* human scientists and inspire new ideas and concepts. . . . [T]he *anomalies* could manifest themselves in a more involved combination of variables, which might be very difficult for humans to grasp. Accordingly, applying advanced statistical methods and machine learning algorithms . . . to this type of problem will be an important future research direction. (Krenn et al., 2022, p. 6; emphasis added)

In terms of our drug discovery example, there is obvious potential for AI to see patterns and anomalies in vast data sets of potential targets and treatments that human scientists would find hard to see. Moreover, a key strength of AI comes in its capacity to use a multiplicity of data types, including numeric, language (say existing publications), visual and network data. However, to generate understandable hypotheses that can be advanced to experimental testing, it will be often necessary to generate an idea that can make sense of those patterns or anomalies. We turn then to idea generation as search over a combinatorial space of existing ideas as the next stage in the scientific process.

2.2 Stage 2: Productivity in idea generation

In some cases, the generation of a research question will automatically lead to the generation of a hypothesis that can be advanced to the design stage or even directly to testing. However, we consider cases where the question must first be intermediated by an “idea” (say in the form of a theory) to advance the hypothesis-generation process. We conceptualize this idea generation process as search over a potentially vast combinatorial search space of existing ideas, A . The feedback from the existing stock of implemented ideas to the generation of new ideas – or “standing on the shoulders of giants effect” – is the source of knowledge growth, \dot{A} , in our model. Again, we consider the potential roles of both AI and human judgment in searching the space of potential combinations of existing ideas.

Although a pure (sometimes called naïve) Baconian induction view would eschew the development of explanatory theories, later empiricist philosophers of science emphasized the importance of ideas in the generation of hypotheses.¹³ An influential example is the English philosopher William Whewell, who stressed the centrality of conceptual thinking in “binding together the facts” as part of an inductive reasoning process:

In order, then, to discover scientific truths, suppositions consisting either of new Conceptions, or of new Combinations of old ones, are to made, till we find one supposition which succeeds in binding together the Facts. . . . It answers its genuine purpose, the Colligation of Facts (Whewell, 1989, p. 137).

For Peirce, following the surprise – or observation of an anomaly – the search for a conceptual explanation involves a process of *abductive* reasoning:¹⁴

¹³Although Francis Bacon is often associated with a pure data-driven scientific method, as Ian Hacking points out, his views were more subtle. This is shown in Bacon’s story of the ant, the spider and the bee. “The men of experiment are like the ant; they only collect and use; the reasoners resemble spiders, who make cobwebs out of their own substance. But the bee takes a middle course; it gathers material from the flowers of the garden and the field; but transforms and digests it by a power of its own. Not unlike this is the true business of philosophy for it neither relies solely or chiefly on the powers of the mind, nor does it take the matter with which it gathers from natural history and mechanical experiments and lay it up in the memory whole, and it finds it; but lays it up in the understanding altered and digested” (quoted in Hacking (1983, p. 247)).

¹⁴“Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value and deduction merely evolves the necessary consequences of a pure hypothesis. Deduction proves that something *must* be, Induction shows that something *actually is* operative, Abduction merely suggests that something *may be*” (Peirce, 1998, p. 216; emphasis added).

The surprising fact, C, is observed;
But if A were true, C would be a matter of course,
Hence there is reason to suppose that A is true (Peirce, 1998, p. 216).

Interestingly, both Whewell and Peirce think of idea generation to make sense of the facts as involving a directed search (or informed *guessing*¹⁵) over some conceptual space that puts a premium on human judgment:

The Conceptions by which Facts are bound together, are suggested by the sagacity of discoverers. This sagacity cannot be taught. It commonly succeeds by *guessing*; and this success seems to consist of framing several hypotheses and selecting the right one. But a supply of appropriate hypotheses cannot be constructed by rule, nor without inventive talent (Whewell, 1989, pp 129-30; emphasis added).

Now, that the matter of no new truth can come from induction or from deduction, we have seen. It can only come from abduction; and abduction is, after all, nothing but *guessing*. We are therefore bound to hope that, although the possible explanations of our facts may be strictly innumerable, yet our mind will be able, in some finite number of guesses, to guess the sole true explanation of them (Peirce, 1994, CP 7.219; emphasis added).¹⁶

Building on these ideas, we next provide a simple model of the idea generation stage as a combinatorial search problem over the space of existing ideas. We make strong assumptions to ensure that the idea-generation equation is linear in the existing stock of ideas. This allows for a tractable model that is consistent with the Romer knowledge production function.

The starting point is the stock of A existing ideas. The scientific team can combine these ideas one-at-a-time, two-at-a-time, up to A -at-a-time. The total number of combinations is thus given

¹⁵Although the term *guessing* may seem to underplay the insight required to generate new ideas, it is noteworthy that this word is central to Richard Feynman's celebrated description of the scientific process:

"In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guess is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science" (Feynman, 1992, p. 156).

¹⁶It is noteworthy that Peirce saw such *guessing* as bound up with the capacity for instinctive judgments: "In examining the reasonings of those physicists who gave to modern science the initial propulsion which has insured its healthful life ever since, we are struck with the great, though not absolutely decisive, weight they allowed to instinctive judgments. Galileo appeals to *il lume naturale* ["the natural light"] at the most critical stages of his reasoning. Kepler, Gilbert, and Harvey – not to speak of Copernicus – substantially rely upon an inward power, not sufficient to reach the truth by itself, but yet supplying an essential factor to the influences carrying their minds to the truth" (Peirce, 1994, CP 1.80; emphasis in original).

by:

$$I = 2^A - 1, \tag{1}$$

which will be a vast number for even relatively modest values of A . We assume, however, there is a limit on the actual number of combinations that can be searched, with the size of the effective search space, J , conveniently given as a constant-elasticity function of the total number of possible combinations:

$$J = \frac{(2^A)^\phi - 1}{\phi}, \tag{2}$$

where ϕ is (approximately) the elasticity of the effective search space with respect to the total size of the search space. As $\phi \rightarrow 0$ (assumed to reflect the challenge of searching a vast combinatorial search space), we can use L'Hopital's rule to show that the effective size of the search space becomes a simple linear function of the existing stock of ideas:

$$J = (\ln 2)A. \tag{3}$$

We next assume that the probability that any given combination in J yields an idea that can be advanced to the design generation stage is equal to $\tilde{\beta}$, where $\tilde{\beta}$ is naturally viewed as a measure of the *quality* of the identified effective search space. Although abduction is often formalized as inference to the *best* alternative, for simplicity we conceive it here as inference to a *sufficiently good* alternative. We can thus think of the team as facing a constant success hazard rate equal to $\tilde{\beta}$ on draws from the effective search space, J , and the team will stop the search when it finds an idea that is good enough to advance to the design stage. Given that the team can search a maximum of J combinations, the probability, p , of finding an idea that can be advanced is then the cumulative probability of achieving a success given J tries:

$$p = 1 - e^{-\tilde{\beta}J} = 1 - e^{-\beta A} \tag{4}$$

where $\beta = (\ln 2)\tilde{\beta}$. Rearranging (4) and taking logs of both sides:

$$\ln(1 - p) = -\beta A. \tag{5}$$

Finally, assuming that p is small, we can approximate (5) by the simple linear function:

$$p \approx \beta A \tag{6}$$

where we henceforth ignore the approximation. Consistent with the Romer knowledge production function, the probability of generating an idea through the combinatorial search process is then a linear function of the existing knowledge stock.

Judgment could potentially matter in two ways in this search process. First, judgment could be required to identify the effective search space, J , that will receive cognitive attention.¹⁷ And, second judgment could be required to classify whether a given searched combination meets the criteria for sufficiency, which could include such criteria as simplicity, coherence with existing theories, interpretability, etc. In terms of Daniel Kahneman’s (Kahneman, 2011) distinction between Type 1 (intuitive) and Type 2 (deliberative) thinking, finding a new idea is naturally thought of as the result of an interplay between the two types of thinking. Type 1 is likely to be most important in discerning the promising subset from the vast set of possibilities; Type 2 is likely to be most important in evaluating the targeted promising combinations.¹⁸ Interestingly, Pierce viewed abductive inferences as being within the class of perceptual judgements, suggesting an important role for what we now call System 1: “The recognition that two objects present belong together as one is a judgment.

¹⁷The need to narrow the search space is emphasized by Jacques Hadamard: “It is obvious . . . that the building up of numerous combinations . . . is only the beginning of creation, even, as we should say, preliminary to it. [A]s Poincaré observes, to create consists precisely in not making useless combinations and in examining only those which are useful and which are only a small minority. Invention is discernment, choice.” (Hadamard, 1945, p. 30).

¹⁸It is unfortunate that System 1 thinking has become mainly associated with biases and errors in human thinking. In addition to highlighting how System 1 can lead us astray, Kahneman also stressed the “marvels” of the powers of intuitive judgment exhibited by System 1, especially its role in creative thought. In the closing chapter of *Thinking, Fast and Slow* he underlines this other side of System 1:

“I have spent more time describing System 1, and have devoted many pages to the errors of intuitive judgment and choice that I attribute to it. However, the relative number of pages is a poor indicator of the balance between the marvels and flaws of intuitive thinking. System 1 is indeed the origin of much that we do wrong, but it is also the origin of much of what we do right – which is most of what we do. Our thoughts and actions are routinely guided by System 1 and are generally on the mark. One of the marvels is the rich and detailed model of our world that is maintained in associative memory; it distinguishes normal from surprising events in a fraction of a second, immediately generates an idea of what was expected instead of a surprise, and automatically searches for some causal interpretation of surprises and events as they take place.” (Kahneman, 2011, pp 415-416).

All ideas arise in judgments” (Peirce, 1994). Human intelligence appears to excel at dealing with situations that require such imaginative discernments and evaluations, while in comparison AI may struggle to deal with subtle distinctions that appear relevant to human intelligence. If the essential use of judgment in the question generation stage is, to use Hanson’s phrase, “seeing that,” the essential use of judgment in the idea generation stage is in effectively “imagining that.”

Philosophers and cognitive scientists have debated whether the search over an idea space where intuitive judgments plays a central part can be viewed as a (boundedly) rational process.¹⁹ The AI pioneer Herbert Simon took the view that a process that employs the appropriate “heuristics” in the search for new ideas is a rational process:

Scientific discoveries seldom, if ever, emerge from random, trial-and-error search; the spaces to be searched are far too large for that. Rationality for the scientist consists in using the best means he has available – the best heuristics – for narrowing the search down to manageable proportions (sometimes at the cost of ruling out good solution candidates). If the scientist’s tools are weak (perhaps because of the novelty of the problem), a great deal of residual search may still be required; but we must regard such a search process as rational if it employs all the heuristics that are known to be applicable to the domain. This is the concept of rationality that is relevant to the creative process and to problem solving in general, and it is with this kind of rationality that a normative theory of creativity and discovery is concerned. (Langley et al., 1987, p. 47)

There still remains the question of whether AI could come to emulate or simulate the full range of the powers of human judgment. Michael Polanyi’s idea of tacit knowledge provides a useful lens with which to view this question. Extending Gestalt psychology, Polanyi argues that the “structure of [tacit knowing] shows that all thought contains components of which we are subsidiarily aware in the focal content of our thinking” (Polanyi, 2009, p. xviii). An implication is that *we can know more than we can tell*. This power of tacit human judgment may be what allows us to combine existing knowledge into new focal combinations, even if we are only dimly aware of how ideas are being combined and selected. The importance of tacit knowing is one explanation for the disappointing

¹⁹In the philosophy of science, a distinction is made between the “context of discovery” and the “context of justification,” with many writers (including Karl Popper) seeing only the latter as the proper normative focus of philosophers. (See Hanson (1958) and Langley et al. (1987) for early cogent dissents from this position.) For the *economics* of AI in science, however, it is essential to consider how AI might impact all stages of the scientific process, including the discovery of new ideas that underpin the hypotheses that are subsequently advanced to the testing stage.

performance of GOFAI (contrary to Simon’s hopes), which relied on being able to program human knowledge in the computer. But it is less obvious that it represents an insurmountable barrier to learning-based AI, where the tacit components of knowledge might be captured in the distributed representations of, say, a deep-learning model. Thus, in principle, it would seem that both human and artificial intelligence could produce the intuitive judgments necessary to navigate a complex combinatorial search space of existing knowledge, though neither might be able to fully articulate how it is coming to its judgments. We therefore do not rule out the possibility of AI developing to match the combinatorial powers of the human mind. However, as richly revealed in Polanyi’s account of the human powers of knowing, the human mind is a remarkable evolutionary adaptation, setting a high bar for the capacity to see the significance of potential combinations. It is noteworthy that Polanyi also emphasized the power of tools to extend the reach of the human mind. Although AI is distinctive compared to other tools in its powers to autonomously navigate complex search spaces, we think it is best, at least for now, to view it as a tool to augment the powers of human judgment rather than as a full replacement for that judgment.

In the specific context of science, Polanyi emphasized another feature of human minds that might be hard to replicate in AI – the human *drive* to seek knowledge that in part explains “the capacity for anticipating the approach of hidden truth” (Polanyi, 2009); a capacity that he sees as reflecting the personal commitment of the scientist in seeking knowledge:

[T]he act of discovery appears personal and indeterminate. It starts with the solitary intimations of a problem, of bits and pieces here and there which seem to offer clues to something hidden. They look like fragments of a yet unknown coherent whole. This tentative vision must turn into a personal obsession; for a problem that does not worry us is no problem: there is no drive in it, it does not exist. This obsession, which spurs and guides us, is about something that no one can tell; it is undefinable, indeterminate, strictly personal. (Polanyi, 2009, pp 75-76)

Polanyi is convinced that this drive, and relatedly the stakes involved in *guessing well* given the personal commitment of the scientist, underpins effective hypothesis generation construed as search over some space of ideas:

[T]here is a widespread opinion that scientists hit on discoveries merely by trying everything as it happens to cross their minds. This opinion follows from an inability to

recognize man’s capacity for anticipating the approach of hidden truth. The scientist’s surmises or hunches are the spurs and pointers of his search. They involve high stakes, as hazardous as their prospects are fascinating. The time and money, the prestige and self-confidence gambled away in disappointing *guesses* will soon exhaust the scientist’s courage and standing. His gropings are weighty decisions. (Polanyi, 2009, p. 76; emphasis added)

Putting the possible importance of this drive (or intrinsic motivation) aside, we acknowledge the potential for AI in the particular form of LLMs to combine existing ideas given that these models are trained on human ideas expressed in language. Indeed, anyone who has experimented with the latest generation LLMs would be hard-pressed not to see impressive powers of combination. Although the tendency of LLMs to “hallucinate” is a problem when it comes to answer factual questions, such imaginative powers could become an advantage in generating new hypotheses (or guesses) that are then subject to testing. As LLMs are trained on a large fraction of recorded human thought, their weights provide users with access to an almost unfathomable store of knowledge. But beyond simply storing knowledge like an encyclopedia, LLMs display a capability to combine ideas in response to complex prompts, thereby giving their users the ability to explore new idea combinations and making them at least a valuable brainstorming tool as a complement to human judgment.

Whether (or when) the interpolative powers of LLMs can fully match the imaginative or creative powers of human intelligence is actively debated by AI scholars (see, e.g., Chollet (2019)).²⁰ These

²⁰The type of challenges that we see facing AI are captured well in the ARC challenge (see Chollet (2019)). The challenge involves puzzles on a two-dimensional grid where participants are given a limited number of input-output pairs where there is a rule for transforming the input into the output. Human participants generally find it easy to solve the puzzles after seeing just two or three instances. Using their powers of judgement, they can quickly form hypotheses about the transformation rule by combining known ideas – including the use of analogies to related problems – and then test the hypotheses by simulating the hypothesized rule. We see this as being a quintessential example of abductive reasoning. However, at least until recently, AI has performed relatively poorly on the challenge when it faces puzzles that are not in its training data. In early 2024, the ARC-AGI benchmark stood as a “reasoning wall” for AI, with top models like GPT-4o struggling to exceed a 9% accuracy rate, but the following two years saw a vertical shift in momentum as the industry pivoted from scaling data to scaling test-time compute. By the end of 2024, specialized systems like OpenAI’s o3 utilized massive search and program synthesis to overcome previous records on the original benchmark, reaching an 87.5% score that nearly matched human performance; however, this success prompted the 2025 launch of ARC-AGI-2, which initially reset AI scores to the low single digits. Through 2025 and into January 2026, the introduction of “refinement loops” and advanced reasoning models like GPT-5.2 Pro and specialized Gemini 3 configurations drove a second surge, with top systems achieving a 54.2% success rate at the time of our final edit of this chapter, rapidly closing the gap with the 60% individual human baseline and setting the stage for the interactive, agent-based challenges of the soon-to-be-released ARC-AGI-3. Although we have little doubt that human-AI gap will have narrowed further by the time this chapter is published, we believe that a gap will persist, at least when viewed over the wide range of domains that human judgment excels.

debates are partly related to a long-standing debate in cognitive psychology about the relationship between language and thought (see, for example, Fodor (1975)). While we do not take a side in this debate, it seems apt to point to possibly the most famous testimony on the relationship between language and thought in the history of science – Albert’s Einstein’s 1945 testimonial to Jacques Hadamard:

The words or language, as they are written or spoken, do not seem to play any role in my mechanism of thought. The psychical entities which seem to serve as elements in thought are certain signs and more or less clear images which can be “voluntarily” reproduced and combined.

There is, of course, a certain connection between those elements and relevant logical concepts. It is also clear that the desire to arrive finally at logically connected concepts is the emotional basis of this rather vague play with the above mentioned elements. But taken from a psychological viewpoint, this combinatory play seem to be the essential feature in productive thought – before there is any connection with logical construction in words other kinds of signs which can be communicated to others. (Hadamard, 1945, p. 142)

While, of course, the methods of combining ideas inherent in LLMs need not parallel those used by human minds, taking Einstein’s introspection at face value suggests that at least some of the most creative human thought involves powers of judgment that transcend the combination of ideas as expressed in language. Whatever the generalizability of this famous introspection, our working assumption is that human judgment will for now remain – notwithstanding the rapid advancement of AI as a tool – indispensable in creative search across the combinatorial space of ideas. However, we next turn to a part of the process where there is already demonstrated potential for AI to aid the scientific process in data-rich environments – hypothesis generation involving search over complex combinatorial spaces of *designs*.

2.3 Stage 3: Productivity in design generation

In Paul Romer’s classic endogenous growth model (Romer, 1990), new knowledge production is modelled as being a function of the existing stock of ideas and the number of researchers. In subsequent writings, he has stressed the combinatorial nature of the knowledge production process (see, e.g., Romer (1992)). Interestingly, however, a central motivating example is the combinatorial

explosion in the number of new compounds that could be generated from the elements using a child’s chemistry set.

Another child’s toy is a chemistry set. For this discussion, the set can be represented by N jars, each containing a different chemical element. From the child’s point of view, the excitement of this toy comes from trying to find some combination of the underlying chemicals that when mixed together and heated, does something more impressive than change colors (explode, for example). In a set with N jars, there are $2^N - 1$ different mixtures of K elements, where K varies between 1 and N . (Romer, 1992, p. 68)

We think it is more useful to think of such combining as taking place in the combinatorial space of designs rather than the combinatorial space of ideas. In Romer’s example, this design space is the space of potential molecular designs for a new compound. For many purposes, the distinction between idea generation and design generation will be of limited importance. However, in understanding the impact of AI on science, we think that many of the early impacts of AI on science have followed from the powerful capabilities of AI in directing search over a design space (e.g., the space of small molecule drugs that bind to a malfunctioning protein). In contrast to the idea generation problem, the design generation challenge will often take place in a data-rich environment where the data is comprised of successful prior instantiations of real world combinations on related problems, making it amenable to an AI-based solution.²¹

In what we call the Hassabis hypothesis (see Agrawal et al. (2024)), Demis Hassabis, the head of Google DeepMind, identified three requirements that make a scientific (design) problem amenable to an AI-aided solution: (i) a combinatorial search space (too large for exhaustive search); (ii) a clear objective function for training the AI model; and (iii) sufficient data (or capability to simulate that data) to train the model. The Hassabis hypothesis is then that the set of amenable problems that cannot be solved by other means is large.

Given the apparent potential of AI in solving such design problems, we next provide a simple model of the productivity of the AI-aided team in generating design hypotheses, where we assume

²¹It is useful to think of both human minds and AI as providing generative models of possible worlds. In our discussion of idea generation, we have underlined the generative powers of human minds – notably in the process of abductive inference – in searching the space of idea combinations. However, such generative powers may be quite limited when faced with the combinatorial space of potential designs. In contrast, AI-based prediction machines may be highly effective in navigating spaces – such as the space of molecular compounds – that are alien to everyday human experience.

that the team has already successfully generated an idea that identifies the appropriate design space. Continuing with our drug discovery example, we can think of the scientist as searching for a small molecule drug that binds effectively to some target protein. Just as with the *idea* for a new drug, the instantiated *design* of the drug can again be thought of as a combination; this time as a combination (or compound) of smaller molecular components.

The resulting combinatorial search space, X , is assumed to be extremely large, with a total number of possible combinations equal to N_X . We denote a particular element of X as x . The result of a test of a candidate combination is y , where $y \in \{0, 1\}$, where $y = 1$ if the test is a success and $y = 0$ if the test is a failure. For simplicity, we treat the outcome of the test as deterministic. Crucially, in selecting combinations for testing, we assume the team has access to a generative AI model that gives the probability of a given x given that the test is a success, $q(x|y = 1)$.

To simplify the analysis, we take it that it is known that there is a single (initially unidentified) small molecule that will effectively bind with the protein. Therefore, the team faces a “needle-in-a-haystack” type problem: it is known there is one successful design but in the absence of the generative model the team has no idea which of N_X combinations would generate that successful design. The probability that a purely randomly chosen test would yield a success is then $q(y = 1) = \frac{1}{N_X}$. Moreover, the probability that a given molecule would be chosen for testing in an undirected trial-and-error search is $q(x) = \frac{1}{N_X}$. From Bayes rule,

$$q(y = 1|x) = \frac{q(x|y = 1)q(y)}{q(x)} = \frac{q(x|y = 1)\frac{1}{N_X}}{\frac{1}{N_X}} = q(x|y = 1). \tag{7}$$

Thus, for a given test “generated” by the AI model (i.e., randomly drawn according to the probabilities $q(x|y = 1)$), the probability of success is simply equal to the probability of drawing that molecule to test using the generative model.

A test is assumed to take one unit of time. Given the expected value of a success resulting from a draw from this distribution, the value of the successful discovery of a design, and the cost of conducting a test, we assume that (conditional on a success not being found on previous draws) the marginal value of an additional test exceeds the marginal cost of the test. Therefore, the team will continue to make draws from the distribution until a success is found. We denote the expected

probability of success given a random draw from this distribution as $q^e(y = 1|x) = \gamma$. A higher value of γ is then taken to be a measure of a more effective generative model. In terms of the probability of achieving a success on the t^{th} test, we can view the team “as if” they are facing an exponential distribution with probability of success on any given test equal to γ . (We assume that draws from the distribution are without replacement, but we assume that the expected probability of success remains constant at γ and that the expected duration of the search is small relative to the total size of the search space.) Given the exponential distribution for the timing of a success, the expected duration of search is $\frac{1}{\gamma}$, and the expected productivity of a given test (measured in terms of the expected probability that the test yields a success) is equal to γ .²²

An improvement in the generative (or predictive) model used to guide search in the design space will lead to an increase in γ and thus in design-stage productivity. Although in this simple model of the design stage we have left out any role for human judgment, in reality scientists are unlikely to take the outputs of the model completely at face value. Some generated outputs might be, for example, problematic (or outright impossible) based on the physical understandings of scientists. Alternatively, given that there may be multiple objectives that the design needs to satisfy – say the efficacy and the safety of a drug – making choices given the tradeoffs could require the judgments of scientists. In terms of our model, rather than simply taking the draws from the generative model at face value, the outputs of the model might be combined with such judgements to create a prioritized list of designs for testing.²³ In other words, the model might provide a “Pareto frontier” given the competing objectives, but human judgement is still required to rank the points on that frontier. While AI seems to offer particular promise in terms of transforming the design stage, it is again best thought of as a tool to be used in concert with human judgment.

²²Instead of assuming a generative model, Agrawal et al. (2024) assume the team has access to a predictive model that gives the probability of success for each combination in the search space. Ordering the combinations by their predicted probabilities of success yields a “hazard function” for the sequential search. We then use survival analysis to determine the maximum duration of search given the cost of conducting a test and the value of a success if found. Given this maximum duration, we determine the expected value of the design search and the expected duration of the search. Improvements in the prediction model are defined in terms of improvements in the cumulative hazard function. The assumption here that the team uses a generative model significantly simplifies the analysis, and improvements in the “design model” are captured simply by γ . In addition to its simplicity, the generative model approach also appears to be consistent with how many scientific teams are using AI to support search over vast and complex design search spaces.

²³See Agrawal et al. (2024) for a model of optimal testing given a prioritized list of potential designs.

2.4 Stage 4: Productivity in testing

Finally, we assume that a scientist can perform δ tests per unit of time. Thus the productivity of a scientist in testing is δ .

While it is possible that AI can directly improve the productivity of scientists in testing – say by improving their capability to analyze experimental data (more on which below) – the indirect effect of AI on the equipment that scientists use to conduct the tests may be of most importance. The obvious example is the use of robotic technologies in testing, which in the extreme case could involve an almost autonomous testing process, with the role of human scientists limited to a supervisory role. To the extent that advances in robotic technologies depend on advances in software as much (or more) than they depend on advances in hardware, advances in AI could be associated with improvements in robotic testing technologies and thus with improvements in the physical productivity of scientists in testing.

In treating the testing of a combination – say whether a particular small molecule effectively binds to a malfunctioning protein – as determinative, we have as a simplification bypassed the role of scientific judgment in inductive inference. In reality inductive inference as practiced by scientists is infused with judgment. In Bayesian inference, for example, although there is a clear rule for how a given hypothesis should be updated given the evidence (Bayes Rule), the initial prior will often reflect judgment about the probability of (or probability distribution for) a hypothesis in the absence of the evidence. Although things may seem more mechanistic under a classic frequentist approach to inference, say the testing of a null hypothesis of a zero effect based on some conventional significance level, in reality scientists make judgments of practical as well as statistical significance. Judgement becomes even more central when using observational data to infer causal effects. For example, in using a natural experiment to test a hypothesis, scientists must make judgments about the credibility of assumptions about the exogeneity of the observed variation in the independent variable. While the flexibility of AI algorithms certainly increases the range of tools available to scientists for conducting inductive inference, especially when dealing with non-conventional data, it is unlikely that its use means that human judgment in the inference process can be dispensed with altogether. Indeed, given the often black-box nature of highly flexible machine-learning-based

function approximators, the judgment of scientists might become even more important.

2.5 Putting the pieces together: The determination of scientific productivity

We assume there are S members of the scientific team. As previously noted, we assume for simplicity that S is set so that the team expects (for given values of γ and δ) to have enough scientists to conduct all necessary tests even if they were to successfully generate a question and an idea each period. There is thus some expected redundancy built into the size of the team.²⁴ Putting the stages together, the expected number of successfully implemented designs per scientist (\dot{A}/S) in a given time period is then:

$$\frac{\dot{A}}{S} = \alpha(\beta A)\gamma\delta. \quad (8)$$

Therefore, the expected number of successfully implemented designs (which then become new ideas, \dot{A} , that are added to the stock of knowledge) per scientist is equal to product of the probability of the team generating a question, the probability of the team generating an idea given a question, the expected productivity of a test generated by the generative model and the physical productivity of scientists in conducting the tests. Increases in γ and δ will raise productivity by lowering the size of the required team; increases in α and βA will raise productivity for any given size of the team.

As previously noted, Bloom et al. (2020) raise the issue of the difficulty of measuring the flow of ideas and instead define constant productivity as delivering constant (i.e., exponential) *growth* in ideas. Applying their approach here, the (total factor) productivity measure becomes:

$$\frac{\dot{A}}{A} = \omega = \alpha\beta\gamma\delta. \quad (9)$$

We thus have a rich framework for thinking about how AI might affect productivity in science in a world where the generation of questions, ideas, designs and tests reflect combinatorial problems that, at least in principle, are amenable to AI-aided solutions. However, we have also identified potentially hard to replace roles of human judgment in different stages of the process. Moreover,

²⁴It follows that an increase in either α or β will not change the size of the scientific team; however, an increase in either γ or δ will cause the size of team to decrease.

we can imagine cases where physical testing can be aided by AI and associated robotic technologies that reduce the need for physical labor in testing.

It should be noted that our idealized description of the scientific process as involving the stages of question, idea, design, and test will unfold quite differently in different domains, with the consequent potential roles for AI tools in supporting the stages also being quite different. Our drug discovery example, which motivated our description of the stages given above, might be best described as *scientific engineering*, where the stages might be: (i) the surprising absence of some artifact in the world leading to a research question/problem; (ii) generation of an idea for creating that artifact (indicating an appropriate design space); (iii) generation of (possibly ranked) designs that might instantiate that artifact; and (iv) testing of the design hypotheses.

The process might unfold quite differently in the case of *explanatory science*, where the end product is, say, an empirically supported explanation for some surprising phenomenon. The stages might be: (i) observation of the surprising phenomenon (“question”); (ii) generation of a potential explanation for that phenomenon (e.g., in the form of a formal theory - “idea”); (iii) designing a controlled (or identifying a natural) experiment for testing observable implications of theory (“design”); and (iv) actual testing of an observable implication (“test”).

The process is clearly different again in the case of *mathematics*. Assuming the mathematics is taking place within some existing formal system, the stages might be: (i) identifying some unproven conjecture within the formal system; (ii) generating an idea for the proof of that conjecture; (iii) designing the structure of the proof (noting that the design of exposition of the proof might be quite different from the actual process discovering the proof); and (iv) verifying that the proof is correct.

In some cases, one or more of the candidate processes above might be sub processes within a larger process. In economics for example, although the initial question might be generated by a surprising empirical phenomenon, the development of an economic model might follow the stages of developing a mathematical proof, which feeds back as the idea in the process of designing and testing an explanatory hypothesis.

To provide a better sense of the scope and limits of AI in supporting heterogeneous scientific

processes, we next briefly consider its use in a number of domains: material science, physics, biology, computer science, economics and mathematics.

3 AI in Science: Domain-Specific Insights

For some tasks, AI tools assist researchers across all disciplines. For example, AI tools may help scientists manage the “burden of knowledge” that results from the massive daily flow of new research (Jones, 2009). AI tools such as Notebook LM, which generates accessible podcasts from the texts of papers, and Gemini Deep Research, which generates research reviews based on user-defined plans, help scientists absorb and prioritize what would otherwise be an overwhelming flow of new literature. At the time of writing this chapter, breakthroughs are also being made in designing AI systems that integrate parts of the scientific process. Using Google’s AI co-scientist, for example, scientists can specify a research goal using natural language; the AI tool can then produce a summary of relevant literature, propose testable hypotheses, and suggest experimental designs.

While some AI tools are general, the overall impact of AI on scientific discovery is not uniform across disciplines. In the previous section, we developed a multi-stage model of scientific productivity that highlights the potential for AI to enhance search over combinatorial spaces while underscoring the enduring role of human judgment in stages requiring abductive inference, contextual nuance, and ethical deliberation. This framework allows us to explore how AI’s impacts may vary across scientific domains, where differences in data availability, the scale of search spaces, and the nature of bottlenecks shape the technology’s applicability. Domains differ not only in their reliance on empirical versus theoretical methods but also in the extent to which tasks are amenable to interpolation from large datasets or require extrapolative reasoning. In this section, we elucidate the virtue of decomposing the steps in the scientific discovery workflow as we have in the model by describing variation across domains such as materials science, physics, computational biology, mathematics, medicine, economics, and computer science (machine learning).

Material Science

We start with a story brought to light by the Nobel Prize in Chemistry (2025) to motivate our examination of the economics of AI as a transformative tool in science. The prize was awarded to

Susumu Kitagawa, Richard Robson and Omar M. Yaghi for their pioneering work on the development of metal-organic frameworks (MOFs) as a new form of material with numerous potential uses. Robson received the prize for his foundational work in developing the idea of MOFs; Kitagawa and Yaghi for their critical subsequent research establishing that MOFs could take a sufficiently stable form to actually work.

Robson’s new idea combined existing ideas. At the most basic level, it combined an analogy with the molecular structure of a diamond (in which each carbon atom bonds to four others) with the idea of attraction between ions and molecules. The combined idea was for a new type of “molecular construction” in which the ions facilitate the bonding of molecules instead of atoms, with distinctive properties.

At that time, most chemists would have assumed that combining copper ions with the four-armed molecules would result in a bird’s nest of ions and molecules. But things went Robson’s way. As he predicted, the ions and molecules inherent attraction to each other mattered and they organized themselves into a large molecular construction. Just like carbon atoms in a diamond, they formed a regular crystalline structure. However, unlike diamond, which is a compact material, this crystal contained a vast number of large cavities (Royal Swedish Academy of Science, 2025).

The idea has turned out to have considerable generality, with a vast set of possibilities for combining metal ions with molecules to create MOFs with a wide range of properties and potential uses. Put differently, Robson’s idea identified a vast combinatorial search space. As demonstrated by AlphaFold for the protein-folding problem, AI has the potential to provide tools to prioritize the search over such vast spaces.

One exciting mooted application is to the Direct Air Capture (DAC) of CO₂. In a recent collaboration, scientists at Meta and Georgia Tech used the results of 40 million quantum chemistry simulations to build a predictive model to screen a vast space of MOFs for given DAC processes (Sriram et al., 2024). However, as reported in the Financial Times (Bryan, 2025), critics have pointed to problems with the identified candidate MOFs in terms of their capacity to bind to CO₂. One critic responded that he wished “they had computed a bit less and thought a bit more.” The architects of the AI model counter that these critiques misunderstand the purpose of the research:

A.J. Medford, an associate professor at Georgia Tech who helped author the paper, said the critique had missed the point of the research. He said the team did not set out to “conclusively identify new materials,” but intended to experiment with more sophisticated techniques, as well as identify new challenges and questions in the field (Bryan, 2025).

While further research is evidently needed to determine the potential for AI to identify candidate MOFs for effective and efficient DAC, we use this example to motivate our approach to thinking about the economics of AI in science. We treat AI as a tool with potential uses across the range of scientific tasks including question generation, idea generation, design generation, and testing. In particular, we stress the transformative effects of AI tools on problems that require search over vast combinatorial spaces. However, we focus on the use of AI as a tool to augment human scientists rather than as necessarily replacing those scientists as tasks are automated. In particular, we assume that human judgment remains central, albeit to varying degrees, across the various stages of scientific inquiry.

Overall, AI tools may be used in each stage of the discovery process in material science. Question generation often stems from applied needs, such as energy storage materials, where AI can mine patents and simulations to refine objectives (Pollice et al. (2021)). Tools such as Elicit (Bernard et al., 2025) or PandaOmics (Kamya et al., 2024) can scan millions of research papers to identify “Peircean surprises,” such as unexplained structural anomalies or gaps in known property-structure relationships. At the same time, human judgment is required to evaluate the normative significance of these questions, such as prioritizing the development of sustainable energy-storage materials over less impactful alternatives.

For idea generation, researchers may use a variety of AI tools to explore new idea combinations for molecular scaffolds and hypothesize novel compounds, such as combining ionic attraction with diamond structures; however, because these models can propose chemically impossible structures, a scientist must use their judgment to prioritize AI-generated ideas to increase the likelihood of discovery. Overall, idea generation involves hypothesizing novel compounds, with generative models like those from Cheng et al. (2021) aiding in exploring chemical spaces exceeding 10^{60} possibilities. Here, AI augments β and γ by enabling rapid iteration, but bottlenecks arise in testing due to costly

physical synthesis and characterization. Human judgment is indispensable for assessing real-world feasibility, including environmental impacts or scalability.

Specialized AI models enhance design generation. AI excels in property prediction, property optimization and chemical space exploration, including applications for the development of new materials for energy generation and storage (Pollice et al., 2021). Modern generative models also make possible the use of “inverse design,” where the model generates candidates based on specified desired properties (Cheng et al., 2021). As an example of another approach, GNoME was used to predict the thermodynamic stability of over two million potential crystal structures, identifying 380,000 stable candidates and significantly accelerating discovery by filtering out unviable designs before performing expensive first-principles simulations (Merchant et al., 2023). However, human judgment is still required to rank these designs and identify candidates that may struggle with real-world feasibility, such as materials that fail to physically bind CO_2 despite predicted efficacy.

Finally, in the testing stage, autonomous robotic laboratories and high-throughput synthesis platforms conduct high-speed characterization of new compounds. Yet, the final causal validation remains a human responsibility, as specialist chemists must interpret complex experimental data to distinguish between the real structure of a phenomenon and visible artifacts of the robotic instrumentation.

Physics

Not only is the role of AI different in physics compared to material science, but even within physics the role of AI is different for theoretical compared to experimental physics. For example, in the question generation stage, experimentalists use tools like AI Feynman to scan massive datasets for Baconian induction, identifying patterns or anomalies that suggest new physical laws (Udrescu and Tegmark, 2020), whereas theorists utilize semantic research agents to surface Peircean surprises (conceptual gaps or contradictions in existing mathematical models). Although AI can flag surprising deviations, human judgment remains essential for the theorist to decide if an anomaly might reflect a foundational insight or for the experimentalist to verify if a signal is simply an artifact of instrument noise.

For idea generation, researchers in both branches use LLMs to explore new idea combinations

and brainstorm abductive guesses. At the same time, because these tools rely on the interpolation of existing knowledge, a human physicist must apply their unique analogical reasoning to evaluate whether a proposed hypothesis represents a paradigm-shifting insight or merely a statistically likely but physically irrelevant guess.

In the design generation phase, experimentalists employ symbolic regression engines to discover interpretable equations from vast data (see, e.g., Udrescu and Tegmark (2020)). AI helps physicists find patterns in vast volumes of data produced by experiments, enhancing α through anomaly identification (Carleo et al., 2019). However, in theoretical physics, where search spaces are conceptual rather than empirical, AI's role in α and β is more limited, as tools like symbolic regression (Udrescu and Tegmark, 2020) rediscover equations but struggle with paradigm-shifting insights requiring analogical reasoning. That said, theorists use quantum simulator AI tools to model the structures of potential physical systems. Still, human judgment is required to rank these designs and resolve trade-offs and judgment is critical for inventing hypotheses from first principles when there is no data to interpolate from.

Finally, in the testing stage, AI-based trigger systems at facilities like the Large Hadron Collider autonomously filter trillions of collisions (Govorkova et al., 2022), whereas theorists use proof-assistant software to verify the logical coherence of a new derivation. However, the final causal validation depends on the human's ability to interpret the statistical significance required to determine whether the results reflect meaningful hidden laws of nature.

Computational Biology

In the question generation stage, AI tools such as Kosmos (Mitchener et al., 2025) or Asta (Bragg et al., 2025) scan millions of abstracts to identify Peircean surprises, such as overlooked correlations between genomic variants and rare metabolic phenotypes. Yet, human judgment is required to determine the ethical and social significance of these gaps, ensuring that the research focuses on questions with patient value rather than technically impressive but trivial patterns.

In idea generation, researchers utilize specialized models like Geneformer to simulate in silico gene manipulations and generate abductive guesses regarding novel regulatory pathways (Theodoris et al., 2023). However, because these models can generate noisy connections, a scientist must use

their judgment to synthesize these fragments into a coherent biological theory that aligns with first principles.

During design generation, generative multimodal models like ESM3 enable the prompt-based design of new protein sequences and structures guided by specific functional constraints (Hayes et al., 2025). At the same time, human researchers must still evaluate the Pareto frontier of these designs to resolve critical trade-offs, such as a binder’s theoretical affinity versus its actual developability and toxicity risks in complex cellular environments.

Finally, in the testing stage, AI agents like Science Machine’s Sam can execute complex differential expression and GSEA analyses instantly (Cavga, 2026), while projects like OPAL orchestrate autonomous laboratories to validate designs (Oak Ridge National Laboratory, 2026); however, the final interpretation of results, distinguishing between a true biological discovery and an artifact of the computational “black box,” remains a uniquely human responsibility that relies on contextual nuance and root-cause analysis.

Mathematics

In mathematics, AI systems like AlphaProof demonstrate prowess in solving competition-level problems, impacting γ by generating verifiable proofs. Yet, question generation (α) often involves aesthetic or foundational pursuits that defy data-driven interpolation, and the field’s emphasis on rigor demands human validation to ensure logical coherence. More specifically, in the question generation stage, theoretical mathematicians use specialized conjecture generation systems to identify Peircean surprises, such as unproven properties in high-dimensional manifolds, while applied mathematicians use RAG-based tools to scan empirical data for gaps in existing physical or economic models. However, human judgment is required for the theorist to determine the aesthetic value of a conjecture and for the applied scientist to evaluate the normative significance of a problem, ensuring it addresses real-world needs.

For idea generation, both theoretical and applied mathematics utilize LLMs like Claude and Gemini 3 Deep Think for neuro-symbolic brainstorming, generating abductive guesses for proof strategies or algorithmic frameworks. At the same time, because these models operate on interpolative patterns and lack a true world model, a mathematician must use their judgment to verify

that a proposed strategy is logically sound.

During design generation, theorists use tools like AlphaProof to auto-formalize natural language whereas applied mathematicians use symbolic regression to discover interpretable equations from vast datasets. Yet, the researcher must still act as the primary architect to provide the high-level intuition that guides the AI through the combinatorial explosion of potential solutions.

Finally, in the testing stage, the AI tools like the Lean 4 kernel verifies the mechanical correctness of a theoretical proof, while applied mathematicians use high-speed simulation to validate models against real-world data. Still, the final validation remains a social process where human reviewers must judge whether the result makes a worthwhile contribution to the extant literature.

Medicine

Medicine represents a domain where AI's promise is particularly pronounced, driven by vast combinatorial search spaces and abundant data from genomics, proteomics, and clinical trials. In question and idea generation (stages characterized by α and β in our model), AI tools, such as large language models, synthesize the literature and propose novel hypotheses, such as identifying gene interactions for rare diseases. However, human judgment remains critical for prioritizing questions with revenue and ethical implications, such as those involving large versus small patient populations (i.e., addressable market size) as well as patient privacy in personalized medicine.

Design generation (γ) has seen dramatic advances, exemplified by AlphaFold's prediction of protein structures, which compresses search over folding possibilities that would otherwise take years. The search space is so vast due to the massive number of ways (combinatorial) a protein can fold. (δ) benefits from AI in analyzing high-throughput screening data, as in the discovery by researchers at McMaster University and MIT of the molecule abaucin as an antibacterial agent for the bacterium *Acinetobacter baumannii* (Liu et al., 2023).²⁵

Overall, in a data-rich field like medicine, AI could substantially boost ω by alleviating bottlenecks in γ , enhancing productivity if complementary investments in data curation and model interpretability are made. Regulatory hurdles and the need for causal validation via clinical testing

²⁵AI-based screening of a large space of chemical compounds led to the identification of a new candidate antibiotic with hypothesized effectiveness against a particularly problematic species of bacteria – *Acinetobacter baumannii*. Another area of biology and medicine being transformed is genomics – which again involves massive search spaces of gene interactions and expressions.

ensure human oversight in the loop and may create a bottleneck at δ .

Recent empirical evidence reinforces the existence of these bottlenecks. For instance, Cavalli (2024) finds that the introduction of AlphaFold led specialist laboratories to expand their workforce by hiring more biologists to validate and interpret AI predictions instead of shrinking it. Similarly, Kim (2025) demonstrates that while such tools accelerate progress in data-rich areas, they may not autonomously solve the problem of exploration in data-sparse domains without directed human inquiry. These findings suggest that in high-complexity domains, AI serves as a complement that increases the returns to domain-specific human judgment.

Economics

Economics benefits from AI in testing (δ) through machine learning for causal inference, but combinatorial spaces in idea generation (β) involve navigating behavioral complexities and ethical considerations, where data sparsity and endogeneity pose bottlenecks. For instance, AI can process survey data for hypothesis testing, but formulating questions about inequality or policy requires contextual sensitivity that AI currently augments rather than replaces. Across these domains, the model predicts that AI will yield uneven productivity gains, with ω increasing more in interpolative, data-abundant fields, while judgment-intensive areas underscore the need for augmentation strategies.

For example, in the question generation stage, empirical economists may use AI tools to scan massive datasets for Baconian induction, identifying unexpected correlations or Peircean surprises in consumer behavior or market trends, while theorists may utilize LLM-powered agents like EDSL to simulate diverse agent rationales (Horton et al., 2023) and identify gaps in existing axiomatic frameworks. Once again, however, human judgment remains essential for both to determine the normative significance of these questions.

For idea generation, both theoretical and empirical researchers may employ LLMs for brainstorming, generating abductive guesses or for setting the initial assumptions that might underpin a model. Yet, because these models operate on interpolative patterns and lack a true world model of human behavior, an economist must use their judgment to verify that a proposed hypothesis is logically coherent and grounded in economic first principles.

In design generation, empiricists may use AI to prioritize search over vast combinatorial spaces for optimal experimental or survey designs, while theorists may use AI tools to explore the implications of different sets of model assumptions. As elsewhere, the researcher must still act as the primary architect to resolve trade-offs between model complexity and interpretability, ensuring the design offers insights rather than an impenetrable black box.

Finally, in the testing stage, AI models provide an “algorithmic baseline” to measure the completeness of a theory by predicting variation in dependent variables, while empirical models are validated through automated high-speed simulations. Still, the final causal validation rests with the human scientist, who must interpret the fragments of understanding provided by the AI and put them into context to make a scholarly contribution or to put them into use for a policy application.

Computer Science (machine learning)

In AI research the discovery process is operating at a high velocity because it is predicated on digital code that can be compiled and tested almost instantaneously, creating a compressed feedback loop that distinguishes it from slower-moving physical sciences. In the question generation stage, researchers may use AI-powered literature tools like Asta (Bragg et al., 2025) or Elicit (Bernard et al., 2025) to identify Peircean surprises, such as unexpected scaling-law deviations or emergent properties in model architectures, which prompt new inquiries into neural efficiency. However, human research judgment is required to evaluate the normative significance of these questions, ensuring that the focus remains on safe, aligned, and societally beneficial intelligence.

For idea generation, LLMs and specialized reasoning models like OpenAI o3 or Gemini 3 Deep Think can be used for abductive brainstorming, suggesting new sets of assumptions for loss functions or attention mechanisms. However, because these models lack a true world model, a computer scientist must apply their tacit knowledge to ensure the proposed architectural guesses are grounded in algorithmic first principles.

During design generation, AI agents like OPAL or Science Machine’s Sam prioritize search over vast combinatorial spaces of hyperparameters and network topologies, iterating through millions of code permutations at a speed that would be impossible in a physical lab. At the same time, the human must still act as the primary architect to resolve trade-offs between computational cost and

interpretability, preventing the resulting systems from becoming completely opaque black boxes.

Finally, in the testing stage, the research is dramatically accelerated because designs can be validated by simply compiling and running the code on high-performance clusters, often optimized by tools like Opal, allowing for thousands of experiments per day. Still, the final causal validation remains a human responsibility, as the scientist must verify that the model’s performance represents a genuine breakthrough in understanding rather than a visible artifact of overfitted training data.

4 AI as a General Purpose Meta Technology

A theme running through this volume is that of meta-science or ideas about how to do science. In turn, meta-science can be viewed as a subset of Paul Romer’s concept of a meta-idea – or an idea about how to generate ideas. A scientific meta-idea might involve a new way to fund science or a new way to conduct peer review; the more general category of meta-idea would also include such ideas as ideas for new way to design the patent or copyright system (e.g., how to design copyright in the age of LLMs trained on large quantities of published material). In Agrawal et al. (2019), we use the term meta-technologies to describe meta-ideas that are embedded in a particular technological form. General purpose meta-technologies (GPMTs) – e.g., the printing press – are then technologically embedded ideas that have general applicability.

It is instructive to consider AI as a GPMT in terms of the broader literature on general purpose technologies. Using both models and historical case studies, this literature highlights the need for complementary investments before the full impact of a GPT is realized. These investments typically need to occur both upstream and downstream of the GPT. Upstream investments might include research to develop a better scientific understanding of the technology; downstream investments might include reorganizations of workflows to take better advantage of the new technology (e.g., redesigning the factory floor to take advantage of distributed electric power) or equipping workers with the expertise to use the technology.

The jagged technological frontier of AI can be thought of as a manifestation of a GPMT in its early stages. Upstream, the revealed limitations of AI in actual uses are incentivizing investments to overcome those limitations (e.g., investments in improving the reasoning capabilities of AI).

Downstream, the users of AI are experimenting with new work designs to improve the productivity gains from integrating AI (e.g., experimenting with how best to integrate AI into work teams: Caplin et al. (2024); Dell’Acqua et al. (2025); Weidmann et al. (2025)) and training workers in the use of AI.

In the specific case of science, the challenge of using AI to solve specific problems is motivating the development of bespoke technologies. This is especially evident in the work of Google DeepMind in the development of technologies such as AlphaFold (protein-folding prediction), AlphaProof (mathematical problem solving) and AlphaGeometry (solving problems in geometry). Although these technologies are developed to solve particular problems, by solving targeted problems and overcoming specific bottlenecks, the resulting technological breakthroughs have potentially a much wider range of application. An example of a downstream investment is the development of high-speed autonomous experimentation to take advantage of AI’s capability to prioritize experiments and rapidly incorporate the results of those experiments into improved prediction models, including more exploratory experiments to produce data on poorly understood regions of the design space.

It is revealing that experts guiding practitioners on the productive use of AI have emphasized that current AI tools are the worst they will ever use and also the need for ongoing experimentation to find the best current use cases (Mollick, 2024). The use of AI in science is no exception. As an early-stage GPMT, the frontier is jagged and currently limits are frequently exposed. Demis Hassabis’ warning seems apt: expectations of progress in the short term may be overdone (due to existing bottlenecks); but the history of GPTs cautions against underestimating the potential to overcome those bottlenecks over the longer term.

The next two sections focus on specific ways in which the shape of the jagged frontier of AI in science might evolve. The next section looks downstream to interpret a simple task-based augmentation model to explore how the availability of expertise in the *use* of AI will impact the productivity gains from advances in AI as a GPMT. The key insight is that the productivity gains from advancements in AI – captured simply as an increase in the range of tasks that can be done by scientists with both normal expertise and expertise in the use of AI – are increasing with the share of the scientific workforce that are equipped with AI expertise. Section 6 then looks upstream

to examine the major forces affecting the development of the GMPT itself. This allows us to look at the trends that are being intensely debated by AI scientists themselves, such as the relative importance of scaling existing models and the role of architectural/algorithmic improvements.

5 AI Skills and Scientific Productivity: Insights from Worker-Augmentation Models

To this point, we have treated AI as a tool available to scientists to help perform various tasks in the scientific workflow. However, this leads to a critical question: do scientists possess the necessary complementary expertise to actually utilize these tools? To address this, we turn to the framework developed in Agrawal et al. (2026b) “Enhancing Worker Productivity Without Automating Tasks: A Different Approach to AI and the Task-Based Model,” which provides two distinct models for how AI might augment human labor rather than replacing it.

5.1 AI Expertise and the Extensive Margin

The first model in this framework focuses on how specialized AI expertise intermediates productivity gains. In Agrawal et al. (2026b), the workforce is divided into “AI-expert” workers, who possess both domain knowledge and the skills to use AI, and “ordinary” workers who lack AI-specific training. Scientific output is generated through a continuum of tasks, where the current state of technology limits the range of tasks that can be performed using AI. Within this constrained equilibrium, AI-expert scientists have a comparative advantage in tasks where the technology is feasible, and firms would ideally utilize more AI-expert labor if the technological boundary expanded. Aggregate productivity is thus determined by the balance between the technological feasibility of using AI and the relative supply of scientists equipped with the expertise to deploy it.

In the context of science, this implies that technological progress in AI occurs at the extensive margin, expanding the range of scientific tasks that can leverage AI tools. A larger supply of “AI-expert” scientists in the workforce acts as a vital lever that amplifies aggregate productivity gains while simultaneously mitigating the exclusionary wage premiums that would otherwise accrue to a

few scarce experts.

For example, in data-rich domains like biology, AI expertise is rapidly “opening” tasks. Scientists equipped with AI skills can now navigate massive combinatorial spaces, such as genomics and protein folding, that were previously limited by human cognitive bandwidth. Conversely, in theoretical physics, the extensive margin expands more slowly. Progress here is more constrained because AI currently struggles with “paradigm-shifting” insights that require deep theoretical judgment and analogical reasoning rather than the data-driven pattern matching seen in interpolative tasks.

The Agrawal et al. (2026b) model predicts that extensive-margin technological improvements in AI typically increase wage inequality because the wages of AI-expert workers rise relative to ordinary workers. AI experts benefit from both a substitution effect as they take over tasks previously done by ordinary workers and a productivity effect, while ordinary workers suffer from a displacement effect that can only be partially offset by aggregate productivity gains. However, the model demonstrates that a larger share of AI-expert workers in the total workforce attenuates this rise in inequality. By increasing the relative supply of experts, their relative wage is driven down, which not only lowers overall inequality, but also enhances the total productivity gain by making AI expertise more economical to deploy across more tasks.

In our scientific context, this reveals a divergence in the income distribution risks between biology and physics. In biology, where the task-opening effect of AI is high, a small cadre of early-adopting AI-expert biologists may capture a large wage premium as they monopolize the newly feasible design spaces. In contrast, theoretical physics may see more stable income distributions because of its expertise barrier; because the extensive margin of AI expands slowly, ordinary physicists remain insulated from displacement for longer. However, this stability comes at the cost of the productivity amplification seen in data-rich fields, suggesting that without intentional training to turn physicists into AI experts, the field may struggle to capture the same transformative growth rates seen in the life sciences.

5.2 Non-AI Skills, Bottlenecks, and Concentration

The second model in Agrawal et al. (2026b) shifts focus to non-AI skills, assuming that while AI tools may be ubiquitous, researchers differ in their underlying domain knowledge. In this model, AI can either “open” task opportunities by reducing the need for narrow technical skills or “close” them where AI is highly complementary only to specialized expertise.

To “open” a task opportunity means that AI tools reduce the need for narrow technical expertise, effectively lowering entry barriers and allowing scientists with general analytical skills to perform research tasks that were previously restricted to specialized experts. An example of this is seen in the social sciences, where AI for causal inference and simulated data can open empirical tasks to a much broader pool of researchers by simplifying data analysis. Conversely, to “close” a task opportunity occurs when AI is highly complementary to a narrow set of specialized human skills, which effectively raises the skill threshold for participation and concentrates the work among a smaller group of practitioners. An example is in theoretical physics or advanced mathematical modeling, where AI tools might only benefit those capable of formulating high-quality hypotheses or complex design constraints, potentially causing others to be pushed out of these critical conceptual stages .

This second model treats AI as effectively available for all tasks but requires workers to allocate themselves based on their heterogeneous non-AI skills and the varying skill requirements across tasks. An improvement in AI tools changes the task-specific productivity parameters and the subset of tasks a worker can compete in, inducing an equilibrium reallocation of scientists across the workflow. A central finding is that aggregate productivity and wage inequality depend on different global properties of this equilibrium distribution: productivity is governed by the alleviation of bottlenecks (measured by geometric mean uniformity), while wage inequality is driven by the concentration of labor in a narrow set of tasks (measured by Shannon differential entropy).

Geometric mean uniformity is a metric that measures the evenness of skill distribution across all tasks, where a high value indicates that no single task is being “bottlenecked” by a critical lack of workers . For example, in data-rich biology, AI can increase this uniformity by filling ”near-holes” in coverage, allowing more scientists to effectively participate in genomic and proteomic tasks that were

once thinly staffed. Conversely, AI can decrease geometric mean uniformity in theoretical physics if the tools remain "task-closing" by only complementing specialized mathematical modeling; this causes talent to drain from critical conceptual tasks, creating new, thinly staffed bottlenecks that weigh down aggregate productivity.

Shannon differential entropy is a measure of how concentrated labor is within a specific subset of tasks, with lower entropy values reflecting a high concentration of workers that typically drives up wage inequality. An example occurs in the social sciences, where AI for data processing may lead to a sharp concentration of human effort in "idea generation," creating a high wage premium for the few scientists with the elite judgment required for that specific stage. However, AI can increase Shannon differential entropy—thereby reducing wage inequality—when it acts as a "task-opening" force that reduces the specificity of skill requirements across a broad range of tasks. For instance, decision-support systems that allow researchers with general analytical skills to perform tasks previously requiring narrow technical expertise will flatten the skill-density schedule, spreading labor more uniformly and raising entropy values.

To clarify, the difference between geometric mean uniformity and Shannon differential entropy lies in how they weight the distribution of workers across tasks and the specific economic outcomes they predict. Geometric mean uniformity governs aggregate productivity and is highly sensitive to thinly staffed regions or near-holes that create bottlenecks, whereas Shannon differential entropy drives wage inequality by responding to concentrations or spikes of labor within a narrow set of tasks. For example, in the social sciences, AI might improve productivity by filling task gaps in data analysis (raising geometric mean uniformity) while simultaneously increasing inequality by concentrating human labor in elite idea generation (lowering Shannon differential entropy).

In the context of science, this implies that technological progress in AI often yields a non-monotonic co-movement between productivity and inequality, as the aggregate outcomes depend on the uniformity of skill densities across the research landscape. Aggregate productivity is governed by geometric mean uniformity and is highly sensitive to "bottlenecks," which are thinly staffed tasks that constrain the entire process. In Materials Science, for instance, while AI opens the design stage, it creates a concentration of workers there, leaving physical synthesis as a persistent,

thinly staffed bottleneck that governs the overall speed of discovery.

The overall implication of advancing AI on productivity depends on whether the technology flattens or spikes the distribution of scientific talent across necessary tasks. This results in a contrast between high and low productivity impacts depending on the field's underlying data structure. In data-rich environments, like biology, AI advancements typically have a high productivity impact by acting as a task-opening force; they fill existing near-holes in coverage by enabling a broader range of scientists to navigate high-dimensional genomic and proteomic spaces, thereby increasing Geometric Mean Uniformity. In contrast, in theoretical physics, AI tools may have a lower productivity impact if they remain task-closing. If these tools only complement a narrow sliver of specialized mathematical modeling, they can cause talent to drain from critical conceptual tasks, creating new bottlenecks that weigh down the overall speed of progress through a reduction in the geometric mean of skill densities.

Inequality is driven by the concentration of researchers into a narrow set of tasks, measured by Shannon differential entropy. In the Social Sciences, AI for causal inference and simulated survey data might open empirical tasks to a much broader pool of researchers, potentially lowering entry barriers to data analysis. However, this induced reallocation may lead to a concentration of human labor in idea generation, where contextual sensitivity and behavioral complexities demand a high wage premium for elite judgment, increasing inequality even as overall productivity rises.

This pattern might contrast sharply with the dynamics suggested by early indications in Biology and Medicine. In these domains, AI breakthroughs like AlphaFold seem to be opening the combinatorial stage of design, compressing years of protein-folding research into negligible timeframes. Instead of concentrating talent at the conceptual start of the workflow, this shift may drive a concentration of researchers toward the testing and validation stage. Because AI-generated designs require causal validation via physical synthesis and clinical trials, early indications suggest specialist laboratories may be responding by expanding their workforces to hire more biologists capable of interpreting and verifying AI predictions.

Ultimately, both fields experience a spiky distribution of skill supply that elevates wage inequality, but the location of the bump differs: in the Social Sciences, the premium may be more likely

to accrue to conceptual ideators, while in Biology, it may shift toward technical validators who manage the physical bottlenecks of the laboratory.

5.3 Frameworks for Policy Development

Drawing on the models presented, the economic impact of AI in science is determined not solely by the technological frontier, but by the *elasticity of the skill supply* in response to shifting task requirements. As AI changes the science of AI upstream and the reorganization of workflows downstream, it raises fundamental questions about the optimal direction of public policy. Rather than providing prescriptive recommendations, the task-based framework suggests a series of empirical questions and considerations for policymakers tasked with shaping the future of scientific labor.

A primary consideration for policymakers is how to balance different types of economic responses to AI-driven change. Under a replacement-centered view, where AI automates most cognitive tasks, policy discussions naturally gravitate toward defensive measures such as universal basic income (UBI) to manage labor obsolescence. Conversely, if AI functions as a tool for augmentation, as the model we have drawn from assumes, a more offensive strategy focusing on human capital investment may be required.

To what extent is contemporary AI displacing scientists versus providing them with “bicycles for the mind” that enhance their productivity (Agrawal et al. (2025))? Empirical evidence on whether AI is primarily task-opening or task-closing is essential to determine the appropriate mix of these strategies.

The models suggest that the nature of the investment in human capital depends heavily on the specific domain of science. In fields like Genomics and Chemistry, where AI is rapidly expanding the range of feasible tasks, how can policy facilitate the broadening of AI-specific expertise across the entire workforce? If expertise remains concentrated in a few elite labs, it risks creating significant exclusionary wage premiums. In fields like Mathematics or Theoretical Physics, where AI struggles with abductive inference and paradigm-shifting insights, what frameworks are needed to identify and foster the higher-order skills required to alleviate critical productivity bottlenecks? For example, in theoretical physics, while AI can assist in identifying patterns in massive experimental

datasets, it currently lacks the capacity for the abductive 'guessing' required to propose a fundamental new law of nature; so, policy must focus on fostering the elite human judgment necessary to navigate these conceptual bottlenecks that pattern-matching tools cannot resolve.²⁶

Policymakers require longitudinal data on scientific workflows to quantify the parameters of these augmentation models and map where the jagged frontier of AI capabilities currently sits across different domains.

The integration of AI into the learning process creates a potential paradox for skill acquisition. While using AI for writing or coding can build AI facility, there are concerns that it may “short-circuit” the development of critical thinking and creative problem-solving – the very higher-order skills that are most complementary to AI over a career. Recent large-scale evidence provides empirical grounding for these concerns. Analyzing 41 million research papers across the natural sciences, Hao et al. (2026) find that while AI adoption is positively associated with individual productivity gains – scientists publish 3.0 times more papers, receive 4.8 times more citations, and become research project leaders 1.4 years earlier than non-adopters – it is associated with a contraction in the scope of what scientists subsequently choose to explore. Specifically, Hao et al. (2026) document a 4.6% contraction in the “knowledge extent” (the breadth of topics studied) of AI-augmented research and a 22% reduction in follow-on scientific engagement, as papers citing AI work interact less with each other. These patterns suggest that AI may be concentrating scientific attention on data-rich, established problems rather than fostering exploration of new research directions. Whether this concentration reflects efficient resource allocation toward tractable problems or a worrying contraction in the diversity of scientific inquiry remains an open question with significant implications for research funding priorities.

How can educational institutions design curricula that encourage facility with AI without undermining the acquisition of the scientific judgment that keeps the human in the loop? And how

²⁶At the India AI Impact Summit in New Delhi (February 2026), Google DeepMind CEO Demis Hassabis proposed a new standard for evaluating Artificial General Intelligence (AGI) which he called the “Einstein Test.” He defined the benchmark as follows: “To test whether a system had truly achieved AGI... training a model with a knowledge cut-off of 1911 and seeing whether it could independently arrive at general relativity, as Einstein did in 1915.” Hassabis noted that while current systems like AlphaFold are powerful tools for navigating known combinatorial spaces, they do not yet possess the abductive reasoning required to “invent their own hypotheses or conjectures about science” from first principles.

might scientific institutions balance the individual incentives favoring AI adoption against potential collective costs in terms of the breadth of scientific exploration? The transition for science into the Age of AI does not mandate a single policy path. Instead, the task-based augmentation framework provides a lens through which to view the dynamic interplay between tool advancement and human skill formation, highlighting the need for empirical research to guide high-stakes decisions in scientific funding and education.

6 The Science of AI

AI is distinctive among general purpose technologies in terms of the degree of dynamism in the development of the underlying technology. Compared to, say, the internet – or even electricity – where the core features of the technology were in place from an early stage, the capabilities of AI, and especially AIs based on an underlying deep-learning technology, are making significant advancements on a monthly basis. These advancements are altering the shape of the jagged frontier, increasing the range of tasks where human capabilities, especially judgment, can be augmented by AI.

A major debate within the AI literature itself is the timeline to AGI. If AGI is thought of as AI that can match human-level capabilities in practically all tasks, this effectively requires that the AI can emulate or simulate human judgment. While we have expressed scepticism that this can be achieved even over the medium term, there is little doubt that the advancements in the underlying AI GPMT will alter the opportunity set for both augmentation and automation in science.

Two ideas predominate in discussions of the future path of AI – scaling and algorithmic improvement. Optimists in particular draw from historical experience to forecast dramatic improvement in AI performance from simply scaling – more parameters, more data, more compute – existing architectures and algorithms. Others see the need for substantial advances in architectures and algorithms to achieve AGI, with varying degrees of confidence that these advances will be forthcoming. Keeping the unusually high level of disagreement within the AI research community in mind, we briefly review the upstream forces – the “science of AI” – that will affect the evolution of jagged frontier of AI as a GPMT in science.

6.1 Scaling

An influential idea in the AI literature is that of empirical scaling “laws.” For the transformer architecture, Kaplan et al. (2020, p. 3) find that:

Model performance depends most strongly on scale, which consists of three factors: the number of model parameters N (excluding embeddings), the size of the dataset D , and the amount of compute C used for training. Within reasonable limits, performance depends very weakly on other architectural hyperparameters such as depth vs. width.

A key finding is that model performance tends to follow a power-law relationship with the key inputs.

To fix ideas in relation to the implied economics, it is useful to think of what conditions would be required to have exponential improvement in performance given an underlying power-law relationship and a single input, compute. If we assume that the relationship between performance (P) and compute (C) is given by the power law $P = C^\mu$, and that available compute is increasing exponentially over time (i.e. $C = e^{\varphi t}$, where φ is the growth rate of compute and t is time), then performance will increase exponentially with time according to: $P = e^{\mu\varphi t}$. Attention then focuses on the resource budget required to sustain exponential growth in the availability of compute. A fixed budget would be sufficient if the price of compute is falling exponentially in line, say, with Moore’s law. However, if the price of compute is falling sub-exponentially (which appears to have been the recent experience), then increasing budgets are required to sustain exponentially improving performance.

In reality, ever-larger budgets are being used to train state-of-the-art AI models. This then focuses attention on the revenue models of the leading AI companies that underpin their capital raising efforts needed to fund the necessary compute. To date the leading companies have been successful in raising the necessary capital to fund their model training (and inference) costs, at least partly underpinned by revenue models (e.g., monthly subscriptions). But there is a question of whether it will remain economical to continue to train ever-larger models. Whether it does will depend, *inter alia*, on performance improvements that result from the larger models and the ability of companies to extract revenues based on that performance.

The foregoing simple account obviously glides over many issues that are prominent in the scaling debate. On the technical side, the nature of the scaling relationships is more complex than our simple single-input power-law type relationship implies, with complex interactions along such dimensions as parameter count, data requirements and computing power (see, for example, the so-called Chinchilla scaling law (Hoffmann et al., 2022)). A second issue is the ease with which follower-models, notably open-source follower models – can come close to matching the performance of expensively trained proprietary leader models at significantly lower costs. The potential implications were brought into focus with the introduction of the DeepSeek’s R1 model in early 2025, which achieved a high level of performance at just a fraction of the training cost of the leader models, causing concerns about the valuations of not just the leading AI companies, but also of companies providing the high-end chips (notably NVIDIA) used to train these models. A third issue is the relevance of computing costs at the inference stage in addition to the training stage. This has come into sharper focus with the development of reasoning models, and their high demands for inference-time compute. Interestingly, new scaling approaches are being developed relating the performance of reasoning models to this use of inference-time compute. Finally, looking beyond compute costs and performance, the energy required for model training and inference has received significant attention, with concerns raised about the capacity of energy infrastructures to supply the needed electricity and the environmental costs of generating that additional electricity assuming it does become available.

6.2 Algorithmic development

Beyond scaling, the other major route to greater AI capabilities is improved architectures and algorithms. Richard Sutton’s brief essay describing what he describes as the “bitter lesson” from 70 years of AI research has had a big impact on this debate.²⁷ In particular, Sutton, a Turing Prize winner, notes the failures of attempts to code human knowledge into AI:

Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in

²⁷The essay is available from Richard Sutton’s website at: <http://incompleteideas.net/IncIdeas/BitterLesson.html>.

the long run is the leveraging of computation.

While Sutton underlines the importance of computation, we do not read his argument as downplaying the importance of algorithmic development in general, but as questioning the utility of methods aimed at incorporating human knowledge into AI. Sutton notes that architectural and algorithmic methods that scale well (e.g., deep learning and reinforcement learning) have been successful.

One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are *search* and *learning*.

Recent algorithmic and architectural innovations that have fueled the development of scalable AI include the transformer architecture that underlies modern LLMs, the method of reinforcement learning from human feedback (RLHF) used to fine-tune LLMs, and the RL techniques applied at inference time to simulate reasoning capabilities.

Not all AI scholars who subscribe to the search and learning paradigm believe that further fundamental architectural innovations will not be required. For example, Yann LeCun, another Turing Prize winner, goes so far as to call LLMs an “off-ramp” on the road to AGI. Instead, he argues that alternative architectures that allow AIs to learn a “world model” will be necessary to advance to models that can match human judgment in tasks such as reasoning and planning. Other AI scientists believe that achieving AGI will require the integration of neural network and symbolic methods (which presumably will require the direct coding of at least some aspects of human knowledge) in the form of “neuro-symbolic” methods. We can think of such methods as coming closer to the type of integration of Kahneman’s System 1 and System 2 that seems central to the effectiveness of human judgment. Whatever the path along which the science of AI develops, it is likely to involve continuing algorithmic and architectural development, notwithstanding that models will also certainly become ever-larger in scale. Even if human judgment stays robustly important, these upstream developments will have profound effects on the shape of the frontier in the application of AI as a GPMT in science.

7 Possible Downsides of AI in Science

We have concentrated on how the availability of AI tools could affect scientific productivity across different stages of the scientific process. As the volume of AI-aided scientific research has expanded, attention is also turning to the potential downsides of AI use in science. Although the study of the effects of AI on science is still in its infancy, two early concerns stand out: the effects on the peer-review system and the implications for the “burden of knowledge” (Jones, 2009).

We draw in particular on two large-scale studies of the use of AI in science that cover different time periods. Hao et al. (2025) detect and examine the effect of AI use – ranging from early machine-learning-based predictive models to modern generative AI – over the period 1980 to 2024. Kusumegi et al. (2026) concentrate on the more recent period 2018 to 2024 associated with the use of deep-learning in science. The first thing to note about these studies is that they each find large increases in measured scientific productivity from the use of AI tools. Using a dataset of 41.3 million papers drawn from across the natural sciences, Hao et al. (2025) find that scientists who use AI tools in their research publish 3.02 times more papers than scientists who do not use these tools. They also find that the AI-using scientists receive 4.84 times more citations and become research project leaders 1.37 years earlier than their non-AI-using-peers. Focusing on the more recent period, Kusumegi et al. (2026) find that that LLM adoption is associated with substantial increases in researchers’ raw productivity based on three preprint repositories: 36.2 per cent for arXiv, 52.9 per cent for bioRxiv, and 59.8 percent for SSRN. Although effect sizes varied by AI-use detection method, all methods showed substantial productivity increases.

Rauch (2021) argues that the success of “liberal science” depends on the observance of two rules that underline the importance of collective checking in what he calls the “ultimate social network” in science. These rules are:

1. “The fallibilist rule: No one gets the final say. You may claim that a statement is established as knowledge only if it can be debunked, in principle, and only insofar as it withstands attempts to debunk it.” (Rauch (2021), p. 88).

2. “The empirical rule: No one has personal authority. You may claim that a statement has been established as knowledge only insofar as the method used to check it gives the same result

regardless of the identity of the checker, and regardless of the source of the statement.” (Rauch (2021), p. 89.)

The achievements of liberal science have depended on the voluntary system of peer review. For the most part, scientists who submit their work to journals accept a responsibility to review the work of others. This review work (including the work of editors) can, at least when done well, be hugely time consuming and takes scarce time away from scientists’ own research. The surge of scientific output creates a risk that this system will be overwhelmed. While the burden on peer reviewers can be lessened by more stringent filtering at the submission stage, the improved writing that is possible using LLMs may make it harder to judge which submissions should be advanced rather than “desk rejected.” AI itself might eventually provide reliable tools for conducting review – which is arguably already the case with the lean proof-assistant system in mathematics – but for the moment AI-assisted reviews may be further working to undermine the system rather acting as an effective counter-weight to the flood of submissions. Kusumegi et al. (2026)(p. 1243) worry that as “traditional heuristics break down, editors and reviewers may increasingly rely on status markers such as author pedigree and institutional affiliation as signals of quality, ironically counteracting the democratizing effects of LLMs on scientific production.”

How are the increased use of AI tools in science and associated increases in scientific productivity affecting the burden of knowledge? On the one hand, the capability of AI tools such as LLMs to aid with literature searches should ameliorate the burden. On the other hand, to the extent that AI tools dramatically increase scientific productivity, the volume of potentially related work could increasingly overwhelm researchers. One open issue is how the use of AI tools will affect the diversity of the knowledge base that researchers combine to generate their contributions. Here, the early evidence is mixed. Kusumegi et al. (2026) find that AI-assisted search behaviour is associated with a shift towards a more diverse knowledge base. However, Hao et al. (2025) find that the use of AI reduces the overall volume of scientific topics studied by nearly 5 percent and reduces scientists’ engagement with other scientists by 22 per cent. They thus point to a paradox: increased individual productivity but a “contraction in collective science’s reach” (Hao et al. (2025), p. 1237).

In a related study, (Acemoglu et al., 2026) examines the long-term impact of AI tools and

develops a model that identifies a dynamic tension where AI tools provide context-specific recommendations that substitute for human learning effort, while society’s shared general knowledge serves as a complement that increases the marginal returns to human cognition. In their model, if AI tools become too accurate and human effort is highly elastic, the economy can spiral into a knowledge-collapse steady state where general knowledge vanishes. This occurs because individuals, relying on high-quality AI inputs into the scientific discovery process, reduce the learning effort that would otherwise generate the public signals necessary to sustain the community’s collective knowledge base. Acemoglu’s theory of knowledge collapse, whereby the substitution of AI-tools for human effort leads to the long-run depletion of collective information implies that if AI displaces the human effort required for abductive inference, the resulting asymmetry could leave scientists unable to comprehend even the fragments of the world models the AI models are designed to represent.

8 Concluding Comments

In this chapter, we explore the implications of AI for scientific discovery and, by extension, for economic growth. Our first main argument centers on the transformational potential of AI in science. By providing powerful prediction machines that facilitate search over vast combinatorial spaces of questions, ideas, designs, and tests, AI affects the knowledge production function. This impact extends beyond mere automation of routine tasks to enabling breakthroughs in domains like biology, materials science, and physics, where traditional human-led methods struggle with complexity. AI will compress timelines for scientific progress as illustrated by examples such as AlphaFold’s acceleration of protein folding predictions and AI-driven drug discovery. Even modest enhancements to the knowledge production function could yield sustained increases in long-term growth rates.

Our second key point emphasizes the augmentation of human scientists by AI, rather than replacement. Although AI excels in data-rich interpolative tasks, human judgment encompassing abductive inference, contextual sensitivity, ethical weighing, causal hypothesizing, and logical reasoning remains indispensable, particularly in data-sparse environments or stages requiring nuanced decision-making. From this perspective, AI complements human capabilities, enhancing produc-

tivity across the multi-stage scientific process. The persistence of human judgment in the loop suggests that policy and strategy may prioritize augmentation, fostering reorganizations in scientific workflows that leverage AI’s strengths while preserving the irreplaceable elements of human intelligence.

Finally, we conceptualize AI as a general purpose meta-technology (GPMT), a technology for inventing new technologies. Drawing on the notion of a jagged frontier, we highlight how AI’s applicability varies across scientific stages and domains, necessitating active exploration and complementary investments. Downstream, this includes building scientists’ expertise in AI use, as we describe in the context of our task-based augmentation framework; upstream, it involves advancing the science of AI itself, such as through scaling laws and algorithmic improvements. This GPMT lens connects AI’s role in science to broader meta-science discussions, positioning it as a catalyst for evolving the jagged frontier and overcoming bottlenecks to realize its full potential.

A recent contribution to this literature develops a task-based framework to evaluate AI’s potential effects on R&D processes. In “Artificial Intelligence in Research and Development,” Jones (2025) emphasizes the role of machines (including AI) alongside human inputs in producing ideas or innovations. The model posits R&D as a continuum of heterogeneous tasks, where progress depends on the share of tasks automatable by machines, the productivity advantage of machines over humans at those tasks, and the degree of complementarity or bottlenecks among tasks. Jones examines how advances in AI such as surging intelligence or expanded automation translate into rates of progress per R&D dollar, highlighting that strong bottlenecks can severely constrain gains even from superintelligent systems. The paper applies this to scenarios like economic growth or health improvements while outlining empirical strategies for parameter estimation and extensions to general equilibrium settings.

Jones’ approach differs from our chapter in its primary focus on automation and substitution, modeling AI as potentially displacing human labor across tasks to quantify marginal returns to intelligence in a production-function style. Where we prioritize augmentation of human judgment within a multi-stage scientific process stressing AI’s role in enhancing combinatorial search and decision-making without fully replacing contextual or ethical reasoning, Jones emphasizes closed-

form solutions for scenarios where AI achieves extraordinary capabilities, potentially leading to large multiples in productivity if bottlenecks are weak. This microeconomic lens on cost-minimizing allocation and task-level trade-offs contrasts with our broader meta-science perspective, which views AI as a GPMT evolving along a jagged frontier through upstream and downstream complementarities.

Nonetheless, Jones’ framework shares similarities with our analysis and offers complementary insights, as both underscore the persistence of human elements amid AI advances and the critical role of complementarities in limiting unbridled progress. His emphasis on measurable parameters like task shares and bottleneck strength aligns with our task-based augmentation model, providing tools to empirically ground assessments of AI’s jagged applicability across scientific domains. By connecting AI-driven R&D accelerations to broader outcomes like growth or health, Jones enriches our GPMT conceptualization, suggesting pathways to integrate quantitative bounds on productivity gains with our focus on expertise-building and frontier evolution.

Another recent contribution offers a more provocative vision of AI’s transformative potential in science. In “Science in the Age of Algorithms,” Mullainathan and Rambachan (2025) draw the familiar analogy to the historical adoption of electricity in manufacturing. Although initial applications merely substituted for steam power with modest productivity gains, the true revolution emerged decades later through a complete redesign of factory layouts (David, 1990). Applied to science, they argue that current AI uses, such as enhancing prediction, testing, or literature review are analogous substitutions that power existing machinery, whereas the deeper opportunity is a factory floor redesign that elevates “off-screen” scientific work and orients workflows around intelligent entities. In their framing, the goal is to move beyond next-token prediction toward world modeling (shifting from pattern-matching what’s likely to be said next to representing how the world works well enough to predict, intervene, and decide what to learn next), reorganizing the pipeline so that models learn and update structured representations in continuous interaction with data and with human scientists.

Mullainathan and Rambachan push this further by illuminating “missing” elements of the current scientific factory floor. They elevate abductive hypothesis generation, anomaly generation (operationalizing doubt by systematically producing counterexamples), and assessing completeness

(benchmarking how much predictable variation a theory captures relative to an algorithmic baseline) as first-class algorithmic targets. They also stress binding (linking formal models to real-world semantics) and the creation of new measurements rather than mere refinement of existing ones. In their perspective, AI evolves from augmentative tool to co-participant, with foundation-model-like world models that encode mechanisms across heterogeneous data streams, learn in use, and support mutual understanding between humans and models through interrogable representations rather than only compact, hand-crafted theories.

While resonant with our emphasis on AI’s capacity to navigate combinatorial complexity and augment judgment, Mullainathan and Rambachan’s perspective diverges in degree and in object of redesign. Where we stress augmentation within a preserved multi-stage workflow, they envision redesigning the architecture of science itself: shifting from static, human-interpretable theories toward learned, high-dimensional representations that blur prediction, exploration, and application. This introduces new trade-offs, most notably between predictive performance and human understanding, and implies fresh norms for validation and interpretability. Our GPMT lens highlights complements that move the jagged frontier; their blueprint suggests world-model-centric workflows in which algorithms become active participants. Read together, the two views invite short-term and long-term perspectives: in the short term, harvest stage-specific gains from AI and in the long term develop representation-centric architectures that embed ideation, anomaly discovery, and completeness assessment directly into the scientific process.

The realization of this re-organized “factory floor” centers on the evolution of research judgment, particularly the transition of abductive inference, the creative initial guess or hypothesis generation, from a purely human endeavor to an algorithmic one. As we argued in our comment on Mullainathan and Rambachan’s vision (Agrawal et al., 2026a), this shift moves the primary “world model” of science from the mind of the scientist to the AI itself. While this promises more complete models that can explain significantly more predictable variation than traditional human-derived theories, it risks a fundamental asymmetry where the scientist can only partially comprehend fragments of the AI’s overarching model. Ultimately, the move toward a self-driving scientific process raises critical questions about whether AI will remain a bicycle for the mind

(Agrawal et al., 2025) or lead to an outsourcing of understanding that fundamentally alters the rate, direction, and human interpretability of scientific progress.

Taken together, these contributions point to a rich agenda for future research on AI in science. Recent empirical work provides early guidance on the magnitudes involved and highlights the measurement challenges that such research must address.

The most immediate evidence of AI’s labor market effects comes from settings where high-frequency data enable near real-time tracking. Using administrative payroll data from ADP covering millions of U.S. workers, Brynjolfsson et al. (2025) document that early-career workers (ages 22–25) in AI-exposed occupations experienced 16% relative employment declines following the widespread adoption of generative AI, while employment for experienced workers remained stable. Critically, these effects were concentrated in occupations where AI *automates* rather than *augments* labor: entry-level employment declined in automating applications but showed growth in augmenting ones. Adjustments occurred primarily through employment rather than compensation, suggesting possible wage stickiness in the short run. These patterns, disproportionate effects on junior workers, the importance of the automation/augmentation distinction, and quantity rather than price adjustment, may presage similar dynamics in scientific labor markets, where graduate students and postdoctoral researchers could face particular displacement pressure in data-intensive, automatable tasks.²⁸

Measuring AI’s productivity effects within science presents distinct challenges, particularly the potentially long lags between adoption and observable outputs. Trišović et al. (2025) provide the first large-scale analysis of foundation model usage in scientific publications, finding that by 2024, approximately 0.9% of papers used foundation models (with an additional 0.4% customizing them), with adoption growing at near-exponential rates. Three findings are particularly relevant. First, scientists systematically adopt older and smaller models than those being built: in 2024,

²⁸Recently, Iscenko and Millet (2026) challenge this technological displacement narrative, arguing that the observed data patterns likely reflect the predictable consequences of a classic macroeconomic shock rather than AI-driven replacement. They find that the downturn in job postings for AI-exposed roles began in early 2022: six months before the public release of generative AI, aligning instead with the Federal Reserve’s most aggressive interest rate tightening cycle in forty years. This suggests that the decline in the 22–25 age group may be a mechanical result of broad hiring freezes, which create a “statistical illusion” of targeted displacement as this entry-level cohort fails to be restocked by new graduates.

the median model adopted was 26 times smaller than the median model released, suggesting that computational constraints or expertise barriers may limit scientists' ability to capture frontier AI capabilities. Second, adoption varies dramatically across fields, with Linguistics (34%), Computer Science (18%), and Engineering (4.6%) leading, while Biology and Chemistry show the fastest recent growth. Third, there is suggestive evidence that papers using larger models appear in higher-impact journals and accrue more citations, pointing to potential returns from expanding scientists' access to state-of-the-art tools.

These early findings suggest specific empirical priorities. First, longitudinal studies tracking scientific workflows before and after AI adoption could quantify productivity parameters at each stage of the research process, enabling direct estimation of the ω terms in our framework. Second, research distinguishing AI's effects on junior versus senior scientists would illuminate whether AI is primarily task-opening (democratizing access) or task-closing (concentrating advantages among those with complementary expertise). Third, studies examining adoption patterns across data-rich versus data-sparse domains would help map the jagged frontier empirically. Early indicators of productivity effects might include changes in publication rates, time-to-discovery metrics, the composition of research teams, and shifts in the topical diversity of research pursued. As AI continues to advance rapidly, integrating insights from these emerging empirical findings with the theoretical frameworks developed here will be essential to understanding not only productivity gains but also the broader transformations in how scientific knowledge is produced.

References

- Acemoglu, D. (2024, May). The simple macroeconomics of ai. NBER Working Paper 32487, National Bureau of Economic Research.
- Acemoglu, D. and S. Johnson (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. New York: PublicAffairs.
- Acemoglu, D., D. Kong, and A. Ozdaglar (2026). Ai, human cognition and knowledge collapse. *Working Paper*.
- Acemoglu, D. and P. Restrepo (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review* 108(6), 1488–1542.
- Acemoglu, D. and P. Restrepo (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives* 33(2), 3–30.
- Aghion, P., B. F. Jones, and C. I. Jones (2019). Artificial intelligence and economic growth. In A. K. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, pp. 237–282. Chicago: University of Chicago Press.
- Agrawal, A., J. Gans, and A. Goldfarb (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Cambridge, MA: Harvard Business Review Press.
- Agrawal, A., J. Gans, and A. Goldfarb (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives* 22(2), 31–50.
- Agrawal, A., J. Gans, and A. Goldfarb (2025). The economics of bicycles for the mind. *Working Paper 34034*.
- Agrawal, A., J. McHale, and A. Oettl (2019). Finding needles in haystacks: Artificial intelligence and recombinant growth. In A. K. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, pp. 149–174. Chicago: University of Chicago Press.

- Agrawal, A., J. McHale, and A. Oettl (2024). Artificial intelligence and scientific discovery: A model of prioritized search. *Research Policy* 53(5), 104989.
- Agrawal, A., J. McHale, and A. Oettl (2026a). Abduction, judgment, and algorithms: Redesigning the factory floor of science? discussion of mullainathan and rambachan. In A. K. Agrawal, E. Brynjolfsson, and A. Korinek (Eds.), *The Economics of Transformative Artificial Intelligence*. Chicago: University of Chicago Press.
- Agrawal, A., J. McHale, and A. Oettl (2026b). Enhancing worker productivity without automating tasks: A different approach to AI and the task-based model. *NBER Working Paper 34781*.
- Altman, S. (2025). Three observations. <https://blog.samaltman.com/three-observations>. Essay.
- Amodei, D. (2024). Machines of loving grace. <https://www.darioamodei.com/essay/machines-of-loving-grace>. Essay.
- Bernard, N., Y. Sagawa, N. Bier, T. Lihoreau, L. Pazart, and T. Tannou (2025). Using artificial intelligence for systematic review: the example of elicitor. *BMC Medical Research Methodology* 25(75), 1–6.
- Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020). Are ideas getting harder to find? *American Economic Review* 110(4), 1104–1144.
- Bragg, J., M. D’Arcy, and B. D. S. F. D. H. J. D. H. P. J. V. K. B. P. M. A. N. S. R. K. R. A. S. H. S. A. T. R. V. G. W. C. A. S. C. J. D. D. E. R. E. M. H. R. H. R. K. M. L. J. L. R. L.-A. C. N. S. R. A. T. B. V. P. C. D. D. Y. G. A. S. D. S. W. Nishant Balepur, Dan Bareket (2025). Astabench: Rigorous benchmarking of ai agents with a scientific research suite. *arXiv:2510.21652*.
- Bryan, K. (2025). Meta’s ai climate tool raised false hope of co removal, scientists say. *Financial Times*.
- Brynjolfsson, E. (2022). The turing trap: The promise & peril of human-like artificial intelligence. *Dædalus* 151(2), 28–31.

- Brynjolfsson, E., B. Chandar, and R. Chen (2025). Canaries in the coal mine? six facts about the recent employment effects of artificial intelligence. Technical report, Stanford Digital Economy Lab Working Paper.
- Caplin, A., D. J. Deming, S. Li, D. J. Martin, P. Marx, B. Weidmann, and K. J. Ye (2024, October). The abc's of who benefits from working with ai: Ability, beliefs, and calibration. Working Paper 33021, National Bureau of Economic Research.
- Carleo, G., I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová (2019). Machine learning and the physical sciences. *Reviews of Modern Physics* 91(4), 045002.
- Cavalli, G. (2024). How scientific organizations adapt to advances in artificial intelligence: The impact of alphafold1. In *Academy of Management Proceedings*, Volume 2024, pp. 21333. Academy of Management Valhalla, NY 10595.
- Cavga, D. (2026). 5 ai tools every life scientist should know.
- Cheng, Y., Y. Gong, Y. Liu, B. Song, and Q. Zou (2021). Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in bioinformatics* 22(6), bbab344.
- Chollet, F. (2019). On the measure of intelligence. arXiv:1911.01547.
- Cockburn, I. M., R. Henderson, and S. Stern (2019). The impact of artificial intelligence on innovation: An exploratory analysis. In A. K. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, pp. 115–148. Chicago: University of Chicago Press.
- David, P. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *The American Economic Review*.
- Dell'Acqua, F., C. Ayoubi, H. Lifshitz, R. Sadun, E. Mollick, L. Mollick, Y. Han, J. Goldman, H. Nair, S. Taub, and K. Lakhani (2025, April). The cybernetic teammate: A field experiment

- on generative ai reshaping teamwork and expertise. Working Paper 33641, National Bureau of Economic Research.
- Dell’Acqua, F., E. I. McFowland, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Kraymer, F. Candelon, and K. R. Lakhani (2023, September). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, Technology & Operations Management Unit.
- Feynman, R. (1992). *The Character of Physical Law*. London: Penguin Books. Originally published 1965.
- Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.
- Govorkova, E., E. Puljak, T. Aarrestad, T. James, V. Loncar, M. Pierini, A. A. Pol, N. Ghielmetti, M. Graczyk, S. Summers, J. Ngadiuba, T. Q. Nguyen, J. Duarte, and Z. Wu (2022). Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 mhz at the large hadron collider. *Nature Machine Intelligence* 4, 154–161.
- Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.
- Hadamard, J. (1945). *The Mathematician’s Mind: The Psychology of Invention in the Mathematical Field*. Princeton, NJ: Princeton University Press. Reissued in the Princeton Science Library series.
- Hanson, N. R. (1958). *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.
- Hao, Q., F. Xu, Y. Li, and J. Evans (2025). Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Science* 18.
- Hao, Q., F. Xu, Y. Li, and J. Evans (2026). Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature*, 1–7.

- Hassabis, D. (2025). The end of disease. <https://www.science.org/content/blog-post/end-disease>. Interview.
- Hayes, T., N. J. S. D. O. Z. L. R. V. V. Q. T. J. D. M. W. R. B. I. S. J. G. A. D. R. S. M. N. T. Y. A. K. C. M. C. K. L. J. B. M. N. P. D. H. T. S. S. C. Roshan Rao, Halil Akin, and A. Rives (2025). Simulating 500 million years of evolution with a language model. *Science* 387, 850–858.
- Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Horton, J. J., A. Filippas, and B. S. Manning (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper 31122*.
- Iscenko, Z. and F. C. Millet (2026). Looking for the ladder: Is ai impacting entry-level jobs? *Working Paper - The Economic Innovation Group*.
- Jones, B. (2025). Artificial intelligence in research and development. *Working Paper*.
- Jones, B. F. (2009). The burden of knowledge and the ‘Death of the Renaissance Man’: Is innovation getting harder? *The Review of Economic Studies* 76(1), 283–317.
- Jones, C. I. (2024). The ai dilemma: Growth versus existential risk. *American Economic Review: Insights* 6(4), 575–590.
- Jones, C. I. (2026, January). A.I. and our economic future. NBER Working Paper 34779, National Bureau of Economic Research. In preparation for the *Journal of Economic Perspectives*.
- Jones, C. I. and C. Tonetti (2026, January). Past automation and future A.I.: How weak links tame the growth explosion. Unpublished manuscript, Stanford GSB.
- Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. (2021). Highly accurate protein structure prediction with alphafold. *nature* 596(7873), 583–589.

- Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books.
- Kamya, P., I. V. Ozerov, F. W. Pun, K. Tretina, T. Fokina, S. Chen, V. Naumov, X. Long, S. Lin, M. Korzinkin, D. Polykovskiy, A. Aliper, F. Ren, and A. Zhavoronkov (2024). Pandaomics: An ai-driven platform for therapeutic target and biomarker discovery. *Journal Chem Inf Model* 64(10), 3961–3969.
- Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Kim, S. (2025). Navigating the rugged data landscape: The impact of data-extrapolation technologies on knowledge production. Technical report, Columbia Business School Working Paper.
- Krenn, M., R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, et al. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics* 4(12), 761–769.
- Kusumegi, K., X. Yang, P. Ginsparg, M. de Vaan, T. Stuart, and Y. Yin (2026). Scientific production in the era of large language models: With the production process rapidly evolving, science policy must consider how institutions could evolve. *Science online January*.
- Langley, P., H. A. Simon, G. L. Bradshaw, and J. M. Zytkow (1987). *Scientific Discovery: Computational Explorations of the Creative Process*. Cambridge, MA: MIT Press.
- Liu, G., D. B. Catacutan, K. Rathod, K. Swanson, W. Jin, J. C. Mohammed, A. Chiappino-Pepe, S. A. Syed, M. Fragis, K. Rachwalski, et al. (2023). Deep learning-guided discovery of an antibiotic targeting acinetobacter baumannii. *Nature Chemical Biology* 19(11), 1342–1350.
- Ludwig, J. and S. Mullainathan (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics* 139(2), 751–827.
- Merchant, A., S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk (2023). Scaling deep learning for materials discovery. *Nature* 624, 80–85.

- Mitchener, L., A. Yiu, B. Chang, M. Bourdenx, T. Nadolski, A. Sulovari, E. C. Landsness, D. L. Barabasi, S. Narayanan, N. Evans, S. Reddy, M. Foiani, A. Kamal, L. P. Shriver, F. Cao, A. T. Wassie, J. M. Laurent, E. Melville-Gree, M. Caldas, A. Bou, K. F. Roberts, S. Zagorac, T. C. Orr, M. E. Orr, K. J. Zvezdaryk, A. E. Ghareeb, L. McCoy, B. Gomes, E. A. Ashley, K. E. Duff, T. Buonassisi, T. Rainforth, R. J. Bateman, M. Skarlinski, S. G. Rodrigues, M. M. Hinks, and A. D. White (2025). Kosmos: An ai scientist for autonomous discovery. *arXiv:2511.02824*.
- Mollick, E. (2024). *Co-Intelligence: Living and Working with AI*. New York: Portfolio.
- Mullainathan, S. and A. Rambachan (2024, May). From predictive algorithms to automatic generation of anomalies. Working Paper 32422, National Bureau of Economic Research.
- Mullainathan, S. and A. Rambachan (2025). Science in the age of algorithms. *Working Paper*.
- Nordhaus, W. D. (2015, August). Are we approaching an economic singularity? information technology and the future of economic growth. NBER Working Paper 21547, National Bureau of Economic Research.
- Oak Ridge National Laboratory (2026). Early breakthrough for ai-enabled biological discovery.
- Peirce, C. S. (1994). *The Collected Papers of Charles Sanders Peirce*. Electronic edition; originally published by Charles Hartshorne and Paul Weiss, edited by Arthur W. Burks, Harvard University Press, 1931–1935.
- Peirce, C. S. (1998). *The Essential Peirce: Selected Philosophical Writings, Volume 2 (1893–1913)*. Bloomington and Indianapolis: Indiana University Press.
- Polanyi, M. (2009). *The Tacit Dimension*. Chicago and London: The University of Chicago Press. Originally published 1966.
- Pollice, R. et al. (2021). Data-driven strategies for accelerated materials design. *Accounts of Chemical Research* 54(4), 849–860.
- Rauch, J. (2021). *The Constitution of Knowledge: A Defense of Truth*. Washington D.C.: Brookings Institution Press.

- Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy* 98(5, Part 2), S71–S102.
- Romer, P. M. (1992). Two strategies for economic development: Using ideas and producing ideas. In *Proceedings of the World Bank Annual Conference on Development Economics 1992*, pp. 63–91. Washington, DC: World Bank.
- Royal Swedish Academy of Science (2025). Nobel prize in chemistry.
- Sriram, A., L. M. B. A. D. Z. U. M. U. A. J. M. Sihoon Choi, Xiaohan Yu, and D. S. Sholl (2024). The open dac 2023 dataset and challenges for sorbent discovery in direct air capture. *ACS Central Science* 10(5), 923–941.
- Theodoris, C. V., L. Xiao, and Z. R. A. S. M. C. H. H. M. E. M. B. Z. Z. X. S. L. . P. T. E. Anant Chopra, Mark D. Chaffin (2023). Transfer learning enables predictions in network biology. *Nature* 618, 616–624.
- Trišović, A., A. Fogelson, J. Sivaloganathan, and N. Thompson (2025). The rapid growth of ai foundation model usage in science.
- Udrescu, S.-M. and M. Tegmark (2020). Ai feynman: A physics-inspired method for symbolic regression. *Science Advances* 6(16), eaay2631.
- Weidmann, B., Y. Xu, and D. J. Deming (2025, April). Measuring human leadership skills with ai agents. Working Paper 33662, National Bureau of Economic Research.
- Weinberg, S. (2015). *To Explain the World: The Discovery of Modern Science*. New York: Harper-Collins.
- Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics* 113(2), 331–360.
- Whewell, W. (1989). *Theory of scientific method*. Hackett Publishing.

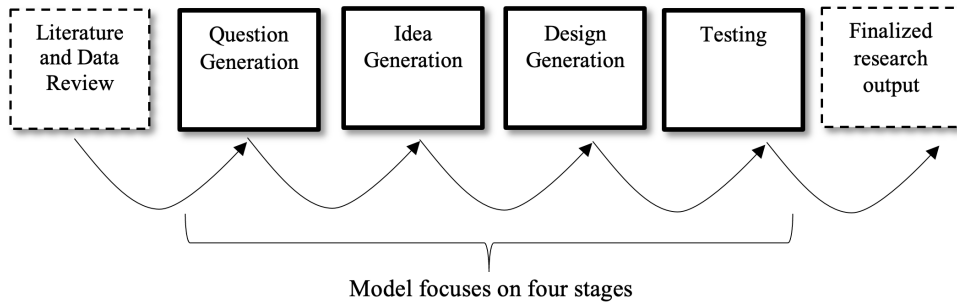


Figure 1: Stylized stages of scientific research