AI in Science*

Ajay Agrawal
University of Toronto and NBER

John McHale
University of Galway

Alexander Oettl
Georgia Tech and NBER

September 25, 2025

DRAFT

Abstract

Much of science involves search over massive combinatorial spaces of possible questions, ideas, designs and tests. AI has the potential to transform science by providing powerful tools to aid these searches. However, human judgment is likely to remain integral to scientific productivity and we argue that the nearer-term promise of AI lies in how it can augment scientists' judgment. We explore the potential impacts of AI as a tool for science that augments judgment across different stages of the scientific process and across different scientific domains. We then consider AI specifically as an example of a general purpose meta-technology (GPMT) – i.e., a technology for discovering new technologies that changes the economy's knowledge production function. We examine the current "jagged frontier" of this technology in science and consider how it might evolve. In particular, we emphasize the importance of scientists' expertise in the use of AI as we look downstream to the evolving uses of AI in science.

*This book chapter was prepared for the 2025 NBER Economics of Science workshop. We thank the organizers, Megan MacGarvie and Reinhilde Veugelers, for the opportunity to contribute this chapter. Contact: ajay@agrawal.ca, john.mchale@nuigalway.ie, and alex.oettl@scheller.gatech.edu

1 Introduction

Artificial intelligence (AI) is already changing how science is done. And, with the rapid rate of improvement in the underlying technology, it has the potential to significantly alter the scientific research process – and scientific productivity – in many, if not all, domains over the next decade. For economists, the transformational effect of AI in science is of particular interest insofar as it changes the economy's knowledge production function, with potentially significant implications for economic growth and the workings of scientific labor markets.

AlphaFold (Jumper et al., 2021) remains the most celebrated example of the use of AI in science, with its invention already resulting in a share of the 2024 Nobel Prize in Chemistry for its Google DeepMind creators. The technology has led to a substantial advance in the capability to predict how proteins fold into their final tertiary shapes based on their amino acid sequences. As AlphaFold demonstrates, AI is proving especially powerful as a tool to aid in search over potentially vast combinatorial spaces – in this case the massive number ways a protein can fold.

Science-based drug discovery companies such as Isomorphic Labs and Insitro are using AI to speed up the drug discovery process. In one striking development, the AI-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. Together with other technological breakthroughs such as gene editing, AI holds the promise to revolutionize medicine in terms of personalized treatments and the tackling of rare diseases.

The use of AI extends far beyond biology and medicine to affect the research process in practically all domains of science. In materials science, AI is being used for 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 are also making possible the use of "inverse design," where the model generates candidates based on specified desired properties (Cheng et al., 2021). In physics, AI is helping physicists find patterns and anomalies in the often vast volumes of data produced by experiments (Carleo et al.,

2019). Tools such as symbolic regression are being used to discover (or rediscover) interpretable equations capturing the relationships between variables in vast datasets (see, e.g., Udrescu and Tegmark (2020)). In mathematics, tools such as AlphaProof show the capability to solve complex mathematical problems, and have achieved success in international mathematics competitions. A notable feature of AlphaProof is its ability to generate formally verifiable proofs, where the proofs are checked by proof assistant software.

At a more systemic level, AI tools are also helping scientists manage the "burden of knowledge" that results from the massive daily flow of new research (Jones, 2009). Innovative utilities such as Notebook LM – which generates accessible podcasts from the texts of papers – and Gemini Deep Research – which generates in-depth research reviews based on user-defined plans – are helping scientists absorb and prioritize what would otherwise be an overwhelming flow of new literature. 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 can then produce a summary of relevant literature, propose testable hypotheses and suggest experimental designs.

With this rapid pace of development, there is good reason to believe that artificial intelligence (AI) will be the most important force affecting the productivity and organization of science over the next decade. As with other aspects of the economy and society, there is a wide spectrum of opinion on how consequential AI will actually be for science. In a recent essay titled "Machines of Loving Grace," Dario Amodei, the CEO of a leading AI company, Anthropic, sets out a highly optimistic scenario[1]:

> [M]y 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 (Amodei, 2024).[2]

Economists have usually tended to be more circumspect. Daron Acemoglu, for example, is more pessimistic about the transformative effects of AI on productivity. Following an exploration

[1] Although clearly towards the optimistic end of the spectrum, Amodei does emphasize the bottlenecks that could attenuate the impacts of AI on science as well as the risks associated with rapid AI development.

[2] Available at: `https://www.darioamodei.com/essay/machines-of-loving-grace`.

of the macroeconomic implications of applying AI over the next decade, he predicts that effects will be "non-trivial but modest" (an upper bound of a cumulative 0.66% increase in total factor productivity (TFP) over 10 years). However, given our focus on AI and science, Acemoglu's macroeconomic evaluation is perhaps most interesting for what it leaves out:

> I . . . do not discuss how AI can have revolutionary effects by changing the process of science (a possibility illustrated by neural network-enabled advances in protein folding and new crystal structures discovered by the Google subsidiary DeepMind), 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, 2024, p. 5).

In this chapter we attempt to steer a middle way between the more optimistic and pessimistic scenarios for the effects of AI on science. A core idea is that AI is, contrary to Acemoglu, 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. However, we also stress the likely continuing importance of human judgment in the scientific process. Our guiding approach, then, 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, 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., reorganizations of scientific workflows) complementary investments.

We thus organize our economics-focused review of AI in science around three main ideas. The first idea is the transformational effects of AI in science and thus the economy. 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.[3] We examine this idea of AI-aided search – and its effects on scientific productivity – across the different stages of the scientific process.

[3]The 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,

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.[4]

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*[5] is likely to remain indispensable in at least some stages of the scientific process that involve search over complex combinatorial search spaces. Put differently, our base assumption is that, although roles of human scientists will change, it is

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 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.

[4]While 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.

[5]We view 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 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.

Relatedly, 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.

likely that given the ongoing importance of human judgment there will remain role for the "human in the loop" of scientific inference and decision making. This shifts attention to the use of AI as a *tool* that augments the capabilities – and thus the productivity – of judgment-exercising human scientists.[6]

The debate over automation versus augmentation has both normative and positive dimensions. On the normative side, there is the question of how public policy (including regulation) and technology company strategy could shift the balance towards the development of AI technology that augments workers' capabilities as opposed to directly 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 argue for the 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 and strategy.[7]

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 applications, innovational complementarities, and dynamism in the development of the GPT itself – a GPMT is a technology for the development of new technology and science (Cockburn et al., 2019; Agrawal et al., 2019). Using the idea of AI as GPMT, we also 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

[6]We 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.

[7]One 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.

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 adapt Acemoglu and Restrepo's task-based model (Acemoglu and Restrepo, 2018, 2019) to develop a simple task-based augmentation model, where the productivity effects of AI on science depends 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 the 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 then explore the usefulness of thinking about AI as a GPMT and the importance of downstream and upstream complementary investments in altering the jagged frontier. Looking downstream, Section 5 develops the simple task-based model of AI augmentation and explores in particular the importance of developing scientists' expertise in the use of AI. We turn our attention upstream to the development of AI technology itself in Section 6, where we consider various driving forces and bottlenecks affecting the science of AI. We conclude in Section 7 with a summary of 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 of the ways in which AI could affect the different stages that make up the scientific process. 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 – say, refining the research question if there is a failure to generate an idea that could answer the 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.[8]

[8] AI 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.

- 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 promise to revolutionize search in one or more of 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 the 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:

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

where $\dot{A}$ is the output of 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).[9] Looking across

[9]By identifying constant scientific productivity with the achievement of constant exponential growth for a given

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 potentially be impacted by AI – the framework gives added structure to allow the consideration of different ways in which AI might impact on science.

To make things concrete, we initially use the example of scientific team that is seeking to find a small molecule drug to bind with a malfunctioning protein for some therapeutic effect. Idea generation is thought of in terms of 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 here is then the physical testing of the binding efficacy of an identified small molecule (though we can also allow "success" to be multi-dimensional, also, 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 helping with the identification of 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 a tool to enhance productivity and the remaining importance of human judgment.

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.

2.1 Stage 1: Productivity in question generation

The first stage of our idealized scientific process is the generation of a question (or problem). 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).[10] For our purposes, we interpret "surprise" broadly to include observations of 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 judge's decision on pre-trial detention. But it is unclear which features of the appearance matter.

[10] A 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, 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).

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 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 at 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 the self-growth 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 have emphasized the importance of ideas in the generation of hypotheses.[11] An influential example is the English philosopher

[11] Although Francis Bacon is often associated with a pure data-driven scientific method, as Ian Hacking points

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:[12]

> 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*[13]) 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).

---

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)).

[12] "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).

[13] Although the term guessing may seem 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).

> 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).[14]

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 that 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 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 by 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 $\phi \longrightarrow 0$ (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:

[14] It 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).

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

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 $\widetilde{\beta}$, where $\widetilde{\beta}$ 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 $\widetilde{\beta}$ 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:

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

where $\beta = (\ln 2)\widetilde{\beta}$. 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 though 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.[15] And,

[15]The 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).

second judgment could be required to identify if 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.[16] 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. 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 in effectively, 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.[17] 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:

[16]It 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 leading 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 in 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 System1 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).

[17]In 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.

> 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 Polayni's idea of tacit knowledge provides a useful lens with which to view this question. Extending Gestalt psychology, Polayni 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 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 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)).[18] These 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 are not qualified to take a firm 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

[18] The 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. Various AI-based approaches have been put forward, generally involving giving the AI the ability to search a solution space at inference time. At the time of writing in May 2025, AI models that are able to do inference time reasoning (such as GPT o3) are beginning to perform much better on the challenge, though it remains an open question as to whether their capability would extend beyond the highly specific domain of puzzles on a two-dimensional grid. 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.

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 in comprises successful prior instantiations of real world combinations on related problems, making it amenable to an AI-based solution.[19]

In what we call the Hassabis hypothesis (see Agrawal et al. (2024)), Demis Hassabis, the head of Google DeepMind, has 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 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 determinative. 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

[19]It 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.

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 given a drawn 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.[20]

[20]Instead 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. They 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, they 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.

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 the physical understandings of scientists. Alternatively, given that there be multiple objectives that the design needs to satisfy – say the efficacy and the safety of a drug – and the 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.[21] In other words, the model might provide a "Pareto frontier" given the competing objectives, but human judgement is still required to ranks 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.

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 combination – say whether a particular small molecule effectively binds

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

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 makes 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.[22] Putting the stages together, the expected number of successfully implemented designs per scientist ($\dot{A}/S$) in a given time period is then:

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

[22]It 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.

Therefore, the expected number of successfully implemented designs (which we equate to successfully implemented new ideas) 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 is 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:

$$\frac{\frac{\dot{A}}{A}}{S} = \omega = \alpha\beta\gamma\delta. \tag{9}$$

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, 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 the particular descriptions 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; (ii) generation of a potential explanation

for that phenomenon (e.g., in the form of a formal theory); (iii) designing a controlled (or identifying a natural) experiment for testing observable implications of theory; and (iv) actual testing of an observable implication.

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: biology, computer science, economics and mathematics.

3 The Use of AI in Science: Examples from Selected Domains

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 demand extrapolative reasoning. In this section, we apply the model to illustrative domains including biology and medicine, materials science, physics, mathematics, and

social sciences to elucidate these variations, drawing on recent examples of AI integration.

Biology and medicine represent domains 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 like large language models can synthesize literature to propose novel hypotheses, such as identifying gene interactions for rare diseases. However, human judgment remains critical for prioritizing questions with ethical implications, such as those involving patient privacy or equity 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. Testing (δ) benefits from AI in analyzing high-throughput screening data, as in Isomorphic Labs' antibiotic discovery against Acinetobacter baumannii. Overall, in these data-rich fields, AI could substantially boost ω by alleviating bottlenecks in γ and δ, potentially doubling productivity if complementary investments in data curation and model interpretability are made. Yet, regulatory hurdles and the need for causal validation ensure human oversight in the loop.

In contrast, materials science illustrates a domain where AI excels in property prediction and inverse design but faces challenges in sparse experimental data. 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). 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. Physics, similarly, leverages AI for pattern detection in massive datasets from experiments like those at the Large Hadron Collider (Carleo et al., 2019), enhancing α through anomaly identification. 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. Mathematics and the social sciences highlight domains where AI's jagged frontier is more pronounced, with greater

reliance on human judgment across stages. 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. Social sciences, such as economics, benefit 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.

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 has highlighted 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 invest-

ments 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 on 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 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 develop 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 result 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 architectural/algorithmic improvements.

5 AI skills and scientific productivity: A task-based model of AI augmentation

To this point in the chapter we have treated AI as a tool that is available to scientists to help them accomplish various tasks in the scientific workflow. This begs the question as to whether scientists have the complementary expertise to actually use AI. To focus on the expertise question, we develop in this section a simple task-based model of the scientific production function that draws on the task-based models of Acemoglu and Restrepo (2018, 2019) and Aghion et al. (2019). In contrast to the automation focus of these models – where AI automates tasks previously undertaken by human workers – our model is explicitly focused on how AI *augments* human scientists. Critically, however, the scientist must possess the requisite skills in the use of AI – what we term AI expertise – to take advantage of developments in AI technology. AI is thus a *tool* used by scientists with the requisite expertise. Using the terminology of Mollick (2024), AI in this world acts as a "co-intelligence" with AI-expert scientists.[23]

The task-based approach offers a powerful and flexible tool for modelling both the labor market and productivity effects of technological change. Important contributions include Zeira (1998); Acemoglu and Autor (2011); Autor et al. (2003); Acemoglu and Restrepo (2018); Autor and Thompson

[23]Mollick (2024) makes a distinction between "Centaurs" – which would have AI and scientists specializing in different elements of a task – and "Cyborgs" – where the AI and scientists fully blend machine and person.

(2025). Aghion et al. (2019) provide an extension in the context of modelling the economy's knowledge production function. The central idea of the paradigmatic task model is that the allocation of workers to tasks is determined in general equilibrium by comparative advantage, with workers and machines having varying relative productivity across a continuum of tasks.

We adapt the *constrained* comparative equilibrium model of Acemoglu and Restrepo (2018, 2019) to model the effects of AI on the science labor market. In the central case studied by Acemoglu and Restrepo, the state of technology provides a constraint on which tasks can be *automated* (i.e., workers replaced by machines) and firm would choose to automate the marginal task if it were technologically feasible. For simplicity, we abstract from machines (and thus the substitution of machines for scientists in a process of automation) to focus on *augmentation*. In our model of augmentation, scientific output is produced by scientists with just "ordinary" expertise and AI-expert scientists (who combine both ordinary expertise and expertise in the use of AI).

Technological change is captured as an exogenous increase in the range of tasks that can be undertaken by AI-expert scientists, where we assume that the starting state of technology provides a binding constraint on which tasks are done using AI. Put differently, as AI technology improves on the extensive margin, more tasks can be done using AI, but only scientists that have the complementary AI expertise are able to use the AI. Technological improvements in AI then augments scientists, but only those scientists who possess the necessary AI expertise, putting the focus on the AI expertise of the scientific workforce.

Specifically, we assume that scientific output, $\dot{A}$, is a function of the number of ordinary scientists, S_O, and AI-expert scientists S_{AI}, where the total number of scientists is $S = S_O + S_{AI}$. We assume that scientific output in the economy is produced by a representative team of scientists using a multi-task (or multi-stage) knowledge production function and that both types of scientific labor are supplied inelasticity with wages determined in competitive labor markets. In the presence of improvements in the AI technology, our focus is how the share of AI-expert scientists ($\frac{S_{AI}}{S}$) in the economy affects the growth of scientific productivity.

The core of the task-based model is the task-based production function, which we assume has the unit-elastic form:

$$\dot{A} = A\left(\exp\left[\int_0^1 \ln z(\tau) d\tau\right]\right), \tag{10}$$

where $z(\tau)$ is the output of task τ. As with the Romer knowledge production function, there is a linear relationship between the existing knowledge stock, A, and the increase in the knowledge stock per unit of time, $\dot{A}$.[24] This Cobb-Douglas functional form allows for substitution between tasks in the production of new knowledge, but restricts the elasticity of substitution to be unity. As a further simplification, we assume there is an unchanging unit measure of tasks, which are indexed on the continuum from 0 to 1 ($\tau \in [0, 1]$).[25] Each task can be done by ordinary scientists or – if technologically feasible – by *either* ordinary scientists or AI-expert scientists. All tasks are necessary to produce output, but the scientific team has flexibility in the level of activity in each task.

The productivity of an ordinary worker in any given task τ is $\theta_O(\tau)$; the productivity of an AI-expert scientist in an AI-supported task is $\theta_{AI}(\tau)$. A key assumption of the model is that $\theta_O(\tau)/\theta_{AI}(\tau)$ is increasing in τ, and so ordinary labour has a *comparative advantage* in higher indexed tasks.

We assume that the share of tasks that it is technologically feasible to perform using AI-expert scientists is $\overline{\tau}$, so that a change in the share of tasks that can be done with AI expertise is the source of AI-related technological change in the model. The current state of technology therefore provides a constraint on the tasks that can be performed by AI-expert scientists, which we assume is binding. In other words, at the current allocation of tasks to ordinary and AI-expert scientists, the scientific team would use AI-expert scientists for the next task in the index if it were technologically feasible. The technological constraint thus prevents the optimal allocation of workers to tasks based on comparative advantage leading to the constrained comparative advantage equilibrium.

We capture AI-related technological progress as an exogenous increase in $\overline{\tau}$ – i.e., the share of tasks that can be done using AI expertise – and $\frac{\partial \overline{\tau}}{\partial t} \geq 0$, where t is time. Our focus is therefore

[24] Alternatively, we can think of increases in the existing stock of knowledge leading to an identical proportional improvements in the productivity of each task.

[25] Acemoglu and Restrepo (2019) also allow for new tasks and consider the potential for new tasks to offset the displacement effects of task-biased technological change on wages. We ignore the possibility of new task for simplicity.

on improvements in AI at the extensive margin – i.e., the range of tasks that can be done with AI-expert scientists – as opposed to the intensive margin, which could be captured by increases in $\theta_{AI}(\tau)$ for tasks that it is already feasible for AI-expert scientists to perform.

At the task level, the task-specific outputs are given by:

$$z(\tau) = \begin{cases} \theta_O(\tau)S_O(\tau) + \theta_{AI}(\tau)S_{AI}(\tau) & \text{if } \tau \in [0, \overline{\tau}] \\ \theta_O(\tau)S_O(\tau) & \text{if } \tau \in (\overline{\tau}, 1] \end{cases}, \tag{11}$$

which allows for the possibility that tasks up to $\overline{\tau}$ can be done by either AI expert or ordinary scientists, but tasks beyond $\overline{\tau}$ are limited to ordinary scientists. However, in the general equilibrium of the model there will be complete specialisation of tasks to one of the scientist types. Similarly to Acemoglu and Restrepo (2019), we assume:

$$\frac{W_O}{W_{AI}} > \frac{\theta_O(\overline{\tau})}{\theta_{AI}(\overline{\tau})} \tag{12}$$

This implies that the marginal task, $\overline{\tau}$, the cost per scientist is higher for ordinary scientists than for AI-expert scientists. This assumption ensures that at the current share of tasks for which it is technologically feasible to use AI, the scientific team would utilize AI expertise if an additional task became technologically feasible. We can therefore think of the state of the AI technology technology as a constraint on the otherwise optimal allocation of scientist types to tasks.

It is useful for what follows to take logs and rearrange (12) to obtain the condition:

$$\ln W_O - \ln \theta_O(\overline{\tau}) - (\ln W_{AI} - \ln \theta_{AI}(\overline{\tau})) > 0. \tag{13}$$

In Appendix 1, we show that the wages for the two types of scientists (AI expert and ordinary) in the (technology-constrained) equilibrium of the economy are given by:

$$W_{AI} = \overline{\tau}\frac{\dot{A}}{S_{AI}}; \text{ and} \tag{14}$$

$$W_O = (1 - \overline{\tau})\frac{\dot{A}}{S_O}. \tag{15}$$

This implies that share of income received by AI-expert scientists is $\frac{W_{AI}S_{AI}}{\dot{A}} = \overline{\tau}$, where the price of a unit of scientific output in normalized to one without loss of generality.

Using the price index for scientific output, we further show in Appendix 1 that aggregate scientific productivity takes the endogenous form:

$$\frac{\dot{A}}{S} = B(\overline{\tau})A\left(\frac{S_{AI}}{\overline{\tau}S}\right)^{\overline{\tau}}\left(\frac{S_O}{(1-\overline{\tau})S}\right)^{1-\overline{\tau}}, \tag{16}$$

where

$$B(\overline{\tau}) = \exp\left(\int_0^{\overline{\tau}} \ln\theta_{AI}(\tau)d\tau + \int_{\overline{\tau}}^1 \ln\theta_O(\tau)d\tau\right). \tag{17}$$

To see that the scientific productivity function in the same general form as our baseline model of Section 2, it is useful to write (16) as:

$$\frac{\dot{A}/A}{S} = \omega\left(\overline{\tau}, \frac{S_{AI}}{S}\right), \tag{18}$$

where

$$\omega\left(\overline{\tau}, \frac{S_{AI}}{S}\right) = B(\overline{\tau})\left(\frac{S_{AI}}{\overline{\tau}S}\right)^{\overline{\tau}}\left(\frac{1}{\overline{\tau}}\left(1-\frac{S_{AI}}{S}\right)\right)^{1-\overline{\tau}}. \tag{19}$$

Thus, our measure of (total factor) productivity in science depends now on both the set of tasks for which it is feasible to use AI-expert scientists and the share of AI-expert scientists in the scientific workforce. Whereas the focus in our baseline model is on how AI affects scientific productivity at different stages of the scientific process, the focus of this task-based model is on how the availability of AI-expert scientists affects scientific productivity and, in particular, how that availability affects the productivity gain from an exogenous improvement in AI.

To identify this effect, it is useful to take logs of (16) and write the log of scientific productivity as the sum of two components:

$$\ln\left(\frac{\dot{A}}{S}\right) = \left[\int_0^{\overline{\tau}} \ln\theta_{AI}(\tau)d\tau + \overline{\tau}\ln\left(\frac{S_{AI}}{\overline{\tau}S}\right) + \overline{\tau}\ln A\right] + \left[\int_{\overline{\tau}}^{1} \ln\theta_O(\tau)d\tau + (1-\overline{\tau})\ln\left(\frac{S_O}{(1-\overline{\tau})S}\right) + (1-\overline{\tau})\ln A\right]. \quad (20)$$

Using the wage equations (14) and (15), the impact on log productivity of a small improvement in AI technology (i.e., $\partial\overline{\tau} > 0$) is given by:

$$\frac{\partial \ln\left(\frac{\dot{A}}{S}\right)}{\partial\overline{\tau}} = \ln W_O - \ln\theta_O(\overline{\tau}) - (\ln W_{AI} - \ln\theta_{AI}(\overline{\tau})) > 0, \quad (21)$$

where the inequality follows from our assumption of the technological constraint implied in (13). That is, an improvement in technology, as measured by an increase in the share of tasks that can be done with AI-expert scientists, leads to an increase in scientific productivity in the constrained equilibrium.

As noted, our primary interest is in how the share of AI-expert scientists affects the size of the gain from an improvement in technology. From (14) and (15), it follows that an increase in the share of AI-expert scientists in the total scientific workforce will lower the relative wage of AI-expert scientists. And from (21) it follows that the impact of an increase in $\overline{\tau}$ on scientific productivity will be larger the larger is the share of AI-expert scientists.

To better grasp the intuition for this result, it is useful to think of the optimal (i.e., scientific-productivity-maximising) allocation of scarce resources to tasks as being based on *comparative* advantage in the absence of a technological constraint on which tasks can be done using AI. But the state of technology (as measured by the share of tasks that can actually be done with AI) is such that this optimal allocation cannot be reached. At the margin, a marginal relaxation in the constraint, in the sense of a marginal increase in the share of tasks that can be done with AI-expert scientists, allows the team to come closer to the optimal allocation of resources to tasks based on unconstrained comparative advantage, and thus to increase its productivity.[26]

[26] Although our focus is on the effect of AI-biased technological change on scientific productivity, it is worth noting the effect of such change on wages. Analogous to the results in Acemoglu and Restrepo (2019) for the wage effects

The reasons behind the increase in productivity can be more easily seen with the help of Figure 2 (see Appendix 1 for details on the construction of the figure). Figure 2a shows how the log of productivity changes as we add additional tasks, where tasks 0 to $\overline{\tau}$ are done by AI-expert scientists and the remaining tasks are done by ordinary scientists. The figure also shows the optimal share of tasks, τ^*, that would be done with AI-expert scientists if there was no technological constraint on using AI (see Appendix 1 for the determination of τ^*). The assumption given in (13) means that there is a discontinuity at $\overline{\tau}$. It follows that an exogenous improvement in the share of tasks that can be done with AI-expert scientists will increase the total area under the relevant portions of the two productivity curves and thus overall scientific productivity.

Moreover, from Figure 2b, we can see that an increase in the share of AI-expert scientists in the scientific workforce will, through its impact on the relative wage, increase the gap between the two productivity curves at $\overline{\tau}$, and thus the size of the increase in productivity gain from an exogenous increase in $\overline{\tau}$. Intuitively, the distortion created by the technological constraint to the allocation of resources based on comparative advantage is increasing the share of AI-expert scientists the workforce (which increases the size of the gap between the curves at $\overline{\tau}$). It follows that the increase in scientific productivity that results from a given exogenous rate of increase in the share of tasks that can be done by AI-expert expert scientists is increasing in the share of those scientists in the scientific workforce. Note also that the optimal share of tasks done by AI-expert scientists also increases (from τ^* to τ^{**}) when the share of AI-experts in the scientific workforce increases. Thus we can think of the technological constraint as generating a bigger distortion the larger is the share of AI-expert workers in the workforce.

It is also useful to consider the effect of an increase in the AI-expert share on the level of productivity at a given $\overline{\tau}$. Taking logs of (16) and then taking the derivative with respect to the digital skills share, we obtain:

of automation, ordinary scientists will be subject to a negative "displacement effect" as well a positive "productivity effect." It is possible that the wage of ordinary scientists will fall – probably the most consequential finding of task-based models. On the other hand, as the use of AI-expert scientists expands, the wage of these scientists will unambiguously rise. As we have seen, it follows that the share of income going to AI-expert scientists will also rise.

$$\frac{\partial \ln\left(\frac{\dot{A}}{S}\right)}{\partial\left(\frac{S_{AI}}{S}\right)} = \overline{\tau}\frac{S}{S_{AI}} - (1-\overline{\tau})\frac{S}{S-S_{AI}}. \tag{23}$$

This derivative will be exactly equal to zero when the share of tasks that are allocated to AI-expert scientists is equal to the share of AI-expert scientists in the workforce: i.e., $\overline{\tau} = \frac{S_{AI}}{S}$. However, if AI expertise is relatively abundant when compared with the technological possibilities for using that expertise (i.e., $\overline{\tau} < \frac{S_{AI}}{S}$), then an increase in the share of AI-expert scientists will actually lower productivity. However, from (14) and (15), a sufficient condition for $\overline{\tau} \geq \frac{S_{AI}}{S}$ is that $W_{AI} \geq W_O$. We assume that this condition holds because AI-expert scientists are free to work as ordinary scientists so that productivity is non-decreasing in $\frac{S_{AI}}{S}$.[27]

The model of the effects of AI on science in this section is one of pure augmentation – science continues to be done by human scientists, but with the proviso that only scientists with the necessary AI expertise are augmented by AI as it develops as a GPMT. This has led us to concentrate on the importance of downstream complementary investments in AI expertise. While we have focused on augmentation and thus on AI as a tool used by scientists, we do not deny that AI may also automate scientific tasks. A number of authors have recently emphasized that the balance in AI development between automation and augmentation – the direction of the upstream investments –is a choice (see, e.g., Acemoglu and Johnson (2023); Brynjolfsson (2022)). Thus, while the picture we paint of how AI can (conditionally) augment scientists is a relatively optimistic one, it is premised on a particular path that keeps human judgment central in the scientific process. Alternative scenarios – such as that set out in the widely discussed AI-2027 scenario forecast (Kokotajlo et al., 2025) – see self-improving, automated AI becoming dominant in the research process in the surprisingly near future. Whatever the merits of such scenario forecasts, they do point to the importance of policy in directing the development of upstream AI technology in a way that supports a human-scientist-driven version of the future of science.

We have stressed the importance of scientist's possessing AI expertise to be in a position to use

[27] One rationale for this assumption is that it is possible for AI-expert scientists to work as ordinary scientists (but not the reverse). A scientist who is capable of working as an AI-expert will not choose to utilise their expertise unless $W_{AI} \geq W_O$. Thus we can think of the relative skill stocks adjusting (i.e., AI-expert scientists choosing to work as ordinary scientists) until this restriction holds.

AI in the scientific process. Although our model does not distinguish between types of expertise, in reality there is obviously a wide range of such expertise, going from the relatively simple (e.g., using a chabot for brainstorming) to the relatively complex (e.g., designing an API to undertake a specialized task with an underlying LLM). Although we have emphasized the importance of acquiring AI expertise, another important force is likely to be the increasing ease with which AI tools can be used, thereby limiting the need for costly investments in expertise. Beyond its impressive capabilities, the fast diffusion of ChatGPT following its launch in later 2022 was made possible its ease of use, allowing people with very limited knowledge of AI to productively use it for productive tasks. In reality, then, the diffusion of relevant AI expertise will be affected by both investments in that expertise and developments that make the AI easier to use without much specialized investment.

One much-debated issue is how access to AI will affect students' ability to acquire non-AI expertise. On the positive side, the use of AI tutors or the integration of lecture notes and other course specific materials into a course-specific chatbot could support the teaching and learning process. On the negative side, however, there are growing fears that access to AI – say, to write essays, produce software code or solve math problems – could undermine the process of expertise acquisition undermining the very skills that are complementary to AI. The concern is that the development the critical-thinking skills that underpin judgment require practice – with the actual practice undertaken being reduced by the easy (if sometimes formally prohibited) access to AI to undertake the learning task. While we are convinced by the importance of expertise in AI use, the practical challenge for both students and teachers will be develop this expertise without undermining the acquisition of the critical-thinking skills that underpin scientific judgment.

Finally, although the model of this section highlights one (market-based) mechanism through which the availability of AI expertise could intermediate the productivity-enhancing effects of improvements in AI, other mechanisms are likely to be operative as well. A strong assumption in the model is that the AI technology diffuses instantly. The supply of AI expertise then affects productivity through its impact on relative wages. Another important mechanism might be how the availability of AI expertise affects the speed of diffusion of innovations in AI. It is easy to imagine

that the lack of such expertise could be a barrier to diffusion (which could be captured in our model by making $\overline{\tau}$ depend on the level of expertise). We plan to explore such broader mechanisms in future work.

6 The Science of AI

[**This is very much a first stab. You guys are probably much closer to how things are developing. But hopefully there are some ideas in here that you can run with**]

AI is distinctive among general purpose technologies in terms of the degree of dynamism is 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 almost a monthly basis. These advancements are altering the shape of the jagged frontier, increasing the range of tasks where human capabilities – and especially human judgment – that 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 to OpenAI's ChatGPT). 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 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 sees as the "bitter lesson" from 70 years of AI research has had a big impact on this debate.[28] In particular, Sutton 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

[28]The 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, Chief AI Scientist at Meta and Turing Award winner, goes so far as to call LLMs as 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 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. In this perspective AI complements human capabilities, enhancing productivity across the multi-stage scientific process. From this perspective, 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 model in 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 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.

Taken together, these contributions point to a rich agenda for future research on AI in science. Key directions include empirical studies to map the jagged frontier of AI capabilities across domains, quantifying the parameters of augmentation and substitution models, and assessing the evolution of bottlenecks through longitudinal data on scientific workflows. Further work could explore the implications for scientific labor markets, such as skill-biased technological change and the demand for AI expertise among researchers, as well as policy interventions to foster complementary investments in data infrastructure and ethical frameworks. As AI continues to advance rapidly, integrating insights from these perspectives will be essential to understanding not only productivity gains but also the broader societal transformations in how knowledge is produced and applied.

References

Acemoglu, D. (2024, May). The simple macroeconomics of ai. NBER Working Paper 32487, National Bureau of Economic Research.

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, Volume 4B*, pp. 1043–1171. Amsterdam: Elsevier.

Acemoglu, D. and S. Johnson (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. New York: PublicAffairs.

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. 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.

Amodei, D. (2024). Machines of loving grace. `https://www.darioamodei.com/essay/machines-of-loving-grace`. Essay.

Autor, D. and N. Thompson (2025). Expertise. *Journal of the European Economic Association*, jvaf023.

Autor, D. H., F. Levy, and R. J. Murnane (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics 118*(4), 1279–1333.

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.

Brynjolfsson, E. (2022). The turing trap: The promise & peril of human-like artificial intelligence. *Dædalus 151*(2), 28–31.

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.

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. Krayer, 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.

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.

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*.

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.

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.

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*.

Kokotajlo, D., S. Alexander, T. Larsen, E. Lifland, and R. Dean (2025, April). Ai 2027. Originally published April 3, 2025.

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.

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.

Ludwig, J. and S. Mullainathan (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics 139*(2), 751–827.

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.

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.

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.

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: HarperCollins.

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics 113*(2), 331–360.

Whewell, W. (1989). *Theory of scientific method*. Hackett Publishing.

Zeira, J. (1998). Workers, machines, and economic growth. *Quarterly Journal of Economics 113*(4), 1091–1117.

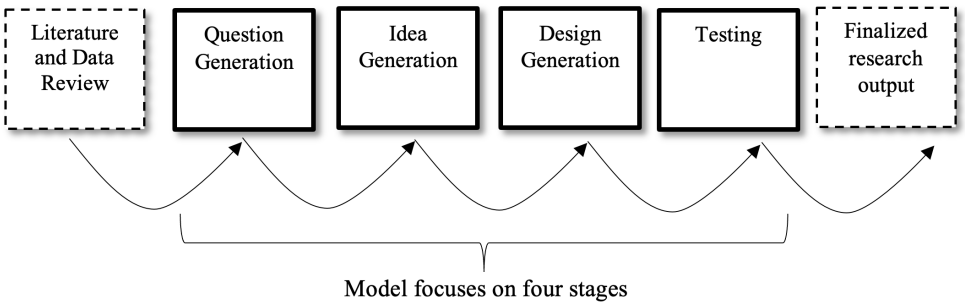
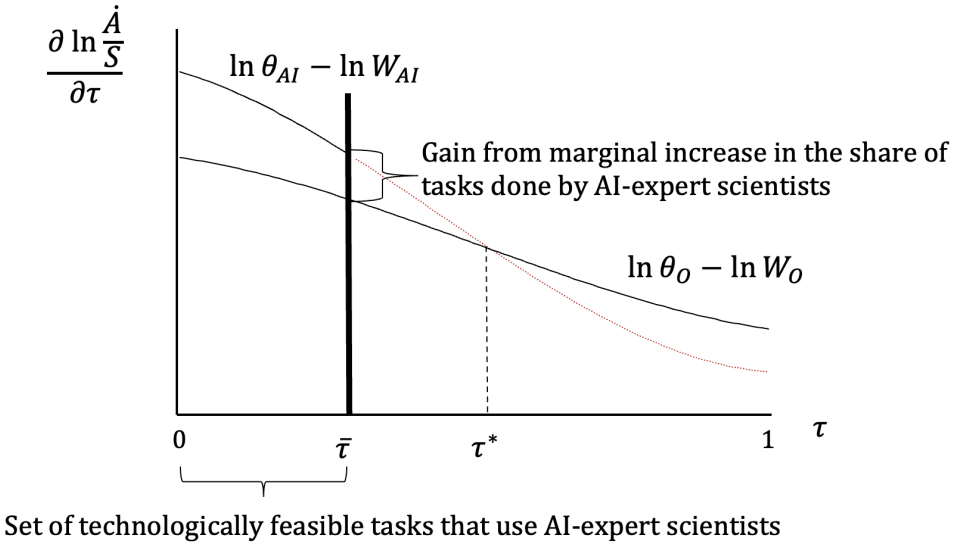
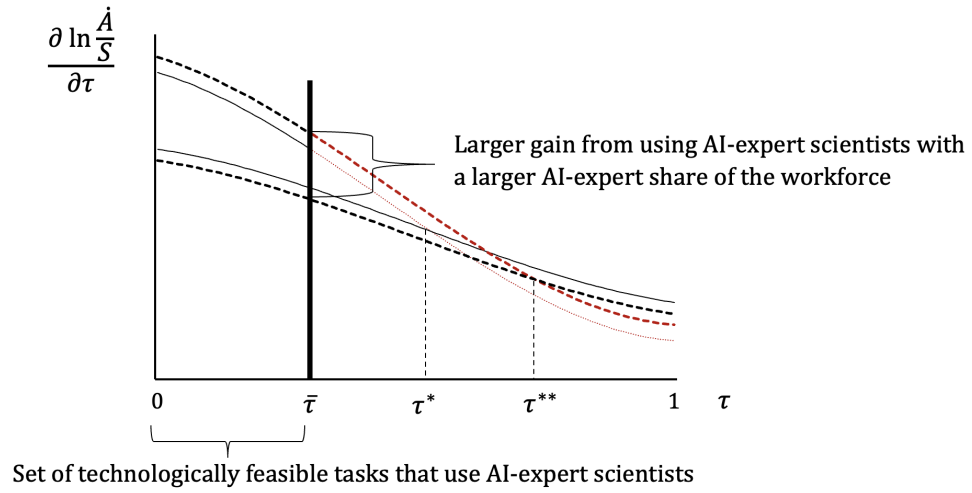


Figure 1: Stylized stages of scientific research



(a) Potential Gain from AI-Biased Technological Change



(b) Impact of an Increase in the AI-Expert Share of the Scientific Workforce on the Gain from AI-Biased Technological Change

Figure 2

Notes: Figure 2a shows the constrained comparative advantage equilibrium where the production of science requires a continuum of tasks indexed from 0 to 1. $\bar{\tau}$ denotes the share of tasks for which AI can be deployed, where $\bar{\tau}$ is viewed as a technological constraint given the current state of AI. The use of AI requires complementary AI-expert scientists. It is assumed that the current equilibrium is constrained in the sense that the marginal task would be reallocated to AI-expert scientists if it were technologically feasible. As shown, the gap between the curves is a measure of the productivity gain from being able to use AI-expert scientists for an additional task. Technological improvements in AI are therefore associated with increased scientific productivity. Figure 2b shows how the size of the gain from a marginal improvement in the share of tasks for which it is feasible to use AI-expert scientists will increase with the share of AI-experts scientists in the workforce. The mechanism is that increase in the relative supply of AI-expert scientists drives down their relative wage, thus increasing the productivity gain from employing these scientists in general equilibrium.

Appendix 1: Additional Details on the Task-Based Model

This appendix develops additional details of the task-based model. We first write down the price of task τ as the marginal (equal to average) cost of delivering that task:

$$p(\tau) = \begin{cases} \dfrac{W_{AI}}{\theta_{AI}(\tau)}, & \text{if } \tau \in [0, \overline{\tau}] \\ \dfrac{W_O}{\theta_O(\tau)}, & \text{if } \tau \in (\overline{\tau}, 1] \end{cases}. \tag{A.1.1}$$

Given the Cobb-Douglas form of the task-based production function, the demand for task τ is given as usual by:

$$z(\tau) = \frac{\dot{A}}{p(\tau)}. \tag{A.1.2}$$

Using (A.1.1) and (A.1.2) we can then write the demand for AI-expert scientists in task τ as:

$$\frac{z(\tau)}{\theta_{AI}(\tau)} = s_{AI}(\tau) = \begin{cases} \dfrac{\dot{A}}{W_{AI}}, & \text{if } \tau \in [0, \overline{\tau}] \\ 0, & \text{if } \tau \in (\overline{\tau}, 1] \end{cases}. \tag{A.1.3}$$

And the demand for ordinary scientists in task τ as:

$$\frac{z(\tau)}{\theta_O(\tau)} = s_O(\tau) = \begin{cases} 0, & \text{if } \tau \in [0, \overline{\tau}] \\ \dfrac{\dot{A}}{W_O}, & \text{if } \tau \in (\overline{\tau}, 1] \end{cases}. \tag{A.1.4}$$

Summing over tasks, the aggregate demands for the factors are then given by:

$$\sum_{\tau=0}^{\overline{\tau}} s_{AI}(\tau) = S_{AI} = \overline{\tau}\frac{\dot{A}}{W_{AI}}, \tag{A.1.5}$$

$$\sum_{\tau=\overline{\tau}}^{1} s_O(\tau) = S_O = (1-\overline{\tau})\frac{\dot{A}}{W_O}, \tag{A.1.6}$$

where we also impose the constraint that the quantity demanded of each worker type must equal to the quantify of that type that is (inelastically) supplied. Rearranging the aggregate

demand functions yields the expressions for the competitively determined wages for each scientist type given in the main text:

$$W_{AI} = \overline{\tau}\frac{\dot{A}}{S_{AI}}, \tag{A.1.7}$$

$$W_O = (1-\overline{\tau})\frac{\dot{A}}{S_O}, \tag{A.1.8}$$

Given the Cobb-Douglas form of the task-based production function, the price index for a unit of output is:

$$P = \exp\left(\int_{\tau=0}^{1} \ln p(\tau) d\tau\right) A = 1, \tag{A.1.9}$$

which is normalized to one without loss of generality. Taking logs:

$$\ln P = \int_{\tau=0}^{1} \ln p(\tau) d\tau + \ln A = 0. \tag{A.1.10}$$

Now substituting for the task prices from (A.1.1) and using the factor wage expressions we have:

$$\int_{\tau=0}^{\overline{\tau}} (\ln W_{AI} - \ln \theta_{AI}(\tau)) d\tau + \int_{\tau=\overline{\tau}}^{1} (\ln W_O - \ln \theta_O(\tau)) d\tau + \ln A = 0. \tag{A.1.11}$$

$$\Rightarrow \int_{\tau=0}^{\overline{\tau}} \left[\ln \dot{A} - \ln\left(\frac{S_{AI}}{\overline{\tau}}\right) - \ln \theta_{AI}(\tau)\right] d\tau + \int_{\tau=\overline{\tau}}^{1} \left[\ln \dot{A} - \ln\left(\frac{S_O}{1-\overline{\tau}}\right) - \ln \theta_O(\tau)\right] d\tau + \ln A = 0. \tag{A.1.12}$$

Rearranging and subtracting $\ln S$ from both sides, we obtain the endogenous form of the productivity function in log form as:

$$\ln \frac{\dot{A}}{S} = \overline{\tau} \ln \left(\frac{S_{AI}}{\overline{\tau} S} \right) + (1 - \overline{\tau}) \ln \left(\frac{S_O}{(1 - \overline{\tau}) S} \right) + \left[\int_{\tau=0}^{\overline{\tau}} \ln \theta_{AI}(\tau) d\tau + \int_{\tau=\overline{\tau}}^{1} \ln \theta_O(\tau) d\tau \right] + \ln A. \quad \text{(A.1.13)}$$

Taking the exponential of both sides:

$$\frac{\dot{A}}{S} = B(\overline{\tau}) A \left(\frac{S_{AI}}{\overline{\tau} S} \right)^{\overline{\tau}} \left(\frac{S_O}{(1 - \overline{\tau}) S} \right)^{1 - \overline{\tau}}, \quad \text{(A.1.14)}$$

$$\text{where } B(\overline{\tau}) = \exp \left[\int_{\tau=0}^{\overline{\tau}} \ln \theta_{AI}(\tau) d\tau + \int_{\tau=\overline{\tau}}^{1} \ln \theta_O(\tau) d\tau \right]. \quad \text{(A.1.15)}$$

Note also that:

$$\ln \frac{\dot{A}}{S} = \int_{\tau=0}^{\overline{\tau}} \frac{\partial \ln (\dot{A}/S)}{\partial \tau} d\tau + \int_{\tau=\overline{\tau}}^{1} \frac{\partial \ln (\dot{A}/S)}{\partial \tau} d\tau = \int_{\tau=0}^{\overline{\tau}} (\ln \theta_{AI}(\tau) - \ln W_{AI}) d\tau + \int_{\tau=\overline{\tau}}^{1} (\ln \theta_O(\tau) - \ln W_O) d\tau, \quad \text{(A.1.16)}$$

which provides the basis for Figure 2.

Finally, it is useful to determine that optimal allocation of tasks to AI-expert workers, which we denote as τ^*. Setting equation (23) in the main text equal to zero and using equations (14) and (15), τ^* is implicitly given by,

$$\tau^* = \frac{1}{\frac{\theta_O(\tau^*)}{\theta_{AI}(\tau^*)} \frac{1 - \frac{S_{AI}}{S}}{\frac{S_{AI}}{S}} + 1}, \quad \text{(A.1.17)}$$

where recall that $\frac{\theta_O(\tau)}{\theta_{AI}(\tau)}$ is rising in τ by assumption. To ensure that an interior unconstrained equilibrium exists and is unique, we assume:

$$0 < \frac{1}{\frac{\theta_O(1)}{\theta_{AI}(1)} \frac{1-\frac{S_{AI}}{S}}{\frac{S_{AI}}{S}} + 1} < 1. \tag{A.1.18}$$

The determination of the unconstrained equilibrium share of tasks allocated to digital skilled workers is shown in Figure A.1. The figure also shows the effect of an increase in the share of the AI-expert scientists on the unconstrained equilibrium allocation of tasks. Not surprisingly, an increase in the share of AI-expert scientists will increase the share of tasks allocated to these scientists from τ^* to τ^{**} where there is no technological constraint on which tasks can be performed by AI-expert scientists. However, as the allocation of tasks is determined by comparative advantage, there will still be tasks will be allocated to ordinary scientists in the unconstrained equilibrium of the model.

In the *constrained* equilibrium of the model we assume that $\frac{S_{AI}}{S} \leq \overline{\tau} < \tau^*$. The first inequality is ensured by the requirement that $W_{AI} \geq W_O$ and is a sufficient condition for the *level* of productivity to be non-decreasing in the AI-expert share of the scientific workforce. The second (strict) inequality is a sufficient condition for the *growth* rate of productivity to be increasing in the AI-expert skills share for a given positive rate of AI-task-biased technological change.

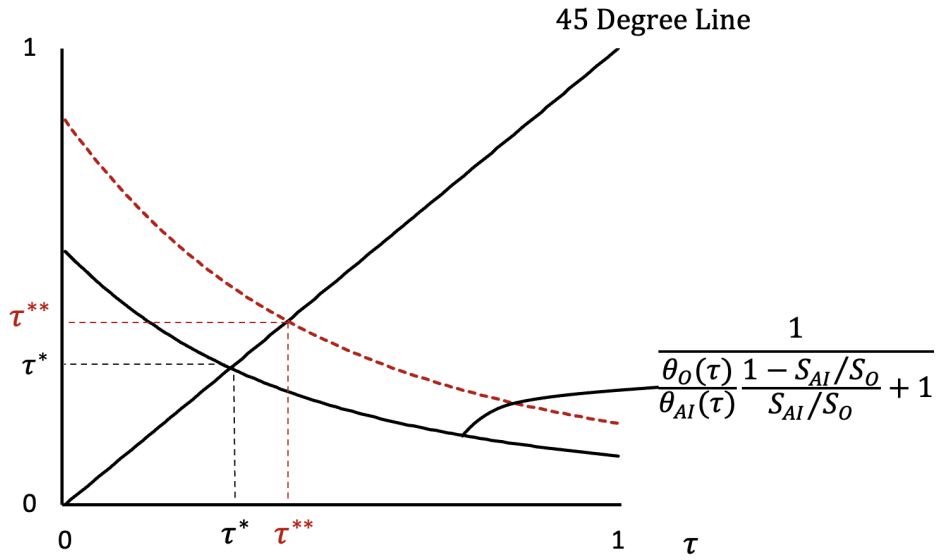


Figure A.1: Determination of unconstrained equilibrium allocation of tasks to AI-expert scientists for two levels of the ratio of AI-expert scientists to ordinary scientists