Artificial Intelligence, Scientific Discovery, and Product Innovation*

Aidan Toner-Rodgers†
MIT

November 3, 2024

This paper studies the impact of artificial intelligence on innovation, exploiting the randomized introduction of a new materials discovery technology to 1,018 scientists in the R&D lab of a large U.S. firm. AI-assisted researchers discover 44% more materials, resulting in a 39% increase in patent filings and a 17% rise in downstream product innovation. These compounds possess more novel chemical structures and lead to more radical inventions. However, the technology has strikingly disparate effects across the productivity distribution: while the bottom third of scientists see little benefit, the output of top researchers nearly doubles. Investigating the mechanisms behind these results, I show that AI automates 57% of "idea-generation" tasks, reallocating researchers to the new task of evaluating model-produced candidate materials. Top scientists leverage their domain knowledge to prioritize promising AI suggestions, while others waste significant resources testing false positives. Together, these findings demonstrate the potential of AI-augmented research and highlight the complementarity between algorithms and expertise in the innovative process. Survey evidence reveals that these gains come at a cost, however, as 82% of scientists report reduced satisfaction with their work due to decreased creativity and skill underutilization.

*I am especially grateful to Daron Acemoglu, David Autor, Jacob Moscona, and Nina Rousille for their guidance and support. I also thank Nikhil Agarwal, Jules Baudet, Vladimir Bulović, Jamie Emery, Sarah Gertler, Julia Gilman, Jon Gruber, Daniel Luo, Florian Mudekereza, Shakked Noy, Julie Seager, Anand Shah, Advik Shreekumar, Scott Stern, Laura Weiwu, Whitney Zhang, and seminar participants at the MIT Applied Micro Lunch for helpful comments. I am indebted to several of the lab's scientists for their generosity in explaining the materials discovery process. This work was supported by the George and Obie Shultz Fund and the National Science Foundation Graduate Research Fellowship Program under Grant No. 2141064. IRB approval for the survey was granted by MIT's Committee on the Use of Humans as Experimental Subjects under ID E-5842. JEL Codes: O31, O32, O33, J24, L65.

†MIT Department of Economics; Email: aidantr@mit.edu; Website: aidantr@github.io

1 Introduction

The economic impact of artificial intelligence will depend critically on whether AI technologies not only transform the production of goods and services, but also augment the process of innovation itself (Aghion et al., 2019; Cockburn et al., 2019). Recent advances in deep learning show promise in generating scientific breakthroughs, particularly in areas such as drug discovery and materials science where models can be trained on large datasets of existing examples (Merchant et al., 2023; Mullowney et al., 2023; DeepMind, 2024). Yet little is known about how these tools impact invention in a real-world setting, where R&D bottlenecks, organizational frictions, or lack of reliability may limit their effectiveness. As a result, the implications of AI for both the pace and direction of innovation remain uncertain. Moreover, the consequences for scientists are ambiguous, hinging on whether AI complements or substitutes for human expertise.

To provide evidence on these questions, I exploit the randomized introduction of an AI tool for materials discovery to 1,018 scientists in the R&D lab of a large U.S. firm. The lab focuses on applications of materials science in healthcare, optics, and industrial manufacturing, employing researchers with advanced degrees in chemistry, physics, and engineering. Traditionally, scientists discover materials through an expensive and time-consuming system of trial and error, conceptualizing many potential structures and testing their properties. The AI technology leverages developments in deep learning to partially automate this process. Trained on the composition and characteristics of existing materials, the model generates "recipes" for novel compounds predicted to possess specified properties. Scientists then evaluate these candidates and synthesize the most promising options. Once researchers create a useful material, they integrate it into new product prototypes that are then developed, scaled, and commercialized.

The lab rolled out the tool in three waves starting in May of 2022. Teams of researchers were randomly assigned to waves, allowing me to identify the effects of the technology by comparing treated and not-yet-treated scientists. The cohorts are balanced on observables like education, experience, and past performance, confirming successful randomization. Using detailed data on each stage of R&D, I study AI's impact on materials discovery and its downstream effects on patenting and product innovation.

AI-assisted scientists discover 44% more materials. These compounds possess superior properties, revealing that the model also improves quality. This influx of materials leads to a 39% increase in patent filings and, several months later, a 17% rise in product prototypes incorporating the new compounds. Accounting for input costs, the tool boosts R&D efficiency by 13-15%. These results have two implications. First, they demonstrate the potential of AI-augmented research. Second,

they confirm that these discoveries translate into product innovations, not fully bottlenecked by later stages of R&D.

AI accelerates the pace of innovation, but how novel are these breakthroughs? A key concern with using machine learning for science is its potential to amplify the "streetlight effect" (Khurana, 2023; Kim, 2023; Hoelzemann et al., 2024). Because models are trained on existing knowledge, they might direct search toward well-understood but low-value areas. Contrary to this hypothesis, I find that the tool increases novelty in all three stages of R&D. I first measure the originality of new materials themselves, following the chemical similarity approach of De et al. (2016). Compared to existing compounds, model-generated materials have more distinct physical structures, suggesting that AI unlocks new parts of the design space. Second, I show that this leads to more creative inventions. Patents filed by treated scientists are more likely to introduce novel technical terms—a leading indicator of transformative technologies (Kalyani, 2024). Third, I find that AI changes the nature of products: it boosts the share of prototypes that represent new product lines rather than improvements to existing ones, engendering a shift toward more radical innovation (Acemoglu et al., 2022a).

Next, I turn to the technology's distributional effects, showing that it disproportionately benefits high-ability scientists. I construct a measure of initial productivity based on discoveries in the pre-treatment period. To account for the possibility that certain compounds are inherently easier to discover, I control for material type and application. Estimating separate treatment effects for each productivity quantile, I document strikingly disparate impacts across the ability distribution. While the bottom third of researchers see minimal gains, the output of top-decile scientists increases by 81%. Consequently, 90:10 performance inequality more than doubles. This suggests that AI and human expertise are complements in the innovation production function.

The second part of the paper investigates the mechanisms behind these results. Combining rich text data on scientist activities with a large language model to categorize them into research tasks, I show that AI dramatically changes the discovery process. The tool automates a majority of "idea generation" tasks, reallocating scientists to the new task of evaluating model-suggested candidate compounds. In the absence of AI, researchers devote nearly half their time to conceptualizing potential materials. This falls to less than 16% after the tool's introduction. Meanwhile, time spent assessing candidate materials increases by 74%. AI therefore has countervailing effects: while it replaces labor in the specific activity of designing compounds, it augments labor in the broader discovery process due to its complementarity with evaluation tasks.

Next, I show that scientists' differential skill in judging AI-generated candidate compounds explains the tool's heterogeneous impact. I collect data on the materials researchers test and the

outcomes of these experiments. Top scientists leverage their expertise to identify promising AI suggestions, enabling them to investigate the most viable candidates first. In contrast, others waste significant resources investigating false positives. Indeed, a significant minority of researchers order their tests no better than random chance, seeing little benefit from the tool. Evaluation ability is positively correlated with initial productivity, explaining the widening inequality in scientists' performance. These results demonstrate the growing importance of a new research skill—assessing model predictions—that complements AI technologies. I therefore provide evidence for the argument in Agrawal et al. (2018) that improvements in machine prediction make human judgment and decision-making more valuable.

To understand the source of these large differences in judgment, I conduct a survey of the lab's researchers. The responses reveal the central role of domain knowledge. Scientists skilled in evaluation credit their training and experience with similar materials as key to their assessment process. Meanwhile, those who struggle to judge AI-suggested compounds report that their background offers little assistance. Supporting this explanation, researchers in the top quartile of evaluation ability are 3.4 times more likely to have published an academic article on their material of focus. While some posit that big data and machine learning will render domain knowledge obsolete (Anderson, 2008; Gennatas et al., 2020), these results show that only scientists with sufficient expertise can harness the power of AI.

I combine my estimates with a simple model to illustrate how organizational adaptation can amplify the tool's impact. AI alters the returns to specific skills, increasing the value of judgment while diminishing the importance of idea generation. Therefore, adjusting employment practices to prioritize scientists with strong judgment implies significant productivity gains. In the final month of my sample—excluded from the primary analysis—the lab fired 3% of its researchers. Consistent with the theory, 83% of these scientists were in the bottom quartile of judgment. The lab more than offset these departures through increased hiring, expanding its workforce on net. Due to this change in the composition of researchers, my estimates may understate AI's longer-run impact.

The consequences of new technologies extend beyond productivity. They can profoundly affect worker wellbeing and the expertise needed to succeed on the job (Nazareno and Schiff, 2021; Soffia et al., 2024). These considerations are particularly salient in the context of innovation, as they mediate AI's impact on who becomes a scientist, the fields they enter, and skills they invest in. The final part of the paper explores these questions using the survey.

Researchers experience a 44% reduction in satisfaction with the content of their work. This effect is fairly uniform across scientists, showing that even the "winners" from AI face costs. Respondents cite skill underutilization and reduced creativity as their top concerns, highlighting the difficulty

of adapting to rapid technological progress. Moreover, these results challenge the view that AI will primarily automate tedious tasks, allowing humans to focus on more rewarding activities. While enjoyment from improved productivity partially offsets this negative effect, especially for high-ability scientists, 82% of researchers see an overall decline in wellbeing.

In addition to impacting job satisfaction, working with the tool changes materials scientists' views on artificial intelligence. Belief in the ability of AI to enhance productivity nearly doubles. At the same time, concerns over job loss remain constant, reflecting the continued need for human judgment. However, due to the changing research process, scientists expect AI to alter the skills needed to succeed in their field. Consequently, the number of researchers planning to reskill rises by 71%. These findings show that hands-on experience with AI can meaningfully influence views on the technology. The responses also reveal an important fact: domain experts did not anticipate the effects documented in this paper.

While my study focuses on materials science, the insights may apply more generally to fields where the discovery process requires search over a vast but well-defined technological space. This characterizes areas where the foundational principles are known, but complexity makes it challenging to identify specific instances. In drug discovery, for example, the properties of atomic bonds are well established, but the large number of possible chemical configurations makes the problem extremely difficult. Deep learning models—which excel in extracting features from complex data—have the potential to transform research in such settings (Hassabis, 2022). Beyond materials science and pharmaceuticals, several economically important fields fall into this category, including structural biology (Jumper et al., 2021; Abramson et al., 2024; Subramaniam, 2024), genomics (Caudai et al., 2021), climatology (Kochkov et al., 2024), and even certain parts of mathematics (Tao, 2024; Trinh et al., 2024).

Related Literature This paper contributes to four related literatures. First, it adds to a large body of evidence on the consequences of new technologies for productivity, labor demand, and organizations (Autor et al., 1998; Athey and Stern, 2002; Bresnahan et al., 2002; Autor et al., 2003; Bloom et al., 2014; Autor, 2015; Garicano and Rossi-Hansberg, 2015; Webb, 2020; Acemoglu and Restrepo, 2022; Agrawal et al., 2019; Acemoglu et al., 2022b; Kogan et al., 2023; Autor et al., 2024). While these studies focus on the production of goods and services, I consider a breakthrough that augments the process of innovation itself. In standard models, technologies that automate production tasks have qualitatively different effects than those performing research tasks, given the compounding social returns to innovation (Aghion et al., 2019; Jones and Summers, 2020).

Therefore, understanding the implications of AI for R&D is of central importance.[1] Indeed, my results suggest that analyses of AI that do not consider effects on science and innovation (e.g., Acemoglu, 2024) may be incomplete.[2]

Second, I contribute to recent evidence on the productivity effects of "generative" AI. This literature studies the application of large language models to mid-skill writing and coding tasks, revealing that LLMs boost average output and compress the productivity distribution (Brynjolfsson et al., 2023; Dell'Aqua et al., 2023; Noy and Zhang, 2023; Peng et al., 2023; Cui et al., 2024). I analyze the adoption of a different type of generative model—one that produces novel material designs rather than words. While I similarly observe large gains in performance, AI is most useful to high-ability scientists in my setting, increasing inequality. This suggests that the so-called "leveling-up" effect is context-dependent and not a general feature of AI technologies. Moreover, since the model unlocks breakthroughs beyond the capabilities of even highly trained scientists, my findings help alleviate concerns that AI will contribute only to low-value tasks (Angwin, 2024).

Third, I relate to work on human-AI collaboration, particularly the relationship between algorithms and expertise (Kleinberg et al., 2017; Chouldechova et al., 2018; Cowgill, 2018; Cheng and Chouldechova, 2022; Donahue et al., 2022; Leitão et al., 2022; Mullainathan and Obermeyer, 2022; Alur et al., 2023; Angelova et al., 2023; Agarwal et al., 2024; Alur et al., 2024; Gruber et al., 2024). Researchers display considerable variation in their ability to evaluate AI-suggested compounds. Discretion boosts discovery rates for two-thirds of scientists, who leverage their domain knowledge to improve upon the model. This shows that these experts observe certain features of the materials design problem not captured by the algorithm. In contrast, a significant minority of scientists see little benefit from the technology due to poor judgment. This dichotomy highlights the need for further research on human-AI collaboration in generative tasks.

Finally, I add to a nascent literature on the implications of artificial intelligence for science and innovation (Cockburn et al., 2019; Agrawal et al., 2023; Besiroglu et al., 2023; Ludwig and Mullainathan, 2024; Manning et al., 2024; Mullainathan and Rambachan, 2024). This paper provides the first causal evidence of AI's impact on real-world R&D. I show that AI can automate the critical ideation step of the materials discovery process, leading to an increase in invention and a rise in product innovation. Moreover, I quantify the effect of this automation on scientists, revealing a

[1] A recent debate centers around AI's macroeconomic implications (Bostrom, 2014; Brynjolfsson and Unger, 2023; Clancy and Besiroglu, 2023; Erdil and Besiroglu, 2023; Ramani and Wang, 2023; Schulman, 2023; Aschenbrenner, 2024; Chow et al., 2024; Korinek and Suh, 2024). While these papers agree that AI's impact on innovation is critical, a lack of empirical evidence on this question has prevented consensus.

[2] Acemoglu (2024) expects that significant impacts of AI on overall science and innovation are more than a decade away. While my results do not directly speak to such aggregate effects, they do show that AI technologies can already accelerate R&D in specific settings.

dramatic reallocation across research tasks. Consequently, I establish a micro-foundation for the finding that investments in AI technologies predict subsequent firm growth (Babina et al., 2024).

Organization The paper is structured as follows. Section 2 provides information on the setting, detailing the materials science R&D process and the AI technology. Section 3 describes the data, measurement strategy, and research design. Section 4 presents the main results on discovery, patenting, and product innovation. Section 5 analyzes heterogeneity in the effect of AI across scientists and Section 6 studies the dynamics of the human-AI collaboration. Section 7 explores the tool's impact on scientists' job satisfaction and beliefs about AI. Section 8 concludes.

2 Setting and Background

2.1 Materials Science R&D Lab

My setting is the R&D lab of a large U.S. firm. The lab specializes in materials science, a field that integrates insights from physics, chemistry, and engineering to create novel substances and incorporate them into products. Often deemed the "unsung hero" of technological progress, advances in materials science underpin many significant breakthroughs (Diamandis, 2020). The purification of silicon in the 1950s enabled the development of integrated circuits, laying the foundation for modern computing. Graphene—a crystalline carbon lattice created in 2004—has transformed numerous products ranging from batteries to desalination filters. More recently, novel photovoltaic structures have enhanced solar panel efficiency, driving the steep decline in renewable energy costs. And in medicine, biocompatible compounds allow implants to merge seamlessly with human tissue, improving drug delivery systems. Moskowitz (2022) estimates that more than two thirds of new technologies rely on innovative materials, highlighting the field's economic importance.

The lab in my study focuses on applications of materials science in healthcare, optics, and industrial manufacturing, producing a wide range of patents and product innovations. To create these technologies, the firm employs thousands of highly trained scientists with advanced degrees in chemical engineering, physics, and materials science. Researchers are organized in teams, specializing in material-application pairs relevant to their expertise. These groups also include support staff such as technicians and administrators.

Figure 1 summarizes the R&D pipeline. First, scientists define a set of target properties and generate ideas for new compounds predicted to satisfy these requirements. Before the introduction of AI, researchers employed a combination of domain knowledge and iterative computational

procedures to create preliminary designs. Given the difficulty of predicting material characteristics, this process is time-intensive and involves many false positives (Reiser et al., 2022).

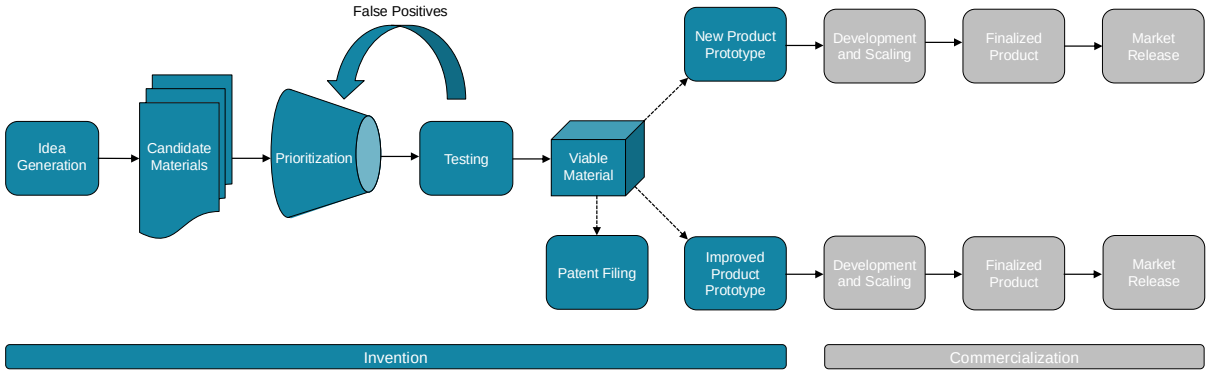


Figure 1. Materials Science R&D Pipeline

Notes: This figure summarizes the materials science R&D pipeline. First, scientists generate new compound designs. Next, they evaluate these candidates and test the most promising. After scientists discover a viable material, they typically file for a patent and incorporate it into product prototypes. These could represent entirely new products or improvements to existing lines. Finally, prototypes are developed, mass-produced, and released to market. My analysis focuses on the "invention" steps, depicted in blue.

Next, scientists evaluate candidate compounds and decide which to test. Experiments are costly, so prioritization is key to efficiency.[3] To identify the most promising candidates, researchers judge compound designs using simulations and prior experience with similar materials. Once the lab selects a candidate for testing, they advance it through a series of experiments. First, they attempt to synthesize the material, ruling out a large share of candidates that do not yield stable compounds. Next, they conduct tests to assess its properties at both the atomic and macro scales. Finally, they subject it to real-world conditions such as heat, pressure, or human interaction.

After discovering a material, scientists often file for a patent. This can pertain to a single compound, a combination of compounds, or a new technology that uses them. Patents require three criteria: novelty, utility, and non-obviousness. As a result, they mark the stage of research where a scientific discovery transforms into a useful invention. Patent applications typically take two years for approval, so my analysis focuses on filings. Historically, more than 80% of the lab's applications have been approved.

Finally, the lab integrates new materials into products. These could represent entirely new product lines or improvements to existing ones. In either case, researchers first develop a prototype that is then advanced through development and mass production. The "commercialization lag"

[3]For example, the machines used to grow crystal structures are extremely capital-intensive and require costly inputs for each use.

between the creation of a working prototype and its market release can be substantial, ranging from years to decades. Consequently, my analysis focuses on the prototyping stage. This paper's results should therefore be understood as reflecting the "invention" component of R&D (Budish et al., 2015).

2.2 Deep Learning for Materials Discovery

Materials discovery is challenging due to its complexity. The space of plausible chemical configurations is vast, requiring scientists to explore many potential compounds. Moreover, while the properties of atomic bonds are well known, it is difficult to predict how they will aggregate into large-scale characteristics. Deep learning models—which excel in extracting features from complex data—have the potential to overcome these challenges (Hassabis, 2022).

Recent years have seen an explosion of large, standardized databases compiling the structure and characteristics of known compounds. Combined with algorithmic progress and increased compute, this has greatly improved the performance of deep learning in materials science (Reiser et al., 2022). As a result, the field has shown rapidly growing interest in these techniques.

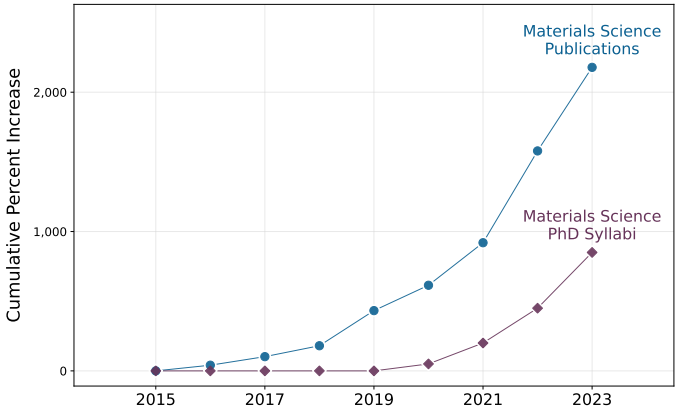


Figure 2. The Rise of Deep Learning in Materials Science

Notes: This figure shows the rise of deep learning in materials science. The blue circles represent the cumulative percent increase in materials science publications mentioning deep learning since 2015, based on data from Web of Science. The purple diamonds plot the cumulative percent increase in materials science PhD syllabi covering deep learning, sourced from Open Syllabus.

Figure 2 illustrates the rise of deep learning in materials science. Since 2015, the number of materials science publications referencing deep learning has grown more than 20-fold. This follows a decade of foundational progress in the computer science literature.[4] At the same time, deep

[4]Seminal papers include Goodfellow et al. (2014); Mirza and Osindero (2014); Radford et al. (2016); Arjovsky et al. (2017); Karras et al. (2018).

learning has become integrated into materials science curricula, evidenced by a sharp increase in PhD syllabi covering these methods.

2.3 AI Tool

The lab's AI technology is a set of graph neural networks (GNNs) trained on the structure and properties of existing materials. Introduced by Scarselli et al. (2009), GNNs extend generative models to settings where geometric structure contains essential information.[5] The GNN architecture represents materials as multidimensional graphs of atoms and bonds, enabling it to learn physical laws and encode large-scale properties.

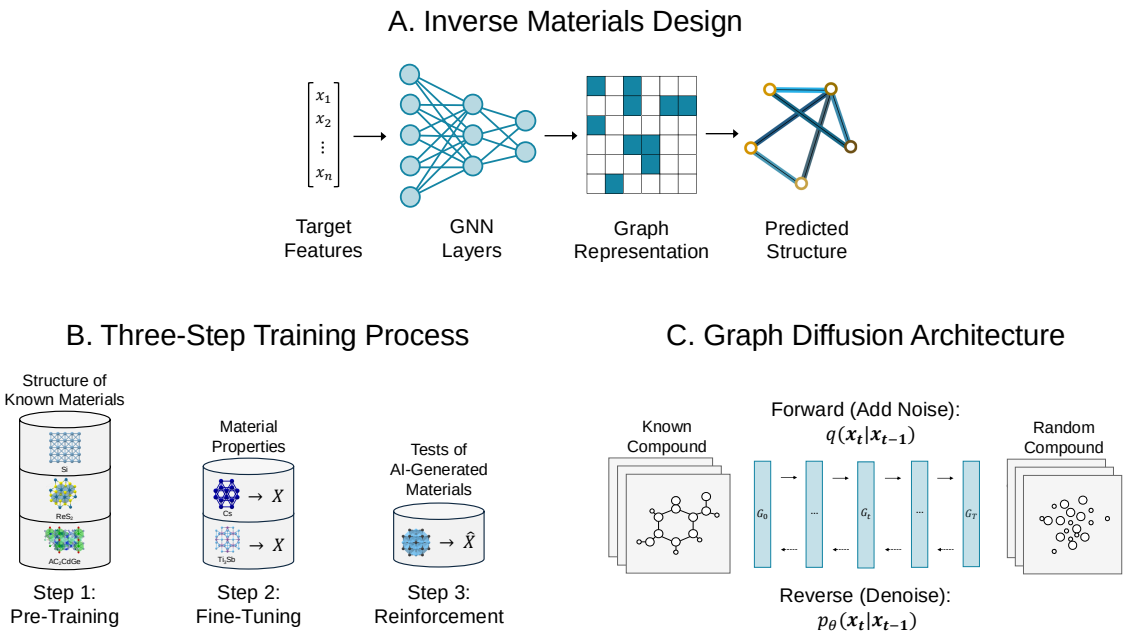


Figure 3. Illustration of Materials Discovery Tool

Notes: This figure shows the structure of the lab's AI tool. The graph neural network is input a set of target features and outputs a predicted structure (Panel A). The model is trained in three steps: pre-training based on the structure of known materials, fine-tuning for a specific application based on material properties, and reinforcement learning based on scientists' experiments on AI-generated compounds (Panel B). It generates stable materials by reversing a corruption process, iteratively denoising an initially random structure (Panel C).

Scientists use the tool for "inverse materials design," illustrated in Figure 3, Panel A. Researchers input a set of desired characteristics and the model generates candidate compounds predicted to possess these properties (Zunger, 2018; Noh et al., 2020). It undergoes three stages of training to optimize its performance: pre-training on a large dataset of known structures, fine-tuning with

[5]Foundational papers on graph neural networks include Henaff et al. (2015), Johnson (2017), Kipf and Welling (2017), Li et al. (2018), and Velickovic et al. (2018).

application-specific material properties, and reinforcement learning using tests of model-suggested compounds (Panel B). The model employs a diffusion-based approach to generate novel materials. It begins with a known structure, adds noise, and then reverses the process to create a new compound (Panel C). While AI designs have the potential to rapidly advance discovery, materials proposed by even frontier models are often unstable or display other undesirable features.[6] Consequently, it is essential to pair the technology with scientists who can evaluate, refine, and test candidate compounds.

2.4 Model Rollout

Following a short pilot program, the lab began a large-scale rollout of the model in May of 2022. They randomly assigned teams of researchers to three adoption waves spaced approximately six months apart. The firm randomized deployment due to uncertainty about the tool's impacts and to determine if they should alter hiring and training practices.[7] The treatment groups were comprised of 404, 419, and 195 scientists, respectively. At the start of each wave, researchers participated in a training program teaching them how to use the technology. My analysis focuses on the three main treatment groups, totaling 1,018 scientists across 221 teams.

3 Data, Measurement, and Research Design

3.1 Data and Measurement

I combine several data sources to create a detailed picture of the R&D process. Appendix B provides a full description of variable definitions, sample construction, and measurement procedures.

Materials I collect data on candidate compounds, synthesized substances, and finalized materials. These include information on compounds' physical structure, captured by the composition and geometric orientation of their atoms and bonds. Additionally, I observe the results from tests of material properties, providing a large set of characteristics at the atomic and macro scales.

I classify new materials as "discovered" once they are added to the lab's internal database of compounds deemed viable for product use. This marks the transition from science to engineering, after which the materials are developed and produced at scale. I link compounds to the scientists

[6] For example, an analysis of 2.2 million crystal structures discovered by DeepMind's AI tool found "scant evidence for compounds that fulfill the trifecta of novelty, credibility and utility" (Cheetham and Seshadri, 2024).

[7] To my knowledge, no changes to hiring and training were made during the experiment. Nevertheless, I exclude new hires from my main analysis.

that create them. Because researchers collaborate in teams, many materials are credited to multiple scientists. However, not all team members work on every material their team discovers.

I measure the quality of compounds based on their properties. Scientists begin the search process by defining a set of target features. These features vary significantly across materials, but often include both atomic and large-scale characteristics determined by the intended application. For example, researchers may target the band gap (the energy difference between the valence and conduction bands) or the refractive index (how much a material bends light). For each feature, I calculate the distance between the target and its realized value. Following the materials science literature, I combine these distances into three indices: one capturing quality at the atomic scale, another at the macro scale, and a third that combines all properties into a single index. Appendix B.1 details the construction of these indices.

To assess the novelty of materials, I employ the chemical similarity approach of De et al. (2016). This method is similar in spirit to the procedure used by Krieger et al. (2021) to measure the novelty of pharmaceutical drugs. However, it includes additional features necessary for studying the inorganic compounds in my setting. Premised on the idea that "form determines function," it defines the innovativeness of a new material by comparing its physical structure to existing compounds. The similarity measure captures both the set of elements in a compound and their geometric orientation. I compare each discovered compound to more than 150,000 materials in the Materials Project and Alexandria databases (Jain et al., 2013; Schmidt et al., 2022). This methodology applies only to crystal structures, solids with orderly and periodic atomic arrangements. The novelty results are therefore based solely on the crystals in my sample, comprising 64% of all materials. Appendix B.2 describes the similarity measure in detail.

Patent Filings I match new materials to patent filings. I include patents for both compounds themselves and technologies that use them. Given that patent applications take approximately two years for approval, my results focus on filings rather than approvals. Historically, more than 80% of the lab's applications have been approved. The patent data are useful for two reasons. First, patents identify inventions as significant, applicable breakthroughs. Second, the text of patent filings allows me to assess the novelty of inventions using similarity measures.

The first similarity measure, proposed by Kelly et al. (2021), captures a broad notion of novelty using the full text of patent filings. It quantifies textual similarity using the cosine similarity between vectors of term frequencies. To appropriately weight terms by their importance, each component is scaled by the inverse frequency of that term in the full sample. The resulting metric lies in the interval $[0, 1]$, where patents using the exact same words in the same proportion have

similarity of one and patents with no common terms have similarity of zero.

Following Kalyani (2024), the second measure of patent novelty focuses solely on the introduction of new technical terms. It converts the text of patent filings into bigrams—word pairs such as "phase transition" or "thermal conductivity." After removing non-technical terms, it defines the novelty of a patent as the share of bigrams that do not appear in any previous patent. As show by Kalyani (2024), this is a leading indicator of transformative technologies. In my sample, patents contain an average of 544 technical bigrams. Out of these, 6.28% are classified as new terms. Appendix B.3 details the construction of the patent similarity measures.

Products To assess downstream innovation, I gather data on products incorporating newly discovered materials. I observe how the material is used and whether the product represents a new line or an improvement to an existing one. I measure product innovations based on the creation of a prototype that incorporates the new compound. Due to the lag between prototyping and market release, the products in my analysis are not yet mass-produced or available to consumers.

Scientist Activities I observe rich text data on scientists' work. Required for administrative purposes, researchers meticulously log their tasks and the duration of each activity. On average, scientists record 7.8 entries per week, totaling more than 1.6 million observations over the four year sample period.

To transform this textual information into a meaningful quantitative metric, I separate the materials discovery process into three categories of tasks: idea generation, judgment, and experimentation. Idea generation encompasses activities related to developing potential compounds, such as reviewing the literature on existing materials or creating preliminary designs. Judgment tasks focus on selecting which compounds to advance, often involving the analysis of simulations or predicting material characteristics based on domain knowledge. Finally, experimentation tasks are dedicated to synthesizing new materials and conducting tests to evaluate their properties. Based on the materials science literature and discussions with the lab's scientists, these categories divide the discovery process into meaningful, high-level groups.

I employ a large language model—Anthropic's Claude 3.5—to classify scientists' activities into these three categories (with a fourth for all other tasks). This enables me to quantify the amount of time scientists dedicate to each stage. I fine-tune the model using manually classified training data, improving its performance substantially. I validate my classification procedure using out of sample predictions and a survey of the lab's scientists, confirming that the model accurately categorizes research activities. Appendix B.4 provides further details on the text data, classification process, and validation.

Scientist Characteristics I also observe information on scientists' education, demographics, employment history, academic publications, and past discoveries. I use these as control variables and to assess treatment group balance.

Other R&D Inputs Finally, I use data on the lab's input costs to construct a measure of R&D efficiency. This includes expenditures on scientist wages, lab equipment, and raw materials, as well as training and inference costs for the model.

3.2 Survey of Scientists

I supplement these data by conducting a survey of the lab's scientists. The survey has three objectives: first, to understand how researchers collaborate with the model and determine why some are more skilled at evaluating AI suggestions; second, to validate the task allocation measures; and third, to assess the tool's impact on scientists' job satisfaction and beliefs about artificial intelligence. I match respondents to the researchers in my panel.

I sent the 15-minute survey to all 1,018 scientists. To incentivize participation, it states that the results will inform the lab's plans regarding future AI use. I received 447 responses (44%). This response rate is relatively high—the median paper in a top economics journal has a response rate of 35% (Dutz et al., 2022) and for surveys of scientists it can be below 10% (Hill and Stein, 2024; Myers et al., 2024). Appendix Table A12 reports the characteristics of respondents. They are representative in terms of educational background, material of focus, and past performance, but skew slightly younger and less experienced. Appendix Table A13 shows the number of responses to each survey question. While there is some attrition, more than two thirds of respondents answered all questions.

I fielded the survey in May and June of 2024, after all scientists had gained access to the model. Consequently, the results should be interpreted as descriptive. Additionally, questions about changes over time rely on respondents' best recollection. The survey instruments are provided in Appendix C.

3.3 Sample Selection and Summary Statistics

My panel covers the period between May 2020 and June 2024, including two years of pre-treatment observations and 25 months after the start of the experiment. I restrict the analysis to ever-treated scientists who remain at the firm during the entire period. I also remove new hires and the small number of researchers who switched teams or roles. Attrition rates are low and do not vary with treatment status. I consider only scientists who focus primarily on materials discovery, excluding

lab technicians, administrators, and managers. Using the task data, I verify that all researchers in my sample indeed perform functions related to materials discovery.

The final sample includes 1,018 scientists across 221 teams. Table 1 reports descriptive statistics on these researchers' age, education, tenure at the firm, material specialization, and innovative output. Scientists are highly educated, with more than 60% holding a PhD, and have worked at the firm for an average of 7.1 years. They are divided roughly evenly between four material categories: biomaterials (including synthetics), ceramics and glasses, metals and alloys, and polymers. The typical researcher produces 0.56 new materials, 0.16 patent filings, and 0.20 product prototypes each year. Discovery rates vary somewhat by material type, with biomaterials having the highest rates and metals the lowest.

3.4 Research Design

The random assignment of research teams to adoption waves provides a favorable setting for identification. I therefore estimate simple two-way fixed effects specifications at the team level:

$$\text{Static:} \qquad y_{it} = \alpha_i + \omega_t + \delta D_{i,t-l} + \varepsilon_{it} \tag{1}$$

$$\text{Event Study:} \qquad y_{it} = \alpha_i + \omega_t + \sum_{k \in [\underline{k}, \bar{k}]} \delta_k D_{it-k} + \varepsilon_{it} \tag{2}$$

where y_{it} is the outcome of interest, D_{it} is a treatment indicator, and α_i and ω_t are team and month fixed effects. The static specification includes a lagged structure to capture the treatment effect once the full impact of AI has been realized.

Randomization ensures strong ignorability, so under a stable unit treatment value assumption δ captures the average treatment effect. To confirm that teams were indeed randomly assigned to treatment groups, Table 1 reports balance tests for the three waves, showing no systematic differences. My main specifications include unit and time fixed effects, but the results are robust to adding covariates or excluding the fixed effects. To estimate individual treatment effect heterogeneity, I run scientist-level versions of these specifications, interacting D_{it} with researcher characteristics.[8]

Ordinary least squares estimates of Equations (1) and (2) may be biased if treatment effects vary over time or across adoption cohorts. The results are similar using the unbiased estimators proposed by de Chaisemartin and D'Haultfœuille (2020), Callaway and Sant'Anna (2021), Sun

[8]Because scientists within a team are treated simultaneously and have correlated outcomes, these individual effects are less well-identified. I therefore also assess treatment effect heterogeneoity at the team level. The results are similar but more muted, since teams display less variation in research productivity than individual scientists

and Abraham (2021), and Borusyak et al. (2024). Given the count nature of my data, I also employ Poisson and Negative Binomial regressions, again yielding nearly identical estimates.[9]

I conduct inference in two ways. In my main results, I report standard t-based confidence intervals, clustering at the team level. Following the econometrics literature on experiments (Athey and Imbens, 2017; Abadie et al., 2020), I also take a design-based approach and perform permutation tests. All findings are robust to the choice of inference strategy.[10]

Table 1. Summary Statistics and Treatment Group Balance

	Full Sample (All Periods)	Wave 1 (Pre-Treatment)	Wave 2 (Pre-Treatment)	Wave 3 (Pre-Treatment)	Normalized Difference
Scientist Characteristics					
Age (Years)	45.78	45.75	45.88	45.60	0.04
Tenure at Firm (Years)	7.12	7.25	6.79	7.06	0.06
Share Bachelor's Degree	0.06	0.07	0.04	0.05	0.06
Share Master's Degree	0.33	0.30	0.33	0.40	0.06
Share Doctoral Degreee	0.61	0.63	0.63	0.56	0.05
Share Chemical Engineers	0.22	0.21	0.23	0.19	0.06
Share Chemists	0.14	0.13	0.14	0.15	0.03
Share Materials Scientists	0.23	0.25	0.22	0.24	0.04
Share Physicists	0.10	0.07	0.11	0.15	0.06
Share Other Fields	0.31	0.34	0.30	0.26	0.05
Material Specialization					
Share Biomaterials	0.25	0.24	0.26	0.26	0.04
Share Ceramics and Glasses	0.19	0.19	0.19	0.21	0.05
Share Metals and Alloys	0.32	0.34	0.34	0.20	0.04
Share Polymers	0.24	0.22	0.21	0.33	0.04
Innovative Output					
New Materials (Yearly)	0.56	0.45	0.40	0.44	0.05
Patent Filings (Yearly)	0.16	0.14	0.09	0.14	0.06
Product Prototypes (Yearly)	0.20	0.17	0.12	0.17	0.04
Sample Size					
Number of Teams	221	89	91	41	–
Scientists per Team	4.61	4.53	4.60	4.78	0.03

Notes: This table presents summary statistics for the scientists in my sample. The first column reports means for the full sample, calculated between May 2021 and June 2024. The middle three columns show pre-treatment means separately by adoption wave, which comprise 89, 91, and 41 teams, respectively. The final column reports the maximum pairwise normalized difference between wave-specific means, calculated as: $|\bar{X}_1 - \bar{X}_2| / \sqrt{(s_1^2 + s_2^2)/2}$.

[9] Results using these alternative estimators are presented in Appendix Tables A6, A7, and A8

[10] Results using randomization inference are reported in Appendix Table A4.

4 The Impact of AI on Scientific Discovery and Innovation

4.1 New Materials, Patent Filings, and Product Prototypes

I begin by presenting descriptive evidence of AI's impact on materials discovery, patenting, and product innovation. Figure 4 plots time series of the three outcomes for treated and untreated scientists, revealing a marked rise in new compounds and patent filings after AI-adoption. Ten to twelve months later, this is followed by an increase in product prototypes incorporating the discovered compounds.

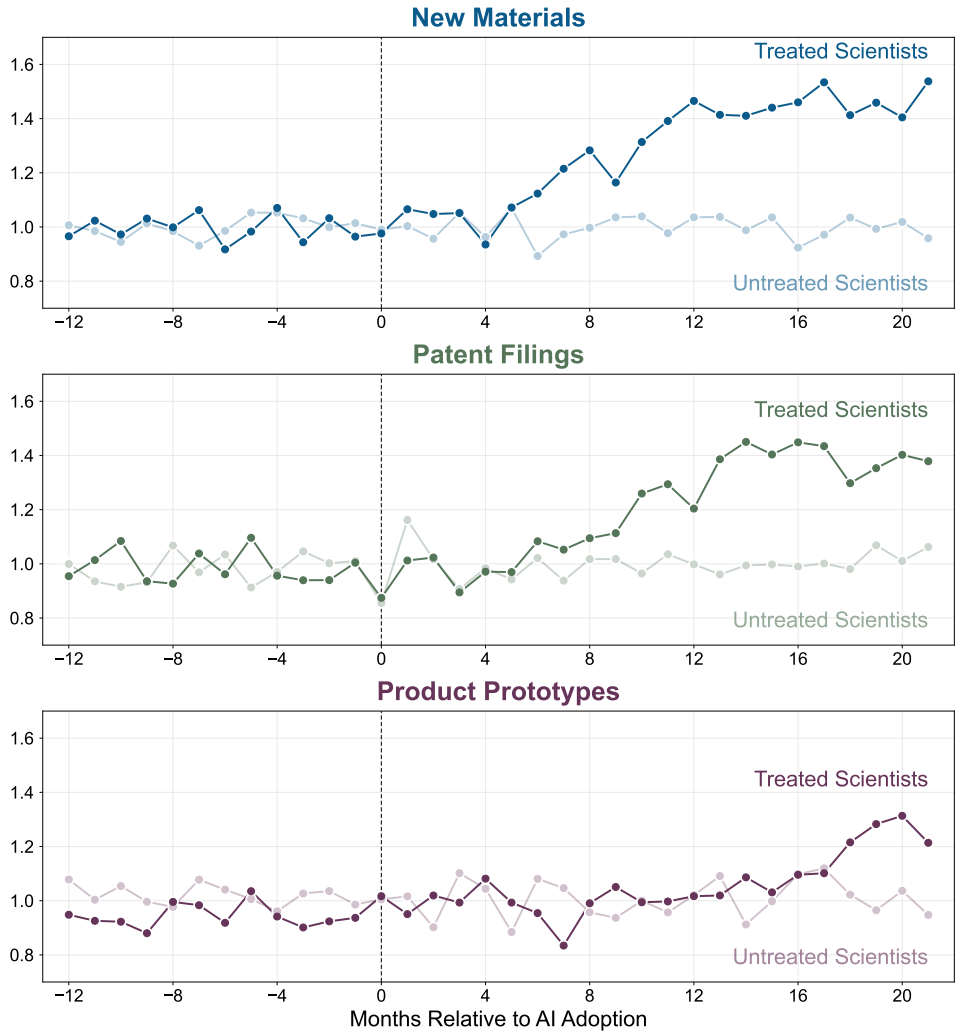


Figure 4. Time Series of Innovation Rates in Treatment and Control Groups

Notes: This figure plots monthly time series of new materials, patent filings, and product prototypes for treated and untreated scientists. In the pre-period, the "untreated" group includes scientists from the third wave (treated in month 12), while the "treated" group includes waves 1 and 2. To facilitate comparison, I normalize the levels of the variables so that they average 1 in the pre-period. This series is in months relative to AI adoption.

Next, I turn to the regression estimates. Figure 5, Panel A reports endline treatment effects for the last five months of the sample. On average, AI-assisted scientists discover 44% more materials, leading to a 39% increase in patent filings and a 17% rise in product prototypes. To investigate dynamics, Panel B plots event study estimates. The results show a similar pattern to the raw time series: the effects on materials discovery and patenting emerge after 5-6 months, while the impact on product innovation lags by more than a year. Both specifications include team and month fixed effects, and I report 95% confidence intervals based on standard errors clustered at the team level. Appendix Tables A1–A5 show that these findings are robust to the inclusion of covariates and the use of alternative estimators.

These effects are large. To put the rise in materials discovery in perspective, the lab's research output per scientist declined by 4% over the preceding five years. This was despite the introduction of several computational tools designed to aid scientists. AI therefore appears to be a different class of technology, with impacts that are orders of magnitude greater than previous methods.

While still representing a substantial increase in innovation, the downstream effect on product prototypes is about half as large. This could reflect several factors. First, the lag between discovery and product development means that some new materials may have yet to be integrated into prototypes. Second, as I show in Section 4.3, AI increases the share of products that represent new lines rather than improvements to existing ones. The smaller effect could thus reflect a compositional shift toward more challenging inventions. Finally, prototyping may simply be a bottleneck, partially attenuating the impact of faster materials discovery.

I conduct a number of exercises testing the robustness of these findings. One scenario that would invalidate my research design is if teams within the firm race to make discoveries. In this case, AI would impact innovation rates for untreated scientists in the opposite direction, leading to upward bias in the estimated treatment effect. While time series evidence on discoveries in the treatment and control groups shows no signs of such spillovers (see Figure 4), I devise an identification strategy robust to the presence of racing. For each treated team, I construct a control group composed of untreated teams not working in the same subspecialty. I implement this design by estimating subspecialty-specific treatment effects that are aggregated into an average impact. Shown in Appendix Table A5, these estimates are nearly identical to the coefficients using the full sample, suggesting that racing dynamics are not a concern in my setting.

Furthermore, the effects are not driven by differential investment. After accounting for the cost of the AI tool, R&D spending per material is similar between treated and not-yet-treated teams. Finally, the results do not reflect a change in the types of compounds scientists focus on.

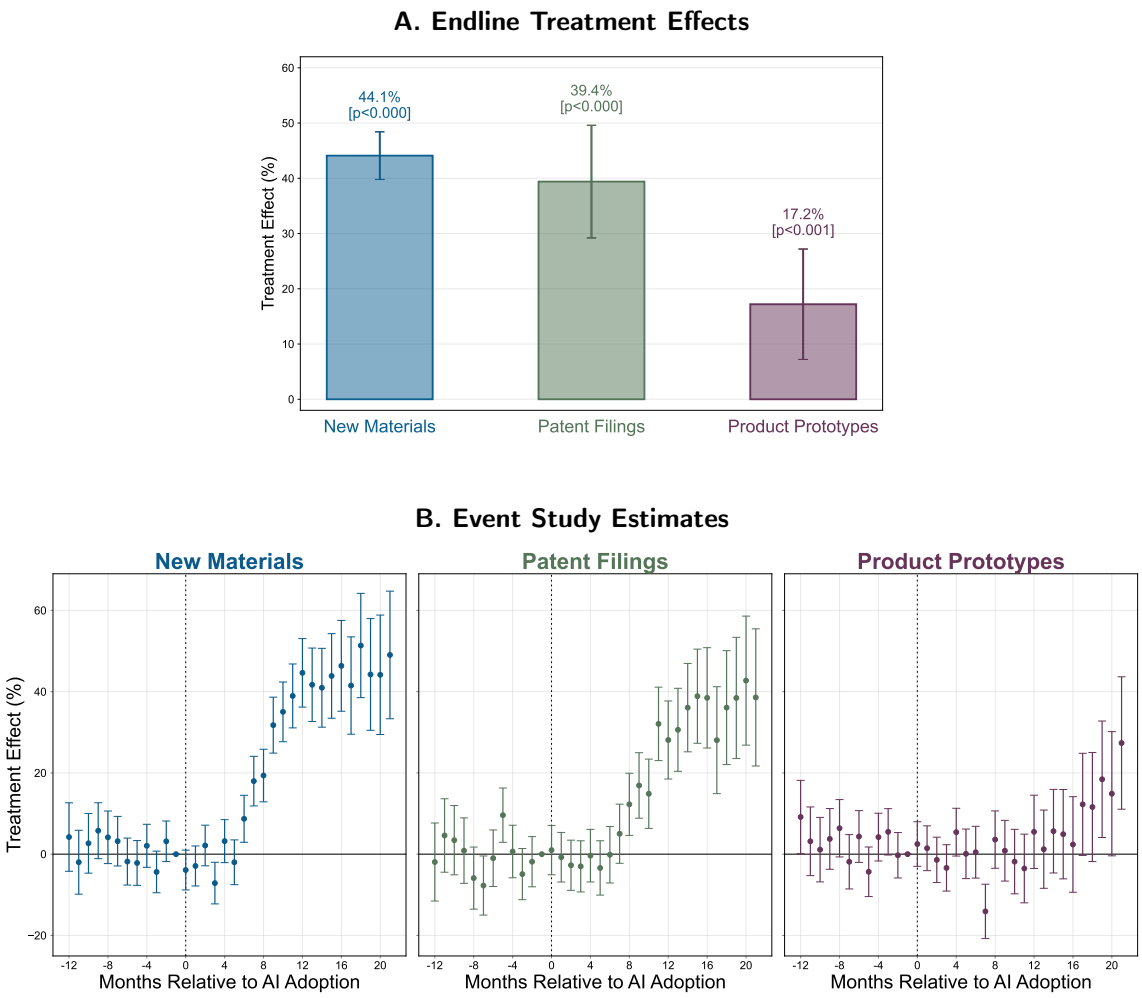


Figure 5. Impact of AI on Materials Discovery, Patent Filings, and Product Prototypes

Notes: This figure shows the impact of AI on new materials, patent filings, and product prototypes. Panel A plots endline treatment effects for the last 5 months of the sample. These are estimated at the team-level using Equation (1), where the outcomes are monthly counts. Panel B present event study estimates, following Equation (2). Both specifications include team and month fixed effects and report 95% confidence intervals based on standard errors clustered at the team level. The coefficients are converted to percentage terms based on pre-treatment means.

4.2 Material Quality

AI increases the number of new compounds. However, it could be the case that it simultaneously lowers material quality. Indeed, human science appears to exhibit such a speed-quality tradeoff (Hill and Stein, 2024). If so, my estimates would overstate the tool's impact. While the results on patents and products help mitigate this concern, I use material properties to test quality directly. As described in Section 3, I construct three quality indices based on the distance between scientists' target characteristics and compounds' observed properties.[11]

[11] Appendix B.1 provides details on the construction of the quality measures.

Table 2 shows AI's impact on these measures. For atomic properties, the tool increases average quality by 13% and raises the proportion of top-decile materials by 1.7 percentage points (Columns 1-2). Large-scale characteristics see a similar but slightly smaller effect (Columns 3-4). Columns (5) and (6) combine both sets of properties into an overall index, showing statistically significant improvements in average quality (9%) and the proportion of high-quality materials (1.5 pp).

The indices combine several characteristics that may vary in importance to the firm, making it difficult to interpret the magnitude of these estimates. However, the results suggest that AI-assisted materials discovery does not compromise quality.

Table 2. Impact of AI on Material Quality

	Atomic Properties Index		*Large Scale Properties Index*		*Overall Quality Index*	
	Average	Share Top 10%	Average	Share Top 10%	Average	Share Top 10%
	(1)	(2)	(3)	(4)	(5)	(6)
Access to AI	0.066***	0.017*	0.034**	0.0014**	0.045***	0.015**
	(0.032)	(0.009)	(0.003)	(0.008)	(0.004)	(0.008)
Month Fixed Effects	✓	✓	✓	✓	✓	✓
Team Fixed Effects	✓	✓	✓	✓	✓	✓
Material Type Fixed Effects	✓	✓	✓	✓	✓	✓
Number of Scientists	1,018	1,018	1,018	1,018	1,018	1,018
Number of Teams	221	221	221	221	221	221

Notes: This table presents the impact of AI access on three material quality metrics. For each metric, I report the average effect and the impact on the share in the top 10% of the distribution. Columns (1) and (2) present the results for atomic properties, columns (3) and (4) for large scale properties, and columns (5) and (6) for an overall index that combines the two. The construction of these measures is described in Appendix B. Standard errors in parentheses are clustered at the team level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

4.3 Novelty

While AI accelerates the pace of innovation, the consequences hinge on the nature of these breakthroughs. Would they have occurred anyway, just at a later date? I present three pieces of evidence that suggest the answer is no. Instead, the tool engenders a shift toward more radical innovation, increasing novelty in all three stages of R&D.

I first measure the novelty of new materials themselves, following the chemical similarity approach described in Section 3. Shown in Table 3, Column (1), AI decreases average similarity by 0.4 standard deviations. Furthermore, it increases the share of highly distinct materials—defined as those in the bottom quartile of the similarity distribution—by 4 percentage points (Column 2). I corroborate these measurements using the survey of scientists. Consistent with my estimates, 73% of researchers report that the tool generates more novel designs than other methods.

While chemical similarity captures a key aspect of scientific novelty, it is important to determine whether more original materials lead to more innovative technologies. To do so, I analyze the textual similarity of patent filings. As described in Section 3, I calculate two similarity metrics. The first is based on the full text of filings and the second on the share of new technical terms. Shown in Column (3) of Table 3, the tool boosts novelty on the first measure by 11%, shifting the average filing from the 48th to 42nd percentile of the similarity distribution. On the second measure, reported in Column (4), AI increases the share of new technical terms—a leading indicator of transformative technologies—by two percentage points (22%).

Finally, I study the tool's impact on the nature of product innovations. In the absence of AI, scientists focus primarily on improvements to existing products, with only 13% of prototypes representing new lines. As shown in Column (6) of Table 3, this share rises by 2 percentage points in the treatment group (18%). This suggests that the firm has a pipeline of product ideas, but a lack of viable materials represents a barrier to their creation. By relaxing this constraint, AI enables a shift toward more radical innovation.

A central concern with applying machine learning to scientific discovery is its potential to amplify the "streetlight effect"—the tendency to search where it is easiest to look rather than where the best answer is likely to be. This worry stems from models being trained on existing knowledge, potentially directing search toward familiar but less promising areas (Khurana, 2023; Kim, 2023; Hoelzemann et al., 2024). In other words, AI might inefficiently favor exploitation over exploration. In materials science, this appears not to be the case. Instead, the tool increases the novelty of discoveries, leading to more creative patents and more innovative products.

The fact that AI increases novelty can be interpreted in two ways. One possibility is that the model is simply adept at generalization, exploring new parts of the materials design space. The extent to which neural networks can generalize is an open question, but recent evidence suggests that such capabilities are possible (Zhang et al., 2024). Alternatively, this finding could primarily reflect human limitations in the absence of AI, with scientists adhering even more closely to familiar templates.

There are three caveats to these results. First, they do not rule out the possibility that AI increases novelty within the neighborhood of what is known but reduces the likelihood of truly revolutionary discoveries. This connects to a broader concern about the quantitative study of science: if progress is driven by outlier breakthroughs, empirical analyses may be misleading due to the inherent scarcity of right-tail discoveries (Nielsen and Qiu, 2022). Second, I do not observe the ultimate impact of the lab's inventions. I track patent filings but lack data on citations or subsequent innovations. Similarly, the utility of the new products remains uncertain, as they are

not yet available to consumers. Third, AI's impact on materials discovery may face diminishing returns. While I see no signs of this pattern over my sample period, it could arise later on.

Table 3. Impact of AI on Novelty

	Material Similarity (Mean) (1)	Material Similarity (Top 25%) (2)	Patent Similarity (Full Text) (3)	Patent Similarity (New Terms) (4)	Share New Product Lines (5)
Access to AI	0.138***	0.042**	0.015**	0.021***	0.030*
	(0.032)	(0.02)	(0.006)	(0.005)	(0.02)
Pre-Treatment Means	0.523	0.250	0.136	0.092	0.132
Month Fixed Effects	✓	✓	✓	✓	✓
Team Fixed Effects	✓	✓	✓	✓	✓
Number of Scientists	651	651	1,018	1,018	1,018
Number of Teams	145	145	221	221	221

Notes: This table shows the impact of AI on novelty. Columns (1) and (2) presents the results for material novelty, which I measure using the similarity approach of De et al. (2016). This method applies only to the 64% of crystal structures in my sample. Column (1) reports the change in average similarity and Column (2) shows the share in the top 25% of similarity based on a cutoff defined in the pre-period. Columns (3) and (4) report the estimates for patent novelty using the textual similarity metrics of Kelly et al. (2021) and Kalyani (2024), respectively. Column (6) shows the tool's impact on the share of product prototypes that are new lines (rather than improvements to existing lines). I report pre-treatment means for each outcome, calculated between May 2021 and May 2022. Appendix B describes the construction of the novelty measures. Standard errors in parentheses are clustered at the team level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

4.4 R&D Efficiency

Finally, I combine my estimates with data on input costs to calculate the tool's impact on R&D efficiency. Using number of product prototypes as the outcome, I define efficiency as:

$$\text{R\&D Efficiency} = \frac{\text{Product Prototypes}}{\text{Labor Costs} + \text{Other Variable Costs} + \text{Amortized Fixed Costs}}$$

Labor costs include salaries and benefits for the lab's scientists, technicians, and other staff. While base salary is identical across treatment waves, bonuses are larger for treated scientists due to increased productivity. Non-labor variable costs are also higher in the treatment group, due to increased spending on testing, product development, and model inference. Importantly, fixed costs include expenditure on training the model, which I conservatively amortize over a 5-year period. The results are reported in Appendix Table A9. Depending on the specific set of costs included, I find that AI boosts R&D efficiency by 13-15 percent.

5 Heterogeneity by Scientist Ability

AI increases the average rate of materials discovery. However, I now document substantial heterogeneity in its effects. The tool disproportionately benefits scientists with high initial productivity, exacerbating inequality. Decomposing productivity into distinct skills, I show that differences in judgment—the ability to identify promising candidate compounds—explain nearly all of this heterogeneity.

5.1 The Impact of AI Across the Productivity Distribution

Figure 6, Panel A plots the distribution of materials discovery rates before and after the introduction of AI. The distribution shifts to the right and becomes more right-skewed, suggesting that high-ability scientists gain more from the tool. To investigate this systematically, I construct a measure of initial productivity based on materials discovered in the pre-treatment period. I control for material type and application to account for the possibility that some compounds are inherently easier to create. Additionally, as discussed in Section 3, all scientists in my sample work primarily in materials discovery roles, so variation in productivity does not reflect a different area of focus. The resulting metric is positively correlated with the ranking of scientists' graduate institutions as well as tenure in the lab, suggesting that it indeed captures a notion of ability. Moreover, discovery rates are persistent over time, showing that they reflect skill rather than luck.

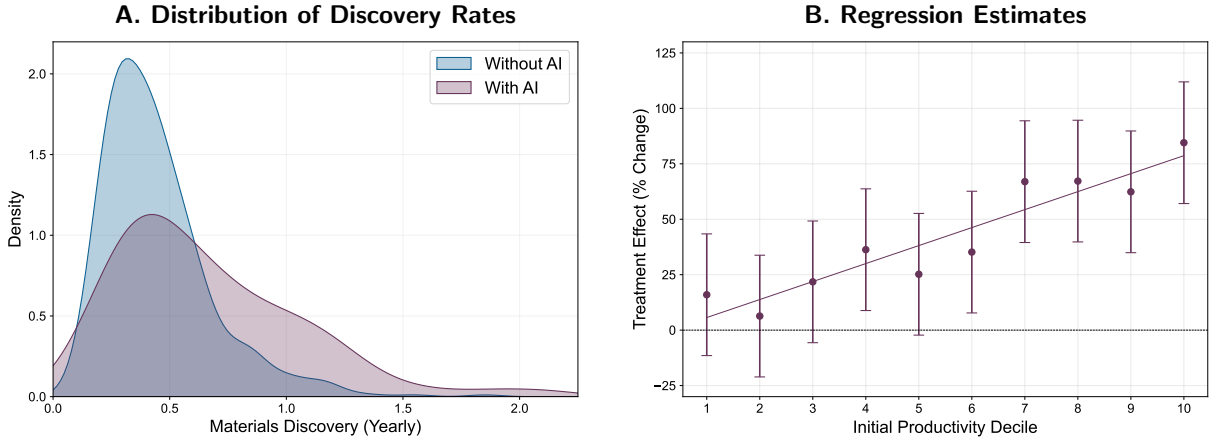


Figure 6. Impact on Materials Discovery Across the Productivity Distribution

Notes: Panel A shows the distribution of yearly materials discovery rates for scientists with and without AI. Panel B shows treatment effects by decile of initial productivity. These estimates come from interacting a treatment indicator with dummy variables for a scientist being in each decile. The outcome is the number of discovered materials, and the coefficients are reported in percentage terms based on pre-treatment means. I report 95% confidence intervals, based on standard errors clustered at the team level.

Figure 6, Panel B presents regression estimates, interacting treatment status with deciles of initial productivity. While the bottom third of researchers see minimal benefit from the tool, the output of top-decile scientists increases by 81%.[12] As a result, the ratio of 90:10 research performance more than doubles. The fact that the tool boosts inequality is in contrast to recent evidence on large language models (Brynjolfsson et al., 2023; Noy and Zhang, 2023; Peng et al., 2023), showing that the distributional consequences of AI depend on the context and type of technology. Appendix Figure A3 reports analogous results for patenting and product prototypes, revealing a similar pattern for downstream innovation.

These findings suggest that AI and expertise are complements in the "scientific production function" (Porter and Stern, 2000). The rest of the paper investigates the source of this complementarity.

5.2 Heterogeneity by Dimension of Skill

Why do highly productive scientists benefit more from AI? As described in Section 2, materials discovery involves three sets of tasks: idea generation, judgment, and experimentation. Thus, differences in productivity reflect scientists' varying abilities in each phase. I now show that skill in a single step—judgment—explains nearly all of the tool's heterogeneous impact.

5.2.1 Estimating Task-Specific Research Abilities

I begin by estimating each scientist's task-specific research ability in the pre-treatment period. Because the experimentation phase consists solely of routine tests, I focus on idea generation and judgment. The ideal experiment to identify these measures would assign a random set of scientists to design candidate materials, another to evaluate them, and then measure the outcomes. Lacking explicit randomization, I exploit the division of labor within teams. The set of scientists assigned to work on a given material—as well as the specific research tasks they perform—varies significantly across compounds. Combined with two key assumptions, this variation allows me to identify task-specific skills.

Let γ_s^g and γ_s^j represent scientist s's skill in idea generation and judgment tasks, respectively. Suppose the probability that candidate material m is viable is given by:

$$\mathbb{P}\left(\text{Viable}\right)_m = \frac{1}{|\,S^g\,|}\sum_{s\in S^g}\gamma_s^g + \frac{1}{|\,S^f\,|}\sum_{s\in S^f}\gamma_s^j + \theta' W_m + \varepsilon_m \tag{3}$$

where S^g and S^f are the sets of scientists that work on idea generation and judgment tasks for

[12] Appendix Figure A2 presents these estimates at the team level, showing a similar but more muted pattern.

material m, respectively. This equation states that expected material viability reflects average skill in idea generation for scientists working on the design step, average skill in judgment tasks for scientists working on the prioritization step, and material-specific controls, W_m. Variation in S^g and S^f across compounds enables me to estimate γ_s^g and γ_s^j for each scientist.[13]

Two assumptions are critical in this design. First, it must be the case that S^g and S^f are orthogonal to ε_m, conditional on W_m. In other words, the set of scientists assigned to work on a particular material are unrelated to its latent viability. This assumption would be violated, for instance, if researchers with low idea generation ability were assigned to materials that are easier to design, even after controlling for observables. Two features of my setting help mitigate this concern. First, scientists within a team are assigned to materials through a queue-based system, so material allocation is primarily determined by availability. Additionally, I observe detailed information on the characteristics of compounds, allowing me to include rich controls for material type and intended application. The second identifying assumption is that γ_s^g and γ_s^j are additively separable. This rules out match effects between scientists working on the same material.

I conduct several tests to validate my ability measures. First, I find that $\hat{\gamma}_s^j$ and $\hat{\gamma}_s^g$ are autocorrelated, suggesting that the estimates reflect underlying skill rather than chance. Second, I document a relationship between academic background and task-specific ability. For a subset of scientists, I observe their journal publications and doctoral dissertations. Scientists who study topics related to a specific research task tend to be especially skilled in that area. For instance, a scientist who developed a novel simulation technique for predicting material properties is more likely to have an advantage in judgment tasks. Third, I observe that researchers are more likely to specialize in tasks where they have a comparative advantage. Finally, I corroborate the estimates with survey data. I ask each scientist to rate their skills in idea generation and judgment on a 1-10 scale and compare their abilities to their peers. These self-assessments align with my empirical findings, in both absolute and relative terms. Indeed, all the results are directionally consistent if I simply use researchers' self-reports rather than estimated ability.

5.2.2 Correlation Between Dimensions of Skill

Figure 7 plots the correlation between scientists' skill in idea generation and judgment. The two measures are positively correlated ($r = 0.42,\ p < 0.00$). This suggests that scientists possess some form of underlying expertise that makes them productive in both sets of tasks. However, the corre-

[13] Given its high-dimensional nature, I estimate Equation (3) using an empirical Bayes approach (Chetty et al., 2014; Kline et al., 2022). Specifically, I assume that the underlying distribution task-specific ability is normal and shrink the estimated fixed effects toward the mean of this distribution.

lation is well below one, revealing a potential role for comparative advantage and specialization. This underscores the importance of moving beyond a one-dimensional notion of "skill-bias" to uncover the specific abilities that are complemented by AI.

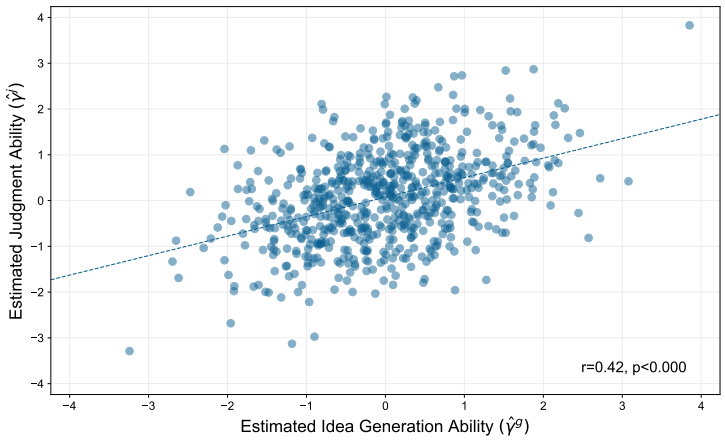


Figure 7. Correlation Between Research Skills

Notes: This figure shows the relationship between scientists' estimated skill in idea generation and judgment tasks, plotting $\hat{\gamma}^g$ against $\hat{\gamma}^j$ for each researcher. These measures are based on the fixed effects specification in Equation 3. The correlation coefficient is 0.42 (p<0.00). The blue dashed line represents the line of best fit.

5.2.3 The Impact of AI by Dimension of Skill

Armed with estimates of task-specific research ability, I examine which skills drive the heterogeneous impact of AI. To do so, I estimate the following regression at the scientist level:

$$y_{st} = \beta_1 D_{st} + \beta_2 \hat{\gamma}_s^g + \beta_3 \hat{\gamma}_s^j + \beta_4 \left(D_{st} \times \hat{\gamma}_s^g \right) + \beta_5 \left(D_{st} \times \hat{\gamma}_s^j \right) + \varepsilon_{st} \tag{4}$$

where y_{st} is the number of materials discovered by scientist s in month t, D_{st} is a treatment indicator, and $\hat{\gamma}_s^g$ and $\hat{\gamma}_s^j$ are estimated research skill in idea generation and judgment tasks, respectively. These measures are normalized to have mean zero and standard deviation one. The coefficients of interest are β_4 and β_5, capturing the differential impact of AI by task-specific skill.

The results are reported in Table 4. The coefficients on both interaction terms are positive and significant, but judgment has a substantially larger effect. A one standard deviation increase in $\hat{\gamma}^j$ corresponds to a 14.8 percentage point rise in the AI treatment effect. Meanwhile, an identical change in $\hat{\gamma}^g$ yields only a 3.5 percentage point increase. Consequently, differences in judgment explain more than 80% of the tool's heterogeneous impact by initial productivity.

These findings demonstrate the central role of judgment in explaining AI's disparate effects across scientists. In the next section, I explore why this is the case.

Table 4. The Impact of AI by Dimension of Research Ability

	Number of Discovered Materials (1)	Number of Discovered Materials (2)	Number of Discovered Materials (3)
Access to AI	0.0015	0.0018	0.0020
	(0.0145)	(0.0178)	(0.0212)
$\hat{\gamma}^g$	0.0158***	0.0160***	0.0163***
	(0.0025)	(0.0028)	(0.0032)
$\hat{\gamma}^j$	0.0218***	0.0221***	0.0225***
	(0.0025)	(0.0028)	(0.0032)
Access to AI $\times\hat{\gamma}^g$	0.0350***	0.0362***	0.0365***
	(0.0106)	(0.0110)	(0.0113)
Access to AI $\times\hat{\gamma}^j$	0.1480***	0.1351***	0.1434***
	(0.0106)	(0.0111)	(0.0114)
Month Fixed Effects	✓	✓	✓
Material Type FE		✓	✓
Application FE			✓
Number of Scientists	1,018	1,018	1,018

Notes: This table reports heterogeneous treatment effects based on scientists' abilities in idea generation and judgment tasks. Following Equation (4), I interact treatment status with pre-period estimates of idea generation ability ($\hat{\gamma}^g$) and judgment ($\hat{\gamma}^j$). These measures come from the fixed effects specification in Equation (3). They are estimated in the pre-treatment period and are constant for each scientist. Column (1) includes only time fixed effects, Column (2) adds material type fixed effects, and Column (3) adds fixed effects for intended application. Standard errors in parentheses are clustered at the scientist level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

6 Scientist-AI Collaboration

To summarize, I have established three facts. First, AI substantially increases the average rate of materials discovery. Second, it disproportionately benefits researchers with high initial productivity. Third, this heterogeneity is driven almost entirely by differences in judgment. To understand the mechanisms behind these results, I investigate the dynamics of human-AI collaboration in science.

I begin by documenting a dramatic change in the content of scientists' work (Section 6.1). AI automates a majority of idea generation tasks, reallocating researchers to the new task of evaluating model-produced candidate materials. Next, I develop a simple prioritized search framework to analyze the implications of this shift (Section 6.2). Taking the model to the data, I find that scientists with strong judgment learn to prioritize promising AI suggestions, while others waste significant resources testing false positives. The resulting gap in discovery rates explains the tool's heterogeneous impact. Turning to the survey, I present evidence suggesting that differences in judgment are driven by domain knowledge. Finally, I show that the lab responds to the changing research process by adjusting its employment practices (Section 6.3). After the conclusion of the

experiment, it redesigned its hiring and firing criteria to favor scientists with strong judgment. My estimates may therefore understate AI's longer-run impact due to organizational adaptation.

6.1 Tasks, Automation, and Reallocation

Figure 8, Panel A plots the share of scientists' research hours dedicated to idea generation, judgment, and experimentation tasks, revealing a substantial reorganization of the discovery process. These measures come from logs of scientist activities, employing the text classification procedure described in Section 3. Without AI, scientists allocate 39% of their time to idea generation tasks. This drops to less than 16% after the model's introduction. In contrast, while judgment tasks initially take up 23% of research hours, they consume 40% by the end of the sample. Finally, the fraction of time dedicated to experimentation tasks increases from 37 to 44 percent. The total number of research hours remains unchanged. Appendix Figure A4 reports event study estimates for these task shares, showing an identical pattern.

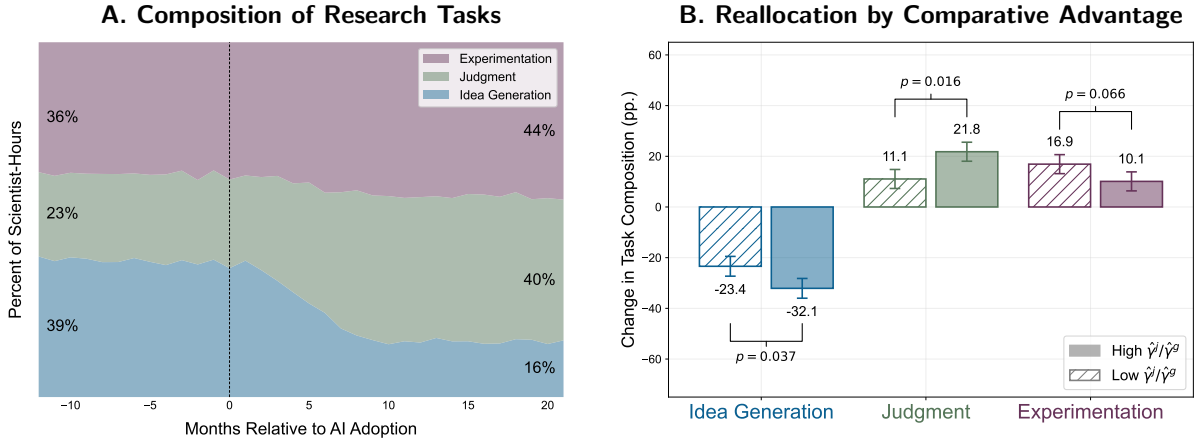


Figure 8. Impact of AI on the Composition of Scientists' Research Tasks

Notes: This figure shows the impact of AI on the composition of scientists' research tasks. Panel A plots the share of research hours dedicated to idea generation (in blue), judgment (in green), and experimentation tasks (in purple). The shares are normalized to exclude tasks not in these three categories. The total number of hours spent on these three categories of tasks is unchanged after the introduction of AI. The x-axis shows months relative to treatment. The shares come from the text classification procedure described in Section 3. Panel B shows the percentage point change in task allocation for scientists in the bottom quartile (hatched bars) and top quartile (solid bars) of comparative advantage in judgment tasks. These measures are estimated in the pre-period using the fixed effects approach described in Section 5.2. The vertical lines represent 95% confidence intervals based on standard errors clustered at the scientist-level and I report p-values for the difference in treatment effects between the two quartiles.

Panel B reports the change in task composition for scientists in the top and bottom quartiles of comparative advantage in judgment, $\hat{\gamma}_s^j/\hat{\gamma}_s^g$. While all researchers significantly adjust their time allocation, those especially skilled in evaluation display a 46% larger shift from idea generation to judgment tasks. Additionally, these scientists see a smaller increase in time spent testing materials.

In Appendix D, I show that these patterns are consistent with a task framework where scientists are assigned to roles based on their comparative advantage and AI automates a fraction of idea generation tasks (Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2022).

I corroborate these results using the survey of scientists. Consistent with the task allocation measures, scientists report initially allocating 36% of their time to idea generation, 26% to judgment, and 42% to experimentation. After the introduction of AI, these shares become 13%, 35%, and 47%, respectively. Matching survey responses to the empirical estimates, I confirm that the shift from idea generation to judgment tasks is increasing in $\hat{\gamma}_s^j / \hat{\gamma}_s^g$.

These results provide granular evidence on automation and reallocation in science. The tool automates a majority of material design tasks, while scientists are reallocated to the new task of evaluating model-generated compounds. AI is thus labor-replacing in the specific activity of creating candidate materials, but labor-augmenting in the broader discovery process due to its complementarity with judgment tasks. Indeed, scientists' total research hours are unchanged after the introduction of AI, revealing that demand for labor remains strong. While the equilibrium implications for employment and earnings are beyond the scope of this paper, my findings highlight the importance of considering task reallocation when assessing the implications of AI for workers.

6.2 The Central Role of Human Judgment

Motivated by these changes to the research process, this section develops a simple framework to study the effect of partially-automated compound design on the rate of materials discovery. I begin by outlining the model and defining the key objects that I measure empirically. Taking it to the data, I show that scientists' differential ability to evaluate AI-generated compounds explains the tool's heterogeneous impact. Finally, I present survey evidence suggesting that these differences in judgment are driven by domain knowledge.

6.2.1 A Model of Materials Discovery as Prioritized Search

Following Agrawal et al. (2023), I conceptualize materials discovery as a prioritized sequential search problem. The research process requires three steps:

1. **Idea Generation:** Generate a set of candidate materials $\mathcal{M} = \{m_1, \ldots, m_M\}$. Each candidate has binary quality $\theta_m \in \{0, 1\}$, which depends on research inputs but is unobserved by the scientist. Let $\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \theta_m$ denote the share of high-quality candidates.

2. **Judgment:** Evaluate candidate materials and rank them in descending order of predicted

quality. Predictions improve with the scientists' judgment, bringing the ranking $r : \mathcal{M} \to \{1, ..., M\}$ closer to the true ordering.

3. **Experimentation:** Conduct costly tests to reveal candidate materials' true quality, where each test requires cost c regardless of its result.

Let $\tilde{\theta}_k$ represent the probability that a material in position k of the ranking is high-quality. $\tilde{\theta}_k$ depends on two factors—the overall share of viable materials and the accuracy of the ranking. Shown in Panel A of Figure 9, the sum of $\tilde{\theta}_k$ across all materials is $\bar{\theta}M$, but its shape reflects prioritization. With no prioritization, tests are conducted at random, resulting in the flat curve $\tilde{\theta}_k = \bar{\theta}$. Perfect judgment, in contrast, produces a step function, with all viable materials tested first. Intermediate judgment yields a curve between these extremes.

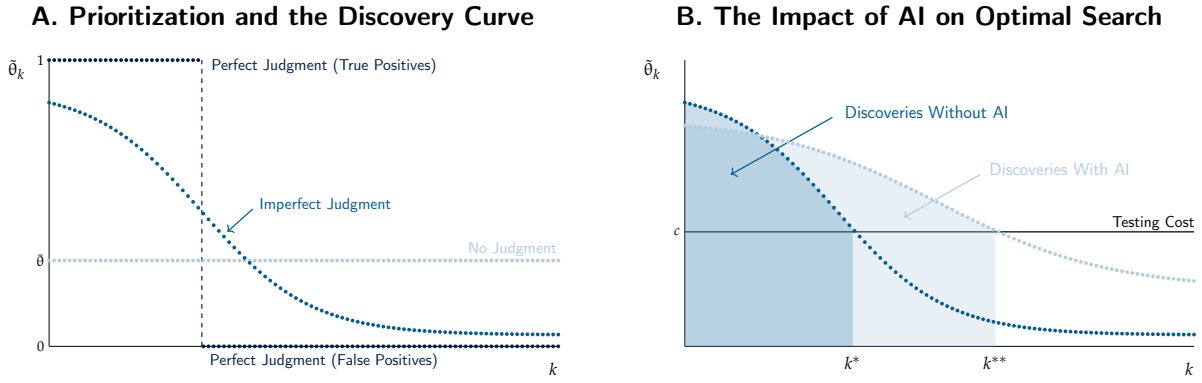


Figure 9. Illustration of Materials Discovery as Prioritized Search

Notes: This figure depicts the materials discovery process as a prioritized search problem. Panel A shows the probability that a material at position k is viable. The light blue line represents no judgment, with a constant discovery rate. The dark blue line illustrates perfect judgment, with all viable materials tested first. The medium blue line shows imperfect judgment. Panel B illustrates optimal search. Without AI, tests continue until k^*, with expected discoveries in the dark blue shaded region. AI increases the share of viable materials, but makes prioritization more difficult. The curve flattens, and the optimal number of tested compounds rises to $k^{**} > k^*$. Expected discoveries are shown by the light blue shaded region.

Given $\tilde{\theta}_k$, optimal search takes a simple form: test candidate materials in descending order of predicted quality, until the cost of an additional test exceeds the expected benefit (Weitzman, 1979; Agrawal et al., 2023). Therefore, scientists test every material such that $\tilde{\theta}_k \geq c$. Let k^* denote the ranking of the last material above this cutoff. The total number of expected discoveries is then:

$$E\,(\text{Discoveries}) = \sum_{k=1}^{k^*} \tilde{\theta}_k$$

I call $\tilde{\theta}_k$ the "discovery curve," as it fully pins down the expected rate of materials discovery.

In this framework, AI could impact materials discovery through two channels. First, the partial automation of compound design may change the average quality of candidates, affecting the area under the discovery curve. Second, AI-generated candidates may prove differentially challenging to evaluate, shifting the curve's slope. Figure 9, Panel B illustrates one potential scenario where AI increases $\bar{\theta}$ but makes candidates more difficult to prioritize. As a result, the optimal number of tests rises from k^* to k^{**}. The effect on the expected rate of materials discovery—defined as the ratio of discoveries to tests—is ambiguous, reflecting the competing forces of higher average quality and worse prioritization.

6.2.2 The Empirical Discovery Curve

Leveraging the granularity of my data, I construct an empirical analog of the discovery curve—the share of high-quality materials at each position in the sequence of tested candidates. The empirical discovery curve, denoted $\hat{\theta}_k$, is defined as:

$$\hat{\theta}_k = \frac{1}{J} \sum_{j=1}^{J} \theta_{k(j)}$$

where $\theta_{k(j)} \in \{0, 1\}$ represents the realized quality of a candidate at position k in the sequence of tested compounds for research cycle j. Each research cycle represents a complete process of idea generation, judgment, and experimentation, and J denotes the total number of cycles. To aggregate across materials with different baseline discovery rates, I normalize $\hat{\theta}_k$ by the average viability of each material category. I then group k into bins, adjusting the bin widths to represent a fixed share of total tested candidates within each category.

Figure 10, Panel A plots $\hat{\theta}_k$ before and after the introduction of AI, revealing a shift along two dimensions. First, the tool boosts the overall share of viable compounds, $\bar{\theta}$, captured by the increased area under the curve. Second, AI-generated compounds prove more challenging to evaluate, evidenced by the shallower slope in the treatment group. As I show in the next section, this reflects the difficulty of evaluating compounds scientists did not themselves create. The combination of these effects results in a higher average discovery rate and a greater number of materials tested overall. Additionally, the nearly identical values of $\hat{\theta}_k$ for the last tested candidate in each curve lend support to the prioritized search framework, as the theory predicts that search should stop when the cost of an additional test exceeds the expected benefit.[14]

Panels B and C compare pre- and post-treatment discovery curves for researchers in the top

[14]Notably, this not the case for scientists in the bottom quartile of judgment (see Figure 10, Panel B), suggesting that they may be at a corner solution where they simply test all AI suggestions.

and bottom quartiles of judgment, revealing a widening gap in their ability to evaluate candidates. While top-quartile scientists experience only a modest decline in prioritization, those in the bottom quartile see a marked decrease. Strikingly, the post-treatment curve for bottom-quartile researchers is essentially flat—their ranking of AI-generated candidates is no better than random chance. Consequently, scientists with strong judgment test significantly fewer candidates but discover more viable compounds. Regressing observed discovery rates in the post-treatment period on the share of tests yielding a high-quality material, I find that differences in prioritization explain over three-quarters of the variation in research productivity. These results explain the central role of judgment in shaping the heterogeneous impact of AI.

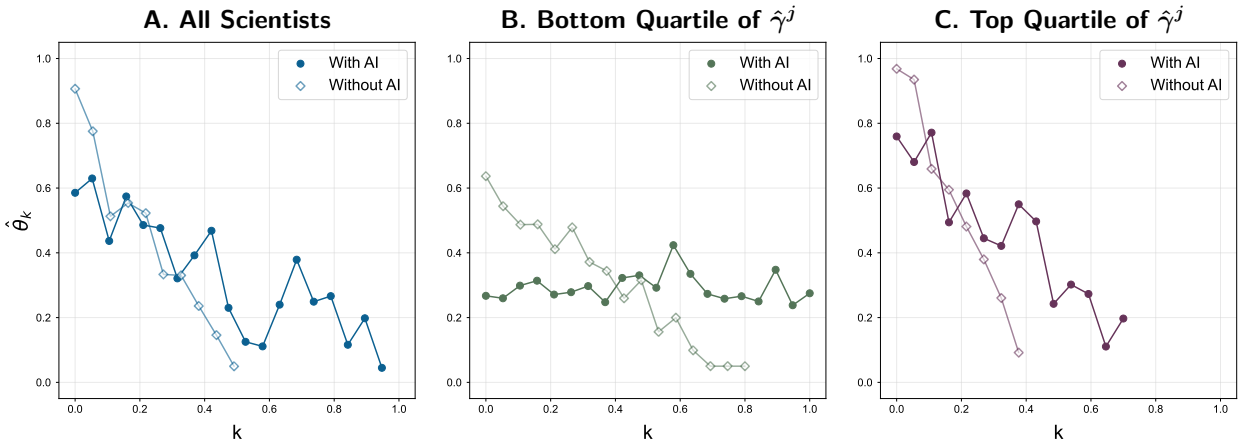


Figure 10. The Impact of AI on the Discovery Curve

Notes: This figure plots the empirical discovery curve, $\hat{\theta}_k$, defined as the share of viable materials at position k in the sequence of tested compounds. To aggregate across different materials types, I normalize discovery rates by the baseline viability of each material category. I group k into bins, adjusting the bin widths to represent a fixed share of total tested candidates for each material type. Panel A shows the discovery curve for treated scientists (solid dots) and not-yet-treated scientists (hollow diamonds). Panels B and C plot these curves separately for scientists in the bottom and top quartile of judgment, respectively. Judgment ability, denoted $\hat{\gamma}^j$, is estimated in the pre-period using the fixed effects specification described in Section 5.2.

Next, I demonstrate that differences in scientists' abilities to evaluate AI suggestions increase over time. Figure 11 presents post-treatment discovery curves relative to time since adoption. For scientists in the bottom quartile of judgment, $\hat{\theta}_k$ remains flat across all periods, indicating no improvement in prioritization. For those in the top quartile, the curve is also flat during the first five months after adoption. However, it becomes increasingly steep over the post-treatment period, reflecting rapid improvement. This shows that top evaluators learn to apply their expertise to the new problem of assessing AI-generated compounds, while others fail to do so.

The prioritized search framework also rationalizes the results on task composition presented in Section 6.1. First, the combination of partially-automated compound design and more challenging

evaluation shifts research effort from idea generation to judgment tasks. Second, the rise in the number of tested candidates explains the increase in time spent on experimentation tasks. Finally, superior prioritization causes scientists with strong judgment to see a smaller change in effort allocated to experimentation.

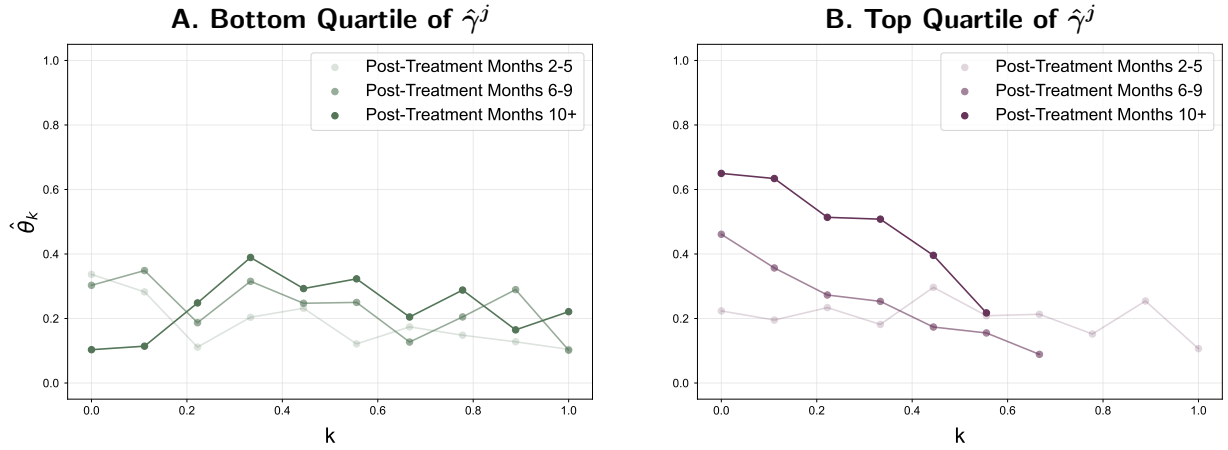


Figure 11. Learning to Evaluate AI Suggestions

Notes: This figure shows the empirical discovery curve, $\hat{\theta}_k$, separated into three post-treatment periods: months 2-5, months 6-9, and months 10+. Panels A and B show these curves for scientists in the bottom and top quartiles of judgment, respectively. Judgment ability, denoted $\hat{\gamma}^j$, is estimated in the pre-period using the fixed effects specification described in Section 5.2. To aggregate across different materials types, I normalize discovery rates by the baseline viability of each material category. I group k into bins, adjusting the bin widths to represent a fixed share of total tested candidates for each material category.

6.2.3 Domain Knowledge and Judgment

Why are some scientists so much better at judging AI-suggested compounds? To investigate this question, I turn to the survey of the lab's researchers. I ask scientists three questions about their evaluation process:

Q1: Compared to previous methods, how would you rate the difficulty of evaluating AI-generated candidate materials (much easier, somewhat easier, about the same, somewhat harder, much harder)?

Q2: How frequently can you identify AI-suggested candidate compounds as false positives without synthesizing the material (never/rarely/sometimes/often)?

Q3: On a scale of 1–10, how useful are each of the following in evaluating AI-suggested candidate materials (scientific training, experience with similar materials, intuition or gut feeling, and experience with similar tools)?

Consistent with my empirical results, 88% of scientists find AI-generated compounds more challenging to evaluate. The reason is simple: designing a material provides information about its quality. However, difficulty is not uniform across researchers. Compared to the bottom quartile, those in the top quartile of judgment are nearly twice as likely to report frequently ruling out false positives without testing (62% vs. 34%, $p < 0.00$).

This disparity suggests that top scientists possess some form of knowledge that allows them to effectively assess AI suggestions. Figure 12 investigates four potential sources of their expertise. Panels A and B show that researchers in the top half of judgment place greater value on their scientific training and experience with similar materials when evaluating model-generated candidates. The utility of "intuition or gut feeling"—a proxy for tacit knowledge—is also positively correlated with judgment (Panel C). In contrast, familiarity with similar AI technologies does not explain the observed differences, as all scientists report minimal prior exposure (Panel D). This is consistent with the fact that differences in prioritization emerge over time. Supporting the importance of domain knowledge, I find that scientists in the top quartile of judgment are 3.4 times more likely to have published an academic article on their material of focus.

I rule out three alternative explanations for differences in prioritization, shown in Appendix Table A10. First, they are not due to confusion about how to use the tool. Second, they are not explained by scientists' differential trust in the model's suggestions. Finally, they do not reflect variation in material of focus, as the relationships in Figure 12 hold within compound type and intended application.

These results demonstrate the value of domain knowledge for assessing AI suggestions. From a machine learning perspective, this shows that top scientists observe certain features of the materials design problem not captured by the model. Consequently, integrating human feedback into algorithmic predictions may be a promising path forward for scientific discovery (Hassabis, 2022; Alur et al., 2023, 2024). From an economic perspective, these findings highlight the complementarity between algorithms and expertise in the innovative process. In particular, they emphasize the growing importance of a new research skill—judging model suggestions—that augments AI technologies.

Some have speculated that big data and machine learning will render domain knowledge obsolete (Anderson, 2008; Klein et al., 2017; Gennatas et al., 2020). In the context of materials science, this appears not to be the case. Instead, only researchers with sufficient expertise can harness the power of the technology. The fact that the tool is ineffective without skilled scientists suggests that demand for human researchers may remain strong, even as AI transforms scientific discovery.

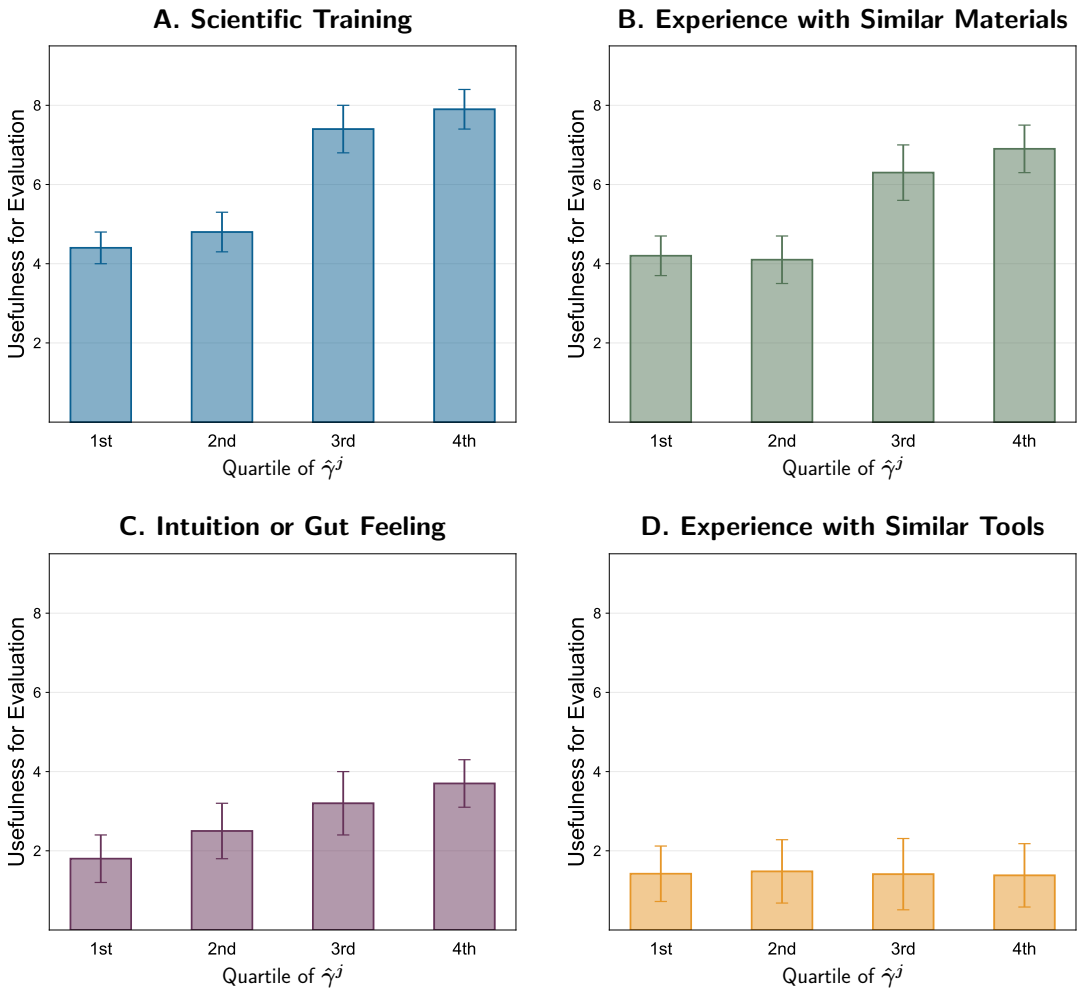


Figure 12. The Role of Domain Knowledge in Assessing AI Suggestions

Notes: This figure presents survey results on the usefulness of different forms of expertise for judging AI-generated candidate materials. The outcomes are reported on a scale of 1 (least useful) to 10 (most useful), separated by quartile of estimated judgment skill, $\hat{\gamma}^j$. Judgment ability is estimated in the pre-period, using the fixed effects approach described Section 5.2. Panel A shows results for scientific training, Panel B shows results for experience with similar materials, Panel C shows results for intuition or gut feeling, and Panel D shows results for experience with similar tools. The vertical lines represent 90% confidence intervals. The relevant survey instruments are provided in Appendix C.

6.3 Organizational Adaptation and Longer-Run Effects

AI changes the returns to specific skills, increasing the value of scientists' judgment while diminishing the importance of idea generation. Therefore, adjusting employment practices to prioritize researchers with strong judgment implies significant productivity gains.

In the final month of my sample—excluded from the primary analysis—the firm restructured its research teams.[15] The lab fired 3% of its researchers. At the same time, it more than offset these departures through increased hiring, expanding its workforce on net. While I do not observe the

[15]The lab made no changes to its employment practices during the experiment.

abilities of the new hires, those dismissed were significantly more likely to have weak judgment. Figure 13 shows the percent fired or reassigned by quartile of $\hat{\gamma}^j$. Scientists in the top three quartiles faced less than a 2% chance of being let go, while those in the bottom quartile had nearly a 10% chance.

The lab's response exemplifies the LeChatelier principle: it reacted more strongly to the tool over time because it could re-optimize a broader set of inputs (Samuelson, 1947; Milgrom and Roberts, 1996). The resulting change in the composition of researchers suggests that my estimates may understate the tool's longer-run impact.

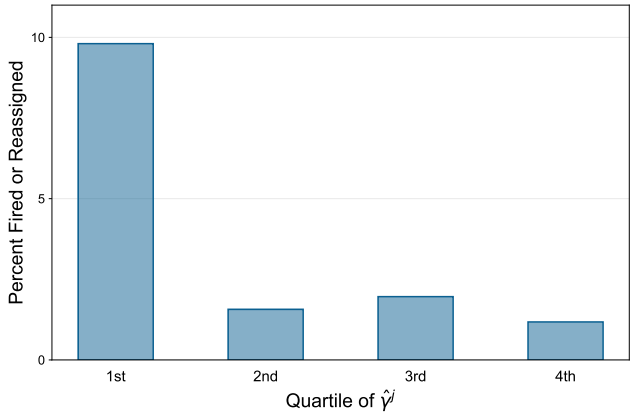


Figure 13. Probability of Firing or Reassignment by Quartile of Judgment

Notes: This figure shows the percent of scientists fired or reassigned by quartile of estimated judgment, $\hat{\gamma}^j$. Judgment ability is estimated in the pre-period, using the fixed effects approach described Section 5.2. Data on firing and reassignment come from the final month of the sample, which is excluded from my primary analysis. The lab made no changes to hiring and firing practices during the experiment.

7 Scientist Wellbeing and Beliefs About AI

AI substantially changes the materials discovery process. Using the survey, I explore how this impacts scientists' job satisfaction and beliefs about artificial intelligence. Beyond the direct welfare implications, these results shed light on AI's potential to shape who becomes a scientist, the fields they select into, and the skills they invest in.

7.1 Job Satisfaction

AI could affect scientists' job satisfaction in contrasting ways. It might boost morale by enhancing capabilities and increasing the rate of scientific discovery. Alternatively, it could make the work less enjoyable by shifting focus to less engaging tasks. To investigate the relative importance of these

channels, I elicit changes in job satisfaction along three dimensions—those due to productivity shifts, those resulting from task reallocation, and the overall impact.

Figure 14, Panel A shows the results on a -10 to 10 scale, grouped by quartile of initial productivity. Two countervailing trends emerge: a negative impact from task changes and a mostly positive effect from improved productivity. The task reallocation effect is consistently negative across quartiles, ranging from -4.1 to -4.8. While enjoyment from increased productivity partially offsets this negative effect—especially for high-ability scientists—82% of researchers see an overall decline in satisfaction.

Panel B reports the primary reasons scientists dislike the changes in their tasks. The most common complaint is skill underutilization (73%), followed by tasks becoming less creative and more repetitive (53%). Concerns over credit allocation and the complexity of the AI tool were cited by 21% and 19% of researchers, respectively. This highlights the difficulty of adapting to rapid technological progress. As one scientist noted: "While I was impressed by the performance of the [AI tool]...I couldn't help feeling that much of my education is now worthless. This is not what I was trained to do."

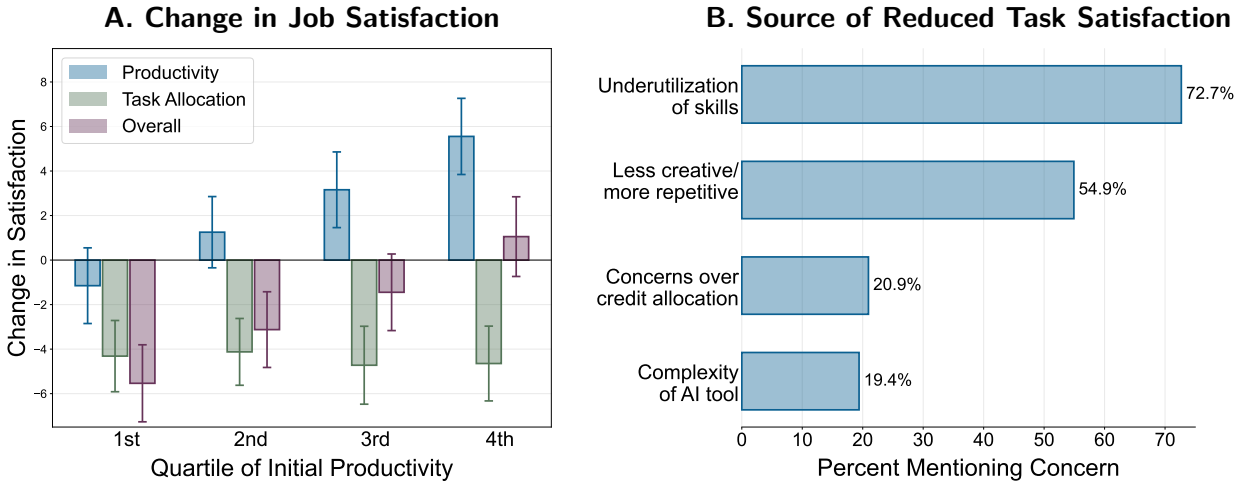


Figure 14. Impact of AI on Job Satisfaction

Notes: This figure shows the impact of AI on scientists' job satisfaction. Panel A displays changes in job satisfaction across quartiles of initial productivity, along with 95% confidence intervals. I report three measures: changes in satisfaction due to shifts in productivity (blue), changes in satisfaction due to shifts in task allocation (green), and overall changes in satisfaction (purple). Change in satisfaction is on a -10 to 10 scale, where negative values indicate decreased satisfaction, zero indicates no effect, and positive values indicate increased satisfaction. Panel B presents the primary sources of reduced task satisfaction. The x-axis shows the percentage of respondents mentioning each concern. The top four concerns are listed: underutilization of skills, less creative/more repetitive work, concerns over credit allocation, and complexity of the AI tool. The relevant survey questions are provided in Appendix C.

These results challenge the view that AI will primarily automate tedious tasks, allowing humans to focus on more rewarding activities (Mollick, 2023; McKendrick, 2024). Instead, the tool automates precisely the tasks that scientists find most interesting—creating ideas for new materials. This reflects a fundamental difference between AI and previous technologies. While earlier innovations excelled in routine, programmable tasks (Autor, 2014), deep learning models generate novel outputs by identifying patterns in their training data.

The responses also illustrate how organizational practices shape the welfare effects of AI. Scientists care not only about their own productivity, but performance relative to their peers (Card et al., 2012). Consequently, despite seeing small gains in research output (see Figure 6), scientists in the bottom quartile experience declining satisfaction with their productivity. This aligns with the firm's promotion practices, as advancement decisions are based on relative performance.

7.2 Beliefs About Artificial Intelligence

In addition to impacting job satisfaction, working with the model changes scientists' beliefs about artificial intelligence. Figure 15 shows researchers' level of agreement with five AI-related statements before and after the tool's introduction, revealing shifts along several dimensions. As discussed in Section 3, I fielded the survey after all scientists had gained access to the model, so pre-treatment beliefs represent respondents' best recollection.

Researchers report a significant increase in their belief that AI will enhance productivity in their field. Concern over AI displacing jobs remains relatively stable, however, potentially reflecting the continued need for human judgment. Due to the changing research process, scientists indicate a significant increase in their belief that AI will alter the skills needed to succeed in their job. Consequently, the number of researchers planning to reskill grows substantially. Finally, scientists report a small reduction in satisfaction with their choice of field, consistent with the decline in job satisfaction documented in the previous section.

These results show that hands-on experience with AI can dramatically influence views on the technology. Furthermore, the responses reveal an important fact: scientists did not anticipate the effects documented in this paper. This fits a recurring pattern of domain experts underestimating the capabilities of AI in their respective fields (Grace et al., 2018; Bostock, 2022; Grace et al., 2022; Steinhardt, 2022; Cotra, 2023).

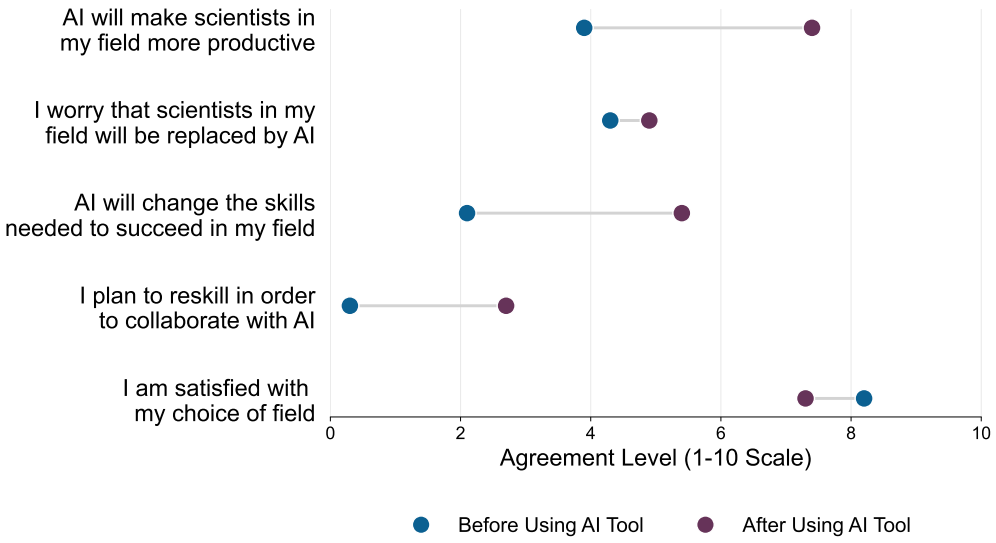


Figure 15. Shifts in Scientists' Beliefs About AI

Notes: This figure reports the change in researchers' beliefs about AI before and after using the tool. It shows five statements related to AI's impact on science and scientists. Agreement levels are measured on a 1-10 scale, shown on the x-axis. Each statement is represented by a pair of dots: blue dots indicate beliefs before using the AI tool, while purple dots represent beliefs after using it. The survey was conducted in May and June of 2024, after all scientists had gained access to the model. Pre-treatment beliefs are therefore based on scientists' best recollection. The relevant survey questions are provided in Appendix C.

8 Conclusion

The long-run impact of artificial intelligence hinges on the extent to which AI technologies transform science and innovation. This paper takes the first step toward answering this question, studying the randomized introduction of a new materials discovery tool in a large R&D lab. I find that AI substantially boosts materials discovery, leading to an increase in patent filing and a rise in downstream product innovation. However, the technology is effective only when paired with sufficiently skilled scientists.

Investigating the mechanisms behind these results, I show that AI automates a majority of idea generation tasks, reallocating researchers to the new task of judging model-produced candidate compounds. Top scientists learn to prioritize promising AI suggestions, while others waste significant resources testing false positives. Interpreted through the lens of a task framework, this shows that AI changes the skills needed to make scientific discoveries.

My findings speak to a broader debate about human expertise and creativity in a world of artificial intelligence (Brynjolfsson and McAfee, 2014; Daugherty and Wilson, 2018; Acemoglu and Restrepo, 2019; Autor et al., 2023). One perspective—often associated with the AI research community—posits that the combination of big data and deep learning will render domain knowl-

edge obsolete, as models automate most forms of cognitive labor (Anderson, 2008; Grace et al., 2022; Schulman, 2023; Aschenbrenner, 2024). In contrast, others are pessimistic about AI's potential to perform economically valuable tasks, particularly in areas such as scientific discovery that require creative leaps (Woodie, 2022; Acemoglu, 2024; Wolfram, 2024). My results suggest an intermediate view. In materials science, I show that AI can meaningfully accelerate invention. However, the model must be complemented by domain experts who can evaluate and refine its predictions.

My analysis leaves several key issues unexplored. First, it is important to investigate the equilibrium effects on the supply and demand for scientific expertise. Second, while I consider one organizational adaptation to AI—hiring and firing practices—I do not study changes in training or incentives. Finally, it is critical to understand how the results change as AI technologies improve.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–296. ↪ [p. 15]

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 2024. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." *Nature* 630 (8016): 493–500. ↪ [p. 4]

Acemoglu, Daron. 2024. "The Simple Macroeconomics of AI." ↪ [pp. 5 and 39]

Acemoglu, Daron, Ufuk Akcigit, and Murat Alp Celik. 2022a. "Radical and Incremental Innovation: The Roles of Firms, Managers, and Innovators." *American Economic Journal: Macroeconomics* 14 (3): 199–249. ↪ [p. 2]

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." 1043–1171. ↪ [pp. 28 and 75]

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022b. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." *Journal of Labor Economics* 40 (S1). ↪ [p. 4]

Acemoglu, Daron, and Pascual Restrepo. 2019. "The wrong kind of AI? Artificial intelligence and the Future of Labour Demand." *Cambridge Journal of Regions, Economy and Society* 13 (1): 25–35. ↪ [p. 38]

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, Automation, and the Rise in U.S. Wage Inequality." *Econometrica*. ↪ [pp. 4 and 28]

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2024. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." *NBER Working Paper 31422*. ↪ [p. 5]

Aghion, Philippe, Benjamin Jones, and Charles Jones. 2019. "Artificial Intelligence and Economic Growth." *The Economics of Artificial Intelligence: An Agenda*. ↪ [pp. 1 and 4]

Agrawal, A., J. S. Gans, and A. Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." *Journal of Economic Perspectives* 33: 31–50. ↪ [p. 4]

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence.*: Harvard Business Review Press. ↪ [p. 3]

Agrawal, Ajay K., John McHale, and Alexander Oettl. 2023. "Artificial Intelligence and Scientific Discovery: A Model of Prioritized Search." *NBER Working Paper No. 31558*. ↪ [pp. 5, 28, and 29]

Alur, Rohan, Loren Laine, Darrick K. Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. 2023. "Auditing for Human Expertise." *Neural Information Processing Systems (NeurIPS)*. ↪ [pp. 5 and 33]

Alur, Rohan, Manish Raghavan, and Devavrat Shah. 2024. "Human Expertise in Algorithmic Prediction." ↪ [pp. 5 and 33]

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine* 16: 16–07. ↪ [pp. 3, 33, and 39]

Angelova, Victoria, Will Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion."Technical report. ↪ [p. 5]

Angwin, Julia. 2024. "Press Pause on the Silicon Valley Hype Machine." *The New York Times*. ↪ [p. 5]

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. "Wasserstein GAN." *Proceedings of Machine Learning Research (PMLR)*. ↪ [p. 8]

Aschenbrenner, Leopold. 2024. "Situational Awareness: The Decade Ahead." ↪ [pp. 5 and 39]

Athey, S., and G.W. Imbens. 2017. "he Econometrics of Randomized Experimentsa." *Handbook of Economic Field Experiments* 1: 73–140. ↪ [p. 15]

Athey, Susan, and Scott Stern. 2002. "The Impact of Information Technology on Emergency Health Care Outcomes." *RAND Journal of Economics* 33 (3): 399–432. ↪ [p. 4]

Autor, David. 2014. "Polanyi's Paradox and the Shape of Employment Growth." *NBER Working Paper No. 20485*. ↪ [p. 37]

Autor, David. 2015. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." *Journal of Economic Perspectives* 29: 3–30. ↪ [p. 4]

Autor, David, Caroline Chin, Anna Salomon, and Bryan Seegmiller. 2024. "New Frontiers: The Origins and Content of New Work, 1940–2018." *Quarterly Journal of Economics*. ↪ [p. 4]

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4): 1279–1333. ↪ [p. 4]

Autor, David, Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *The Quarterly Journal of Economics* 113 (4): 1169–1213. ↪ [p. 4]

Autor, David, David A. Mindell, and Elisabeth B. Reynolds. 2023. *The Work of the Future: Building Better Jobs in an Age of Intelligent Machines*.: The MIT Press. ↪ [p. 38]

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2024. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics* 151: 103745. ↪ [p. 6]

Besiroglu, Tamay, Nicholas Emery-Xu, and Neil Thompson. 2023. "Economic Impacts of AI-augmented R&D." ↪ [p. 5]

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen. 2014. "The Distinct Effects of Information Technology and Communication Technology on Firm Organization." *Management Science* 60 (12): 2859–2885. ↪ [p. 4]

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event Study Designs: Robust and Efficient Estimation." *Review of Economic Studies*. ↪ [pp. 15 and 56]

Bostock, J. 2022. "A Confused Chemist's Review of AlphaFold 2." ↪ [p. 37]

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*.: Oxford University Press. ↪ [p. 5]

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." *The Quarterly Journal of Economics* 117 (1): 339–376. ↪ [p. 4]

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2023. "Generative AI at Work." ↪ [pp. 5 and 23]

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, NY: W. W. Norton Company. ↪ [p. 38]

Brynjolfsson, Erik, and Gabriel Unger. 2023. "The Macroeconomics of Artificial Intelligence." ↪ [p. 5]

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review* 105 (7): 2044–85. ↪ [p. 8]

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230. ↪ [pp. 14 and 56]

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. 2012. "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." *American Economic Review* 102 (6): 2981–3003. ↪ [p. 37]

Caudai, Claudia, Antonella Galizia, Filippo Geraci, Loredana Le Pera, Veronica Morea, Emanuele Salerno, Allegra Via, and Teresa Colombo. 2021. "AI applications in functional genomics." *Computational and Structural Biotechnology Journal* 19: 5762–5790. ↪ [p. 4]

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96. ↪ [pp. 14 and 56]

Cheetham, Anthony K., and Ram Seshadri. 2024. "Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery." *Chemistry of Materials* 36 (8): 3490–3495. ↪ [p. 10]

Cheng, Lingwei, and Alexandra Chouldechova. 2022. "Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening." *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2). ↪ [p. 5]

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher

Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9). ↪ [p. 24]

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." *Proceedings of Machine Learning Research (PMLR)* 81: 134–148. ↪ [p. 5]

Chow, Trevor, Basil Halperin, and Zachary Mazlish. 2024. "Transformative AI, Existential Risk, and Asset Pricing." ↪ [p. 5]

Clancy, Matt, and Tamay Besiroglu. 2023. "The Great Inflection? A Debate About AI and Explosive Growth." *Asterisk Magazine*. ↪ [p. 5]

Cockburn, Iain M, Rebecca Henderson, and Scott Stern. 2019. "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis." *The Economics of Artificial Intelligence: An Agenda*. ↪ [pp. 1 and 5]

Cotra, Ajeya. 2023. "Language models surprised us." ↪ [p. 37]

Cowgill, Bo. 2018. "Bias and productivity in humans and algorithms: Theory and evidence from resume screening." ↪ [p. 5]

Cui, Kevin Zheyuan, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. 2024. "The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers." ↪ [p. 5]

Daugherty, Paul R., and H. James Wilson. 2018. *Human + Machine: Reimagining Work in the Age of AI*. Boston, MA: Harvard Business Review Press. ↪ [p. 38]

De, Sandip, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. 2016. "Comparing molecules and solids across structural and alchemical space." *Physical Chemistry Chemical Physics* 18: 13754–13769. ↪ [pp. 2, 11, 21, 47, 64, and 65]

DeepMind. 2024. "AlphaFold 3 predicts the structure and interactions of all of life's molecules." ↪ [p. 1]

Dell'Aqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *HBS Working Paper 24-013*. ↪ [p. 5]

Diamandis, Peter. 2020. "Materials Science: The Unsung Hero." ↪ [p. 6]

Donahue, Kate, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. "Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ↪ [p. 5]

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie van Dijk. 2022. "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias." ↪ [p. 13]

Erdil, Ege, and Tamay Besiroglu. 2023. "Explosive growth from AI automation: A review of the arguments." ↪ [p. 5]

Garicano, Luis, and Esteban Rossi-Hansberg. 2015. "Knowledge-Based Hierarchies: Using Organizations to Understand the Economy." *Annual Review of Economics* 7 (1): 1–30. ↪ [p. 4]

Gennatas, Efstathios, Jerome Friedman, Lyle Ungar, Romain Pirracchio, Eric Eaton, Lara Reichmann, Yannet Interian, José Luna, Charles Simone, Andrew Auerbach, Elier Delgado, Mark van der Laan, Timothy Solberg, and Gilmer Valdes. 2020. "Expert-augmented machine learning." *Proceedings of the National Academy of Sciences* 117 (9): 4571–4577. ↪ [pp. 3 and 33]

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." *Advances in Neural Information Processing Systems (NeurIPS)* 27: 2672–2680. ↪ [p. 8]

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts." *Journal of Artificial Intelligence Research* 62: 729–754. ↪ [p. 37]

Grace, Katja, Zach Stein-Perlman, Benjamin Weinstein-Raun, and John Salvatier. 2022. "2022 Expert Survey on Progress in AI." ↪ [pp. 37 and 39]

Gruber, Jonathan, Benjamin R. Handel, Samuel H. Kina, and Jonathan T. Kolstad. 2024. "Managing

Intelligence: Skilled Experts and AI in Markets for Complex Products." *NBER Working Paper No. 27038*. ↪ [p. 5]

Hassabis, D. 2022. "Using AI to accelerate Scientific Discovery." Speech at the Institute for Ethics in AI, University of Oxford. Available at: https://youtu.be/44DLDemisHassabis. ↪ [pp. 4, 8, and 33]

Henaff, Mikael, Joan Bruna, and Yann LeCun. 2015. "Deep Convolutional Networks on Graph-Structured Data." ↪ [p. 9]

Hill, Ryan, and Carolyn Stein. 2024. "Scooped! Estimating Rewards for Priority in Science." *Journal of Political Economy*. ↪ [pp. 13 and 18]

Hoelzemann, Johannes, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero. 2024. "The Streetlight Effect in Data-Driven Exploration." ↪ [pp. 2 and 20]

Imbens, Guido, and Donald Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*.: Cambridge University Press. ↪ [p. 54]

Jain, Anubhav, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin Persson. 2013. "The Materials Project: A materials genome approach to accelerating materials innovation." *APL Materials* 1 (1): 011002. ↪ [p. 11]

Johnson, Daniel D. 2017. "Learning Graphical State Transitions." *International Conference on Learning Representations (ICLR)*. ↪ [p. 9]

Jones, Benjamin, and Lawrence Summers. 2020. "A Calculation of the Social Returns to Innovation." *Innovation and Public Policy*: 13–59. ↪ [p. 4]

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596 (7873): 583–589. ↪ [p. 4]

Kalyani, Aakash. 2024. "The Creativity Decline: Evidence from US Patents." ↪ [pp. 2, 12, 21, 46, 65, and 66]

Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." *arXiv preprint arXiv:1710.10196*. ↪ [p. 8]

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring Technological Innovation over the Long Run." *American Economic Review: Insights* 3 (3): 303–20. ↪ [pp. 11, 21, 46, 65, and 66]

Khurana, Mark. 2023. *Streetlight Effects: How Does the Streetlight Influence What We Choose to Study?*. 23–26: Cambridge University Press. ↪ [pp. 2 and 20]

Kim, Soomi. 2023. "Shortcuts to Innovation: The Use of Analogies in Knowledge Production." ↪ [pp. 2 and 20]

Kipf, Thomas N., and Max Welling. 2017. "Semi-Supervised Classification with Graph Convolutional Networks." *International Conference on Learning Representations (ICLR)*. ↪ [p. 9]

Klein, Gary, Ben Shneiderman, Robert Hoffman, and Kenneth Ford. 2017. "Why Expertise Matters: A Response to the Challenges." *IEEE Intelligent Systems* 32 (6): 67–73. ↪ [p. 33]

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–293. ↪ [p. 5]

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic Discrimination Among Large U.S. Employers." *The Quarterly Journal of Economics* 137 (4): 1963–2036. ↪ [p. 24]

Kochkov, Dmitrii, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. 2024. "Neural general circulation models for weather and climate." *Nature*. ↪ [p. 4]

Kogan, Leonid, Dimitris Papanikolaou, Lawrence D. W. Schmidt, and Bryan Seegmiller. 2023. "Technology

and Labor Displacement: Evidence from Linking Patents with Worker-Level Data." ↪ [p. 4]

Korinek, Anton, and Donghyun Suh. 2024. "Scenarios for the Transition to AGI." *NBER Working Paper 32255*. ↪ [p. 5]

Krieger, Joshua, Danielle Li, and Dimitris Papanikolaou. 2021. "Missing Novelty in Drug Development." *The Review of Financial Studies* 35 (2): 636–679. ↪ [p. 11]

Leitão, Diogo, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. 2022. "Human-AI Collaboration in Decision-Making: Beyond Learning to Defer." *International Conference on Learning Representations (ICLR)*. ↪ [p. 5]

Li, Yujia, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. 2018. "Learning Deep Generative Models of Graphs." *International Conference on Machine Learning (ICML)*. ↪ [p. 9]

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation." *The Quarterly Journal of Economics*. ↪ [p. 5]

Manning, Benjamin, Kehang Zhu, and John Horton. 2024. "Automated Social Science: A Structural Causal Model-Based Approach." ↪ [p. 5]

McKendrick, Joe. 2024. "Artificial Intelligence Will Replace Tasks, Not Jobs." *Forbes*. ↪ [p. 37]

Merchant, Amil, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. "Scaling deep learning for materials discovery." *Nature* 624 (7990): 80–85. ↪ [p. 1]

Milgrom, Paul, and John Roberts. 1996. "The LeChatelier Principle." *The American Economic Review* 86 (1): 173–179. ↪ [p. 35]

Mirza, Mehdi, and Simon Osindero. 2014. "Conditional Generative Adversarial Nets." ↪ [p. 8]

Mollick, Ethan. 2023. "In Praise of Boring AI." *One Useful Thing*. ↪ [p. 37]

Moskowitz, Sanford. 2022. *Advanced Materials Innovation: Managing Global Technology in the 21st Century*.: Wiley. ↪ [p. 6]

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics* 137 (2): 679–727. ↪ [p. 5]

Mullainathan, Sendhil, and Ashesh Rambachan. 2024. "From Predictive Algorithms to Automatic Generation of Anomalies." ↪ [p. 5]

Mullowney, Michael W., Katherine R. Duncan, Somayah S. Elsayed, Neha Garg, Justin J. J. van der Hooft, Nathaniel I. Martin, David Meijer, Barbara R. Terlouw, Friederike Biermann, Kai Blin, Janani Durairaj, Marina Gorostiola González, Eric J. N. Helfrich, Florian Huber, Stefan Leopold-Messer, Kohulan Rajan, Tristan de Rond, Jeffrey A. van Santen, Maria Sorokina, Marcy J. Balunas, Mehdi A. Beniddir, Doris A. van Bergeijk, Laura M. Carroll, Chase M. Clark, Djork-Arné Clevert, Chris A. Dejong, Chao Du, Scarlet Ferrinho, Francesca Grisoni, Albert Hofstetter, Willem Jespers, Olga V. Kalinina, Satria A. Kautsar, Hyunwoo Kim, Tiago F. Leao, Joleen Masschelein, Evan R. Rees, Raphael Reher, Daniel Reker, Philippe Schwaller, Marwin Segler, Michael A. Skinnider, Allison S. Walker, Egon L. Willighagen, Barbara Zdrazil, Nadine Ziemert, Rebecca J. M. Goss, Pierre Guyomard, Andrea Volkamer, William H. Gerwick, Hyun Uk Kim, Rolf Müller, Gilles P. van Wezel, Gerard J. P. van Westen, Anna K. H. Hirsch, Roger G. Linington, Serina L. Robinson, and Marnix H. Medema. 2023. "Artificial intelligence for natural product drug discovery." *Nature Reviews Drug Discovery* 22 (11): 895–916. ↪ [p. 1]

Myers, Kyle, Wei Yang Tham, Jerry Thursby, Marie Thursby, Nina Cohodes, Karim Lakhani, Rachel Mural, and Yilun Xu. 2024. "New Facts and Data about Professors and their Research." ↪ [p. 13]

Nazareno, Luísa, and Daniel S. Schiff. 2021. "The impact of automation and artificial intelligence on worker well-being." *Technology in Society* 67: 101679. ↪ [p. 3]

Nielsen, Michael, and Kanjun Qiu. 2022. "The trouble in comparing different approaches to science funding." ↪ [p. 20]

Noh, J., G. H. Gu, S. Kim, and Y. Jung. 2020. "Machine-enabled inverse design of inorganic solid materials: promises and challenges." *Chemical Science* 11: 4871–4881. ↪ [p. 9]

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381 (6654): 187–192. ↪ [pp. 5 and 23]

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot." ↪ [pp. 5 and 23]

Porter, Michael E, and Scott Stern. 2000. "Measuring the "Ideas" Production Function: Evidence from International Patent Output." *NBER Working Paper No. 7891*. ↪ [p. 23]

Radford, Alec, Luke Metz, and Soumith Chintala. 2016. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." *International Conference on Learning Representations (ICLR)*. ↪ [p. 8]

Ramani, Arjun, and Zhengdong Wang. 2023. "Why Transformative Artificial Intelligence is Really, Really Hard to Achieve." *The Gradient*. ↪ [p. 5]

Reiser, Patrick, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. 2022. "Graph neural networks for materials science and chemistry." *Nature* 3 (1). ↪ [pp. 7 and 8]

Samuelson, Paul. 1947. *The Foundations of Economic Analysis.*: Harvard University Press. ↪ [p. 35]

Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. "The Graph Neural Network Model." *IEEE Transactions on Neural Networks* 20 (1): 61–80. ↪ [p. 9]

Schmidt, J., H.C. Wang, T.F. Cerqueira, S. Botti, and M.A. Marques. 2022. "A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals." *Scientific Data* 9 (1): 64. ↪ [p. 11]

Schulman, Carl. 2023. "Intelligence Explosion." ↪ [pp. 5 and 39]

Soffia, Magdalena, Rosa Leiva-Granados, Xiaowei Zhou, and Jolene Skordis. 2024. "Does technology use impact UK workers' quality of life? A report on worker wellbeing."Technical report, Institute for the Future of Work. ↪ [p. 3]

Steinhardt, Jacob. 2022. "AI Forecasting: One Year In." ↪ [p. 37]

Subramaniam, Sriram. 2024. "Structural biology in the age of AI." *Nature Methods* 21 (1): 18–19. ↪ [p. 4]

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199. ↪ [pp. 14 and 56]

Tao, Terence. 2024. "Machine Assisted Proof." ↪ [p. 4]

Trinh, Trieu H., Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. "Solving olympiad geometry without human demonstrations." *Nature* 625 (7995): 476–482. ↪ [p. 4]

Velickovic, Petar, Arantxa Casanova, Pietro Liò, Guillem Cucurull, Adriana Romero, and Yoshua Bengio. 2018. "Graph Attention Networks." *International Conference on Learning Representations (ICLR)*. ↪ [p. 9]

Webb, Michael. 2020. "The Impact of Artificial Intelligence on the Labor Market." ↪ [p. 4]

Weitzman, M. L. 1979. "Optimal search for the best alternative." *Econometrica*: 641–654. ↪ [p. 29]

Wolfram, Stephen. 2024. "Can AI Solve Science?" *Stephen Wolfram Writings*. ↪ [p. 39]

Woodie, Alex. 2022. "Why AI Alone Won't Solve Drug Discovery." ↪ [p. 39]

Zhang, Edwin, Vincent Zhu, Naomi Saphra, Anat Kleiman, Benjamin L. Edelman, Milind Tambe, Sham M. Kakade, and Eran Malach. 2024. "Transcendence: Generative Models Can Outperform The Experts That Train Them." ↪ [p. 20]

Zunger, A. 2018. "Inverse design in search of materials with target functionalities." *Nature Reviews Chemistry* 2: 1–16. ↪ [p. 9]

Appendices

A Additional Tables and Figures **47**

B Data Appendix **64**

B.1 Measuring Material Quality 64

B.2 Measuring Structural Similarity 64

B.3 Measuring Patent Novelty 65

B.3.1 Kelly et al. (2021) Measure 66

B.3.2 Kalyani (2024) Measure 66

B.4 Classifying Research Tasks 67

B.4.1 Scientist Activities Data 67

B.4.2 Large Language Model Prompt and Fine-Tuning 67

B.4.3 Validation 69

C Survey Details **70**

C.1 Survey Questions 70

C.1.1 Introduction 70

C.1.2 Background Questions 70

C.1.3 Questions About Use of AI Tool 71

C.1.4 Questions About Job Satisfaction 73

C.1.5 Questions About Views on Artificial Intelligence 74

C.2 Respondent Characteristics and Attrition 75

D Task Model **75**

A Additional Tables and Figures

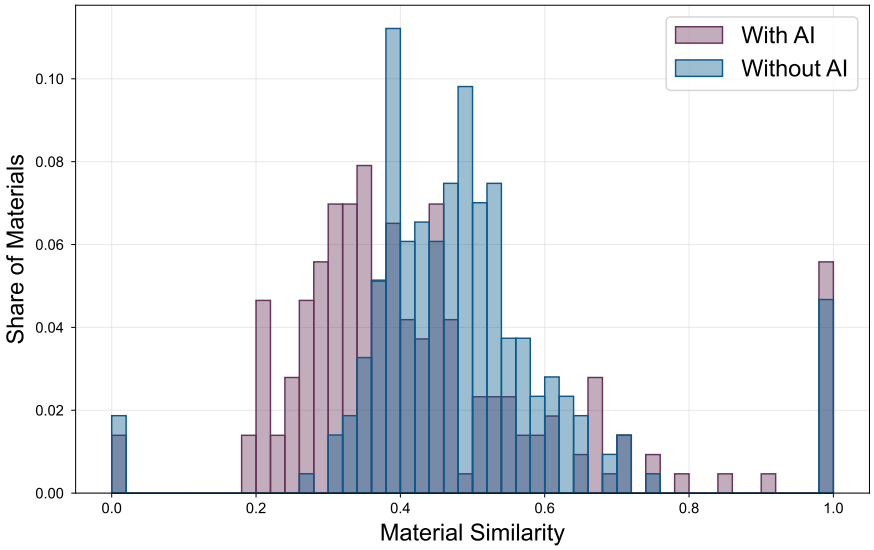


Figure A1. Distribution of Material Similarity

Notes: This figure shows the distribution of material similarity for compounds created before and after the introduction of AI. The similarity scores are calculated using the approach of De et al. (2016) described in Appendix B. These scores are only calculated for the crystals in my sample, which comprise 64% of all materials.

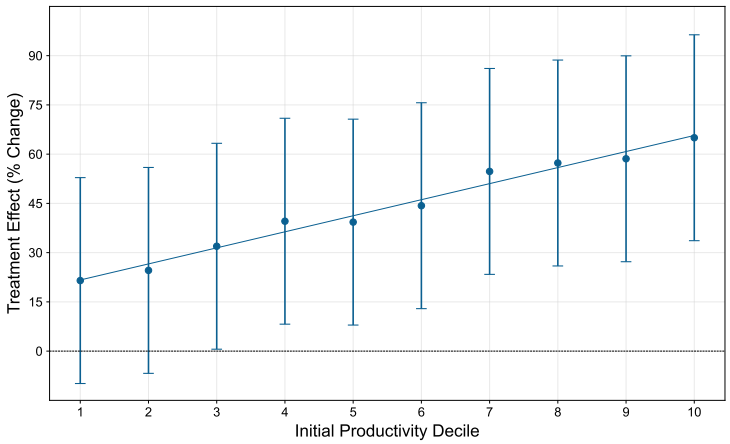


Figure A2. Team-Level Heterogeneity Estimates for Materials Discovery

Notes: This figure shows the heterogeneous impact of AI on materials discovery at the team level. I use an equivalent specification as the one described in Section 5, interacting AI access with quantile dummies of initial productivity. However, all variables are now aggregated to the team-level. I report 95% confidence intervals based on standard errors clustered at the team level.

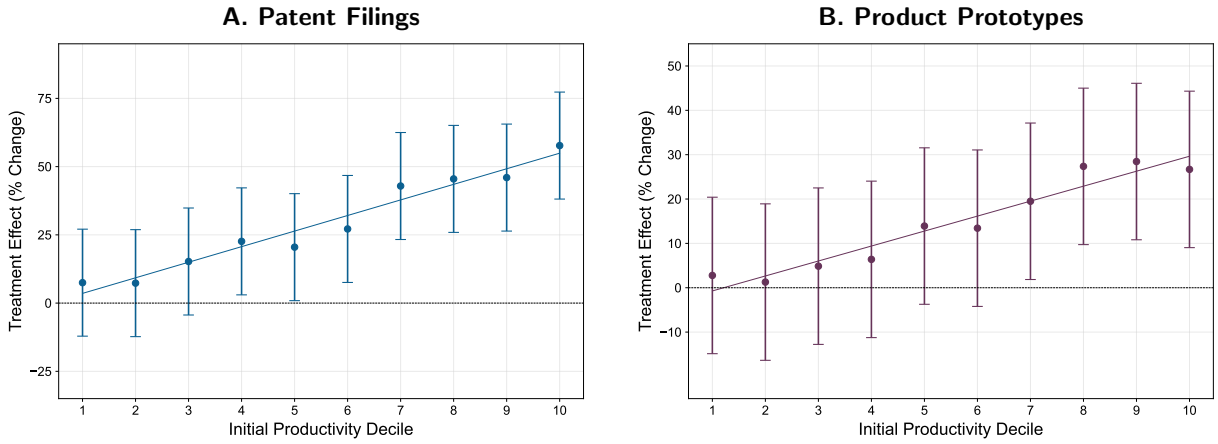


Figure A3. Heterogeneous Impact on Downstream Outcomes

Notes: This figure shows the heterogeneous impact of AI on downstream outcomes across the scientist productivity distribution. Panel A shows estimates for patent filings and Panel B shows estimates for product prototypes. I use an identical specification as the one described in Section 5, interacting AI access with quantile dummies of initial productivity. The vertical lines represent 95% confidence intervals based on standard errors clustered at the scientist level.

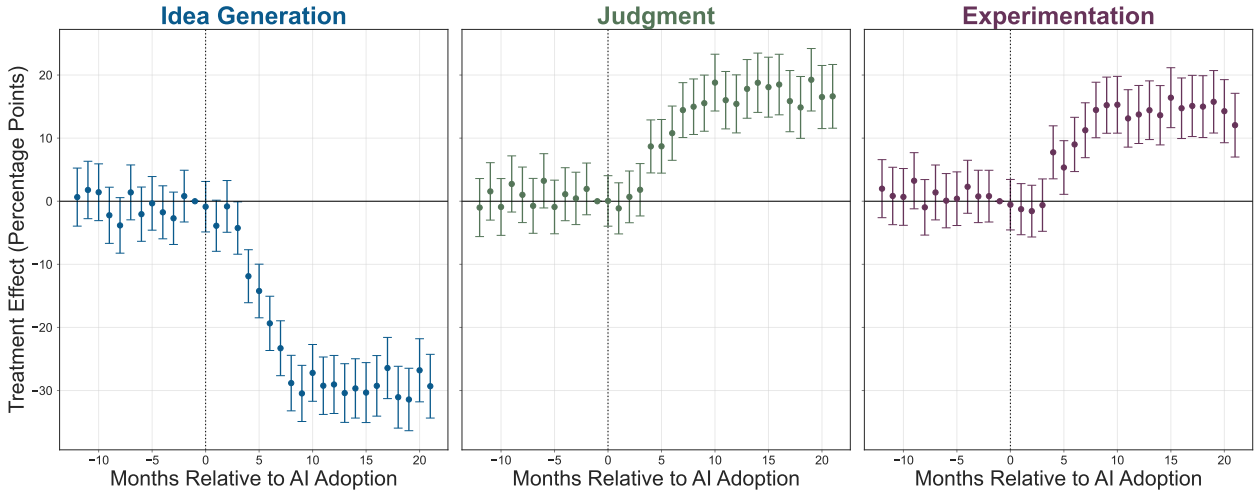


Figure A4. Impact of AI on the Composition of Research Tasks

Notes: This figure shows event studies for the impact of AI on the composition of research tasks. The estimates come from the following specification: $y_{it} = \alpha_i + \omega_t + \sum_{k \in \mathcal{K}} \delta_k D_{i,t-k} + \beta' X_{it} + \varepsilon_{it}$ where I include a full set of fixed effects and controls. The outcomes are the percentage of scientist-hours spent on idea generation tasks, judgment tasks, and experimentation tasks, respectively. These are normalized to exclude tasks not in these three categories. The process for classifying research tasks into these categories is discussed in Appendix B.4. I report 95% confidence intervals based on standard errors clustered at the team level.

Table A1. Impact on Materials Discovery Under Alternative Specifications

	Materials Discovery (1)	Materials Discovery (2)	Materials Discovery (3)	Materials Discovery (4)
Access to AI	0.195***	0.212***	0.204***	0.208***
	(0.105)	(0.092)	(0.078)	(0.065)
Team Fixed Effects		✓	✓	✓
Month Fixed Effects		✓	✓	✓
Time-Varying Team Characteristics			✓	✓
Material Type Fixed Effects				✓
Intended Application Fixed Effects				✓
Number of Scientists	1,018	1,018	1,018	1,018
Number of Teams	221	221	221	221

Notes: This table reports average treatment effect estimates on materials discovery under alternative specifications. Standard errors in parentheses are clustered at the team level. Column (1) includes no controls. Column (2) adds team and month fixed effects. Column (3) further includes controls for time-varying team characteristics. Column (4) adds fixed effects for material type and intended application. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A2. Impact on Patent Filings Under Alternative Specifications

	Patent Filings (1)	Patent Filings (2)	Patent Filings (3)	Patent Filings (4)
Access to AI	0.049***	0.051***	0.048***	0.050***
	(0.018)	(0.016)	(0.014)	(0.012)
Team Fixed Effects		✓	✓	✓
Month Fixed Effects		✓	✓	✓
Time-Varying Team Characteristics			✓	✓
Patent Category Fixed Effects				✓
Number of Scientists	1,018	1,018	1,018	1,018
Number of Teams	221	221	221	221

This table reports average treatment effect estimates on patent filings under different model specifications. Standard errors in parentheses are clustered at the team level. Column (1) includes no controls. Column (2) adds team and month fixed effects. Column (3) further includes controls for time-varying team characteristics. Column (4) adds fixed effects for patent type and technology field. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A3. Impact on Product Prototypes Under Alternative Specifications

	Product Prototypes (1)	Product Prototypes (2)	Product Prototypes (3)	Product Prototypes (4)
Access to AI	0.024**	0.026**	0.023**	0.025**
	(0.015)	(0.014)	(0.013)	(0.012)
Team Fixed Effects		✓	✓	✓
Month Fixed Effects		✓	✓	✓
Time-Varying Team Characteristics			✓	✓
Product Category Fixed Effects				✓
Number of Scientists	1,018	1,018	1,018	1,018
Number of Teams	221	221	221	221

This table reports average treatment effect estimates on product prototypes under different model specifications. Standard errors in parentheses are clustered at the team level. Column (1) includes no controls. Column (2) adds team and month fixed effects. Column (3) further includes controls for time-varying team characteristics. Column (4) adds fixed effects for product category. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A4. Estimates Using Randomization Inference

	Materials Discovery	Patent Filings	Product Prototypes
	(1)	(2)	(3)
Access to AI	0.231**	0.055**	0.025**
	(0.092)	(0.022)	(0.010)
Team Fixed Effects	✓	✓	✓
Month Fixed Effects	✓	✓	✓
Number of Scientists	1,018	1,018	1,018
Number of Teams	221	221	221

Notes: This table shows average treatment effects with 95% confidence constructed using permutation tests, following Imbens and Rubin (2015). All regressions include team and month fixed effects. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A5. Estimates Excluding Untreated Teams Working on Same Material-Application

	Number of Discovered Materials (1)	Number of Patent Filings (2)	Number of Product Prototypes (3)
Access to AI	0.18***	0.09***	0.08**
	(0.02)	(0.03)	(0.04)
Team Fixed Effects	✓	✓	✓
Month Fixed Effects	✓	✓	✓
Number of Scientists	1,018	1,018	1,018
Number of Teams	221	221	221

Notes: This table shows average treatment effect estimates robust to the presence of racing dynamics. I estimate team-level treatment effects excluding untreated teams with the same material-application focus. I then aggregate these estimates into an overall average effect. All regressions include team and month fixed effects. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A6. Robustness to Alternative Estimators

	Point Estimate (1)	Standard Error (2)	Lower Confidence Bound (95%) (3)	Upper Confidence Bound (95%) (4)
Panel A: Materials Discovery				
Borusyak et al. (2024)	0.205	0.028	0.150	0.260
Callaway and Sant'Anna (2021)	0.212	0.025	0.163	0.261
de Chaisemartin and D'Haultfœuille (2020)	0.198	0.021	0.157	0.239
Sun and Abraham (2021)	0.208	0.027	0.155	0.261
Panel B: Patent Filings				
Borusyak et al. (2024)	0.052	0.018	0.017	0.087
Callaway and Sant'Anna (2021)	0.054	0.015	0.025	0.083
de Chaisemartin and D'Haultfœuille (2020)	0.048	0.011	0.026	0.070
Sun and Abraham (2021)	0.051	0.017	0.018	0.084
Panel C: Product Prototypes				
Borusyak et al. (2024)	0.026	0.008	0.010	0.042
Callaway and Sant'Anna (2021)	0.028	0.005	0.018	0.038
de Chaisemartin and D'Haultfœuille (2020)	0.022	0.001	0.020	0.024
Sun and Abraham (2021)	0.027	0.007	0.013	0.041

Notes: This table shows the impact of AI on materials discovery (Panel A), patent filings (Panel B), and product prototypes (Panel C) using robust difference-in-differences estimators proposed by Borusyak et al. (2024), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020), and Sun and Abraham (2021). All regressions include team and month fixed effects. Standard errors are clustered at the team level.

Table A7. Poisson Regression Estimates

	Materials Discovery	Patent Filings	Product Prototypes
	(1)	(2)	(3)
Access to AI	0.34**	0.31**	0.22**
	(0.092)	(0.022)	(0.010)
Team Fixed Effects	✓	✓	✓
Month Fixed Effects	✓	✓	✓
Number of Scientists	1,018	1,018	1,018
Number of Teams	221	221	221

Notes: This table shows average treatment effects using a Poisson regression. All specifications include team and month fixed effects. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A8. Negative Binomial Regression Estimates

	Materials Discovery	Patent Filings	Product Prototypes
	(1)	(2)	(3)
Access to AI	0.098**	0.050**	0.022**
	(0.009)	(0.004)	(0.005)
Team Fixed Effects	✓	✓	✓
Month Fixed Effects	✓	✓	✓
Number of Scientists	1,018	1,018	1,018
Number of Teams	221	221	221

Notes: This table shows average treatment effects using a Negative Binomial regression. All specifications include team and month fixed effects. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A9. Impact of AI on R&D Efficiency

	R&D Efficiency (1)	R&D Efficiency (2)	R&D Efficiency (3)
Access to AI	0.151***	0.140***	0.132***
	(0.033)	(0.032)	(0.031)
Include new equipment costs			✓
Include bonuses in labor costs		✓	✓
Month Fixed Effects	✓	✓	✓
Team Fixed Effects	✓	✓	✓
Number of Scientists	651	651	651
Number of Teams	145	145	145

Notes: This table presents the impact of AI access on R&D efficiency. I define efficiency as: Product Prototypes/(Labor Costs + Other Variable Costs + Amortized Fixed Costs). Labor costs include salaries and benefits for the lab's scientists, technicians, and other staff. Non-labor variable costs are higher for treated scientists, due to increased spending on testing, product development, and model inference. Fixed costs include expenditure on training the model, which I amortize over a 5-year period. Column (1) does not include costs for new equipment or bonus salary. Column (2) adds costs for new equipment. Column (3) includes costs for both bonuses and new equipment. All specifications include team and month fixed effects. Standard errors in parentheses are clustered at the team level. *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table A10. Alternative Explanations for Differences in Judgment

Survey Question	Correlation with $\hat{\gamma}^j$	p-value
Confusion using AI tool	0.03	0.68
Trust in AI suggestions	-0.02	0.79
Materials of focus	0.04	0.60

Notes: This table tests alternative explanations for the observed differences in scientists' ability to prioritize AI-suggested compounds. It reports the correlation between scientists' estimated skill in judgment tasks, denoted $\hat{\gamma}^j$, and their responses to survey questions regarding their confusion using the tool and trust in the tool's suggestions, as well as their material of focus. $\hat{\gamma}^j$ is estimated in the pre-period using the fixed effects approach described in Section 5.2.

	True Positive Rate (1)	False Positive Rate (2)	True Negative Rate (3)	False Negative Rate (4)
Idea Generation	0.92	0.08	0.92	0.08
Judgment	0.94	0.05	0.95	0.06
Experimentation	0.91	0.07	0.93	0.09
Other	0.90	0.09	0.91	0.10

Table A11. Task Classification Accuracy

Notes: This table shows the classification accuracy of the large language model for each category of research task. Columns 1 through 4 show true positive, false positive, true negative, and false negative rates. The true values are based on a manually classified training set of 2,000 observations.

Table A12. Characteristics of Survey Sample

	Respondents	Non-Respondents
Scientist Characteristics		
Age (Years)	42.87	46.12
Tenure at Firm (Years)	8.33	10.58
Education		
Share Bachelor's Degree	0.04	0.07
Share Master's Degree	0.32	0.34
Share Doctoral Degree	0.64	0.59
Education Field		
Share Chemical Engineers	0.24	0.20
Share Chemists	0.15	0.13
Share Materials Scientists	0.22	0.24
Share Other Fields	0.27	0.34
Share Physicists	0.12	0.09
Material Type		
Share Biomaterials	0.25	0.25
Share Ceramics and Glasses	0.19	0.19
Share Metals and Alloys	0.31	0.32
Share Polymers	0.24	0.24
Innovative Output		
New Materials (Yearly)	0.53	0.59
Patent Filings (Yearly)	0.15	0.27
Product Prototypes (Yearly)	0.27	0.32
Treatment Waves		
Share Treatment Wave 1	0.43	0.40
Share Treatment Wave 2	0.41	0.39
Share Treatment Wave 3	0.16	0.20
Sample Size		
Total Number of Scientists	447	571

Table A13. Responses by Survey Question

Question Number	Responses
Question 1	447
Question 2	447
Question 3	447
Question 4	447
Question 5	445
Question 6	445
Question 7	442
Question 8	440
Question 9	440
Question 10	438
Question 11	435
Question 12	435
Question 13	432
Question 14	428
Question 15	425
Question 16	425
Question 17	420
Question 18	415
Question 19	415
Question 20	408
Question 21	402
Question 22	395
Question 23	388

B Data Appendix

B.1 Measuring Material Quality

I measure the quality of materials based on their properties. Scientists begin the search process by defining a set of target features for each compound. These features vary significantly across materials, but often include both atomic and large-scale characteristics determined by the intended application. For example, researchers may target the band gap (the energy difference between the valence band and the conduction band) or the refractive index (how much a material bends light). For each feature, I calculate the absolute distance between the target and its realized value. I combine these distances into two indices capturing quality at the atomic and macro scales, as well as an overall index.

Concretely, given a material m with K target features, I follow the chemical informatics literature and define the quality index as:

$$Q_m = \left(\frac{1}{K} \sum_{k=1}^{K} \frac{\left(p_{k,m} - t_k \right)^2}{\sqrt{\frac{1}{M-2} \sum_{m' \neq m} \left(p_{k,m'} - \bar{p}_{k,-m} \right)^2}} \right)^{-1}$$

where $p_{1,m}, \ldots, p_{k,m}$ are the relevant properties of material m and $t_1, \ldots, t_k$ are the targets.[16] The distance metric is rescaled by the leave-out standard deviation of each property across all materials tested, allowing aggregation across different types of features. Additionally, the index is normalized by the number of characteristics to support comparisons across materials with different numbers of target features. Across materials, the average number of target features is 4.7.

B.2 Measuring Structural Similarity

Following De et al. (2016), the similarity between crystal structures is calculated in four steps:

1. Define the substructure of crystal i to be the minimal repeating configuration of atoms and bonds, denoted s_j. Let $\mathcal{A}_i$ be the set of atoms in the substructure. Let $d_i(a_i)$ be the physical distance of atom $a_i \in \mathcal{A}_i$ from the centroid of s_i.

2. For two atoms a_i and a_j, define their chemical replaceability as $r(a_i, a_j) \in [0, 1]$. This is a scalar capturing the change in material properties when atom a_i is replaced with atom

[16] For features with unbounded targets, I set $t_k \in \{-\infty, \infty\}$.

a_j. The substitutability of two identical atoms is 1, and it decreases to a minimum of 0. Following De et al. (2016), the values of r are taken from the experimental literature.

3. Given substructures s_i and s_j, define a matching between the atoms in each substructure as a one-to-one function $m : \mathcal{A}_i \to \mathcal{A}_j$. If the substructures have different numbers of atoms, the surplus atoms in one set will be unmatched.

4. The similarity score for substructures s_i and s_j is defined as:

$$\text{Score}(s_i, s_j) = \max_{m} \sum_{\substack{a_i \in \mathcal{A}_i, a_j \in \mathcal{A}_j \\ \text{s.t. } a_j = m(a_i)}} r(a_i, a_j)^{\omega_r} \min\left(d_i(a_i), d_j(a_j)\right)^{\omega_c} \left\| d_i(a_i) - d_j(a_j) \right\|_2^{\omega_d}$$

   The sum is maximized over all possible matching of atoms in s_i and s_j. The first term captures the chemical similarity of atoms a_i and a_j. The second term puts additional weight on atoms that are closer to the centroid, reflecting the fact that central atoms are more predictive of material properties. The third term is the squared difference in the physical distance of a_i and a_j from their respective centroids, capturing geometric similarity between the structures. The weights ω_r, ω_d, and ω_d capture the relative importance of chemical similarity, centrality, and geometric structure, respectively.

5. Finally, the similarity measure is normalized to ensure that the metric is independent of the number of atoms in a substructure and has a scale between 0 and 1:

$$\text{Similarity}(s_i, s_j) = \text{Sigmoid}\left(\frac{\text{Score}(s_i, s_j)}{\sqrt{\text{Score}(s_i, s_i)\text{Score}(s_j, s_j)}}\right)$$

Using this method, I calculate the similarity between newly discovered compounds and existing materials. The structure of known materials comes from the Materials Project and Alexandria Databases, including more than 150,000 unique crystals. Figure A1 shows the similarity distribution of the new crystals in my sample. The measure displays a somewhat trimodal distribution: the majority of materials follow a right skewed distribution centered around 0.45, while the rest are concentrated in a large mass near 1 and smaller peak near 0.

B.3 Measuring Patent Novelty

I employ two patent similarity measures, proposed by Kelly et al. (2021) and Kalyani (2024).

B.3.1 Kelly et al. (2021) Measure

Kelly et al. (2021) construct a measure of patent similarity based on the full text of patents. It follows a modified term frequency-inverse document frequency (TF-IDF) approach. For a given patent p and term w, the term frequency is defined as:

$$\text{TF}_{pw} = \frac{c_{pw}}{\sum_k c_{pk}}$$

where c_{pw} is the count of term w in patent p. To account for the temporal nature of patent filings, it employs backward inverse document frequency (BIDF) for term w in year t:

$$\text{BIDF}_{wt} = \log\left(\frac{\#\text{ patents prior to } t}{1 + \#\text{ documents prior to } t \text{ that include term } w}\right)$$

The modified TF-BIDF for patent i and term w is then defined as:

$$\text{TFBIDF}_{w,i,t} = \text{TF}_{w,i} \times \text{BIDF}_{w,t}$$

where $t = \min(\text{filing year for } i, \text{filing year for } j)$ when comparing patents i and j. The TFBIDF vectors are normalized to unit length:

$$V_{i,t} = \frac{\text{TFBIDF}_{i,t}}{||\text{TFBIDF}_{i,t}||}$$

Finally, the similarity between patents i and j is computed as the cosine similarity of their normalized vectors:

$$\rho_{i,j} = V_{i,t} \cdot V_{j,t}$$

$\rho_{i,j}$ lies in the interval $[0, 1]$, with higher values indicating greater similarity.

B.3.2 Kalyani (2024) Measure

Kalyani (2024) constructs a measure of patent creativity based on the share of previously unused technical terms. The process involves two main steps: identifying technical bigrams and measuring whether they appear in existing patents.

Each patent is decomposed into bigrams, considering only those that appear at least twice in a patent and do not contain filler words. Technical bigrams are identified by removing non-technical bigrams (derived from the Corpus of Historical American English pre-1900) from each patent's

bigram list. For a patent p filed in year t, a bigram b is classified as creative at time t if:

$$\text{CreativeBigram}_{b,t} = \vec{1}\{b \notin \mathbb{B}'_{t-5,\ldots,t-1}\}$$

where $B'_{t-5,\ldots,t-1}$ is the set of bigrams appearing in patents filed in the previous five years. Patent creativity is then calculated as:

$$\text{PatentCreativity}_p = \frac{1}{|B_p|} \sum_{b=1}^{B_p} \vec{1}\{b \notin \mathbb{B}'_{t-5,\ldots,t-1}\}$$

where B_p is the set of bigrams in patent p. The measure is standardized by the average in a technology class throughout the sample.

B.4 Classifying Research Tasks

This section details the procedure for classifying textual data on scientist activities into categories of research tasks.

B.4.1 Scientist Activities Data

Scientists record their activities in logs required for administrative purposes. Each entry contains three components: entry date, duration of activity, and textual description. On average, scientists record 7.8 entries per week. Across scientists, this amounts to more than 1.6 million entries over my sample period.

To transform this textual information into a meaningful quantitative metric, I separate the materials discovery process into three categories of tasks: idea generation, judgment, and experimentation. Idea generation encompasses activities related to developing potential compounds, such as reviewing the literature on existing materials or creating preliminary designs. Judgment tasks focus on selecting which compounds to advance, often involving the analysis of simulations or predicting material characteristics based on domain knowledge. Finally, experimentation tasks are dedicated to synthesizing new materials and conducting tests to evaluate their properties. Based on the materials science literature and discussions with the lab's scientists, these categories divide the discovery process into meaningful, high-level groups.

B.4.2 Large Language Model Prompt and Fine-Tuning

I employ a large language model—Anthropic's Claude 3.5—to classify the textual descriptions of research activities. Compared to other frontier models, Claude 3.5 displays considerably better

performance. The prompt is shown in Figure A5.

Prompt: You are a materials science research activity classifer AI. Classify the given materials discovery research activity and output in JSON format with keys:

- classification: (Idea Generation/Evaluation/Testing/Other)
- certainty: (integer from 1 to 10, where 1 is least certain and 10 is most certain)

Description of categories:

- Idea generation: Activities relating to coming up with designs for novel candidate compounds, including brainstorming new molecular structures, applying design rules, using generative models, exploring composition spaces, or proposing modifications to existing materials based on structure-property relationships
- Evaluation: Activities relating to evaluating candidates to determine which to test, including computational screening, density functional theory calculations, molecular dynamics simulations, using machine learning models to predict properties, analyzing crystal structure stability, or reviewing literature on similar compounds
- Testing: Activities relating to physically synthesizing the candidates and testing their properties, including lab synthesis, characterization techniques (XRD, SEM, TEM, etc.), performance testing, stability analysis, and experimental validation of predicted properties
- Other: Any activity that does not fall into one of the previous three categories

Research activity to classify: [Research Activity Description]

Analyze the activity in the context of materials science and materials discovery, then provide your classification and certainty level in the specified JSON format. Always choose one of the four categories for the classification, even if you're uncertain.

Claude 3.5:

```
{
  "classification": "[Category]",
  "certainty": [1-10]
}
```

Figure A5. Large Language Model Prompt for Classifying Research Activities

Notes: This figure shows the large language model prompt used for classifying research tasks. The output is in JSON format with keys for classification and certainty.

The prompt incorporates three features that improve the LLM's accuracy. First, I include high-level descriptions of each task category, selecting the phrasing that yielded the best performance after testing multiple options. Second, I request a certainty measure. This improves classification

even without considering the score, but also allows me to exclude observations below a threshold. In the main results, I omit observations with certainty scores below 3. However, the findings are robust to different thresholds, including using all observations. Third, I ask for the response in JSON format to ensure consistency.

I fine-tune the model using 2,000 manually classified observations, created in consultation with lab scientists to ensure accurate representation of researchers' intentions in their activity descriptions. Although the LLM performs reasonably well without fine-tuning, this process significantly improves its accuracy by allowing it to learn relevant details of the setting, including technical terms and acronyms frequently found in activity logs.

B.4.3 Validation

I validate the accuracy of the LLM classification on a set of 1,000 manually classified observations. Importantly, these observations are held out from the LLM fine-tuning. The results are presented in Table A11, showing that the model accurately classifies research tasks.

C Survey Details

C.1 Survey Questions

C.1.1 Introduction

Welcome and thank you for participating! We appreciate your willingness to share your experience with the newly implemented AI tool in your lab. This 15-minute survey aims to gather feedback on how the tool has been integrated into the materials discovery processes and how it has impacted your work.

Your accurate responses will play an important role in our lab's future decisions about AI use, such as changes to training and management practices. Please note that all responses to this survey are anonymous. We value your honest feedback as it is essential to improve our scientists' experience with the AI tool.

C.1.2 Background Questions

1. We will start by asking you a few questions about your background and role in the lab. Which of the following, if any, best describes your role in the lab?

    - Research Scientist
    - Lab Technician
    - Administrator
    - Other (please specify)

2. Is your research primarily focused on discovering or synthesizing new materials?

    - Yes
    - No

3. We will now ask a few questions about your use of the AI tool. Throughout the survey, "AI tool" refers to the new deep learning model recently introduced in your lab.

    Please indicate the date when you first gained access to the AI tool.

4. Please indicate the date when you first started using the AI tool in your work.

C.1.3 Questions About Use of AI Tool

5. Think about your current process for discovering materials. How important is the AI tool in this process?

    - Not at all important
    - Slightly important
    - Moderately important
    - Very important
    - Extremely important

6. Based on your experience working with the AI tool, indicate your level of agreement with the following statements:

    - The complexity of the AI tool makes it difficult to understand.
    - The complexity of the AI tool makes it hard to trust its results.
    - Compared to other methods I have used, the AI tool generates potential materials that are more likely to possess desirable properties.
    - The AI tool generates potential materials with physical structures that are more distinct than those produced by other methods I have used.

    (Scale: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree)

7. Consider the tasks you typically perform in your job. Think about the percentage of your work hours dedicated to the following categories:

    - *Generating ideas for potential materials:* This category includes activities like reviewing literature on existing materials or creating preliminary designs.
    - *Deciding which potential materials to test:* This category includes activities like analyzing simulation results or predicting which materials will be easy to synthesize.
    - *Synthesizing materials and testing their properties:* This category includes activities like preparing compounds or conducting experiments to test material properties.
    - *Other:* This category includes any tasks involved in the process of discovering materials that do not fall into the previous categories.

Answer the following questions as accurately as you can, based on your best recollection.

In the last three months, what percentage of your work hours were dedicated to the following categories?

- Generating ideas for potential materials
- Deciding which potential materials to test
- Synthesizing materials and testing their properties
- Other

8. Before the introduction of the AI tool, what percentage of your work hours were dedicated to the following categories?

    - Generating ideas for potential materials
    - Deciding which potential materials to test
    - Synthesizing materials and testing their properties
    - Other

9. On a scale of 1 to 10, how effective are you in each category of task?

    - Generating ideas for potential materials
    - Deciding which potential materials to test
    - Synthesizing materials and testing their properties

10. On a scale of 1 to 10, how effective are you in each category of task relative to other members of your team?

    - Generating ideas for potential materials
    - Deciding which potential materials to test
    - Synthesizing materials and testing their properties

11. Think about how you select which AI-generated material candidates to synthesize and test. Answer the following questions based on your experience working with the AI tool.

    Compared to previous methods, how would you rate the difficulty of evaluating AI-generated candidate materials?

    - Much easier

- Somewhat easier
- About the same
- Somewhat harder
- Much harder

12. How frequently can you identify an AI-suggested candidate compound as a false positive without attempting to synthesize the material?

    - Never
    - Rarely
    - Sometimes
    - Often

13. On a scale of 1 to 10, how useful are each of the following for evaluating AI-generated candidate materials?

    - Past scientific training
    - Past experience with similar materials
    - Intuition or gut feeling
    - Experience with similar AI tools

14. Please provide any additional comments you would like to share on your experience selecting AI-generated candidate materials for testing. (Text entry)

C.1.4 Questions About Job Satisfaction

15. Next, we will ask a few questions about how you enjoyed working with the AI tool.

    Compared to your previous workflow, how enjoyable is your work after gaining access to the AI tool on a scale of 1 to 10?

16. Think about the impact of the AI tool on your productivity. On a scale of 1 to 10, how enjoyable is your work as a result of this shift in productivity?

17. Think about the impact of the AI tool on the tasks you perform. On a scale of 1 to 10, how enjoyable is your work as a result of this shift in tasks?

18. What are the primary reasons that the AI tool makes your tasks less enjoyable? Please select all that apply.

- My tasks are more repetitive
- My tasks are less creative
- My tasks underutilize my skills
- I experience technical issues using the AI tool
- I don't understand the AI tool
- The AI tool makes it harder to assign credit for discoveries
- Other (please specify)

19. Based on your best understanding, are promotion and compensation decisions in your lab primarily based on:

    - Absolute performance
    - Performance relative to other scientists
    - Other
    - Not sure

20. Please provide any additional comments you would like to share on how you enjoyed working with the AI tool and why. (Text entry)

C.1.5 Questions About Views on Artificial Intelligence

21. Finally, we will ask a few questions about your views on artificial intelligence.

    Please report your current level of agreement with the following statements on a scale of 1 to 10:

    - AI will make scientists in my field more productive
    - I worry that scientists in my field will be replaced by AI
    - AI will change the skills needed to succeed in my field
    - I plan to reskill in order to collaborate with AI
    - I am satisfied with my choice of field

22. Based on your best recollection, please report your level of agreement with the following statements before you started using the AI tool on a scale of 1 to 10:

    - AI will make scientists in my field more productive

- I worry that scientists in my field will be replaced by AI
- AI will change the skills needed to succeed in my field
- I plan to reskill in order to collaborate with AI
- I am satisfied with my choice of field

23. Please provide any additional comments on how working with the AI tool has changed your views. (Text entry)

C.2 Respondent Characteristics and Attrition

Appendix Table A12 compares the characteristics of respondents and non-respondents, including their age, tenure, education, material of focus, innovative output, and treatment wave. Appendix Table A13 shows the number of respondents to each survey question. While there is some attrition, more than two thirds of respondents answered all questions.

D Task Model

There is a unit mass of scientists. Each is endowed with a unit of labor and possesses heterogeneous skills in idea generation and judgment, denoted γ_s^g and γ_s^j. Following Acemoglu and Autor (2011), the number of discovered compounds Y is produced by combining a continuum of research tasks $y(i)$ with $i \in [0,1]$:

$$Y = \left(\int_0^1 y(i)^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}} \tag{A1}$$

where $\sigma \in (0,1)$ is the elasticity of substitution between tasks. Shown in Figure A6, the interval $\left[0, I^j\right]$ represents idea generation tasks, the interval $\left(I^j, I^e\right]$ represents judgment tasks, and the interval $\left(I^e, 1\right]$ represents experimentation tasks. judgment and experimentation tasks are produced solely by scientist labor, ℓ. However, idea generation tasks in the interval $[0, \tilde{I}]$ can be produced with either scientist labor or AI, where $\tilde{I} < I^j$. Accordingly, the production function for task i is:

$$y(i) = \begin{cases} \gamma^g \ell(i) + \gamma^a a(i) & \text{if } i \leq \tilde{I} \\ \gamma^g \ell(i) & \text{if } i \in \left(\tilde{I}, I^j\right] \\ \gamma^j \ell(i) & \text{if } i \in \left(I^j, I^e\right] \\ \ell(i) & \text{if } i > I^e \end{cases}$$

where a is compute allocated to task i and γ^a represents the productivity of AI in each automatable task. I assume γ^a is sufficiently high such that it is optimal to assign all automatable tasks to AI.

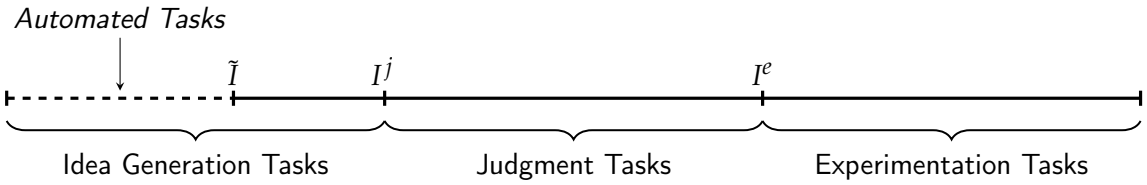


Figure A6. Illustration of Task Framework

This framework provides three predictions about how an increase in $\tilde{I}$ will impact task assignment. Let ℓ_s^g and ℓ_s^j denote labor allocated by scientist s to idea generation and judgment tasks, respectively. Let $L^g = \int_0^1 \ell_s^g ds$ and $L^j = \int_0^1 \ell_s^j ds$. Then:

1. $\frac{\partial L^j}{\partial \tilde{I}} > 0$ and $\frac{\partial L^g}{\partial \tilde{I}} < 0$. In other words, scientists reallocate labor from idea generation to judgment tasks after the introduction of AI.

2. Let $\Gamma_s^j \equiv \frac{\gamma_s^j}{\gamma_s^g}$. Then, $\frac{\partial \ell_s^j}{\partial \tilde{I} \partial \Gamma_s^j} \geq 0$ and $\frac{\partial \ell_s^g}{\partial \tilde{I} \partial \Gamma_s^j} \leq 0$. In other words, the reallocation effect is increasing in scientists' comparative advantage in judgment tasks.

3. Let $\gamma^g = \mathbb{E}\left(\gamma_s^g\right)$ and $\gamma^j = \mathbb{E}\left(\gamma_s^j\right)$. Then, $\frac{\partial Y}{\partial \tilde{I} \partial \gamma^j} > \frac{\partial Y}{\partial \tilde{I} \partial \gamma^g}$. In other words, automating idea generation tasks makes the returns to judgment relatively larger.

The intuition for these effects is straightforward. AI automates a fraction of idea generation tasks, making it optimal to reallocate scientist effort to judgment tasks. Because researchers are assigned to tasks based on their comparative advantage, $\frac{\hat{\gamma}_s^j}{\hat{\gamma}_s^g}$, this shift is larger for those with a comparative advantage in judgment. Both effects are consistent with the empirical results in Section 6.1.

As scientist labor becomes increasingly concentrated on judgment tasks, it raises the importance of skill in these tasks. The third prediction thus rationalizes the fact that AI disproportionately benefits scientists with strong judgment. Importantly, this implies that the lab could boost productivity by adjusting the composition of its workforce to prioritize judgment ability, which I provide evidence for in Section 6.3.