PNAS PERSPECTIVE OPEN ACCESS

Automating the practice of science: Opportunities, challenges, and implications

Sebastian Musslick[a,b,1], Laura K. Bartlett[c], Suyog H. Chandramouli[d,e,f], Marina Dubova[g], Fernand Gobet[c,h], Thomas L. Griffiths[f,i], Jessica Hullman[j], Ross D. King[k,l], J. Nathan Kutz[m], Christopher G. Lucas[n], Suhas Mahesh[o], Franco Pestilli[p,q], Sabina J. Sloman[r], and William R. Holmes[g]

Edited by Gordon Logan, Vanderbilt University, Nashville, TN; received March 8, 2024; accepted August 29, 2024

Automation transformed various aspects of our human civilization, revolutionizing industries and streamlining processes. In the domain of scientific inquiry, automated approaches emerged as powerful tools, holding promise for accelerating discovery, enhancing reproducibility, and overcoming the traditional impediments to scientific progress. This article evaluates the scope of automation within scientific practice and assesses recent approaches. Furthermore, it discusses different perspectives to the following questions: where do the greatest opportunities lie for automation in scientific practice?; What are the current bottlenecks of automating scientific practice?; and What are significant ethical and practical consequences of automating scientific practice? By discussing the motivations behind automated science, analyzing the hurdles encountered, and examining its implications, this article invites researchers, policymakers, and stakeholders to navigate the rapidly evolving frontier of automated scientific practice.

automation | computational scientific discovery | metascience | AI for science

> *"Though the world does not change with a change of paradigm, the scientist afterward works in a different world."*
> *- Thomas S. Kuhn, The Structure of Scientific Revolutions*

Automation is transforming every domain of scientific inquiry, from the study of functional genomics in biology (1, 2) to the derivation of conjectures in mathematics (3, 4). Recent advances in automation are accelerating hypothesis generation in chemistry (5–8), material discovery in materials science (9, 10), and theory development in psychology (11). These breakthroughs are not only garnering attention but also an uptick in funding and prizes dedicated to the automation of scientific practice (12–14). Furthermore, concurrent advancements in AI, software, and computing hardware are setting the stage for even more extensive automation within the scientific process (15–17).

The impact of automation in industry serves as a parallel to its potential in science. In the early 20th century, industrial automation began with mechanized assembly lines, revolutionizing manufacturing efficiency and output. The introduction of robotics and computer-aided manufacturing marked another leap, enabling precision and consistency previously unattainable by human labor. Today, industry-wide automation facilitates not just cost-efficient mass production, but also customized, adaptable, and intelligent manufacturing processes. This evolution demonstrates the capacity of automation to radically redefine operational paradigms.

Drawing parallels to scientific practice, one can anticipate a similar trajectory of profound change, where automation could accelerate discovery, reshape research methodologies, and redefine the very nature of scientific inquiry. At the same time, automation in industry had significant impacts on workers and the kind of products that dominate the marketplace. It is thus important to consider parallel impacts in the scientific setting which may have negative consequences for science and society.

In this perspective, we evaluate what automation should and can achieve for scientific practice. In doing so, we outline the current state of science automation, drawing on recent examples from different domains of science. Furthermore, we examine technological advancements that open new avenues for automation in science and discuss current bottlenecks. Finally, we highlight a selection of practical and ethical considerations and discuss how automation may lead scientists to work in a different world, one where traditional methodologies are redefined and new meta-paradigms for science emerge.

What Are the Bounds of Automating Scientific Practice?

Scientific practice can be defined as the set of methods and processes used by scientists to acquire knowledge about the natural world. Automation, in its broadest sense, refers to the use of technology to perform tasks with minimal human intervention. In the context of scientific practice, automation

Author affiliations: [a]Institute of Cognitive Science, Osnabrück University, 49090 Osnabrück, Germany; [b]Department of Cognitive and Psychological Sciences, Brown University, Providence, RI 02912; [c]Centre for Philosophy of Natural and Social Science, The London School of Economics and Political Science, London WC2A 2AE, United Kingdom; [d]Department of Information and Communications Engineering, Aalto University, FI-00076 Espoo, Finland; [e]Department of Computing Science, University of Alberta, Edmonton, AB T6G 2S4, Canada; [f]Department of Psychology, Princeton University, Princeton, NJ 08544; [g]Cognitive Science Program, Indiana University, Bloomington, IN 47405; [h]School of Psychology, University of Roehampton, London SW15 4JD, United Kingdom; [i]Department of Computer Science, Princeton University, Princeton, NJ 08544; [j]Department of Computer Science, Northwestern University, Evanston, IL 60208; [k]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, United Kingdom; [l]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 412 96, Sweden; [m]Department of Applied Mathematics and Electrical and Computer Engineering, University of Washington, Seattle, WA 98195; [n]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom; [o]Department of Materials Science and Engineering, University of Toronto, Toronto, ON M5S 3E4, Canada; [p]Department of Psychology, The University of Texas, Austin, TX M5S 3E4; [q]Department of Neuroscience, The University of Texas, Austin, TX M5S 3E4; and [r]Department of Computer Science, University of Manchester, Manchester M13 9PL, United Kingdom

Author contributions: S. Musslick drafted the paper; and S. Musslick, L.K.B., S.H.C., M.D., F.G., T.L.G., J.H., R.D.K., J.N.K., C.G.L., S. Mahesh, F.P., S.J.S., and W.R.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: sebastian.musslick@uos.de.

Published January 27, 2025.

specifically denotes the use of technological tools and systems to carry out scientific tasks or processes traditionally performed by human scientists.

The bounds of automation within scientific practice hinge on at least two questions: first, is there a desire and justification for automating a given scientific practice? This question touches upon *goal-related bounds*—the alignment of automation with the overarching goals of science. Second, what factors characterizing a scientific practice influence the feasibility of automating that practice? This aspect focuses on the *technological bounds*, assessing the practicality and potential constraints of applying automation in science.

Goal-Related Bounds: What Automation Should (Not) Achieve. Science is driven by normative and epistemic goals. Here, we discuss arguments for and against automation serving these goals.

The normative goals of science involve ethical, moral, and societal values guiding both basic and applied science. One such goal may be to enable cheap and fast discoveries that advance human health. Along these lines, automation can serve to yield faster scientific discoveries with fewer resources. This is particularly desirable in the applied sciences, e.g., for identifying novel drugs or treatments. Thus, automation can aid scientific practice if societal needs are clear and research questions are well defined. However, the process of identifying a research question itself requires considering societal needs or the interests of the scientific community. As noted in the *Opportunities* section below, generative AI can integrate large bodies of literature to identify societally and scientifically important gaps in our knowledge that are worth filling. However, since the relevant normative considerations inherently depend on evolving human contexts, it can be argued that humans ought to always be involved in and monitor the degree to which scientific practices achieve these objectives (18). Consequently, full automation in these areas might not only be impractical but also undesirable, underscoring the indispensable role of human scientists in addressing the normative dimensions of science.

The epistemic goal of science is to understand the natural world through description, prediction, explanation, and control. As discussed in the sections that follow, advances in machine learning can aid in automating the description or explanation of natural phenomena. Such automation can help reduce human errors and biases, leading to more accurate predictions and better control of natural phenomena. Even more so, automation may help bypass or augment the cognitive capacities of human researchers (19), enabling degrees of prediction and control unachievable for human cognition alone. For example, machine learning models can generate millions of proposals for novel materials that lie beyond human intuition (9). Yet, the increase in precision achieved through automation presents an epistemic dilemma, as automation can limit human understanding. In the basic sciences, advancement of human understanding may be more desirable than merely improving predictability through automation. The complexity of a machine learning model, for example, might enhance its ability to accurately predict new stable materials, but concurrently obscure the process by which these predictions are made for human scientists. This scenario illustrates a potential conflict between the scientific objectives of enhancing prediction, on the one hand, and enabling human understanding, on the other (*Practical Implications*). This suggests keeping human scientists involved in the scientific process rather than minimizing their involvement. Meanwhile, in applied sciences and engineering, the focus might shift toward maximizing prediction and control, providing a stronger case for automation of scientific practice.

Technological Bounds: What Automation Can (Not) Achieve. The technological bounds of automation hinge on the difficulty of automating scientific tasks. Here, we discuss four factors characterizing this difficulty (Fig. 1). These factors indicate both opportunities and barriers to automation, thereby guiding the identification of areas within scientific practice where automation can be most effectively implemented or where it may face challenges.

The first factor concerns the *availability and quality of inputs* that a scientific task requires. Some tasks, such as identifying a research question, rely on diverse and sometimes subjective inputs, including peer opinions, news articles, or funding announcements. Such inputs may not be trustworthy, widely accessible, or structured for machine processing, posing a challenge to automation.

Another limiting factor for automation is the *computational complexity* of algorithms available to perform a scientific task. For example, identifying an appropriate experiment for testing a research question may require taking into account numerous decision variables (e.g., internal validity, resources needed, novelty) and searching an exponentially increasing space of possible experimental paradigms, which can be computationally intractable.

A related, yet often overlooked, factor influencing the automation of scientific tasks is the *complexity of required hardware engineering*. As stated in Moravec's paradox, sensorimotor tasks, like executing invasive brain recordings or social experiments, require advanced solutions in robotics to facilitate automation, which can pose more significant challenges to automation compared to cognitive tasks (20).

Finally, some tasks are difficult to automate because of the *subjectivity of the task goal*. Some scientific goals cannot be easily turned into a well-defined objective, which is required to communicate it to a machine. For instance, choosing between scientific models can be a matter of personal preference (21).

While the four factors collectively dictate the automatability of scientific tasks, they can be considered interdependent. For example, the automated discovery of scientific equations long relied on search methods with high computational complexity, such as evolutionary computation or brute force search, to identify a set of equations that best describes a given dataset (22, 23). However, the ability to collect large datasets cheaply, paired with improvements in computing hardware, enables the application of "data-hungry" but computationally tractable machine learning algorithms for equation discovery (24–27). This approach reduces computational complexity, illustrating how enhancements in one factor can compensate for limitations in another.

Automation in Current Scientific Practice

Existing approaches to automating science target tasks with readily available inputs, computational complexity, and hardware demands that align well with current technological capabilities and clear task goals. Accordingly, efforts at

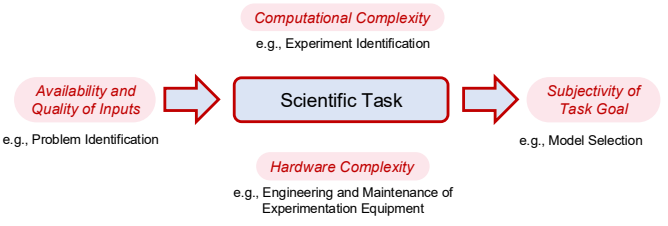

Fig. 1. Factors determining the technological reach of automation in scientific practice.

automatization in science have mostly been confined to tasks characterized by clearly specified objectives and well-defined subtasks, which include instances of quantitative hypothesis generation, experimental design, data collection, and quantitative analysis and inference. While covering all advances is out of the scope of this article, we highlight a subset of these approaches, focusing on cases that facilitated novel discoveries.

Hypothesis Generation. Hypothesis generation is the development of testable statements that are based on observations, existing knowledge, or theory. Advances in automated hypothesis generation were primarily driven by two factors: improvements in computer algorithms, and the availability of large datasets.

Initial automated hypothesis formation approaches relied on symbolic reasoning systems. For example, in organic chemistry, logical deduction based on existing knowledge was employed to formulate hypotheses about the chemical constituents of body fluids (28). Furthermore, quantum simulations, facilitated through cloud computing, became the backbone of hypothesis generation for materials properties (29, 30). The development of efficient search algorithms further expanded the scope of automated hypothesis formation to areas with large hypothesis spaces (3). For instance, hypothesis generation in mathematics leveraged efficient machine learning algorithms to identify novel conjectures about fundamental constants (3). Finally, deep learning enabled more breakthroughs in chemistry. A landmark achievement in this area is the nobel-prize winning AlphaFold, which predicts 3D protein structures from amino acid sequences, facilitating the development of drugs (6).

The availability of large datasets led to further advances in automated hypothesis formation. One example is the field of biomedicine, where large gene databases led to a surge in hypothesis generation with computational methods, e.g., using data mining and network analysis to propose genes that may be linked to diseases (31, 32). Similarly, existing materials databases provided sufficient information for machine learning methods to generate over 2.2 million proposals for novel materials that, so far, escaped human intuition (9).

Experimental Design. The problem of automated experimental design is to systematically identify the most informative experiment to address a particular hypothesis or scientific question. The informativeness of an experiment can be evaluated in various ways. Some automated experimental design methods are geared toward identifying the experimental conditions that minimize the influence of nuisance variables–experimental variables that are not of interest but can pollute the informativeness of intended experimental manipulations (33, 34). Other methods aim to find experimental conditions that are well suited to identify a scientific model of interest (35–37). This problem of experimental design is closely related to the problem of active learning in machine learning research (2, 38–40), which seeks to identify data points that can best inform a machine learning model when included as training data. A prominent active learning method used for scientific practice is Bayesian optimal experimental design, which has been successfully applied in various fields, including psychology (36, 37, 41, 42), neuroscience (43), physics (44, 45), biology (46, 47), chemistry (48, 49), materials science (50–52), and engineering (53). For example, in the domain of psychology, Bayesian optimal experimental design led to the discovery of novel models of how humans discount the future relative to the present (54).

While automated experimental design methods can facilitate efficient data collection and strong inferences, their efficacy can be compromised if the underlying assumptions are violated or if the scientific model is incorrectly specified (55–57). This limitation led to unexpected findings in simulation studies, where random sampling of experimental conditions outperformed theory-driven approaches to experimental design (38, 58), and where uniform sampling outperformed adaptive approaches in learning continuous relationships (59).

Another limitation of current approaches to automated experimental design pertains to their scope, as they focus on navigating a predefined space of experimental manipulations. Exploring novel research directions, however, often involves identifying completely new experimental manipulations (60).

Data Collection. Data collection, often a time-consuming and costly aspect of empirical research, is a significant bottleneck in scientific discovery. Accordingly, automated tools for data collection emerged as some of the most impactful innovations in accelerating the pace of science. These tools span a wide range of applications and fields: fitness trackers revolutionized public health studies (61), continuous glucose monitors are providing critical insights into nutrition and diabetes research (62), and automated weather stations enhanced meteorological predictions (63). In addition to providing streams of real-time data for ongoing analysis, these automated systems can minimize human observation and experimenter biases. Experimenter bias occurs when the beliefs, expectations, or preferences of the researcher unconsciously influence the conduct or outcome of an experiment. Automating data collection in animal studies helped to eliminate experimenter bias, resulting in refutations of previous results, such as the evidence for statistical learning ability in newborn chicks (64). A particularly noteworthy advancement in the behavioral sciences was the adoption of web-based experiments, especially during the COVID-19 pandemic. Online platforms and interfaces for recruiting and conducting experiments did not only facilitate the collection of behavioral data at a time when traditional lab-based studies were impractical, but they also broadened the scope and diversity of participants (65–67). Automating data collection also generated opportunities for automating other elements of behavioral research, such as adopting adaptive experimental designs that change based on the responses of participants (68) or collecting larger datasets that can support the use of machine-learning algorithms (11).

Statistical Inference. The automation of statistical inference transformed dramatically from the era of manual computations, a reality echoed in old statistical textbooks filled with computation-simplifying shortcuts. The introduction of computers altered statistical methodologies, sometimes even leading to their replacement by machine learning techniques. For example, modern statistical inference engines, like Stan, leverage techniques such as Markov Chain Monte Carlo (MCMC) for efficient sampling of model parameters (69). Tools for likelihood-free inference enable the analysis of statistical models that are not mathematically tractable. Furthermore, frameworks such as Bayesian Workflow (70) and platforms such as the Automatic Statistician (71) are streamlining complex processes like Bayesian inference and the construction of traditional statistical models. The automation of statistical inference, however, is mostly confined to the deduction of new knowledge based on prespecified statistical models.

Scientific Inference and Model Discovery. Scientific inference, unlike statistical inference, involves generating hypotheses about observations (abduction) and generalizing from observations to laws or broader theories (induction). The automation of

scientific inference is termed computational scientific discovery and has so far centered on identifying models or laws that elucidate specific phenomena (22, 23, 72). One instance of computational scientific discovery involves the identification of equations ("symbolic regression") to uncover quantitative laws governing a given dataset. Early efforts relied on heuristic search techniques to rediscover insights from mathematics (73, 74) or physics (75). Advances in machine learning and high-performance computing facilitated equation discovery, building on reinforcement learning (26), genetic algorithms (25, 76, 77), MCMC sampling (78), mixed-integer nonlinear programming (79), or gradient-based search techniques (24, 27, 80, 81). However, most forms of computational model discovery are limited to the rediscovery of existing knowledge. Possible exceptions include the discovery of scaling laws and boundary equations in plasma physics (82) and novel models of human decision-making (11).

Closed-Loop Automation Spanning Multiple Scientific Practices. Demonstrations of successful closed-loop automation in empirical research—implementing iterations between experimental design, data collection, and model discovery—mark a significant progression for automated scientific practice. One pioneering example is the robot scientist Adam (Fig. 2*A*), which was the first fully automated machine to discover novel scientific knowledge (2). Adam investigated the functional genomics of the yeast *Saccharomyces cerevisiae* and discovered the function of locally orphan enzymes—enzymes known to be in yeast but for which the gene(s) encoding them were unknown. The successor of Adam, Eve, is a robot scientist designed for early-stage drug development (39), which identified chemical compounds that outperformed standard drug screening. Eve's most significant discovery is that triclosan (an antimicrobial compound commonly used in toothpastes) may aid against malaria (39, 83, 84). Another example of a closed-loop discovery system in biology is Wormbot-AI, a platform designed to autonomously conduct experiments on the longevity of worms, capable of testing thousands of interventions annually (85, 86).

Complete automation also gained momentum in materials science and chemistry, where efforts are focused on integrating hypothesis generation, decentralized experimentation, and cloud-based decision-making. For instance, modular robotic platforms, driven by machine learning algorithms, were used to optimize material properties by varying synthesis conditions (87–89). One notable example is A-Lab (Fig. 2*B*), an autonomous laboratory for the solid-state synthesis of inorganic powders, which leverages a combination of active learning and machine learning models trained on the literature, to propose and synthesize novel material candidates (10).

Additionally, behavioral research became amenable to closed-loop automation with the ability to collect data via online experiments. Open-source tools like AutoRA (90) facilitate closed-loop research by integrating automated model discovery, experimental design, and experimentation in empirical research. AutoRA effectively interfaces with web-based platforms for automated data collection, integrating the acquisition of behavioral data from human participants. In addition to facilitating the discovery of novel behavioral phenomena and cognitive mechanisms, AutoRA served as a computational testbed for philosophy of science, exposing cases where random experimentation outperforms model-guided experimentation (38).

Finally, researchers introduced a Large Language Model (LLM)-based agent for automating empirical machine learning research, from idea development and experimental design to execution and data analysis, e.g., for improving existing machine learning models (91). Notably, this system also leveraged LLMs to automate the writing and peer review of the resulting research manuscript, with the computational cost of one article estimated to be just 15 USD.

Despite their potential to accelerate scientific discovery, it is important to recognize that the pioneering examples of closed-loop automation are currently confined to specific, automatable research steps and operate within a constrained range of experimental design and model spaces as delineated by human researchers (cf. Fig. 2).

Future Opportunities

Existing approaches for automating scientific practice primarily target tasks for which a) high-quality data is available, b) the computational complexity can be addressed by current algorithms, c) hardware complexity is manageable, and d) task goals are well-defined (cf. Fig. 1). The most promising prospects for future automation in scientific practice are found in tasks traditionally limited by human cognitive capacities. This includes areas requiring the processing of large volumes of high-dimensional data or exhaustive literature searches. In this section, we highlight a few technological trends that promise to push the boundaries of science automation along these lines.

Data Collection, Standardization, and Sharing. Advancements in cost-effective data collection, standardization, and sharing significantly boost the automatability of scientific practices, particularly those dependent on empirical data. For example, in the behavioral sciences, the utilization of crowd-sourced experimentation platforms like Amazon Mechanical Turk and Prolific revolutionized the efficiency of behavioral data collection. Additionally, LLMs that can mimic human behavior were proposed as proxies for participants, aiding in the acquisition of large-scale datasets (92). Once acquired, such large—yet cost-efficient—datasets can empower data-hungry machine learning algorithms, enabling them to uncover novel, and more precise models of human behavior (93–96). Large-scale data collection, however, still bears significant hardware challenges, e.g., for collecting biological samples from a large number of participants (*Future Challenges*). Nevertheless, the data quality needed for automated analysis techniques should be complemented by data standardization and sharing.

Scientific data-sharing platforms, such as the Open Science Framework, facilitated the availability and accessibility of data needed for automated analyses and computational discovery. The potential of data sharing and standardization is perhaps best illustrated in materials science, where databases for stable materials enabled the prediction of large quantities of new materials (9). Other scientific domains profit from similar efforts. For example, in neuroscience, archives like DANDI, OpenNeuro, DABI, and BossDB allow researchers to share data using community standards (97), such as BIDS for neural data (98).

Combining Data-Driven and Knowledge-Driven Discovery. A particularly promising approach to automating scientific discovery is the integration of preexisting human knowledge into the discovery process. Traditionally, data-driven discovery methods operated with minimal prior knowledge about the specific domain of scientific inquiry. This pure data-driven approach makes such methods particularly susceptible to noisy data. However, recent work demonstrates that incorporating prior theoretical knowledge can significantly aid in recovering scientific models from noisy datasets. For example, Bayesian symbolic regression exhibits greater efficacy in recovering equations from noisy

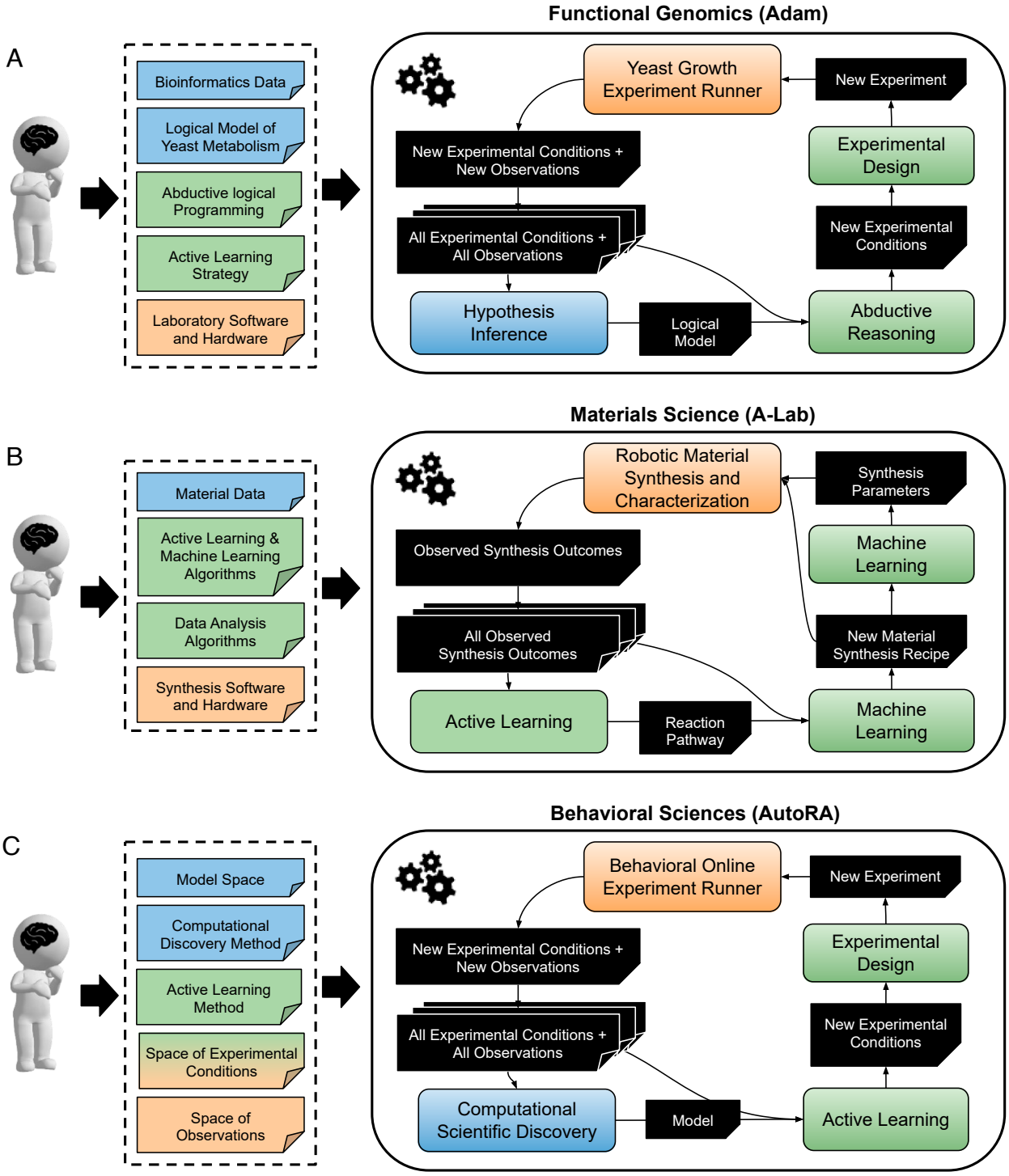

Fig. 2. Closed-loop automation systems. (*A*) Adam for functional genomics. (*B*) A-Lab for materials science. (*C*) AutoRA for behavioral sciences. Dashed boxes list knowledge and processes provided by human researchers.

data when given priors about scientific equations extracted from Wikipedia (78, 99). Similarly, embedding prior knowledge in the form of general logical axioms proved instrumental in rediscovering complex scientific laws, including Kepler's third law of planetary motion and Einstein's relativistic time-dilation law (79, 100). Furthermore, experiments with the BacterAI, which uses active learning for the automated study of microbial metabolisms, have demonstrated the advantage of leveraging relevant prior knowledge (101). Specifically, when the metabolic model trained on one bacterial species was retrained for the species of interest, it more efficiently discovered its metabolic model compared to starting the learning process from scratch, despite the two species differing in their metabolic capabilities. These examples highlight the benefits of combining data-driven and knowledge-driven approaches for automated model discovery.

The benefits of knowledge-driven discovery are, however, fundamentally limited by the quality of prior knowledge. For

example, Bayesian adaptive experimentation can be misled if prior knowledge mischaracterizes the data (102, 103). Thus, data-driven approaches to computational model discovery become particularly beneficial when dominant scientific models in the empirical sciences are more informed by (wrong) theory versus data. This is evident in computational models of human reinforcement learning, which predominantly rely on classic machine learning algorithms (104). Recent work demonstrated that a data-driven model discovery can uncover novel reinforcement learning models that better explain human learning than traditional models (95).

Finally, a notable area of progress in automated model discovery is the analysis of high-dimensional datasets, such as fluid dynamics captured in video format, through reduced-order modeling. This process involves learning a low-dimensional representation of the dynamics inherent in complex data and then decoding the governing equations of these latent dynamics (105–108). Similar approaches were developed to automate the discovery of neural data embeddings correlating with behavioral dynamics (109). These approaches promise to extend the reach of automated model discovery to high-dimensional naturalistic datasets. beyond experimental control.

Generative AI and LLMs. Generative AI and LLMs offer paths toward automating scientific practices that have historically been challenging due to their computational complexity and qualitative nature (8, 16, 91, 110). Among these are the synthesis and integration of literature, and documentation of findings.

Researchers argued that LLMs show promise in enhancing literature reviews, a task currently limited by the cognitive constraints and language barriers of human scientists (111, 112). Whereas humans may only be able to parse and integrate a few hundred articles into a literature review—the scope of which is heavily influenced by the expertise and biases of the researcher—LLMs may accomplish literature synthesis in the order of thousands or millions of articles. Critically, LLMs can take into account articles written in different languages, thus helping to counter the dominance of Western perspectives in scientific literature. Thus, LLMs can assist in extending or even bypassing human researchers' cognitive limitations. A notable application of LLMs for the purpose of literature synthesis is Elicit, which utilizes LLMs trained on paper abstracts to support and help researchers extract relevant information from the scientific literature (112). Another instance of such assistance is an LLM-based "co-scientist" for chemical research, which improved the planning of chemical syntheses based on information available on the internet, and aided in the navigation of extensive hardware documentation (8). Additionally, BrainGPT—an LLM fine-tuned to the neuroscience literature—demonstrated the capability to outperform human experts in predicting the results of neuroscience experiments (113).

Combined with their capability for literature synthesis, LLMs can foster the discovery of new research directions and hypotheses (91). Along these lines, LLMs have the potential to expand experimental design spaces, addressing a common bottleneck in automated scientific practice. While traditional automated experimentation is confined to researcher-defined variables (cf. Fig. 2), LLMs could identify novel experimental variables of interest, thus broadening the scope of scientific inquiry. However, it can be argued that LLMs risk rediscovering already known hypotheses and experiments (18).

Once experiments are designed, LLMs may aid in the balanced documentation and communication of the research study, including the automated documentation of research code (114, 115). Apart from aiding in the construction of research articles, LLMs can enable automated translation into multiple languages. This advancement is particularly beneficial for nonnative English speakers and is an example of how automation and AI can address ethical challenges in science. Nevertheless, literature reviews conducted by human scientists serve not only to synthesize knowledge but also to build and refine the conceptual frameworks of evolving scientists—a process that is critical to scientific training and that is challenged by the overuse of LLMs for literature synthesis.

Future Challenges

Despite recent advances and opportunities for the automation of science, there remain substantial obstacles. This section examines technological bounds rooted in four bottlenecks (cf. Fig. 1): limited availability and quality of data, intractable computational complexity of certain scientific tasks, lack of required hardware, and subjectivity in assessing the outputs of scientific tasks. These bottlenecks highlight why barriers to automation remain difficult to surmount in the basic sciences (as opposed to engineering), at least with the technologies and methodologies currently at our disposal. Addressing these challenges will require significant interdisciplinary efforts to identify solutions that enable automation beyond a few selected domains of scientific inquiry.

Limited Availability and Quality of Inputs. Prior applications of computational discovery, such as in chemistry (5, 7, 116) and materials science (9, 10), relied on standardized formats for both data and scientific hypotheses that are easily parsed by machine learning algorithms. However, most tasks of scientific practice rely on a diversity of representations for scientific knowledge. For example, computational models in the natural sciences are expressed in various formats, such as equations embedded in scientific articles or computer code written in different programming languages. Without standardization across disciplines, automated systems face significant challenges in drawing parallels or applying concepts from one domain to another. Efforts to standardize the representation of scientific models and other forms of scientific knowledge promise to ease the automation of scientific practices relying on such knowledge (117). However, even if data are standardized and widely available, ensuring their quality remains critical. For instance, literature synthesis enabled by LLMs may be unfruitful or even misleading if fraudulent or unreproducible papers are included as inputs to these models. Therefore, robust quality control measures must accompany standardization efforts to maintain the integrity and usefulness of automated systems.

Computational Complexity. One of the fundamental bottlenecks in the automation of scientific practice lies in the computational complexity of many scientific tasks. For example, complexity analyses within the realm of cognitive science indicate that scientific discovery in cognitive science may be computationally intractable in principle, even with unlimited availability of data (118). These theoretical results suggest that uncovering a definitive "ground-truth" theory may be beyond the reach of computation.

One potential critique of leveraging computational methods for scientific discovery hinges on the incomplete comprehension of the cognitive processes, and the concomitant computational complexity underlying it. One may argue that without a full grasp of how humans tackle scientific inquiries, designing algorithms capable of similar feats seems implausible. However, at least two counterarguments challenge this perspective. First, replicating natural processes is not a

prerequisite for solving problems. For instance, modern airplanes achieve superior lift not by emulating the flapping motion of birds but through aerodynamically efficient designs. Second, a deep understanding of cognitive phenomena is not a strict requirement for automation, as evidenced by the capabilities of LLMs to produce coherent natural language sequences without humans having a complete scientific understanding of language generation. Nonetheless, this gap in understanding underscores the importance of implementing robust evaluation methods to ensure accuracy and mitigate any potential negative impacts of automating scientific processes.

Hardware Engineering. The advancement of automated science is significantly hindered by current limitations in laboratory robotics and hardware engineering. For instance, executing complex biological or physics experiments remains challenging. Moreover, while robotic automation has been successfully implemented in certain areas, such as with the robot scientist concept (1, 2, 101, 119), its application is primarily limited to clearly defined engineering problems. Yet, even well-defined engineering problems must manage the noise and variability inherent in the data collected by sensors, which can dramatically affect the reliability of scientific outcomes. Therefore, while progress has been made in automating scientific practice, developing more sophisticated hardware to handle complex, noisy data is crucial for its broader adoption and effectiveness.

The automation of hardware tasks in scientific practice is also hindered by the need for highly specialized equipment, leading to significant capital expenditures, often exceeding millions of dollars. Such custom-built hardware is typically field-specific and lacks versatility for reuse in other scientific domains. This challenge is evident in the limited cross-utilization of hardware between disciplines, as seen in the relatively small amount of equipment that materials scientists have been able to adapt from the more heavily automated field of drug discovery. Addressing this issue requires a strategic approach where, for each scientific field, scientists identify and develop a core set of automated hardware that can deliver the greatest impact. This not only involves designing equipment that meets the unique needs of each field but also balancing specificity with adaptability, to maximize utility and cost-effectiveness.

Subjective Goals of Scientific Tasks. More than in engineering, practices in basic science are inherently subjective in how the outcomes of those practices are evaluated. This challenge is particularly evident in developing AI capable of generating novel and impactful scientific ideas. Novelty and impact involve a high degree of subjectivity and variability, making it difficult for these systems to replicate human judgment in the space of scientific inquiry (16). This issue is compounded by the personal aspect of scientific practice. The selection of scientific projects is guided by the personal experience and perspective of human scientists. Diversity in such perspectives paired with interdisciplinary exchange can lead to a greater diversity of ideas in human scientific systems (120)—a dimension that AI currently cannot emulate without explicit instruction. Furthermore, the lack of standardized solutions in many scientific areas means that automating these tasks risks constraining exploration, which is vital for scientific advancement.

Moreover, interpretation of data patterns and hypothesis generation often necessitates human judgment to translate statistical regularities into meaningful scientific interpretations. Techniques like topic modeling, while effective in identifying text co-occurrence patterns, require human insight to align these patterns with relevant scientific constructs (121). The role of human judgment is perhaps best exemplified in serendipitous discovery, often stemming from unexpected failures or results. For example, Alexander Fleming's discovery of penicillin began with the accidental contamination of a Petri dish. Instead of discarding it, his observation of the bacteria being killed by the mold led to the development of the first antibiotic. These aspects highlight the crucial role of human judgment in scientific discovery.

Implications

Although the automation of science currently faces significant limitations, the extent to which it will evolve in the mid- to long-term remains an open empirical question. As advancements in hardware and algorithms continue, the range of practices subject to automation is likely to expand. In this section, we explore the practical and ethical consequences of this trend.

Practical Implications.

The role of human scientists and the paradox of automation. The advancement of automation in scientific practice raises considerations regarding the future role of human scientists. On the one hand, it can be argued that automation reduces the need for human involvement. Scientific discovery systems may become able to monitor themselves and tune themselves to optimal performance—potentially excluding humans from the scientific discovery loop. On the other hand, it can argued that the greater the efficiency of an automated system, the more vital the role of human oversight (122). A critical assumption underlying this "paradox of automation" is that automation is not perfect; the potential for accumulating errors necessitates human intervention. If automation were flawless, human oversight would be unnecessary, and the paradox would not exist. However, for tasks with sufficient complexity and uncertainty, this paradox suggests that, in highly automated environments, human contributions, though less frequent, are more critical. This may specifically apply to tasks that demand subjective assessment or the synthesis of complex data, such as reviewing scientific literature, as well as high-level responsibilities such as strategic allocation of funds for scientific inquiry.

Even in the absence of subjective assessment, there are inherent risks associated with automation. For instance, an error within an automated system can lead to a cascade of compounded errors, persisting and potentially amplifying until the system is either corrected or deactivated. This may be particularly problematic for automation methods whose decision-making processes are not completely predictable, as is the case for many machine learning algorithms. This unpredictability raises the issue of responsibility for unintended consequences such as injuries. Given the potential severe legal and financial implications of compounding errors in automation, the involvement of human scientists, even in areas where automation is technically feasible, may prove to be more efficient, practical, and safe in the near future. Thus, the paradox of automation underscores the lasting importance of human expertise and the need for a balanced approach that combines automated systems with human judgment.

Research training. With increased automation of science, there arises a need to reevaluate and adapt scientific education. This new landscape calls for training that encompasses not only traditional scientific knowledge but also skills for effectively working alongside automated scientific discovery systems. For instance, obtaining valuable outputs from LLMs is becoming an essential skill. Moreover, scientists will need to develop competencies in understanding and evaluating the functioning

and outputs of automated systems, as is already demanded for statistical software (47). This shift implies a growing demand for engineers, scientists, and technicians proficient in advanced STEM skills.

Research evaluation. The current pace of science is primarily determined by our capacity to carry out the research itself. Laboratory studies in fields like biology and chemistry can take years, contrasting with the relatively quick peer review process. However, if advancements in automation enable research to be conducted and documented several magnitudes faster (91), this could lead to a substantial increase in the rate of research article submissions. Such a scenario would further strain the already pressured peer review system. One potential solution could be the automation of peer review, possibly through the use of LLMs; however, this approach has already faced restrictions and bans in certain contexts due to concerns about its efficacy, reliability, and confidentiality (123). Another potential solution is for journals to require that articles generated by automated systems be accompanied by critical evaluations from corresponding human authors. This ensures that human researchers retain comprehension and oversight of what is being submitted while also serving as initial reviewers of the work generated by their automated systems. Either way, this shift would necessitate a reevaluation of the peer review process, ensuring it remains rigorous and effective in the face of increased scientific productivity.

Scientific methods. The automation of scientific practice has the potential to bring about a shift in scientific methods that goes beyond mere acceleration of scientific discovery. As discussed above, the use of machines for scientific discovery allows us to move beyond the cognitive and physical constraints inherent to human scientists (19). Consider, for example, the principle of parsimony in the construction of scientific models. Traditionally, parsimonious models have been favored for their superior generalization, ease of interpretation, and communicability among human scientists. However, as discussed in ref. 21, recent studies suggest that highly complex models can, under certain conditions, surpass the generalization capabilities of simpler ones (124), leading to unprecedented advances in scientific research (e.g., for 3D protein folding (6) or material discovery (9)). Moreover, as explored in ref. 21, the development of such complex models is often a prerequisite for discovering successful parsimonious models (e.g., refs. 125–127). This ability of machines to explore and develop models with a level of complexity beyond what is readily interpretable by humans opens up new avenues for scientific progress, less constrained by human cognitive limitations. However, as discussed above, for basic science, there is epistemic value in *human* understanding that may outweigh the predictive power of AI scientists.

Another consequence of automation concerns the ways in which empirical research is conducted. For example, automated systems can hypothesize and experiment in design spaces far beyond the reach of human cognitive capabilities (9, 119). Furthermore, the ability to collect large amounts of data cheaply may obviate frequent iterations between hypothesis generation, experimental design, and data collection. Instead, with the availability of large datasets, the problem of scientific discovery can be transformed into a model discovery problem more amenable to machine learning (11, 94, 128). However, it is important to recognize that the success of a one-time large-scale data collection hinges on a well-defined experimental design space and the stability of the system under study, as constant changes in the system can undermine the effectiveness of this approach. Accordingly, adaptive experimental design may be needed to identify suitable design spaces (58).

Ethical Implications.

Biases. While human biases influence every aspect of scientific work, automated systems are not immune to bias. They can inherit biases from their creators, the construction process, the data they use, and their training format (129). Examples include discriminatory biases in facial recognition technology (130), unrepresentative sampling in psychological experiments (116), and discrimination in automated participant recruitment processes (131). Moreover, automated literature reviews do not escape the biases inherent to the existing literature. These biases can be democratized and exacerbated by the pace of these systems, especially when they are uninterpretable or operate as "black boxes." However, a potential advantage is that biases in automated systems may be easier to correct than in humans, such as by using more diverse data, or by aligning automated systems with societal norms.

Value alignment and responsibility. The risk of harmful biases and outcomes of automated processes call for their value alignment with broader societal norms. This is particularly crucial as automation could potentially ease the path for malevolent entities to conduct research detrimental to society, such as developing chemical or biological weapons. Such outcomes underscore the necessity of ethics dedicated to addressing these issues, ensuring that automated scientific advancements align with human values.

Consequences of automation also bring about the issue of responsibility: if a scientific discovery that affects the wider society is based on an automated process, who is responsible? The accountability for effects arising from harmful scientific practice remains ambiguous-whether it lies with the system's creator, its user, or the implementer of societal changes based on the system's output. This issue parallels broader debates in AI, such as liability in self-driving car accidents or the creation of automated artwork. Additionally, the potential misuse of powerful systems (e.g., a system suggesting harmful drug treatments) necessitates robust safeguards. The same applies to potential violations of data privacy. When automated systems generate contentious theories or design ethically questionable experiments, human oversight and responsibility are imperative. Importantly, ethical guidelines are often formulated by the institutions developing the systems (132), highlighting the need for an external framework that can hold institutions accountable.

Conclusion

While the automation of scientific practice is currently confined mostly to well-defined engineering and discovery problems, there is the potential for automation to pervade a large part of scientific practice. We suggest that this trend represents not merely a series of quantitative changes, such as increased efficiency or precision in science, but brings about a fundamental shift in the conduct of science. The integration of AI into scientific practice has the potential to overcome human cognitive limitations, thereby expanding our capabilities for discovery. Yet, this advance is not without challenges—data availability, computational complexity, engineering demands, and subjectivity of scientific task goals mark the technical boundaries of current automatability. Furthermore, normative goals of science—anchored on societal values—potentially make complete automation of scientific practice neither desirable nor feasible. Finally, this qualitative shift comes with practical and ethical challenges that call for interdisciplinary and collective efforts from researchers, policymakers, and the broader community to navigate the future of science.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. S. Musslick and S. Mahesh were supported by Schmidt Science Fellows, in partnership with the Rhodes Trust. S. Musslick was also supported by the Carney BRAINSTORM program at Brown University and the NSF (2318549). S. Mahesh also acknowledges the support of the Acceleration Consortium fellowship. S.J. Sloman acknowledges support from the UK Research and Innovation (UKRI) Turing AI World-Leading Researcher Fellowship [EP/W002973/1]. S.H.C. was supported by the Finnish Center for Artificial Intelligence, and Academy of Finland (328813); he also acknowledges the support from the Jorma Ollila Mobility Grant by Nokia Foundation. L.K.B. and F.G. were supported by European Research Council Grant ERC-ADG-835002-GEMS. T.L.G. was supported by a grant from the NOMIS Foundation. R.D.K. was supported by the Wallenberg AI, Autonomous Systems and Software Program funded by the Knut and Alice Wallenberg Foundation, by Chalmers Artificial Intelligence Research Centre, and by the UK Engineering and Physical Sciences Research Council (EPSRC) Grants EP/R022925/2 and EP/W004801/1. W.R.H. was supported by the NSF (SES-2242962). J.N.K. acknowledges support from the National Science Foundation AI Institute in Dynamic Systems (2112085). We thank Solomon Oyakhire for valuable feedback. F.R. acknowledges support by National Institutes of Health, National Institute of Biomedical Imaging and Bioengineering (R01EB029272, R01EB030896NSF and R01EB030896), National Science Foundation Behavior and Cognitive Science (1734853, 1636893), Advanced Cyberinfrastructure (1916518), and Information and Intelligent Systems (1912270).

1. R. D. King *et al*., Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252 (2004).
2. R. D. King *et al*., The automation of science. *Science* **324**, 85–89 (2009).
3. G. Raayoni *et al*., Generating conjectures on fundamental constants with the Ramanujan Machine. *Nature* **590**, 67–73 (2021).
4. A. Davies *et al*., Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
5. A. F. de Almeida, R. Moreira, T. Rodrigues, Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **3**, 589–604 (2019).
6. J. Jumper *et al*., Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
7. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project* (McGraw-Hill, New York, 1980).
8. D. Boiko *et al*., Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
9. A. Merchant *et al*., Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
10. N. J. Szymanski *et al*., An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
11. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
12. U.S. Department of Energy, *Scientific Machine Learning for Complex Systems* (Funding Announcement, 2023).
13. A. Velasquez, *Foundation Models for Scientific Discovery (FoundSci)* (Defense Advanced Research Projects Agency DARPA Program Solicitation, 2023).
14. H. Kitano, Nobel Turing Challenge: Creating the engine for scientific discovery. *npj Syst. Biol. Appl.* **7**, 29 (2021).
15. H. Zenil *et al*., The future of fundamental science led by generative closed-loop artificial intelligence. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.07522.
16. A. Birhane, A. Kasirzadeh, D. Leslie, S. Wachter, Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
17. H. Wang *et al*., Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
18. M. Binz *et al*., How should the advancement of large language models affect the practice of science? *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401227121 (2025).
19. M. Dubova, M. Galesic, R. L. Goldstone, Cognitive science of augmented intelligence. *Cogn. Sci.* **46**, e13229 (2022).
20. H. Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press, 1988).
21. M. Dubova *et al*., Is Ockham's razor losing its edge? New perspectives on the principle of model parsimony. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401230121 (2025).
22. P. Langley, *Scientific Discovery: Computational Explorations of the Creative Processes* (MIT Press, 1987).
23. S. Džeroski, P. Langley, L. Todorovski, "Computational discovery of scientific knowledge" in *Computational Discovery of Scientific Knowledge: Introduction, Techniques and Applications in Environmental and Life Sciences*, S. Džeroski, L. Todorovski, Eds. (Springer, 2007), pp. 1–14.
24. S. M. Udrescu *et al*., AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Adv. Neural Inf. Process. Syst.* **33**, 4860–4871 (2020).
25. M. Cranmer, Interpretable machine learning for science with PySR and SymbolicRegression.jl. arXiv [Preprint] (2023). https://arxiv.org/abs/2305.01582.
26. M. Landajuela *et al*., *Discovering Symbolic Policies with Deep Reinforcement Learning* (PMLR, 2021), pp. 5979–5989.
27. S. Li, I. Marinescu, S. Musslick, *Symbolic Regression with Generative Flow Networks* (GFN-SR, 2023).
28. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artif. Intell.* **61**, 209–261 (1993).
29. J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
30. A. Jain *et al*., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
31. A. M. Liekens *et al*., BioGraph: Unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol.* **12**, 1–12 (2011).
32. J. B. Voytek, B. Voytek, Automated cognome construction and semi-automated hypothesis generation. *J. Neurosci. Methods* **208**, 92–100 (2012).
33. S. Musslick *et al*., A standard language for factorial experimental design. *Behav. Res. Methods* **54**, 805–829 (2020).
34. M. van Casteren, M. H. Davis, Mix, a program for pseudorandomization. *Behav. Res. Methods* **38**, 584–589 (2006).
35. D. J. Navarro, M. A. Pitt, I. J. Myung, Assessing the distinguishability of models and the informativeness of data. *Cogn. Psychol.* **49**, 47–84 (2004).
36. J. I. Myung, M. A. Pitt, Optimal experimental design for model discrimination. *Psychol. Rev.* **116**, 499 (2009).
37. D. R. Cavagnaro, J. I. Myung, M. A. Pitt, J. V. Kujala, Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Comput.* **22**, 887–905 (2010).
38. S. Musslick *et al*., "An evaluation of experimental sampling strategies for autonomous empirical research in cognitive science" in *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, M. Goldwater, F. K. Anggoro, B. K. Hayes, D. C. Ong, Eds. (Cognitive Science Society, 2023), pp. 1386–1392.
39. K. Williams *et al*., Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J. R. Soc. Interface* **12**, 20141289 (2015).
40. A. Coutant *et al*., Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 18142–18147 (2019).
41. A. B. Watson, QUEST+: A general multidimensional Bayesian adaptive psychometric method *J. Vis.* **17**, 10 (2017).
42. S. Valentin *et al*., Designing optimal behavioral experiments using machine learning. *eLife* **13**, e86224 (2024).
43. B. Shababo, B. Paige, A. Pakman, L. Paninski, Bayesian inference and online experimental design for mapping neural microcircuits. *Adv. Neural Inf. Process. Syst.* **26**, 1304–1312 (2013).
44. S. Dushenko, K. Ambal, R. D. McMichael, Sequential Bayesian experiment design for optically detected magnetic resonance of nitrogen-vacancy centers. *Phys. Rev. Appl.* **14**, 054036 (2020).
45. X. Huan, Y. M. Marzouk, Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys.* **232**, 288–317 (2013).
46. G. N. Kanda *et al*., Robotic search for optimal cell culture in regenerative medicine. *eLife* **11**, e77007 (2022).
47. S. Stanton *et al*., *Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders* (PMLR, 2022), pp. 20459–20478.
48. K. Korovina *et al*., *Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations* (PMLR, 2020), pp. 3393–3403.
49. R. R. Griffiths, J. M. Hernández-Lobato, Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.* **11**, 577–586 (2020).
50. Y. Zhang, D. W. Apley, W. Chen, Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Sci. Rep.* **10**, 4924 (2020).
51. A. G. Kusne *et al*., On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Commun.* **11**, 5966 (2020).
52. Q. Liang *et al*., Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Comput. Mater.* **7**, 188 (2021).
53. C. Papadimitriou, Optimal sensor placement methodology for parametric identification of structural systems. *J. Sound Vib.* **278**, 923–947 (2004).
54. J. Chang, J. Kim, B. T. Zhang, M. A. Pitt, J. I. Myung, Data-driven experimental design and model development using Gaussian process with active learning. *Cogn. Psychol.* **125**, 101360 (2021).
55. P. Grünwald, T. van Ommen, Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12**, 1069–1103 (2017).
56. T. Rainforth, A. Foster, D. R. Ivanova, F. B. Smith, Modern bayesian experimental design. *Stat. Sci.* **39**, 100–114 (2024).
57. S. J. Sloman, "Towards robust bayesian adaptive design methods for the study of human behavior," PhD thesis, Carnegie Mellon University (2022).
58. M. Dubova, A. Moskvichev, K. Zollman, Against theory-motivated experimentation in science. MetaArXiv [Preprint] (2022). https://doi.org/10.31222/osf.io/ysv2u.
59. R. Gelpi, N. Saxena, G. Lifchits, D. Buchsbaum, C. G. Lucas, "Sampling heuristics for active function learning" in *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (2021), https://cognitivesciencesociety.org.
60. M. Dubova, S. J. Sloman, B. Andrew, M. R. Nassar, S. Musslick, Explore your experimental designs and theories before you exploit them! *Behav. Brain Sci.* **47**, e40 (2024).
61. G. Shin *et al*., Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. *J. Biomed. Inf.* **93**, 103153 (2019).
62. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, Effectiveness of continuous glucose monitoring in a clinical care environment: Evidence from the Juvenile Diabetes Research Foundation continuous glucose monitoring (JDRF-CGM) trial. *Diabetes Care* **33**, 17–22 (2010).
63. M. A. Lazzara, G. A. Weidner, L. M. Keller, J. E. Thom, J. J. Cassano, Antarctic automatic weather station program: 30 years of polar observation. *Bull. Am. Meteorol. Soc.* **93**, 1519–1537 (2012).

64. S. M. Wood, S. P. Johnson, J. N. Wood, Automated study challenges the existence of a foundational statistical-learning ability in newborn chicks. *Psychol. Sci.* **30**, 1592–1602 (2019).
65. T. M. Gureckis *et al*., psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* **48**, 829–842 (2016).
66. W. Mason, S. Suri, Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* **44**, 1–23 (2012).
67. S. Palan, C. Schitter, Prolific.ac–A subject pool for online experiments. *J. Behav. Exp. Fin.* **17**, 22–27 (2018).
68. B. Thompson, B. Van Opheusden, T. Sumers, T. Griffiths, Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* **376**, 95–98 (2022).
69. B. Carpenter *et al*., Stan: A probabilistic programming language. *J. Stat. Software* **76** (2017).
70. A. Gelman *et al*., Bayesian workflow. arXiv [Preprint] (2020). https://arxiv.org/abs/2011.01808.
71. C. Steinruecken, E. Smith, D. Janz, J. Lloyd, Z. Ghahramani, *The Automatic Statistician. Automated Machine Learning: Methods, Systems, and Challenges* (Springer, 2019), pp. 161–173.
72. F. Gobet, M. Addis, P. C. Lane, P. D. Sozou, "Introduction: Scientific discovery in the social sciences" in *Scientific Discovery in the Social Sciences*, M. Addis, P. C. R. Lane, P. D. Sozou, F. Gobet, Eds. (Springer, 2019), pp. 1–7.
73. B. C. Falkenhainer, R. S. Michalski, Integrating quantitative and qualitative discovery: The ABACUS system. *Mach. Learn.* **1**, 367–401 (1986).
74. D. B. Lenat, The ubiquity of discovery. *Artif. Intell.* **9**, 257–285 (1977).
75. P. Langley, Data-driven discovery of physical laws. *Cogn. Sci.* **5**, 31–54 (1981).
76. L. Bartlett, A. Pirrone, N. Javed, P. Lane, F. Gobet, "Genetic programming for developing simple cognitive models" in *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, M. Goldwater, F. K. Anggoro, B. K. Hayes, D. C. Ong, Eds. (Cognitive Science Society, 2023), pp. 2833–2839.
77. E. Frias-Martinez, F. Gobet, Automatic generation of cognitive theories using genetic programming. *Minds Mach.* **17**, 287–309 (2007).
78. R. Guimerà *et al*., A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci. Adv.* **6**, eaav6971 (2020).
79. C. Cornelio *et al*., Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat. Commun.* **14**, 1777 (2023).
80. M. Landajuela *et al*., A unified framework for deep symbolic regression. *Adv. Neural Inf. Process. Syst.* **35**, 33985–33998 (2022).
81. S. Musslick, "Recovering quantitative models of human information processing with differentiable architecture search" in *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, T. Fitch, C. Lamm, H. Leder, K. Teßmar-Raible, Eds. (Cognitive Science Society, 2021), pp. 348–354.
82. A. Murari *et al*., Data driven theory for knowledge discovery in the exact sciences with applications to thermonuclear fusion. *Sci. Rep.* **10**, 19858 (2020).
83. E. Bilsland *et al*., Yeast-based automated high-throughput screens to identify anti-parasitic lead compounds. *Open Biol.* **3**, 120158 (2013).
84. E. Bilsland *et al*., Plasmodium dihydrofolate reductase is a second enzyme target for the antimalarial action of triclosan. *Sci. Rep.* **8**, 1038 (2018).
85. M. B. Lee, B. Blue, M. Muir, M. Kaeberlein, The million-molecule challenge: A moonshot project to rapidly advance longevity intervention discovery. *GeroScience* **6**, 3103–3113 (2023).
86. J. N. Pitt *et al*., WormBot, an open-source robotics platform for survival and behavior analysis in *C. elegans*. *GeroScience* **41**, 961–973 (2019).
87. M. J. Tamasi *et al*., Machine learning on a robotic platform for the design of polymer-protein hybrids. *Adv. Mater.* **34**, 2201809 (2022).
88. B. P. MacLeod *et al*., Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
89. E. A. Pogue *et al*., Closed-loop superconducting materials discovery. *npj Comput. Mater.* **9**, 181 (2023).
90. S. Musslick *et al*., AutoRA: Automated research assistant for closed-loop empirical research. *J. Open Source Softw.* **9**, 6839 (2024).
91. C. Lu *et al*., The AI scientist: Towards fully automated open-ended scientific discovery. arXiv [Preprint] (2024). https://arxiv.org/abs/2408.06292.
92. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).
93. B. K. Petersen *et al*., Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. arXiv [Preprint] (2021). https://arxiv.org/abs/1912.04871. Accessed 5 November 2024.
94. A. Almaatouq *et al*., Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.*, 1–55 (2022).
95. M. K. Eckstein, C. Summerfield, N. D. Daw, K. J. Miller, "Predictive and interpretable: Combining artificial neural networks and classic cognitive models to understand human learning and decision making" in *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*, M. Goldwater, F. K. Anggoro, B. K. Hayes, D. C. Ong, Eds. (Cognitive Science Society, 2023), pp. 928–935.
96. L. Ji-An, M. K. Benna, M. G. Mattar, Automatic discovery of cognitive strategies with tiny recurrent neural networks. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.04.12.536629.
97. P. Subash *et al*., A comparison of neuroelectrophysiology databases. *Sci. Data* **10**, 719 (2023).
98. K. J. Gorgolewski *et al*., The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3**, 1–9 (2016).
99. J. Hewson, Y. Strittmatter, I. Marinescu, C. Williams, S. Musslick, Bayesian machine scientist for model discovery in psychology. NeurIPS 2023 AI for Science Workshop. https://openreview.net/forum?id=XHFfvzlQ1n. Accessed 5 November 2024.
100. R. Cory-Wright, C. Cornelio, S. Dash, B. El Khadir, L. Horesh, Evolving scientific discovery by unifying data and background knowledge with AI Hilbert. *Nat. Commun.* **15**, 5922 (2024).
101. A. C. Dama *et al*., BacterAI maps microbial metabolism without prior knowledge. *Nat. Microbiol.* **8**, 1018–1025 (2023).
102. S. J. Sloman, D. M. Oppenheimer, S. B. Broomell, C. R. Shalizi, Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias. arXiv [Preprint] (2022). https://arxiv.org/abs/2205.13698.
103. S. J. Sloman, D. R. Cavagnaro, S. B. Broomell, Knowing what to know: Implications of the choice of prior distribution on the behavior of adaptive design optimization. *Behav. Res. Methods* **56**, 1–24 (2024).
104. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).
105. J. N. Kutz, "Machine learning methods for reduced order modeling" in *Model Order Reduction and Applications: Cetraro, Italy 2021*, M. Falcone, G. Rozza, Eds. (Springer, 2023), pp. 201–228.
106. P. Conti, G. Gobat, S. Fresca, A. Manzoni, A. Frangi, Reduced order modeling of parametrized systems through autoencoders and SINDy approach: Continuation of periodic solutions. *Comput. Methods Appl. Mech. Eng.* **411**, 116072 (2023).
107. A. Mendible, S. L. Brunton, A. Y. Aravkin, W. Lowrie, J. N. Kutz, Dimensionality reduction and reduced-order modeling for traveling wave physics. *Theor. Comput. Fluid Dyn.* **34**, 385–400 (2020).
108. B. Chen *et al*., Automated discovery of fundamental variables hidden in experimental data. *Nat. Comput. Sci.* **2**, 433–442 (2022).
109. S. Schneider, J. H. Lee, M. W. Mathis, Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).
110. Y. Zheng *et al*., Large language models for scientific synthesis. Inference and explanation. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.07984.
111. N. Guler, S. Kirshner, R. Vidgen, Artificial intelligence research in business and management: A literature review leveraging machine learning and large language models. SSRN. https://ssrn.com/abstract=4540834. Accessed 5 November 2024.
112. S. Whitfield, M. A. Hofmann, Elicit: AI literature review research assistant. *Public Serv. Q.* **19**, 201–207 (2023).
113. X. Luo *et al.*, Large language models surpass human experts in predicting neuroscience results. *Nat. Hum. Behav.*, in press.
114. M. Chen *et al*., Evaluating large language models trained on code. arXiv [Preprint] (2021). https://arxiv.org/abs/2107.03374.
115. Y. Su *et al*., *How to Make Software Documentation More Useful with a Large Language Model?* (Association for Computing Machinery (ACM), Washington, 2023), pp. 87–93.
116. J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world? *Behav. brain sciences* **33**, 61–83 (2010).
117. P. Gleeson *et al*., Integrating model development across computational neuroscience, cognitive science, and machine learning. *Neuron* **111**, 1526–1530 (2023).
118. P. Rich, R. de Haan, T. Wareham, I. van Rooij, "How hard is cognitive science?" in *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, T. Fitch, C. Lamm, H. Leder, K. Teßmar-Raible, Eds. (Cognitive Science Society, 2021), pp. 3034–3040.
119. B. Burger *et al*., A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
120. L. Messeri, M. Crockett, Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
121. J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, D. Blei, Reading tea leaves: How humans interpret topic models. *Adv. Neural Inf. Process. Syst.* **22**, 288–296 (2009).
122. L. Bainbridge, "Ironies of automation" in *Analysis, Design and Evaluation of Man–Machine Systems*, G. Johannsen, J. E. Rijnsdorp, Eds. (Elsevier, 1983), pp. 129–135.
123. National Institutes of Health, The use of generative artificial intelligence technologies is prohibited for the NIH peer review process. NIH Grants Guide. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html. Accessed 5 November 2024.
124. M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15849–15854 (2019).
125. J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv [Preprint] (2018). https://arxiv.org/abs/1803.03635. Accessed 5 November 2024.
126. Z. Li *et al*., "Train big, then compress: Rethinking model size for efficient training and inference of transformers" in *International Conference on Machine Learning* (PMLR, 2020), pp. 5958–5968.
127. M. Agrawal, J. C. Peterson, T. L. Griffiths, Scaling up psychology via scientific regret minimization. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8825–8835 (2020).
128. T. L. Griffiths, Manifesto for a new (computational) cognitive revolution. *Cognition* **135**, 21–23 (2015).
129. L. Daston, P. Galison, *Objectivity* (Princeton University Press, 2021).
130. J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (PMLR, 2018), pp. 77–91.
131. L. Yarger, F. Cobb Payton, B. Neupane, Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Inf. Rev.* **44**, 383–395 (2020).
132. T. Hagendorff, The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* **30**, 99–120 (2020).