Artificial intelligence in science: Overview and policy proposals

A. Nolan, Organisation for Economic Co-operation and Development

Introduction

This book addresses the current and emerging roles of artificial intelligence (AI) in science. Accelerating the productivity of research could be the most economically and socially valuable of all AI's uses. AI and its various subdisciplines are pervading every field and stage of the scientific process. Advances in AI have led to an outpouring of creative uses in research. However, AI's potential contribution to science is far from realised, and the impact of some widely hailed achievements may be less than is generally thought. AI, for instance, contributed little to research and treatment during the COVID-19 pandemic. Moreover, policy makers and other actors in research systems can do much to speed and broaden the uptake of AI in science, and to magnify its positive contributions to science and society.

The book's main contributions are to:

- Describe, in terms amenable to non-technical readers, AI's current and possible future uses in science.
- Help raise awareness of the roles that public policy could play in amplifying AI's positive impact on science, while also managing governance challenges.
- Draw attention to applications of AI in science and related topics that may be unfamiliar to some lay readers. Such applications include, among others, AI and collective intelligence, AI and laboratory robotics, AI and citizen science, developments in scientific fact-checking, and the emerging uses of AI in research governance. Related topics include the thematic narrowing of AI research and the reproducibility of AI research.
- Assess what AI cannot yet do in science, and areas of progress still required.
- Examine empirical claims of a slowdown in the productivity of science, engaging the views of domain experts and economists.
- Consider the implications of AI in science for developing countries, and the measures that could be taken to expedite uptake in developing-country research.

This chapter proceeds as follows: the opening sections discuss why raising research productivity is important, whether through using AI or other means. The key issues concern economic effects, addressing critical knowledge gaps, summarising the evidence for and countering possible sources of drag on research productivity. In so doing, the text outlines why some scholars have argued that the productivity of science may be stagnating. To be clear, the claim is not that progress in science is slowing, but that it is becoming harder to achieve. The chapter continues with summaries of the book's 34 essays. The summaries are presented under five broad headings. These correspond to the five parts of the book:

- Is science getting harder?

- Artificial intelligence in science today
- The near future: Challenges and ways forward
- Artificial intelligence in science: Implications for public policy
- Artificial intelligence, science and developing countries.

The salient policy implications and suggestions are highlighted in text boxes.

AI and the productivity of science: Why does this matter?

The productivity of science is of critical interest for many reasons. Three are described here: economic; the need to close gaps in significant areas of scientific knowledge; and claims of slowing research productivity.

Economic implications of research productivity

Economists have established a fundamental relationship between innovation, which draws from basic research, and long-term productivity growth. The economic effects of COVID-19, sluggish macro-economic conditions in most OECD countries, burgeoning public debt and population ageing have all added urgency to the quest for growth.

The sheer scope of science's role in modern economies is easily underestimated. By one assessment, industries reliant just on physics research, including electrical, civil and mechanical engineering, as well as computing and other industries, contribute more to Europe's economic output and gross value added than retail and construction combined (European Physical Society, 2019). The scope of any feedthrough from changes in research productivity will be correspondingly broad. Recent analysis by the International Monetary Fund (IMF) based on patents data suggests that basic scientific research diffuses to more sectors in more countries and for a longer time than commercially oriented applied research (IMF, 2021).

Theory also suggests that growth stemming from more productive R&D will be more lasting than that spurred by automation in final goods production, which can yield a one-time increase in the rate of growth (Trammell and Korinek, 2020).

Much basic and essential scientific knowledge is lacking

In many domains, science is advancing rapidly. In 2022, there was widely publicised progress in fields as diverse as astronomy, with unprecedented images from the James Web telescope, the development of a nasal vaccine for COVID-19 and the first laboratory-based controlled fusion reaction. However, it is also the case that both old scientific questions endure and new ones arise continually. To take just three examples:

- After decades of climate modelling, uncertainty persists. Important uncertainties exist on such issues as tipping points (e.g. inversion of the flows of cold and hot oceanic waters), when changes could become irreversible (e.g. melting of West Antarctic or Greenland ice-shelves), and the quantitative role of plants and microbes in the carbon cycle (plants and microbes cycle some 200 billion tons of carbon a year, compared to anthropogenic production of around 6 billion tons).
- Many elementary cellular processes are not understood. For instance, the process by which *Escherichia coli* (a bacterium) consumes sugar for energy is one of the most basic biological functions. It is also important for industry in designing microbial biocatalysts that use carbohydrates in biomass. However, how the process operates has not been fully established (even though research on the subject was first published over 70 years ago).

- Around 55 million people worldwide currently suffer from Alzheimer's disease or other dementias. While studies have identified several risk factors for Alzheimer's disease – from age, to head injury, to high cholesterol – the cause of the disease is still unknown (and treatments are missing).

More productive science will also set foundations for breakthroughs in innovation, especially in some crucial fields. For instance, many of the antibiotics in use today were discovered in the 1950s, and the most recent class of antibiotic treatments was discovered in 1987. Innovation in the energy sector is also essential for achieving low-emission economic growth. But today's leading energy generation technologies were mostly invented over a century ago. The combustion turbine was invented in 1791, the fuel cell in 1842, the hydro-electric turbine in 1878 and the solar photo-voltaic cell in 1883. Even the first nuclear power plant began operating over 60 years ago (Webber et al., 2013) (although the performance of these technologies has of course improved over time).

By accelerating science and innovation, AI could help to find solutions to global challenges such as climate change (Boxes 1 and 2), and the diseases of ageing.

Box 1. Artificial intelligence, materials science and net zero

Materials science is central to new technologies needed to address climate change. Among many possibilities, new materials promise more efficient solar panels, better batteries, lightweight metal alloys for more fuel-efficient vehicles, carbon-neutral fuels, more sustainable building materials and low-carbon textiles. Progress in materials science may also create substitutes for materials with fragile supply chains, including rare earth elements.

Assisted by an open-source research community and open-access databases, AI is ushering in a revolution in materials science, quickly and efficiently exploring large datasets for arrangements of atoms that yield materials with user-desired properties, while optimising aspects of experimentation.

Materials discovery has traditionally been slow and uncertain, based on trial-and-error examination of many – sometimes millions – of candidate samples. The research sometimes takes decades. However, the new combinations of high-performance computing, AI and laboratory robots can greatly accelerate discovery (later essays in this book explore robotics in science). Service (2019) describes some materials discovery processes being compressed from months to just a few days. One lab robot conducts 100 000 experiments a year, producing five years of experiments in just two weeks (Grizou et al., 2020).

The urgency of achieving net zero underscores the importance of accelerating materials discovery. Faster discovery can also encourage the private sector to invest in materials R&D, as returns are more likely to be had within commercial timeframes. Lowering costs per experiment can encourage more creative research, as the risk of failure is mitigated if a broad and fast-running portfolio of experiments is possible. In addition, faster discovery might help junior researchers to establish themselves (Correa-Baena et al., 2018).

These advances in materials science require contributions from many disciplines, including computer scientists, roboticists, electronics engineers, physical scientists and materials researchers. Policies and approaches that facilitate cross-disciplinary research and exchange of ideas could help.

Box 2. Catalysing research at the intersection of climate change and machine learning

Climate Change AI (CCAI)[1] is a not-for-profit organisation bringing together volunteers from academia and industry. One of its most significant offerings is a catalogue[2] of numerous research questions across many areas in science, engineering, industry and social policy where AI could make a dent in climate problems. CCAI also cultivates a community of many researchers, engineers, policy makers, investors, companies and non-governmental organisations, many of which are applying AI techniques to scientific problems.

1. See https://www.climatechange.ai/.
2. See https://www.climatechange.ai/summaries.

AI also matters because science itself may be becoming harder

Claims of a slowdown in science are not new. More than 50 years ago, Bentley Glass, former President of the American Academy for the Advancement of Science, asserted that "There are still innumerable details to fill in, but the endless horizons no longer exist" (Glass, 1971). Recently, attention to a purported stagnation in research productivity has been spurred by Bloom et al. (2020) and other papers. Matt Clancy, in this book, reviews the relevant economic and technology-specific studies, and concludes that while quantification of research productivity is conceptually and methodologically complex, and not uncontentious, science has by some measures become harder.

If science were indeed to become harder then, other conditions unchanged, governments would be forced to spend more to achieve existing rates of growth of useful scientific output. Timeframes could be lengthened for achieving scientific progress needed to address today's global challenges. And for investments in science equivalent to today's, ever-fewer increments of new knowledge will be available with which to counter unforeseen events with negative global ramifications, from new contagions to novel crop diseases.

It is helpful to consider the arguments made by the scholars who contend that science is getting harder. These are summarised in Box 3. Examining the explanations why this might be can help to pinpoint how AI could help. Essays in this book examine various issues relevant to the effects of bad incentives in science systems, argument (1) in Box 3. Those essays explore such issues as AI in scientific fact-checking, and AI in governance processes (see the contributions of Varoquaux and Cheplygina; Flanagan, Ribeiro and Ferri; and Gundersen Wang). In connection with argument (2) in Box 3 – a more limited involvement of the private sector in basic research – AI can incentivise some areas of private research and development. This is because AI can help conduct some parts of science more rapidly, better aligning with commercial investment horizons. AI has also spurred the creation of firms specialised in doing basic science for larger corporates (see essays by Szalay; Ghosh; and by King, Peter and Courtney).

AI in science is also relevant to argument (3) – the economic limits on discovery – as it can lower costs in some stages of science, especially laboratory experimentation. In addition, potentially large savings of scientists' time could come from compressing the duration of research projects – for instance by using increasingly capable AI-driven research assistants (the subject of the essay by Byun and Stuhlmüller). Argument (4) in Box 3 relates to the need for larger teams in science. The essay on AI and collective intelligence by Malliaraki and Berditchevskaia considers how to harness the capabilities of such teams, as does the essay on AI and citizen science by Ceccaroni and his colleagues. Furthermore, arguments relating to the burden of knowledge – arguments (5) and (6) – are explored from different viewpoints in essays on natural language processing applied to scientific texts (see the contributions of Dunietz; Wang; Byun and Stuhlmüller; and Smalheiser, Hahn-Powell, Hristovski and Sebastian).

Box 3. Why might science get harder?

Researchers have posited reasons for an alleged decline in the productivity research. While not exhaustive, the main arguments concern the following:

1. *Changes in scientific incentives.* Among others, Bhattacharya and Packalen (2020) explore the role of citations in performance measurement and in shifting scientists' rewards and behaviour toward incremental science, with high rates of retraction, non-replicability and even fraud.
2. *A more limited engagement of the private sector in basic science* (Arora et al., 2019).
3. *Economic limits on discovery.* For example, the cost of the next generation LHC supercollider is estimated at EUR 21 billion. To generate energies needed to probe smaller subatomic phenomena would be orders of magnitude more costly.
4. *As more prior and diverse science must be absorbed to make new breakthroughs, larger teams are needed.* But larger teams seem less prone to make fundamental discoveries than small teams (Wu, Wang and Evans, 2019).
5. *Scientists have reached "peak reading".* By one account, 100 000 articles on COVID-19 were published in the first year of the pandemic. Tens of millions of peer-reviewed papers exist in biomedicine alone. However, the average scientist reads about 250 papers a year (Noorden, 2014).
6. *The sheer size of the corpus of scientific literature in different fields.* In larger corpora, potentially important contributions cannot garner field-wide attention through gradual processes of diffusion (Chu and Evans, 2021).
7. *As science progresses, it branches into new disciplines.* Some breakthroughs require more inter-disciplinarity, but there is friction at the boundaries between disciplines.
8. *There are a finite number of scientific laws*. Once a law or artefact is discovered, science has to proceed to the next challenge. DNA, for example, can only be discovered once.

Is science getting harder?

Are ideas getting harder to find? A short review of the evidence

Reviewing multiple studies, Matt Clancy concludes that, using diverse methodological and conceptual approaches, a constant supply of research effort (such as numbers of scientists) does not lead to a constant proportional increase in various proxies for technological capabilities (e.g. doubling the number of transistors on an integrated roughly every two years). There are few exceptions to the general finding that a constant proportional increase in metrics of interest has tended to require an increasing supply of research effort.

Clancy also points to other measurement approaches based on the idea that progress is not just about squeezing the last drop of possibility from each technology, it is also, and perhaps mostly, about the creation of entirely new branches of technology. However, acknowledging this perspective, Bloom et al. (2020) showed that, at least in health, despite successive waves of new technologies, from antibiotics to mRNA vaccines, etc., saving a year of life has needed increasing research effort measured by the number of clinical trials or biomedical articles.

Another measure of the effects of R&D relates to performance outcomes in private sector companies. Bloom et al. (2020) examine sales, number of employees, sales per employee and market capitalisation

and find here, too, that on average it takes more and more R&D effort by firms to maintain growth in these measures.

Clancy likewise discusses total factor productivity (TFP) – the efficiency with which an economy combines inputs to create outputs – as a broad measure of technological progress. Bloom et al. (2020) found that for the US economy, going back to the 1930s, growing R&D effort has been required to keep TFP increasing at a constant exponential rate. Miyagawa, in this book, arrives at a similar result for Japan, as do Boeing and Hünermund for Germany and the People's Republic of China (hereafter "China").

Another way to examine research productivity is to look at measures from science. Clancy discusses one approach which looked at the share of Nobel Prize winning awards that go to discoveries described in papers published in the preceding 20 years. Across all fields, this has fallen significantly. Clancy also describes studies that show a steady decline since the 1960s in the share of citations to more recent papers (those published in the preceding five or ten years), possibly suggesting a declining impact of recent scientific output. Patents share this pattern, and increasingly cite older scientific work.

Clancy also explains why conceptual and methodological caveats apply to all the analyses. TFP, for instance, can vary for reasons unrelated to science and technology, such as changes in the geographic mobility of workers. However, many papers employing diverse approaches arrive at converging conclusions. Nevertheless, Clancy closes by acknowledging that even if ideas are getting harder to find, society also seems to be trying harder to find them, causing science to advance.

Other essays in this volume – summarised below – examine three fields of technology where Bloom et al. (2020) compared performance metrics with measures of research input and thereby argued for a decline in research productivity: namely Moore's Law, agriculture and the biopharmaceuticals sector. However, the picture that emerges in the essays below is not quite as clear-cut as Bloom et al. (2020) suggest.

The end of Moore's Law?

Moore's Law, which has held since the 1960s, posits that transistor chip density doubles roughly every two years, with a corresponding decline in unit transistor cost. Bloom et al. (2020) suggest that an apparent slowing of Moore's Law indicates a decline in the pace of innovation in electronics. Such a decline would have serious consequences, as microelectronics are central to practically all industrial products and systems.

However, Henry Kressel shows that while the ability to shrink transistors is reaching physical limits, fears of stagnation or decline in the power of computing systems are premature. He shows that other innovations – additional to those tracked by Moore's Law – continue to improve the economic and technical performance of electronic systems. For instance, manufacturers are findings ways to improve energy efficiency, and developing three-dimensional architectures that make better use of the chip area. Good ideas are not running out. Nor is there evidence of declining interest in such research.

At base, Kressel's essay contains an important generalisable message: measuring the progress of a technology-driven field with a single metric can mislead. Indeed, at present, while non-specialists focus on Moore's Law, no reliable general metric of progress is available today because computing systems range so greatly in scale and functionality.

Is technological progress in US agriculture slowing?

Matt Clancy examines innovation in US agriculture and concludes that the case for a slowdown seems to hold whether measured with growth in yields over time or using more sophisticated methods, such as changes in TFP. The slowdown may stem from agriculture-specific factors, such as stagnating levels of R&D through much of the late 20th century. It may also be influenced by broader forces, such as slowing technological progress in non-farm domains that supply critical inputs to agriculture. Moreover, while this

essay examines US agriculture, Clancy cites research suggesting that global productivity growth in agriculture fell from an average of 2% per year over the 2000s to 1.3% per year over the 2010s.

Echoing Kressel's point on the need for care in selecting metrics of progress, Clancy observes that changes in agricultural yield – a focus of Bloom et al. – has drawbacks. For example, almost all of US corn is genetically modified to confer resistance to a key pesticide (glyphosate). This helps farmers by making it less costly to control weeds, a benefit not captured in measures of yield. Similarly, an important dimension of agricultural innovation not typically included in TFP is the environmental sustainability of agricultural production, which may be improving.

Eroom's law and the decline in the productivity of biopharmaceutical R&D

Jack Scannell explores Eroom's law, the observation that drug development becomes slower and more expensive over time. Scannell examines various metrics that show a significant decline in the productivity of biopharmaceutical R&D since the late 1990s (although with a slight uptick since 2010). He points out that DNA sequencing, genomics, high-throughput screening, computer-aided drug design and computational chemistry, among other advances, were widely adopted and/or became orders of magnitude cheaper between 1950 and 2010. However, over the same period, the number of new drugs approved by the US Food and Drug Administration (FDA) per billion US dollars of inflation-adjusted R&D fell roughly a hundredfold.

Scannell suggests that levels of innovation in biopharma have fallen for several reasons. Arguably of greatest importance is the progressive accumulation of an inexpensive pharmacopoeia of effective generic drugs. When drugs' patents expire, they become much cheaper but no less effective. An ever-expanding catalogue of cheap generic drugs progressively raises the competitive bar for new drugs in the same therapy area, eroding incentives for R&D. Such therapy areas hold meagre returns for investment in "new ideas", even if the ideas themselves have not become harder to find (there are many unexploited drug targets and therapeutic mechanisms and a vast number of chemical compounds).

Scannell explains that R&D investment has been squeezed towards diseases where R&D has for long been less successful, such as advanced Alzheimer's, some metastatic solid cancers, etc. He observes that novel chemistry – where AI can play a big role - is the most investible form of biopharmaceutical innovation because it can be protected by strong patents. However, the lack of good screening and disease models is a key constraint on drug discovery (a disease model is a biological system in the laboratory that mirrors a disease and its processes). A major reason for this shortage is economic: once the mechanism identified by a new disease model is publicly proven in trials in human patients, the information becomes freely available to competitors.

AI will be incrementally helpful but not revolutionary in drug discovery

Scannell considers that AI will help in drug R&D. However, its overall impact on industry-level productivity will likely be modest in the near term. This is because the areas with the most progress in using AI – such as drug chemistry – are rarely relevant to the rate-limiting steps in drug development. Meanwhile, AI is less likely to yield solutions where gains in R&D productivity are most needed. A main reason for this is that much of the critical data is of insufficient quality. For example, too much of the published biomedical literature is false, irrelevant or both. Generating better biological data will help take advantage of AI, but doing so is costly and takes time.

Is there a slowdown in research productivity? Evidence from China and Germany

Philipp Boeing and Paul Hünermund provide evidence for a decrease in research productivity in recent decades for China and Germany, following the methodology developed by Bloom et al. (2020) – where it

was argued that R&D efficiency, measured by economic productivity growth divided by the number of researchers, has declined in the United States.

For Germany, R&D expenditures increased by an average of 3.3% per year during the period 1992-2017. Averaged over firm-level outcome measures, research productivity fell by 5.2% per year. This number is similar to that reported by Bloom et al. (2020) for the United States. These negative compound average growth rates imply that research effort must be doubled every 13 years to support constant rates of economic growth.

The authors find that research productivity in China has declined much faster. The effective number of researchers employed by publicly listed firms in the sample used increased by, on average, 21.9% per year between 2001 and 2019. This significant expansion is not matched by increases in economic growth. The findings entail a drop in research productivity of 23.8% per year. However, if analysis is restricted to the most recent decade (when China began large-scale R&D activities) research productivity fell by only 7.3% a year, a number closer to those found for Germany and the United States.

Declining R&D efficiency: Evidence from Japan

Tsutomu Miyagawa notes that while Japan has maintained a ratio of R&D to gross domestic product (GDP) of around 3% for some time, R&D efficiency growth appears to have slowed. Adopting the methodology used in Bloom et al. (2020), Miyagawa and Ishikawa (2019) found that the efficiency of R&D in Japanese manufacturing and information services had fallen. Using more recent data, Miyagawa's essay in this volume examines two measures of R&D efficiency. The first is derived from a simple production function in which productivity depends on the stock of R&D. The second again follows the method of Bloom et al. (2020). Both measures show that R&D efficiency in Japan in the 2010s declined compared to the 2000s.

Quantifying the "cognitive extent" of science and how it has changed over time and across countries

Staša Milojević approaches the measurement of research productivity in an entirely different way. She discusses trends in the "cognitive extent" of knowledge in scientific literature. Milojević quantifies the cognitive extent of scientific fields by using information on the number of unique phrases contained in the titles of journal articles. In a given body of literature, a smaller number of unique phrases would indicate a lot of repetition, and a smaller cognitive extent. A larger number of unique phrases suggests a wider range of concepts and a greater cognitive extent.

Milojević finds stagnation in cognitive extent since the mid-2000s. She also examines individual fields of research, showing that cognitive extent in physics, astronomy and biology is expanding, whereas medicine is stagnating or even contracting. In addition. Milojević compares cognitive extent across countries. She finds that while China was the biggest producer of scientific publications in 2019, its papers covered a smaller cognitive extent than many individual West European countries and Japan.

What can bibliometrics contribute to understanding research productivity?

Giovanni Abramo and Ciriaco Andrea D'Angelo discuss the strengths and weaknesses of the most popular bibliometric indicators used to assess research performance. They describe the well-known limits of evaluative bibliometrics: 1) publications may not be representative of all knowledge produced; 2) bibliographic repertories do not cover all publications; and 3) citations are not always a certification of use. However, the authors underscore that bibliometrics is primarily concerned with research outputs. Understanding changes in research productivity also requires measures of the associated research inputs, namely labour and capital.

Abramo and Andrea D'Angelo present a proxy bibliometric indicator of research productivity that includes data on research inputs. They describe the first results of a longitudinal analysis of academic research productivity at a national level using such an indicator. This shows that productivity is increasing over time for Italian academics in most research fields.

The authors call on governments to support more useful national and international research productivity assessments by establishing mechanisms by which bibliometricians are provided with data on labour and capital inputs to research institutions.

Artificial intelligence in science today

How can artificial intelligence help scientists? A (non-exhaustive) overview

Aishik Ghosh observes that AI is being taken up in every domain and stage of science, from hypothesis generation to experiment design, monitoring and simulation, all the way to scientific publication and communication. In the future, AI may optimise many scientific workflows end-to-end – from data collection to final statistical analysis (see the essay on laboratory robots by King, Peter and Courtney). Nonetheless, Ghosh explains that the potential impact of AI on science is a long way from being realised.

The author sets out the main categories of AI's use in science. While typical machine-learning models are difficult to interpret – a point repeated in other essays in the book – they remain useful for tasks such as hypothesis generation, experiment monitoring and precision measurements. Models that create new data – generative AI – can assist with simulations, removing unwanted features from data and converting low-resolution, high-noise images into high-resolution, low-noise images, with many useful applications. In materials science, for example, AI can correctly enhance cheaper, low-resolution electron microscopic images into otherwise more expensive high-resolution images.

Unstructured data (e.g. satellite images, global weather data) have traditionally been a challenge because dedicated algorithms need to be developed to handle them. Deep learning (a class of machine learning, or ML) has been enormously effective in handling such data to solve unusual tasks. Innovations in developing causal models – to disentangle correlation from causation – will provide huge benefits for the medical and social sciences.

AI can also keep track of multiple uncertainties that accumulate through long scientific pipelines. One benefit of this is to make data acquisition more efficient by prioritising data gathering where there is uncertainty. AI is also benefiting science in indirect ways, for instance by advancing mathematics. For example, towards the end of 2022 DeepMind announced it had used a technique known as reinforcement learning to discover how to multiply matrices more rapidly.

Beyond the main stages of research, AI is also more broadly useful to science. For example, some AI models have been developed to summarise research papers and a few popular Twitter bots regularly tweet these automated summaries. Ghosh also points to recent research on an AI-based method to present experimental measurements in physics to theoretical physicists more effectively. Box 4 considers AI in peer review.

Box 4. AI and peer review: Semi-automating time-consuming processes

Peer review consumes enormous scientific resources. By one estimate, just in the United States, and in 2020 only, the time cost of peer review was USD 1.5 billion (Aczel, Szaszi and Holcombe, 2021). Experiments are underway to assess potential uses of AI in multiple aspects of research governance. Checco et al. (2022) describes one such study of AI-assisted peer review. The authors trained an AI

model on 3 300 past conference papers and the associated review evaluations. When shown unreviewed papers the AI model could often predict the peer review outcome. Semi-automated peer review raises ethical and institutional challenges. One possible problem is bias, for instance in propagating cultural and organisational features in the papers on which the AI is trained. However, AI can also reveal biases already operating in human-only peer review. Some uses of AI in peer review would be time saving and relatively uncontroversial, such as in pre-peer review screening to detect early superficial problems in papers. This could be helpful to authors. In addition, removing such problems could lower the impact of first-impression bias and help peer reviewers to focus on papers' scientific content. As Checco et al. explain, more study is needed of AI-enabled decision support. However, as the volume of scientific literature rapidly expands, the practical benefits of emerging AI systems could outweigh their potential disbenefits.

Ghosh also describes possible dangers raised by AI in science. AI models sometimes malfunction in different ways than do traditional algorithms. Using deep learning, a robot trained to work with red, blue and green bottles in a laboratory, for example, may not generalise correctly to black bottles. Deep-learning models pick up subtle patterns in training data, including biases in simulations. And some bias mitigation techniques can lead to further unintended harm. In addition, the trend has been to develop large AI models that require enormous computing resources to train. As other authors in this book also note, this can create problems for research groups with smaller budgets.

In November 2022, following Ghosh's essay, OpenAI released ChatGPT. Many professions are now debating how ChatGPT and other large language models (LLMs) will affect their futures. Uses to increase the productivity of knowledge work are many: quickly and automatically writing diverse materials, from presentations to essays; improving the quality of written language; reducing language barriers for non-native speakers; rapid summarisation; writing computer code; and fostering creativity through dialogue. Evidently, such benefits are also available to science.

However, as Byun and Stuhlmüller discuss later in this book, LLMs like ChatGPT and Galactica often gets things wrong. These authors emphasise the need for processes of evaluation to ensure accuracy as applications are scaled up. They also observe that LLMs risk making superficial work more abundant, as well as creating inequalities, for instance between English-speaking and other users. In a commentary in *Nature*, van Dis et al. (2023) draw attention to the need for research systems to address governance challenges posed by LLMs (Box 5).

Box 5. What do ChatGPT and future LLMs imply for the research community?

Van Dis et al. (2023) call for an international forum on the development and use of LLMs for research. The goal would be to answer questions essential to research governance. Among the questions they highlight are the following:

- Which academic skills remain essential for researchers, and in what ways might scientists' training need to change?
- Which steps in an AI-assisted research process should require human verification?
- How should research integrity and other policies change? (for example, ChatGPT does not reliably cite original sources, and researchers might use it without giving credit to earlier work. This might be unintentional).
- Most LLMs are proprietary products of large tech companies. Should this spur public investment in open-source LLMs? How could this best be done, given the much larger resources available to tech companies?

- What quality standards should be expected of LLMs (such as source crediting and transparency)? Which stakeholders should be responsible for the standards?
- How should LLMs be used to enhance principles of open science?
- How can researchers ensure that LLMs do not create inequities in research?
- What legal implications do LLMs have for scientific practice (for example, laws and regulations related to patents, copyright and ownership)?

A framework for evaluating the AI-driven automation of science

Ross King and Hector Zenil hold that the future of science, especially experimental science, lies in AI-led closed-looped automation systems. Automation has accelerated productivity in many industries, and could do so again in science. Citing a prediction of the physics Nobel Laureate Frank Wilczek that in 100 years the best physicist would be a machine, the authors underscore the importance of developing autonomous systems to improving human welfare (King himself co-developed the robot scientist "Adam", the first machine to autonomously discover scientific knowledge, generating a hypothesis which it then tested using laboratory automation, King et al. 2009). Robotic systems are already accelerating science in genetics and drug discovery (the essay by King, Peter and Courtney explores the role of robot scientists in greater depth).

The authors describe a possible future in which human scientists will decide how to work with the AI scientists and how much scope AI will have to define its own problems and solutions. Synergies could arise in which AI identifies research where humans have been biased or else highlights areas of research that human scientists have failed to explore.

A progressive scale of automation in science

King and Zenil set out a framework of automation levels in science based on the quantity and quality of input and execution required from human scientists. An analogy they draw is to the 1 to 5 classification of automation in cars set by The Society of Automotive Engineers. In science, at Level 1, humans still describe a problem in full, but machines do some data manipulation or calculation. A case might be made for dating the achievement of Level 1 to the 1950s and 1960s, with the advent of the first theorem provers. Level 5 corresponds to full automation, covering all levels of discovery with no human intervention. Today, in certain areas of laboratory-based science, some systems have reached Level 4. This is the stage where science can be greatly accelerated. For instance, a robot chemist developed at the University of Liverpool moves about the laboratory guided by Lidar and touch sensors. An algorithm lets the robot explore almost 100 million possible experiments, choosing which to do next based on previous test results. The robot can operate for days, stopping only to charge its batteries. For such machines, there is almost no human intervention except for providing consumables.

The authors are part of the "Nobel Turing Challenge". This challenge is exploring how to develop AI systems capable of making Nobel-quality scientific discoveries highly autonomously by 2050. As they report, participants at the first workshop on the Turing Challenge, in 2020, estimated that widespread uptake of Level 2 and Level 3 systems will happen within the following five years. Level 4 systems could become widespread in the next 10-15 years, and Level 5 in the next 20-30 years. Concluding, King and Zenil cite the example of a fully automated experiment that recently tested systematic research reproducibility from literature papers for the first time, illustrating progress towards Levels 4 and 5.

Using machine learning to verify scientific claims

Lucy Wang explores the current state and limitations of ML systems for scientific claim verification. She notes that there is a renewed urgency to successfully automate claim verification, driven by the significant

extent of misinformation spread on line during the COVID-19 pandemic, the sensitivity of topics such as climate change and the sheer abundance of scientific output.

Platforms like Twitter, Facebook and others engage in both manual and automated fact-checking. These companies may employ teams of fact-checkers and ML models. However, Wang notes that scientific claims pose a unique set of challenges for fact- checking due to the abundance of specialised terminology, the need for domain-specific knowledge and the inherent uncertainty of findings at the knowledge frontier.

Automated scientific claim verification has made significant advances in recent years, but technical and other challenges require further progress. Wang describes areas where more work is needed, including integrating external sources of information into veracity prediction, such as information on funding sources and sources' historical trustworthiness; how to generalise specific domains (scientific claim verification datasets are limited to a few select domains, most notably biomedicine, public health and climate change); widening the space of potential evidence documents, for example expanding from a sample of trusted scientific articles to all peer-reviewed scientific documents; and, achieving claim verification that accounts for the beliefs and needs of users.

Wang notes that questions remain around how to integrate the outputs of claim verification models with the decisions of human fact-checkers. In addition, there is little study so far on the social issues or consequences of automated scientific claim verification. For example, that the outputs of models built to assist manual fact-checking might have to be different from models built to increase the ability of lay people to engage in scientific discourse.

Robot scientists: From Adam to Eve to Genesis

Ross King, Oliver Peter and Patrick Courtney discuss the rapid pace of development in combining robotics with AI to automate aspects of the scientific process. Materials scientists, chemists and drug designers have increasingly taken up integration of AI with laboratory automation.

AI systems and robots can work more cheaply, faster, more accurately and longer than human beings (i.e. 24/7). But they have other advantages besides. As the authors explain, robot scientists can do the following:

- Flawlessly collect, record and consider vast numbers of facts.
- Systematically extract data from millions of scientific papers.
- Perform unbiased, near-optimal probabilistic reasoning.
- Generate and compare a vast number of hypotheses in parallel.
- Select near-optimal (in time and money) experiments to test hypotheses.
- Systematically describe experiments in semantic detail, automatically recording and storing results along with the associated metadata and procedures employed, in accordance with accepted standards, at no additional cost, to help reproduce work in other labs, increase knowledge transfer and improve the quality of science.
- Increase the transparency of research (fraudulent research is more difficult), standardisation and exchangeability (by reducing undocumented laboratory bias).

Furthermore, once a working robot scientist is built, it can be easily multiplied and scaled. Robotic systems are also immune to a range of hazards, including pandemic infections. All of these capabilities remain complementary to the creativity of human scientists.

Emerging laboratories in the "cloud"

King, Peter and Courtney also describe new experimentation services in the biopharmaceutical industry whereby researchers access automated labs through a user interface or an API, designing and executing

their experiments remotely. Such services could enable biopharmaceutical enterprises to operate without needing to own a laboratory. However, global cross-platform standards for cloud-based laboratories must be adopted. The authors suggest various roles for public support for robotics in science (Box 6).

Box 6. Laboratory automation: Suggestions for policy

Foster interaction between roboticists and domain experts. Industrial robotics has developed rapidly but not always in ways that meet the needs of science. Collaborative research programmes and centres could help to bridge these needs by bringing together materials scientists, chemists, AI experts and roboticists to help, for example, develop next-generation battery materials. Collaborative programmes could also facilitate road-mapping across disciplines to identify gaps, opportunities and funding priorities. Governments are best placed to create such programmes, bringing together players that otherwise rarely co-ordinate their activities.

Strengthen data governance. Laboratory instruments need to become interoperable via standardised interfaces. At present the controls and data produced are presented in a proprietary format and lack the digital metadata around an experiment. This stifles exchange and re-use of data. Laboratory users, suppliers and technology developers could be brought together and incentivised to co-operate from the moment when data are generated by funders and publishers. This might take place under open science initiatives, such as the European Open Science Cloud, that support data curation and sharing through the FAIR principles.

Support long-term collaboration across scientific disciplines. The development of cross-disciplinary research and development centres can serve as a focus for such collaboration, setting medium-term goals and providing formal training that combines engineering (robotics, AI, data, etc.) and science. For example, engineers are seldom exposed to modern, data-rich life science. When linked together, such centres (often national in reach) can also support common interests such as training and evolving research practice. OECD (2020) reviews good practice in designing and implementing cross-disciplinary research.

The Centre for Rapid Online Analysis of Reactions (ROAR), at Imperial College London, is an example of such an approach. ROAR aims at digitising chemistry, providing the missing cross-disciplinary exposure and training. Similarly, the CAT+ centre is an open-access facility for Swiss scientists combining cutting-edge high-throughput and automated experimentation equipment, as well as AI, to develop sustainable catalysts. The centre also provides training and enables collaborative work.

Support visionary initiatives with long-term impact. Initiatives such as the Nobel Turing Challenge (see the essay by King and Zenil) can galvanise and inspire collaboration and co-ordination in science and should be supported at an international level. This could help focus efforts on addressing global challenges. It could help to drive agreement on standards and attract young scientists to such ambitious endeavours.

From knowledge discovery to knowledge creation: How can literature-based discovery accelerate progress in science?

Neil Smalheiser, Gus Hahn-Powell, Dimitar Hristovski and Yakub Sebastian describe prospects for generating new scientific insight from “undiscovered public knowledge” (UPK) and literature-based discovery (LBD). UPK refers to scientific findings, hypotheses and assertions that exist within the published literature without anyone being aware of them. They may be undiscovered for many reasons. Perhaps, for instance, they were published in obscure journals or lack Internet indexing. Or perhaps multiple types of

evidence exist across different studies that address the same issue but are not integrated readily with each other (e.g. epidemiologic studies vs. case reports).

Entirely new, plausible and scientifically non-trivial hypotheses can be found by combining findings or assertions across multiple documents. If one article asserts that "A affects B" and another that "B affects C", then "A affects C" is a natural hypothesis. LBD differs from AI data mining efforts to identify explicitly stated findings or associative trends in the data. LBD attempts to identify *unknown* knowledge that is implicitly rather than explicitly stated. The problems that LBD tools are solving (generating potentially novel hypotheses) are inherently more difficult and specialised than searching the research literature (as done by PubMed and Google Scholar). And LBD is distinct from to meta-analysis, which attempts to collate comparable studies.

To date, most research on LBD has come from practitioners in computer science, information science and bioinformatics. Indeed, the authors note that LBD launched the entire field of drug repurposing. But LBD can be used much more widely. The authors show that less than 6% of all LBD publications can be mapped to at least one of the United Nations Sustainable Development Goals, even though the techniques could facilitate progress in relevant fields.

The next-generation LBD systems are also likely to use information in non-natural language forms, such as numerical tables, charts and figures, programming codes, etc. The authors suggest that advances in AI are key to improving LBD systems. Proposals for better exploiting LBD in science are set out in Box 7.

Box 7. Better utilising LBD systems in science: Suggestions for policy

Train students to search systematically for new hypotheses. The biomedical curriculum, for example, provides no such training. LBD analyses should be undertaken in dialogue or partnership between biomedical end-users and informatics consultants in response to specific research questions. For example, what molecular pathways are most promising to study in Alzheimer's disease?

Increase the availability of open research data. Platforms such as Figshare (https://figshare.com) and Zenodo (https://zenodo.org) provide open access to research data as figures, datasets, images or videos. Cloud-based bibliography management solutions (Mendeley, Zotero) and academic social networking sites (ResearchGate, Academia.edu) could open exciting possibilities for more author and community-centric LBDs. Such sites could serve as platforms for new initiatives and/or co-ordination mediated by research funders and/or policymaking bodies.

Help integrate LBD analyses into everyday science. There is no LBD tool similar to Google Scholar used by the general scientific community. Instead, LBD tools are more specialised and require some training, not unlike the training required to use statistics packages or computer programming environments. Perhaps the best way forward is not to require bench and clinical investigators to become LBD experts themselves but rather to create partnerships and collaborations with informatics consultants fluent with LBD tools. One might also envision holding workshops and conferences that address specific problems (e.g. climate change) and carry out brainstorming in conjunction with domain experts assisted by LBD analyses.

Advancing the productivity of science with citizen science and artificial intelligence

Luigi Ceccaroni, Jessica Oliver, Erin Roger, James Bibby, Paul Flemons, Katina Michael and Alexis Joly explain how AI can enhance citizen science. Advances in communication and computing technologies have enabled the public to collaboratively participate in new ways in science projects. To date, the most significant impacts of citizen science have been in data collection and processing, such as classifying

photographic images, video and audio recordings. However, citizen scientists are engaged in projects across scientific domains such as astronomy, chemistry, computer science and environmental science.

The authors describe how citizen science systems in combination with AI are advancing science by increasing the speed and scale of data processing; collecting observations in ways not achievable with traditional science; improving the quality of data collected and processed; supporting learning between humans and machines; leveraging new data sources; and diversifying engagement opportunities.

Future applications, emerging now, will include more accessible ways for non-experts to use AI techniques, along with autonomous systems of all types, such as drones, self-driving vehicles, and other robotic and remote sensing instrumentation integrated with AI. All these and other emerging applications will aid data collection and the automatic detection and identification of items in images, audio recordings or videos.

More generally, citizen science needs to find ways to break complex research projects into discrete tasks that citizen scientists can then undertake. AI might assist in this partitioning of tasks. It is also foreseeable that AI could help ensure adherence to the scientific method and assist in quality assessment (concerns over data quality remain prevalent in citizen science). The authors also describe how policy makers can help advance the use of AI in citizen science (Box 8).

Box 8. AI to help raise the productivity of science using citizen science: Suggestions for policy

Develop guidance on proper application of AI. Each use of AI in citizen science needs to carefully consider risks, traceability, transparency and upgradability. Traceability is essential to reproduce, qualify and revise the data generated by AI algorithms (e.g. through version control and accessibility of the AI models). Transparency is crucial for understanding and correcting biases in AI models (e.g. by making training data fully accessible). Without appropriate transparency, errors by AI algorithms cannot be understood or, in some cases, even detected. Upgradability – the ability of AI algorithms to be upgraded over time – is necessary to accommodate new inputs and corrections made by experts and citizen scientists.

What can artificial intelligence do for physics?

Sabine Hossenfelder observes that ML has spread to every part of physics. Furthermore, physicists themselves have been at the forefront developments in ML. The behaviour of magnets, to take one example, sheds light on some properties of machines that learn. Hossenfelder groups the applications of AI in physics into three main categories:

- Data analysis. For example, achieving fusion power requires AI-enabled solutions to the challenge of suspending super-hot unstable plasma in a ring of powerful magnets.
- Modelling. For instance, simulating some physical systems – such as how subatomic particles scatter – takes a long time. However, ML can learn to extrapolate from existing simulations without re-running the full simulation each time.
- Model analysis. For example, the theory for materials' atomic structure is known in principle. However, many calculations needed to operationalise the theory are so vast that they have exceeded computational resources. ML is beginning to change that.

Hossenfelder reiterates what other contributors to this volume also draw attention to, namely that current algorithms are not a scientific panacea. They rely heavily on humans to provide suitable input data and cannot yet formulate their own goals.

AI in drug discovery

Kristof Szalay explains that ML has been integral to parts of the process of drug development for decades. Recent improvements in AI have allowed it to enter other areas in the drug discovery. As major pharmaceutical companies have adopted a business model aimed at decreasing risk in the early parts of drug discovery – by in-licensing trial-ready compounds from smaller biotech companies – it is in small biotechnology companies where an explosion in the use of AI technologies has happened.

Szalay observes, in line with Jack Scannell's essay in this volume, that the main challenge of bringing a new drug to market is that a lot of time and money are needed before a drug's efficacy is determined by testing on patients. AI's main impact will be in selecting experiments with the best chance of yielding drugs that pass clinical testing. However, predicting which patients will respond well enough to a drug is a challenge for AI. Each patient is unique, with slightly different biochemistry. In addition, each patient can be dosed only once. If they return to the clinic, whether the drug has worked or not, their condition may have changed, essentially rendering them – for training purposes – a different patient.

Szalay also highlights a tension between the dynamic creativity of software development and the safety needs of the drug industry. Explainable AI could address this problem, and help with others, for instance in detecting biases against ethnic minorities in the composition of genomic databases. However, the leading AI models – deep-learning systems – are not explainable, and other AI approaches are not yet good enough.

AI infrastructure and the financial burden on smaller academic groups

Szalay explains that large modern AI set-ups must move all the pieces of data and the code together at large scales. AI companies have a dedicated team of engineers building the necessary scaffolding (data processing pipelines, orchestrating compute resources, database partitioning, etc.). In this way, every piece of code and data is in the right place at the right time on all the dozens of machines training the AI. This requires expertise and human resources that only make sense to gather if AI is a main focus of a business. Early discovery requires large AI systems and many training runs, with costs running from hundreds of thousands to millions of US dollars. Szalay suggests a role for policy in addressing the infrastructure challenges (Box 9).

Box 9. Access to computational infrastructure for small academic groups: Suggestions for policy

Academic groups would need a stronger AI backbone like, for example, that proposed by the National Artificial Intelligence Research Resource Task Force in the United States (NAIRR Task Force, 2022). Similar consortia such as the European Open Science Cloud (EC, n.d.) have been established recently in the European Union to support collaboration in the field. However, they are mostly focused on sharing data and tools rather than solving the problem of scaling AI in academia. One step might be to offer research grants that require universities to pool their AI resources into one single effort. Access to supercomputing centres – possibly subsidised – should include the involvement of data engineers who could help researchers get their data through the computing system.

Data-driven innovation in clinical pharmaceutical research

Joshua New explains that a major barrier to developing new treatments is the cost of evaluating candidate drugs for safety and efficacy. He cites estimates that, as of 2018, the average cost of an individual clinical trial was USD 19 million. A promising way to reduce costs is through improved use of data and AI in clinical

trial design, particularly to increase patient recruitment and engagement. Selecting a site to perform a clinical trial can be a significant financial commitment. To minimise this risk, some companies have developed AI systems that can guide site-selection decisions. Several companies are using AI to improve patient recruitment directly. They analyse structured and unstructured clinical data to better identify patients that match trial criteria, allowing trial organisers to conduct more targeted recruitment. In some cases, patients may end their participation in a trial due to the negative side effects of a treatment. Therefore, researchers have developed ML algorithms that can identify the fewest and smallest doses of a treatment, to reduce overall toxicity.

The author suggests, among other recommendations, that policy makers should expand access to institutional and non-traditional data. For example, they could reduce regulatory barriers to data sharing, better enforce publication of clinical trial results and promote data sharing with international partners.

Applying AI to real-world health-care settings and the life sciences: Tackling data privacy, security and policy challenges with federated learning

Mathieu Galtier and Darius Meadon explain that ML in health care will not successfully transition from research settings into everyday clinical practice without large, diverse and multimodal data (i.e. digital pathology, radiology and clinical). However, patient and other important data are usually stored in silos, for instance in different hospitals, companies, research centres, and across different servers and databases. Health data are also tightly regulated. While necessary, this can also hinder research. For instance, completely removing information on a patient's identity can decrease the performance of an algorithm.

The authors discuss how federated learning (FL) can overcome the challenge of fragmented health data. With FL, algorithms are dispatched to different data centres where they train locally. Once improved, the algorithms return to a central location. The data themselves do not need to be shared (FL is one part of broader family of "privacy-enhancing technologies" that can be applied to AI. Other examples include differential privacy, homomorphic encryption, secure multiparty computation and distributed analytics).

Many start-ups now provide FL platforms, but few have managed to apply these in real-world settings at scale. The public sector has started to become active. The UK government, for example, has outlined a plan to set up a federated infrastructure for managing UK genomics data. The authors set out suggestions for policy (Box 10).

Box 10. Expanding the use of federated learning across research centres: Suggestions for policy

Governments can assist through public financing, especially in helping research centres to adopt a decentralised approach and to create shared infrastructure. Public funding is important because the level of co-operation needed would otherwise emerge slowly. Any funding should be conditional on the recipient infrastructure being governed on the basis of a shared set of rules and protocols for, for example, interoperability, data portability and security. More broadly, governments can take steps to harness the power of data across various fields, from health to climate. For example, in 2022 the European Commission presented its Health Data Space (HDS) (EC, 2022). The HDS aims to create a trustworthy and efficient context for the use of health data for research, innovation, policy making and regulation. More broadly, the OECD *Recommendation of the Council concerning Access to Research Data from Public Funding* provides guidance to governments on enhancing access to research data (OECD, 2021).

AI and science in the near future: Challenges and ways forward

Artificial intelligence in scientific discovery: Challenges and opportunities

Hector Zenil and Ross King consider challenges and opportunities in using AI for science. Their key insights concern the differences between the two main forms of ML learning: statistical ML, the most used and successful form, which is based upon complex pattern learning, and model-driven ML.

As the authors explain, the ability of human scientists to reason rationally, to do abstract modelling and to make logical inferences (deduction and abduction) is central to science. However, these abilities are handled poorly by statistical ML. Statistical ML operates differently from the human mind. Humans build abstract models of the world that allow mental simulations on the fly of how an object can be modified. They can also generalise even if they have never encountered the same situation before. Humans do not need to drive millions of miles to pass a driving test, for example. Model-driven methods can explain more observations with less training data, just as human scientists do when they derive models from sparse data. For instance, Newton and others derived the classical theory of gravitation from relatively few observations.

Pointing to limitations in statistical ML the authors draw attention to the large amounts of data it requires, which are often unavailable in some realms of science; problems associated with data annotation and labelling (for example, it takes time and resources to label large databases by hand, and those doing the labelling might have different levels of competence); variation in features of the data across some areas of science, which may not allow generalisation across fields; and, the black-box character of statistical ML approaches.

No matter how abundant the supply of data, the problem of understanding and transfer learning (generalisation) cannot be solved simply by applying ever-more powerful statistical computation.

Too little attention, research effort, conference venues, journals and funds are available to AI approaches that differ from statistical ML, such as deep learning. This is a consequence of the dominant role of some academic actors and corporate AI research and development (see the essay in this volume by Mateos-Garcia and Klinger).

Computers are still unable to formulate interesting research questions, design proper experiments, and understand and describe their limitations. More resources are needed to develop the methodological frameworks most relevant to the AI required for further progress in scientific discovery.

Machine reading: Successes, challenges and implications for science

Jesse Dunietz examines the capabilities of state-of-the-art natural language processing (NLP). NLP, researchers hope, could assist scientists by automating some of the reading of scientific papers. Dunietz lays out a variety of reading comprehension tasks that NLP systems might perform on scientific literature, placing these on a spectrum of sophistication based on how humans comprehend written material.

The author shows that current NLP techniques grow less capable as tasks require more sophisticated understanding. For example, today's systems excel at flagging names of chemicals. However, they are only moderately reliable at extracting machine-friendly assertions about those chemicals, and they fall far short of, say, explaining why a given chemical was chosen over plausible alternatives.

The fundamental problem is that NLP techniques lack rich models of the world to which they can ground language (the essay by Ken Forbus explains the importance of knowledge bases and graphs in addressing this problem). They have no exposure to the entities, relationships, events, experiences and so forth that a text speaks about. As a result, even the most sophisticated models still often generate fabrications or outright nonsense.

The author observes that a surprisingly large fraction of research on NLP applied to science has focused only on the surface structure of texts, such as finding key words. Research policies may be able to facilitate progress towards machines capable of sophisticated comprehension of what they read, including scientific papers. To that end, Dunietz proposes two possible ways forward (Box 11).

Box 11. Making progress in machine reading of scientific texts: Suggestions for policy

Foster new, interdisciplinary, blue-sky thinking: NLP research is often driven by the pursuit of standardised metrics, by expectations of quick publications and by the allure of the low-hanging fruit from the past decade's progress. This environment produces much high-quality work, but it offers limited incentives for the sort of high-risk, speculative ideation that breakthroughs may need. Research centres, funding streams and/or publication processes could be set up to reward novel methods – even if at a nascent stage. These steps could be taken without prioritising publishing speed, performance metrics and immediate commercial applicability.

Support under-studied research: Policy makers can fund specific areas of under-studied research. To this end, prioritising and funding selected techniques may prove less important than funding aimed at achieving specific tasks. The most sophisticated forms of machine reading seem likeliest to emerge where systems must communicate with humans to perform tasks in a real or simulated physical environment.

Interpretability: Should – and can – we understand the reasoning of machine learning systems?

Hugh Cartwright examines the inability of the most powerful ML systems to explain their output, and what means for science, where elucidating the link between cause and effect is fundamental. He notes that not all forms of AI lack interpretability: tools, such as decision trees or reverse engineering offer some insight into their own logic. However, most scale poorly with software complexity and are of value only to experts.

Cartwright describes why interpretation in science poses particular conceptual challenges, even if ML could explain its own logic. As science continues to evolve, some topics may become so intellectually demanding that no one can understand them (he gives an example from the mathematics of string theory, understandable perhaps to only a few specialists). If an AI system were to discover such knowledge, it is unclear what an explanation for human scientists would look like. Similarly, translating into human-digestible form what an AI system has learnt in a hugely dimensional data space may yield hard-to-understand lines of reasoning, even if individual parts of the argument are clear.

In some cases, explanations need to be illustrated by images. However, Cartwright points out that while image recognition applications have progressed, it is challenging for AI systems to construct images to assist explanation. In addition, explanation mechanisms may not port well from one application area to another.

A risk exists, in Cartwright's view, that the demand for useful, commercially valuable, AI may outstrip progress on explanation.

Combining collective and machine intelligence at the knowledge frontier

Eirini Malliaraki and Aleks Berditchevskaia highlight that while AI has greatly advanced, humans have unique abilities such as intuition, contextualisation and abstraction. Consequently, novel AI and human collaborations could advance science in new ways. Properly orchestrated, the capabilities of collaborating

individuals can exceed the sum of the capabilities of the same individuals working in isolation. This is "collective intelligence".

Malliaraki and Berditchevskaia observe that a robust understanding of how to make the most of collective intelligence in science is only beginning to emerge. In addition, progress in combining human collective intelligence and AI is important because science is now carried out by ever-larger teams and international consortia. The authors describe how AI-human collaborations can improve upon current approaches to mapping the knowledge frontier in a number of ways, including those described below.

Encoding and discovering knowledge

Today's science communication infrastructure does not help researchers make the best use of predominantly document-centric scholarly outputs. For example, words and sentences may be searched for, but images, references, symbols and other semantics are mostly inaccessible to current machines. Recent advances in language models can help but do not work well outside the domains where they are developed. Harnessing complementary expertise from among scientists and policy makers would assist.

Connecting and structuring knowledge

Once relevant public knowledge is encoded and discovered it needs to be organised and synthesised. With recent advances in knowledge representation and human-machine interaction, scholarly information can be expressed as knowledge graphs (see Ken Forbus' essay on knowledge bases and graphs). Current automatic approaches to create these graphs have limited accuracy and coverage. Hybrid human-AI systems help.

Oversight and quality control

A knowledge synthesis infrastructure will not be complete without ongoing curation and quality assurance by domain experts, librarians and information scientists. Automated systems to check scientific papers are helpful, but they require augmentation by distributed peer review or the crowdsourced intelligence of experts.

Malliaraki and Berditchevskaia suggest how policy could accelerate the integration of combined AI-human systems into mainstream science (Box 12).

Box 12. Integrating combined AI-human systems into mainstream science: Suggestions for policy

Develop tools to enhance AI and collective intelligence combinations: Co-operative human-AI systems will have to navigate problems where the goals of different actors and organisations are in tension with one another, as well as those where actors have common agendas. For instance, some academic groups are in competition. They may not be incentivised to share for fear of being scooped or may simply have conflicting approaches to a method or a problem. While there has been some research in this area –such as www.cooperativeai.com/ – investment in this field of research has lagged other topics in AI.

Make use of existing social networks to experiment with human-AI collaboration: Social platforms such as Academia.edu and the Loop community support knowledge exchange between academics and provide an infrastructure for literature discovery. Some of these platforms already use AI-enabled recommendation systems. Such platforms could become testbeds for experimenting with combined human-AI knowledge discovery, idea generation and synthesis. The benefit of these platforms is that they already have an engaged community united around a common interest/purpose. An extended

functionality would need to align with or enhance that common purpose. Working together with researchers, funding and/or incentives provided by research funders might catalyse progress. Such investment could also be connected to mission-oriented research agendas.

Re-think incentives for knowledge mapping and synthesis: Several institutional and educational conditions inhibit work on knowledge integration. Existing measures of publishability motivate discoveries built on individual disciplines rather than knowledge synthesis. New integrative PhD programmes and/or industry research programmes based on knowledge synthesis might help. Research councils and academic institutions should experiment with these proposals and support new roles and career paths. They could support the development of expertise in curating and maintaining information infrastructure, which could also help to build bridges between the public, academia and industry.

Elicit: Language models as research tools

Jungwon Byun and Andreas Stuhlmüller examine how ML could change research over the next decade. Intelligent research assistants could increase the productivity of science, for instance by enabling qualitatively new work, making research accessible to non-experts, and reducing what can be extraordinary and sometimes fruitless calls on scientists' time (for example, one study in Australia found that 400 years of researchers' time was spent preparing unfunded grant proposals for support from a single health research fund, Herbert, Barnett and Graves, 2013).

Byun and Stuhlmüller observe that existing research tools are not designed to direct the researcher quickly and systematically to research-backed answers. In response, the authors have helped to build Elicit, a research assistant that uses language models – including GPT-3, an LLM trained on hundreds of billions of words on the Internet. Researchers today primarily use Elicit for literature search, review, summarisation and rephrasing, classification, identifying which papers are randomised controlled trials, and automatically extracting key information, such as a study's sample population, study location, measured outcomes, etc.

As the authors explain, LLMs are text predictors. Given a text prefix, they try to produce the most plausible completion, calculating a probability distribution on the possible completions. For example, given the prefix "The dog chased the", GPT-3 assigns 12% to the probability that the next word is "cat", 6% that it is "man", 5% that it is "car", 4% that it is "ball", etc. LLMs can complete many tasks without specific training, including question answering, summarisation, writing computer code and text-based classification. Hundreds of applications have been built on top of GPT-3, for purposes such as customer support, software engineering and ad copywriting.

The enormous public interest in ChatGPT has drawn attention to the power of LLMs. Through Elicit, progress in LLMs such as ChatGPT directly translates into better tooling for researchers. Better language models mean Elicit finds more relevant studies, more correctly summarises them and more accurately extracts details from them to help evaluate relevance or trustworthiness. It is expected that newer language models will help with tasks like giving practical guidance on promising avenues of research.

The launch of models like ChatGPT and Galactica has emphasised the need for processes of evaluation to ensure accuracy as applications are scaled up. Their abstractive intelligence directly trades off with accuracy and faithfulness. These models are not fundamentally trained to speak accurately or stay faithful to some ground truth.

Byun and Stuhlmüller point out that as of early 2022 there are no guarantees that LLMs will help substantially with research, which requires deep domain expertise and careful assessment of arguments and evidence. However, on the assumption that their performance will continue to improve, the authors sketch an intriguing picture of what LLM-based research assistants might be capable of in a medium-term future (Box 13).

Box 13. AI research assistants in a medium-term future

In the future, researchers might generate a team of their own AI research assistants, each specialising in different tasks. Some of these assistants will represent the researcher and the researcher's specific preferences about things like which questions to work on and how to phrase conclusions. Some researchers are already fine-tuning language models on their own notes.

Some of the assistants will do work that researchers today might delegate to contractors or interns, like extracting references and metadata from papers. Other assistants will use more expertise than the researcher. For example, they might help a researcher evaluate the trustworthiness of findings by aggregating the heuristics of many experts.

Some assistants might help the researcher think about effective delegation strategies, sub-delegating tasks to other AI assistants. Some will help the researcher evaluate the work of these other assistants. This sub-delegation support would allow the researcher to zoom into any sub-task and troubleshoot, using assistants for help if needed. Human researchers could oversee the work of such a team of assistants to ensure it is aligned with their intent.

Byun and Stuhlmüller suggest that LLMs in research could also bring risks. To help policy makers prepare, two of these possible risks are described in Box 14.

Box 14. A note for policy makers: Possible risks from the use of language models in research

A risk that shallow work becomes easier: Language models might become good enough to be widely used to speed up content generation but not good enough to evaluate arguments and evidence well. In that case, the publish-or-perish dynamics of academia may reward researchers who (ab)use language models to publish low-quality content. This could disadvantage researchers who take more time to publish higher quality research. Language models might also favour certain types of research over others. The scientific community will need to monitor and respond to such dynamics.

Risks from data-dependent performance: Language models are trained on text on the Internet by (to date) companies mostly headquartered in English-speaking countries. They therefore demonstrate English- and Western-centric biases. Without measures that let users control this bias, these language models may exacerbate a "rich get richer" effect. More generally, broad adoption of language models requires infrastructure that enables users to understand and control what the models do and why.

Democratising AI to accelerate scientific discovery

As Joaquin Vanschoren and other authors in this volume explain, developing well-performing AI models often requires large interdisciplinary teams of excellent scientists and engineers, large datasets and significant computational resources. The current intense competition for highly trained AI experts makes it hard to scale such projects across thousands of labs. Vanschoren's essay explores progress in automating the design of ML models – AutoML – enabling more and smaller teams to use it effectively in breakthrough scientific research.

Advances in self-learning AutoML are accelerated by the emergence of open AI data platforms like OpenML. Such platforms host or index many datasets representing different scientific problems. For each dataset, one can look up the best models trained on them and the best ways to pre-process the data they use. When new models are found for new tasks they can also be shared on the platforms, creating a

collective AI memory. Vanschoren suggests that, as has been done for global databases of genetic sequences or astronomical observations, information should be collected and placed on line on how to build AI models. Data should also be put through tools that help structure them to facilitate analysis using AI.

Work to automate AI has only scratched the surface of what is possible. Fully realising this potential will require co-operation between AI experts, domain scientists and policy makers. The authors suggests policy measures to help bring this about (Box 15).

Box 15. Automating the design of machine learning models for science: Suggestions for policy

Support AutoML for real-world problems. Most AutoML researchers only evaluate their methods againts technical performance benchmarks instead of on scientific problems where they could have much more impact. Challenges around AutoML for science could be organised, or research could be funded that involves directly applying AutoML in AI-driven science.

Encourage more collaboration. On a larger scale, support should be given for the development of open platforms such as OpenML and DynaBench that track which AI models work best for a wide range of problems. While these platforms are already having an impact in AI research, public support is needed to make them easier to use across many scientific fields, and to ensure their long-term availability and reliability. For instance, interlinking scientific data infrastructure would link the latest scientific datasets to the best AI models known for that data in an easily accessible way. In the past, agreements around rapid public sharing of genome data – the Bermuda principles –led to the creation of global genome databases critical to research. Doing the same for AI models, and building databases of the best AI models for all kinds of scientific problems, could dramatically facilitate their use to accelerate science.

In addition, to create new incentives for scientists, such platforms could track dataset and model re-use, much like existing paper citation tracking services. That way, researchers would receive proper credit for sharing datasets and AI models. This would require analysis of all AI literature to identify the use of datasets and models inside papers, which is non-trivial. It would also require new ways to reference datasets and models in the literature. Commercial entities have few incentives to work on this (Google Dataset Search is valuable and shows some usage metrics for datasets, but this is based on proprietary information that cannot be shared.) Hence, a public initiative is needed to collect and publish this information on datasets and model re-use and provide true incentives for researchers to share their datasets and models. The public funding required would be small.

Is there a narrowing of diversity in AI research?

Juan Mateos-Garcia and Joel Klinger examine changes in the diversity of AI research. They note that recent advances in AI have in great part been driven by deep-learning techniques developed and/or deployed at scale by large technology companies. Many of the ideas underpinning these advances originated in academia and public research labs. At the same time, researchers in universities and the public sector are increasingly adopting powerful software tools and models developed in industry.

However, the authors point out that the short-term benefits of rapid advances in deep learning and the tighter intertwining of public and private research agendas is not without risks. Indeed, several scientists and technologists have expressed concerns about the possible downsides of the data and compute-intensive deep-learning methods that dominate AI research. For instance, with significantly larger models available to industry, academics could find it difficult to develop competing models, interpret industry models and develop public use alternatives. Some evidence also suggests that industry is draining

researchers from academia. In 2004, for example, 21% of AI PhDs in the United States went to industry, compared to almost 70% in 2020 (Ahmed, Wahed and Thompson, 2023). Similarly, Mateos-Garcia and Klinger cite evidence of skewed research priorities in public research labs that receive private funding from and/or collaborate with industry to access the large datasets and infrastructures required for cutting-edge research.

Klinger et al. (2020) conducted a quantitative analysis of 1.8 million articles from *arXiv*, a preprint repository widely used by the AI research community. They showed the following:

- There is evidence of a recent stagnation and even decline in the diversity of AI research.
- Private AI research is thematically narrower and more influential than academic research, and it focuses on computationally intensive deep-learning techniques.
- Private companies tend to specialise in deep learning and applications in online search, social media and ad-targeting. They tend to be less focused on health applications of AI and analyses of the societal implications of AI.

Some of the largest and most prestigious universities have lower levels of thematic diversity in AI research than would be expected given their volume of activity and public nature. Such influential universities tend to be the top collaborators of private companies.

The authors make various policy suggestions (Box 16).

Box 16. Increasing the thematic diversity AI research: Suggestions for policy

Universities tend to produce more diverse AI research than the private sector, so bolstering public R&D might make the field more diverse. This could be done by increasing the levels of research funding, the supply of talent, computational infrastructure and data for publicly oriented AI research. A larger talent pool would reduce the impact of a migration of AI researchers from universities to industry. Better public cloud and data infrastructures would also make academic researchers less reliant on collaboration with private companies.

Funders should pay special attention to projects that explore new techniques and methods separate from the dominant deep-learning paradigm. This may require patience and a tolerance of failure.

New datasets, benchmarks and metrics could highlight the limitations of deep-learning techniques and the advantages of their alternatives. In so doing, they could help steer the efforts of AI research teams. Mission-driven innovation policies could encourage deployment of AI techniques to tackle big societal challenges, which could in turn spur development of new techniques more relevant for domains where deep learning is less suitable.

While funding institutions often engage the research community in their decision making, policy makers may need more expertise and know-how to help them decide what sort of technology initiatives to support. Policy makers could also help to further examine and quantify any losses of technological resilience, creativity and inclusiveness brought about by a narrowing of AI research.

Lessons from shortcomings in machine learning for medical imaging

Gaël Varoquaux and Veronika Cheplygina note that the application of ML to medical imaging has attracted much attention in recent years. Yet, for various reasons, progress remains slow and the impact on clinical practice has not met expectations. Studies for many clinical applications of ML – including COVID 19 – have failed to find reliable published prediction models.

Varoquaux and Cheplygina show that progress is not guaranteed by having larger datasets and developing more algorithms. For example, analysis of predictions of Alzheimer's disease from more than 500 publications shows that studies with larger sample sizes tend to report worse prediction accuracy. The authors suggest reasons for this. Not all clinical tasks translate neatly into ML tasks. In addition, creating large datasets often relies on automatic methods that may introduce errors and bias into the data. For example, a machine might wrongly label x-rays as showing the presence or non-presence of pneumonia based on wording in the associated radiology reports.

Norms should be established whereby datasets include a report of the data's characteristics, and the potential implications for models trained on the data. Benchmarking the performance of algorithms alone is also not sufficient to advance the field. Papers focusing on understanding, replication of earlier results and so forth are also valuable.

The authors stress the importance of open science and highlight the need to make work on curated datasets and open-source software that everybody can use more attractive. They note it is difficult to acquire funding, and often to publish, when working on such projects. Many team members are therefore volunteers. More regular funding and more secure positions would help to improve on the status quo. Other policy-relevant suggestions relate to the need for greater, quality and evaluation of research. These observations – set out in Box 17 – are also relevant to ML in science more generally, as the growth of methods is rapid and institutional incentives sometimes prize novelty.

Box 17. Machine learning in medical imaging and other fields of science: policies to avoid the primacy of novelty

Set incentives to encourage research on methods with greater validation: As research positions and funding are often tied to the output of publications, researchers have strong incentives to optimise for publication-related metrics. Metrics that prize novelty and state-of-the-art results create incentives to submit papers using novel methods that are under-validated. External incentives are needed to accelerate the change towards methods with greater validation.

Provide funding for rigorous evaluation: Funding should focus less on perceived novelty, and more on rigorous evaluation practices. Such practices could include evaluation of existing algorithms, and replication of existing studies. This would provide more realistic evaluations of how algorithms might perform in practice. Ideally, such funding schemes should be accessible to early career researchers, for example, by not requiring a permanent position at application.

Artificial intelligence in science: Further implications for public policy

Artificial intelligence for science and engineering: A priority for public investment in research and development

Tony Hey reviews the evolving history of data-led science. He observes that greatly increased data volumes are expected for the next generation of scientific experiments. AI will be needed to automate the data collection pipelines and enhance the analysis phase of such experiments.

Hey asks if academic researchers can compete with recent breakthroughs in science achieved by large tech companies using powerful and expensive computational resources and large multidisciplinary teams. He holds that a number of publicly driven actions are needed to address this situation, along with investments in R&D on foundational topics in the science of AI itself (Box 18).

Box 18. Public research initiatives and R&D priorities for AI in science: Suggestions for policy

Broad multidisciplinary programmes are needed to enable scientists, engineers and industry to collaborate with computer scientists, applied mathematicians and statisticians to solve challenges using a range of AI technologies. This needs dedicated government funding with processes that encourage such collaboration rather than stove-piped funding allocated to individual disciplines. In the United States, the National Science Foundation recently established 18 National AI Research Institutes involving research partnerships in 40 states.

New AI hardware is being developed in industry for data centres, autonomous driving systems and gaming, among others. The research community could work with industry to co-design heterogeneous compute systems that use the new architectures and tools.

Multidisciplinary programmes should create a shared cloud infrastructure that allows researchers to access the necessary computing resources for AI R&D. In the United States, the planned National AI Research Resource is intended to be a shared research infrastructure that will provide AI researchers with significantly expanded access to computational resources, high-quality data, user support and educational tools (NAIRR, 2022).

Prioritise areas of public R&D support. DOE (2020) – which Hey helped prepare – describes topics on which research breakthroughs are needed to broaden and deepen AI's uses in science and engineering. They include the need for the following:

- Go beyond current models driven only by data or simple algorithms, laws and constraints.
- Automate the large-scale creation of findable, accessible, interoperable and reusable (FAIR) data from a diverse range of sources, ranging from experimental facilities and computational models to environmental sensors and satellite data streams.

Advances are also needed in foundational topics in the science of AI itself. This includes developing frameworks and tools to help establish: that a given problem is solvable by AI/ML methods; the limits of AI techniques; the quantification of uncertainties when using AI; and, the conditions that give assurance of an AI system's predictions and decisions.

The importance of knowledge bases for artificial intelligence in science

Knowledge bases and graphs are foundational to human interaction with much of the digital world. Everyday use of a search engine or recommender system typically draws on a knowledge base or graph. They organise the world's knowledge by mapping the connections between different concepts, using information from many sources. Ken Forbus explains that for AI systems to realise their full potential to increase the productivity of science they need knowledge bases so as to understand individual domains of science, the world in which each domain is embedded, and how domains connect with each other.

There are many kinds of knowledge. For some types, the commercial world has already deployed knowledge bases (like Microsoft's Satori and Google's Knowledge Graph) with billions of facts to support web search, advertising placement and simple forms of question answering. Forbus describes the state of the art in knowledge bases and graphs and the improvements needed to support broader uses of AI in science. These improvements include the creation of bases that capture:

- Commonsense knowledge, to tie scientific concepts to the everyday world and to provide common ground for communication with human partners.
- Connections across domains of science, to help address problems which span multiple areas.

- Professional knowledge, to connect professional concepts with each other and the everyday world.
- Robust reasoning techniques that go beyond simple information retrieval.

While a large-scale high-quality graph of commonsense knowledge would benefit everyone, the effort needed to build one is beyond the usual research horizons of the private sector, and public action is needed (Box 19).

Box 19. Building knowledge bases for AI in science: A suggestion for policy

Governments should support an extensive programme to build knowledge bases essential to AI in science. This will not be done by the private sector. Support could aim to create an open knowledge network to serve as a resource for the whole AI research community. Open licensing of such a resource – such as Creative Commons Attribution Only – matters. However, to maximise utility to the scientific community, in terms of impact, reusability, replicability and dissemination, funding is needed for the construction of open knowledge graphs.

Relatively small amounts of public money could bring together scientists from AI and other domains of science to build the knowledge bases essential for AI to utilise and communicate professional and commonsense knowledge. In biology, for example, efforts could focus beyond biochemistry or genetics to produce everyday knowledge about animals and plants that connects professional concepts to the everyday world. Other efforts should use community testbeds where commonsense reasoning is needed, e.g. robotics.

Funding teams through professional societies could help enlist talent in each field to help. Funding teams in multiple disciplines which interact (e.g. climatology, biology, and chemistry) could help ensure better interoperability in the knowledge bases produced. More than most other professions, scientists recognise the value of knowledge bases and would likely be willing to contribute. As with Wikipedia, enlisting volunteer efforts to help develop commonsense knowledge graphs will be essential. Some distant curation, as found in citizen-science crowdsourcing projects, would be useful.

Among other outputs, new knowledge graphs could develop machine understandable vocabulary to integrate knowledge across sub-areas within a scientific field and across scientific fields.

The ultimate aim is a federation of knowledge graphs, ideally continually updated as research progresses and eventually encompassing all scientific knowledge.

High-performance computing leadership to enable advances in artificial intelligence and a thriving compute ecosystem

From the Oak Ridge Leadership Computing Facility (OLCF) – a part of the United States Department of Energy – Georgia Tourassi, Mallikarjun Shankar and Feiyi Wang note that high-performance computing (HPC) is essential in leading-edge science. The importance of HPC is only likely to grow as – as seems probable – the performance of ML systems improves. Countries are competing to develop ever-more powerful HPC systems. To increase HPC capabilities in the United States, Congress passed the Department of Energy High-End Computing Revitalization Act of 2004 (DOE, 2022), which called for leadership in computing systems.

The power of new computing systems, combined with the concentration of AI talent, could limit research opportunities for developing countries and lesser-resourced universities. Partly to address this risk, the OLCF allocates compute resources using two competitive programmes. Extramural panels decide on the allocations, including to users in developing countries. The requests typically exceed the available

resources by up to five times. Allocations of computing resources are typically 100 times greater than routinely available for university, laboratory, and industrial scientific and engineering environments.

The AI compute ecosystem: Gaps and opportunities

The authors explain that major corporations have developed software and specialised hardware for AI. Tools such as TensorFlow (originating in Google) and PyTorch (originating in Facebook) have been distributed in the open-source community. However, while cloud vendors such as Google Colab and Microsoft Azure also offer free allocations of computing resources, these offerings have limitations. For example, to maintain maximal schedule flexibility, Colab resources are not guaranteed and not unlimited. Access to the graphics processing units (GPUs) – essential for AI – may also be limited. Such practices hinder even moderate scientific and technical R&D.

The authors identify two main areas where systematic approaches led by nations at the forefront of this field can help in alleviating computing and data availability constraints (Box 20).

Box 20. Increasing access to high-performance computing for advances in AI and science: Suggestions for policy

Computing infrastructure and software availability could be stewarded to support open science. The open-source ecosystem is a thriving location for these tools and capabilities. However, curating best practices and applications that may be shared in a rapidly changing field is critical for the global community to benefit from emerging advances. How applications must be scaled up – which is crucial to AI – cannot be the sole province of a handful of large firms.

Nationally funded laboratories and their computing infrastructures, in collaboration with industry and academia, could also nurture the AI ecosystems for tertiary educational entities and partner countries (especially those that are only beginning to build core competencies in this field). Step-up guides from basic skills to scalable data and software management will be needed in tutorial-accessible form. This would enable students and practitioners to begin on their personal computers or small-scale cloud resources. They could then advance to larger cloud or institutional-scale resources, and then to national-scale resources.

Countries at the forefront of the field, including the United States and leaders in the European Union, may collaborate on policy frameworks to make resources available in a shared pool for deserving entities. Major commercial providers today offer computing grants to academic institutions. This model could be expanded to share computing resources and frameworks, potentially across all OECD countries. Such sharing could assist nascent and growing initiatives, help prevent reinvention and provide secondary benefits such as workforce development and fast knowledge dissemination. Frameworks could address co-ordination, sequencing of efforts, agreement on respective resource allocations among partners, and how pooling and sharing can be done while accounting for different national policies on data access or the use of sensitive data, and the need to ensure ethical AI. Under the EU-Japan Digital Partnership an action is launching – as of early 2023 – to provide mutual access to HPC. This could hopefully provide insights on how such a pool of shared resources may be implemented in future.

Improving reproducibility of artificial intelligence research to increase trust and productivity

Odd Erik Gundersen addresses the problem of limited reproducibility of AI research and scientific research more generally. He points to studies suggesting that up to 70% of AI research may not be reproducible

(the highest level of reproducibility is in physics). Irreproducibility has been documented in many of the technical subfields of AI, as well as in such application domains as medicine and social sciences. Increasing the rate of published reproducible findings will increase the productivity of science, and more importantly, increase trust in it.

Gundersen illustrates the major sources of irreproducibility as they affect AI research. These include how studies are designed (e.g. if comparing a state-of-the-art deep-learning algorithm for a given task to one that is not state of the art); the choices of ML algorithms and training processes; choices related to the software and hardware used; how data are generated, processed and augmented; the broader environment in which studies are located (e.g. a system might fail to recognise images of coffee mugs simply because some have handles pointing in different directions than others); how researchers evaluate and report their findings; and, how well the study documentation reflects the actual experiment.

Suggesting that an achievable goal is to reduce the proportion of irreproducible studies in AI to the level of physics, Gundersen describes measures that could be adopted in research systems (Box 21).

Box 21. Improving the reproducibility of AI research: Suggestions for research systems and policy

Research institutions: Research institutions should ensure that best practices for AI research are followed. This includes training employees and providing quality assurance processes. They should ensure that research projects set aside enough time for quality assurance. Adherence to quality and transparent research practices should also play a role in hiring researchers.

Publishers: Few publishers standardise the review process and provide instructions that reviewers should follow. This contrasts with the peer review that occurs as part of AI conferences, which involves checklists and structured information that reviewers should provide. It would help if journals used formal structures to check different sources of irreproducibility. Furthermore, journals should encourage publishing code and data in scientific articles.

Funding agencies: Funding agencies can select evaluators with a good track record of open and transparent research. They can also require that funded research be published in open-access journals and conferences. Finally, and most importantly, they can require both code and data to be shared freely with third parties, allowing them to run experiments on different hardware (although, for reasons Gundersen explains, this will not solve all issues with reproducibility).

AI and scientific productivity: Considering policy and governance challenges

Kieron Flanagan, Barbara Ribeiro and Priscilla Ferri explore various science policy and governance implications of AI, drawing in part on lessons from previous waves of automation in science. The authors highlight that scientific work involves many diverse roles. Some labour-intensive, routine and mundane practices may be replaceable by automated tools. However, the adoption of new tools can also create a demand for new routine and mundane tasks that must be incorporated into the practice of science (e.g. from preparing and supervising robots to checking and standardising large volumes of data).

The authors note that early career researchers are likely to perform the tasks created by adoption of new AI tools. Such tasks include data curation, cleaning and labelling. Deeper automation of scientific work might pose employment-related risks to such scientific workers.

In one key observation, the research environment is also the environment in which researchers are trained. Graduate students and post-docs learn not only lab and analytical skills and practices but – like apprentices

– they also learn the assumptions and cultures of the communities they are embedded in. Wider adoption of AI in science could affect the quantity and quality of those training opportunities.

The authors draw attention to the possibility that automating manual or cognitive practices might risk that some scientific skills are lost. If critical scientific techniques and processes become "black-boxed", students, as well as early career and other researchers, may not get the opportunity to fully learn or understand them. In a similar way, the earlier black-boxing of statistical analysis in software packages may have contributed to misapplications of statistical tests.

Questions also arise about how future automation in the public research base will be funded. The authors observe that funding and governance processes must often adapt to new scientific tools. Overall, the cost effects of the adoption of new tools may be difficult to predict. Some AI tools entail little or no cost. However, AI tools are part of wider systems of data collection, curation, storage and validation, skilled technical and user support staff, preparation and analysis facilities and other complementary assets. Some robotic systems may be particularly expensive. Evidence exists that competitive project-based grant funding systems struggle to fund mid-range and generic research equipment that may be used across many projects and grants. Thus, research policies need to consider both how to fund new tools and how to ensure support for complementary assets.

Flanagan, Ribeiro and Ferri also consider AI's roles in research governance, including in funding body processes. Experiments have used AI to identify peer reviewers for grant proposals, with the promise of speeding up the matching of reviewers with applications as well as avoiding lobbying or networks of influence. However, policymakers need to be alert to the risk that these uses of AI could introduce new biases into review processes. For example, an AI system might select reviewers who have conflicts of interest. There has also been much interest in tools to partially automate aspects of the funding or journal peer review process. This has raised similar concerns about the consequences of hidden biases within black-boxed processes. It has also raised questions around the implications for sensitive funding decisions of even small inaccuracies in machine predictions (for a recent example, published after this essay was completed, see Thelwall et al. (2023). Box 22 describes possible implications for policy makers and research systems from the authors' analyses.

Box 22. Governance challenges raised by AI in science: Suggestions for policy

Conduct *ex ante* and real-time assessments of the impacts of technological change on research. The potential impacts of AI on everyday scientific practice and the structures and dynamics of science, including work and training, must be better understood. Requirements for such assessments should be embedded within funding calls and made conditional in inviting plans for capital investments in infrastructure. Assessment should never be left to the promoters of new technologies, and should draw meaningfully on interdisciplinary expertise, including from the social sciences and humanities.

Following from the above suggestion, funders and policy makers should establish response mechanisms to act on insights from ex-ante and real-time assessments. This is a key dimension of the practice of responsible research and innovation, but one that is often forgotten. Policies that support AI must consider and learn from real-world experiments as they are developed and revised. This should be done transparently and in dialogue with the scientific community. Funders and policy makers could do this in part by establishing and supporting new independent fora for ongoing dialogue about the changing nature of scientific work and its impacts on research productivity and culture.

A further point on governance – a danger of dual use of AI in science

An additional point on governance (not raised by Flanagan, Ribeiro and Ferri) concerns the possible dual use of AI in drug discovery. Urbina et al. (2022) describe their biopharma company's exploration of how AI models originally created to avoid toxicity in drug discovery could also be used to design toxic molecules.

The authors show that by drawing on publicly available databases they could design compounds more lethal than the most lethal chemical warfare agents available. Indeed, in just six hours their model generated 40 000 molecules similar to the nerve agent VX. The primary purpose of this work was to draw attention to dangers inherent in the diffusion of AI and molecule synthesis (the authors did not synthesise the molecules they designed but noted that many companies offer synthesis services and that these are poorly regulated). Work on autonomous synthesis – the laboratory robots discussed elsewhere in this book – could soon lead to an automatic closed-loop cycle designing, making and testing toxic agents. Furthermore, the intersection of AI and autonomous systems lowers the need for domain-specific expertise in chemistry and toxicology. It is unclear how to control for these dangers, which have been little discussed in the broader context of AI governance. However, the issue is urgent, and the authors offer some initial suggestions (Box 23).

Box 23. The dangers of dual use of AI-powered drug discovery: Preliminary ideas for policy and research system governance

- Scientific conferences and learned societies should foster a dialogue involving industry academia and policy makers on the implications of emerging dual use tools in drug discovery.
- Requirements for impact statements might be set for authors submitting work involving the relevant technologies to conferences, institutional review bodies and funding agencies.
- Inspired by existing frameworks for responsible science – such as the Hague Ethical Guidelines – a code of conduct might be developed and agreed to by pharmaceutical and other companies. Such a code would contain articles on employee training, preventing misuse and unauthorised access to critical technologies, among others.
- Develop a reporting structure or hotline to alert authorities should persons or companies seek to develop toxic molecules for non-therapeutic purposes.
- Create a public facing API for AI models, with code and data available on request, to help control how models are used.
- Redouble efforts in universities to provide ethics training for science students, particularly those in computer science, and raise awareness of the possible misuse of AI in science.

Artificial intelligence, science and developing countries

It is unclear thus far what the effects of AI will be in developing countries, and whether AI will widen gaps in scientific capabilities between rich and poor countries. However, researchers in Europe, North America and China clearly dominate research on AI, and the use of AI in science. In 2020, East Asia and the Pacific accounted for 27% of all conference publications, North America 22%, and Europe and Central Asia 19%. By contrast, sub-Saharan Africa accounted for just 0.03% of conference publications (Zhang et al., 2021). As noted in a number of essays in this volume, the computational resources required for cutting-edge AI research favour well-resourced universities, large tech companies and rich countries more generally. The following essays explore remedial initiatives.

Artificial intelligence and development projects

John Shawe-Taylor and Davor Orlič draw on lessons from emerging networks of excellence in developing countries, particularly AI4D Africa. Established in 2019 with financial support from Canada's International Development Research Centre, AI4D Africa helped build capacity in a network of institutions and individuals working on and researching AI from across sub-Saharan Africa.

A significant AI community has grown up in Africa in recent years, with initiatives such as Deep Learning Indaba2022 and Data Science Africa (DSA, 2022). Among other actions, these self-mobilising expert communities have introduced funding for a range of micro-scale research projects. The authors show how such a bottom-up approach with small-scale investments has resulted in significant research on different scientific, non-scientific, engineering and educational topics, including a profile of African languages. Among others, a call for micro-projects helped create the first African Grand Challenge in AI. It focused on curing leishmaniasis, a neglected disease that affects the region. Projects have had budgets in the range of USD 5 000-8 000 each.

Building on the experience of initiatives in developed countries, such as the PASCAL networks of excellence, the authors note that co-ordinating micro-projects as part of a larger coherent programme might deliver still greater benefits. The PASCAL networks used a bottom-up and small-scale agile funding structure built around a co-ordinated research and collaborative theme of pattern analysis and ML. Shawe-Taylor and Orlič conclude that, on first impression, independently of the funding mechanism, there is a case for sub-Saharan Africa to receive much greater funding for AI in science.

Artificial intelligence for science in Africa

Gregg Barrett observes that greater use of AI in research in Africa will deepen African science, broaden global research agendas, incentivise the location of corporate R&D labs and, indirectly, help upgrade the capabilities of civil society.

Barrett points out that while world-class research does take place at African institutions, African researchers lack the computing infrastructure and engineering resources to develop and apply the more powerful and critical AI methods.

New capabilities are needed in most of Africa involving engineering personnel to prepare data, and configure hardware, software and ML algorithms. In addition, the ad hoc mix of campus computers and commercial clouds that Africa's educators and researchers rely on today are inadequate. Simply providing underserved academic and research organisations with the data, hardware, software and engineering resources is also insufficient. To truly reduce barriers to AI-enhanced research, underserved institutions need access to experts who can implement best practices in approaching problems, in methods of learning, selection of tools for tasks and optimisation of workflows.

Based out of Wits University in Johannesburg, South Africa, Cirrus and the AI Africa Consortium aim to respond to the AI deficit in African science. Cirrus is designed to provide data, dedicated compute infrastructure and engineering resources at no cost to academic and research institutions through the AI Africa Consortium. Providing a data management platform is a priority for Cirrus. Such a platform will enable users to store, manage, share and find data with which to develop AI systems. A high priority must be to identify and use existing and potential scientific programmes to produce AI-ready data repositories.

The Africa AI Consortium fosters collaboration agreements with parties across the African R&D ecosystem. Over five years, the legal groundwork has been laid to operationalise Cirrus and the AI Africa Consortium. Some activities have already begun, including the rollout of ML for embedded devices.

Artificial intelligence, developing-country science and bilateral co-operation

Peter Martey Addo considers how bilateral and multilateral development co-operation could help address AI deficits in low-income countries, specifically in relation to science, and suggests a series of practical measures and goals (Box 24).

Box 24. Bilateral and multilateral co-operation to strengthen AI in developing-country science: Suggestions for policy

Strengthening AI readiness: Development co-operation can help countries advance data protection legislation, improve data infrastructures and strengthen AI readiness overall. An example is the collaboration between The GovLab (an action research centre based at New York University's Tandon School of Engineering) and the Agence Française de Développement (French Development Agency, or AFD). Together, they launched the recent #Data4COVID19 Africa Challenge. This supported Africa-based organisations to use innovative data sources to respond to the COVID-19 pandemic.

Fostering collaboration: Bilateral co-operation can also help plan, finance and assist implementation of research and technological development in an environment favouring multidisciplinary and multi-stakeholder collaboration. For instance, in 2021, France's Agence Nationale de la Recherche, in partnership with the AFD, launched the IA-Biodiv Challenge, aimed at supporting AI-driven research in biodiversity (AFD, n.d.) This initiative helps scientists working on AI and biodiversity in France and Africa to mutually learn, share and engage.

Supporting open science, centres of excellence and networking: Development co-operation can go beyond sharing data to supporting open science initiatives. In addition, grants could support investments in AI R&D in developing countries. This could include the creation and support for centres of research excellence like the African Research Centre on Artificial Intelligence in the Democratic Republic of Congo.

Supporting private-public collaborations: Stakeholders in developing countries could also consider formulating research questions relevant to local priorities and amenable to analysis using AI. The 100 Questions Initiative, launched by the GovLab, could provide inspiration (The 100 Questions, n.d.). This initiative seeks to map the world's 100 most pressing, high-impact questions that could be addressed if relevant datasets were available.

Conclusion

This chapter has shown why deepening the use of AI in science matters for raising economic productivity, fostering critical areas of innovation, and addressing global challenges, from climate change to future contagions to the diseases of ageing. Few applications of AI are as socially and economically significant as its use in science. This chapter has also synthesised the main policy messages and insights contained in the essays that follow. AI is pervading research. Recent rapid progress in the capabilities of AI systems is also spurring an outpouring of creative uses in science. However, AI's potential contribution to science is far from realised. Public policy can help to materialise this potential.

References

Aczel, B, B. Szaszi and A.O. Holcombe (2021), "A billion-dollar donation: Estimating the cost of researchers' time spent on peer review", *Research Integrity and Peer Review*, Vol. 6/14, https://doi.org/10.1186/s41073-021-00118-2.

AFD (n.d.), "IA-Biodiv Challenge: Research in Artificial Intelligence in the Field of Diversity", webpage, www.afd.fr/en/actualites/agenda/ia-biodiv-challenge-research-artificial-intelligence-field-biodiversity-information-sessions (accessed 24 January 2023).

Arora, A. et al. (2019), "The changing structure of American innovation: Some cautionary remarks for economics growth", in *Innovation Policy and the Economy*, Lerner, J. and S. Stern (eds.), Vol. 20, University of Chicago Press.

Bhattacharya, J. and M. Packalen (2020), "Stagnation and scientific incentives", *Working Paper*, No. 26752, National Bureau of Economic Research, Cambridge, MA, www.nber.org/papers/w26752.

Bloom, N. et al. (2020), "Are ideas getting harder to find?", *American Economic Review*, Vol. 110/4, pp. 1104-1144, https://doi.org/10.1257/aer.20180338.

Checco, A. et al. (2021), "AI-assisted peer review", *Humanities and Social Sciences Communications*, Vol. 8/25, https://doi.org/10.1057/s41599-020-00703-8.

Chu, Johan S.G. and J.A. Evans (2021), "Slowed canonical progress in large fields of science", *PNAS*, 12 October, Vol. 118/41, e2021636118, https://doi.org/10.1073/pnas.2021636118.

Correa-Baena, J-P. et al. (2018), "Accelerating materials development via automation, machine learning, and high performance computing", *Joule* Vol. 2, pp. 1410-1420, https://doi.org/10.1016/j.joule.2018.05.009.

DOE (2020), *AI for Science, Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*, US Department of Energy, Office of Science, Argonne National Laboratory, Lemont, https://publications.anl.gov/anlpubs/2020/03/158802.pdf.

DSA (2022), "African AI Research Award 2022", webpage, www.datascienceafrica.org (accessed 11 September 2022).

EC (n.d.), "European Open Science Cloud", webpage, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en (accessed 12 January 2023).

EC (2022), "European Health Data Space", webpage, https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (accessed 25 November 2022).

European Physical Society (2019), "The importance of physics to the economies of Europe", *European Physical Society*, eps_pp_physics_ecov5_full.pdf (ymaws.com).

Glass, B. (1971), "Science: Endless horizons or golden age?", *Science*, 8 Jan, Vol. 171/3966, pp. 23-29, https://doi.org/10.1126/science.171.3966.23.

Grizou, J. et al. (2020), "A curious formulation robot enables the discovery of a novel protocell behavior", *Science Advances*, 31 Jan, Vol. 6/5, https://doi.org/10.1126/sciadv.aay4237.

Herbert, D.L, A.G. Barnett and N. Graves (2013), "Australia's grant system wastes time", *Nature*, Vol. 495, 21 March, *Nature Research*, Springer, pp. 314, www.nature.com/articles/495314d.

IMF (2021), "World Economic Outlook: Recovery during a pandemic", International Monetary Fund, Washington, DC, www.imf.org/en/Publications/WEO/Issues/2021/10/12/world-economic-outlook-october-2021.

King, R.D. et al. (2009), "The automation of science", *Science*, Vol. 324/5923, pp. 85-89, https://doi.org/10.1126/science.1165620.

Klinger, J. et al. (2020), "A narrowing of AI research?", *arXiv*, preprint arXiv:2009.10385, https://doi.org/10.48550/arXiv.2009.10385.

NAIRR (2022), "National AI Research Resource (NAIRR) Task Force", webpage, www.nsf.gov/cise/national-ai.jsp (accessed 23 November 2022).

Miyagawa, T. and T. Ishikawa (2019), "On the decline of R&D efficiency", *Discussion Paper*, No. 19052, Research Institute of Economy, Trade and Industry, Tokyo, https://ideas.repec.org/p/eti/dpaper/19052.html.

Noorden, R.V. (5 February 2014), "Scientists may be reaching a peak in reading habits", Nature News blog, www.nature.com/news/scientists-may-be-reaching-a-peak-in-reading-habits-1.14658.

OECD (2021), *Recommendation of the Council concerning Access to Research Data from Public Funding*, OECD, Paris, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347.

OECD (2020), "Addressing societal challenges using transdisciplinary research", *OECD Science, Technology and Industry Policy Papers*, No. 88, OECD Publishing, Paris, https://doi.org/10.1787/0ca0ca45-en.

Service, R.F. (2019), "AI-driven robots are making new materials, improving solar cells and other technologies", *Science*, December, www.sciencemag.org/news/2019/12/ai-driven-robots-are-making-new-materials-improving-solar-cells-and-other-technologies#.

Thelwall, M. et al. (16 January 2023), "Can artificial intelligence assess the quality of academic journal articles in the next REF?", London School of Economics blog, https://blogs.lse.ac.uk/impactofsocialsciences/2023/01/16/can-artificial-intelligence-assess-the-quality-of-academic-journal-articles-in-the-next-ref/.

The 100 Questions (n.d.), The 100 Questions website, https://the100questions.org (accessed 20 January 2023).

Trammell, P. and A. Korinek (2021), "Economic growth under transformative AI: A guide to the vast range of possibilities for output growth, wages, and the labor share", Center for the Governance of AI, www.governance.ai/research-paper/economic-growth-under-transformative-ai-a-guide-to-the-vast-range-of-possibilities-for-output-growth-wages-and-the-laborshare.

Urbina, F. et al. (2022), "Dual use of artificial-intelligence-powered drug discover", *Nature Machine Intelligence* Vol. 4, pp. 189-191, https://doi.org/10.1038/s42256-022-00465-9.

Webber, M.E., R.D. Duncan and M.S. Gonzalez (2013), "Four technologies and a conundrum: The glacial pace of energy innovation", *Issues in Science and Technology*, Winter, National Academy of Sciences, National Academy of Engineering, Institute of Medicine, University of Texas at Dallas, www.issues.org/29.2/Webber.html.

van dis, E. et al., "ChatGPT: Five priorities for research", *Nature*, Vol. 614/7947, pp. 224-226, https://doi.org/10.1038/d41586-023-00288-7.

Wu, L., D. Wang and J.A. Evans (2019), "Large teams develop and small teams disrupt science and technology", *Nature*, Vol. 566, pp. 378-382, https://doi.org/10.1038/s41586-019-0941-9.

Zhang, D. et al. (2021), *The AI Index 2021 Annual Report*, AI Index Steering Committee, Human-Centred AI Institute, Stanford University, Stanford, https://aiindex.stanford.edu/report.