Artificial Intelligence in Science

CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH

OECD

Artificial Intelligence in Science

CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH

OECD
BETTER POLICIES FOR BETTER LIVES

The Executive Summary and Chapter entitled “Artificial intelligence in science: Overview and policy proposals” were approved by the Committee on Scientific and Technological Policy at its 122nd Session on 22-24 March 2023 and prepared for publication by the OECD Secretariat.

The essays set out in Parts I to IV of this document are under the responsibility of the authors named and the opinions expressed and arguments employed therein are their own. The essays benefited from input and comments from the OECD Secretariat and CSTP delegates. The essays should not be reported as representing the views of the OECD or of its member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Please cite this publication as:
OECD (2023), *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, Paris, https://doi.org/10.1787/a8d820bd-en.

ISBN 978-92-64-44154-5 (print)
ISBN 978-92-64-44621-2 (pdf)
ISBN 978-92-64-92820-6 (HTML)
ISBN 978-92-64-33228-7 (epub)

Photo credits: Cover © PopTika/Shutterstock.com.

Corrigenda to OECD publications may be found on line at: www.oecd.org/about/publishing/corrigenda.htm.

Preface

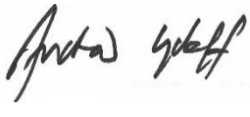
Rarely a week passes without announcements that artificial intelligence (AI) has achieved new capabilities. Since the arrival of generative AI, ChatGPT and subsequent large language models – after many of the contributions to this book were written - discussion of AI's proliferating uses and their implications is increasingly visible in mainstream media. The economic, business, labour market and societal ramifications of AI now occupy the attention of firms, professional bodies, governmental and non-governmental organisations. Indeed, most governments in OECD countries have national AI strategies.

Amid these developments, and except for specialised journals, less consideration has been given to the role of AI in research. This may be inevitable, as science is a specialised field. However, raising the productivity of research may be the most valuable of all the uses of AI. Being able to discover more scientific knowledge, helping science become more efficient, and doing this more quickly, will strengthen the foundations critical to addressing global challenges. Applying AI to research could be as transformative as the rise of systematised and institutionalised research and development in the post-war era. Preparing for new contagions, generating technologies that elevate living standards, countering the diseases of ageing, producing clean energy, creating environmentally benign materials, and other overarching goals, all require technologies and innovations that emerge from science.

In this context, it gives us great pleasure to present this publication, *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*. Gathering the views of leading practitioners and researchers, but written in non-technical language, this publication is addressed to a wide readership, including the public, policymakers, and stakeholders in all parts of science. Among other topics examined are: AI's current, emerging and possible future uses in science, including a number of rarely discussed applications; where progress in AI is needed to better serve science; changes in the productivity of science; and, measures to expedite the uptake of AI in developing-country research.

A distinctive contribution is the book's examination of policies for AI in science. Policymakers and actors across research systems can do much to maximise the society-wide benefits of AI in science, deepening AI's use in science, while also addressing the fast-changing implications of AI for research governance.

This publication is the fruit of a collaboration between our two organisations. The OECD's Directorate for Science, Technology and Innovation undertook the substantive work, under the aegis of its Committee for Scientific and Technological Policy. The publication and the wider project of which it is a part have been made possible thanks to financial and other support from the Fondation IPSEN (https://www.ipsen.com/our-company/ipsen-foundation/), which works to improve living conditions by disseminating scientific knowledge to the public and promoting exchanges within the scientific community.



James A. Levine,

President,

Fondation IPSEN

Andrew Wyckoff,

Director,

OECD Directorate for Science, Technology and Innovation

Foreword

In late 2019 the OECD concluded an agreement with the Fondation IPSEN, which would provide financial support to work on artificial intelligence (AI) and the productivity of science. The context was one in which some scholars had argued that the productivity of science may be stagnating, or even in decline. One aim of the project was to update and significantly expand previous work on AI in science conducted under the aegis of the Committee on Scientific and Technological Policy (CSTP). This prior work included a chapter in the 2018 edition of the OECD *Science, Technology and Innovation Outlook*, titled "Artificial intelligence and machine learning in science". A session on the growing importance of AI in science was also organised on 23 February 2022 the second OECD AI WIPS Conference.

The first output of the project was a workshop – "AI and the Productivity of Science" – held from 29 October to 5 November 2021. The workshop gathered over 80 leading experts to explore topics highlighted in this book. The workshop was filmed and can be viewed here https://www.youtube.com/watch?v=V8ZlGpb0f3c. A project update was discussed at the 120th Session of the CSTP on 6-7 April 2022.

Analysis of numerous issues underpinning a discussion of policies for AI in science necessarily draws on prior CSTP examinations of topics bearing on data-intensive science. These topics include, among others:

- The changing demand for and nature of digital skills in the scientific workforce (see, in particular, the report "Building digital workforce capacity and skills for data-intensive science", https://doi.org/10.1787/e08aa3bb-en).
- Access to public research data (see the *Recommendation of the Council concerning Access to Research Data from Public Funding*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347).

Many of the issues raised in this publication are also relevant to CSTP's current and upcoming work streams, especially in connection with the role of science and technology in sustainable transitions, as well as technology governance, skills, and citizen engagement in science.

Work on AI in science is one among a wide set of AI-related topics being examined by the OECD, overviews of which can be found at the OECD AI Policy Observatory.

Acknowledgements

This publication was edited by Alistair Nolan from the OECD Directorate for Science Technology and Innovation. Alistair Nolan also wrote the opening overview and synthesis of policy recommendations.

Special thanks are due to several experts who provided ideas and advice through much of the process of preparing this book, and for commenting on various of the essays. Among this group are Marjorie Blumenthal, William Clements, Jeremy Frey, Aishik Ghosh, Dominique Guellec, Ross King, Isabelle Ryle, and Hector Zenil.

Valuable comments on parts of the book were had from Jesus Anton, Jonathan Brooks, Alessandra Colecchia, Diogo Machado, Daniel Opalka, Carthage Smith, and Pierre Warnier.

Thanks are due to all the authors of the papers in this publication, who gave freely of their time and insights. Some of the essays also benefitted from the assistance of third parties, as follows.

For helping to develop the Elicit model described in the essay “Elicit: Language models as research tools”, the authors thank Ben Rachbach, Amanda Ngo, Eli Lapland, Justin Reppert, Luke Stebbing, Melissa Samworth, and James Brady.

As concerns the essay “AI in drug discovery”, Kristof Zsolt Szalay extends thanks to Andreas Bender, Krishna Bulusu, Abraham Heifets and Aviad Tsherniak for insights into the state of the field, as well as to Andreas Bender and Daniel V. Veres for expert reviews.

With respect to the essay “Declining R&D efficiency – Evidence from Japan”, Tsutomu Miyagawa expresses gratitude to Takayuki Ishikawa.

With regards to the essay “The end of Moore’s Law? Innovation in computer systems continues at a high pace”, Henry Kressel acknowledges valuable discussions with William Janeway.

Sylvain Fraccola helped generate graphics, and Mark Foss copy edited the entirety of the text, with support from Céline Colombier-Maffre.

Thanks are due to all the participants and contributors to the workshop “AI and the Productivity of Science”, held from 29 October to 5 November 2021. Celia Valeani managed the organisation of that event.

Angela Gosmann, Beatrice Jeffries and Blandine Serve kindly made the text ready for publication.

Lastly, the essay “Quantifying the ‘cognitive extent’ of science and how it has changed over time (and across countries)” was made possible in part thanks to support from the United States Air Force Office of Scientific Research, and a Grant Thornton Fellowship.

Table of contents

Executive summary

Accelerating the productivity of research could be the most economically and socially valuable of all the uses of artificial intelligence (AI). While AI is penetrating all domains and stages of science, its full potential is far from realised. Policy makers and actors across research systems can do much to accelerate and deepen the uptake of AI in science, magnifying its positive contributions to research. This will support the ability of OECD countries to grow, innovate and address global challenges, from climate change to new contagions.

Ambitious multidisciplinary programmes can promote progress

Broad multidisciplinary programmes are needed that bring together computer and other scientists with engineers, statisticians, mathematicians and others to solve challenges using AI. Among other measures, dedicated government funding is required. It needs to be allocated using processes that encourage broad collaboration, rather than siloed funding for individual disciplines. One priority is to **foster interaction between roboticists and domain experts.** Laboratory robots could revolutionise some domains of science, lowering the cost and hugely increasing the pace of experimentation.

Governments can encourage and support visionary initiatives with long-term impact. Initiatives such as the Nobel Turing Challenge – to build autonomous systems capable of world-class research – can inspire collaboration and co-ordination in science, to help focus efforts on global challenges, drive agreement on standards and attract young scientists to such ambitious endeavours.

It is important to **increase access to high-performance computing (HPC) and software for advances in AI and science.** The provision of computing resources by large tech companies is helpful, but this has important gaps, and less well-funded research groups could fall behind. For academics to be competitive using state-of-the-art HPC/AI computing resources from commercial cloud providers is in most cases unrealistically expensive. **National laboratories and their computing infrastructures, in collaboration with industry and academia, could address the gaps and help to develop training materials for institutions of tertiary education**. Countries at the forefront of the field, including the United States and leaders in the European Union, may also collaborate on policy frameworks to make resources available from a shared pool.

Updating curricula could assist. **For example, using already proven AI-enabled techniques, students could be taught how to search for new hypotheses** in existing scientific literature. The standard biomedical curriculum provides no such training. New integrative PhD programmes and/or industry research programmes based on knowledge synthesis – aided by AI – could also help.

Governments can take steps to **increase the availability of open research data and to harness the power of data across various fields, from health to climate**. Examples include Europe's Health Data Space, and GAIA-X, which aims to build a federated data infrastructure for Europe. **Research centres can be helped to adopt systems such as federated learning that can apply AI to sensitive data held by multiple parties without compromising privacy**. Another challenge is to make laboratory instruments

more interoperable via standardised interfaces. **Governments could bring laboratory users, instrument suppliers and technology developers together** and incentivise them to achieve this goal.

Public R&D can be used to advance the field

Public research and development (R&D) can target areas of research where breakthroughs are needed to deepen AI's uses in science and engineering. Research goals include going beyond current models based on large datasets and high-performance computing, and to find ways to automate the large-scale creation of findable, accessible, interoperable and reusable (FAIR) data. Another target could be to advance AutoML – automating the design of machine-learning models – to help address the scarcity and high cost of AI expertise. Research challenges could be organised around AutoML for science, and research could be funded that involves applying AutoML in AI-driven science.

Support should also 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. Public support is needed to make such platforms easier to use across many scientific fields.

Public R&D could help foster new, interdisciplinary, blue-sky thinking. For instance, natural language processing (NLP) can help to work with the enormous growth of scientific literature. However, current performance claims are overstated. Today's research in NLP also 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 these are at a nascent stage.

Knowledge bases organise the world's knowledge by mapping the connections between different concepts, drawing on information from many sources. **Governments should support an extensive programme to build knowledge bases essential to AI in science, a need that will not be met by the private sector.** Research could work towards creating an open knowledge network to serve as a resource for the whole AI research community. Relatively small amounts of public funding could help bring together AI scientists, scientists from multiple domains and professional societies – along with volunteers – to build the foundations for AI to utilise and communicate professional and commonsense knowledge.

The thematic diversity of research on AI appears to be narrowing and is increasingly driven by the compute- and data-intensive approaches that dominate in large tech companies. **Bolstering public R&D might make the field more diverse and help to grow the talent pool.** Funders could pay special attention to projects that explore new techniques and methods separate from the dominant deep-learning paradigm. Meanwhile, **policy makers could support research to examine and quantify losses of technological resilience, creativity and inclusiveness brought about by a narrowing of AI research and the possible implications of the increasing dominance of industry in AI research.**

Much of AI in science involves teaming with people, but **funders could also help develop specialised tools to enhance collaborative human-AI teams, and to integrate these tools into mainstream science**. Combining the collective intelligence of humans and AI is important, not least because science is now carried out by ever-larger teams and international consortia. Investment in this field of research has lagged other topics in AI.

Among other fields, progress is needed in applying machine learning to medical imaging. Failures during COVID-19 were considerable. As in other uses of machine learning in science, incentives are needed to encourage research on methods with greater validation. Funding should involve more rigorous evaluation practices.

Research governance matters

Policy bodies should systematically evaluate the impacts of AI on everyday scientific practice, including on human-AI teaming, work, career trajectories and training – where important changes could occur. Funding calls could **require such assessments, and funders and policy makers should establish response mechanisms to act on the insights gathered.** Among other measures, funders and policy makers could establish and support new independent fora for ongoing dialogue about the changing nature of scientific work and its impacts on research productivity and culture.

The deployment of large language models (LLMs), such as ChatGPT, demands attention from policy makers as their consequences are currently uncertain. LLMs could lead to more shallow work by making this easier, blur concepts of authorship and ownership, and possibly create inequalities between speakers of high- and low-resource languages. However, LLMs and other forms of AI could also aid governance processes, for instance in supporting peer review – a possibility that requires more study and testing.

Policy should address the potential dangers entailed in dual use of AI-powered drug discovery. Little attention has been paid to the imminent dangers of being able to automate the design, testing and making of extremely lethal molecules (and there will be other dual use research to consider, too). Policy makers and other actors in the research system need to assess which of the possible governance arrangements will best protect the public good.

Policy makers and their staff need more know-how to help decide what sort of technology initiatives to support

Existing social networks and platforms could be used to help spread emerging practices. Social platforms such as Academia.edu and the Loop community could be used as testbeds for experimenting with combined human-AI knowledge discovery, idea generation and synthesis, and for propagating and evolving such approaches as literature-based discovery.

Steps are likewise needed to improve the reproducibility of AI research. Among other actions, public funding agencies can require code, data and metadata to be shared freely with third parties, allowing them to run experiments on their own hardware.

There is a strong case for sub-Saharan Africa, and possibly other developing regions, to receive much greater funding for AI in science. Development co-operation can help countries to advance open science, frame data protection legislation, improve digital infrastructures, strengthen overall AI readiness and support Africa's own emerging initiatives, including indigenous development of data, software and technology. Projects with developing countries for AI in science can be mutually beneficial, and low-cost models of support have been proven. Development co-operation can also help create and support centres of research excellence.

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.

Part I Is science getting harder?

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

M. Clancy, Institute for Progress, United States

Introduction

Some believe that technological progress between the 1970s and 2020s was significantly slower than during the preceding 50 years. In recent work, various well-known economists have drawn on evidence from a broad range of fields – computer chips, agriculture, health, national productivity statistics and firm-level data – to argue that ideas are becoming harder to find. This essay reviews the flurry of work generated in response to those claims. It finds, despite the intensity of the debate, considerable consensus about what the data show. Measured in a particular way, a constant supply of research effort does not lead to a constant proportional increase in various proxies for technological capabilities (e.g. doubling the number of transistors on an integrated circuit every 2 years, or doubling US corn yields every 20 years). Instead, a constant proportional increase in metrics of interest has tended to require an increasing supply of research effort.

The terms of the debate

Bloom et al. (2020) provoked much discussion. At the heart of the dispute is whether their choice of metric for measuring changes in the productivity of research is appropriate. Was the focus on such metrics relevant for assessing whether ideas really are getting "harder to find"? Some argue that so long as a constant supply of research effort – e.g. the same number of researchers working on the same problem area year after year – leads to a constant absolute increase in some technology metric, then ideas are not getting harder to find.

To illustrate the disagreement, suppose US corn yields is used as the technology metric – an example Bloom et al. (2020) does use, as will be discussed shortly. Between 1980 and 2008, US agricultural research remained roughly constant and annual corn yields increased by a fairly consistent 1.5 bushels per acre per year. Since annual corn yields increased from roughly 100 bushels per acre in 1980 to 150 bushels per acre over the time period, a consistent annual increase of 1.5 bushels per acre per year implies the growth rate slowed from 1.5% per year to 1.0% per year (see the author's essay on agricultural productivity in this book).

Bloom et al. (2020) would argue this is evidence that ideas are getting harder to find: constant research effort led to declining growth rates in the technology metric. However, others disagree since constant research effort still led to constant absolute increases in crop yields. The actual calculations are not quite so simple, but this example still illustrates one point of disagreement about how to interpret the data.

For their part, the reason why Bloom and his co-authors frame the question in the way they do derives from long-standing models of economic growth. These models show that constant exponential growth in

gross domestic product (GDP) per capita (i.e. the same percentage growth every year) can be achieved with constant exponential growth in technology (defined and measured in different ways, depending on the theoretical framework – Acemoglu, 2009).

Bloom et al. (2020) make a point often missed by readers unfamiliar with the literature on economic growth. Namely, they do not actually assume that a constant (i.e. unchanging) supply of researchers must deliver constant exponential increases in technological capability. Their paper instead uses a subtly different input measure: the effective number of researchers.

The effective number of researchers is computed by dividing some measure of total research and development (R&D) spending by the wage rate for scientists in the relevant country or industry. This is the number of scientists who could be employed if R&D was spent exclusively on the hiring of scientists. It is not an actual headcount of the number of scientists because R&D spending is also spent on research equipment, materials and other non-labour inputs.

More precisely, according to Bloom and co-authors, if ideas are not getting harder to find, then a constant effective number of researchers should be able to generate exponential increases in technological capability.[1] That said, it is not necessary to adopt Bloom's framing of the question, and many do not. The next section, then, looks at what some of the data reveal.

Empirics of technological progress

To make any headway, some measure of "ideas" is needed. Many papers look at the evolution of various metrics of change related to specific technologies. They then compare them to various measures of the inputs to innovation. This section looks at several potential metrics.

Technology: Doubling integrated circuits

Bloom et al. (2020) look at Moore's law, which holds that the number of transistors that can fit on an integrated circuit doubles approximately every two years. While this doubling has held constant over the last half century, Bloom et al. (2020) point out that research effort devoted to achieving these doublings has grown nearly twentyfold.

Kressel's essay in this book largely agrees that doubling the number of transistors on an integrated circuit has gotten harder. He points out that few expect this to continue at the same pace. However, he disagrees on the relevance of this metric as a measure of technological progress. While shrinking transistors is one way to enhance the performance of integrated circuits, it is not the only way. For example, much research effort is dedicated to improving energy efficiency in integrated circuits. Kressel points to advances on a range of parameters, besides transistor density, to argue that overall progress remains robust.

Agriculture: Growth in agricultural yields

Bloom et al. (2020) also show that growth in the yields of a variety of agricultural crops has only been sustained by a substantial expansion of agricultural R&D effort. In another essay in this book, Clancy reviews complementary evidence that supports the general thrust of this conclusion. More theoretically grounded (but empirically challenging) measures of technological progress in agriculture in the United States indicate a slowdown in the rate of progress that extends beyond yields. It suggests this slowdown cannot be easily pinned on confounding factors such as a changing climate or pest burdens.

Health: Years of life saved

When studying health, Bloom et al. (2020) diverge from their earlier metrics, choosing to measure research effort by the number of scientific articles or clinical trials. They also measure improvements in health outcomes by the number of years of life saved. They again document that a constant supply of research effort results in steadily lower incremental improvements in a selection of key health outcomes. Scannell, in this report, provides a variety of complementary types of evidence. He shows, for instance, that new molecular entities discovered per US dollar has fallen dramatically, as has the financial return on investment in health R&D in general.

Creation of new technologies

Others argue the focus on individual technologies misses the point. Guzey and Rischel (2021), for example, argue that technological progress is mostly about the creation of entirely new technological categories, not just improvement of existing technologies. It may well be that for any given technology, improvements get harder to eke out. However, this is more than counterbalanced by the creation of new technologies. Rather than breeding faster horses, we invent automobiles. Or more broadly, we invent telegrams, telephones and the Internet so that we do not need to travel at all to communicate. If the creation of new technologies is the main way technological progress occurs, then metrics of this progress that help capture the creation of these entirely new technologies should be considered.

Bloom et al. (2020) look at a few alternative metrics that might help answer the concern that technological progress in part involves the generation of entirely new technologies. For example, health outcomes (years of life saved) derive in part from successive waves of new forms of technology: antibiotics, chemotherapy, gene therapy, mRNA vaccines, etc. Each of these medical interventions may have diminishing returns to R&D. However, medical R&D overall could remain just as productive by successively hopping from mature to new and emerging technologies. But here, too, Bloom et al. (2020) show it takes increasing research effort to save a year of life. In this case, effort is measured by the number of clinical trials or biomedical articles.

An even broader measure of the fruits of R&D can be computed for private sector companies. After all, they invest in R&D presumably because they believe it will help their firms become more profitable. Profit is agnostic as to the underlying technology; if firms find it more profitable to invent new categories of technology, then they will seek to do that.

Profit itself can be challenging to measure properly. Therefore, Bloom et al. (2020) examine a host of related metrics: sales, number of employees, sales per employee and market capitalisation. They find here, too, that on average it takes more and more R&D effort by firms to increase any one of these profit proxies by an exponential amount.

Boeing and Hünermund, in this volume, examine the same metrics for the People's Republic of China and Germany. They confirm that Bloom et al. (2020) findings are not limited to the United States; it takes more R&D effort for firms to proportionally increase proxies for profit in these two countries as well.

That said, a lot more than R&D affects profit, which presents a conceptual challenge. A firm entering a larger market, for example, might have higher profits, even though its underlying technologies are the same.

Total factor productivity

An alternative measure of broad technological progress is total factor productivity. The idea here is that, rather than trying to directly observe and measure technology, one should measure economic outputs (for example, total GDP) and inputs (for example, labour and capital), and then examine how the ability to produce more outputs from the same or fewer inputs changes over time. Economists assume that one

driver of an economy or firm being able to squeeze more outputs from each unit of input is technological progress, which can include the sequential creation of new technological categories.

Bloom et al. (2020) do use data on total factor productivity for the US economy, going back to the 1930s, and find that more and more R&D effort is required to keep total factor productivity growing at a constant exponential rate. Miyagawa and Ishikawa (2019) perform a similar exercise over a much shorter time frame (1996-2015) but over a wider set of countries and industries. Again, R&D productivity, framed in the way Bloom et al. (2020) prefer, has generally declined, although with some exceptions.

Total factor productivity has its own problems as a measure. Like profit, it can change for reasons that have little connection to technological progress. Vollrath (2019), for example, decomposes the decline of the growth rate of total factor productivity in the United States since 2000 into several categories with little or no relation to technological progress. These include rising consumer spending on services and declining geographic mobility of the workforce. Likewise, total factor productivity is a statistical estimate of how well inputs can be translated into outputs. It thus requires good data on both inputs and outputs. Since most firms use a complicated mix of inputs and often produce a complicated mix of outputs, measurement is challenging.

Empirics of scientific progress

Another way to tackle the question of research productivity is to look more directly at “ideas” themselves – or at least something closer to ideas than the technologies derived from them. In this connection, another line of work looks at scientific research specifically rather than technologies. Several metrics are noted below.

Number of papers per author

A seemingly natural place to start is with trends in the production of scientific publications. While the number of scientific publications produced each year has grown rapidly over the last century, it turns out this is mostly driven by an equally rapid increase in the number of authors (Wang and Barabási, 2021). The number of papers produced per author has been remarkably stable during the 20th century but has begun to increase slightly since then. However, this does not necessarily imply there has been no slowdown in scientific research productivity as papers vary substantially in their contribution to knowledge. It could be that papers today contribute less than in the past. Indeed, several studies suggest this is the case.

Numbers of exceptional people

Cauwels and Sornette (2020) propose to count exceptional ideas by counting exceptional people, since the discoverers of new ideas tend to be recognised and celebrated. Using the Krebs Encyclopedia of Scientific Principles and Asimov’s Chronology of Science and Discovery, they assemble a count of exceptional scientists in physics and the life sciences going back to 1750. They then look to see if this number has been rising or falling, especially as a share of the total population. They hypothesise that ideas are becoming harder to find if a bigger population is failing to discover new ideas (and thereby generate celebrated scientists). They find the number of exceptional scientists as a share of the population increased from 1750 to roughly 1950 but has been falling since.

There are two concerns with their analysis. First, it may take time for the importance of new ideas to be recognised. Their sources end in 2008 so they will miss any ideas developed before that date but only recognised for their importance afterwards. This creates the appearance of a decline in the number of scientists towards the end of the sample (Cauwels and Sornette attempt to statistically correct for this bias).

Second, science is increasingly a team endeavour (Wuchty, Jones and Uzzi, 2007), rather than an individual pursuit. This complicates efforts to assign individuals to ideas.

Numbers of exceptional ideas and discoveries

Rather than counting exceptional people, it may also be possible to identify trends in the production of exceptional ideas and discoveries. For example, one could examine trends in characteristics of the Nobel Prize for physics, chemistry and medicine. At least in theory, the Nobel Prize recognises the most important discoveries in the respective fields, and has been awarded for long enough to observe long-run trends.

One simple measure of scientific progress is to see the share of awards that go to discoveries described in papers published in the preceding 20 years. Across all fields, this has fallen from an average of nearly 90% prior to the 1970s to closer to 50% today (calculations based on Li et al., 2019). Alternatively, Collison and Nielsen (2018) survey scientists and ask them to select the more important discovery from pairs of randomly selected Nobel Prizes. Since the 1940s, there is no clear evidence that scientists prefer discoveries made in more recent decades.

Numbers of citations

Nobel Prizes have their own idiosyncrasies that make them less than ideal for measuring the rate of progress. Another potential measurement approach, then, is to look at features of scientific publications, such as their citations. Citations also provide a potentially informative window into how the scientific community receives new scientific work. Diverse citation-based metrics suggest a slowdown in scientific progress. Chu and Evans (2021) document that as fields have grown larger, the turnover among top-cited papers has slowed. They argue this slowdown shows the scientific canon is increasingly ossified.

Another indicator focuses on research papers and the share of academic citations that go to recent work. Lariviere et al. (2007) and Cui, Wu and Evans (2022) collectively show a steady decline since the 1960s in the share of citations to recent papers (those published in the preceding five or ten years). Patents also increasingly cite older scientific work (Park et al., 2022). One possible explanation for these trends is that more recent work is proving less useful to today's scientists and inventors compared to earlier work. Finally, Park et al. (2022) show that a citation-based measure of a paper's level of disruption also indicates the typical paper has become steadily less disruptive over time.

Numbers of unique phrases

Alas, citations can also be biased by a range of factors and likely measure scientific impact with (possible substantial) noise. Milojević (see her essay in this book) provides another alternative measure of the rate of progress in science. She counts the number of unique phrases in paper titles as a way of gauging the number of distinct research concepts that a field is investigating. In every year, for each field, she samples the same number of phrases from paper titles. A decline in the number of unique phrases is taken to indicate that more papers are working on the same ideas. This may indicate that progress along any one idea has gotten harder.

For this book, Milojević applied the method to the entire Web of Science database covering 1900 to 2020. It encompasses the literature for science as a whole, as well as individual research fields. She finds the number of unique phrases in constant samples of articles has grown (at variable rates) since 1900. Since 2005, however, this trend has begun to reverse. This 2005-20 period seems to be the longest stretch in history during which the number of ideas explored by science has declined.

Conclusion

To sum up, across a range of approaches and measures, there are few exceptions to the general finding that constant proportional progress in technology requires ever- larger investments of research effort. In several of the (uncommon) exceptions to this rule, the exceptions were present in the past but have been absent for at least a decade.

Does this mean ideas really are literally getting harder to find? Not necessarily, for a few reasons. With regards to the first question, none of the papers referenced here have tried to directly measure "ideas" themselves, only proxies. Indeed, it is not clear how one would begin to count actual ideas in the first place.

It does at least seem clear that exponential growth in most things requires increasing effort directed at improvement. Why this might be the case is also an important question. It may simply be a constraint imposed on research by the way nature works – insights, discoveries and applications yielding a proportional increase simply get harder and harder to achieve.

Conversely, no such constraint on research may exist and the decline in research productivity could be driven by changes in the institutions supporting research. Several possible factors have been suggested on this front. Arora et al. (2019), for example, look at the retreat of the private sector from basic science. Cui, Wu and Evans (2022) study the impact of ageing on the scientific labour force. Meanwhile, Bhattacharya and Packalen (2020) blame the increasing importance of citations as a measure of scientists' output. However, this is not an exhaustive list of posited explanations.

If, for whatever reason, research productivity is falling, has technology stagnated since the 1970s? Again, it does not necessarily follow that technological process slows because of declining R&D productivity. This is because declining research productivity has been at least partially counterbalanced by increased R&D effort. If ideas are getting harder to find, society also seems to be trying harder to find them.

References

Acemoglu, D. (2009), *Introduction to Economic Growth*, Princeton University Press.

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, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3539319.

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.

Cauwels, P. and D. Sornette (2020), "Are 'flow of ideas' and 'research productivity' in secular decline?", *Research Paper*, No. 20-90, Swiss Finance Institute, Zurich, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3716939.

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

Collison, P. and M. Nielsen (2018), "Science is getting less bang for its buck", 16 November, *The Atlantic*, www.theatlantic.com/science/archive/2018/11/diminishing-returns-science/575665.

Cowen, T. (2011), *The Great Stagnation: How America Ate All the Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better*, Dutton, New York.

Cui, H., L. Wu and J.A. Evans (2022), "Aging scientists and slowed advance", *arXiv,* 2202.04044, https://doi.org/10.48550/arXiv.2202.04044.

Gordon, R. (2017), *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*, Princeton University Press.

Guzey, A. and E. Rischel (2021), "Issues with Bloom et al.'s 'Are ideas getting harder to find?' and why total factor productivity should never be used as a measure of innovation", webpage, https://guzey.com/economics/bloom (accessed 28 November 2022).

Larivière, V. et al. (2007), "Long-term patterns in the aging of the scientific literature, 1900–2004", in *Proceedings of ISSI* 2007, Torres-Salinas D. and H.F. Moed (eds.), www.issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2007.

Li, J. et al. (2019), "A dataset of publication records for Nobel Laureates", *Scientific Data*, Vol. 6/33, https://doi.org/10.1038/s41597-019-0033-6.

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.

Park, M. et al. (2022), "The decline of disruptive science and technology", *arXiv*, 2106.11184, https://doi.org/10.48550/arXiv.2106.11184.

Vollrath, D. (2019), *Fully Grown: Why a Stagnant Economy is a Sign of Success*, Chicago University Press.

Wang, D. and A. Barabási (2021), *The Science of Science,* Cambridge University Press, https://doi.org/10.1017/9781108610834.

Wuchty, S., B.F. Jones, and B. Uzzi (2007), "The increasing dominance of teams in production of knowledge", *Science, Vol.* 316/5827, pp. 1036-1039, www.science.org/doi/10.1126/science.1136099.

Note

[1] Being precise about these concepts matters. Based on economic theories, measuring research effort in terms of the effective number of researchers is meant to recognise that a larger economy can do more things than a smaller one. To illustrate the intuition, suppose (inflation-adjusted) annual salaries for US scientists in 1950 are USD 20 000 and USD 80 000 in 2000. If R&D spending in a given industry is USD 20 million in 1950 and USD 80 million in 2000, then in each year the effective number of scientists is 1 000, which is obtained by dividing R&D spending by annual salaries. However, suppose only half of R&D spending is actually spent on labour in each year. Then the number of research scientists hired by this industry is 500 in both 1950 and 2000. Suppose the other half of R&D spending is spent on non-labour research inputs. Even though the effective and actual number of scientists is the same in 1950 and 2000, USD 10 million is spent in 1950 on non-labour research inputs, and USD 40 million is spent on non-labour inputs in 2000. In other words, the scientists working in the year 2000 have access to many more non-labour research resources than the ones working in 1950. Measuring research in terms of the number of effective researchers (1 000 in each year, in this example) is meant to reflect that economic growth should allow scientists to bring more resources to bear on problems over time. The key point of Bloom et al. (2020) is to document that the increased research productivity of scientists over time, as they work in an increasingly advanced economy, is not sufficient to maintain a proportional growth rate in various technology metrics.

The end of Moore's Law? Innovation in computer systems continues at a high pace

H. Kressel, Warburg Pincus, United States

Introduction

Are ideas to improve computing systems getting harder to find? A key metric tracking the dramatic progress of electronic technology (denoted Moore's Law) suggests that progress is nearing its end. This has stimulated fears of a serious decline in the pace of innovation in electronics. Such a decline could have many ramifications; microelectronics are central to practically all industrial products and systems – from kitchen appliances to power generators. However, while the ability to shrink transistors is reaching physical limits, fears of stagnation or decline in the power of computing systems are premature. As this essay discusses, other innovations – additional to those tracked by Moore's Law – continue to improve the productivity and processing power of computing systems.

Measuring and predicting the progress of a technology-driven field with a single metric can generate inadequate results. Eventually, technologies that advance along established lines of innovation reach a point of diminishing returns. However, rates of progress can then increase because of unanticipated innovations that shift established sources of development. Technical developments in electronics are a great example of this (Kressel and Lento, 2007).

While Moore's Law has been useful, no reliable and general metric of progress is available because computing systems range so greatly in scale and functionality.

A short description of computing systems

A short review of what makes computing systems function is needed to understand the shifting sources of innovation in electronics technology. The transistor has enabled the digital electronic world. Invented in 1946, it has developed in current amplifiers, storage elements and switches. Interconnected tiny transistors form the integrated circuit chips that underpin computing systems.

The computing power of a system is a function of the available transistor capacity, the speed of transistor switching (i.e. the opening and closing of a circuit), memory volume and interconnection speed. A smartphone today has more computing capacity than a typical mainframe computer in the 1960s.

Moore's Law tracks the progress of the transistors in integrated circuit chips (also called "integrated circuits"). Such chips perform different functions in computing systems – either signal processing, data storage or combinations of both. The first integrated circuits, built in 1962 at Intel, incorporated only a few transistors. Today, as many as 16 billion transistors can be interconnected on a chip not much larger than

a thumbnail. Minimum transistor feature sizes on chips have declined from about 30 microns to about 5 nanometres (for reference, a human hair is around 100 microns thick). In 1960, a single transistor sold for USD 1.00.

All of the above developments mean that computing capacity per unit cost has increased enormously. Today, an integrated circuit chip with 1 billion transistors costs under USD 3.00. Technological wonders had to be invented to achieve such reductions in scale and cost, all while maintaining extraordinary reliability.

The slowing of Moore's Law

Moore's Law – an observation rather than a physical law – has held for about five decades. It posits that transistor chip density doubles roughly every two years, with a corresponding decline in unit transistor cost. With declining size, transistor switching speed also improves as power dissipation declines.

As discussed below, progress in increasing transistor density in keeping with the historical pattern is reaching an end. Hence, improvements in electronic systems driven solely by shrinking transistors are at risk. This has led to fears of an end to innovation in computing systems.

However, other innovations continue to improve the economic and technical performance of electronic systems. Good ideas are not running out. Nor is there evidence of declining interest in such research. The key technical issues – much simplified – are summarised in what follows.

Physical limits impact transistor scaling due to the relationship of gate width, transistor performance and photolithography. Gates control the flow of current in a circuit. As active gate width shrinks, it becomes a challenge to maintain transistor performance. It is also challenging to create films of materials at very small dimensions on silicon (using a process known as photolithography).

Below a certain gate width, which now approaches near atomic dimensions, the switching properties of the conventional transistor structure deteriorate. Furthermore, as the problem of patterning becomes increasingly severe, special ultraviolet laser light sources, multiple exposures and extraordinary control of the photolithographic equipment are required. Together, they greatly raise cost.

As a result, unit production costs for transistors start to rise, as contrasted with the decline described as Moore's Law. Accordingly, today, the most advanced, smallest-feature chips are justified economically only by applications requiring the highest logic and memory performance. This limits design decisions to products expected to sell in high volumes, such as smartphones or large cloud computing systems.

It is important to note that chips represent a declining share of overall system cost (except for several memory chips). This is because the costs of software and peripheral hardware, but particularly software, are rising as a share of system cost. The functional performance of the chips is a more important consideration; in this, there is great progress.

Where innovations are driving progress in electronic systems

Improving chip performance by shrinking the size of transistors is clearly reaching an end. However, electronic system performance is improving due to a number of innovative approaches.

Three-dimensional architecture

New three-dimensional structures are extending transistor performance, while shrinking some of its dimensions. These three-dimensional architectures make better use of the chip area. This is accomplished as follows. All chips incorporate dozens of thin layers (stacks) of different materials. Tailoring the layers in

new ways can increase switching speed and lower power dissipation. Furthermore, within the stacks, chips can include sophisticated interconnections of memory and logic elements to increase interconnection speed. Industry sources believe this approach will double the operating chip transistor density over the next decade.

However, progress will be costly because new vertical architectures are hard to manufacture. New state-of-the-art chip production facilities cost more than USD 10 billion each and require a highly skilled workforce. Only a few plants in the world can produce such chips in volume (current manufacturers include TSMC, Samsung and Intel, Micron, SK Hynix and Western Digital).

Integration

Another area of innovation, which can also reduce costs, is in packaging chips to bring logic processing functions, memory and external communications closer together. Packages have been developed where optical fibre and lasers are close to the chip output. In this way, self-contained sub-systems can be integrated economically into a larger system. For example, a new company, AyerLabs, has developed modules that replace internal copper interconnections with optical links. These reduce power dissipation and allow faster inter-chip communications. Novel packaging technology makes it cheaper to build systems compared to assembling individual chips on a circuit board.

These innovations aim to continue to reduce computing costs as data volumes mushroom. In addition, cloud computing is enabling ever-more powerful computing power at reduced cost. The emergence of cloud-centralised computing capacities supports massive computing needs in a cost-effective manner and on an as-needed basis. With cloud technology from Google, Amazon and other providers, enterprises can muster the computing power needed from organisations that have built (and continue to build) large-scale, cost-effective computing centres.

Finally, it is highly likely that quantum computing systems using technologies different from classical computers will one day become practical for large-scale computing systems. This will raise computing capabilities to new heights of performance. Worldwide research is attempting to solve great engineering challenges facing developers of quantum computers. Laboratories in the United States and elsewhere are making good progress and providing practical demonstrations with small systems.

Alternative metrics to changes in transistor density

Given the diversity in types, scales and functionalities of electronic systems, no reliable general metrics of progress in computing systems (or even integrated circuits) exist yet. However, various attempts have been made to find metrics. In *IEEE Spectrum*, Moore (2020) described approaches to develop a useful metric for monitoring progress in the field. These approaches combined measures of progress in changing chip parameters such as interconnections. However, these attempts are hindered by changes in system performance occurring along many parameters. The development of metrics that define performance of specific systems, such as smartphones or cloud computing services, is more likely.

The ongoing debate about the value of imperfect metrics compared to none at all is healthy and invigorating. A field that appears to have reached its limits will not attract the best students. Moreover, the question of metrics may attract great students to a career in microelectronics. It is important, then, that the reality of continued progress, and the corresponding opportunities, is widely understood.

Which parts of the world will deliver the needed innovations?

All major chip manufacturers invest substantially in advanced product development. However, the fruits of this work are not always published. Published results tend to come from academic institutions and government-funded laboratories.

In the past, when the United States dominated semiconductor manufacture, it produced most semiconductor process innovations. Intel was the clear leader. Extensive academic research in places like the University of California, funded by government, was an important source of innovations that found their way into industry. However, offshoring of the semiconductor industry has changed the innovation landscape, as Korea and Chinese Tapei have developed companies with equivalent or possibly superior capabilities. Meanwhile, the People's Republic of China has also made significant advances (Badaroglu and Gargini, 2021).

In addition, as previously noted, no innovation described in this paper can reach the market without production equipment of increasing sophistication and cost. Only a few companies in Europe or North America are left in this market. Applied Materials, in the United States, is a clear leader in process equipment, along with Lam Research. In advanced photolithography, a Dutch company, ASML, has a near monopoly. These companies maintain high levels of research to maintain their positions and profitability.

Conclusion

For 50 years, the world has benefited from an extraordinary level of innovation in electronics. The ability to scale and manufacture transistors at ever-decreasing unit cost has been a key enabler. However, falling unit cost turns out to have been a simplistic measure of innovation in this field. Industry sources predict a doubling of chip transistor density over the next ten years, not the two years described by Moore's Law. This does not mean the end of innovation in electronic systems based on semiconductors.

There are many creative ideas for development. As one reason for optimism, innovations have moved from a focus on chip transistor density to the elimination of bottlenecks in system performance. To that end, they have reduced "parasitic losses" (e.g. those related to peripheral capacitance), and decreased inter-system signal delays that reduce processing speed. Furthermore, new transistor and chip architectures extend switching performance limits as device dimensions get closer to the atomic scale.

References

Badaroglu, A. and P.A. Gargini (2021), "System and high volume manufacturing driven more Moore scaling roadmap", *IEEE Electron Society Newsletter*, January, Vol. 28, pp. 1-9, www.ieee.org/ns/periodicals/EDS/EDS-JANUARY-2021-HTML-V5/InnerFiles/LandPage.html.

Bloom, N. et al. (2020), "Are ideas getting harder to find," *American Economic Review*, Vol. 110/4, April, pp. 1104-44, www.aeaweb.org/articles?id=10.1257/aer.20180338.

Kressel, H. and T.V. Lento (2007), *Competing for the Future: How Digital Innovations are Changing the World*, Cambridge University Press, Cambridge.

Moore, S.K. (2020), "A better way to measure progress in semiconductors", *IEEE Spectrum*, 21 July, https://spectrum-ieee.org/a-better-way-to-measure-progress in semiconductors.

Is technological progress in US agriculture slowing?

M. Clancy, Institute for Progress, United States

Introduction

This essay reviews evidence that technological progress in US agriculture – here understood as the increasing efficiency by which inputs are transformed into outputs – is slowing, at least relative to the common benchmark of constant exponential growth. The case for a slowdown seems to hold whether measured with yields or more sophisticated methods, such as total factor productivity (TFP). The slowdown may stem from agriculture-specific factors, such as stagnating levels of research and development (R&D) through much of the late 20th century. It may also be influenced by broader factors, such as slowing technological progress in other domains and a general tendency for innovation to get harder.

Four reasons to focus on US agriculture

Why focus on US agriculture? Surprisingly, perhaps, agriculture is a good subject for the study of long-run changes in technological progress. Four reasons are elaborated below.

First, any study of progress needs data over time. US agriculture is unique in providing quite good data over a long period.

Second, the United States is traditionally seen as operating on the technological frontier. US public funding for agricultural R&D is by far the largest in the world, accounting for 25% of publicly funded agricultural R&D in high-income countries (Heisey and Fuglie, 2018).

Third, economists usually think of “technology” as the processes that convert inputs into outputs. In agriculture, the nature of an output has not changed much, at least compared to sectors such as communication, transportation and manufacturing. Corn is corn and the way it was measured 150 years ago is not that different from today. For something like automobiles, where accounting for the changing nature of goods produced is important, counting the number produced in a year might be controversial. However, comparing the annual number of corn bushels produced in 1920 and 2020 is not controversial in the same way.

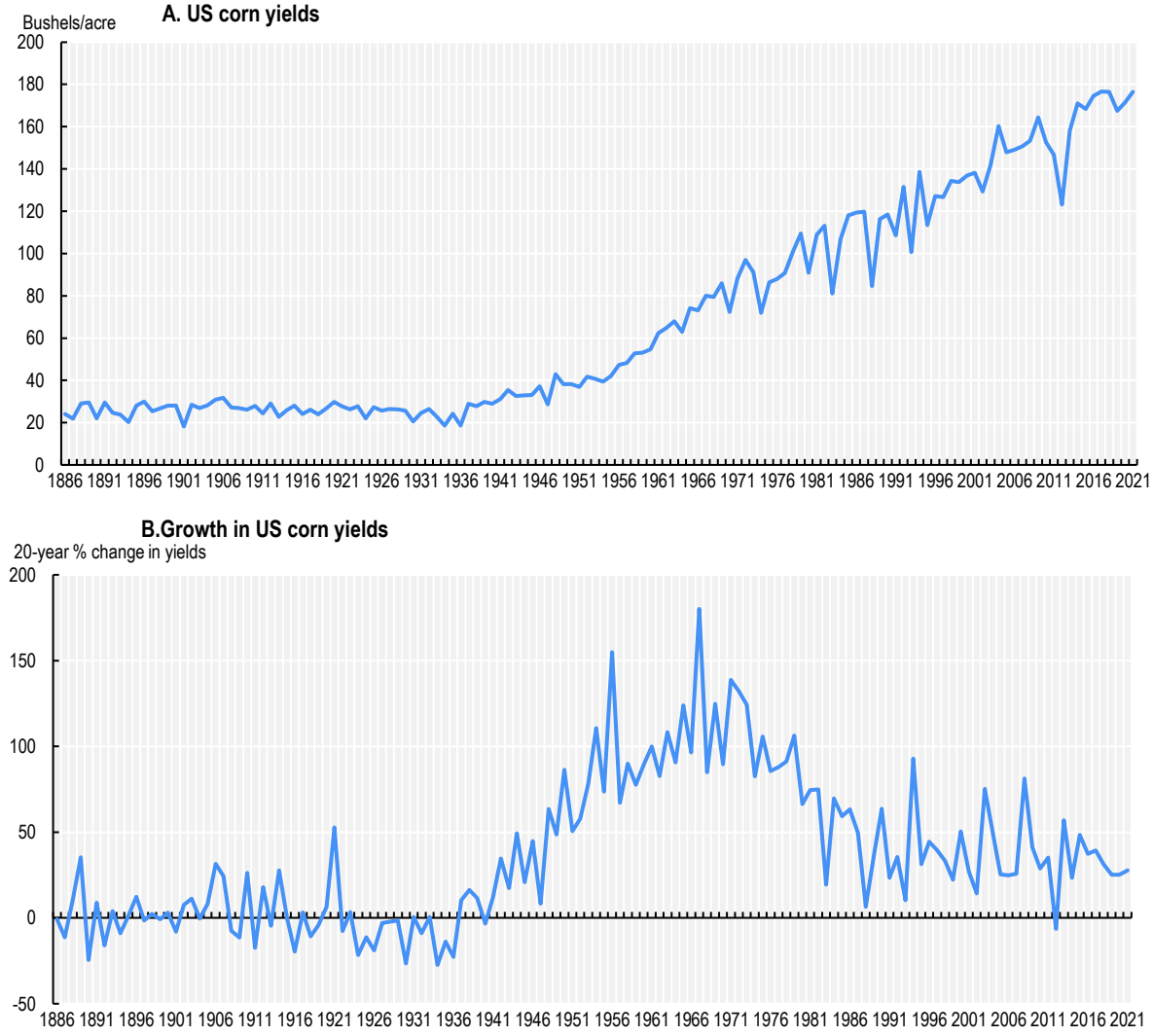
Fourth, agriculture is full of technological progress. Two of the most important inputs to agricultural production, over the long run, have been labour and land. The increase of annual corn production in the United States by more than sevenfold between 1948 and 2022 was accomplished without any significant increase in these two inputs. Land use has stayed roughly the same, while labour has fallen dramatically.

One can already derive a crude but common measure of technological progress in agriculture using these data on output and land, namely “yield” (e.g. bushels of corn per acre). In economics, “constant technological progress” is typically defined to mean constant exponential growth in the efficiency with which

inputs are converted into outputs – e.g. yields increasing by 2% per year. Since agricultural output fluctuates a lot due to weather, this essay uses a long-run indicator of progress – the percentage change in a measure of technological efficiency over 20 years.

Figure 1 below plots average US corn yields on the left and the 20-year growth rate of those yields on the right. It shows a dramatic slowdown in the rate of technological progress by this measure (though progress in yield growth today remains significantly higher than the stagnation that prevailed prior to 1940).

Figure 1. Trends in corn yields in the United States, 1886-2001



Source: Author's calculations based on data from USDA NASS Quickstats.

Yield as a measure of technological progress

Measuring technological progress in agriculture with yields has the advantage of not depending much on theoretical constructs. This simple set of data shows a sharp slowdown in yield growth. This is mostly (but not entirely) because growth has been constant in absolute terms (growing by about 38.5 bushels every 20 years between 1970 and 2021), and therefore must be decreasing in exponential terms.

However, yield is also an unsatisfying measure of technological progress because it misses so many aspects of agricultural production. A way is needed to account for the varied agricultural products (not just corn); the diversity of inputs used; and other factors that affect agricultural production and that might have changed, such as the climate.

Ciliberto, Moschini and Perry (2019) nicely illustrate some of the ways yield is an unsatisfying measure of technological progress. A prominent technological innovation in US agriculture has been the genetic modification of crops. For example, in 2014, nearly 90% of US corn was genetically modified with a gene that confers resistance to the chemical glyphosate, a key pesticide. This modification makes it easier and less costly to control weeds. Another common genetic modification confers resistance to various species of corn rootworms, which reduces the need for insecticides. Both innovations are only indirectly connected to yield but are highly valued by farmers. By comparing demand for these seeds at various prices, relative to comparable seeds without genetic modification, Ciliberto, Moschini and Perry (2019) estimate farmers are willing to pay an extra USD 5-17 per acre for one of these traits.

To capture these kinds of improvements, a measure of technological progress needs to be created. It must account for progress that keeps yield the same but reduces farm labour (for example, related to weed control), or use of other inputs (such as insecticides). Moreover, to study technological progress in the agricultural sector as a whole literally requires a way to compare apples and oranges and everything else farmers grow. Fortunately, economists have many theories for aggregating baskets of goods over time by using data on spending and price changes. Using these techniques, from 1949 to 2017 (the last year for available data), growth in total agricultural output in the United States did not look that different from the trends seen in corn production: total US agricultural output, in inflation-adjusted terms, nearly tripled.

Calculating what happened on the input side is trickier. Technological progress in agriculture has involved waves of new technologies, which are gradually adopted by a larger and larger share of farmers (Pardey and Alston, 2021). It is a lot harder to measure these new kinds of inputs because they come in such a variety of forms (fertiliser, pesticides, tractors, silos, etc.).

Moreover, the quality of these inputs evolves over time due to technological progress. For example, one cannot just count the gallons of pesticides used over time since the nature of that pesticide changes. Instead, economists at the US Department of Agriculture (USDA) attempt to adjust for the changing quality of pesticides to measure the farm sector's use of some kind of "constant-quality" pesticide. Similar adjustments are made for the other inputs used in agricultural production.

Those inputs are then aggregated in a way that weights their share of the value of intermediate inputs. This reveals at least two places where agriculture has increased, rather than decreased, its use of inputs. According to USDA measures, quality-adjusted fertiliser use more than tripled between 1948 and 2017, while quality-adjusted pesticide use increased more than fiftyfold over the same period (from a low base).

To some extent, the invention of cost-effective fertiliser and pesticides is itself a story of technological improvement since effective versions of these inputs were themselves inventions. However, taking the existence of fertiliser and pesticide for granted, if more intense use drives the rise in yields noted above, it calls into question the story of technological progress observed so far. Maybe increased output in agriculture arose simply from using more (non-land) inputs, not from any increased ability to get more from less.

It will take several steps to continue measuring technological progress as the ability to translate ever-fewer inputs into ever-more outputs. First, the basket of different inputs, use of which grew in some cases and shrank in others, should be aggregated into a single index of inputs. Second, the basket of different outputs should be aggregated into a single index of outputs.

Dividing an index of all agricultural outputs by an index of all agricultural inputs is analogous to how yield was one output (corn) divided by one input (land). It gives a more comprehensive measure of technological

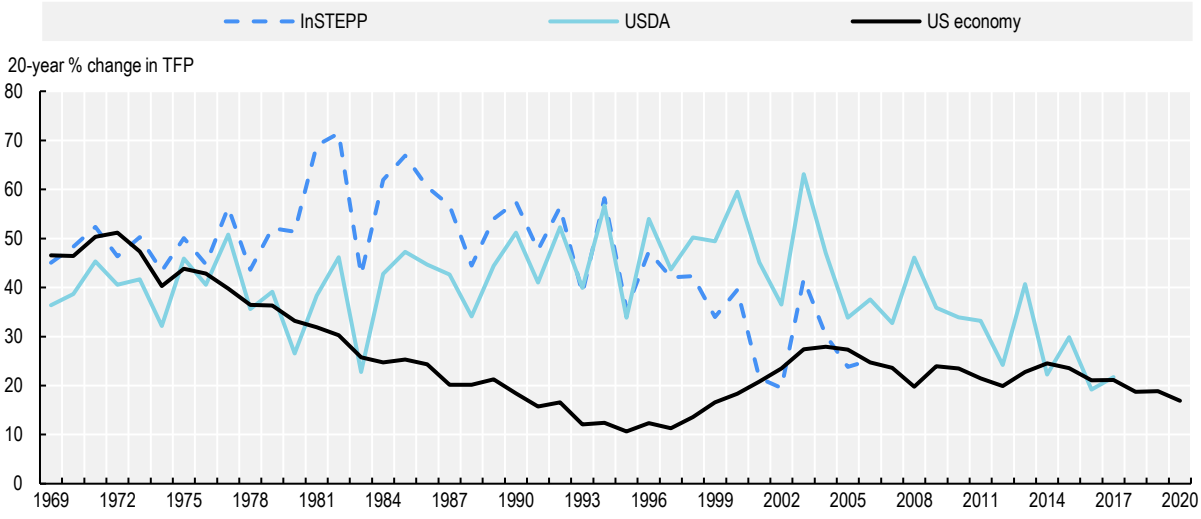
progress – TFP (sometimes called multi-factor productivity), which measures the capacity to produce more with less.

Economic theory provides a framework for constructing these estimates. Given data on all the different inputs used in production, under some standard assumptions, it is possible to weight indices by their share of total costs and add them up to generate an aggregate index measure. Many methodological choices go into constructing these data series, however. Ultimately, there is no simple and objective metric to check that these choices are right. These measures are ultimately theoretical constructs.

However, US agriculture is fortunate in having two different teams of economists – one composed of government economists affiliated with the USDA, the other led by academics affiliated with the International Science and Technology Practice and Policy group (InSTEPP). Together, they have tackled this measurement challenge using somewhat different methodologies (see Fuglie et al., 2017 for a discussion). The extent to which the two different approaches converge on the same findings gives some confidence in the results.

The different estimates can be seen in Figure 2, which plots estimates from InSTEPP and the USDA for the 20-year growth rate of agricultural TFP, and for comparison, the 20-year growth rate of TFP for the entire US economy. Among economists, it is not controversial to assert a slowdown in US TFP growth since the 1970s.

Figure 2. Twenty years of TFP growth in US agriculture and the US economy, 1969-2020



Source: Data are from USDA TFP series and InSTEPP MFP series. US economy is utilisation adjusted TFP from Federal Reserve Bank of San Francisco (2021).

Figure 2 clearly indicates that TFP growth in agriculture has slowed. Notably, the magnitude of the decline in TFP is similar across all three series, though there is some considerable disagreement between InSTEPP and the USDA in the 1980s. However, setting this decade aside, each series hangs around the 40-50% range (total over each preceding 20-year period) in the first half of the series. Each series then ends in the 20-30% range (total over the preceding 20-year period).

There is some significant disagreement about when this slowdown began. InSTEPP showed declines beginning in the 1990s, while USDA placed them in the 2000s. In any event, the two different TFP

estimates suggest a slowdown comparable to the slowdown in the entire US economy in the late 20th century and early 21st but with a later onset.

On the other hand, perhaps using TFP as a measure of technological progress is misleading. Like yields, TFP misses important contemporaneous factors that affected how much output was produced from agricultural inputs. For example, a worsening climate or pest burden could reduce agricultural output from a given set of inputs. This would also reduce measure of TFP. However, in this case, the decline in TFP growth would not be caused by a slowdown in technological progress. As discussed in Clancy (2021), these considerations do not appear to alter the core claim made above that the growth rate of agricultural TFP slowed in the late 20th century and early 21st century.

Another important dimension of agricultural production not typically included in TFP relates to the environmental sustainability of agricultural production. Unsustainable forms of production, which drew down natural resources (for example, soil quality) or produced harmful pollutants, may have shifted towards more sustainable practices. Since TFP does not typically measure use of natural resources and production of pollutants, a move to greater sustainability could result in lower TFP growth. Again, this would not imply any actual reduction in the rate of technological progress.

Expanding the scope of TFP to include these non-marketed inputs and outputs is an active area of research in agricultural economics (e.g. Bureau and Antón, 2022). However, long-run data are needed to determine if these measures also indicate a slowdown in technological progress.

Why has technological progress slowed?

If the available data are correct, and technological progress has indeed slowed, the next question is: why might this have occurred?

To begin with, note the following coincidence: over the entire 20th century the growth rate of agricultural TFP (as estimated by InSTEPP) has followed the same basic trajectory as the growth rate of non-farm TFP. However, it has a multi-decade lag. Pardey and Alston (2021) argue TFP growth in the non-farm sector increased through the 1940s, then declined through 1990. Conversely, agricultural TFP rose unevenly through the 1980s and then fell.

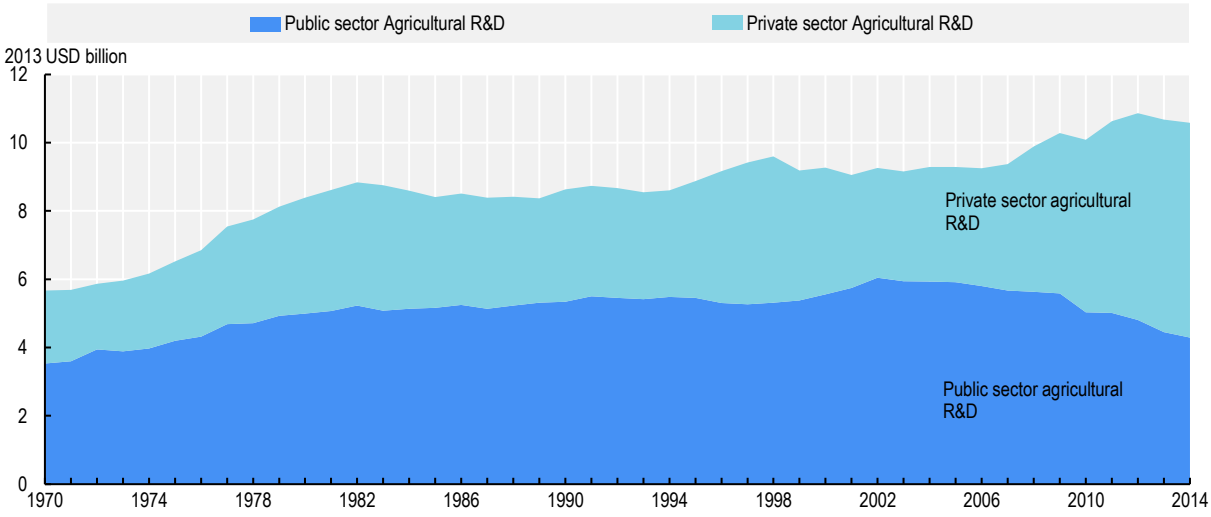
There are a few reasons to think this rise and fall, separated by decades, might not be a coincidence. Agricultural economists generally agree that new and better technologies have been the fundamental driver of US agricultural productivity gains. These technologies themselves emerged from earlier R&D (in some cases, many decades earlier). For reasons described below, in the short term (which can still be pretty long), specifically agricultural R&D might be most useful. Conversely, non-agricultural R&D is likely to be useful in the longer run.

Figure 3 plots US agricultural R&D over roughly the same time period as Figure 2.

At first glance, this does not look much like the TFP growth series. Rather than staying constant and then declining, R&D rises, stagnates and then rises (though its composition changes a lot). However, movements in contemporaneous R&D should not be expected to match those in agricultural TFP growth for two reasons. First, R&D only affects productivity with a lag. Second, falling TFP growth alongside stagnant R&D would be expected if the productivity of R&D is declining over time.

With respect to lags, there is a long literature in agricultural economics that attempts to pin down the correct time lag between R&D and productivity. Researchers look for correlations between R&D spending and productivity, both at the national and state level. Baldos et al. (2019), for example, adopts a Bayesian methodology to explicitly model uncertainty about this process. It finds a peak impact of R&D on productivity around 20 years. This is consistent with much other literature in this field, which tends to find a multi-decade gap between the onset of R&D and its impact on productivity.

Figure 3. US agricultural R&D, 1970-2014



Source: Data are from USDA ERS.

With respect to declining research productivity, R&D effort is defined as current R&D expenditure levels divided by the prevailing wage for scientific labour. Constant levels of R&D effort yield diminishing proportional increases in any given R&D target across many fields. Thus, a time path of agricultural R&D that initially grows, then flattens, then grows again, might well generate a path of TFP growth that is constant, then declines, then remains constant.

Given at least a 20-year lag between R&D and its effects on productivity, one might then predict agricultural TFP growth to remain constant for at least 20 years after agricultural R&D began to stagnate in 1980. Beginning sometime around 2000, agricultural TFP growth might then begin to decline as stagnating agricultural R&D catches up to it with a delay. For the period illustrated in Figure 3, this actually fits the time path for TFP growth (as measured by the USDA). However, the decline in TFP growth as measured by InSTEPP appears to occur too early for changes in the productivity of R&D to be the whole story (though it could certainly exacerbate declines in later years).

Still, agricultural R&D itself builds significantly on R&D elsewhere. Clancy et al. (2021) measure the extent to which patented agricultural technologies rely on knowledge developed outside of agriculture. This includes, for example, farm machinery, fertilisers, pesticides, veterinary medicine, plant varieties and plant breeding techniques. They measure this by looking at the share of citations made by patents for agricultural technology to other patents and academic journals. They also look at the share of novel technological concepts in the text of agricultural patents that originate in other non-agricultural patents. They find, in most cases, that most ideas used in patented agricultural technologies do not originate in agricultural research.

Their findings suggest a model where technological developments in the non-farm economy seed promising avenues for agricultural research to adapt. Indeed, Gordon (2017) argues the surge in TFP growth for the non-farm economy is a crude measure of widespread technological innovations in the US economy. Such innovation begins with the automobile, roads and electrification, and is later followed by a revolution in chemical technology. To a large extent, the story of agricultural productivity in the 20th century is the story of these economy-wide innovations gradually diffusing out from cities into rural America. At that point, they are adapted for agricultural use via agricultural R&D and then diffuse across farms over the course of decades.

However, as Pardey and Alston (2021) show, changes in productivity on the farm only occur after the uptake of these innovations. They argue a process of economic reorganisation is necessary to reap the benefits of these new technologies. For example, if technology disproportionately improves the productivity of larger farms, the full benefits of technology will only be realised after a protracted period of farm consolidation (which is, in fact, observed). Furthermore, as Costinot and Donaldson (2016) show, better infrastructure may allow farms to sell on distant markets. If this happens, farms can specialise in growing crops for which they have a comparative advantage rather than a diverse set of crops to satisfy local market demand. This adjustment also takes time.

Conclusion

It does appear that technological progress in US agriculture has begun to slow, at least compared to a benchmark of constant exponential growth. This is visible with crude but robust measures like the growth in yields over time, but the result also holds true after incorporating changes in the mix and quality of inputs. Moreover, while this essay is focused on US agriculture, a slowdown in agricultural productivity growth is a global phenomenon. ERS/USDA has also produced an internationally comparable TFP database using a simplified methodology applicable across countries.[1]

As noted in Fuglie, Jelliffe, and Morgan (2021), these data show global productivity growth in agriculture fell from an average of 2% per year over the 2000s to 1.3% per year over the 2010s. Developing countries experienced much steeper declines. As in the United States, these declines may stem from slowing technological progress but also from non-technological factors such as climate change.

Moreover, TFP growth in countries further from the technological frontier is more likely to reflect the adoption of frontier technologies and practices, as well as the transition to efficient scales given new technologies, rather than the rate of frontier technological advance per se. Reviewing these factors, however, is beyond the scope of this essay.

In the United States, stagnating agricultural R&D in the late 20th century, in an environment where innovation gets harder, may well explain part of the country's decline in agricultural productivity. However, at a deeper level, it may well be that slowing progress in agriculture is a long-delayed echo of a slowdown in innovation across the wider non-farm economy.

References

Baldos, U.L.C. et al. (2019), "R&D spending, knowledge capital, and agricultural productivity growth: A Bayesian approach", *American Journal of Agricultural Economics,* Vol. 101/1, pp. 291-310, https://doi.org/10.1093/ajae/aay039.

Bureau, J. and J. Antón (2022), "Agricultural Total Factor Productivity and the environment: A guide to emerging best practices in measurement", *OECD Food, Agriculture and Fisheries Papers*, No. 177, OECD Publishing, Paris, https://doi.org/10.1787/6fe2f9e0-en.

Ciliberto, F., G. Moschini and E.D. Perry (2019), "Valuing product innovation: Genetically engineered varieties in US corn and soybeans", *RAND Journal of Economics*, Vol. 50/3, pp. 615-644, https://doi.org/10.1111/1756-2171.12290.

Clancy, M. (2021), "Is technological progress slowing? The case of American agriculture", 24 November, *New Things Under the Sun,* www.newthingsunderthesun.com/pub/0i50ju3x.

Clancy, M. et al. (2021), "The roots of agricultural innovation: Patent evidence of knowledge spillovers", in *Economics of Research and Innovation in Agriculture, P. Moser (ed.) University of Chicago Press.*

Costinot, A. and D. Donaldson (2016), "How large are the gains from economic integration? Theory and evidence from U.S. agriculture, 1880-1997", *Working Paper*, No. 22946, National Bureau of Economic Research, Washington, DC, https://doi.org/10.3386/w22946.

Federal Reserve Bank of San Francisco (2021), "Total Factor Productivity", webpage, www.frbsf.org/economic-research/indicators-data/total-factor-productivity-tfp (accessed 28 November 2022).

Fuglie, K.O. et al. (2017), "Research, productivity, and output growth in U.S. Agriculture", *Journal of Agricultural and Applied Economics*, Vol. 49/4, pp. 514-554, https://doi.org/10.1017/aae.2017.13.

Fuglie, K., J. Jelliffe, and S. Morgan (2021), "Slowing productivity reduces growth in global agricultural output", 28 December, *Amber Waves*, www.ers.usda.gov/amber-waves/2021/december/slowing-productivity-reduces-growth-in-global-agricultural-output.

Gordon, R. (2017), *The Rise and Fall of American Growth*, Princeton University Press.

Heisey, P.W. and K.O. Fuglie (2018), "Agricultural research investment and policy reform in high-income countries"*, Economic Research Report*, No. 249, US Department of Agriculture, Economic Research Service, Washington, DC.

Pardey, P. and J. Alston (2021), "Unpacking the agricultural black box: The rise and fall of American farm productivity growth", *The Journal of Economic History,* Vol. 81/1, pp. 114-155, https://doi.org/10.1017/S0022050720000649.

Note

[1] See ERS/USDA website for further information: www.ers.usda.gov/data-products/international-agricultural-productivity.

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

J.W. Scannell, University of Edinburgh, United Kingdom

Introduction

There is a historical case for describing biomedical innovation from around 1940 to 1970 as a "golden age", which followed the maturation of medicinal chemistry and the application of physiological science to people. Levels of innovation have since fallen for several reasons. Arguably of greatest importance is the progressive accumulation of an excellent and inexpensive pharmacopoeia of generic drugs. When drugs' patents expire, they become much cheaper but no less effective. The ever-expanding catalogue of cheap generic drugs progressively raises the evidential, regulatory and competitive bar for new drugs in the same therapy area, eroding incentives for research and development (R&D). Such therapy areas hold meagre returns for investment in "new ideas", even if the ideas themselves have not become harder to find.

The catalogue of generic medicines, now over 90% of prescriptions in the United States, has therefore squeezed R&D investment towards diseases where R&D has been less successful over the last hundred or so years; diseases that may be pharmacologically intractable and/or hard to model effectively in the laboratory (e.g. advanced Alzheimer's, some metastatic solid cancers, etc.). Again, it is relatively easy to propose therapeutic "ideas" for these diseases. However, the lack of predictive laboratory models and/or the inherent pharmacological intractability creates a very low "innovation yield" from the human trials required to identify the small subset of ideas that are any good.

The fact of the decline in innovative efficiency in the drug industry is relatively uncontroversial. Steward and Wibberley (1980) asked in *Nature,* the leading science journal: "Drug innovation: What's slowing it down?" Two years later in the same journal, Weatherall (1982) speculated on "an end to the search for new drugs". By 1997, rapid progress against AIDS was celebrated as a return to a golden age of innovation that ran through the middle third of the 20th century (Richard and Wurtman, 1997; Le Fanu, 1999). There has been a distinct uptick in some research productivity measures since 2010, whose causes are considered later, but this is modest compared with the prior fall.

The causes of the decline are more obscure than the decline itself. The literature describes a great many possible causes, but they have not been sufficiently prioritised. Widely touted productivity "fixes" have generally failed to change the downward trend. There has also been a notable failure to explain the large divergence in the efficiency trends of R&D inputs and outputs. DNA sequencing, genomics, high-throughput screening, computer-aided drug design, x-ray crystallography and computational chemistry, among other advances, were created and widely adopted, and/or became orders of magnitude cheaper between 1950 and 2010. The efficiency gains resemble, and in the case of DNA sequencing exceed, the performance gains of computer chips described by the now famous Moore's Law. In contrast, the number of new drugs approved by the US Food and Drug Administration (FDA) per billion US dollars of inflation-adjusted industrial R&D investment fell roughly a hundredfold over the same period (Figure 1A). The term

"Eroom's Law" (Eroom is "Moore" backwards) was coined to draw attention to the contrasting trends in input efficiency relative to output efficiency (Scannell et al., 2012).

This essay reviews some of the data that point to a decline in the productivity of biopharmaceutical R&D. It summarises some of the causes. It then draws attention to two important papers (Bender and Cortés-Ciriano, 2021a, 2021b) and some technical blog posts (Lowe, 16 December 2021, 9 December 2021, 8 November 2021, 23 July 2021, 30 November 2020, 25 September 2019). These provide a realistic assessment of the impact of artificial intelligence (AI) on drug discovery, which is likely to be modest in the near term. The essay finishes with some comments on financial incentives for biopharmaceutical innovation. At present, private sector investment in novel chemistry may be over-incentivised. Conversely, investment in scientific tools that help decide whether the novel chemistry is likely to benefit sick people is likely under-incentivised.

Diverse measures of declining R&D productivity

Before looking at the trends, it is worth reflecting on the practical challenges of measuring biopharmaceutical R&D productivity.

The first challenge is deciding what to measure. Any productivity measure divides an output by an input, and there is a wide choice of both. Output choices could concentrate, for example, on any of the following: drug industry profits, the number of new drugs approved, the number of new patients treated by those new drugs, the number of healthy life-years gained by those patients, etc. Input choices could include the amount of money spent on R&D each year, the quantity of labour involved, etc. Productivity measures could also consider either all or parts of the R&D process (e.g. the academic work, the medicinal chemistry or antibody creation, experimental efficacy testing or the clinical trials).

As the next challenge, some of the most appealing output measures, such as the number of healthy life-years gained, are not practically measurable with any degree of precision. New drugs are adopted into changing health systems and their real-world use is optimised over years or even decades. Drugs, diagnosis, surgery and other aspects of patient-management co-evolve.

There are practical problems with the data needed to measure temporal trends. These can be a Frankenstein's monster, crudely stitched together from datasets that have changed over time. An FDA drug approval today is not the same as in the 1960s, but many productivity measures treat it as if it were.

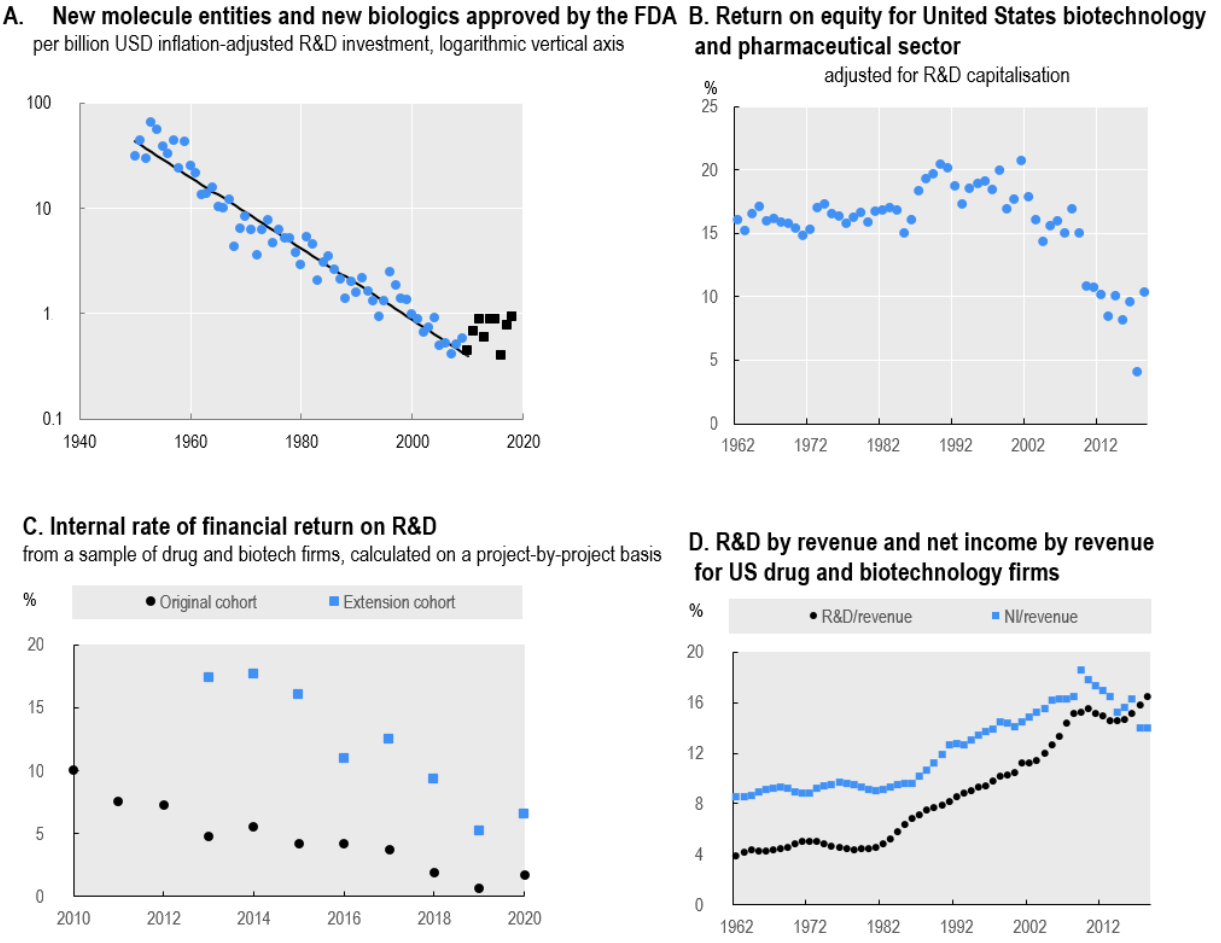
There is the problem of extreme lumpiness, or skew, in the financial and therapeutic value of new drugs. The mRNA-based COVID vaccines have allowed billions of people to return to normal life, will generate hundreds of billions of US dollars of revenue, and have transformed future vaccine innovation. But most new drugs offer marginal clinical gains and generate little revenue. Averages and trends calculated from such skewed data can be misleading.

There are also the problems of survivor bias and mean reversion. Drug R&D has the characteristics of a lottery. Companies attract scrutiny, and corporate outcomes enter analytic datasets; an R&D success is a lottery win. However, performance then tends to revert to the industry average. This means that the productivity of samples of "obvious" companies will tend to decline over time, even if industry productivity does not.

Having said all that, a diverse range of metrics shows a declining R&D productivity trend. Several measures are shown in Figure 1. Panel A shows the number of new drugs – defined as "new molecular entities" and "new biologics" – approved by the world's major drug regulator, the FDA, per billion US dollars of inflation-adjusted R&D in the drug and biotechnology industries. This measure fell roughly a hundredfold between 1950 and 2010 (Scannell et al., 2012). The vertical axis is logarithmic. In other words, a real-terms US dollar of R&D spending in 1950 made a contribution to generating new drugs that was around 100 times

greater than a real-terms US dollar in 2010. The downward trend broke around 2010, with a modest uptick (Ringel et al., 2020). The uptick in drug approvals is, however, associated with a decline in the number of eligible patients per new drug. This is due, in part, to greater focus on rare diseases (Ringel et al., 2020). Note that the absolute number of drug approvals has not declined. Rather, R&D spending per drug has grown. Approvals bounced around at 20-30 per year until around 2010 and have roughly doubled since then.

Figure 1. Trends in selected R&D productivity measures for the drug and biotechnology industries



Notes: **A**. For details of data and methodology, see Scannell et al. (2012) and Ringel et al. (2020). **B.** Calculated from Compustat data that provide nearly comprehensive coverage of US drug and biotechnology companies. Data from different companies were aggregated on a value-weighted basis. Standard methods were used to adjust the accounting data, treating R&D as long-term capital expenditure. **C.** Data in the graph are from Deloitte (2021, 2019). **D.** As with panel B, calculated from Compustat data that provide nearly comprehensive coverage of US drug and biotechnology companies. Data from different companies were aggregated on a value-weighted basis.
Source: Grabowski and Vernon (1990), Damodaran (2007), SSR Health (2014), Scannell et al. (2015, 2012), Deloitte (2021, 2019).

Figures 1B and 1C are financial productivity measures. Investors are typically interested in the unit of profit, per unit of capital employed, per unit of time (Damodaran, 2007). Return on equity (ROE) captures this idea (Figure 1B), and can be easily computed from public accounting data. ROE is annual net profit divided by the average annual equity balance. Equity is a measure of the quantity of the owners' capital tied up in the business. Equity tends to increase the more the owners invest in long-term assets (such as factories and intellectual property), and the less they take out of the business (in terms of dividends or stock buybacks). Financial analysis of R&D-intense industries, such as the drug industry, should treat spending

on multi-year R&D programmes in the same way as long-term capital expenditure is treated in other industries. In this way, R&D investment shows up in the equity balance.[1] With this treatment, the drug industry's ROE becomes a proxy for financial returns on its R&D investment. This is because its profits are largely R&D-dependent and because R&D investment dominates the equity balance. Taking this approach, ROE for publicly traded US drug and biotech firms has fallen since around 2000. Biopharma ROE is now roughly comparable to that of other industries.

Various authors have published an alternative financial measure called the "internal rate of return" (IRR) on R&D investment, also derived from public accounting data. Think of IRR as the aggregate interest rate earned on a series of cashflows, in this case the series of profits from an initial series of R&D investments. For a sample of large US drug and biotech firms, the IRR paints a similar picture to the ROE (Figure 1B) (SSR Health, 2014; Scannell et al., 2015).

A third financial measure, less transparent but closer to the reality of individual project-level R&D investment decisions is an IRR number calculated from companies' own internal project data rather than their published accounts (Grabowski and Vernon, 1990) (Figure 1C). This measure matches project-level R&D spending to the profits that those projects yield (making suitable allocations for the cost of failed projects, etc.). Recent time series (Deloitte, 2021, 2019) show a decline in this R&D productivity measure (Figure 1C).

How did the drug and biotech industries do so well financially for so long (Figure 1B) despite a large fall in innovative efficiency (Figure 1A)? The short answer is that profit growth offset rising R&D costs.[2] However, profit growth could not keep pace with R&D cost growth indefinitely (Figure 1D), and this has depressed financial returns on R&D investment since around 2000. In the early 1960s, the industry's net income was roughly twice its spending on R&D. Today, for the industry as a whole, aggregate R&D spending is higher than net income. Ironically, the "golden age" of pharmaceutical innovation occurred when R&D investment in the pharmaceutical sector was much less intense than today (Figure 1D).

Other published analyses also suggest a decline in the productivity of R&D using a variety of measures. These include SSR Health LLC (2014), Barker and Scannell (2015) and Bloom et al. (2020).

Causes of declining R&D productivity

Good explanations of the productivity decline should be able to account for the large scale of the productivity change and its progressive nature. Two broad classes of mutually non-exclusive explanation are key (Scannell et al., 2012):

The exhaustion of opportunities for pharmaceutical innovation. These opportunities might include as-yet-untreated diseases, unexploited biological mechanisms or unexplored regions of chemical space.

The gradual abandonment of more productive methods of R&D in favour of less productive ones (Horrobin, 2003). For example, many patients were treated as "experimental material" during the 1950s and 1960s. This may have been extremely productive but would cause horror today (Le Fanu, 1999).

These classes of explanation are causally linked. The depletion of certain opportunities has forced the abandonment of more productive R&D methods and the adoption of less productive ones. Consider a therapy area where R&D is particularly successful. The patents on the successful drugs eventually expire, and generic versions become available at a fraction of the price of new branded drugs. Generics often settle at around 10% of the price of the branded versions they replace. Since the 1980s, health systems have increasingly used these cheap generics before using more expensive, newer, patent-protected drugs. The ever-growing catalogue of cheap generic drugs progressively raises the competitive bar for new drug candidates in the same therapy area and so deters R&D investment. In 1994, around one-third of US prescriptions were for generic drugs. Today, over 90% of US prescriptions are for generics. Antibiotics,

nearly all discovered and launched before 1970, are one obvious example. Antidepressants, nearly all discovered and launched by the early 1990s, are another. The improving generic pharmacopoeia is, of course, good for health systems. However, it pushes R&D towards diseases that have proven less tractable over the last 80 years – diseases for which the established R&D methods have proven less productive.

The second of the two main proposed explanations for declining R&D productivity relates to the progressive abandonment of more productive R&D methods (Scannell and Bosley, 2016; Scannell et al., 2022). The important factor here is the adequacy of the models used to test both new therapeutic hypotheses and drug candidates. These include animal, *in vitro*, computational and AI models, and even certain kinds of experimental medicine in human subjects. These are the tools used to evaluate new drugs and therapeutic mechanisms to decide if they are likely to work in patients.

Decision theory suggests that the ability to detect effective therapeutic candidates is extremely sensitive to a model's "predictive validity". Models have high predictive validity if they rank a set of therapeutic candidates in a way that matches the ranking generated if one could test all the candidates in patients. Furthermore, predictive validity is generally more important than simply being able to test tens or even hundreds of times as many drug candidates. In other words, quality beats quantity.

The history of drug R&D suggests that screening and disease models with high predictive validity (e.g. animal models of bacterial infection, animal models of hypertension, etc.) correctly identified drugs that worked well in people. When their patents expired, the drugs became the generics that undermine economic incentives for further R&D in the disease area. This rendered the best models commercially redundant (Scannell and Bosley, 2016; Shih, Zhang and Aronov, 2018, Scannell et al., 2022). This leaves the diseases for which the models are widely acknowledged to lack predictive validity (e.g. advanced solid cancers, Alzheimer's, etc.).

Ironically, bad screening and disease models often remain in academic and commercial use for decades (Horvath et al., 2016; Scannell et al., 2022). There are several possible reasons for this. First, there may be nothing obviously better. Second, there may be strong tradition and availability biases (Veening-Griffioen et al., 2021). Third, they may not identify the useful drugs that would lead to their redundancy (Scannell and Bosley, 2016). Furthermore, as argued later, the private sector incentives for developing better screening and disease models are relatively weak. A progressive decline in the predictive validity of the stock of industrially relevant screening and disease models offsets the large gains in brute force efficiency.

The post-2010 uptick in drug approvals is also explicable within the framework of screening and disease models' predictive validity. Modern genetic methods have made it easier to match pathological mechanisms with the patients who share the pathology. For genetically identifiable groups of patients, often with rare diseases that have a simple genetic basis, one can create or identify screening and disease models with relatively high predictive validity (Ringel et al., 2020). Modern genetic methods have been much less useful in creating good models of common diseases where single gene errors play a less important role in human pathology (Joyner and Paneth, 2019).

Researchers are clearly not exhausting some other important resources or opportunities. For example, the number of different chemical compounds that could be produced is, for practical purposes, infinitely large. There also appear to be plenty of as-yet-unexploited drug targets and therapeutic mechanisms (Finan et al., 2017; Rodgers et al., 2018; Shih et al., 2018).

AI will be incrementally helpful but not revolutionary in drug discovery

AI technologies will help in drug R&D. They will be important in certain niches, particularly with respect to drug chemistry. However, their overall impact on industry-level productivity will likely be modest in the near term. The areas with the most progress in using AI are rarely relevant to the rate-limiting steps in drug

R&D. Meanwhile, the biggest gaps in the ability to raise R&D productivity tend to be less amenable for AI solutions. For a longer and more technical discussion of these points, see Bender and Cortés-Ciriano, (2021a, 2021b) and a series of blog posts by Lowe (16 December 2021, 9 December 2021, 8 November 2021, 23 July 2021, 30 November 2020, 25 September 2019).

Much of what is called AI, in fact, falls within the broader field of statistical pattern recognition (and increasingly, pattern generation). However, any statistical pattern recognition technology is likely to perform poorly without sufficient training data that closely resemble the real-world problems to which the technology will be applied. This is the case with many of the as-yet-poorly treated diseases that still have commercial value for the drug industry. By AI standards, the data are lousy (Bender and Cortés-Ciriano, 2021a, 2021b). It may be, for example, that most of the published biomedical literature is false, irrelevant or both (Horrobin, 2003, 2001; Ioannidis, 2005; Prinz, Schlange and Asadullah, 2011; Begley and Ellis, 2012). It is unreasonable to expect screening and disease models that struggle to identify good therapeutic mechanisms and good drugs to suddenly generate the reliable and unbiased data required to train accurate AI algorithms. Generating better biological data will help take advantage of AI (and a wide range of older pattern recognition methods). However, that is costly and takes time.

The general enthusiasm for AI also means that a wide set of activities now fall under the AI banner. This includes many disciplines that have been important in drug R&D for decades, such as chemoinformatics, bioinformatics, computational chemistry, structural biology and biostatistics (Bender and Cortés-Ciriano, 2021a, 2021b). For example, the "protein folding problem" of structural biology – where there has been both fanfare and real progress in applying AI – is not new. It reflects "a seventy-year symbiotic relationship between molecular biology and computer science" (Singh, 2020). Similar long-standing relationships exist between computer science and genomics, virtual drug screening, etc. These long-standing relationships overlap in time with a huge decline in R&D productivity (Figure 1A).

Concluding thoughts on rate-limiting steps in drug discovery

There is a degree of consensus that the lack of valid screening and disease models is a major constraint on drug discovery. Researchers also generally agree that novel chemistry is the most appropriable and investible form of biopharmaceutical innovation because it can be protected by strong patents.

Chemistry has become easier over time, in part because of the contribution of computational and data analytic methods. It can be difficult, however, for the private sector to appropriate much of the potential economic value by investing in better screening and disease models (Scannell and Bosley, 2016; Billette de Villemeur and Versaevel, 2019). The models behave as what economists call "public goods" with substantial knowledge spillovers to those who have not invested in them. Once the mechanism identified by the new model is publicly proven in early trials in human patients, for example, the information becomes freely available to competitors. Competitor firms can then exploit the mechanism without investing in the novel models that led to its discovery. This situation skews private sector investment. Cancer drug candidates, for example, have among the highest clinical failure rates of any major therapy area (Shih, Zhang and Aronov, 2018; Wong, Siah and Lo, 2019). Nonetheless, as of 2018, around 1 500 compounds were in human trials (Moser and Verdin, 2018). They were progressed on the basis of *in vitro* and animal-based cancer models that nearly everyone involved believes to be inadequate. In relative terms, investing in chemical "roulette" appears to be too profitable. Meanwhile, investing in the screening and disease models that might improve the odds of the game is not profitable enough.

References

Barker, R.W. and J.W. Scannell (2015), "The life sciences translational challenge: The European perspective", *Therapeutic Innovation & Regulatory Science,* Vol. 49/3, pp. 415-424, https://doi.org/10.1177/2168479014561340.

Begley, C.G. and L.M. Ellis (2012), "Raise standards for preclinical cancer research", *Nature*, Vol. 483/7391, pp. 531-533, https://doi.org/10.1038/483531a.

Bender, A. and I. Cortés-Ciriano (2021a), "Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet", *Drug Discovery Today*, Vol. 26/2, pp. 511-524, https://doi.org/10.1016/j.drudis.2020.12.009.

Bender, A. and I. Cortés-Ciriano, I. (2021b), "Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data", *Drug Discovery Today*, Vol. 26/4, pp. 1040-1052, https://doi.org/10.1016/j.drudis.2020.11.037.

Billette de Villemeur, E. and B. Versaevel (2019), "One lab, two firms, many possibilities: On R&D outsourcing in the biopharmaceutical industry", *Journal of Health Economics*, Vol. 65, pp. 260-283, https://doi.org/10.1016/j.jhealeco.2019.01.002.

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.

Damodaran, A. (2020), "Data: History and Sharing", webpage, http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datahistory.html (accessed 4 May 2020).

Damodaran, A. (2007), "Return on Capital (ROC), Return on Invested Capital (ROIC) and Return on Equity (ROE): Measurement and Implications", *SSRN*, 1105499, https://doi.org/10.2139/ssrn.1105499.

Deloitte (2021), *Seeds of Change: Measuring the Return on Pharmaceutical Innovation 2021*, Deloitte Centre for Health Solutions, London, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-uk-measuring-the-return-from-pharmaceutical-innovation-2021.pdf.

Deloitte (2019), "Pharma R&D return on investment falls to lowest level in a decade", Deloitte, London, 18 December, Press Release, https://www2.deloitte.com/uk/en/pages/press-releases/articles/pharma-r-n-d-return-on-investment-falls-to-lowest-level-in-a-decade.html.

Finan, C. et al. (2017) "The druggable genome and support for target identification and validation in drug development", *Science Translational Medicine*, Vol. 9/383, https://doi.org/10.1126/scitranslmed.aag1166.

Goncharov, I., J.C. Mahlich and B.B. Yurtoglu (2018), "Accounting profitability and the political process: The case of R&D accounting in the pharmaceutical industry", *SSRN*, 2531467, https://doi.org/10.2139/ssrn.2531467.

Goncharov, I., J.C. Mahlich and B.B. Yurtoglu (2014), "R&D investments, intangible capital and profitability in the pharmaceutical industry", *Value in Health*, Vol. 17/7, p. A419, https://doi.org/10.1016/j.jval.2014.08.1025.

Grabowski, H. and J. Vernon (1990), "A new look at the returns and risks to pharmaceutical R&D", *Management Science*, Vol. 36/7, pp. 804-821, https://doi.org/10.1287/mnsc.36.7.804.

Horrobin, D.F. (2003), "Modern biomedical research: An internally self-consistent universe with little contact with medical reality?", *Nature Reviews Drug Discovery*, Vol. 2/2, pp. 151-154, https://doi.org/10.1038/nrd1012.

Horrobin, D.F. (2001), "Realism in drug discovery – could Cassandra be right?", *Nature Biotechnology*, Vol. 19/12, pp. 1099-1100, https://doi.org/10.1038/nbt1201-1099.

Horvath, P. et al. (2016), "Screening out irrelevant cell-based models of disease", *Nature Reviews Drug Discovery*, Vol.15/11, pp. 751-769, https://doi.org/10.1038/nrd.2016.175.

Ioannidis, J.P.A. (2005), "Why most published research findings are false", *PLOS Medicine*, Vol. 2/8, p. e124, https://doi.org/10.1371/journal.pmed.0020124.

Joyner, M.J. and N. Paneth (2019), "Promises, promises, and precision medicine", *The Journal of Clinical Investigation*, Vol. 129/3, pp. 946-948, https://doi.org/10.1172/JCI126119.

Le Fanu, J. (1999), *The Rise and Fall of Modern Medicine*, Little Brown, Boston.

Lowe, D. (16 December 2021), "AI improvements in chemical calculations", In the Pipeline blog, www.science.org/content/blog-post/ai-improvements-chemical-calculations.

Lowe, D. (9 December 2021), "Another AI drug announcement", In the Pipeline blog, www.science.org/content/blog-post/another-ai-drug-announcement.

Lowe, D. (8 November 2021), "AI-generated clinical candidates, so far", In the Pipeline blog, www.science.org/content/blog-post/ai-generated-clinical-candidates-so-far.

Lowe, D. (23 July 2021), "More protein folding progress – what's it mean?", In the Pipeline blog, www.science.org/content/blog-post/more-protein-folding-progress---what-s-it-mean.

Lowe, D. (30 November 2020), "Protein folding, 2020", In the Pipeline blog, www.science.org/content/blog-post/protein-folding-2020.

Lowe, D. (25 September 2019), "What's crucial and what isn't", In the Pipeline blog, www.science.org/content/blog-post/s-crucial-and-isn-t.

Moser, J. and P. Verdin (2018), "Burgeoning oncology pipeline raises questions about sustainability", *Nature Reviews Drug Discovery*, Vol. 17/10, pp. 698-699, https://doi.org/10.1038/nrd.2018.165.

Prinz, F., T. Schlange and K. Asadullah (2011), "Believe it or not: How much can we rely on published data on potential drug targets?, *Nature Reviews Drug Discovery*, Vol. 10/9, pp. 712-712, https://doi.org/10.1038/nrd3439-c1.

Richard, J. and M.D. Wurtman (1997), "What went right: Why is HIV a treatable infection?", *Nature Medicine*, Vol. 3/7, pp. 714-717, https://doi.org/10.1038/nm0797-714.

Ringel, M.S. et al. (2020), "Breaking Eroom's law", *Nature Reviews Drug Discovery*, preprint, https://doi.org/10.1038/d41573-020-00059-3.

Rodgers, G. et al. (2018), "Glimmers in illuminating the druggable genome", *Nature Reviews Drug Discovery*, Vol. 17/5, pp. 301-302, https://doi.org/10.1038/nrd.2017.252.

Scannell, J. et al. (2022), "Predictive validity in drug discovery: What it is, why it matters and how to improve it", *Nature Reviews Drug Discovery*, Vol. 21/12, https://doi.org/10.1038/s41573-022-00552-x.

Scannell, J.W. and J. Bosley (2016), "When quality beats quantity: Decision theory, drug discovery, and the reproducibility crisis", *PLOS ONE*, Vol. 11/2, pp. e0147215, https://doi.org/10.1371/journal.pone.0147215.

Scannell, J.W., S. Hinds and R. Evans (2015), "Financial returns on R&D: Looking back at history, looking forward to adaptive licensing", *Reviews on Recent Clinical Trials*, Vol. 10/1, pp. 2-43, https://doi.org/10.2174/1574887110666150430151751.

Scannell, J. et al. (2012), "Diagnosing the decline in pharmaceutical R&D efficiency", *Nature Reviews Drug Discovery*, Vol. 11/3, pp. 191-200, https://doi.org/10.1038/nrd3681.

Shih, H.-P., X. Zhang and A.M. Aronov (2018), "Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications", *Nature Reviews Drug Discovery*, Vol. 17/1, pp. 19-33, https://doi.org/10.1038/nrd.2017.194.

Singh, J. (2020), "The history of the protein folding problem: A seventy year symbiotic relationship between molecular biology and computer science", 12 June, Medium, https://medium.com/@jaguarsingh/the-history-of-the-protein-folding-problem-a-seventy-year-symbiotic-relationship-between-483afc9f704c.

SSR Health LLC (2014), "Biopharmaceuticals R&D productivity: Metrics, benchmarks and rankings for the 22 largest (by R&D spending) US-Listed firms", SSR Health LLC.

Steward, F. and G. Wibberley (1980), "Drug innovation: What's slowing it down?", *Nature*, Vol. 284/5752, pp. 118-120, https://doi.org/10.1038/284118a0.

Veening-Griffioen, D. et al. (2021). "Tradition, not science, is the basis of animal model selection in translational and applied research", ALTEX - Alternatives to Animal Experimentation, https://doi.org/10.14573/altex.2003301.

Weatherall, M. (1982), "An end to the search for new drugs?", *Nature*, Vol. 296/5856, pp. 387-390, https://doi.org/10.1038/296387a0.

Wong, C.H., K.W. Siah and A.W. Lo (2019), "Estimation of clinical trial success rates and related parameters", *Biostatistics*, Vol. 20/2, pp. 273-286, https://academic.oup.com/biostatistics/article/20/2/273/4817524.

Notes

[1] This graph has not been published elsewhere, although the analysis is easy to replicate and the required accounting data are widely available. Damodaran routinely puts similar analyses and datasets in the public domain (Damodaran, 2020). An important technical point here, often neglected in the public policy debate on drug industry profits and productivity, is that one needs to adjust both the profits and the equity balance for R&D capitalisation if one wants to compare ROE in R&D-intense industries, such as the drug and biotechnology industries, with industrial sectors (Damodaran, 2020, 2007; Goncharov, Mahlich and Yurtoglu, 2018, 2014). Without the adjustment, one overstates the financial performance of R&D-intensive sectors.

[2] In recent decades, expensive drug classes (e.g. cancer drugs and rare disease drugs) have grown as a share of drugs launched. Furthermore, newly launched drugs in these expensive classes generally launch at higher prices than similar drugs launched in prior years. This phenomenon – "mix inflation" – largely superseded like-for-like price inflation (i.e. the same drug getting more expensive in real terms each year). At an aggregate industry level, like-for-like price inflation has become less important. Payers have become much more effective at engendering price competition between branded drugs that are therapeutic substitutes. Like-for-like price inflation itself superseded prescription volume growth as the driver of revenue growth for branded drugs. Patent expiry and generic substitution have been a major drag on branded drug volumes in the United States since the mid-1980s. Expensive branded drugs are now less than 10% of US prescriptions. Note that these pricing comments apply to drug prices net of rebates and discounts and are also US-focused.

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

P. Boeing, ZEW – Leibniz Centre for European Economic Research, Germany

P. Hünermund, Copenhagen Business School, Denmark

Introduction

This essay provides evidence for a decrease in research productivity in the last decades for the People's Republic of China (hereafter "China") and Germany. Estimates imply that research productivity falls, on average, by 5.2% per year in Germany and by 23.8% per year in China, which corresponds to a reduction by half in 13 years and in 3 years, respectively. Results indicate that policy measures to increase the productivity of research and development (R&D) are important for curbing the ongoing global productivity slowdown.

Innovation, productivity and spending

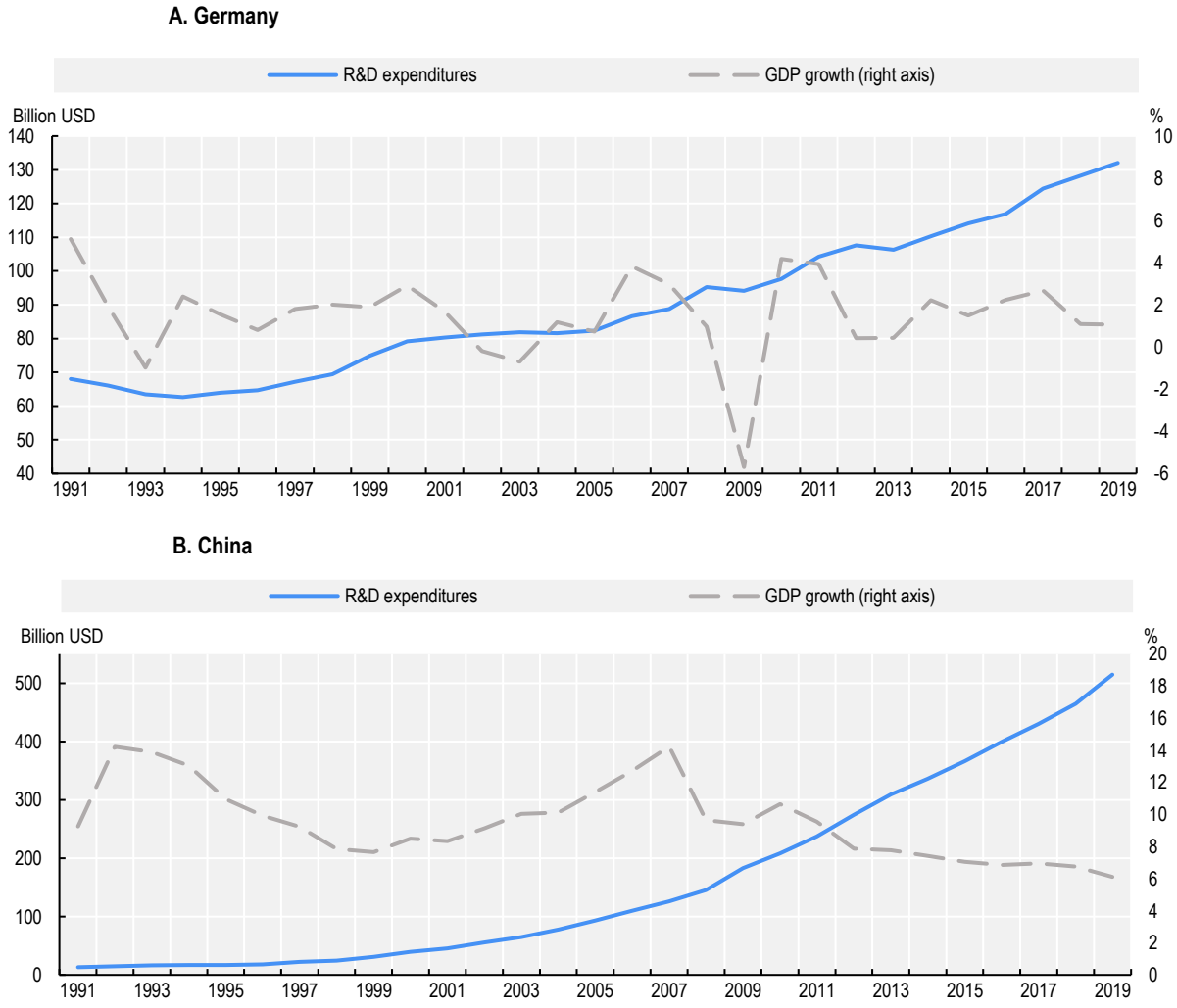
In contrast to assumptions invoked by endogenous growth theory (Romer, 1990), the productivity of R&D might not be constant. Breakthrough innovations may be getting more difficult to achieve over time. This would result in higher levels of R&D spending being needed to maintain constant rates of economic growth.

Figure 1 plots aggregate R&D spending and gross domestic product (GDP) growth rates in China and Germany for the last three decades. It shows that while aggregate spending on R&D was steadily increasing, growth rates of GDP remained constant or even decreased. Similar trends can be observed for the United States (Bloom et al., 2020).

This finding is inconsistent with assumptions invoked in standard endogenous growth models, which usually posit a one-to-one link between R&D spending and growth rates. Several economists have proposed that a decline in R&D productivity over time could explain this decoupling.

Cowen (2011) argues the US economy has benefited from low-hanging fruit in science and technology for the last two centuries; this started to run out in the second half of the 20th century. Gordon (2016) makes a similar argument, stating that new technologies such as electrification, indoor plumbing, home appliances and the rise of motor vehicles created exceptional drivers of economic growth unlikely to be replicable in the coming decades. Jones (2009) describes a "burden of knowledge"; as science is generally cumulative, researchers at the scientific frontier must keep up with an increasing body of knowledge. This, in turn, prolongs training times and renders scientific breakthroughs harder to achieve.

Figure 1. R&D investment and annual GDP growth rates in Germany and China, 1991-2019



Notes: Gross domestic spending on R&D measured in USD billion at constant prices, using 2010 as a base year. Growth domestic product growth rates expressed in percentages.
Source: OECD data, https://data.oecd.org.

Bloom et al. (2020) provide ample empirical evidence, derived from industry case studies and firm-level analyses, that research productivity has fallen over time in the United States. Consequently, and if correct, R&D activities today create a much smaller growth impulse for a given level of spending than 40 years ago.

It is also important methodologically to examine micro-level evidence next to the aggregate data in Figure 1. If new industries are added in an expanding economy, aggregate R&D spending could in principle be increasing while the level of spending per industry is kept constant. Macro trends such as the ones presented in Figure 1 would then suggest a decline in research productivity, although it remains stable at the micro level. To avoid such a misleading picture, researchers need to zoom in and estimate research productivity trends over time for individual sectors and firms. Bloom et al. (2020) analyse the semiconductor industry, agriculture, health care and the US manufacturing sector as a whole. They find evidence for a substantial decline in R&D productivity in all these domains.

Measuring the productivity of R&D

Boeing and Hünermund (2020) replicate the findings of Bloom et al. (2020) for the two largest R&D spending economies in Asia and Europe: China and Germany. They focus on firm-level data and analyses since these provide the most generalisable evidence across different sectors. In other words, because the microdata cover firms in most sectors, the results are likely to be representative of the business sector as a whole rather than just a specific industry or technological field.

As a starting point for measuring the productivity of R&D over time, the analysis uses the following idea production function, which is standard in the endogenous growth literature (Romer, 1990; Aghion and Howitt, 1992):

$$\textit{Economic growth} = \textit{Research productivity} \times \textit{Number of researchers}$$

In line with the approach of Bloom et al. (2020), this allows computation of research productivity at the firm level by dividing growth rates by the number of R&D employees. They compute growth rates, in the numerator, based on several commonly used output measures: sales revenue, employment level, market capitalisation and revenue labour productivity (i.e. sales revenue per worker). The number of researchers, in the denominator, is proxied by R&D spending divided by the average wage of high-skilled workers in the economy. This operationalisation has the advantage of also accounting for complementary capital expenditures within the R&D process.

To smooth out short-term business cycle fluctuations, numbers are averaged over a ten-year period and the change in research productivity is computed over two consecutive decades. This methodology requires detailed firm-level panel data over a long period. For Germany, the analysis relies on data for 64 902 firms from the Community Innovation Survey (Peters and Rammer, 2013; OECD/Eurostat, 2018) for 1992-2017. For China, it takes the universe of 3 947 Chinese firms listed on the Shanghai and Shenzhen stock exchanges – China's so-called A-share market – in 2001-19.

Since firms need to be observed over two consecutive decades and figures are averaged per decade, the sample size drops considerably (to 1 121 in Germany and 516 in China). This is consistent with the original analysis in Bloom et al. (2020). Moreover, as with this essay's study of China, Bloom and co-authors also analysed research productivity trends in US publicly listed firms based on Compustat data. Compared to firms listed on a stock exchange, the Community Innovation Survey contains a larger number of privately held small and medium-sized enterprises. This must be considered when comparing results across countries.

Research productivity trends in China and Germany

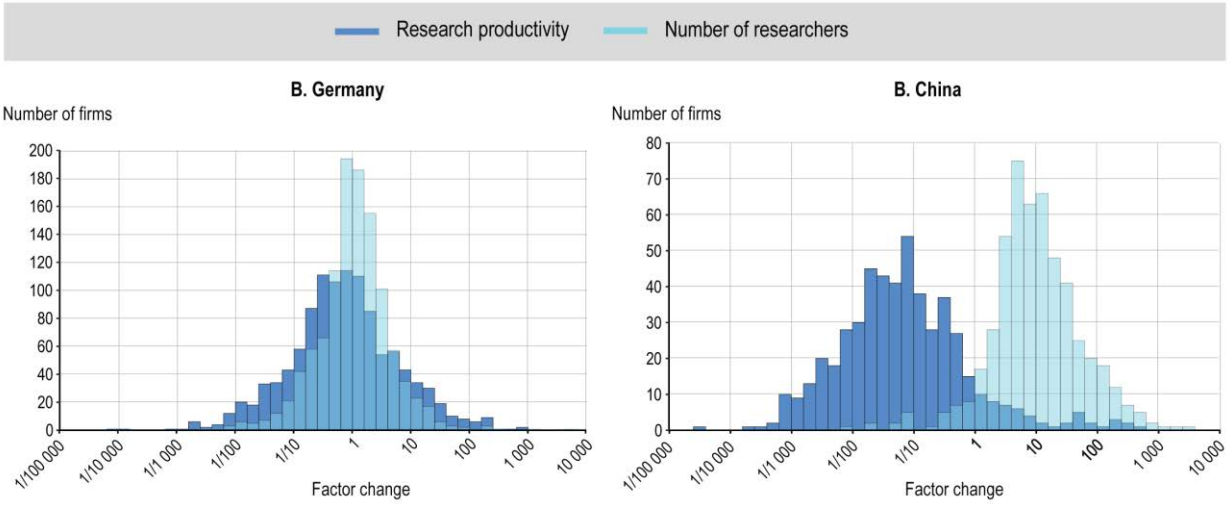
For Germany, R&D expenditures, measured in head counts, increased by an average of 3.3% per year during the period of investigation. At the same time, this expansion of research activities is not accompanied by a similar increase in growth at the firm level. Averaged over all the firm-level outcome measures previously discussed, research productivity declines by 5.2% per year. This is remarkably similar to numbers that Bloom et al. (2020) report for the aggregated US economy. These negative compound average growth rates imply that research productivity reduces by half roughly every 13 years. In other words, research efforts must be doubled every 13 years to support constant economic growth rates.

For China, research productivity has declined even more drastically. This implies the initial growth of significant R&D activities in China that generated high returns in the 2000s subsequently diminished. Here, the effective number of researchers employed by publicly listed firms in the sample increases by, on average, 21.9% per year between 2001 and 2019. Meanwhile, economic growth rates again do not increase proportionally. Arithmetically, this entails a drop in research productivity of 23.8% per year.

The implied half-life of research productivity of around three years constitutes a swift decline. However, if analysis is restricted to the last decade (when China began large-scale R&D activities), and growth rates are compared in five-year intervals (2010-14 and 2015-19), research productivity declines by only 7.3%. These numbers are closer to the ones found for Germany and the United States. They may reflect China's progression from when it was aiming to catch up to developed economies to one where it has been operating closer to the research frontier in many fields.

Figure 2 plots the histograms of changes in number of researchers (light blue bars) and research productivity (dark blue bars) across the two samples. A constant research productivity and a constant number of researchers over time would correspond to a factor change equal to one. As the blue bars in the histogram illustrate, most firms in the sample are located to the left of one, which implies declining research productivity over time. Many firms experienced positive growth rates in research productivity during the last three decades, especially in Germany. This substantial degree of heterogeneity is in line with what Bloom et al. (2020) find for the United States.

Figure 2. The heterogeneity of R&D productivity change across firms in Germany (1992-2017) and China (2001-19)



Note: Aside from the light blue and the dark blue bars, the third colour denotes where the distributions overlap.
Source: Peters and Rammer, 2013; OECD/Eurostat, 2018; authors' calculations based on data from Shanghai and Shenzhen stock exchanges.

Discussion

According to this analysis and that of Bloom et al. (2020), the leading R&D-performing countries in North America, Asia and Europe have all been experiencing a decline in average research productivity over the last two decades. One implication is that further increases in global R&D inputs will be required to avoid contracting GDP growth. For example, according to China's 14th Five-Year Plan for 2021-25, gross R&D spending is expected to increase by at least 7% annually. This is well above the projected 5% annual GDP growth and implies that the current R&D-to-GDP ratio of about 2.2% will increase further. China already accounted for 24.4% of global spending on R&D in 2018, while the United States accounted for 25.6% (in purchasing power parity terms).[1]

However, there may be limits to the increase in actual R&D input. This is because an inelastic supply of researchers tends to increase the cost of R&D (e.g. through higher wages for scientists) but not the amount of R&D activity (Goolsbee, 1998). In the past, education reforms in China led to a steady growth in the number of college graduates. This resulted in a notable increase in the supply of students and researchers

in China and abroad. The United States also benefited greatly from its attraction of foreign scientific talent from around the world.

However, the long-term decline in population growth in industrialised countries, coupled with shocks to international mobility (e.g. during the COVID-19 pandemic) may take a toll on the number of deployable researchers. Thus, in addition to ensuring the quantitative supply of new researchers, policy makers must improve the quality of education and research and the optimal allocation of ideas to slow (or even reverse) the decline in research productivity.

As one limitation, the methodology in this paper cannot distinguish between a productivity effect and a business-stealing effect on firm growth. That is, a firm's R&D and the associated process of creative destruction might influence not only its own productivity but also the market share of its competitors. In that case, output growth would not be the result of increased productivity but come at the cost of competitors. This, in turn, would lead to an overestimation of research productivity at the firm level. Even though both effects could occur at the same time, the productivity effect of R&D has been shown to dominate the business-stealing effect empirically (Bloom Schankerman and Van Reenen, 2013).

Likewise, the methodology does not explicitly consider technology spillovers. Since spillovers lead to productivity growth without direct R&D investments, they enter the measure of research productivity in the numerator. Thus, if technology spillovers had been slowing during the period of observation, it could partly explain the decline in research productivity found over time.

Conclusion

Ideas are not only getting harder to find in the United States but also in the European and Asian countries that spend the most on R&D. Although the estimates are difficult to compare due to different data sources, negative growth rates in Germany and the United States are remarkably similar. China has experienced a much higher decline in research productivity, but growth rates appear to be converging to those of the United States and Germany in recent years. This convergence coincides with a shift in innovation-driven growth in China. This growth has evolved from the imitation of inventions developed elsewhere to genuine innovation at the global technology frontier.

References

Aghion, P. and P. Howitt (1992), "A model of growth through creative destruction", *Econometrica*, Vol. 60/2, pp. 323-351, https://doi.org/10.2307/2951599.

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.

Bloom, N., M. Schankerman and J. Van Reenen (2013), "Identifying technology spillovers

and product market rivalry", *Econometrica*, Vol. 81/4, pp. 1347-1393, https://doi.org/10.3982/ECTA9466.

Boeing, P. and P. Hünermund (2020), "A global decline in research productivity? Evidence from China and Germany", *Economics Letters*, Vol. 197/109646, https://doi.org/10.1016/j.econlet.2020.109646.

Cowen, T. (2011), *The Great Stagnation: How America Ate All the Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better*, Dutton, New York.

Goolsbee, A. (1998), "Does government R&D policy mainly benefit scientists and engineers?", *American Economic Review*, Vol. 88, pp. 298-302, www.jstor.org/stable/116937.

Gordon, R.J. (2016), *The Rise and Fall of American Growth: The US Standard of Living Since the Civil War,* Princeton University Press.

Jones, B.F. (2009), "The burden of knowledge and the 'death of the renaissance man': Is innovation getting harder?", *The Review of Economic Studies*, Vol. 76/1, pp. 283- 317, www.jstor.org/stable/20185091.

OECD/Eurostat (2018), *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th. Edition*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg, https://doi.org/10.1787/9789264304604-en.

Peters, B. and C. Rammer (2013), "Innovation panel surveys in Germany", in *Handbook of Innovation Indicators and Measurement,* Gault, F. (ed.), Edward Elgar Publishing, Cheltenham.

Romer, P.M. (1990), "Endogenous technological change", *Journal of Political Economy*, Vol. 98/5, pp. 71-102, www.jstor.org/stable/2937632.

Note

[1] See https://data.oecd.org.

Declining R&D efficiency: Evidence from Japan

T. Miyagawa, Gakushuin University, Japan

Introduction

Ordinary Japanese people, who believed their country was one of the most technologically advanced, expected Japan would quickly develop a vaccine for COVID-19. The failure to do so shocked Japanese society, raising questions about research and technological progress in the country. Building on earlier research, this essay examines two key measures that show research and development (R&D) efficiency in Japan did indeed decline in the 2010s from the 2000s.

The recent weak contribution of R&D activities to productivity growth is a major issue in the analysis of secular economic stagnation in Japan. Although Japan has maintained a ratio of R&D to gross domestic product (GDP) of 3% for some time, R&D efficiency growth appears to have slowed. Bloom et al. (2020) point to one possible solution to the puzzle. They argued that R&D efficiency, measured by economic productivity growth divided by the number of researchers, has declined in the United States. Following their work, Miyagawa and Ishikawa (2019) found that the efficiency of R&D in Japanese manufacturing and information services has also declined.

Using more recent data, this essay 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, developed by Bloom et al. (2020), is expressed as economic productivity growth divided by labour input allocated to R&D.

Both measures show that R&D efficiency in Japan in the 2010s declined compared to the 2000s. These results suggest the Japanese government should consider further supporting investment in human resources and organisational change, both of which are complementary to R&D.

Secular stagnation and the decline of R&D efficiency

The secular stagnation in economic productivity growth in the United States after the global economic crisis of 2008 is a topic of active discussion. An optimistic view, expressed by Brynjolfsson and McAfee (2014) and Aghion et al. (2019), is that slow labour productivity growth results from a mismeasurement of quality in new technology-generated services. A pessimistic view, articulated by Gordon (2016), is that acceleration of economic productivity growth due to the information and communication technology (ICT) revolution has ended. In another interpretation, Bloom et al. (2020) argue that a decline of R&D efficiency plays a role.

Recognising the importance of R&D for economic productivity, Japan has kept the ratio of R&D to GDP at around 3% for over 20 years. Nevertheless, the Japanese economy has stagnated for a long time and has

not returned to the growth rates seen in the 1980s. Miyagawa and Ishikawa (2019) measured R&D efficiency at the industry level for the 20 years from 1995 to 2015 using the Japan Industrial Productivity database (JIP database)[1] and the EUKLEMS database, following the approach of Bloom et al. (2020). They found that R&D efficiency in Japan declined over this period.

Two approaches to measuring R&D efficiency

This essay considers two approaches for measuring R&D efficiency. The first, developed by Griliches (1979), recognises that accumulation of R&D expenditures constitutes a stock of knowledge that contributes to economic productivity. Assuming a standard production function, R&D efficiency in this approach is expressed as follows:

Economic productivity growth (total factor productivity growth) = R&D efficiency (marginal efficiency of knowledge stock) x R&D intensity (R&D expenditures/GDP) (1)

From equation 1, using time-series data for total factor productivity (TFP) growth and R&D intensity, changes in R&D efficiency can be examined.

The second approach was developed by Bloom et al. (2020). In the context of endogenous growth theory, economic productivity growth here depends on the number of researchers. However, instead of an actual headcount, Bloom et al. (2020) use a measure they term “effective R&D”. They obtain effective R&D by dividing R&D expenditure by an appropriate wage rate for researchers. They do this as a way to address the difficulty of measuring the total number of researchers. Accordingly, from equation 2, Bloom et al. (2020) derive R&D efficiency as follows:

R&D efficiency x effective R&D = TFP growth rate (2)

R&D efficiency in manufacturing

Table 1 shows data on changes in variables associated with measured R&D efficiency in Japanese manufacturing in the late 1990s, 2000s and 2010s. It indicates that TFP growth declined across the three periods, while R&D intensity and effective R&D increased (as noted, effective R&D is measured by dividing R&D expenditure by labour compensation per hour). When these data are included in equations (1) and (2), above, the result suggests a falling trend in R&D efficiency.

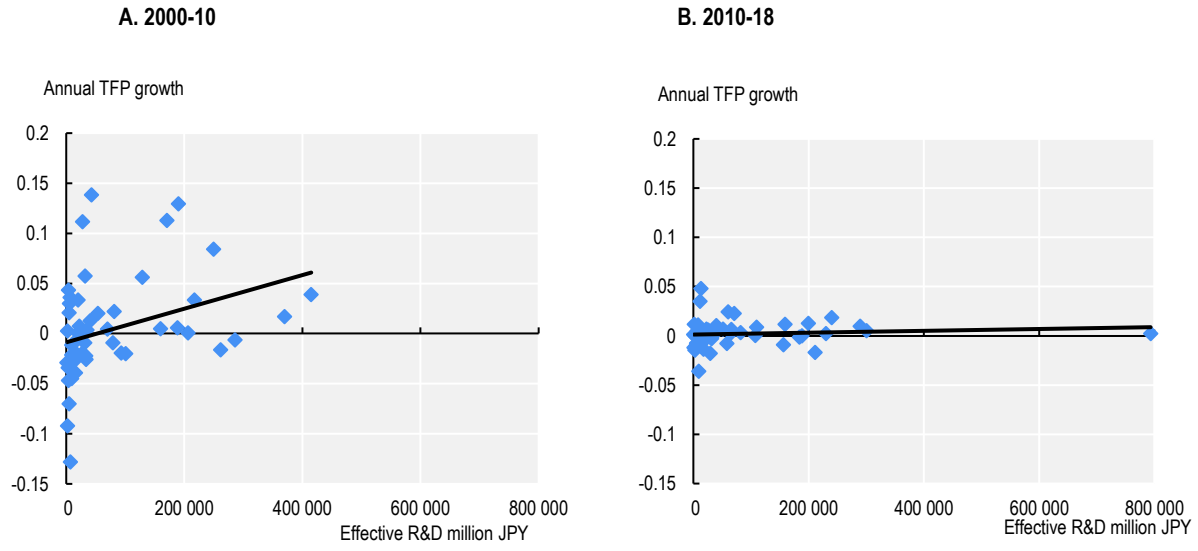
Table 1. Changes in TFP and R&D in manufacturing in Japan, 1995-2018

	1995-2000	2000-10	2010-18
Average annual rate of TFP growth in manufacturing	1.96%	1.80%	0.85%
R&D intensity (R&D/value added in manufacturing)	8.54%	10.64%	12.12%
Effective R&D (year 2000 = 1)	0.98	1.08	1.13

Source: JIP database (2021), www.rieti.go.jp/en/database/JIP2021/index.html.

The analysis also examined the hypothesis of declining R&D efficiency using cross-sectional data. The manufacturing sector in the JIP 2021 database consists of 54 industries, which the analysis divided into two periods: from 2000-10 and 2010-18. TFP growth and effective R&D by industry were measured and plotted in Figures 1 and 2. The slope of the tangent indicates changes in R&D efficiency. The slope in the 2010s (Figure 2) is flatter than that in the 2000s (Figure 1), again suggesting declining R&D efficiency in Japan.

Figure 1. The relationship between annual average TFP growth and effective R&D by industry, 2000-18



Source: JIP database (2021), www.rieti.go.jp/en/database/JIP2021/index.html.

As a main limitation of this approach to R&D efficiency, TFP growth is affected not only by R&D activities but also by several other drivers such as human and organisational capital. Therefore, the analysis includes the number of patents as an outcome of R&D activity because this is a closer proxy of R&D than TFP growth (Hall et al., 2005).

Accordingly, the ratio of new patent applications to the total number of patents are divided into two periods: 1996-2005 and 2006-15. Effective R&D by industry is also measured from the JIP database. R&D efficiency in each period was obtained by dividing the average number of patents in each period by the average value of effective R&D in the corresponding period. This measure showed that R&D efficiency in manufacturing in the second period was 56% of that in the first period.

R&D efficiency in information services

R&D in manufacturing accounts for over 70% of all R&D spending in Japan. However, R&D expenditure in information services is the largest in the service sector overall (excluding the research and education industries). In addition, in information services, software investment has a similar role to R&D investment.

According to the JIP database, average annual TFP growth was negative in Japan's information services industry from 2000 to 2017. As effective R&D has increased since 1995, there was negative R&D efficiency growth in information services. One possible reason for this is the slow growth of the information services market in Japan. In the United Kingdom and the United States, TFP growth rates in information services became positive in the 2000s, after negative rates in the late 1990s. In this case, companies in information services invested aggressively in R&D and software early in the ICT revolution. As the technology-productivity J-curve developed by Brynjolfsson et al. (2021) suggests, this investment in new technology in the late 1990s likely contributed to high productivity growth in the 2000s.

Conclusion

Using several measures, this essay shows that R&D efficiency in Japanese manufacturing, and in the information services industry, has declined. These findings are consistent with Bloom et al. (2020), who pointed to a decline in R&D efficiency in the United States. Japan has spent around 3% of GDP on R&D for many years. However, these results imply the scale of investment is not enough to achieve required improvements in economic productivity.

References

Aghion, P. et al. (2019), "Missing growth from creative destruction", *The American Economic Review*, Vol. 109/8, pp. 2795-2822, https://doi.org/10.1257/aer.20171745.

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

Brynjolfsson, E. and A. McAfee (2014), The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, W.W. Norton, New York.

Brynjolfsson, E. et al. (2021), "The productivity J-curve: How intangibles complement general-purpose technologies", *American Economic Journal: Macroeconomics*, Vol. 13/1, pp. 333-372, https://doi.org/10.1257/mac.20180386.

Gordon, R.J. (2016), The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War, Princeton University Press.

Griliches, Z. (1979), "Issues in assessing the contribution of research and development to productivity growth", *Bell Journal of Economics*, Vol. 10/1, pp. 92-116, https://doi.org/10.2307/3003321.

Hall, B. et al. (2005), "Market value and patent citations", *The RAND Journal of Economics*, Vol. 36/1, pp. 16-38, www.jstor.org/stable/1593752.

Miyagawa, T. and T. Ishikawa (2019), *On the Decline of R&D Efficiency*, RIETI Discussion Paper Series, No. 9-E-052, Research Institute of Economy, Trade and Industry, Tokyo.

Note

[1] The JIP database is a KLEMS-type database in Japan. The authors used the 2021 version, www.rieti.go.jp/en/database/JIP2021/index.html (accessed 10 August 2022).

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

S. Milojević, Indiana University, United States

Introduction

This essay presents and discusses recent trends in the growth of science in terms of the extent of knowledge in scientific literature rather than from the standpoint of publication volumes. It applies an existing method for evaluating the cognitive content of scientific publications to the entire Web of Science database for the period 1900-2020. It compares the growth dynamics of science based on productivity measures with those based on a concept of "cognitive territory", finding stagnation in the latter since the mid-2000s. Growth dynamics are also examined for individual fields of research, showing that physics, astronomy and biology are expanding, whereas medicine is stagnating or even contracting. Cognitive extent is compared for different countries. While the People's Republic of China (hereafter "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.

Is science in decline? This seemingly simple question has so far eluded a definitive answer because it depends on how science is measured. The development of science indicators and metrics dates back to the 1962 publication of the so-called *Frascati Manual*, now in its sixth edition (2015), by the Organisation for European Economic Co-operation (later OECD).

These pioneering efforts enabled assessment of and better informed science policies. However, they also emphasised economic input-output measures stripped of any context rather than the content of science and the context of its creation. Thus, the common measures of scientific progress mostly focus on volume, capturing either the number of papers or the number of researchers.

The focus on the number of journal articles assumes that individual papers represent "a unit" of knowledge, and that all papers contribute equally to the advancement of science. To address the issue of an unequal contribution of individual papers (and/or authors), the focus on volume was complemented by impact measures, most commonly defined in terms of the number of citations received. However, neither volume nor impact is a good measure of the breadth or extent of knowledge produced. To assess the extent of knowledge, the focus needs to shift to the content of publications as exhibited in their text.

Measuring cognitive extent

The use of text in the quantitative study of science has a rich history, especially in mapping the structure of scientific fields. Milojević (2015) quantifies the cognitive extent of scientific fields by using information

contained in the titles of journal articles. The method is statistical and uses natural language processing to extract phrases from titles in English. The phrases are combinations of words that describe specific concepts, such as the methods of study, scientific instruments, and objects of study and their properties. For example, "scanning tunnelling microscopy" would be identified as a phrase, as would "high-temperature superconducting". General words, such as "study" or "observation", would not be identified as a phrase.

Cognitive extent might be easiest to envision as a measure, at any point in time, of the intellectual territory covered by science or its different subunits (e.g. broad research areas, specialisations, journals, countries). For the measure to be unbiased by the volume of output, and to facilitate comparisons across articles with titles of unequal lengths, the cognitive extent is calculated using samples that always contain the same number of title phrases (10 000, corresponding to about 3 000 articles).

Once the phrases in the titles are identified, the number of unique phrases is counted. A smaller number of unique phrases among 10 000 title phrases (e.g. 5 000) would indicate a lot of repetition in the given volume of literature. One could say this body of literature covers a smaller cognitive territory. On the other hand, a large number of unique phrases (e.g. 8 000) means the body of literature examined covers a wider range of concepts and a larger cognitive territory.

Depending on the object of study (the entirety of scientific output, a specific field of science, etc.), the batch of 10 000 title phrases comes from either one of two sources. First, it could come from around 3 000 randomly selected articles from any field, as long as the articles are all published in the same period (typically the same year). Second, it might come only from articles in some predefined field, again so long as they are all published in the same period.

The cognitive extent measure at any time is static (not cumulative) and therefore not influenced by previous states – each measurement is independent. Although many of the specific phrases change from one batch to another, the resulting degree of diversity is remarkably stable from one year to the next. It is true that not all phrases are equally useful or relevant. However, this is not a significant limitation of the method. The measure is applied in a relative sense: comparing one country to another, or one period to another. In each body of literature to which the measure is applied there will be a distribution of phrases across the spectrum of relevance.

Milojević (2015) applied the above measure to follow developments in three research areas (physics, astronomy and biomedicine). It found that while the number of papers grew exponentially in all three fields, between 1900 (1945 for biomedicine) and 2010, their cognitive extent grew linearly. Furthermore, the measure was applied to literature produced by teams of differing sizes, where an inverse relationship was found between cognitive extent and team size. This finding suggests that small teams play a particularly important role in expanding the cognitive territory.

For the purposes of the OECD work on artificial intelligence and the productivity of science, this method was applied to the entire Web of Science database covering 1900 to 2020. It encompasses the literature for science as a whole, as well as individual research fields. It also examined the cognitive extent of research produced in different countries. Driving these analyses was the question: "Is science (still) expanding, and if so how fast?"

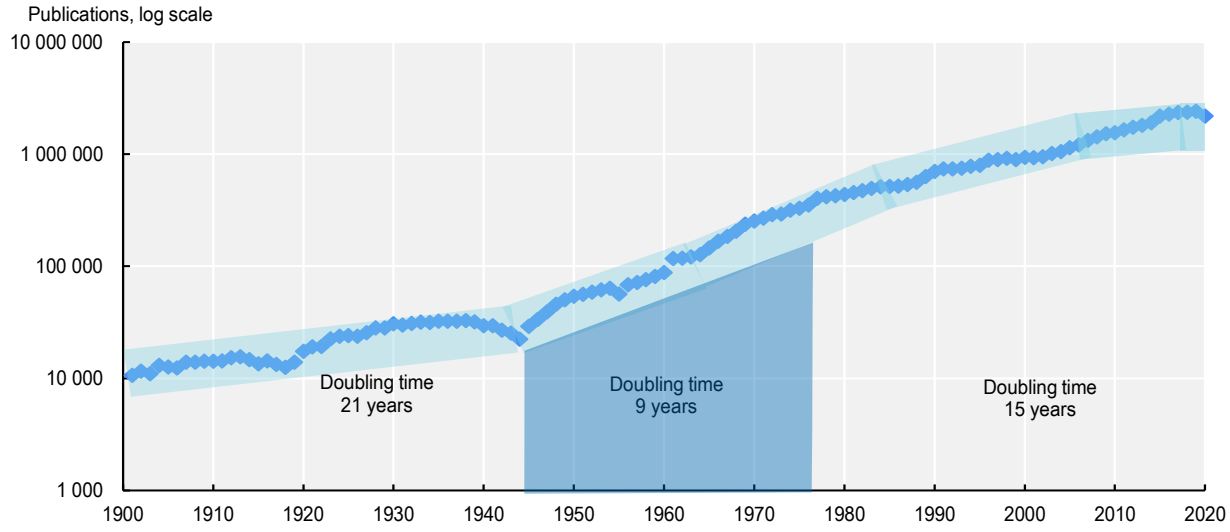
Is the knowledge produced by science expanding?

Several studies have shown the number of published scientific papers is doubling every 9 to 15 years (Larsen and von Ins, 2010; Bornmann and Mutz, 2015). Using more recent data, Figure 1 shows the growth of scientific output as indexed in the Web of Science from 1900 to 2020. Interestingly, although scientific output grew approximately exponentially, the rate of exponential growth varied. The fastest growth (with a

doubling time of nine years) was in the immediate aftermath of the Second World War until the mid-1970s. Since then, scientific output has doubled every 15 years.

Figure 1. The exponential growth of scientific publications, 1900-2020

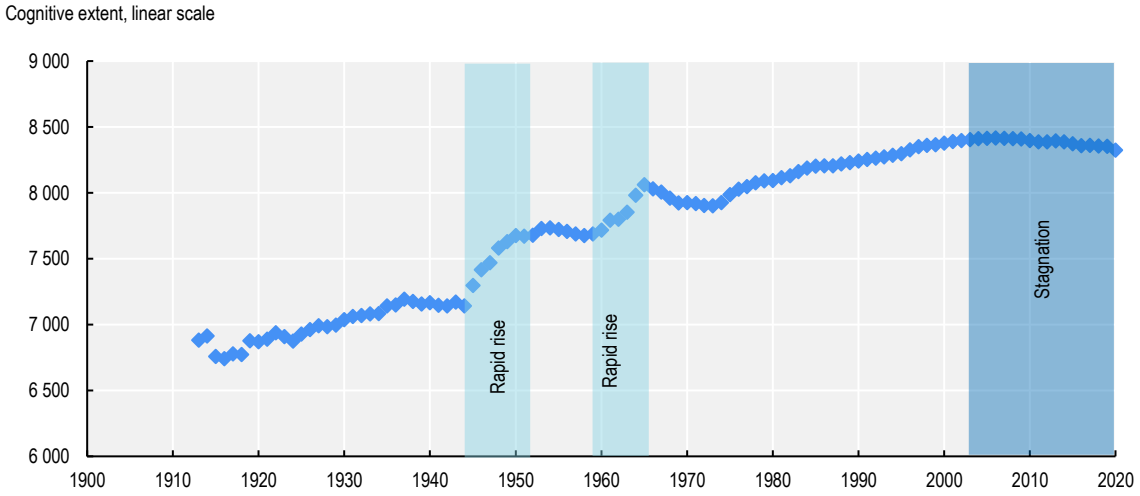
All of science publications



Source: Author's calculations based on Web of Science data, www.webofknowledge.com.

Figure 2. Growth of cognitive extent across all of science, 1900-2020

All of science publications



Note: "Cognitive extent" refers to the number of unique phrases among 10 000 article title phrases.
Source: Author's calculations based on Web of Science data, www.webofknowledge.com.

While the volume of scientific output has grown exponentially, Figure 2 shows that the cognitive territory of science has only grown linearly (Milojević, 2015; Fortunato et al., 2018). Recall that cognitive extent, the vertical axis on the figure, is expressed as the number of unique phrases among 10 000 article title phrases. Figure 2 is based on the same articles used for Figure 1 but paints a different picture. The fastest

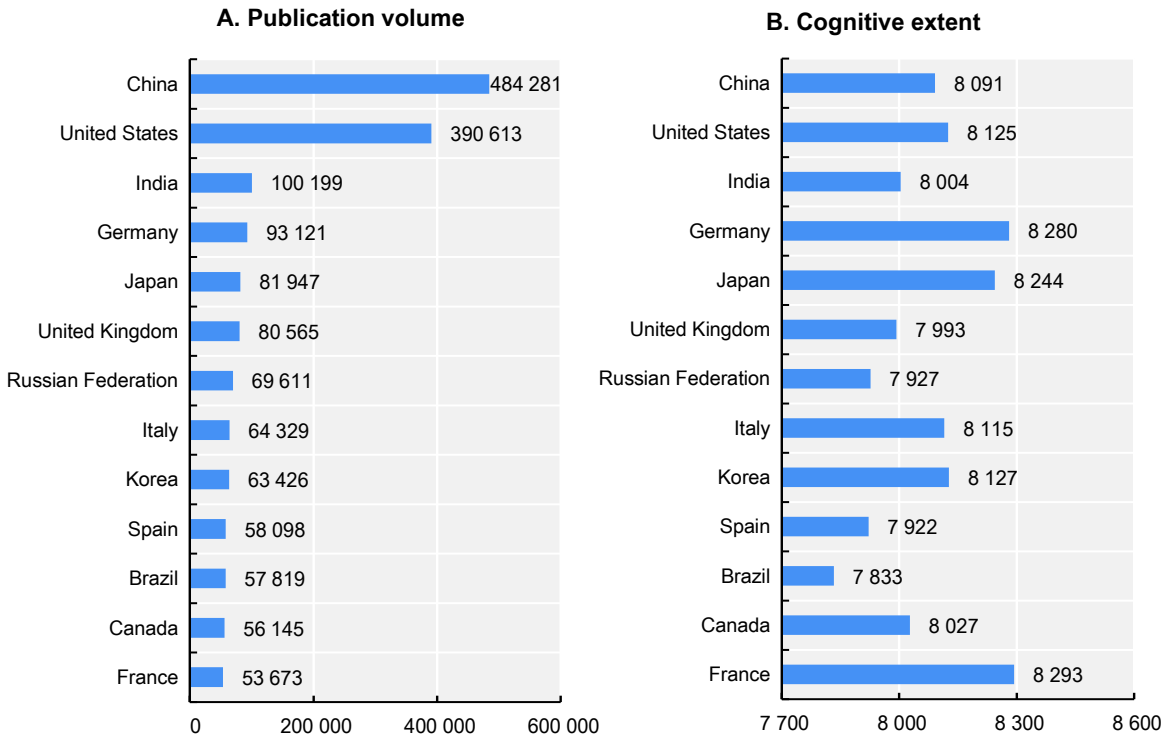
expansion of cognitive extent occurred immediately after the Second World War (1945-51) and then after the Soviet Union launched Sputnik (1958-65). In addition, the continuing increase in literature production (Figure 1) is accompanied by a slowing, and in some years stagnation, in the cognitive extent of science. This starts in 2004 and continues until the time series ends in 2020.

It is well known that different scientific fields develop at a different pace. The work described here shows that not all fields have stagnated. Physics, astronomy and biology are expanding in cognitive extent, whereas mathematics, social sciences, computer science and psychology show slower expansion. Earth sciences, chemistry, agriculture and engineering appear to be stagnating. Meanwhile, medicine has even experienced a contraction since around 2009. In general, basic sciences are expanding, while applied sciences are not.

Assessing differences across countries

After the launch of Sputnik in 1957, strategic science policies and increased investment in science assumed unprecedented importance in many national policies. It was what Johnson (1972) described as a kind of science Olympics. Figure 3 shows, for 2019, a ranked list of countries with respect both to scientific output and cognitive extent. China has become the major producer of scientific literature globally, overtaking the United States. However, China's scientific output still covers a smaller cognitive territory than countries with a longer tradition of modern science, such as France, Germany and the United States. (Reminder: for all countries to calculate the cognitive extent, this essay used batches of publications of the same size – 10 000 phrases or roughly 3 000 articles.)

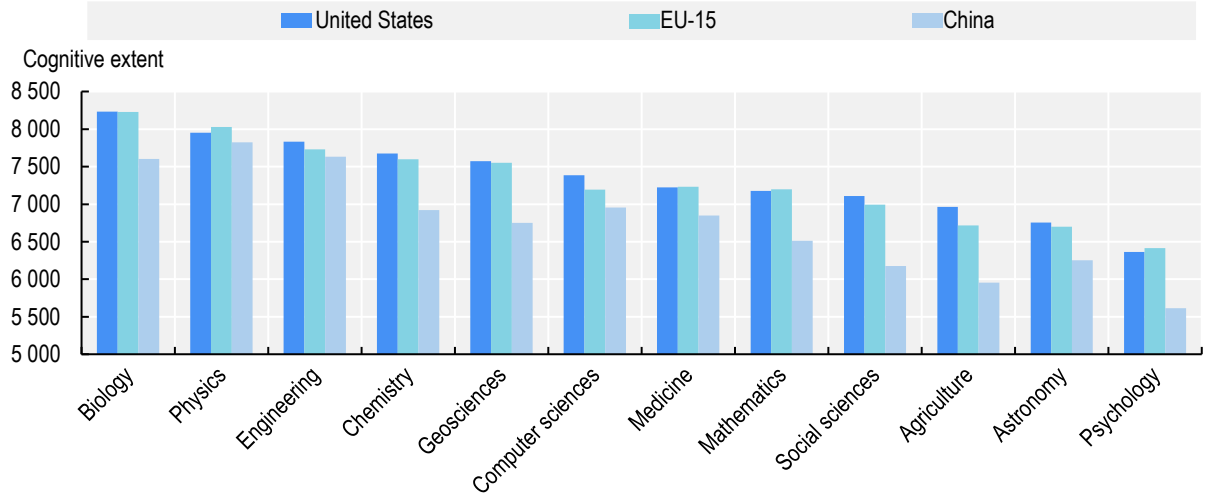
Figure 3. Ranked list of countries based on number of scientific publications and the cognitive extent of science, 2019



Source: Author's calculations based on Web of Science data, www.webofknowledge.com.

Figure 4 shows the cognitive extent of broad fields of science in different countries, in this instance, the United States, the first 15 members of the European Union and China. It focuses on EU-15 rather than the current EU members to consider countries that have traditionally contributed most to overall scientific input. Figure 4 suggests these countries appear to follow different strategies in terms of how broadly they cover different research areas (Figure 4). In terms of cognitive extent, China is approaching the United States and the EU-15 countries in physics and engineering, followed by computer science and medicine. However, China lags significantly in psychology, agriculture and social sciences. It is beyond the scope of this essay to speculate on why this may be the case. However, as one possible factor, the literature in these fields is primarily published in national languages, and therefore not included in this study.

Figure 4. Cognitive extent by broad research area for the United States, EU-15 countries and China, for publications in 2019



Source: Author's calculations based on Web of Science data, www.webofknowledge.com/.

Conclusion

Cognitive extent is an interesting and important measure to add to other measures of science. As suggested in Milojević et al. (2017), "… rather than thinking of individual publications as accreting into an ever-greater understanding of the natural world, it may be better to think of science as building a ladder to the sky. Some publications add new rungs to the ladder, while others primarily increase the width of the ladder."

Following this insight, cognitive extent can be viewed as a measure of the productivity of science not based on a count of publications (however weighted) but rather as an indicator of the pace at which new rungs are added to the ladder. Increasing the width of the ladder may also be important to ascend, but the two still need to be distinguished.

When combined with other measures, cognitive extent can shed new light on the dynamics of science as a whole, as well as its individual fields. For example, we might not expect rapid advances in fields that cover a large territory or extent but have a small number of researchers working in this domain. Conversely, fields with a relatively small territory or extent and many researchers would be more likely to have a clearly defined research frontier. This helps them to form consensus and resolve open questions quickly.

Some results here indicate that science as a whole may be stagnating. This is occurring, to some extent, in terms of the expansion of frontiers of knowledge rather than in terms of sheer output. Some areas may

be in decline. The decline may indicate areas with increased concentration on existing problems using current approaches rather than the introduction of novelty. Such concentration might lead to faster short-term solutions for existing problems, due to intensified effort. However, it might also make future progress slower, if new ideas that might serve to seed progress appear more slowly. More research that includes both qualitative and quantitative approaches will be needed to give a definitive answer to these questions.

References

Bornmann, L. and R. Mutz (2015), "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references", *Journal of the Association for Information Science and Technology,* Vol. 66/11, pp. 2215-2222, https://doi.org/10.1002/asi.23329.

Fortunato, S. et al. (2018), "Science of science", *Science*, Vol. 359/6379, p. eaao0185, https://doi.org/10.1126/science.aao0185.

Johnson, H.G. (1972), "Some economic aspects of science", *Minerva*, Vol. 10/1, pp. 10-18, www.jstor.org/stable/41822128.

Larsen, P.O. and M. von Ins (2010), "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index", *Scientometrics*, Vol. 84/3, pp. 575-603, https://doi.org/10.1007/s11192-010-0202-z.

Milojević, S. (2015), "Quantifying the cognitive extent of science", *Journal of Informetrics*, Vol. 9/4, pp. 962-973, https://doi.org/10.1016/j.joi.2015.10.005.

Milojević, S. et al. (2017), "Team composition and the pace of science: An ecological perspective", presented at the LEI-BRICK Workshop, Organization, Economics and Policy of Scientific Research, Turin.

What can bibliometrics contribute to understanding research productivity?

G. Abramo, National Research Council of Italy

C.A. D'Angelo, University of Rome "Tor Vergata", Italy

Introduction

Measuring academic research productivity is a formidable task, mainly because of the lack of data on inputs. Most of the bibliometric approaches proposed to get around the problem are questionable since they are based on assumptions that invalidate them as supports for policy or management decisions. This essay presents a proxy bibliometric indicator of research productivity that overcomes most of the assumptions and limits that affect the more popular indicators.

In today's knowledge-based economy, governments strive to continuously improve the effectiveness and efficiency of scientific systems to support competitiveness and socio-economic development. Countries are therefore increasingly moving to strengthen competitive mechanisms in public research, mainly through selective funding and merit-based access to resources (Hicks, 2012). Of the members of the (former) EU28, for example, 16 countries use some form of "performance-based research funding", or PBRF (Zacharewicz et al., 2019). PBRF systems are generally associated with national research assessment exercises. These resort more or less extensively to evaluative bibliometrics for measuring research performance and ranking universities and public research organisations.

The next section discusses the most popular bibliometric indicators used to assess research performance. The essay then presents what is arguably, to date, the most accurate bibliometric indicator of research performance. It stresses the need for governments to provide bibliometricians with input data (on labour and capital) to research institutions, the lack of which hinders precise measurements. It also presents the first results of a longitudinal analysis of academic research productivity at a national level. It shows that productivity is increasing over time for Italian academics in most research fields and overall.

Evaluative bibliometrics

Evaluative bibliometrics builds on two pillars of information: 1) publications indexed in bibliographic repertories, as a measure of research output; and 2) citations received, as a measure of their value, known as "scholarly impact". The underlying rationale is that for research results to have an impact, they must be used and citations must certify their use. The intrinsic limits of evaluative bibliometrics are apparent: 1) publications are not representative of all knowledge produced; 2) bibliographic repertories do not cover all publications; [1] and 3) citations are not always a certification of real use and need not reflect all use.

The past two decades have seen a proliferation of research performance indicators and their variants. This has disoriented decision makers and practitioners, who are no longer able to discriminate the relative pros and cons. The next subsection analyses the most popular categories of these indicators.

Number of publications per researcher

One indicator of research productivity is simply the number of publications per researcher. This would be an acceptable metric if the resources used for all research were the same and if all papers, once published, were to have the same impact. However, these assumptions could not be further from the truth.

Mean normalised citation score

Another category consists of "citation size-independent indicators", which are based on a ratio of citations to publications. The most popular representative of this type of indicator is the "mean normalised citation score", or MNCS. The MNCS measures the average number of (normalised)[2] citations of the publications of an individual or institution (Waltman et al., 2011). Within the MNCS category, another indicator of research performance is the share of publications belonging to the top X% of highly cited articles (HCAs).

Such "size-independent" indicators were probably devised to get around the lack of data on inputs to the research process, in particular the names and affiliations of research staff. While relatively easy to measure, both indicators – MNCS and HCAs – are invalid for practical uses. Imagine two universities of precisely the same size, resources and research fields. Two simple questions can be asked:

- Which one performs better: the first university with 100 articles each earning 10 citations (1 000 total), or a second university with 200 articles, of which 100 have 10 citations, and the other 100 have 5 citations (1 500 total)?
- Which performs better: the first university with 10 HCAs out of 100 publications (10% of the total) or a second university with 15 HCAs out of 200 (7.5% of the total)?

In the first example – using MNCS – the second university performs worse than the first (the first has a 25% higher mean citation count). However, using common sense, the second university is the better performer because its higher number of total citations has been produced using the same research resources available to the first university.

In the second case, the first university performs better, as the rate at which it produces HCAs is higher. However, again, using common sense, the second is the better performer as it produces a 50% higher number of HCAs from the same research spending.

This category of indicators violates the self-evident fact that if output increases under equal inputs, performance cannot be considered to have diminished. Paradoxically, an organisation (or individual) will receive a worsened MNCS should it produce an additional publication with a normalised impact even slightly below the previous value for the MNCS.

h-index

Another well-known performance indicator is the *h*-index. In the words of the originator, the *h*-index "represents the maximum number *h* of works by a scientist that have at least *h* citations each" (Hirsch, 2005). Hirsch's intuitive breakthrough was to represent, with a single whole number, a synthesis of both the quantity and impact of the entire portfolio of a scientist's published work.

However, the *h*-index also has drawbacks. First, it ignores the impact of works with a number of citations below *h* and all citations above *h* of the *h*-score works, which is often a very considerable share. Second, it fails to field-normalise citations, favouring publications in citation-intensive fields. Three; it fails to account for the years of life of publications, favouring older ones. Fourth, it also does not adjust for the number of

co-authors and their order in the byline. Lastly, because of the different intensity of publications across research fields, comparing *h*-indexes for researchers across fields can lead to wrong conclusions. Each of the proposed *h*-variant indicators tackles one of the many drawbacks of the *h-index* while leaving the others unsolved. Therefore, none can be considered entirely satisfactory (Iglesias and Pecharromán, 2007; Bornmann et al., 2008).

However, all of the above performance indicators share a common problem: they all focus on outputs and ignore the inputs to research.

Ways forward

Research performance evaluations based on the above indicators are, at best, of little value. Indeed, they could be dangerous due to the distortions embedded in the information provided to decision makers.[3] Some years ago, to overcome the limitations of these indicators, the authors conceived, operationalised and applied a proxy indicator of research productivity derived from the microeconomic theory of production: "Fractional Scientific Strength", or FSS (Abramo and D'Angelo, 2014).

In simple terms, the FSS of a researcher is the ratio of the value of research output, in a given period, to the cost of the inputs used to produce it. The output consists of researchers' contributions to their publications indexed in bibliographic repertoires. Citation-based metrics measure the value of each publication.[4] The cost of inputs consists of the researcher's wage (labour) and other resources (capital) used to carry out the research.[5]

Unlike the most popular indicators mentioned above, the FSS accounts for input data in addition to output. Nevertheless, all the usual limits of evaluative bibliometrics apply here too. First, publications are not representative of all knowledge produced. Second, bibliographic repertories do not cover all publications. Finally, citations are not always a certification of real use or representative of all use. Furthermore, results are sensitive to the classification schemes adopted for both publications and professors.

Because the intensity of publication varies across research fields, researchers' productivity is compared to that of others in the same fields.[6] For the same reason, productivity at the aggregate level (university, department, research group, discipline or field) cannot be measured by simply averaging the productivity of individual researchers (of each university, department, etc.). A three-step procedure is required: measuring the productivity of each researcher in a field; normalising the individual's productivity by the average in the field (for instance, an FSS value of 1.10 means the researcher's productivity is 10% above average); and finally, averaging the normalised productivities.

The FSS can be applied to the Italian academic context thanks to access to input data not readily available in other countries.[7]

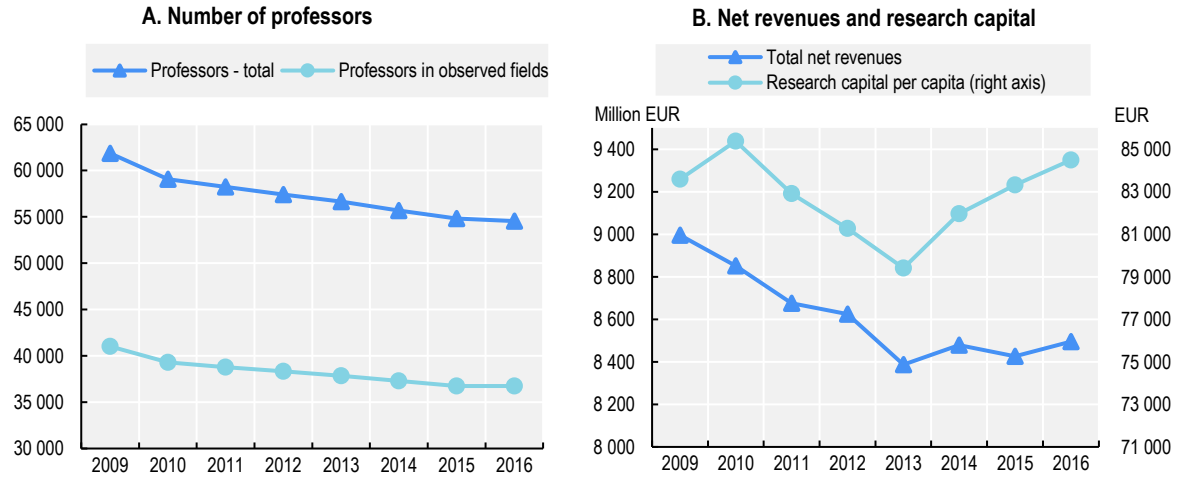
Evidence of variation in academic research productivity over time in Italy

The authors assessed variation in research productivity of all Italian professors in the sciences between two consecutive periods: 2009-12 and 2013-16. The analysis was restricted to Italian professors because data were lacking on inputs for public research organisations other than universities in Italy, and for universities and public research organisations in the rest of the world. The choice of four-year observation periods helps assure the robustness of the results. Input data refer to the two periods, while output data refer to a period one year later. This assumes it takes, on average, a year from knowledge production to its publication.

In the Italian academic system, professors are classified as working in "scientific disciplinary sectors" (SDSs) – e.g. experimental physics, physics of matter, analytical chemistry, organic chemistry, etc. SDSs, in turn, are grouped into "university disciplinary areas" (UDAs), e.g. physics, chemistry, etc. Analysis was

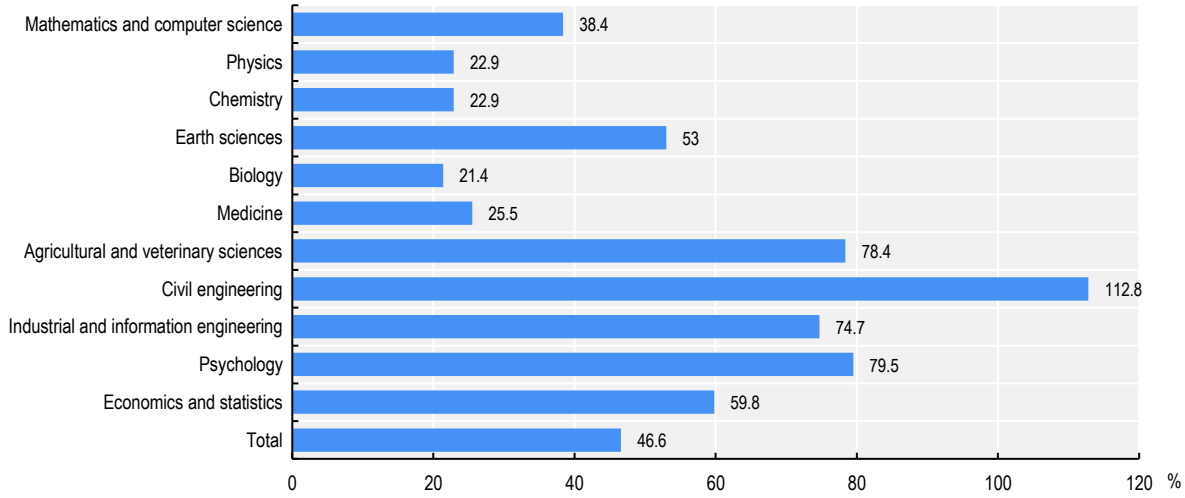
limited to the UDAs where bibliometrics can be applied (10 in all, containing 215 SDSs). The analyses were carried out at the SDS level and then aggregated to the UDA and overall levels.

Figure 1. Number of professors in the Italian academic system; and universities' net revenues and research capital per capita at constant prices, 2009-16



Note: The 2009 and 2010 values are inferred.
Source: Ministry of University and Research, http://cercauniversita.cineca.it/php5/docenti/cerca.php, for number of professors; ba.miur.it, for total net revenues.

Figure 2. Variations in research productivity in the 2013-16 period as compared to the 2009-12 period, by university disciplinary area



Source: Data elaborations by the authors based on Web of Science Italian publications.

With respect to research inputs, Figure 1 shows that between 2009 and 2016 the total number of professors decreased by over 10%. In the same period, the overall net revenues of public universities decreased significantly until 2013, with a slight recovery after that. This implies the U-shaped plot of the yearly resources per capita used for research (dotted line in the right panel of Figure 1).

Figure 2 reports by UDA the change in research productivity in the later four-year period, compared to the earlier four-year period. The overall average variation is +46.6%, led by civil engineering (+112.8%), psychology (+79.5%), and agricultural and veterinary sciences (+78.4%). The lowest average increases occurred in biology (+21.4), physics (+22.9%) and chemistry (+22.9%).

Only 12 SDSs of the 215 total registered a decrease in productivity. The average decrease is -7.6%. Conversely, 203 SDSs registered a productivity increase (+49.0%).

Conclusion

Bibliometrics can inform large-scale assessments of academic research productivity in the sciences. However, the lack of input data in most countries has led bibliometricians to develop indicators based on assumptions that largely invalidate them as a support to policy or management decisions. Governments and research institutions may well expect precise and reliable performance evaluations in support of decisions and policy making. If so, they must be prepared to provide bibliometricians, wherever possible, with the data necessary for the assessments (i.e. name and affiliation of scientists, field of research, wage or academic rank, other resources allocated, etc.).

Such data are largely available in Italy, which has permitted conception and application of an output-to-input indicator of research performance at individual and aggregate levels. An intertemporal analysis between two consecutive periods, 2009-12 and 2013-16, showed that Italian academics in science increased research productivity overall and in about 95% of research fields.

The reasons behind these widespread and noticeable increases in productivity are arguably to be found in the competitive mechanisms introduced by the Italian government in the public research sector. These comprise selective funding, merit-based access to public resources and performance-based access to academia. In particular, the government set up a national agency to evaluate research. It completed the first national research assessment exercise in July 2013 with the publication of the university performance ranking lists. Finally, it set up the national scientific accreditation scheme for professorships, based on bibliometric performance indicators and relevant thresholds.

A key question for all countries is how to conduct national or international research productivity assessments without relevant data on research inputs. There are two possible research trajectories in evaluative bibliometrics. The classic one, followed by most, keeps ignoring input data and tries to improve the old output-based indicators (or propose new ones). Conversely, a new paradigm tries to find ways to identify and account for input data. Among the latter, one stream of research is trying to trace the research personnel of the institutions indirectly, through their publications, using bibliographic repertories that embed the possible affiliation of each author.

The authors are currently investigating the extent of bias in the productivity rankings of research organisations. They are looking for bias when assessments are based on research personnel identified through such “indirect” methods. This aims to inform policy makers’ decisions on whether to invest in building national research staff databases instead of settling for indirect methods, which have their own measurement biases.

References

Abramo, G. and C.A. D’Angelo (2018), “A comparison of university performance scores and ranks by MNCS and FSS”, *Journal of Informetrics*, Vol. 10/4, pp. 889-901, https://arxiv.org/abs/1810.12661

Abramo, G. and C.A. D’Angelo (2014), “How do you define and measure research productivity?” *Scientometrics,* Vol. 101/2, pp. 1129-1144, http://doi.org/10.1007/s11192-014-1269-8.

Abramo, G. et al. (2020), “Comparison of research productivity of Italian and Norwegian professors and universities”, *Journal of Informetrics*, Vol. 14/2, pp. 101023, https://doi.org/10.1016/j.joi.2020.101023.

Abramo, G. et al. (2019), "Predicting long-term publication impact through a combination of early citations and journal impact factor", *Journal of Informetrics,* Vol. 13/1, pp. 32-49, https://doi.org/10.1016/j.joi.2018.11.003.

Archambault, É. et al. (2006), "Benchmarking scientific output in the social sciences and humanities: The limits of existing databases", *Scientometrics*, Vol. 68/3, pp. 329-342, https://doi.org/10.1007/s11192-006-0115-z.

Bornmann, L. et al. (2008), "Are there better indices for evaluation purposes than the h-index? A comparison of nine different variants of the h-index using data from biomedicine", *Journal of the American Society for Information Science and Technology*, Vol. 59/5, pp. 830-837, https://doi.org/10.1002/asi.20806.

Hicks, D. (2012), "Performance-based university research funding systems", *Research Policy*, Vol. 41/2, pp. 251-261, https://doi.org/10.1016/j.respol.2011.09.007.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", in *Proceedings of the National Academy of Sciences*, Vol. 102/46, pp. 16569-16572, https://doi.org/10.1073/pnas.0507655102.

Iglesias, J.E. and C. Pecharromán (2007), "Scaling the h-index for different scientific ISI fields", *Scientometrics,* Vol. 73/3, pp. 303-320, https://doi.org/10.1007/s11192-007-1805-x.

Waltman, L. et al. (2011), "Towards a new crown indicator: Some theoretical considerations", *Journal of Informetrics*, Vol. 5/1, pp. 37-47, https://doi.org/10.1016/j.joi.2010.08.001.

Zacharewicz, T. et al. (2019), "Performance-based research funding in EU member states – a comparative assessment", *Science and Public Policy*, Vol. 46/1, pp. 105-115, https://doi.org/10.1093/scipol/scy041.

Notes

[1] A corollary is that evaluative bibliometrics should not be applied to the arts and humanities, due to the scarce coverage of these fields in bibliographic repertories (Archambault et al., 2006).

[2] Citations are normalised to the average citations of all world publications of the same year and field. This aims to avoid favouring older publications, which would accumulate more citations simply because there has been more time for them to be cited, or publications falling in fields with a high intensity of citation.

[3] To appreciate the magnitude of such distortions, see Abramo and D'Angelo (2018).

[4] Weighted combinations of normalised citations and normalised impact factor (i.e. the prestige) of the hosting journal are used. These are the best predictors of publications' future total citations (see Abramo et al., 2019).

[5] Abramo et al. (2020) explains the limits and assumptions embedded in the operationalisation of the measurement.

[6] In Italy, all academics are officially classified in one and only one field. In countries where this classification is missing, the field of research might be identified as the field in which the scientist's publications are most frequent.

[7] Nevertheless, it is still necessary to make several assumptions that limit the final results. Abramo and D'Angelo (2014) describe the data, FSS formula and methods used to assess research productivity in Italy.

Part II Artificial intelligence in science today

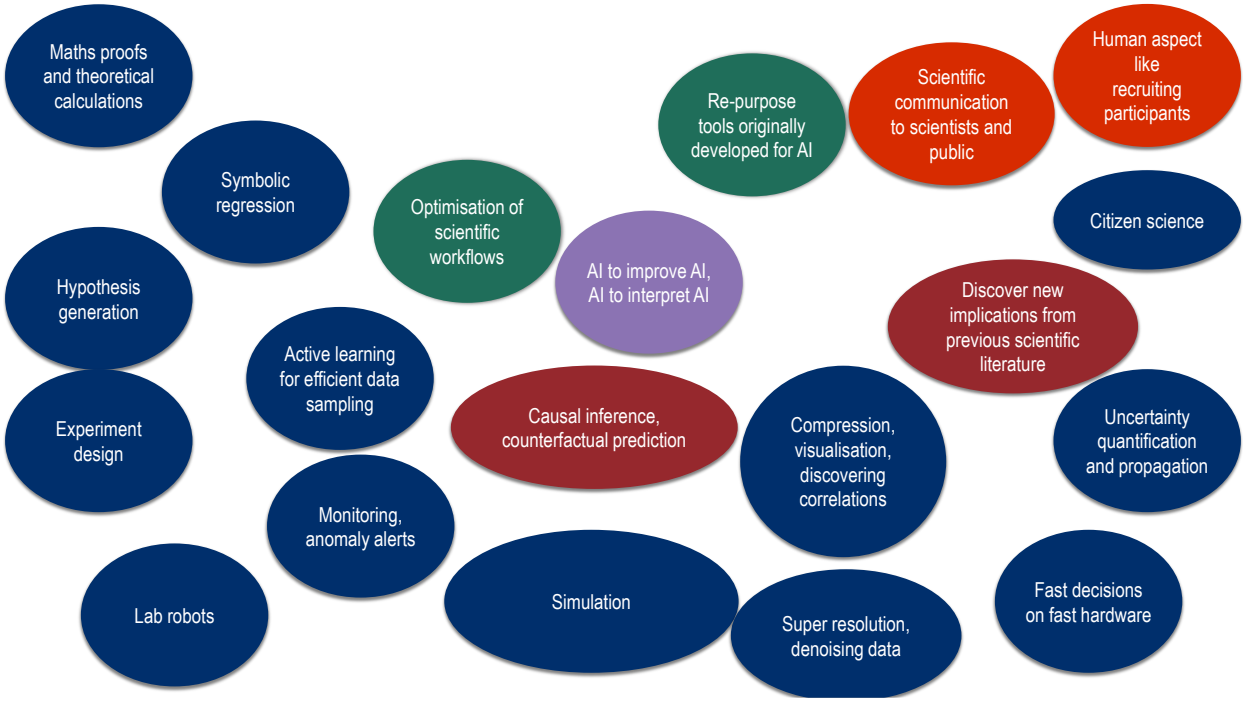
How can artificial intelligence help scientists? A (non-exhaustive) overview

A. Ghosh, Lawrence Berkeley National Laboratory, United States

Introduction

The diversity and ingenuity of ways in which artificial intelligence (AI) is helping scientists are sometimes shocking, even to domain experts. AI has already left a mark on every stage of the scientific process: from hypothesis generation and mathematical proof building to experiment design and monitoring, data collection, simulation and rapid inference, among others. Some intriguing cases coming up include AI helping to find new scientific insight from old scientific literature, simulate different teaching methodologies for education, write clearer scientific papers and even help with research on AI itself. This essay discusses the role that AI can play in science, with an eye on potential impacts in the near future. It touches upon some key challenges to overcome for AI to be more widely adopted in science, such as causal inference and the treatment of uncertainties.

Figure 1. A (non-exhaustive) plethora of AI uses in science



Note: Blue examples show where AI is directly used to improve a core aspect of the scientific progress; light red examples show uses that help set up studies or communicate results to peers or to the public. Green bubbles represent the benefit to science not from AI directly, but from the software and hardware infrastructure developed primarily for AI uses. Dark red bubbles refer to frontiers of AI for science. Violet signifies AI for AI research.

Scientists are a strange breed of professionals, one actually encouraged by the prospect of AI "taking away their jobs". The search for knowledge never ends: for every question that AI helps answer, scientists grow curious about many more. Once a discovery is made, one may seek a more fundamental understanding of why the finding is what it is. One may also want to know how to use this newfound knowledge to help humanity.

Researchers are intrigued by connections between seemingly unrelated disciplines of science. As science has become extremely specialised over the past century, research on how to use AI in science has formed a natural oasis for knowledge sharing and cross-disciplinary work. AI tools developed to create super-resolution images of celebrities, for example, actually find applications in materials science. Meanwhile, innovative applications in automated drug discovery have found parallels in AI for theoretical physics.

The transfer of technology between fields has never been quicker. This is because seemingly unrelated problems in different domains appear to have a unifying theme through the lens of AI applications (e.g. clustering of data, anomaly detection, visualisation and experiment design, regardless of scientific domain, have common characteristics). To date, AI has had a wide range of applications in different stages of the scientific process (Figure 1).

The most typical uses of AI in science

Supervised learning

The most typical uses of AI in science over the past decade have involved supervised learning, where a model is "trained" (optimised with an automatic algorithm) on data already annotated with the right answers. The data may have been painstakingly annotated by humans or come pre-annotated from simulations. AI might classify objects into some predefined set of categories like identifying Higgs bosons from the vast amount of particle collision data collected at the Large Hadron Collider (LHC). It could also regress some property of an object to, for example, predict the energy of a particle recorded in a detector from its image.

Once it learns the patterns from annotated data, AI can make predictions about new data where the correct answers are not already known. For instance, having learnt about what different household waste products look like from annotated images, AI could then assign the correct trash bin (recyclable vs. non-recyclable) for a new product that is not human annotated.

Anomaly detection

In "anomaly detection", AI aims to identify novel objects that look different from what the AI model is used to seeing. For example, it is difficult to have an exhaustive, annotated list of brain scan images spanning all the possible categories of abnormality. However, anomaly detection models need only see examples of healthy brains in their training to subsequently flag abnormalities in images of new patients. Such models do not require annotated training data.

The need for interpretability

As science demands interpretability, an opaque AI model that gives the right answers without any further explanation has limited use. For instance, anomaly detection models can highlight regions in medical images that are a cause for concern. This, in turn, points medical practitioners to regions for further investigation.

In fundamental physics, there is value in finding the simplest description of a phenomenon, often in the form of a concise, easy-to-understand formula. On the other hand, the power of deep learning comes from the ability to build enormous statistical models, often comprising millions of parameters. These are inherently difficult to interpret (see also the essay in this book by Hugh Cartwright on interpretability).

Graph neural networks and symbolic regression

In certain cases, physicists have found a way to use the power of deep learning, while retaining interpretability. In one instance, they do this with the help of graph neural networks. These can be designed so that individual components of the model describe specific physical attributes, such as the interaction between two celestial bodies.

Once the network has performed the more challenging task of learning these relationships directly from data, symbolic regression can be used to distil the information learnt by the network into an easy-to-interpret formula. This is a less powerful technique than deep learning, but it can automatically find simple formulas to describe data. Symbolic regression has been used recently to describe the concentration of dark matter from the mass distribution of nearby cosmic structures (Cranmer et al., 2020) with the help of an easy-to-understand formula.

Reinforcement learning

In the mathematical sciences, the need for interpretability might be greater still. Mathematicians would like to be able to say: “AI, please write the entire proof of this theorem, and remember to show every step of your work!” How would that work? High school students of calculus are well aware of how useful a single hint can be to arrive at a solution. Mastering all the integration tactics in the syllabus is not enough. There are too many tactics to try for a given problem.

The key to acing a calculus exam is to develop an intuition for what tactic might work in what kind of situation. This intuition develops through practice, or, in the case of AI, through training. Researchers have developed AI that can hint at the tactic most likely to work in each situation. This approach has been used to automate the formalisation of mathematical proofs. The AI suggests a tactic, a classical theorem prover implements it and together they get the job done.

An exciting form of AI in this field is reinforcement learning. This has gained much publicity recently for mastering the rules of chess, the game of Go and popular computer games, and then beating the best human players. Reinforced learning is excellent at learning a long sequence of actions to reach a desired goal.

In knot theory, for example, many open questions revolve around whether two knots can be considered equivalent, and whether one might be transformed into the other using a specific sequence of actions. If “yes”, reinforcement learning can often find the exact path from the first knot to the second, providing a clear proof of equivalence (Gukov et al., 2020). While these sorts of fully verifiable solutions are interesting, they are usually restricted to the mathematical sciences.

Treatment of uncertainties when using AI in science

When it comes to trusting the products of science, whether measurements based on recorded data or complex simulations that make simplifying assumptions, scientists care a lot about uncertainties. For a while, many scientists refrained from using AI because it was difficult to quantify the uncertainty in the results. Recently, however, the tide has turned as scientists have found that AI can help more accurately quantify uncertainties.

Uncertainty-aware networks

AI can keep track of multiple uncertainties that accumulate through long scientific pipelines, while traditional methods could only keep track of certain summary information about the uncertainties. AI can even help reduce such uncertainties, allowing scientists to make more confident measurements. For

instance, particle physicists have developed uncertainty-aware networks. These AI models are explicitly shown potential biases in data measurements when they are being trained. In this way, the model can automatically find the best way of handling every potential bias (Ghosh, Nachman and Whiteson, 2021).

The same technique has allowed astrophysicists to track uncertainties from high-dimensional telescope images (raw images at very high resolution, which usually require summarisation to apply traditional statistical techniques). The uncertainties can be tracked all the way to the final step of statistical inference, for instance, to deduce the nature of matter inside neutron stars from x-ray telescope images (Farrell et al., 2022).

This process makes it possible to have a comprehensive final measurement without leaving out vital information at intermediate steps. Such end-to-end models have grown in popularity. The quantification of a model's own uncertainty has an added benefit; it can then be used to acquire data more efficiently. Consider medicine, where a vast amount of data is often available but with only a small fraction of it labelled (because labelling data requires a lot of human labour). AI can help figure out which samples are the most important for humans to annotate.

Active learning

Active learning models can iteratively ask humans to annotate data points in such a way as to reduce their overall uncertainty about the data. For instance, having annotations for one image may allow the AI to learn a general pattern among similar images. In this case, asking for annotations for the first image is valuable, but subsequent similar images do not need annotation for the model to make accurate predictions about them.

An AI system that is more uncertain about some type of data indicates that there is presently less recorded knowledge about such data. Investing human time to label these uncertain data, then, will add more to recorded knowledge than spending the same resources labelling data for which the AI's uncertainties are already small. In one instance of drug discovery, a similar approach helped cut down the number of required experiments from the 20% of possible experiments needed for a traditional algorithm to 2.5% (Kangas, Naik and Murphy, 2014).

Beyond inference

The scientific process has many stages – from hypothesis generation, experiment design, monitoring and simulation all the way to publication. Until now, this essay has only discussed the use of AI in providing final results. However, AI is expected to contribute to every stage of science.

In drug discovery, for example, when there are too many possible chemical combinations to try, AI can narrow them down to the most promising options. In theoretical physics, if the researcher has a hunch that two kinds of mathematical tool might have some underlying equivalence, AI can help determine a correlation. This, in turn, encourages the mathematician to invest time to discover a rigorous mathematical connection. Another key component in science is simulation, and deep learning has had an enormous impact here.

Simulation in science with generative models

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 has been sensationally effective in handling such data to solve unusual tasks. It has made its way into popular culture through a variety of applications. One model, for example, uses a person's image and shows how that person may look in 30

years. As another example, GitHub Copilot writes entire blocks of code for a software developer based only on a description in plain English of what the code needs to do.

Generative AI

Models that can create new data in this manner are called “generative”. In science, such generative networks are used to simulate physical systems. Sometimes they can improve over the state-of-the-art traditional simulation algorithms in terms of accuracy. More often they are useful because they consume orders of magnitude fewer computing resources. In this way, they relieve scientists from the burden of creating specialised simulation algorithms for each physical process.

Generative AI models with similar structure can learn to simulate the evolution of the universe, certain biological processes and so on, making them a general-purpose tool. In another use case, generative models can remove noise or unwanted objects from data, for example, to un-blend images of galaxies.

A particularly exciting feature of such models is their ability to provide “super-resolution” data, that is, data with higher resolution than in the original recorded data. In materials science, for example, super-resolution models can correctly enhance cheaper, low-resolution electron microscopic images into high-resolution images that would otherwise have been more expensive to capture (Qian et al., 2020). The trick is to have the system view small areas in high resolution and compare those to the same areas in low resolution, and learn the differences. The system can then convert all areas in the low-resolution image – the entire field of view – into a high-resolution image. In the biological sciences, aside from saving money, this approach can also protect some of the objects of research. For instance, flashing high intensity light can help to image cellular structures but can also damage the specimen. Super-resolution techniques can circumvent this problem.

A curious reader may wonder at the power of these AI simulation models. Given some initial conditions, can one always train some model to simulate some system far into the future? Classical mechanics shows this should become increasingly difficult for chaotic systems. While AI does not magically circumvent this fundamental limitation, it can improve on previous best practice. This is what makes the use of AI in climate simulation (or simulations of any other chaotic system) fascinating. Naive applications of generative AI models may not succeed in accurately predicting weather patterns over a long period. However, Pathak et al. (2020) have shown that a hybrid simulation engine that combines the power of AI with fundamental physics computations can indeed predict such patterns.

In principle, the physics equations can be computed using a traditional algorithm (an algorithm handcrafted by scientists rather than learnt automatically by AI) to make accurate predictions. However, it is prohibitively expensive in terms of the computing resources required to run it at high resolution. The low-resolution version of the algorithm is cheaper but inaccurate. However, by enhancing the predictions of the cheaper algorithm with AI, these researchers achieved accurate simulations of weather over a long period. This combination is much cheaper to run than the high-resolution traditional algorithm. The key to this technique is to recursively make predictions across small time periods using the cheap solver, enhance the prediction with the AI model and then repeat for the next time step.

There is a general trend in incorporating domain knowledge into AI systems to help push the boundaries of what was once thought possible. In the above climate example, knowledge about the specialised scientific field was expressed in the physics equations used for the handcrafted traditional algorithm. The world will see many more innovations in weather and climate modelling with AI over the coming years, especially given the growing impact of climate change (such as increasingly erratic and extreme weather patterns).

Using AI to compress data

AI can also be used for data compression, finding ways to summarise the same information using fewer attributes. Consider a dataset full of 256x256 pixel images of circles. Instead of storing the value of every

pixel, one could store only the location of the circle and its radius, and still retain all the relevant information. This can make data storage and transmission memory efficient.

Recently, AI has also been used to compress multi-dimensional data into two dimensions (data summarised in two attributes) so it can be visualised on a screen or paper. The compressed representation of the data can itself reveal underlying but otherwise hard-to-detect patterns in the data. For instance, the compressed representation might show that certain data points clump into distinct clusters, which usually indicates some unifying characteristics in each cluster. If scientists can identify a characteristic that unites a cluster, they may also notice new data points in the cluster for which that characteristic has not yet been discovered. This could help scientists find items – from chemicals to materials to mathematical groups – with the desired characteristic. In one example, this line of study has created a growing interest in theoretical particle physics in how it could help find new theories that could describe the universe.

Compression also helps by producing resource-efficient algorithms. AI can be self-optimised in a way to find smaller models that can more easily be deployed on fast hardware for speed-critical applications such as at the LHC.

Indirect benefits of the deep-learning revolution to science

The advent of deep learning has also benefited science in indirect ways. It spurred development of software that automatically performs differential calculus, known as automatic differentiation (AD) software. It also contributed to the need for advanced parallel processing hardware like Graphics Processor Units (GPUs) and more efficient data storage technology. These developments have allowed scientists to replace older optimisation algorithms with AD, optimise complicated traditional algorithms and leverage the power of parallel programming. Increasingly ambitious efforts are also emerging to use the new optimisation algorithms for elaborate experiment design. Relying on open-source AD software reduces the burden on scientists to maintain their own software or to upgrade them to run on modern hardware such as GPUs.

AI-supporting scientific communication

Beyond the main stages of research, AI is also more broadly useful to science. In terms of communication, some AI models have been developed to summarise research papers (see also the essays in this book by Dunietz, and by Byun and Stuhlmüller), and a few popular Twitter bots regularly tweet these automated summaries. Certain AI models highlight aspects of a draft research paper that make it either easier or less easy to comprehend (Huang, 2018). For example, the model favours articles that contain conceptual diagrams early on, presumably to help guide the reader.

Recently, an AI-based method has been proposed to present experimental measurements in physics to theoretical physicists more effectively (Arratia et al., 2022). Using data from large experiments at CERN effectively often requires a team of physicists familiar with the detector. To combine results from multiple large experiments requires a specialised team, including physicists from each experiment. This is not feasible each time a new theory needs to be tested against data. Consequently, large experimental physics collaborations try to present their results in a way for a theorist to re-use easily. Traditionally, this has involved summarisation of the results at the expense of leaving out details. The newly proposed AI method to present experimental results allows theorists unfamiliar with the experiment to access the detailed measurements more easily to explore, combine and re-use measurements from multiple large experimental collaborations, such as ATLAS, CMS, LHCb (at CERN), Belle II (in Japan), and even cosmological observations. This would enhance the impact of each measurement by making the information more accessible to the larger scientific community.

These are examples of AI helping to better disseminate scientific results, even to subject-matter experts. In the future, AI-powered virtual or augmented reality is expected to help visualise and explore scientific concepts, from the structure of DNA to particle collisions at the LHC.

Robotics

Although AI today is mostly talked about in the context of digitised data, the use of AI-enhanced laboratory robots is growing (see the essay in this book by King, Peter and Courtney). Laboratory robotics can help automate precise repetitive tasks such as handling test tubes and cell cultures, among others, and avoid human exposure to harmful chemicals or radiation. Moreover, as other chapters in this book show, increasingly intelligent laboratory robots will have growing roles in experiment design and analysis. Curiosity, the Mars rover, is endeared to many. Future space and ocean exploration will see AI-powered robots in a multitude of applications.

Dangers and weaknesses of AI in science

The discussion in this essay has so far been optimistic. However, it would be an oversight to skip over the weaknesses of AI-powered research tools and the potential dangers of indiscriminate adoption.

Data-driven AI can malfunction

Data-driven 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. Validating the behaviour of the AI model under different circumstances therefore needs to be rigorous. There is ongoing work on developing AI that can be fully validated and where the maximum risk of failure can be quantified. However, significant innovation is needed before such models become useful for real-world tasks.

Efforts to mitigate bias may lead to further harm

Deep-learning models pick up subtle patterns in training data, including any biases in simulations. This is similar to how a model trained on some types of historical human data can learn social biases (such as sexism and discrimination against minorities). An often-discussed solution to this problem is to force a model's predictions to be de-correlated from protected features (e.g. race, gender, age). This means the AI would on average have a similar response regardless of an individual's race, gender and age. However, such attempts at de-correlation can actually lead to further unintended harm, especially when it is not easy to list all the sources of potential bias.

It is sometimes easier to demonstrate the unintended consequences of such bias mitigation techniques not on human data but on well-understood and fully controlled scientific data. For instance, in the context of particle physics, Ghosh and Nachman (2022) show that decorrelation techniques sometimes *hide biases* instead of getting rid of them. In some cases, the true bias is difficult or impossible to measure so physicists use proxy metrics to estimate it. For instance, when the exact mathematical computation of some theory cannot be done, they may use the best known approximation technique. To estimate the potential bias of this technique, they also compute approximations using a series of alternate techniques and treat the difference in the results as an estimate of the uncertainty.

In this physics study performed by Ghosh and Nachman (2022), the true bias in the model was found to be even greater after applying a debiasing solution that minimises the proxy metrics for bias. This led to

vastly underestimated uncertainties on the final measurement. It is therefore advisable to consider the possibility of such unintended consequences before attempting to de-correlate away biases in AI models.

Technological solutions cannot solve all problems

More generally, it should be considered carefully whether using the same metric to evaluate the performance of a model makes sense if it was already used for optimisation of that model. Further, sometimes a policy solution may be needed rather than a technological solution. For example, in theory, AI models could try to predict which students are more likely to succeed in science, technology, engineering and mathematics research careers. However, data are likely to be plagued by existing biases in society. A more effective solution to improving success may lie in better policies in terms of access to resource material, mentorship and creating inclusive work environments.

Causal models are needed to disentangle correlation from causation

AI models simply learn correlations in data, not the causal relationships involved. Causal models are needed to disentangle correlation from causation. For example, if a study indicates that levels of vitamin D in a population correlate with depression, does that mean one caused the other, or are they both simply symptoms of an (as yet unknown) underlying problem?

An interesting line of research in cognitive science focuses on human-AI interactions, which illustrates one way that AI can help shed light on causation. Researchers realised they can generate situations using AI that are difficult to create in real life, and then study their impact in the real world. For example, children in the United Kingdom interacted with an AI-driven virtual teacher who spoke first in a working-class British accent and then in the different accent of the real teacher. This allowed the researchers to study the impact of a teacher's accent on learning in children from diverse backgrounds. The ability to study these alternate situations is useful in establishing causation.

There is also growing interest in interfacing probabilistic programming (algorithms that account for the probabilistic nature of certain processes in science) with scientific simulators (such as particle physics simulators) to infer causation. These programs can run through a number of scenarios that might explain some observed data. The intersection of AI and causal inference is a nascent field, which has recently become a hot topic of research. Progress in this field will help accelerate progress in science.

Large AI models are expensive and more harmful for the environment

The trend has been to develop large AI models that require enormous computing resources to train. This can create problems for research groups with smaller budgets, particularly compared to large AI companies. Such models also leave a large carbon footprint that is harmful for the environment.

Innovation will be required to improve the resource efficiency of AI models. Besides this, governments may have to invest in computing resources that can be shared among research groups nationally. In the United States, a task force has already been set up to look into the feasibility of a National AI Research Resource (NAIRR, 2022).

Conclusion

The ways in which AI is accelerating science is growing rapidly. In certain cases, giant leaps in science made possible by AI have attracted public attention. The "AlphaFold" model (a deep-learning solution) made headlines, for example, by demonstrating an extraordinary ability to predict 3D-protein structures from their amino-acid sequence. Nonetheless, the potential impact of AI on science is a long way from being realised.

In this current "AI overhang", many innovations have potential, but there has not been enough time to explore them all. The last decade has seen a flurry of proof-of-concept innovations, but in the next one it will become common to incorporate AI into large scientific workflows. In some cases, such as at the LHC, automated workflows have already been established (Simko et al., 2021). The future may see scientific workflows optimised end-to-end – from data collection to final statistical analysis – using AI. The entire scientific process in certain cases – from hypothesis generation to the communication of scientific results – may also be fully automated.

Innovations in AI for science are often easy to transfer across different scientific domains. This has led to unifying approaches that cut across scientific disciplines. In simulation-based inference, scientific inference relies on the use of precise simulators to optimise some measurement. Differentiable programming optimises scientific workflows with software that performs automatic differential calculus. These and other unifying approaches like anomaly detection and generative models have re-energised the need for interdisciplinary experts.

Typical machine-learning models are difficult to interpret, but remain useful for tasks such as hypothesis generation, experiment monitoring and precision measurements. More interpretable models are useful for mathematical proof building. Generative models assist with tasks such as simulations, removing unwanted features from data and providing super-resolution data. Uncertainty-aware and uncertainty-quantifying models are extremely useful in providing trustworthy, reliable results. Such models can also help with efficient data acquisition by prioritising data acquisition in regions of uncertainty.

There are dangers in the indiscriminate use of AI because such models break down in unexpected ways. Therefore, it is important for AI experts to be vocal proponents of AI adoption and also caution against the unintended consequences of ill-informed applications. Future innovations may enhance interpretability and allow developers to provide more algorithmic restrictions on a mode to avoid catastrophic failures. These could occur, for example, if an AI system that controls scientific machinery were to behave erratically when it encounters a situation it has never experienced in training. The specific needs of science have fuelled interesting AI innovations, some of which have already found uses outside of science. As with other technologies developed for science, it is reasonable to expect an increasing number of innovations in AI for science to eventually benefit humanity in broader ways.

The future will likely bring growing use of AI-powered robots in laboratories and other sectors, such as space and the oceans, where scientific data are gathered. Innovations in developing causal inference models will provide huge benefits for the medical and social sciences. By accelerating science, innovations in AI are expected to help find solutions to global challenges such as clean energy generation and storage, improved climate models and treatments for disease.

References

Arratia, M. et al. (2022), "Publishing unbinned differential cross section results", *Journal of Instrumentation*, Vol. 17, https://iopscience.iop.org/article/10.1088/1748-0221/17/01/P01024.

Cranmer, M. et al. (2020), "Discovering symbolic models from deep learning with inductive biases", *arXiv*, arXiv:2006.11287 [cs.LG], arXiv:2006.11287v2.

Farrell, D. et al. (2022), "Deducing neutron star equation of state parameters directly from telescope spectra with uncertainty-aware machine learning", *arXiv*, arXiv:2209.02817 [astro-ph.HE], https://arxiv.org/abs/2209.02817.

Ghosh, A. and B. Nachman (2022), "A cautionary tale of decorrelating theory uncertainties", *The European Physical Journal C*, Vol. 82/46, https://link.springer.com/article/10.1140/epjc/s10052-022-10012-w.

Ghosh, A., B. Nachman and D. Whiteson (2021), "Uncertainty-aware machine learning for high energy physics", *Physical Review D*, Vol. 104/056206, https://journals.aps.org/prd/abstract/10.1103/PhysRevD.104.056026.

Gukov, S. et al. (2020), "Learning to unknot", *arXiv*, arXiv:2010.16263 [math.GT], https://arxiv.org/abs/2010.16263.

Huang, J.-B. (2018), "Deep paper gestalt", *arXiv*, arXiv:1812.08775 [cs.CV], https://arxiv.org/abs/1812.08775.

Kangas, J.D., A.W. Naik and R.F. Murphy (2014), *BMC Bioinformatics,* Vol. 15/143, "Efficient discovery of responses of proteins to compounds using active learning", https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-143.

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

Pathak, J. et al. (2020), "Using machine learning to augment coarse-grid computational fluid dynamics simulations", *arXiv*, arXiv:2010.00072 [physics.comp-ph], https://arxiv.org/abs/2010.00072.

Qian, Y. et al. (2020), "Effective super-resolution methods for paired electron microscopic images", *IEEE Transactions on Image Processing*, Vol. 29, pp. 7317-7330, https://ieeexplore.ieee.org/document/9117049.

Simko, T. et al. (2021), "Scalable declarative HEP analysis workflows for containerised compute clouds", 7 May, *Frontiers in Big Data*, https://doi.org/10.3389/fdata.2021.661501.

A framework for evaluating the AI-driven automation of science

R. King and H. Zenil, Cambridge University, United Kingdom

Introduction

The fundamental goal of science is to construct models that predict what will happen in the real world. This provides a natural objective function for artificial intelligence (AI) systems used in science to optimise how well they predict what happens in experiments. This essay looks at the future of AI-led science, presenting a roadmap of challenges for AI in scientific discovery and then proposing a framework for evaluating AI in science.

The traditional name for the application of AI to science is "discovery science", which dates to the 1960s and the work of Joshua Lederberg (Herzenberg, Rindfleisch and Herzenberget, 2008). Lederberg won the Nobel Prize in Physiology or Medicine but taught himself to program. He was very interested in AI and how to formalise science using logic. Lederberg's Meta-Dendral project was designed as part of the Viking probes to Mars (Klein et al., 1976); Mars was so distant that an automated system was needed to do the science there. While computer science and computers were not up to the task, this initiative turned out to be influential on AI. Also closely involved in Meta-Dendral was Ed Feigenbaum who won the Turing Award for his work on expert systems (Feigenbaum, 1992). Carl Djerassi was the main chemist, famous for his work on the birth control pill. The machine learning pioneer Bruce Buchanan was also involved.

Another highlight in the development of discovery science is the *Bacon* system. The driving force here was Herbert Simon, the only person to have won both a Nobel Prize and a Turing Award. It was claimed that *Bacon* rediscovered scientific laws such as Kepler's laws of planetary motion (Qin and Simon, 1990). This was controversial because the data were in effect cleaned up, and *Bacon* fitted equations to these clean data; this was very different from what Kepler had to do in reality. Nevertheless, *Bacon* was an important and influential project.

Science is a well-suited task for AI systems (Kitano, 2021). Science is abstract like the games of chess and Go, where AI systems can beat the best human players. Scientific problems are also restricted in scope. If an AI system is working on a scientific problem, it does not need to know about vegetables or politics or anything else. It just needs to know about the scientific domain in question. Moreover, nature is honest. Whether a human or a robot does a scientific experiment, for instance, the real world can be trusted; it is not trying to fool us about how nature works. This is quite different from AI systems in business or war, where many agents are dishonest, in nature or by design.

Much of the AI developed for science to date has limitations (Castelvecchi, 2016). There are many examples of "black-box" applications of AI in science which lead to little understanding. AlphaFold (and AlphaFold 2), for example, produced impressive results on the problem of protein folding. However, they did not generate much understanding of the underlying mechanisms of protein folding (Pinheiro et al., 2021; David et al., 2022). Most current AI applications to science have only contributed indirectly to

understanding; domain experts, in fact, do the crucial prior work (Hedlund and Persson, 2022). This may have negative consequences. For example, if the problem of protein folding is considered solved, funding for basic science in molecular dynamics may be reduced. This problem is compounded by the fact that measuring scientific progress is both difficult and often controversial (Zenil and King, forthcoming).

The future of AI-led science

How much knowledge is gained by applying AI to science? This is an important aspect of measuring scientific progress. The issue is not about the speed of discovery or how this may be increased but rather about how much is actually gained. This is difficult to know because there are no universally accepted criteria for gauging originality or relevance of a scientific discovery. Nobel Prizes, for example, reward contributions to science of the highest order, and yet bias and subjectivity are at play in the award process.

In this section, the essay explores issues that related papers do not seem to address, particularly the inclusion of lab automation in a closed-loop experimental cycle led by AI, and its full implications. AI is making important contributions to many aspects of science. However, there are few examples of AI being used to complete the experimental cycle (King et al., 2018), or to fully automate it from beginning to end in a generalised manner.

Overcoming bottlenecks

One of the main bottlenecks in the automation of science is the question of knowledge extraction and knowledge representation, both in a domain-specific and a general sense (Ataeva et al., 2020). If an artificial general intelligence (AGI) existed, it could extract and represent knowledge from all domains. AGI is yet to happen, though it likely will in time.

Currently, even the most automated systems are usually given a hypothesis to test. In other words, the best available AI today cannot enable systems capable of defining their own hypothesis space and own experiment design. Some forms of AI-driven instrument automation for experimental acceleration do exist and have proven fruitful (King et al., 2018; Frueh, 2021). These closed systems require contextual rejection, validation and verification of hypotheses and models. Ideally, they should be capable, just as humans are, of interpreting accidents as a source of inspiration and innovation.

Several groups have done research on “computational serendipity” (Niu and Abbas, 2017; Abbas and Niu, 2019). Such research may prove to be essential to reproduce human scientific practice. It may even improve its efficiency (negative results are often transformed into positive discoveries under a different context). In a closed-loop approach to scientific discovery, and as part of its definition, it is important to consider how to really close the experimental loop. A system should incorporate a result into a knowledge database and consider it for the next iteration of the discovery cycle.

AI-led closed-looped automation is arguably the future of science. In other words, in future, a robot scientist (AI scientist, self-driving lab or AI robotic system) will do simple forms of scientific research autonomously. Such a system has background knowledge about an area of science, which it represents in the best way possible – through logic and probability theory.

A robot scientist can autonomously discover and form a new hypothesis about the area of science in question. It can also autonomously identify efficient experiments to test these hypotheses. It may then control and program a laboratory robot to physically execute experiments.

The robotic system can examine the results of these experiments, analyse them and change the probabilities of hypotheses being correct based on the experimental observations. It can then repeat the cycle until some resources run out or only one theory is consistent with the background knowledge and

experimental evidence. Such robotic systems area already accelerating science in genetics and drug discovery (King et al., 2019; 2018; Frueh, 2021).

Motivations for building robot scientists

The motivations for building robot scientists are both epistemological and technological (King et al., 2018).

The epistemological motivation is to better understand how science works. If one can create an engine that can do human-like science, then this is informative about how the practice of human science may work. As was written on Richard Feynman's blackboard at the time of his death, "what I cannot create I do not understand".

The technological motivation is to increase the efficiency and quality of science. Robot scientists can work faster, more cheaply, more accurately and for longer than human beings: 24 hours a day, 7 days a week (King et al., 2018). Robot scientists can also be more easily multiplied than human scientists. If one robot scientist can be built, then thousands, even millions, could also be quickly built.

The science produced by robot scientists is expected to be of higher quality and more reproducible than that of human scientists. With robot scientists, the whole scientific cycle is semantically explicit, and potentially also declarative (i.e. expressing the logic of computations). Robot scientists also record experiments in much greater detail than is possible for most human scientists. This makes the experiments more likely to be reproducible in other laboratories.

Finally, robot scientists are more robust to pandemics than human scientists. For example, a "robot chemist" at Liverpool University made headlines in the United Kingdom by working through the COVID pandemic (Burger et al., 2020).

Evidence for the feasibility of the hugely ambitious project of building such robot systems comes from the success of AIs at playing the most intellectual of games (chess, Go, poker, etc.), and the analogy between such games and science. In chess and Go, for example, there is a continuum of playing ability ranging from novices to grand masters. Over AI history, AI game-playing programs repeated this path: they began playing poorly and went on to easily beat the human world champions (Strogatz, 2018).

The analogy between games and science suggests that AI systems for science may follow the same trajectory as game-playing systems. They may move from simple forms of science that existing autonomous systems can do through to science that average human scientists can do, and end as "Grand Masters" of science (Newton, Darwin, Einstein). If one accepts a continuum of ability in science, then AI systems will likely get better and better at science as hardware and AI software improve and more data become available. Indeed, 10 years ago, physics Nobel laureate Frank Wilczek predicted that in 100 years the best physicist would be a machine (Wilczek, 2016).

Achieving success in the Turing Challenge requires overcoming huge technical challenges. AI systems would need the capacity to:

- make a strategic choice about its research goals
- form exciting and novel hypotheses that move beyond a restricted area
- design novel protocols and experiments to test hypotheses beyond use of prototypical experiments
- notice and characterise a significant discovery in terms that human scientists can comprehend.

A roadmap of challenges for AI in scientific discovery

Is it necessary to solve the problem of general-purpose AI (GPAI) to develop AI systems that can do Nobel Prize-level scientific research? It was once widely believed that building machines able to beat the world

chess champion would require GPAI. Indeed, that was the main motivation for studying computer chess. GPAI turned out to be unnecessary, and it was possible to build machines that are world class at chess/Go but able to do nothing else intelligent. Any future AI system that could do Nobel Prize- quality research would certainly have to be much more general and intelligent than chess/Go playing machines. However, it is not clear if they need to be as general and intelligent as a GPAI.

Achieving the Nobel Turing Challenge would have profound effects on almost everything. Modern society is built on a foundation of science and technology. Most people in developed countries now live better than kings did in the past: they have better food, medical care, transport, etc. This miracle has been made possible through better technology based on science. Success in the Nobel Turing Challenge would result in almost unlimited amounts of new science and technology. This power could then be used for the benefit of all the world's inhabitants, human and non-human.

The problem of communicating scientific results has to do with language-based modelling. Future milestones may range from generating summaries of scientific articles to producing a critique of a whole scientific field. In other words, AI will perhaps be able to pinpoint where humans have been biased or else highlight areas of a domain that humans have failed to explore. If AI can explore a full hypothesis space, and even enlarge the space itself, then it may show that humans have only been exploring small, delimited areas of the hypothesis space, perhaps as a result of their own scientific biases.

Explorations of regions of science could be encouraged that are neither entirely favoured for attention by humans nor random. Instead, they would be AI-led or a hybrid of human-guided and computer- proposed (e.g. as in assisted theorem proving). It could be that areas humans have chosen to explore are the only ones relevant to human challenges. Over time, it may become apparent that humans have neglected areas of discovery that could have positive social impacts.

Reasoning needs to be able to move from capabilities such as generating scientific questions to passing any open-ended scientific or professional exam. This kind of investigation should also be able to explore some of the algorithms in use. This need not be in great depth, but it must at least be aware of broad categories of algorithms. Statistical data-driven approaches (Kim et al., 2020) dominate the current AI and machine-learning scene. Consequently, model-driven approaches (Shlezinger et al., 2021) would be a small subset of the broader categories.

As with many such areas of science, much remains speculative. However, a combination of methods will probably help achieve the relevant goals. This will include approaches similar to what currently happens in hybrid human-AI interaction but doing so more explicitly (Maadi et al., 2021). Some researchers have also found the best results may come from combining the best of both worlds. Statistically data-driven approaches, such as most of the deep-learning space, could go together with methods based on cognitive or symbolic computing (Pisano et al., 2020). Among other things, this would permit a better handling of aspects of causality (Zenil et al., 2019).

Levels of automation in science

Measuring acceleration or deceleration of progress in science is difficult because each is likely to be highly domain- and method-dependent. Imagine an attempt to quantify the acceleration of research in a single scientific domain. Obviously, there could be different methods for addressing the problem, and each could find a different rate of acceleration. Unfortunately, today, there is no universally agreed-upon way to measure progress or productivity in science.

Assessing the contribution of AI to science and the evolution of science towards full automation requires identifying and advancing measures for evaluating progress. The Society of Automotive Engineers, for example, developed a classification from Zero to Five to assess progressive degrees of autonomy in cars.

One way to think of these levels is how much human input the car requires to navigate. The higher the level, the less human input is required. Thus, Zero signifies no automation: these are regular cars where the human driver is in charge of every aspect of driving. Level One signifies driver assistance. Level Two involves automated steering and acceleration. Some may consider having a GPS and other such aides as amounting to partial automation, but this combines automation and human effort, comparable to the process of acceleration. The same could be said of cruise control, which requires human intervention. At any rate, autonomy between One and Two signifies assistance plus automation (Badue et al., 2021). In Level Three, some responsibility of driving is transferred to an AI system. Level Five is full automation (with no human intervention). This remains elusive, despite the hype of the last decade. Level Four is Level Five but restricted in scope, e.g. in time and space, or to certain situations.

A proposal for evaluating AI in science

This essay proposes a similar scheme for evaluating AI in science, since this too involves a transfer of responsibility from humans to machines. It too is about progressively consigning aspects of the scientific endeavour to machines, until humans are no longer involved. Any adopted classification needs to be useful, understandable, specifically measurable, achievable, relevant and robust. In other words, assigning a level must be easy and classification would not require constant updating.

The proposed classification has an associated staged process that the automation of science by machines and AI might follow. However, the framework is itself a work in progress.

Level Zero

Level Zero is simple because it designates the absence of automation in science. Most traditional human science, before the advent of computers, belongs here. It is led, driven and undertaken by human minds.

Level One

In Level One, human scientists still describe the problem in full, but machines do some data manipulation or calculation. Some commentators date Level One, machine assistance, to the beginning of the last century. Others trace it to the advent of data science in the 1980s, or even to the 1990s with the emergence of statistical machine learning. A case might also be made for dating the achievement of Level One to the 1950s and 1960s, when the first theorem provers appeared (Harrison, Urban and Wiedijk, 2014).

Level Two

Level Two would signify that an important aspect of the discovery cycle is fully automated. This could include, for example, the simulation or extraction of knowledge, or the testing of propositions. This means that humans are still required for some of the most important aspects of the full experimental cycle but that at least one pathway has been fully automated. For example, some AI systems are able to read databases and provide this input to another human or to a machine system.

Level Three

Level Three would signify a state where AI can perform model selection and generation (Hecht, 2018). This would be equivalent to having a knowledgeable system appear, with some agency, which could receive a set of hypotheses and then follow the consequences. For example, a scientist might be able to provide the system with a selection of problems and data, and the AI would then match them to provide a solution. In this case, human scientists are still giving the AI the hypothesis and solution spaces.

Theorem proving may belong to Level Three, being quite advanced in some ways. Still, today's theorem provers are also limited because they do not learn over time; they are deterministic. In other words, they

start from scratch every time, unless humans add new knowledge to the theorem database. In a few systems, this may have been automated, but some level of human curation still needed.

Level Four

Level Four would entail closing the loop – as it were – with AI being able to generate and explore the hypothesis space. However, at least one aspect of the discovery cycle would not be fully automated: humans would still need to feed the AI system with all the initial information and data it needs. A Level Four system, for example, would be a theorem prover. Without new data inputs after the initial cycle of analysis, the theorem prover could continue to explore a hypothesis space to generate new theorems without human intervention.

Level Five

Level Five corresponds to full automation, covering all levels of discovery and with no human intervention. An automated system operating at Level Five will be equivalent, if not superior, to a human scientist. This type of system would not require any human input. What follows are examples illustrating the various levels.

The state of automation

Level One is perhaps the state of the art today, with one or two processes entrusted to machines, data science, data analytics, etc. The Kepler space telescope, for example, generates a lot of data. Use of a computer to analyse the data is needed to extract all the information about exoplanets embedded in the data. Given the sheer quantity of data and the weakness of the signals, little to nothing might be accomplished without computers.

Machine learning arguably fits Level Two. Weather forecasting, which is very much based on dynamic systems, would be a good example of more physics-driven (Chowdhury and Subramani, 2020) and model-driven approaches. With weather forecasting, a physical representation requires little human intervention because obviously these weather sensors put out information almost in real time. Nearly the entire process has been automated, except for model creation. The model is already determined and of human devising. Humans generate a model and implement it and then let the system do the simulation and ingest the data, almost all of this occurring in real time.

The placement of AlphaFold 2 at either Level One, Two or Three is an open question. Level Three seems inappropriate because it would designate something more like auto machine learning (He et al., 2021), which is about trying to pick the model that best fits the observations. Auto machine learning is in an early stage of development and is perhaps also domain-specific. The authors believe that capabilities in science are moving towards Level Three where AIs can choose the best model for the data.

Only the robot scientists may be said to have reached Level Four. This is the stage where science, especially experimental science, can be greatly accelerated. For such machines, this involves almost no human intervention except for providing consumables.

Participants at the first workshop on the Nobel Turing Challenge, organised by the Alan Turing Institute in 2020, estimated that widespread uptake of Level Two and Level Three systems will happen within the next five years. They considered that Level Four systems could become widespread in the next 10-15 years, and Level Five in the next 20-30 years. Indeed, a fully automated experiment recently tested systematic research reproducibility from literature papers for the first time (Roper et al., 2022). It shows higher Levels (4-5) are becoming possible. If the estimates of the experts cited here are even broadly correct, then science will shortly be transformed.

Conclusion

This essay argues that the future of science lies in AI-led closed-looped automation systems. These run the full scientific cycle autonomously, iterating continuously from hypothesis generation to experimental validation and re-interpretation of results. These systems will emulate the human scientific process but work faster and more precisely. They will be less biased and able to open up ever-larger regions of scientific discovery. To achieve this requires well-defined key performance indicators grounded in a framework of automation levels based on the quantity and quality of input and execution required from human scientists. Human scientists will decide how to work with the AI scientists, and how much room AI will have to define its own problems and solutions.

References

Abbas, F. and X. Niu (2019), "Computational serendipitous recommender system frameworks: A literature survey", in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1-8, https://ieeexplore.ieee.org/abstract/document/9035339

Ataeva, O. et al.(2020), "Ontological approach: Knowledge representation and knowledge extraction", *Lobachevskii Journal of Mathematics*, Vol. 41/10, pp. 1938-1948, www.azooov.ru/index.php/ljm/issue/view/81.

Badue, C. et al.(2021), "Self-driving cars: A survey", *Expert Systems with Applications,* Vol. 165/113816, https://doi.org/10.1016/j.eswa.2020.113816.

Burger, B. et al.(2020), "A mobile robotic chemist", *Nature*, Vol. 583, pp. 224-237, https://doi.org/10.1038/s41586-020-2442-2.

Castelvecchi, D. (2016), "Can we open the black box of AI?", *Nature News,* 5 October, Vol. 538/7623, www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

Chowdhury. R. and D.N. Subramani (2020), "Physics-driven machine learning for time-optimal path planning in stochastic dynamic flows", in *International Conference on Dynamic Data Driven Application Systems*, pp. 293-301, https://dl.acm.org/doi/abs/10.1007/978-3-030-61725-7_34.

David, S. et al.(2022), "The alphafold database of protein structures: A biologist's guide", *Journal of Molecular Biology*, Vol. 434/2, p. 167336, https://doi.org/10.1016/j.jmb.2021.167336.

Feigenbaum, E.A. (1992), "A personal view of expert systems: Looking back and looking ahead", Knowledge Systems Laboratory, Department of Computer Science, Stanford, https://stacks.stanford.edu/file/druid:dp864rk0005/dp864rk0005.pdf.

Frueh, A. (2021), "Inventorship in the age of artificial intelligence", *SSRN,* https://dx.doi.org/10.2139/ssrn.3664637.

Harrison, J., J. Urban and F. Wiedijk (2014), "History of interactive theorem proving", *Computational Logic*, Vol. 9, pp. 135-214, www.cl.cam.ac.uk/~jrh13/papers/joerg.pdf.

He, X. et al.(2021), "Automl: A survey of the state-of-the-art", *arXiv,* arXiv:1908.00709 [cs.LG], https://doi.org/10.1016/j.knosys.2020.106622.

Hecht, J. (2018), "Lidar for self-driving cars", *Optics and Photonics News*, Vol. 29/1, pp. 26-33, https://doi.org/10.1364/OPN.29.1.000026.

Hedlund, M. and E. Persson (2022), "Expert responsibility in AI development", *AI and Society*, https://doi.org/10.1007/s00146-022-01498-9.

Herzenberg, L., T. Rindfleisch and L. Herzenberget (2008), *The Stanford Years* (1958-1978), *Annual Review of Genetics,* Vol. 42, pp. 19-25, https://doi.org/10.1146/annurev.genet.072408.095841.

Kim, Y. and M. Chung (2019). "An approach to hyperparameter optimization for the objective function in machine learning", *Electronics*, Vol. 8/11, pp. 1267-2019, http://dx.doi.org/10.3390/electronics8111267.

Kim, H. et al. (2020), "Artificial intelligence in drug discovery: A comprehensive review of data-driven and machine learning approaches", *Biotechnology and Bioprocess Engineering*, Vol. 25/6, pp. 895-930, https://doi.org/10.1007/s12257-020-0049-y.

King, R.D. et al. (2018), "Automating sciences: Philosophical and social dimensions", *IEEE Technology and Society Magazine*, Vol. 37/1, pp. 40-46, http://doi.org/10.1109/MTS.2018.2795097.

Kitano, H. (2021), "Nobel Turing Challenge: Creating the engine for scientific discovery", *NPJ Systems Biology and Applications*, Vol. 7/1, pp. 1-12, https://doi.org/10.1038/s41540-021-00189-3.

Klein, H.P. et al. (1976), "The Viking mission search for life on Mars", *Nature*, Vol. 262/5563, pp. 24-27, https://doi.org/10.1038/262024a0.

Maadi, M. et al. (2021), "A review on human–AI interaction in machine learning and insights for medical applications", *International Journal of Environmental Research and Public Health,* Vol. 18/4, https://doi.org/10.3390/ijerph18042121.

Niu, X. and F. Abbas (2017), "A framework for computational serendipity", in *Adjunct Publication of the 25th Conference* on *User Modeling, Adaptation and Personalization*, Association for Computing Machinery, New York, pp. 360-363, https://doi.org/10.1145/3099023.3099097.

Pisano, G. et al. (2020), "Neuro-symbolic computation for XAI: Towards a unified model", in *WOA*, Vol. 1613, pp. 101-117, https://ceur-ws.org/Vol-2706/paper18.pdf.

Pinheiro, F. et al. (2021), "Alphafold and the amyloid landscape", *Journal of Molecular Biology*, Vol. 433/20, pp. 167059, https://doi.org/10.1016/j.jmb.2021.167059.

Qin, Y. and H.A. Simon (1990), "Laboratory replication of scientific discovery processes", *Cognitive Science*, Vol. 14/2, pp. 281-312, https://doi.org/10.1016/0364-0213(90)90005-H.

Roper, K. et al. (2022), "Testing the reproducibility and robustness of the cancer biology literature by robot," *Royal Society Interface,* Vol. 19/189, http://dx.doi.org/10.1098/rsif.2021.0821.

Shlezinger, N. et al. (2021), "Model-based deep learning: Key approaches and design guidelines" in *2021 IEEE Data Science and Learning Workshop (DSLW)*, pp. 1-6, https://arxiv.org/pdf/2012.08405.pdf.

Strogatz, S. (2018), "One giant step for a chess-playing machine". 26 December, *New York Times*, pp. 1-6, www.nytimes.com/2018/12/26/science/chess-artificial-intelligence.html.

Wilczek, F. (2016), "Physics in 100 years", *Physics Today*, Vol. 69/4, pp. 32-39, https://doi.org/10.1063/PT.3.3137.

Zenil, H. and R. King (forthcoming), "The far future of AI in scientific discovery", in *AI For Science*, Choudhary F. and T. Hey (eds.), World Scientific Publishing Company/Imperial College Press.

Zenil, H. et al. (2019), "Causal deconvolution by algorithmic generative models", *Nature Machine Intelligence*, Vol. 1/1, pp. 58-66, https://doi.org/10.1038/s42256-018-0005-0.

Using machine learning to verify scientific claims

L.L. Wang, University of Washington, United States

Introduction

The verification of scientific claims – also known as scientific fact-checking – is an important application area for machine learning (ML) and natural language processing (NLP). This essay explores the current state and limitations of ML systems for scientific claim verification. It begins with some background and motivation for the task, followed by an overview of technological progress and future directions.

There is a sense of renewed urgency around automated methods for claim verification. This has been driven by the abundance of misinformation spread on line during the COVID-19 pandemic, as well as in relation to sensitive topics such as climate change. Indeed, during the COVID-19 pandemic, there were reports of conflicting findings in the literature and early preprints with results that were subsequently disproved. In addition, high-profile retractions found avid uptake by news and social media organisations (Abritis, Marcus and Oransky, 2020).

To combat misinformation, platforms like Twitter, Facebook and others engage in both manual and automated fact-checking. These companies may employ teams of fact-checkers to search for and validate uncertain claims. At the same time, they deploy ML models to identify check-worthy claims, retrieve relevant evidence or predict factuality to different degrees of accuracy and success (Guo, Schlichtkrull and Vlachos, 2022).

Manual fact-checking is laborious and resource-intensive, and difficult to scale to the growing size of content on social media. Science faces a similar rapid growth in output. Millions of papers are written annually, and hundreds published each day in notable and sometimes contentious areas such as COVID-19 and climate change.

Additionally, scientific claims pose a unique set of challenges for fact-checking. This is due to the abundance of specialised terminology, the need for domain-specific knowledge and the inherent uncertainty of scientific findings. In other words, results can go through a long process of theory, experimentation, validation and replication before being accepted as scientific canon. Indeed, some claims may remain contentious due to small effect sizes and difficulties obtaining measurements from humans at a population level. Nonetheless, automated ML-based methods for claim verification are desirable and valuable. They are needed to assist and reduce efforts for human fact- checkers and improve the coverage of fact-checking systems.

Automated scientific claim verification has made significant advances in recent years due to progress in NLP methods. This includes the introduction of pretrained language models, and task-specific gains through the release of new datasets, models and applications to support the study of scientific claim verification. Though results are promising, several key challenges remain:

1. Scientific discourse does not lend itself easily to claim verification.
2. Claim verification methods suitable for the news or political domains may not be appropriate in the scientific domain.

3. Research systems for scientific claim verification do not yet tackle a realistic version of the problem.
4. The social implications of automated claim verification for science are unclear (i.e. what are the desired results from applying fact-checking methods to scientific discourse in social media and elsewhere?).

Few works on automated scientific claim verification engage deeply with the social issues or consequences of such automation. Do these models help assist or replace manual fact-checking? Or are they built to increase scientific literacy and the ability of lay people to engage in scientific discourse? One would expect the outputs of models serving these two goals to be quite different.

Similarly, many questions remain around how to integrate the outputs of claim verification models with the decisions of human fact-checkers. The focus on modelling progress is justifiable given the current technological state and limitations of automated scientific claim verification systems. However, as model performance improves and prototype systems are deployed, the possible social implications of these developments must be addressed.

Background

The task of scientific claim verification begins with a claim. The claim should be a statement about some entity or process, and it must be verifiable – it should not be a statement of opinion. Additionally, some definitions require the claim to be atomic (about a single aspect of the entity or process), decontextualised (able to be understood on its own without additional context) and check-worthy (to confirm its veracity for a target audience) (e.g. Wadden et al., 2020).

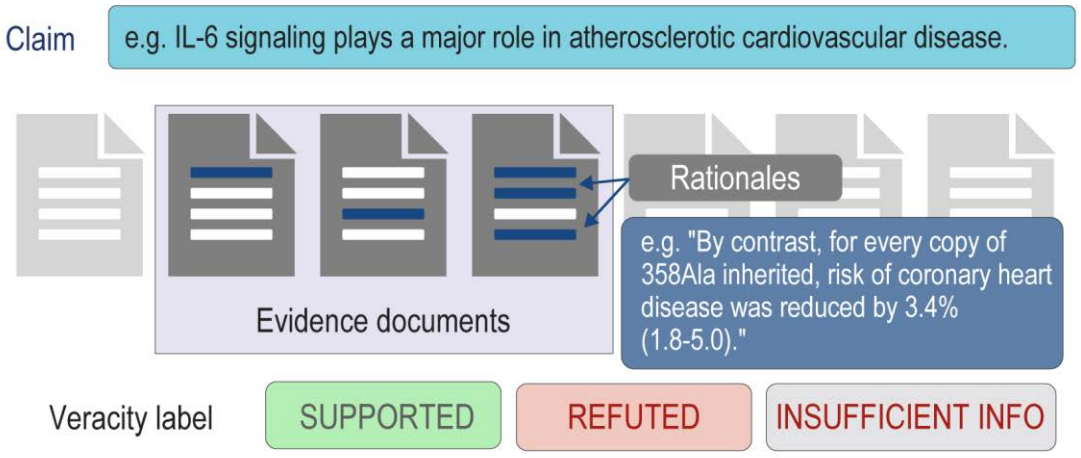
Given a valid claim, the goal is to predict its veracity. Is the claim supported or refuted by the evidence? Is there insufficient information to make a prediction? The model's prediction is known as the "veracity label".

In many cases, the task also requires identifying evidence from trusted sources to support the veracity label. Documents providing evidence towards or countering the prediction are referred to as "evidence documents". Specific spans of text from the evidence documents that support or refute the claim can be provided optionally as "rationales" towards the decision. Figure 1 shows how these components relate to one another with an example claim and its associated evidence that is identified from one of a set of scientific documents.

Pretrained contextual language models lie at the foundation of many state-of-the-art systems for natural language understanding; claim verification as a task is no different. These language models are pretrained on a large amount of unlabelled text in a self-supervised manner, allowing the model representations to capture the meaning and relationships between words. The models are then adapted to various downstream tasks, usually by fine-tuning on a small, labelled dataset specific to the task. Pretrained language models can and have been adapted to perform fact-verification in this manner. For example, they have been fine-tuned on datasets such as FEVER (Thorne et al., 2018) to produce a general domain fact-checker. They have also been fine-tuned on SciFact (Wadden et al., 2020) to produce a fact-checking model adapted for scientific claims. In addition to using textual evidence, some fact-checking models also investigate how source metadata and other information can be used to improve veracity predictions.

Claim verification in science faces several unique challenges. Scientific text contains an abundance of specialised terminology, which can be challenging for language models if the terms are rarely observed in pretraining data. Readers are also assumed to have the background to understand text in various domains such as anatomy and physiology and the functional pathways of various tissues, as well as common acronyms to understand typical sentences in scientific literature. Finally, scientific claims are not clearly true or false. Science, as a process, is designed to help us arrive at increasing certainty through iteration on hypotheses and controlled experiments.

Figure 1. Components of the scientific claim verification task



During this process, each result may only provide limited evidence towards a claim. Contradictory evidence is prevalent, as observed in DeYoung et al. (2021). Consequently, positing this task as claim verification rather than fact-checking casts the goal as identifying evidence to both support and refute the claim. In other words, it is not about making a summative judgement on the truth or falsehood of a particular claim. Given the uncertainties of many scientific outcomes, this is a pragmatic choice. It allows outputs of these models to be less brittle and more suitable for consumption by human fact-checkers and downstream users.

Current state of development

Automated tools can help researchers and the public evaluate the veracity of scientific claims. In recent years, automated fact-verification has advanced significantly in the domain of news, politics and social media. A number of datasets and shared tasks (FEVER; CheckThat!) have been created to support research in these areas. Several shared tasks address fact-verification in science, such as the TREC Health Misinformation Track and the SciVer Scientific Claim Verification. These have helped move the state of the field forward.

Scientific claim verification has received more attention in the last couple of years due to misinformation and disinformation related to COVID-19. However, collecting labelled data at scale for training ML models remains a challenge. Datasets like FEVER use large bodies of crowdsourced factual knowledge – articles on Wikipedia – to produce training data at scale. FEVER consists of many hundreds of thousands of instances, describing claims about a similar order of magnitude of entities. The largest scientific fact-checking datasets released to date are on the order of thousands or tens of thousands of claims and paired evidence documents (see Table 1 for a comparison).

Datasets for claim verification

Datasets are more difficult to construct in the scientific domain, requiring domain expertise to identify or write claims, and classify evidence. By way of illustration, two claim verification datasets in the scientific and health domains and their construction procedure are described here. Others are referenced in Table 1, where readers can also find references to more detailed information.

SciFact

Citation sentences from biomedical papers were rewritten into claims by a group of trained expert annotators. These claims were verified against the cited evidence articles by a different group of annotators. Refuted or negative claims were created by manually negating some of the written claims. The dataset consists of 1 409 claims verified against over 5 000 scientific paper abstracts, along with rationale sentences identified from evidence documents.

COVID-Fact

This derives claims and evidence from the r/COVID19 subreddit. Claims are verified against the text of linked scientific papers and against documents retrieved through Google Search. Claim negations are created automatically by detecting and replacing salient entity spans in the original claim. COVID-Fact contains naturally occurring claims written by their original authors, which are often complex, describing more than one facet of an entity or process. The dataset consists of 4 086 claims and their associated evidence documents on the subject of COVID-19.

Expert annotation

Unlike FEVER, where crowdsourced annotations were used to construct the claim and evidence dataset, SciFact, COVID-Fact and other scientific claim verification datasets required expert annotators. Expert annotation has been employed for components of dataset construction such as claim extraction, claim rewriting, claim negation, evidence classification, rationale extraction, explanation writing and/or veracity labelling.

In some cases (Saakyan, Chakrabarty and Muresan, 2021), little to no expert rewriting of claims takes place. However, the natural claims in these cases tend to be complex. It can be difficult to evaluate model performance (e.g. how to award credit if evidence provides support for only part of the claim).

Automatically derived claims

Manually writing claims and claim negations is a laborious process that can introduce biases into the data. An emerging trend in dataset construction is exploring techniques for automatically deriving claims and evidence from documents for training without labelled data. One example is the automatic production of claim negations (Wright and Augenstein, 2020; Saakyan, Chakrabarty and Muresan, 2021), which are needed to train fact-verification models. Table 1 compares datasets for scientific claim verification using FEVER as a reference for general domain fact-checking.

Table 1. A comparison of datasets for scientific claim verification

Dataset	Domain	Size	Description
FEVER (Thorne et al., 2018)	Wikipedia	185 000 claims	Claims and evidence from Wikipedia with crowdsourced annotations
SciFact (Wadden et al., 2020)	Biology, Medicine	1 409 claims	Claims rewritten by annotators from scientific papers; evidence manually curated by annotators from papers; negative claims manually generated
PubHealth (Kotonya and Toni, 2020)	Public health	11 832 claims	Claims from fact-checking and news websites; evidence and explanations from journalists
Climate-FEVER (Diggelmann et al., 2020)	Climate change	1 535 claims	Claims from web search, evidence from Wikipedia
COVID-Fact (Saakyan, Chakrabarty and Muresan, 2021)	COVID-19	4 086 claims	Claims and evidence from COVID-19 subreddit; negative claims automatically generated
HealthVer (Sarrouti, Ben Abacha and Mrabet, 2021)	COVID-19	14 330 claims	Claims from web search, verified by expert annotators against retrieved scientific articles

Model performance for claim verification

System performance is improving rapidly. However, more real-world case studies are needed to understand the error tolerance of fact-checkers and downstream users. Wadden et al. (2020) conducted a COVID-19 case study with a baseline system trained on SciFact. They found their system produced plausible outputs for around two-thirds of input claims. In this case, plausibility is defined as more than 50% of retrieved evidence and classifications being judged correct by an expert with medical training. Since then, model performance on scientific claim verification has improved considerably. More work is needed to understand how improvements in ML model performance map to system and user gains in real-world settings, especially when considering potential performance degradation on emerging and unseen scientific topics.

Future directions

This section outlines some possible future directions for automated scientific claim verification. The first four directions – bootstrapping training data, integrating additional sources of information, generalisation and robustness, and open-domain fact-checking – propose improvements in the scope and performance of models. They can be thought of as extensions to current tasks and systems. The latter two directions – user-centric fact-checking and characterising social implications – aim to understand how ML technologies are applied or ought to be applied in this domain in practice.

Many automated science claim verification tools and prototypes have user interfaces that present veracity labels for each claim-evidence pair. This may not be the optimal interface for browsing and understanding evidence. As scientific knowledge is always evolving, the best ways to communicate the uncertainty and contradictions of scientific claims and evidence must be studied.

Bootstrapping data at scale

Due to the difficulty and expense of creating training data to verify scientific claims, methods that leverage distant supervision or that generalise well in the few- or zero-shot settings are desirable. Recent work introduces methods for learning general domain fact-checking without any labelled data (Pan et al., 2021). Variants of this method have been adapted to the scientific domain with good results (Wright and Augenstein, 2020). Optimistically, recent findings (Wadden et al., 2022) also demonstrate that training on weakly labelled data may be sufficient for domain transfer. This suggests that model generalisation could be achieved with fewer instances of expensive labelled data.

Integrating additional sources of information

Much of the discussion thus far focuses on scientific claim verification as a pure language modelling task, though this is only part of the picture. Metadata about the source of information – such as the authors, institutions, funding sources and the source's historical trustworthiness – could be useful indicators of veracity. Additionally, the text of evidence articles is not the only viable source of evidence for predicting veracity. Other structured and semi-structured resources such as curated knowledge bases (see the essay on knowledge bases in this volume by Ken Forbes), patient data or experimental data could also be used as sources of evidence. The integration of these external sources of information into veracity prediction is an important direction for future work.

Generalisation and robustness

Scientific claim verification datasets are limited to a few select domains, most notably biomedicine, public health and climate change. This is due in part to the prevalence and high negative costs of misinformation

in these domains. However, the tide of public interest can shift unpredictably; scientific findings will be called into question whenever they interface with policy and the public.

Therefore, scientific claim verification tools need to perform well and generalise beyond select domains. There are at least two directions to explore: understanding the fact verification needs of users in underexplored scientific domains; and developing evaluation benchmarks to assess the performance and suitability of claim verification models in these other domains.

Model robustness is a related direction. For example, Kim et al. (2021) showed that performance of fact verification models degrades when given colloquial claims as inputs. Methods to improve scientific claim verification model robustness are important avenues of future study.

Open-domain fact-checking

Another direction for scientific fact verification is the exploration of open- vs. closed-domain retrieval. Closed-domain retrieval predefines a set of documents that may provide evidence, e.g. a set of 10 000 trusted scientific articles. In open-domain retrieval, the space of potential evidence documents is significantly larger. It may be defined, for example, as all peer-reviewed scientific documents or all indexed websites on the Internet. The more realistic setting of open-domain retrieval is significantly more challenging. The scope of retrieval is orders of magnitude larger, requiring improvements in retrieval efficiency and a different model training regimen. However, this setting also better approximates real-world claim verification, where fact-checkers do not presuppose a limited set of sources for evidence.

User-centric fact-checking

Real-world claim verification must account for the beliefs and needs of users. Individuals may hold varying beliefs about the same claim – from strongly supportive to uncertain and in search of evidence. Knowledge of such stances may be important for selecting how best to communicate model outputs to these users. Nguyen et al. (2018) in their work on human-AI collaborative fact-checking, found that humans tended to trust model predictions even when they are incorrect. They concluded that some communication around model internals is needed to produce better outcomes.

Another aspect of modelling users involves understanding their role and intent. Are they a fact-checker, a journalist, a health-care consumer or some combination of many roles? The model must adjust its goal depending on the user's intended actions and the intended goal. For example, if the goal of verifying claims is to convince rather than inform, it may be important to expose both evidence documents and the rationales within those documents to support or justify the veracity label.

Assessing the social implications of science fact verification

Finally, as noted earlier, there has been limited engagement with the social implications of ML models for scientific fact verification. Work has focused on technological challenges such as improving the performance of models in increasingly real-world settings. At the same time, social science researchers have documented confirmation bias – the likelihood of individuals to seek out or focus on information that confirms their existing beliefs (Bronstein et al., 2019; Park et al., 2021). They also observed how fact-checking can induce questioning of scientific findings that undermines trust in the scientific process (Roozenbeek et al., 2020). These types of cognitive biases and responses can lead to counterproductive results in the application of automated claim verification. These social phenomena require consideration and should help guide the development and evaluation of machine-assisted claim verification systems in the wild.

Conclusion

Significant progress has been made in defining and executing on automated systems for scientific claim verification. Advancements in data, modelling, analysis and evaluation continue at a rapid pace. However, the community must address how claim verification models should present uncertainty and assess the soundness of a claim in the face of contradictory evidence. Work also remains in assessing whether state-of-the-art systems are ready for wide-scale deployment. To progress on these goals, more emphasis is needed on the potential social implications and ramifications of automated science claim verification systems. Through such improvements, these technologies could help improve understanding of the consistency and replicability of science, and transform people's trust and understanding of emerging scientific topics.

References

Abritis, A., A. Marcus and I. Oransky (2020), "An 'alarming' and 'exceptionally high' rate of COVID-19 retractions?", *Accountability in Research*, Vol. 28/1, pp. 58-59, https://doi.org/10.1080/08989621.2020.1793675.

Bronstein, M. et al. (2019), "Dual process theory, conflict processing, and delusional belief", *Clinical Psychology Review*, Vol. 72, pp. 101748, https://doi.org/10.1016/j.cpr.2019.101748.

DeYoung, J. et al. (2021), "MS^2: Multi-Document Summarization of Medical Studies", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, on line and Punta Cana, Dominican Republic, pp. 7494-7513, https://doi.org/10.18653/v1/2021.emnlp-main.594.

Diggelmann, T. et al. (2020), "CLIMATE-FEVER: A dataset for verification of real-world climate claims", *arXiv*, arXiv abs/2012.00614, https://doi.org/10.48550/arXiv.2012.00614.

Guo, Z., M. Schlichtkrull and A. Vlachos (2022), "A survey on automated fact-checking", *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 178-206, https://doi.org/10.1162/tacl_a_00454.

Kim, B. et al. (2021), "How robust are fact checking systems on colloquial claims?", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, on line, pp. 1535-1538, http://dx.doi.org/10.18653/v1/2021.naacl-main.121.

Kotonya, N. and F. Toni (2020), "Explainable automated fact-checking for public health claims", *arXiv*, arXiv:2010.09926 [cs.CL], https://doi.org/10.48550/arXiv.2010.09926.

Nguyen, A.T. et al. (2018), "Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking", in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, https://doi.org/10.1145/3242587.3242666.

Pan, L. et al. (2021), "Zero-shot fact verification by claim generation", *arXiv*, arXiv:2105.14682 [cs.CL], https://doi.org/10.48550/arXiv.2105.14682.

Park, S. et al. (2021), "The presence of unexpected biases in online fact-checking", 27 January, *Misinformation Review*, https://misinforeview.hks.harvard.edu/article/the-presence-of-unexpected-biases-in-online-fact-checking/.

Roozenbeek, J. et al. (2020), "Susceptibility to misinformation about COVID-19 around the world", *Royal Society Open Science*, Vol. 7, https://doi.org/10.1098/rsos.201199.

Saakyan, A., T. Chakrabarty and S. Muresan (2021), "COVIDFact: Fact extraction and verification of real-world claims on COVID-19 pandemic", *arXiv*, arXiv:2106.03794 [cs.CL], https://doi.org/10.48550/arXiv.2106.03794.

Sarrouti, M., A. Ben Abacha and Y. Mrabet (2021), "Fact-checking of health-related claims", in *Findings of EMNLP*, Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.297.

Thorne, J. et al. (2018), "FEVER: A largescale dataset for fact extraction and VERification", *arXiv*, arXiv:1803.05355 [cs.CL], https://doi.org/10.48550/arXiv.1803.05355.

Wadden, D. et al. (2022), "MultiVerS: Improving scientific claim verification with weak supervision and full-document context", in *Findings of the Association for Computational Linguistics: NAACL,* Association for Computational Linguistics, Seattle, United States, pp. 61-76, https://doi.org/10.18653/v1/2022.findings-naacl.6.

Wadden, D. et al. (2020), "Fact or fiction: Verifying scientific claims", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, on line, pp. 7534-7550, http://dx.doi.org/10.18653/v1/2020.emnlp-main.609.

Wright, D. and I. Augenstein (2020), "Claim check-worthiness detection as positive unlabelled learning", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, on line, pp. 476-488, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.43.

Robot scientists: From Adam to Eve to Genesis

R. King, Chalmers University, Sweden

O. Peter, Idorsia, Switzerland

P. Courtney, tec-connection, Germany

Introduction

This essay addresses the concept of the robot scientist, a technology that combines robotics with artificial intelligence (AI) to automate the scientific process. It traces the origins, enabling technologies and possible future directions for robot scientists, as well as the potential impact in advancing science and health care via their use in the biopharmaceutical industry. In identifying key trends, it makes recommendations for continued investment in the development of both AI and robotics and their interface across the medium to long term; incentives to develop and adopt interoperability standards and ontologies to support exchange and collaboration via open science; increased opportunities for collaboration between disciplines, including skills development; and international support for broad initiatives such as the Nobel Turing Challenge that can galvanise and inspire researchers, and even the general public.

Robot scientists

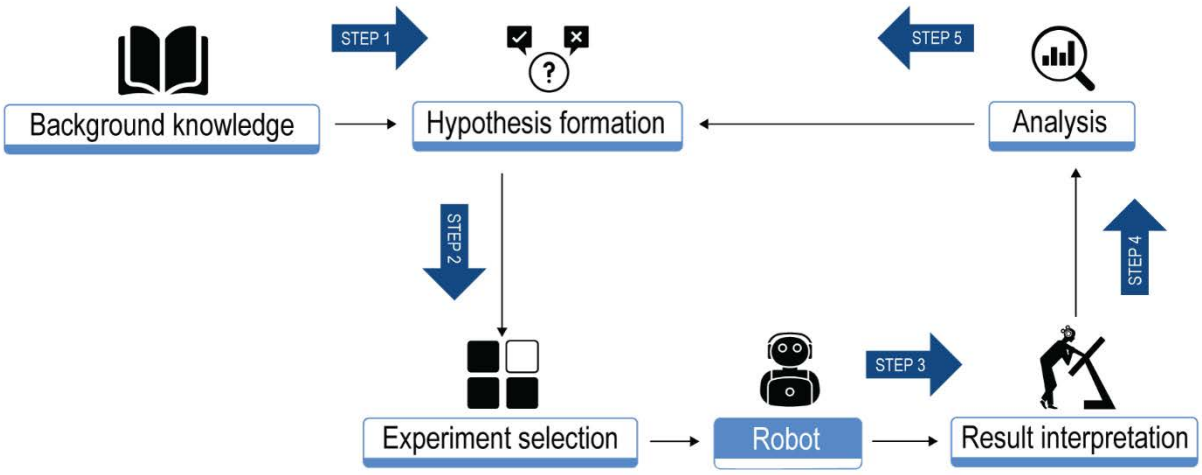
A robot scientist builds an artefact to perform basic scientific research autonomously by combining AI and machine learning (ML) algorithms and laboratory robots. A robot here is defined as an entity that interacts with the world. Once beyond the simplest mechanical devices, such as a pump, the possibilities of combining existing and new technologies in the service of robotics quickly become interesting. With suitable sensors such as cameras, a robot can perceive the world and make decisions. With suitable manipulators, it can interact with objects. With suitable locomotion, it can navigate around the world. Finally, with suitable cognitive or reasoning ability, it can adapt and learn.

For AI to have a significant impact on science, it needs to get into the laboratories where experimental research is actually done. Science is not just thinking about the world – it is about testing hypotheses by experiment.

A robot scientist – as distinct from other sorts of robot – is provided with background knowledge about an area of research. This knowledge is best represented using established tools from logic and probability theory. The robot scientist can automatically form a novel hypothesis about its area of science (Figure 1). That is, it considers existing knowledge available in databases and annotated datasets. It also considers published literature in the form of papers and patents, although to a lesser extent given limitations of natural language processing technology. With this knowledge, it formulates a hypothesis (step 1 in Figure 1); it

devises experiments to test the hypothesis (step 2); physically runs the experiments using laboratory robotics (step 3); interprets the results (step 4) to change the probability of different hypotheses (step 5); and then repeats the cycle (King et al., 2009, 2004). It also has an automated way of selecting efficient experiments (in terms of time and money) to decide between alternative hypotheses.

Figure 1. The robot scientist closed-loop cycle of experiments



Several advanced features distinguish robot scientists from other complex laboratory systems (such as high-throughput drug-screening platforms). These are their integral AI software, their many complex internal cycles (such as hypothesis generation, selection, evaluation and refinement) and their ability to execute individually planned cycles of experiments at high throughput.

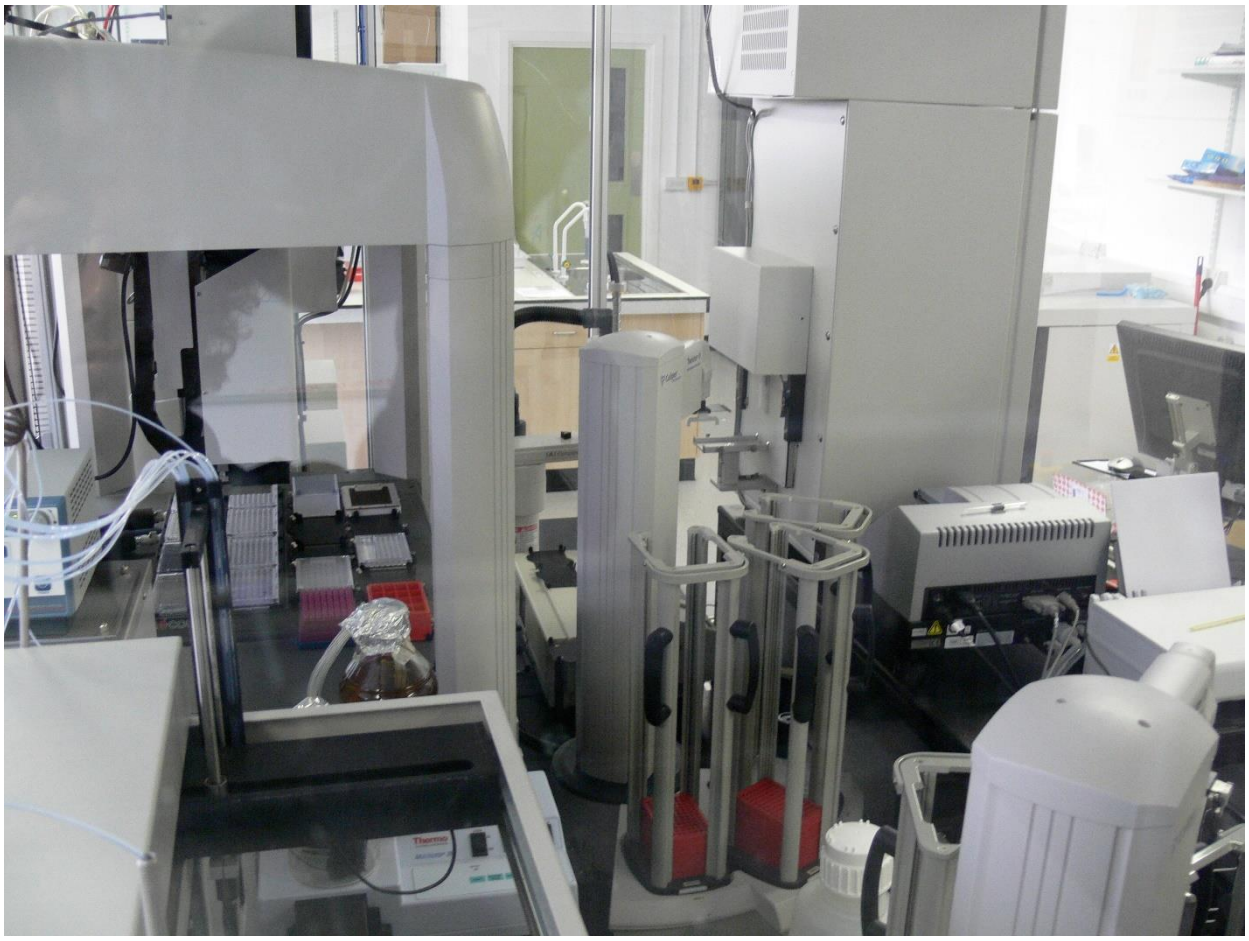
Materials scientists, chemists and drug designers have increasingly taken up integration of AI and laboratory automation. Robot scientists now go by different names, such as closed-loop platform, AI scientist, high-throughput experimentation platform and self-driving labs. The different names reflect developments in the different scientific communities (see later for more examples).

Adam

The original robot scientist was Adam (Figure 2), the first machine to discover novel scientific knowledge autonomously (King et al., 2009). It was designed to identify the relationship between genes and enzymes in yeast metabolism. Adam discovered the function of locally orphan enzymes; i.e. enzymes known to exist because of the functions they perform but for which the gene(s) encoding them are unknown (King et al., 2009).

Adam was designed to study how different strains of bacteria grow in a wide range of conditions. It efficiently generates the growth data needed to compare predictions made by the system. In scientific terms, a growth data maps the bacterial strain (genotype) to its behaviour (phenotype). Being fully automated, Adam required no technicians, except to periodically add laboratory consumables and remove waste. Each experiment ran for many days and around the clock. Adam designed and initiated over 100 new experiments a day from a selection of thousands of yeast strains and growth conditions. Every 20 minutes, accurate optical measurements were made for each experiment. This resulted in more than 10 000 reliable measurements per day of the growth of the different bacterial strains. Adam also automatically recorded the metadata for the experiments.

Figure 2. Robot scientist Adam



Eve

A second robot scientist, Eve (Figure 3), was designed to automate early-stage drug development (Williams et al., 2015). The design of Eve was motivated by the need to make drug discovery cheaper and faster. This was meant to promote development of treatments for diseases neglected for economic reasons, such as tropical and orphan diseases, and more generally to increase the supply of new drugs.

Eve integrates two advances in drug discovery automation. First, a laboratory automation system uses AI techniques to discover scientific knowledge through cycles of experimentation. Second, synthetic biology – an area of research designed to create new biological parts and systems – constructs cellular analogue computers (i.e. computers within living cells based on continuously changing biological processes rather than binary 0s and 1s).

In a novel development, Eve has three integrated modes, each corresponding to a stage in discovering what is known as a "lead" drug (i.e. a candidate-useful drug molecule). This integration aims to save time, avoiding the need to switch between different equipment:

1. In its library-screening mode, Eve systematically tests (assays) each compound in a library of compounds in the standard brute force way of conventional mass screening. While simple to automate, such mass screening is slow. It also wastes resources since every compound in the library is tested. It is also unintelligent, making no use of what is learnt during screening.

2. In confirmation mode, Eve then re-tests the promising compounds found by library screening. To minimise false positives, it uses multiple repeats and iterations.
3. In its final step, it executes cycles of statistics and ML to hypothesise so-called quantitative structure activity relationships (QSARs) and to test these QSARs on new compounds. A QSAR is a mathematical/computational function that predicts activity in an assay from a compound's structure. Eve's "QSAR-mode" is designed to execute such cycles of QSAR learning and testing, thus checking and improving performance.

Figure 3. Robot scientist Eve

The authors demonstrated that, under most circumstances, such intelligent library screening out-performs standard mass screening economically as it saves on time and compound use.

To validate Eve's performance on these processes, the authors used the system to quickly and cheaply find drugs known to be safe and that target multiple human and parasite enzymes. For several of these drugs, Eve helped provide new insight into their mode of action and helped indicate new uses for safe drugs. Furthermore, using econometric modelling, they demonstrated that Eve's use of AI to select compounds outperformed standard drug screening. Eve's most significant discovery is that triclosan (an anti-microbial compound commonly used in toothpaste) inhibits an essential mechanism in the malaria-causing parasites *Plasmodium falciparum* and *P. vixax* (Bilsland et al., 2018).

General advantages of robot scientists

The general motivation for using robot scientists is to increase the productivity of science. AI systems and robots can work more cheaply, faster, more accurately and longer than human beings (i.e. 24/7). More specifically, 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[1] to help reproduce work in other labs, increasing knowledge transfer and improving the quality of science
- increase 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. Robot scientists are also immune to a range of hazards, including pandemic infections. Importantly, all these remarkable capabilities remain complementary to the creative power of human scientists.

Advances and limitations in robotics

Advances in robotics as an enabling technology

Laboratory automation is already a multi-billion-US-dollar industry with strong contributions from Germany and Switzerland, as well as Japan, the United Kingdom and the United States. Laboratory robotics technology is steadily advancing. Today, many (but not all) tasks that a human can do in the laboratory can be automated.

Robotics has been the subject of considerable **public investment**. This is notably the case in Europe, where public investments have amounted to around EUR 100 million per year over the last 20 years (EC, 2008). Many countries have robotics programs, such as the United Kingdom's Robotics and Autonomous Systems initiative. These are set to grow as more investment is committed to development of AI.

AI-inside: AI is generally thought of as the abstract analysis of off-line data, at some distance from actual physical action. Conversely, robotics can be considered as "embodied AI", directly linked to action in the world. AI technology and robotics support one another at several levels – from sensing and actuation to reacting and planning. Classical engineering increasingly adopts and adapts ideas and elements from AI. The relationship between AI and robotics is not simple, but dialogue between the disciplines is ongoing, especially in Europe.

Robotic systems are already widely installed in labs, in particular for **handling liquids.** Many larger labs in the pharmaceutical industry use these routinely. Most clinical analysis, such as that of blood, is fully automated. Recently, some interesting robot assistants have also appeared (Burger, 2020). Developments in robotics in other sectors – such as robot chefs – also have clear implications for laboratory operations.

Limitations of robots to be overcome

Robots still have significant **limitations**. Many tasks in most labs remain largely un-automated. Robots today operate in protective boxes and can be hard for scientists to program. Often, logistics tasks still fall to lab technicians and scientists, who provide the robots with consumables, such as plates and chemicals, and remove waste. The average lab is still a long way from the digitalisation familiar in homes, exemplified by smartphone apps and robot vacuum cleaners.

Some developments can be observed in important components of robotic systems: **robotic arms** have become cheaper, easier to use and safer thanks to the development of collaborative robots with force-sensing capability. However, it is not uncommon to see industrial robots designed to lift 5 kg metal payloads moving plastic tubes weighing 50 g. **Physical manipulators** remain clumsy and not well suited to grasp the range of tubes and other devices commonly used in the lab; new ones are needed. **Mobile platforms** are the subject of intense interest and experimentation. However, they cannot be used to the extent they are in industrial warehouse operations without adaptation. All of the above indicate opportunities that are increasingly being recognised and addressed.

Advances enabling the robot scientists

Road-mapping as a route to collaboration

Given the importance of robotics and AI, there are potentially significant benefits in sketching out future needs. There have been considerable efforts to co-ordinate priorities and funding between researchers and firms working in AI and robotics through public-private partnerships such as euRobotics (n.d.). Through consultations, interviews and brainstorming meetings, for example, laboratory robotics has identified a range of use cases and challenges. These span diverse applications and capabilities, as well as a range of integrating (or platform) and modular approaches. Automated discovery platforms, for example, combine advances in AI and robotics to build on knowledge across many disciplines. Indeed, some groups have started to assemble such systems as open platforms (see below).

The importance of interoperability, ontologies and standards

Road-mapping exercises identified **interoperability** as an important barrier. Mutually agreed **ontologies** of concepts are needed to feed and train the AI algorithms so that semantic information can be shared and understood. Laboratories have benefited greatly from the widespread adoption of the SBS well plate format (Wikipedia, n.d.) as a standard carrier of biological samples. Adopting the SBS format has produced between a hundred- and thousandfold increase in productivity, a transformative impact comparable to the huge efficiency benefits in global trade and logistics from adoption of the 40-foot shipping container. A similar advance in the digital laboratory could come from standardised human- and machine-readable data formats. Initiatives to do this are well underway (SiLA, n.d.). One further consideration is how to share data in a findable, accessible, interoperable and reusable (FAIR) manner to support open science (Wilkinson et al., 2016).

Accelerating drug discovery by robot-supported AI

The innovative core of any new therapeutic drug is a molecule with intricately tuned properties: the active pharmaceutical ingredient (API). This may be a small molecule, typically for oral antibiotic, or a larger biological entity, such as an antibody therapeutic against cancer or an mRNA vaccine.

Novel drug molecules are created in two stages. First, a large substance collection is screened to identify chemical structures with some initial activity (“hits”). This process is called (ultra) high-throughput screening (HTS). The demand by the pharmaceutical industry to make HTS technically possible was a key driver of biological laboratory automation, beginning in the 1980s. HTS also contributed to enabling full genome sequencing, starting in the 2000s. In contrast, the second stage of drug discovery is much less automated. It typically takes several years to optimise the structures of candidate drug molecules. It also takes countless cycles of structure design, manual chemical synthesis and biological property testing before a promising drug molecule is found.

Automating medicinal chemistry was long considered technically impossible, largely because of the complexity of the task and human insight needed. With the emergence of ML/AI, this has started to change. Computer-aided drug design is increasing and might become widespread. However, actually making the molecules is still limited by cost and the capability and capacity of traditional chemistry labs. The pharmaceutical industry has already outsourced much of the work of making molecules to lower-cost countries. Expensive lab space is only occupied 25% of the time during a typical work week, considering eight-hour days and a five-day week. Manual optimisation of all the necessary work on promising candidate molecules inevitably consists mostly of unproductive waiting times, even when the work is outsourced across suppliers in different time zones.

Closed-loop design-make-test platforms for the biopharmaceutical industry

In future, competitive biopharmaceutical drug discovery and development will involve novel, fully automated, closed-loop design-make-test (DMT) platforms. These will integrate iterative design of the structure of molecules by ML/AI algorithms and the synthesis and testing of physical molecules. The goal is to eventually push down the time for optimisation of good candidate molecules from weeks to hours, producing valuable preclinical development candidates in months rather than years.

The “cloud lab” concept is also emerging within the biopharmaceutical industry. This development recognises that laboratory automation systems are still expensive to build and difficult to use. It also reflects how providing automation at scale as a service in the cloud can address challenges. In this way, customers access automated labs through a user interface or an API, designing and executing their experiments remotely. A few companies have started to offer such services, including Strateos (n.d.) and Emerald Cloud Lab (n.d.), both in California.[2] Since 2021, and based on Eli Lilly’s decade-long experience automating medicinal chemistry, Strateos built and operates the Lilly Life Science Studio in San Diego. This is the most ambitious generalised medicinal chemistry platform publicly available today (Mullin, 2021).

Automated lab infrastructure might be built into standard shipping containers to be scalable and relocatable. By analogy with virtualised computing infrastructure like Amazon Web Services, such remote experimentation services could enable the emergence of “virtual” biopharmaceutical enterprises. This means individual companies would not need to own a laboratory. However, to avoid the “balkanisation” of APIs, global cross-platform standards must be adopted.

Most of the physical and computational elements required to build closed-loop platforms exist today. One challenge is to automate milligram-scale transfer of thousands of solid chemicals (the building blocks of molecular synthesis), which are often sticky or viscous. Companies such as Chemspeed offer solutions and are working to improve them. However, systems must be modular to make their inherent complexity manageable and their process control efficient. Modularisation must itself happen in iterative cycles of platform improvement.

The next steps in automation in the biopharmaceutical industry

A stepwise approach towards fully integrated closed-loop DMT platforms could be based on independent islands of automation with manageable yet limited functionality. These would be connected by autonomous lab robots to transport samples. Such robots are just emerging (but manually transferring samples might suffice initially).

These loosely coupled functional modules would be used in a workpiece-centric way, much like a modern digitally connected factory. Rather than centrally orchestrating every move, the system responsible for each job (such as transferring a rack of samples from one station to another) would request whatever processing step is required next as the overall process moves to completion.

The authors’ experience of building large high-throughput systems has shown that technical standards for physical and logical interfaces between laboratory devices are crucial to constantly improve and adapt

such platforms to changing requirements. Such standards relate to the physical dimensions of items consumed by the platforms during operation, reagent packaging, device commands, method descriptions and scientific data formats. Open standards will foster a healthy ecosystem of co-operating and productively competing suppliers and consumers of future DMT platform modules and services.

In summary, better design algorithms will have limited impact on experimentation in the biopharmaceuticals sector without similarly efficient, robotically automated, physical synthesis and testing. They also need interfaces to engage humans, who bring their unique ingenuity.

The next generation robot scientist

The robot scientist concept is increasingly recognised as a general platform for accelerating science. Beyond Adam-Eve, a number of other automated discovery platforms operate in a range of disciplines. Each has specific tools and configurations, and each reveals the next bottlenecks to resolve in automation. These include, for example, initiatives in the following:

- chemistry in Canada (Kebotix n.d.), in the United Kingdom (Imperial College London, n.d.), in the United States (The Cernak Lab, n.d.; MIT, n.d.), and in Switzerland (IBM Research, n.d.) and material science, such as the material genome initiatives in Canada and Cambridge in the United Kingdom (BIG-MAP, n.d.) and in the United States (MGI, n.d.), which call upon large databases of known reactions to create desired molecular forms
- catalysis in France (Realcat, n.d.) and Switzerland (Swiss CAT+, n.d.), which assess the functional performance of the new materials in specific tasks
- metallurgy, which explore new alloys by the combination of existing metals.

In each of these cases, the lack of data on failed experiments, which human scientists are not incentivised to record, is revealed to be a knowledge gap, one that such platforms are well placed to address. One should also mention initiatives in cell culture and bioprocessing, such as the KIWI-biolab (n.d.) from TU Berlin, which leverage developments in genomics. Meanwhile, at labdroids at Riken in Japan, humanoid robots automate operations using tools designed for human hands to improve reproducibility.

Chalmers University in Sweden aims to take the robot scientist to a new level in terms of the number of experiments per step, and the generation and use of more and better quality data. The Chalmers robot, known as Genesis, aims to achieve a detailed and complete understanding of the functioning of complex cells such as yeast. This goal remains a fundamental challenge for 21st century science, and a solution could help answer many questions in the life sciences, biotechnology and medicine.

The new hardware for Genesis will be equipped with 10 000 miniature fermenters (technically, chemostats that control the culture of microorganisms). They will be able to carry out detailed analysis of how biological function relates to metabolism and active genes. The original technology was developed through government-funded research at Vanderbilt University in the United States. Genesis represents a thousandfold scale-up on a traditional manual laboratory that has around ten chemostats. Only an AI system can control so many different experiments, where every day each chemostat will run a separately designed experiment to test a hypothesis.

Conclusion

Science requires experiments involving physical actions, which creates a critical role for robots. This will require support for development of both AI and robotics, and the interface between them (see above section on advances in robotics technology). The newly formed AI, Robotics and Data Association, which focuses on this AI-robotics interface, is a recent example of work in this direction (Adra, n.d.).

The Nobel Turing Challenge underlines the importance of interdisciplinary collaboration at the international level. It challenges researchers to build a system by 2050 that can perform scientific discoveries at a level that merits a Nobel Prize (Kitano, 2021).[3] This bold and inspiring challenge is gaining support in the United States, Japan and Europe.

A few suggestions for the role of public support are noted below.

Robots for laboratory science

Robots are developing fast for industrial applications but not always in ways that meet the needs of laboratories. Since laboratory users are often highly skilled and collaborative, it may not be the most productive path to replace them directly with existing robots. There are deep intellectual challenges in developing the necessary technologies in partnership with laboratory users. As a consequence, more interaction is required between the roboticists and the domain experts, perhaps in collaborative research programmes and centres. Such programmes could, for example, bring together materials scientists, chemists, AI experts and roboticists to help develop next-generation battery materials (BATT4EU, n.d.; Stein and Gregoire, 2019). Collaborative programmes could also facilitate road-mapping activities across disciplines to identify future gaps and opportunities, and thus guide funding priorities (euRobotics, n.d.). Governments are best placed to create such programmes because of the broad reach required. They can bring together players that otherwise rarely co-ordinate their activities.

Data governance

Ontologies are necessary for AI/ML, but the fragmented ones must be consolidated and aligned. Laboratory instruments need to become interoperable via standardised interfaces. Laboratory users, suppliers and technology developers could be brought together and incentivised to co-operate from the moment where the data are generated by funders and publishers. This might take place under open science initiatives, for example, that support data curation and sharing through the FAIR principles, as well as appropriate data governance processes, including ethics.

Bridging disciples to overcome educational gaps

Ongoing long-term collaboration across scientific disciplines is essential but is still too weak. The development of closed-loop research and development centres is encouraging and 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. When linked together, such centres (often national in reach) can also support common interests such as training and evolving research practice. For example, biologists are increasingly exposed to applied mathematics and statistics. However, engineers are still seldom exposed to modern, data-rich life science. All need greater exposure to the many issues relating to data governance, as noted above. Again, governments have a role here, one unlikely to be pursued by the private sector alone.

A visionary initiative with long-term impact

Initiatives such as the Nobel Turing Challenge can galvanise and inspire collaboration and co-ordination in science and should be supported at an international level. This support can help focus efforts on addressing long-term global challenges such as climate change and cancer. At the same time, it can drive agreement on standards and, not least, attract the young talent needed to make this ambition a reality.

These suggestions align well with those made in a recent US report regarding research practices, educational gaps, long-term support and data governance (National Academies of Sciences, Engineering and Medicine, 2022).

References

Adra (n.d.), The AI Data Robotics Association website, https://ai-data-robotics-partnership.eu (accessed 10 January 2023).

BATT4EU (n.d.), Batteries European Partnership website, https://bepassociation.eu (accessed 10 January 2023).

BIG-MAP (n.d.), Batteries Interface Genome – Material Applications Platform website, www.big-map.eu (accessed 10 January 2023).

Bilsland, E. et al. (2018), "Plasmodium dihydrofolate reductase is a second enzyme target for the antimalarial action of triclosan", *Scientific Reports*, Vol. 8/1, pp. 1-8, www.nature.com/articles/s41598-018-19549-x.

Bilsland, E. et al. (2011), "Functional expression of parasite drug targets and their human orthologs in yeast", *PLoS Neglected Tropical Diseases*, Vol. 5/10, pp. e1320, https://doi.org/10.1371/journal.pntd.0001320.

Burger, B. et al. (2020), "A mobile robotic chemist", *Nature,* Vol. 583/7815, pp. 237-241, https://doi.org/10.1038/s41586-020-2442-2.

EC (2008), "EU doubles investment in robotics", European Commission CORDIS Research Results, Brussels, https://cordis.europa.eu/article/id/29537-eu-doubles-investment-in-robotics.

Emerald Cloud Lab (n.d.), Emerald Cloud Lab website, www.emeraldcloudlab.com (accessed 10 January 2023).

euRobotics (n.d.), euRobotics website, www.eu-robotics.net (accessed 10 January 2023).

Gromski, P. et al. (2020), "Universal chemical synthesis and discovery with 'the chemputer'", *Trends in Chemistry*, Vol. 2/1, pp. 4-12, www.sciencedirect.com/journal/trends-in-chemistry/vol/2/issue/1.

IBM Research (n.d.), "IBM RoboRXN", webpage, https://research.ibm.com/science/ibm-roborxn (accessed 10 January 2023).

Imperial College London (n.d.), "Centre for Rapid Online Analysis of Reactions (ROAR)", webpage, www.imperial.ac.uk/rapid-online-analysis-of-reactions# (accessed 10 January 2023).

Kebotix (n.d.), Kebotix website, www.kebotix.com (accessed 11 January 2023).

KIWI-biolab (n.d.), KIWI-biolab website, https://kiwi-biolab.de (accessed 11 January 2023).

Kitano, H. (2021), "Nobel Turing Challenge: Creating the engine for scientific discovery", *npj Systems Biology and Applications*, Vol. 7/1, pp. 1-12, https://doi.org/10.1038/s41540-021-00189-3.

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

King, R.D. et al. (2004), "Functional genomic hypothesis generation and experimentation by a robot scientist", *Nature*, Vol. 427/6971, pp. 247-252, https://doi.org/10.1038/nature02236.

MGI (n.d.), Materials Genome Initiative website, www.mgi.gov (accessed 11 January 2023).

MIT (n.d.), Jensen Research Group website, https://jensenlab.mit.edu (accessed 11 January 2023).

Mullin, R. (2021), "The lab of the future is now", *Chemical & Engineering News* Vol. 99/11, pp. 28, https://cen.acs.org/business/informatics/lab-future-ai-automated-synthesis/99/i11.

National Academies of Sciences, Engineering, and Medicine (2022), *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*, The National Academies Press, Washington, DC, https://doi.org/10.17226/26532.

Realcat (n.d.), Realcat website, www.realcat.fr (accessed 11 January 2023).

SiLA (n.d.), SiLA website, https://sila-standard.com (accessed 11 January 2023).

Stein, H.S. and J.M. Gregoire (2019), "Progress and prospects for accelerating materials science with automated and autonomous workflows", *Chemical Science*, Vol. 10/42, pp. 9640-9649, https://doi.org/10.1039/C9SC03766G.

Strateos (n.d.), Strateos website, https://strateos.com (accessed 11 January 2023).

Swiss CAT+ (n.d.), Swiss CAT+ website, https://swisscatplus.ch (accessed 11 January 2023).

The Cernak Lab (n.d.), The Cernak Lab website, https://cernaklab.com (accessed 11 January 2023).

Wikipedia (n.d.), "Microplate", webpage, https://en.wikipedia.org/wiki/Microplate (accessed 11 January 2023).

Wilkinson, M.D. et al. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3/1, pp. 1-9, https://doi.org/10.1038/sdata.2016.18.

Williams, K. et al. (2015), "Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases", *Journal of the Royal Society Interface*, Vol. 12/104, pp. 20141289, https://doi.org/10.1098/rsif.2014.1289.

Notes

[1] Unlike for humans, the recording of data, metadata and procedures adds up to 15% of the total costs of experimentation. Moreover, despite the recording of experimental data being widespread, it is still uncommon to fully document used procedures, errors and all the metadata.

[2] Other private initiatives outside the United States include Arctoris and LabGenius in the United Kingdom.

[3] For further details, see www.turing.ac.uk/research/research-projects/turing-ai-scientist-grand-challenge.

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

N.R. Smalheiser, University of Illinois at Chicago, United States

G. Hahn-Powell, University of Arizona, United States

D. Hristovski, University of Ljubljana, Slovenia

Y. Sebastian, Charles Darwin University, Australia

Introduction

This essay gives an overview and describes prospects for generating new scientific knowledge from disparate datasets, as viewed by four active practitioners from around the globe (Illinois, Arizona, Slovenia and Australia). Although artificial intelligence (AI) and machine learning (ML) are central techniques employed in the field, the key concepts in this essay are undiscovered public knowledge (UPK) and literature-based discovery (LBD). These comprise a variety of situations, including some not yet tackled via ML.

UPK was originally coined by Swanson (1986) and expanded by Davies (1989). It suggests that scientific findings, hypotheses and assertions may exist within the published literature without anyone being aware of them. They may be undiscovered because no one alive has read the articles (e.g. they were published in obscure journals or lack Internet indexing). In other cases, different snippets of evidence or assertions may be scattered among multiple documents and need to be pieced together. For example, one article may raise a hypothesis that is tested in another, without any one individual being aware the two are related. As another example, multiple types of evidence may exist across different studies that address the same issue but are not integrated readily with each other (e.g. epidemiologic study vs. case reports); this is in contrast to meta-analyses, which attempt to collate comparable studies.

LBD generally refers to the fascinating possibility that one can create entirely new, plausible and scientifically non-trivial hypotheses 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. The potential number of such transitive relations in the literature is astronomical. Thus, the LBD problem is to filter or identify which assertions of the type “A affects C” are novel, scientifically plausible, non-trivial and sufficiently interesting that a scientist would find them worthy of study. LBD differs from AI data mining efforts such as Knowledge Discovery from Databases that use statistics and interestingness[1] measures to

identify explicitly stated findings or significant associative trends in the data. In contrast, LBD attempts to identify *unknown* knowledge that is implicitly rather than explicitly stated.

Advances in AI, ML and computational linguistics are key to improving such systems. For example, better extraction of entities and relations, and better natural language inference and causality models, will improve the precision of "A affects B" and "B affects C" assertions. This, in turn, will greatly help assess whether a potential link can be explained in terms of known mechanisms. Advances in machine reading (teaching computers to read and comprehend natural language text), and especially "deep learning" neural network architectures applied to text, show great potential in identifying assertions in scientific articles and implicit relationships.

What LBD tools are available?

The first computer-assisted tools for carrying out LBD analyses were the following:

Arrowsmith 1-node and 2-node search tools (http://arrowsmith.psych.uic.edu) (Swanson and Smalheiser, 1997; Torvik and Smalheiser 2007). In the 2-node search, users define two sets of biomedical articles (hereby termed literature sets A and C) by carrying out two searches within the PubMed search engine. The Arrowsmith software then identifies title words from both literature sets to identify one or more connecting terms/phrases ($B_{i=1, 2, 3, ...}$) in common. These phrases are then ranked according to their predicted relevance for linking A and C in a meaningful manner. For each connecting term B_i, the system displays the instances of B_i in the A literature next to instances of B_i in the C literature, making it easy to see if there is an interesting A - B_i - C relationship. In the 1-node search, the user defines a single literature A that studies a given problem (e.g. Alzheimer's disease). The system then identifies disparate literatures C_i ranked by how many intermediate terms or concepts they share with A.

BITOLA (https://ibmi.mf.uni-lj.si/en/node/253) is based on co-occurrences of medical subject headings integrated with genetic background knowledge. This makes it especially useful for identifying candidate genes (Hristovski, 2005).

SemBT (http://sembt.mf.uni-lj.si) uses semantic relations extracted from the biomedical literature combined with microarray results (microarrays are used in laboratory settings to detect simultaneous expressions of thousands of genes). LBD in this case can be used both for microarray results interpretation (Hristovski, 2009) and for drug repurposing.

Mine the Gap! (accessible in https://h2020-minethegap.eu/) is a variant approach in which the user specifies a given set of literature, whereupon the software identifies "gaps" within that field. These gaps are pairs of topics that separately are studied frequently within the field, yet have never been discussed in the same article in that field.

Influence Search provides direct and indirect search over a graph of influence relations mined from English and Portuguese scholarly documents indexed by PubMed and SciELO. Each edge is weighted by how frequently its corresponding relation is discussed across documents. It is also weighted as a measure of the certainty of the relationship based on the degree of hedging in its description (Hahn-Powell, Valenzuela-Escárcega and Surdeanu, 2017; Barbosa et al., 2019).

Finally, **Lion-LBD** (https://lbd.lionproject.net) looks for relationships among instances of diseases, genes, mutations, chemicals, cancer hallmarks and species mentioned within biomedical articles rather than among documents or sets of documents.

All of these systems are implemented as free, public biomedical web tools. As well, proprietary systems include IBM Watson for Drug Discovery and Biovista's Biolab Experiment Assistant.

New and emerging models of LBD

To date, most research on LBD has come from practitioners in computer science, information science and bioinformatics. It has largely dealt with methodological questions that employ the ABiC model. For example, should Bi terms be extracted from title, abstract, specific document sections or the full text of literatures A and C? Should Bi terms represent text or ontological concepts? How can LBD be modelled on knowledge graphs, for example, by predicting which unlinked nodes are likely to become connected in the future?

Emerging approaches extend the ABC model in various ways. For example, instead of A – Bi – C, one may wish to create longer paths or chains of assertions (A – B1 -B2 – B3 – C) bridging any two literatures or concepts (Hossain et al., 2012). Alternatively, instead of connecting textual artefacts (documents or concepts), one may envision connecting investigators to identify potential collaborators (or potential reviewers). “Dr Smith” and “Dr Jones”, for example, may not know each other or attend the same meetings. However, they may be implicitly linked if they published on similar topics or even co-authored with some of the same scientists. If they share certain common interests or attributes, they might be expected to collaborate fruitfully, perhaps synergistically, on a particular hypothesis or scientific problem.

Even more interesting is when the collaborators come from complementary domains. Recently, a semantics-based methodology for cross-domain collaboration recommendations has been proposed (Hristovski et al., 2015) and later implemented with a graph database (Hristovski et al., 2016). This methodology proposes not only pairs of potential collaborators but also an explanation for why such a collaboration makes sense. Another approach along these lines is to define research communities, looking for links across disparate fields of research (Hahn-Powell, 2018). This would identify knowledge gaps and key ideas that can bridge disciplines and foster the kind of collaboration that accelerates scientific progress.

Early LBD studies focused on identifying novel links that represent potential new hypotheses. However, it is increasingly clear that the real goal is not novelty per se but rather finding hypotheses that domain scientists will find interesting, non-trivial and worthy of further study. Assertions that represent small increments from current knowledge may be very likely to be true. For example, if dexamethasone helps patients with COVID-19, then similar steroids may help COVID-19 as well. Yet these assertions are the least surprising and, because they are obvious, perhaps the least interesting from the standpoint of investigators in the field. Thus, there is an apparent trade-off: the more divergent a predicted hypothesis is from current knowledge, the more surprising it is but (all things being equal) the least likely it is to be true. On the other hand, previously published findings that were neglected or apparently refuted using the methods available at the time may actually represent the raw material for new hypotheses and even new paradigms. This is especially true if one looks at them in the light of more recent findings and methods (Swanson, 2011; Smalheiser, 2013; Smalheiser and Gomes, 2014; Peng, Bonifield and Smalheiser, 2017).

The above suggests a need for “interestingness” measures that can automatically score and rank hypotheses in terms of their surprisingness and potential impact on science. These would help guide users to focus on those that have “bang for the buck”. While no such dataset exists, judgements on aspects of “interestingness” could be collected through user interaction with an information retrieval system. These judgements could then be used to train a personalised recommendation system that learns to combine features derived from knowledge graph structures with user profiles and behaviours (Zhao, Wu and Liu, 2016; Guo et al., 2020). Such a system could be continuously improved through a virtuous cycle of human-machine collaboration.

Future LBD systems may also need to consider radically new approaches in synthesising knowledge that assess multiple weak findings across disparate sources. For example, multiple medical case reports sometimes publish quite similar findings, albeit in different contexts (Smalheiser, Shao and Yu, 2015). Under this scenario, one cannot undertake a conventional meta-analysis. Yet, intuitively, the presence of multiple independent reports should point to a real signal among the “noise” of individual cases. In materials

sciences, scientific documents commonly report a limited number of material samples being synthesised and characterised from non-comparable experiments. Again, conventional meta-analysis is not appropriate. Instead, new ways of combining information across disparate contexts are needed (Tshitoyan et al., 2019; Szczypiński et al., 2021).

LBD can be fruitfully integrated with other AI methods, such as neural networks, to provide explanation capabilities. In Zhang et al. (2021), several methods for knowledge graph completion (link prediction) using neural networks were used for drug repurposing for COVID-19. The medical doctor responsible for the evaluation may not always find it easy to directly interpret the rationales behind the proposed drug repurposing. In these cases, LBD's 2-node searches (such as those provided by the Arrowsmith system) can be used to provide explanations for paths in the knowledge graph between the drugs and COVID-19.

How can informatics scientists best collaborate with bench scientists, especially in biology and medicine?

Many biomedical hypotheses emerging from LBD analyses have been published. The earliest examples suggested that magnesium supplementation could prevent or treat migraine headaches, and that fish oil could treat Raynaud's disease. Indeed, the entire field of drug repurposing owes its underlying strategy to LBD. For example, one may rank drugs according to whether they elicit changes in gene expression that occur in directions opposite to those that occur in a given disease. LBD or bioinformatics practitioners themselves carried out most of the published analyses. Roughly 25 specific hypotheses have appeared among the Swanson-Smalheiser group, the Hristovski-Rindflesch group, the Wren-Garner group and a few others. Some have been experimentally tested and confirmed. As well, several independent biomedical investigators have employed Arrowsmith software to generate and assess hypotheses related to their laboratory studies (Kell, 2009; Manev and Manev, 2010).

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). This may partially explain their limited use by the biomedical community to date. As well, LBD tools may need to become more user-friendly, fast responding and interactive. Perhaps they could display only the few best hypotheses generated by these systems that need to be investigated and evaluated. The tools should also be able to explain why the proposed novel hypotheses are attractive. In other words, explainability is also essential for wider adoption. Finally, LBD tools need to be publicly accessible as web-based tools (not merely code archived on Github) that operate over a continuously updated document collection.

There are also social and organisational obstacles to wider LBD adoption by the biomedical community. Most LBD research is published and presented at venues attended by biomedical informaticians, who are not the real end-users. Conversely, the biomedical curriculum does not train students to search systematically for new hypotheses. Moreover, in general, many investigators have little expertise or formal training with computer programming, data provenance issues and so forth. Investigators design and conduct their experiments (or should do) in collaboration with statisticians. In the same way, 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?

Extending LBD analyses beyond text

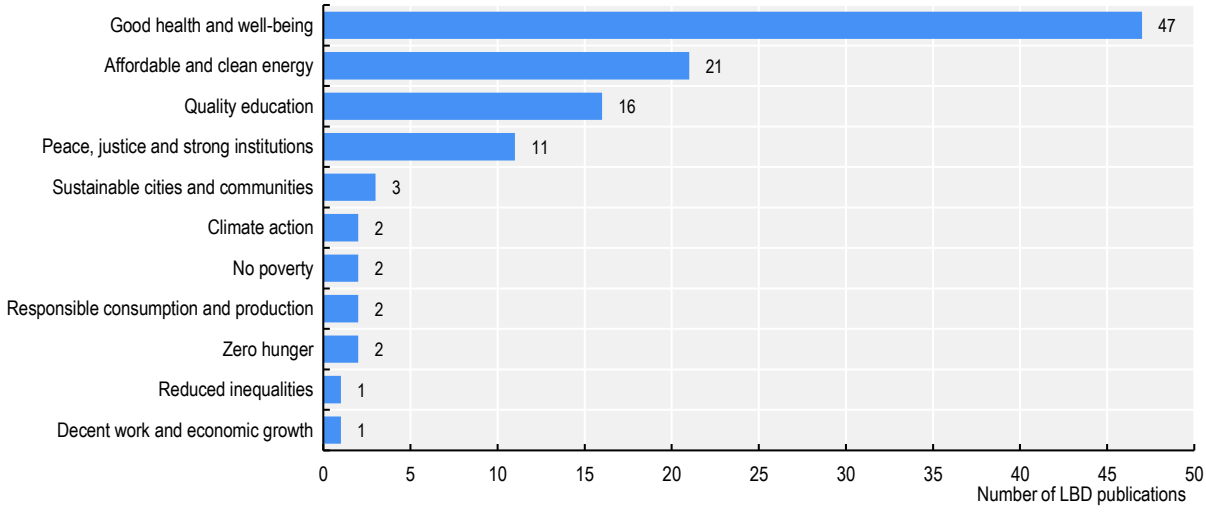
Besides linking assertions or findings in articles and other documents, the next-generation LBD systems are likely to use information in non-natural language forms. These could include numerical tables, charts

and figures, programming codes, microarrays, next-generation sequencing results, phenotypes, clinical data, etc. This is in parallel with the increasing awareness that different scientific fields communicate differently, implying diverse emphases on various formats of information (National Academies of Sciences, Engineering and Medicine, 2017). In fact, progress is being made in this direction that makes non-textual information more amenable to text mining (Pyarelal et al., 2020; Suadaa et al., 2021).

Prospects for LBD accelerating scientific progress outside biomedicine

An increasing number of LBD applications are being reported outside of biomedicine. In materials sciences, a group at the Lawrence Berkeley National Laboratory (Tshitoyan et al., 2019) recently demonstrated an LBD-style computer algorithm for discovering new materials. The algorithm uses static word embeddings to discover latent associations between an existing material (e.g. a crystal structure) and its previously unexplored thermoelectric applications. Word embeddings are vector representations of words built from millions of materials science publications. These vector representations can capture complex relationships among materials concepts without requiring explicit chemical knowledge to be specified *a priori* (e.g. periodic table). Using this method, computers can be used to automatically recommend new or existing materials for novel applications *long before* their discoveries. This saves money and time given that conventional materials engineering approaches typically rely on slow and arduous experimentations to discover or repurpose new materials (Szczypiński et al., 2021). As an example, researchers at MIT recently demonstrated the discovery time of new materials can be dramatically reduced from 50 years by conventional analytical methods to merely 5 weeks with the help of artificial neural networks (Janet et al., 2020).

Figure 1. The distribution of literature-based discovery research publications according to their alignment with selected UN Sustainable Development Goals (1989-2021)



Note: Bars indicate the number of publications containing the keyword "literature-based discovery" published between 1989 and 2021.
Source: www.dimensions.ai (accessed on 9 October 2021).

Figure 1 illustrates the far-reaching potentials of LBD in terms of the UN Sustainable Development Goals (SDGs). LBD researchers have previously attempted analyses on 10 of 17 goals. However, Figure 1 also points to a problem: less than 6% of all LBD publications (108 of 1 928) can be mapped to at least one SDG. Limitations of bibliographic indexing aside, this may suggest that the practicality of new LBD methods

and algorithms needs to be better contextualised within real-world problems (Mejia and Kajikawa, 2021). Doing so could help increase the uptake of LBD by the scientific and non-scientific community at large.

Future developments of AI-driven knowledge creation tools must be accompanied by the increasing availability of open research data. Platforms such as Figshare (https://figshare.com), Dryad (https://datadryad.org/stash) 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 also open exciting possibilities for more author and community-centric LBDs. Finally, catalysts can also be found in public data initiatives such as The Australian Research Data Commons (https://researchdata.edu.au), the US Government's Open Data (https://www.data.gov) and the EU ORD Pilot (https://data.europa.eu).

Conclusion

UPK and LBD are simple, intuitive concepts that have profound implications for the philosophy and practice of science. Investigators now realise that publications are not simply archives of prior studies. They can also be a fertile raw material for making new and testable hypotheses that represent potential discoveries. LBD techniques work hand in hand with AI methods in machine learning, ontologies, knowledge graphs and computational linguistics, which are themselves making rapid progress. Thus, LBD analyses should continue to expand in biomedicine, the physical and social sciences, and even the humanities.

The greatest challenge is to integrate LBD analyses into real-life scientific workflows. There is no "killer app" akin to Google Scholar used by the general scientific community on a daily basis. Instead, 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. Maybe, in the not-so-distant future, AI software agents could serve the role of an intermediary between LBD tools and their intended users.

References

Barbosa G.C.G. et al. (2019), "Enabling search and collaborative assembly of causal interactions extracted from multilingual and multi-domain free text", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations),* Minneapolis, https://doi.org/10.18653/v1/N19-4003.

Davies, R. (1989), "The creation of new knowledge by information retrieval and classification", *Journal of Documentation*, Vol. 45/4, pp. 273-301, https://doi.org/10.1108/eb026846.

Guo, Q. et al. (2020), "A survey on knowledge graph-based recommender systems", *arXiv,* abs/2003.00911, https://doi.org/10.48550/arXiv.2003.00911.

Hahn-Powell, G. (2018), "Machine reading for scientific discovery", PhD dissertation, University of Arizona, Tucson, https://repository.arizona.edu/handle/10150/630562.

Hahn-Powell, G., M.A. Valenzuela-Escárcega and M. Surdeanu (2017), "Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph", in *Proceedings of ACL 2017, System Demonstrations*, Association for Computational Linguistics, Vancouver, https://doi.org/10.18653/v1/P17-4018.

Hristovski, D. et al. (2016), "Implementing semantics-based cross-domain collaboration recommendation in biomedicine with a graph database", in *Proceedings of the Eighth International Conference on Advances in Databases, Knowledge, and Data Applications*, Lisbon, https://www.iaria.org/conferences2016/DBKDA16.html.

Hristovski, D. et al. (2015), "Semantics-based cross-domain collaboration recommendation in the life sciences: Preliminary results", in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, Calgary, https://doi.org/10.1145/2808797.2809300.

Hristovski, D. et al. (2013), "Using literature-based discovery to identify novel therapeutic approaches", *Cardiovascular & Hematological Agents in Medicinal Chemistry,* Vol. 11/1, pp. 14-24, https://doi.org/10.2174/1871525711311010005.

Hristovski, D. et al. (2009), "Semantic relations for interpreting DNA microarray data" in *AMIA Annual Symposium Proceedings,* Vol. 255/9, American Medical Informatics Association, Rockville.

Hristovski, D. et al. (2005), "Using literature-based discovery to identify disease candidate genes", *International Journal of Medical Informatics*, Vol. 74/2-4, pp. 289-298, https://doi.org/10.1016/j.ijmedinf.2004.04.024.

Janet, J.P. et al. (2020), "Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization", *ACS Central Science,* Vol. 6/4, pp. 513-524, https://doi.org/10.1021/acscentsci.0c00026.

Kell D.B. (2009), "Iron behaving badly: Inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases", *BMC Medical Genomics* Vol. 2/2, https://doi.org/10.1186/1755-8794-2-2.

Manev, H. and R. Manev (2010), "Benefits of neuropsychiatric phenomics: Example of the 5-lipoxygenase-leptin-Alzheimer connection", *Cardiovascular Psychiatry and Neurology* 2010:838164, https://doi.org/10.1155/2010/838164.

Mejia, C. and Y. Kajikawa (2021), "Exploration of shared themes between food security and Internet of Things research through literature-based discovery", *Frontiers in Research Metrics and Analytics,* Vol. 6/25, https://doi.org/10.3389/frma.2021.652285.

National Academies of Sciences, Engineering, and Medicine (2017), *Communicating Science Effectively: A Research Agenda,* National Academies Press, Washington, DC, https://doi.org/10.17226/23674.

Peng, Y., G. Bonifield and N.R. Smalheiser (2017), "Gaps within the biomedical literature: Initial characterization and assessment of strategies for discovery", 22 May*, Frontiers in Research Metrics and Analytics,* https://doi.org/10.3389/frma.2017.00003.

Pyarelal, A. et al. (2020), "Automates: Automated model assembly from text, equations, and software", *arXiv,* arXiv:2001.07295, https://arxiv.org/abs/2001.07295v1.

Sebastian, Y., E.G. Siew and S.O. Orimaye (2017), "Emerging approaches in literature-based discovery: Techniques and performance review", *The Knowledge Engineering Review*, Vol. 32, p. e12, https://doi.org/10.1017/S0269888917000042.

Smalheiser, N.R. (2017), "Rediscovering Don Swanson: The past, present and future of literature-based discovery", *Journal of Data and Information Science*, Vol. 2/4, pp. 43-64, https://doi.org/10.1515/jdis-2017-0019.

Smalheiser, N.R. (2013), "How many scientists does it take to change a paradigm? New ideas to explain scientific observations are everywhere – we just need to learn how to see them", *EMBO Reports*, Vol. 14/10, pp. 861-865, https://doi.org/10.1038/embor.2013.125.

Smalheiser, N.R (2012), "Literature-based discovery: Beyond the ABCs", *Journal of the American Society for Information Science and Technology*, Vol. 63/2, pp. 218-224, https://doi.org/10.1002/asi.21599.

Smalheiser N.R. and O.L. Gomes (2014), “Mammalian Argonaute-DNA binding?”, *Biology Direct*, Vol. 10/27, https://doi.org/10.1186/s13062-014-0027-4.

Smalheiser, N.R., W. Shao and P.S. Yu (2015), “Nuggets: Findings shared in multiple clinical case reports”, *Journal of the Medical Library Association*, Vol. 103/4, pp. 171-176, https://doi.org/10.3163/1536-5050.103.4.002.

Suadaa, L.H. et al. (2021), "Towards table-to-text generation with numerical reasoning", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, on line, https://doi.org/10.18653/v1/2021.acl-long.115.

Swanson, D.R. (2011), “Literature-based resurrection of neglected medical discoveries”, *Journal of Biomedical Discovery and Collaboration*, Vol. 6, pp. 34-47, https://doi.org/10.5210/disco.v6i0.3515.

Swanson, D.R (1986), “Undiscovered public knowledge”, *The Library Quarterly: Information, Community, Policy,* Vol. 56/2, p. 103118, https://doi.org/10.1086/601720.

Swanson, D.R. and N.R. Smalheiser (1997), “An interactive system for finding complementary literatures: A stimulus to scientific discovery”, *Artificial Intelligence*, Vol. 91/2, pp. 183-203, https://doi.org/10.1016/S0004-3702(97)00008-8.

Szczypińsk, F.T. et al. (2021), “Can we predict materials that can be synthesised?”, *Chemical Science*, Vol. 12/3, pp. 830-840, https://doi.org/10.1039/D0SC04321D.

Torvik, V.I. and N.R. Smalheiser (2007), “A quantitative model for linking two disparate sets of articles in MEDLINE”, *Bioinformatics*, 1 July, Vol. 23/13, pp.1658-1665, https://doi.org/10.1093/bioinformatics/btm161.

Tshitoyan, V. et al. (2019), “Unsupervised word embeddings capture latent knowledge from materials science literature”, *Nature,* Vol. 571/7763, pp. 95-98, https://doi.org/10.1038/s41586-019-1335-8.

Zhang, R. et al. (2021), “Drug repurposing for COVID-19 via knowledge graph completion”, *Journal of Biomedical Informatics*, Vol. 115, p. 103696, https://doi.org/10.1016/j.jbi.2021.103696.

Zhao, W., R. Wu and H. Liu (2016), “Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target”, *Information Processing and Management,* Vol. 52/5, pp. 976-988, https://dl.acm.org/doi/abs/10.1016/j.ipm.2016.04.004.

Note

[1] “Interestingness” in data mining is a broad concept encompassing such ideas as reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability.

Advancing the productivity of science with citizen science and artificial intelligence

L. Ceccaroni, Earthwatch, United Kingdom

J.L. Oliver, University of Sydney, Australia

E. Roger, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

J. Bibby, Industry and Environment, Australia

P. Flemons, Australian Museum, Australia

K. Michael, Arizona State University, United States

A. Joly, Institut national de recherche en sciences et technologies du numérique (INRIA), France

Introduction

Citizen science is a form of scientific inquiry where members of the public engage in scientific investigations, often in collaboration with, or under the direction of, professional scientists and scientific institutions. It supports scientific research and applied sciences through a wide range of activities and across diverse topics. Thanks to advances in communication and computing technologies, the public can collaboratively participate in new ways in citizen science projects. For example, participants submit observations and samples about the environment via eBird, iNaturalist or the EchidnaCSI project, among other platforms. They also engage on line by transcribing historical documents or classifying photographs, audio and video via platforms such as DigiVol or Zooniverse. In other cases, participants collaboratively solve mathematical problems via the Polymath Project, or play online games via Foldit, to inform medical research. The public disseminates project outcomes as well.

To date, the most significant impact of citizen science in accelerating scientific discoveries has been in relation to data collection and processing activities (Bonney et al., 2016). Citizen science continues to gain support and acceptance, delivering positive societal, economic and environmental impacts. Many projects actively support learning about specific topics, increase understanding of science and inform decision making (Bonney et al., 2014). Citizen scientists are involved in projects across scientific domains such as astronomy, chemistry, computer science, environmental science, mathematics, medicine and social science. However, the vast majority of citizen-science projects support the understanding of biodiversity, wildlife, plants and environmental processes (Kullenberg and Kasperowski, 2016).

Intelligence demonstrated by machines, known as artificial intelligence (AI), is widely applied across various scientific domains. Citizen science is no exception, and is increasingly being enhanced by the integration of AI (Ceccaroni et al., 2019). This essay examines the synergy of AI and citizen science to improve the productivity of science. It concludes by exploring future opportunities and considerations in this emerging area, including policy implications.

How citizen science coupled with artificial intelligence can increase the productivity of science

Over the past decade, there has been huge growth in the capabilities and applications of AI in citizen science. These applications can take an unsupervised or supervised machine-learning approach. In the former, data do not have to be annotated accurately by people first. In the latter, which occurs more frequently, data labelled by humans are needed to train the AI algorithms. At present, citizen science systems using AI are advancing science through a variety of mechanisms:

- increasing the speed and scale of data processing
- increasing projects' temporal and geographical scope
- improving the quality of data collected and processed
- supporting learning between humans and machines
- leveraging new data sources
- diversifying engagement opportunities.

These mechanisms are detailed below, and current examples are provided.

Increasing the speed and scale of data processing

Cameras triggered by motion typically capture many photos of moving vegetation rather than the intended animals moving past. Audio, video and other media can often be filtered using similar machine-learning techniques. In this case, AI algorithms can reduce how much data need to be processed by humans. AI is used to filter out false positives in images so that citizen scientists are more likely to see photographs of animals that need identification (Willi et al., 2018). More robust integration of AI and citizen science applied to the ever-growing volume of measures used by ecological studies will lead to more conclusive environmental insights at scale (Tuia et al., 2022). A similar filtering technique is applied in Galaxy Zoo, an online citizen science project. In this project, participants classify types of galaxies based on visible features in satellite photographs. The analysis of large amounts of data is facilitated by image pre-processing performed by AI. Here, the combination of humans and machines, often referred to as human-machine teaming, increases the rate of data processing (Beck et al., 2018).

Increasing projects' temporal and geographical scope

There is growing awareness of the potential of citizen science (integrated with AI) to expand environmental monitoring programmes. These include projects where solutions depend on large numbers of observations distributed across space and time (McClure et al., 2020). The Pl@ntNet citizen-science platform, for instance, includes tools to identify plants automatically. This has resulted in citizen scientists contributing more accurate data to global repositories and monitoring projects (Bonnet et al., 2020). Similarly, the iNaturalist project include tools to automatically identify most species, respectively. This has enabled the collection of observations at temporal and spatial scales not achievable with traditional science.

Improving the quality of data collected and processed

Several highly successful projects use AI to improve the quality of data collected and processed. Through the global platform eBird, birdwatchers have submitted copious bird observations, which have informed development of species distribution models (Sullivan, et al., 2014). These models have subsequently been applied to improve data quality by automatically filtering out observations of bird species residing outside of the birdwatcher's location (Kelling et al., 2012).

Supporting learning between humans and machines

Citizen scientists can contribute to training AI to solve complex analytical tasks usually carried out by experts. Human-in-the-loop processes are systems built with human supervision at different stages of the project cycle. For example, humans create and label datasets that are then used to train AI algorithms and models, with humans overseeing the models and fine-tuning them. Humans can also test and validate these models, resulting in high-quality AI systems. Several large citizen science projects focused on identifying species, such as iNaturalist, Pl@ntNet and BirdNet, are strongly enhanced by adopting a human-in-the-loop approach. In some cases, these types of human-AI systems can train AI algorithms to recognise species almost as accurately as humans with species expertise (Bonnet et al., 2018). In the online Gravity Spy project, participants identify glitches in visual representations of data from interferometers to assist scientists' search for gravitational waves; AI is used to train newcomers to learn more quickly (Jackson et al., 2020). Such AI integrations make projects more efficient.

Another example is a monitoring project called Penguin Watch, in which humans analyse time-lapse images of penguin colonies (Jones et al., 2020). This analysis by volunteers greatly helps assess the reliability of the AI algorithm used to identify species. It also helps refine it in different conditions (day and night) and for the different species.

In iNaturalist, AI provides participants with immediate feedback, derived from computer-vision models, about the organisms (plants, animals or fungi) in the photographs submitted. This feedback is an opportunity for citizen scientists to learn more about biodiversity, and has the potential to maintain their engagement in the project (Van Horn et al., 2018). Other, more expert members of the community identify species more specifically or validate AI identifications. Such contributions are used to refine the computer-vision algorithms (Van Horn et al., 2018).

Leveraging new data sources

Tapping into non-traditional data sources, such as social media, with the support of AI (data filtering), can vastly enhance the temporal and geographic availability of data and collect real-time information (MacDonald et al., 2015). In the Aurorasaurus project, participants submit observations and verifications of aurora sightings. The project is relatively novel in aggregating observations from both direct submissions through the project website and social media. Several other projects (particularly weather observation projects) are also starting to harvest data from social media platforms such as Twitter to increase the amount of available data for analysis (MacDonald et al., 2015).

Diversifying engagement opportunities

The use of AI offers more ways for participants to take part, and increased engagement provides more information for scientific investigations. Some people enjoy searching through a lot of data to find something uncommon. Participants may, for example, hope to see wildlife captured in photographs from

motion-triggered cameras (Bowyer et al., 2015) or hear the calls of a rare bird species (Oliver et al., 2019). In some cases, AI can be trained to quickly perform tasks that might be considered time-consuming or uninteresting to some participants. This allows the citizen science community to engage with tasks that are considered more exciting and challenging (Ceccaroni et al., 2019). In some camera trap projects, AI is used to remove false positives in images. This enables citizen scientists to focus on identifying animals just in the pictures where an animal occurs, saving their time. In the iNaturalist application, AI assists species identification and increases biodiversity knowledge in participants using the platform (Unger et al., 2021).

Future applications

Opportunities exist for further growth of AI-supported citizen science. These include developing new AI applications; more accessible ways for non-experts to use AI techniques; and increased private investment in AI, similar to Microsoft's existing investment in “AI for Earth” (Joppa, 2017). Realising these opportunities will likely result in more participants using AI-assisted citizen science applications (Rzanny et al., 2022). It can also lead to including more citizen science data in international data repositories. This is useful because these data are generally more accessible to the public, researchers and policy makers. In the future, AI will be increasingly applied in citizen science. Applications will include autonomous systems of all types, such as drones, autonomous vehicles, and other robotic and remote sensing instrumentation that is integrated with AI. It will also include improvements in mobile applications and hardware, and communication technologies such as wireless broadband networks and cloud computing. All these emerging applications will give rise to new capabilities, particularly in data collection and in the automatic detection and identification of items in images, audio recordings or videos.

In integrating AI and citizen science, risks, traceability, transparency and upgradability of AI algorithms and AI-assisted information systems must be carefully considered (Ceccaroni et al., 2019; Ponti et al., 2021). 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.

Additionally, quantifying uncertainty is essential. In the case of citizen science, uncertainty originates from any error or bias in the data collection, classification or processing resulting from AI algorithms (e.g. results, predictions) or participants, and from natural data variance. It is crucial to maintain meta-information on how the data have been treated throughout the data’s life cycle. Tracking uncertainty can ensure that the related variables and biases (e.g. errors in an observation map that may affect subsequent decisions) are findable, accessible, interoperable and reusable (Wilkinson et al., 2016). A first step to achieving this, in relation to biodiversity, could be integrating the uncertainty associated with species identification into Darwin Core (i.e. a broadly accepted biodiversity data standard). This information could then be made searchable in biodiversity data applications. The allowable uncertainty in data ultimately depends on how the data are being used. Data quality cannot be reduced to a binary attribute (usable vs. unusable). For example, the construction of a species distribution model can tolerate a certain percentage of error in the input data (Botella et al., 2018). However, a single erroneous observation can severely impact a warning system based on the early detection of certain species (Botella et al., 2018).

Policy considerations

As technology improves, machines will perform more of the heavy data processing and time-consuming aspects of citizen science projects. This provokes several questions. How will citizen scientists be motivated to maintain their involvement in projects? How can they be engaged with learning? How can they be educated? How can their contributions be appropriately attributed and acknowledged? How can their time and effort be rewarded? Finally, how can data exploitation and ownership be managed? (Franzen et al., 2021; Ponti et al., 2021). Without resolving these challenges, the interest and participation in citizen science might decrease. It will be an ongoing challenge to ensure these issues are adequately considered. At the same time, these issues should not hinder or limit citizen science and detract from its appeal. Indeed, AI could attract more people to citizen science because some (youth, for example), especially curious about AI, might be drawn to the domain.

Policy makers should dedicate resources to generating creative ideas on how AI could help advance science productivity with citizen science around the following issues:

- **Expansion of the range of science project types that can use citizen science**. To date, the area has been primarily dominated by projects on biodiversity, wildlife, plants and environmental processes. Typically, these research domains have a more extended history of readily engaging the public. AI's contribution to these areas has henceforth evolved the most.
- **Best practice guidance for scientists, technologists and broader groups so they can adopt a citizen science approach**. Guidance is especially needed for breaking complex research projects into discrete tasks that citizen scientists can then undertake. AI could assist in this partitioning of tasks.
- **Validation of citizen science contributions by quantifying the accuracy of output**. AI could help ensure adherence to the scientific method and assist in quality and impact assessment, whose metrics continue to be challenging for citizen science projects to report on (Wehn et al., 2021). Improved reporting measures could help alleviate long-running concerns over data quality that remain prevalent in citizen science and science more broadly.
- **Proper application of AI**. Joppa (2017) suggests that, for every problem, two questions should be asked: "How can AI help solve this?" and "How can we facilitate its application?" An additional question should also be asked: "How can we ensure that each use of AI in citizen science carefully considers risks, traceability, transparency and upgradability?"

Conclusion

Citizen science at local, national and global scales represents an opportunity for a shift in the ability to inform scientific inquiries, enrich lives and engage diverse communities in science. As citizen science grows, new technologies will likely proliferate, supporting people in learning, exchanging information and solving problems collaboratively. With these new technologies making data more accessible and interpretable, new opportunities for synergies between citizen science and AI are likely to emerge. This essay has described how AI, coupled with citizen science, can enhance the productivity of science.

The new technologies, which will integrate AI into citizen science and facilitate automation, also come with potential risks. Project leaders will need to consider these risks and how best to mitigate them to ensure transparency and positive outcomes. The success of this integration, in terms of increasing the scientific and public benefit and enhancing the productivity of science, will require continued investment. It will also demand consideration in areas such as ethics, motivations and attribution for diverse groups of participants, system development, system optimisation, data quality and impact assessment.

References

Beck, M.R. et al. (2018), "Integrating human and machine intelligence in galaxy morphology classification tasks", *Monthly Notices of the Royal Astronomical Society*, Vol. 476/4, pp. 5516-5534, https://doi.org/10.1093/mnras/sty503.

Bonnet, P. et al. (2020), "How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools", *Ecological Solutions and Evidence*, Vol.1/2, p. e12023, https://doi.org/10.1002/2688-8319.12023.

Bonnet, P. et al. (2018), "Plant identification: Experts vs. machines in the era of deep learning", in *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, Springer, Cham.

Bonney, R. et al. (2016), "Can citizen science enhance public understanding of science?", *Public Understanding of Science*, Vol. 25/1, pp. 2-16, https://doi.org/10.1177/0963662515607406.

Bonney, R. et al. (2014), "Next steps for citizen science", *Science*, Vol. 343/6178, pp. 1436-1437, https://doi.org/10.1126/science.125155.

Botella, C. et al. (2018), "Species distribution modeling based on the automated identification of citizen observations", *Applications in Plant Sciences*, Vol. 6/2, p. e1029, https://doi.org/10.1002/aps3.1029.

Bowyer, A. et al. (2015), "Mundane images increase citizen science participation", presentation to 2015 Conference on Human Computation & Crowdsourcing, San Diego, https://doi.org/10.13140/RG.2.2.35844.53121

Ceccaroni, L. et al. (2019), "Opportunities and risks for citizen science in the age of artificial intelligence", *Citizen Science: Theory and Practice*, Vol. 4/1, p. 29, http://doi.org/10.5334/cstp.241.

Franzen, M. et al. (2021), "Machine learning in citizen science: Promises and implications" in *The Science of Citizen Science*, Springer, Cham.

Jackson, C. et al. (2020), "Teaching citizen scientists to categorize glitches using machine learning guided training", *Computers in Human Behavior*, Vol. 105/106198, https://doi.org/10.1016/j.chb.2019.106198.

Jones, F.M. et al. (2020), "Processing citizen science- and machine-annotated time-lapse imagery for biologically meaningful metrics", *Scientific Data*, Vol. 7/102, https://doi.org/10.1038/s41597-020-0442-6.

Joppa, L.N. (2017), "The case for technology investments in the environment", 19 December, *Nature*, www.nature.com/articles/d41586-017-08675-7.

Kelling, S. et al. (2012), "eBird: A human/computer learning network for biodiversity conservation and research", in *Proceedings of the Twenty-Fourth Innovative Applications of Artificial Intelligence Conference*, Vol. 26/2, AAAI Press, Palo Alto, https://doi.org/ 10.1609/aaai.v26i2.18963.

Kullenberg, C. and D. Kasperowski (2016), "What is citizen science? A scientometric meta-analysis", *PLOS ONE*, Vol. 11/1, p. e0147152, https://doi.org/10.1371/journal.pone.0147152.

MacDonald, E.A. et al. (2015), "Aurorasaurus: A citizen science platform for viewing and reporting the aurora", *Space Weather*, Vol. 13/9, pp. 548-559, https://doi.org/10.1002/2015SW001214.

McClure, E.C. et al. (2020), "Artificial intelligence meets citizen science to supercharge ecological monitoring", *Patterns*, Vol. 1/7, p. 100109, https://doi.org/10.3389/fmars.2022.918104.

Oliver, J.L. et al. (2019), "Listening to save wildlife: Lessons learnt from use of acoustic technology by a species recovery team", in *Proceedings of the 2019 Designing Interactive Systems Conference (DIS'19), 23-28 June, San Diego*, pp. 1335-1348, https://doi.org/10.1145/3322276.3322360.

Perry, T. et al. (2022), "EchidnaCSI: Engaging the public in research and conservation of the short-beaked echidna", *Proceedings of the National Academy of Sciences*, Vol. 119/5, p. e2108826119, https://doi.org/10.1073/pnas.2108826119.

Ponti, M. et al. (2021), “Can't we all just get along? Citizen scientists interacting with algorithms”, *Human Computation*, Vol. 8/2, pp. 5-14, https://doi.org/10.15346/hc.v8i2.128.

Rzanny, M. et al. (2022), “Image-based automated recognition of 31 Poaceae species: The most relevant perspectives”, *Frontiers in Plant Science*, Vol. 12, 26 January, https://doi.org/10.3389/fpls.2021.804140.

Sullivan, B.L. et al. (2014), “The eBird enterprise: An integrated approach to development and application of citizen science”, *Biological Conservation*, Vol. 169, pp. 31-40, https://doi.org/10.1016/j.biocon.2013.11.003.

Tuia, D. et al. (2022), “Perspectives in machine learning for wildlife conservation”, *Nature Communications*, Vol. 13/1, pp. 1-15, https://doi.org/10.1038/s41467-022-27980-y.

Unger, S, et al. (2021), “iNaturalist as an engaging tool for identifying organisms in outdoor activities”, *Journal of Biological Education*, Vol. 55/5, pp. 537-547, https://doi.org/10.1080/00219266.2020.1739114.

Van Horn, G. et al. (2018), “The iNaturalist species classification and detection dataset”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Institute of Electrical and Electronic Engineers, Piscataway, https://authors.library.caltech.edu/87114/.

Wehn, U. et al. (2021), “Impact assessment of citizen science: State of the art and guiding principles for a consolidated approach”, *Sustainability Science*, Vol. 16/5, pp. 1683-1699, https://doi.org/10.1007/s11625-021-00959-2.

Wilkinson, M.D. et al. (2016), “The FAIR Guiding Principles for scientific data management and stewardship”, *Scientific Data*, Vol. 3/1, pp. 1-9, https://doi.org/10.1038/sdata.2016.18.

Willi, M. et al. (2018), “Identifying animal species in camera trap images using deep learning and citizen science”, *Methods in Ecology and Evolution*, Vol. 10/1, pp. 80-91, https://doi.org/10.1111/2041-210X.13099.

What can artificial intelligence do for physics?

S. Hossenfelder, Frankfurt Institute for Advanced Studies, Germany

Introduction

In recent years, governments all over the world have launched research initiatives for artificial intelligence (AI). These range from Australia, Canada and the United States to the People's Republic of China, Denmark, the European Commission, France, Germany and the United Kingdom. Everyone suddenly has a strategy for "AI made in", whatever happens to be their own part of the planet. In the coming decades, it is likely that tens of billions of public and private dollars, euros and Yuan renminbi will flow into the field. However, ask physicists what they think of AI, and they will probably be surprised. For them, AI was trendy in the 1980s. They prefer to call it "machine learning" and pride themselves on having used that term for decades. This essay summarises different applications for which AI physicists use AI, classifying them roughly into data analysis, modelling and model analysis.

The evolution of machine learning in physics

Already in the mid-1980s, researchers working in statistical mechanics – a field concerned with the interaction of large numbers of particles – set out to better understand how machines learn. They noticed that magnets with disorderly magnetisation (known as "spin glasses") can serve as a physical realisation for certain mathematical rules used in machine learning. This, in turn, means the physical behaviour of these magnets sheds light on some properties of machines that learn, such as their storage capacity (Peretto, 1984). Back then, physicists also used techniques from statistical mechanics to classify the learning abilities of algorithms.

Particle physicists, too, were at the forefront of machine learning. The first workshop on Artificial Intelligence in High Energy and Nuclear Physics was held as early as 1990. Workshops in this series still take place but have since been renamed to Advanced Computing and Analysis Techniques. This may be because the new acronym, ACAT, is catchier. However, it also illustrates the phrase "artificial intelligence" is no longer in common use among researchers in physics.

Physicists avoid the term "artificial intelligence" because it reeks of hype and because the analogy to natural intelligence is superficial at best, misleading at worst. True, the current models are loosely based on the human brain's architecture. The term "neural networks" refers not to an actual structure, such as a neuron, but to algorithms based on mathematical representations of "neurons" connected by "synapses". Using feedback about its performance – the "training" – the algorithm then "learns" to optimise a quantifiable goal, such as recognising an image or predicting a data-trend.

This type of iterative learning is certainly one aspect of intelligence, but it is far from complete. The current algorithms rely heavily on humans to provide suitable input data. They do not formulate their own goals.

They do not propose models. They are, for what physicists are concerned, simply elaborate ways of fitting and extrapolating data.

So, what novelty can AI bring to physics? A lot, it turns out. The techniques are not new – even deep learning, a neural network with three or more layers, dates back to the early 2000s. However, today's ease of use and sheer computational power mean that computers can perform tasks previously reserved for humans.

Developments in AI have also enabled scientists to explore entirely new research directions. Until a few years ago, other computational methods often outperformed machine learning, but now it leads in many different areas. This is why, in recent years, interest in machine learning has spread into seemingly every niche of physics.

Most applications of AI in physics loosely fall into three main categories: data analysis, modelling and model analysis.

Data analysis

Data analysis is the most widely known application of machine learning. Neural networks can be trained to recognise specific patterns, and can also learn to find new patterns on their own. In physics, this is used in image analysis, such as when astrophysicists search for signals of gravitational lensing. Gravitational lensing happens when space-time around an object is deformed so much that it noticeably distorts the light coming from behind it. The recent, headline-making, black hole image is an extreme example. However, most gravitational lensing events are more subtle, resulting in smears or partial arcs of light. AIs can learn to identify them.

Particle physicists also use neural networks to find patterns, both specific and unspecific. Highly energetic particle collisions, like those done at the Large Hadron Collider, produce huge amounts of data. Neural networks can be trained to flag interesting events. Similar techniques have been used to identify certain types of gamma-ray bursts (Chen and Bo-Qiang, 2021). They may also soon help to find gravitational waves (George and Huerta, 2018).

Data analysis is not necessarily passive. Achieving fusion power requires solutions to the challenge of suspending a super-heated plasma in a torus of powerful magnets. Using AI to analyse the dynamics of the plasma and predict instabilities can help control a potentially chaotic system (Degrave et al., 2022).

Modelling

Machine learning aids the modelling of physical systems by both speeding up existing calculations and enabling new types. For example, simulations for the formation of galaxies take a long time even on the current generation of supercomputers. However, neural networks can learn to extrapolate from the existing simulations without re-running the full simulation each time. This technique was successfully used to match the amount of dark matter to the amount of visible matter in galaxies (Moster et al., 2021). Neural networks have also been used to reconstruct what happens when cosmic rays hit the atmosphere (Erdmann et al., 2018), or how elementary particles are distributed inside composite particles (Forte et al., 2002).

Model analysis

Machine learning is applied to better understand the properties of known theories that cannot be extracted by other mathematical methods or to speed up computation. For example, the interaction of many quantum particles can result in a variety of phases of matter beyond the commonly known gases, liquid, solids and superfluids. However, existing mathematical methods have not allowed physicists to calculate these phases. Neural nets can encode the many quantum particles and then classify the different types of behaviour.

Similar ideas underlie neural networks that seek to classify the properties of materials, such as conductivity or compressibility. 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. Machine learning is beginning to change that. Many hope it may one day allow physicists to find materials that are superconducting at room temperature. Success in this search would have major practical applications, from medicine to computing. Another fertile area for applications of neural nets is "quantum tomography", i.e. the reconstruction of quantum state from the measurements performed on it, a problem of high relevance for quantum computing.

Machine learning advances physics, but physics can in return advance machine learning. At present, it is not well understood just why neural nets work as well as they do. Since some neural networks can be represented as physical systems, knowledge from physics may shed light on how they operate.

Conclusion

The use of AI in physics is not new. However, today's ease of use, technical progress and enormous computational power mean that machine learning can rather suddenly allow physicists to tackle a lot of problems that were previously intractable. What does this mean for the future of physics? Will we see the "end of theory" predicted by Chris Anderson in his much-cited paper (Anderson, 2008)?

It is unlikely. There are many different types of neural networks, which differ in their architecture and learning schemes. Physicists have to understand which algorithm works for which situation and how well, the same process they went through for theory. Rather than spelling the end of theory, machine learning will take it to the next level.

References

Anderson, C. (2008), "The end of theory: The data deluge makes the scientific method obsolete", *Wired*, Vol. 16/7, www.wired.com/2008/06/pb-theory.

Chen, Y. and M. Bo-Qiang (2021), "Novel pre-burst stage of gamma-ray bursts from machine learning", *Journal of High Energy Astrophysics*, Vol. 32, pp. 78-86, https://doi.org/10.1016/j.jheap.2021.09.002.

Degrave, J. et al. (2022), "Magnetic control of tokamak plasmas through deep reinforcement learning", *Nature,* Vol. 602/7897, pp. 414-419, https://doi.org/10.1038/s41586-021-04301-9.

Erdmann, M. et al. (2018), "A deep learning-based reconstruction of cosmic ray-induced air showers", *Astroparticle Physics*, Vol. 97, pp. 46-53, https://doi.org/10.1016/j.astropartphys.2017.10.006.

Forte, S. et al. (2002), "Neural network parametrization of deep-inelastic structure functions", *Journal of High Energy Physics*, Vol. 5/062, http://doi.org/10.1088/1126-6708/2002/05/062.

George, D. and E.A. Huerta (2018), "Deep learning for real-time gravitational wave detection and parameter estimation: Results with advanced LIGO data", *Physics Letters B* 778, pp. 64-70, https://doi.org/10.1016/j.physletb.2017.12.053.

Moster, B.P. et al. (2021), "GalaxyNet: Connecting galaxies and dark matter haloes with deep neural networks and reinforcement learning in large volumes", *Monthly Notices of the Royal Astronomical Society*, Vol. 507/2, pp. 2115-2136, https://doi.org/10.1093/mnras/stab1449.

Peretto, P. (1984), "Collective properties of neural networks: A statistical physics approach", *Biological cybernetics*, Vol. 50/1, pp. 51-62, https://doi.org/10.1007/bf00317939.

AI in drug discovery

K.Z. Szalay, Turbine.AI, Hungary

Introduction

Artificial intelligence (AI) promises to de-risk the discovery process for new drugs. This essay explores how the pharmaceuticals industry has adopted a new business model to decrease risk in the early parts of drug discovery. It looks at how AI could speed up drug discovery through cost and time savings, and the role of "explainable AI" to bridge the gap between the pharma and software industries. Finally, as early discovery shifts from academia and large pharmaceutical companies to smaller start-ups and biotech spin-offs, it looks at the need for dedicated infrastructure.

AI will change drug discovery. The main challenge of bringing a new drug to market is that a lot of time and money are required before the drug's efficacy is revealed by testing on patients. As AI is integrated into ever-more steps in drug discovery, its main impact is in selecting experiments with the best chance of success, thereby de-risking the discovery process. Even a modest increase in efficiency can result in major savings by the time a drug gets to market.

The ability of AI systems to enhance drug discovery depends on the step in question. Explainable AI could have a major impact on the steps in drug discovery where AI is not yet widely used. However, AI approaches that are explainable by design are not yet good enough. Technical advances in explainability are still needed for "black-box" forms of AI, whatever their field of application.

Adoption of AI in the pharmaceuticals industry has been surprisingly rapid. AI solutions have a reputation for providing quick but unreliable predictions. To the degree this is the case, there is a tension with the safety focus of the pharma industry. Indeed, bringing AI directly to the bedside has not worked well (Herper, 2017). As long as AI stays within the confines of the R&D process, experiments can confirm its predictions before patients are involved.

Meticulous experiments to ensure patient safety will always be needed. However, the potential impact of AI is not to eliminate the need for clinical trials. Rather, it could create a situation in which new drugs fail less often when they eventually do get to clinical trials.

Starting in the late 1990s, productivity in the drugs industry saw a major decline (see the essay by Jack Scannell in this book). This decline continued well into the 2010s.

Fortunately, new technologies – particularly CRISPR and better prediction of drug safety – are helping avert a crash in drug discovery. The cost of new drug approvals had largely stabilised by the end of the last decade (Figure 1).[1] However, getting new drugs to market remains risky; failure rates are well above 60%, even for drugs that reach clinical trials (Wong, 2019).

Major pharmaceutical companies found a new business model to decrease risk in the early parts of drug discovery: in-licensing interesting compounds from smaller biotech companies. Pharmaceutical companies paid a premium for these compounds in exchange for small companies bearing the risks faced in the early

phases of discovery. The large companies did what they do best: capital-intensive clinical trials and commercialisation. This trend sped up during the last decade.

Large companies have complex processes in place, and change is hard for both individuals and organisations. Consequently, it has been in the agile small biotechnology companies where an explosion in the use of AI technologies has happened. AI solutions might already add significant value to drug discovery. However, many applications have not yet reached a level of maturity where they could be adopted inside the complex, optimised (and thus necessarily more rigid) processes of large pharmaceutical companies.

The promise of AI in drug discovery

More creative and faster experiments

How much of an impact on drug discovery should be expected from AI? Most experimentation is done in laboratories, measuring cell cultures, or drug binding using specialised assays[2] (often called wet-labs in contrast to the "dry-lab" work of experimenting on a computer). Even high-throughput wet-lab experimental assays are relatively slow and expensive. Human expertise is thus always required to pick the experiments that make sense to run, based on state-of-the-art science. As no such pre-selection is necessary to run AI predictions, machine-learning systems can help come up with novel ideas no sane human drug hunter would expect to work. While the value of such novelty generation is hard to quantify, the AI-generated list of experiments can be used to run in the lab to help derive savings. This approach would be shorter and possibly more successful than one without an AI system's guidance.

Reduced costs and time associated with failures at each stage

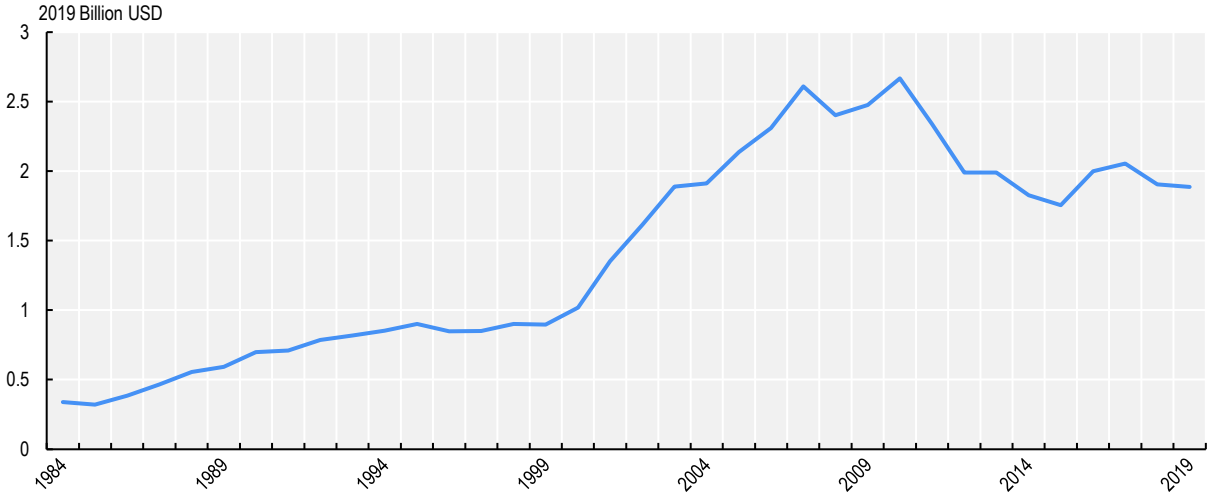
Besides finding novelty, the other major impact expected from AI is a decrease in the cost and time associated with failures at each stage of drug discovery. These reductions are quantifiable. Just decreasing the failure rate by 20% (e.g. from 30% to 24%) in each step of the discovery process would mean halving the total cost of any single project.

The data also show that, generally, decreasing the failure rate is the best option. However, in the earliest phases of drug discovery, saving costs through fewer experiments is more important than decreasing failure rates. Figure 2 shows the estimated cost savings from state-of-the-art AI guidance in all steps of drug discovery. It assumes that AI tool developers focus on applications that deliver the most impact. These savings would total slightly over a billion US dollars per new drug (Bender and Cortés-Ciriano, 2021).

Smaller size of minimum patient population needed to develop drugs

Decreasing the cost of R&D could lower the price of novel drugs, which are often prohibitively expensive for patients, and/or place major burdens on public budgets. It would also become feasible to start developing drugs for smaller patient populations, making drug discovery possible for rare diseases, as well as common diseases that need personalised approaches. For example, DeSantis, Kramer and Jemal (2017) highlighted that 20% of all cancers are rare and thus not covered by drugs targeted towards the most common cancers. Shrinking the size of the minimum patient population required to develop a new drug will be arguably the largest benefit of AI-based drug discovery over time.

Figure 1. Average real cost per new drug approval, USD billion, 1984-2019



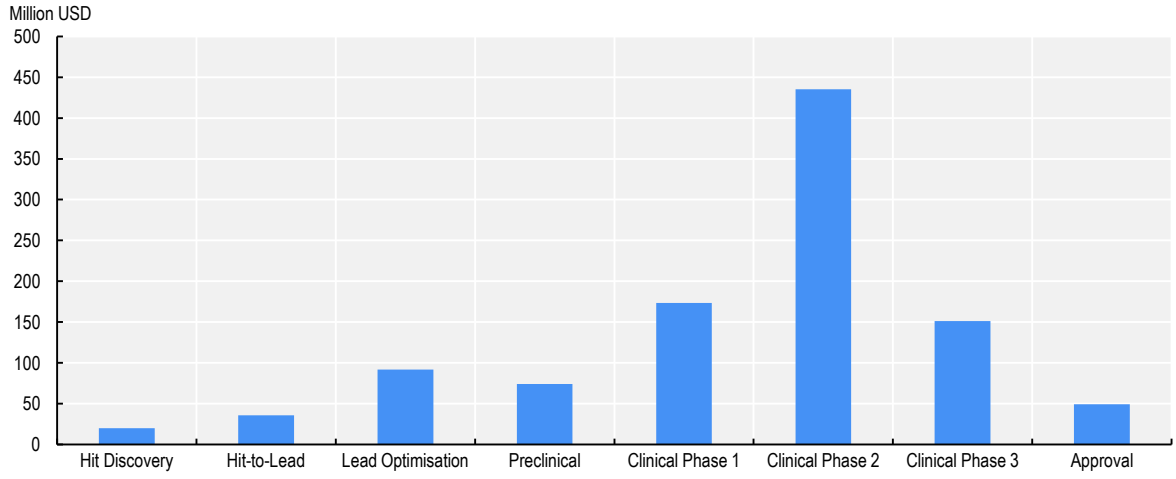
Note: Calculated from the annual R&D spending by member firms of the Pharmaceutical Research Manufacturers of America per Food and Drug Administration (FDA) approvals of new molecular entities (five-year moving average).
Source: Congressional Budget Office (2021).

How AI is used today in the different phases of drug discovery

Understanding how proteins fold

One AI application that made headlines recently was to better understand the structure and dynamics of potential drug targets. Deepmind's Alphafold2 is an AI system that helps researchers understand the 3D structure of proteins. This is an immensely important problem in biology because form defines function in the world of proteins. Indeed, most drug discovery processes start with finding the right protein to target for a given disease. Good experimental data exist on the shape of a small subset of proteins in the human proteome (the entirety of proteins that make up humans). However, science has no idea what the working, folded form of most human proteins looks like.

Figure 2. The net savings for a typical drug discovery project from decreasing the failure rate or cost by 20% (whichever is most impactful) in each phase of drug development



Source: Based on Bender and Cortés-Ciriano (2021).

With extensive training on experimental data, and a smart AI architecture,[3] Alphafold2 developed predictions that approximated experimental results in the CASP14 (CASP14, n.d.) challenge for a set of previously unpublished protein structures (Jumper et al., 2021). This paves the way to predicting with confidence the structure of previously unknown proteins.

Just a few months after the publication of the system, databases containing the Alphafold-predicted protein structures for all human proteins were made available to scientists (Varadi et al., 2022). This advance greatly helps researchers in the drug industry to find molecules to target a protein of interest as it is now possible to have a good understanding of what the target protein actually looks like.

Targeting the right protein ("target discovery")

The next step is to find which protein to target for a given disease. There is no single best way to find good drug targets in the lab. Different experimental assays have different strengths and weaknesses, which is also the case with AI methods in drug discovery. Different machine-learning systems – depending on the data they are trained on – will excel in addressing different analytic problems. For this reason, AI companies working on target discovery are proliferating, each developing its own discovery platform. These companies use diverse data gleaned from cell microscopy, electronic medical records, genetic databases and scientific literature, among others, creating their own drug target pipelines. While some of these methods may make their way into drug discovery, no drug based on a target discovered by AI has yet received approval from the US Food and Drug Administration (FDA). The first such drugs have only just entered human clinical trials (Jayatunga, 2022). This makes target discovery an exciting, fast-moving but still nascent area of AI in drug discovery.

Finding the right molecule ("hit identification")

Having an established protein target, all eyes are on the chemists to find the right molecule to inhibit the protein of interest – and preferably not much else. While established wet-lab methods can screen hundreds of thousands to millions of small molecules in just a few days, finding good hits (molecules that selectively bind to a given protein target) is still a daunting task. This is because the number of drug-like chemicals could be in the order of 10^{60}, a million times the number of the atoms on Earth (Bohacek, McMartin and Guida, 1996).

Indeed, finding good hits was one of the earliest frontiers for machine learning in drug discovery. Virtual screening methods involve computational discovery of molecules that might bind to a specific target of interest. "Molecular docking", for example, tries to find matching surfaces of the candidate molecule and the target protein.

Virtual screening can search a space much larger than is possible with wet-lab screening. This ranges from 10^9 molecules in commercially available virtual screening platforms to 10^{15} or more in proprietary pharma libraries. This, in turn, represents 4 to 9 orders of magnitude of difference compared to the 10^6 molecules in wet-lab screening (Hoffmann and Marcus, 2019).

Virtual screening tools have become increasingly sophisticated in the past two decades (Goodsell et al., 2021), with deep-learning methods recently joining the field (Wallach, Dzamba and Heifets, 2015). While virtual screening might be the most established subfield in which AI helps in drug discovery, the field is still far from able to exploit most of the drug-like chemical space. In general, AI techniques work well in predicting how new combinations of already measured building blocks behave. However, they cannot predict the behaviour of new building blocks (in this case, the behaviour of novel chemical structures).

Generating a more refined molecule ("lead optimisation")

In drug discovery, a long iterative chemistry process follows after the first promising molecules are found. This process aims to generate a more refined molecule (called a lead) – a molecule that has better

selectivity, absorption and distribution properties. This is done to arrive at a molecule that can eventually be administered *in vivo*.[4] AI can assist the lead optimisation process as well, but the next quantitatively different stage is for scientists to start planning experiments in animals. The first two key research tasks are to ensure the compound is not toxic and that it is efficacious (improves the disease status).

Assessing the toxicity and metabolic properties of drugs has also been a mainstay of computational drug discovery. In recent decades, models have improved greatly thanks to large-scale public data generation efforts (Kleinstreuer et al., 2014), as well as general progress in AI. While surprises still happen, most pharma companies have already integrated some metabolism and toxicity modelling solutions into their main pipelines.

Identifying biomarkers through drug repurposing ("preclinical and clinical stages")

The final step, and unfortunately the hardest, is predicting *in vivo* efficacy before administering the molecule in animals, and thereafter humans. The aim is to tell, with some reasonable degree of accuracy, which patients will respond well enough to a drug using "biomarkers" of efficacy. Traditional biomarkers are measures from blood tests or microscopic findings from a biopsy, but molecular genetic tests are being used more frequently, illustrating how medicine is increasingly personalised. However, good biomarkers are hard to find.

Any veteran drug hunter will readily say that getting a good biomarker for a drug is hard, even with support from AI, even if biomarkers are crucially important for success in the clinic (Wong, 2019). Alas, finding reliable biomarkers is not a problem well suited to most AI methods. 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 tumour composition has likely changed. This essentially renders them – for training purposes – a different patient. Both considerations make it extremely hard to generate the data with which to train an AI system to find strong biomarkers without having patient data in advance for that specific drug.

One way drug discovery teams are circumventing the problem of finding good biomarkers is through *drug repurposing*. This involves taking a drug (either an approved drug or one that has failed but was shown to be safe) and using its trials' data to train the AI. The goal is to identify a new biomarker that the original research team missed or deprioritised. In practice, however, AI-based repurposing approaches have not been wildly successful. How much AI would eventually be able to contribute here remains to be seen.

Explainable AI is key to bridging the pharma and the software industry

Bridging the gap between pharma and software

One difficulty of introducing new AI methods to drug discovery is a deep cultural divide. AI comes from the software world where a practice of "move fast and break things" is doable and mostly works well. On the other hand, as evident from the above discussion, safety is deeply embedded in the culture of drug discovery. Bringing any new drug to market is already extremely risky. Consequently, novel, unproven drug ideas understandably gain relatively little traction when companies need to commit years and hundreds of millions of US dollars to prove their efficacy.[5]

Moving towards "explainable AI" is one way to bridge the gap between the dynamics of software development and the safety needs of the drug industry. Explainable AI is a concept introduced in response to the realisation that the best-performing AI systems (such as neural networks) yield results that are generally not explainable.

To help understand explainable AI, visual recognition is a useful analogy. Visual recognition is much more complex, opaque and less conscious than it seems. People, for example, cannot define what exactly triggers the internal, instantaneous visual recognition of a cat. They might rationalise they are seeing pointy

ears, prominent whiskers and so on. However, other animals meeting those criteria can easily be found. This is exactly how a non-explainable AI system (like an artificial neural network) works.

In other machine-learning architectures, such as decision trees,[6] the decision process and the learnt rules are clearly understandable, even for an untrained human observer. Alas, these interpretable architectures are widely believed to offer worse prediction performance (Gunning and Aha, 2019). Explainable AI systems are not, in theory, necessarily worse than non-explainable black-box ones. However, the leading models – deep-learning systems – are not explainable. Thus, choosing an explainable AI architecture for any problem is not straightforward.

Explainability in drug discovery

Two arguments underscore the importance of explainability in drug discovery.

First, scientists in drug discovery already make use of statistical rules of thumb like Lipinski's Rule of Five (a simple set of four rules for what a candidate drug molecule with good bioavailability should look like). The community accepts those rules are not 100% accurate. However, every individual rule makes scientific sense (the molecule should not be too big or too charged, for example).

Second, discovery of a drug is not finished until the molecule gets regulatory approval. Unintended side effects or undesired metabolic properties of the candidate molecule often surface, requiring constant tweaking of either the molecule structure or the target patient population. That is not possible unless the discovery team has a good understanding of why the molecule works the way it does, where it binds and the biological mechanism that the drug should inhibit.

It is possible to get some of the "why" from black-box AI models by using additional external interpretation algorithms. However, a simple ranked list of, for example, indicated targets (e.g. 1. colon carcinoma, 2. non-small-cell lung cancer) is not good enough for drug discovery. The AI team always needs to provide an explanation, one way or another, to ensure the smooth adoption of their results.

Explainability is also important as it enables detection in the data of "minority bias". As most of the genomic data in published databases comes from Caucasians, learning algorithms often have a hard time picking out disease patterns specific for some other ethnic groups. For instance, there are slight but significant differences in the way African Americans metabolise certain drugs, requiring a different dosing schedule.

Recognising a pressing need, the FDA is already working to develop a good regulatory framework for AI in medicine (FDA, 2021). Specifically mentioned in the action plan are a) transparency to users; b) recognition and minimisation of minority bias in data; and c) honouring what is known as "good machine-learning practice". Choosing an AI that is interpretable by design would most likely result in a shorter time to compliance than it would if using a black-box model.

Modern AI needs dedicated infrastructure

As previously noted, early discovery is shifting from academia and large pharmaceutical companies to smaller start-ups and biotech spin-offs. One reason behind that shift might be the infrastructure needs of modern AI. The much-publicised breakthrough models like Alphafold2 and GPT-3 contain billions of parameters and are trained for weeks on hundreds of specialised processors (Jumper, 2021). In such huge AI systems, every training run costs tens of thousands of US dollars, and one has to run training sessions continuously to keep improving the models. This puts a large financial burden on smaller academic groups.

Another challenge for large modern AI set-ups is moving all the pieces of data and the code together at such 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. Otherwise, this would be a major paradigm shift in previously biology-intensive firms.

To address these challenges, academic groups would need a stronger AI backbone like, for example, 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 (n.d.) or ELIXIR (n.d.) have been established recently in the European Union to further collaboration in the field. However, they are mostly focused on sharing data and tools rather than solving the problem of scaling AI in academia.

Conclusion

AI in drug discovery is not a new phenomenon. Machine learning has been an integral part of generating small molecule targets for decades. Recent and ongoing improvements in AI have allowed it to enter other parts of the drug discovery process, to cut costs and improve efficiency. Besides molecular docking and toxicity prediction, which were already staples of state-of-the-art drug discovery workflows, small biotech companies are piloting many new ways of using AI. This is accelerating the shift in the business model of big pharma. Rather than doing all the research in-house, these firms buy trial-ready compounds from external parties. While some of the needed steps are still unsolved, successful adoption of AI in the entire drug discovery pipeline could dramatically decrease drug development costs. This would enable the industry to make drugs for patient populations previously considered far too small to justify the expense.

References

Bender, A and I. Cortés-Ciriano (2021), "Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet", *Drug Discovery Today*, Vol. 26/2, pp. 511-524, https://doi.org/10.1016/j.drudis.2020.11.037.

Bohacek, R.S., C. McMartin C and W.C. Guida (1996), "The art and practice of structure-based drug design: A molecular modeling perspective", *Medicinal Research Reviews*, Vol. 16/1, pp. 3-50, https://doi.org/10.1002/(sici)1098-1128(199601)16:1%3C3::aid-med1%3E3.0.co;2-6.

CASP14 (n.d.), *14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction* website, https://predictioncenter.org/casp14/ (accessed 12 January 2023).

CBO (2021), *Research and Development in the Pharmaceutical Industry,* 8 April, Congressional Budget Office, Washington, DC, www.cbo.gov/publication/57025.

DeSantis, C.E., J.L. Kramer and A. Jemal (2017), "The burden of rare cancers in the United States", *CA: A Cancer Journal for Clinicians*, Vol. 67/4, pp. p. 261-272, https://doi.org/10.3322/caac.21400.

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).

ELIXIR (n.d.), ELIXIR website, https://elixir-europe.org/ (accessed 12 January 2023).

FDA (2021), *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*, US Food and Drug Administration, Washington, DC, www.fda.gov/media/145022/download.

Goodsell, D.S. et al. (2021), "The AutoDock suite at 30", *Protein Science*, Vol. 30/1, pp. 31-43, https://doi.org/10.1002/pro.3934.

Gunning, D. and D.W. Aha, (2019), "DARPA's explainable artificial intelligence (XAI) program", *AI Magazine*, Vol. 40/2, pp. 44-58, https://doi.org/10.1609/aimag.v40i2.2850.

Herper, M. (2017), "MD Anderson benches IBM Watson in setback for artificial intelligence in medicine", 19 February, *Forbes*, www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/.

Hoffmann, T. and G. Marcus (2019), "The next level in chemical space navigation: Going far beyond enumerable compound libraries", *Drug Discovery Today*, Vol. 24/5, pp. 1148-1156, https://doi.org/10.1016/j.drudis.2019.02.013.

Jayatunga, K.P. et al. (2022), "AI in small-molecule drug discovery: A coming wave?" *Nature Reviews Drug Discovery*, Vol 21/3, pp. 175-176, https://doi.org/10.1038/d41573-022-00025-1.

Jumper, J. et al. (2021), "Highly accurate protein structure prediction with AlphaFold", *Nature* 2, Vol. 596, pp. 583-589, https://doi.org/10.1038/s41586-021-03819-2.

Kleinstreuer N.C. et al. (2014), "Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms", *Nature Biotechnology*, Vol. 32, pp. 583-591, https://doi.org/10.1038/nbt.2914.

Morgan, P. et al. (2018), "Impact of a five-dimensional framework on R&D productivity at AstraZeneca", *Nature Reviews Drug Discovery*, Vol. 17, pp. 167-181, https://doi.org/10.1038/nrd.2017.244.

NAIRR Task Force (2022), *Envisioning a National Artificial Intelligence Research Resource (NAIRR): Preliminary Findings and Recommendations*, National Artificial Intelligence Research Resource Task Force, Washington, DC, www.ai.gov/wp-content/uploads/2022/05/NAIRR-TF-Interim-Report-2022.pdf.

Varadi, M. et al. (2022) "AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models", *Nucleic Acids Research*, Vol.7/50(D1), pp. D439-D444, https://doi.org/10.1093/nar/gkab1061.

Wallach, I., M. Dzamba and A. Heifets (2015), "AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery", *arXiv,* arXiv:1510.02855v, https://doi.org/10.48550/arXiv.1510.02855.

Wikipedia (n.d.), "Flowchart", webpage, https://en.wikipedia.org/wiki/Flowchart (accessed 10 September 2022).

Wong, C.H. et al. (2019), "Estimation of clinical trial success rates and related parameters", *Biostatistics*, Vol. 20/2, pp. p. 273-286, https://doi.org/10.1093/biostatistics/kxx069.

Notes

[1] Methodological advancements like AstraZeneca's 5R framework (Morgan et al., 2018), along with the FDA making changes in the approval process also played a role in increasing the chances of approval.

[2] An assay is a test of a substance to determine its quality or ingredients.

[3] An AI architecture is the set of trainable transformations used inside the AI that maps the input of the AI system (features of the drug and the target protein, for example) to the output (the predicted binding strength).

[4] The exact process is not as simple and linear as described here: there are in fact many iterations between the *in vitro* biology and chemistry and work on animal models before a drug is ready to enter clinical trials.

[5] The datasets used for training in drug discovery also differ significantly from most well-known AI problems like computer vision or natural language processing. In image processing, for example, labels are *unconditional* – a cat is a cat no matter the orientation or lighting. In drug discovery, most data are *conditional* – the presence of protein X is a good marker for the efficacy of drug Y, *but only in some specific forms of breast cancer*, for example. This makes data re-use much harder in drug discovery.

[6] A decision tree is a flowchart-like AI architecture in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes) (Wikipedia, n.d.). The paths from root to leaf represent classification rules.

Data-driven innovation in clinical pharmaceutical research

J. New, Center for Data Innovation, United States

Introduction

From screening chemical compounds to optimising clinical trials and improving post-market surveillance of drugs, the increased use of data and better analytical tools such as artificial intelligence (AI) could transform drug development. This will lead to new treatments, improved patient outcomes and lower costs. This essay examines the way data-driven innovation, particularly AI, is transforming the drug development life cycle and recommends policies to accelerate this transformation.

Several recent developments have already begun to transform the entire drug development life cycle. These include widespread adoption of electronic health records, availability of new data sources thanks to technologies like genetic sequencing and smart technologies, and maturation and increased adoption of AI technologies. This transformation is particularly apparent in the clinical research phase of drug development (a 2019 white paper, "The promise of data-driven drug development", from which this essay is an excerpt, examines the way data-driven innovation, particularly AI, is transforming the drug development life cycle, and recommends policies to accelerate this transformation (New, 2019)).

The US Food and Drug Administration (FDA) categorises the drug development life cycle into five stages: discovery and development, preclinical research, clinical research, FDA review and FDA post-market safety monitoring. This excerpt from New (2019) highlights the role data-driven innovation can play in improving the clinical research phase. The phase focuses on studying how a drug interacts with the human body through studies and clinical trials.

Transforming clinical research with AI and smart technologies

A major barrier to developing new treatments is the cost of evaluating candidate drugs for safety and efficacy. As of 2018, the average cost of an individual clinical trial was USD 19 million (JHSPH, 2018). This is consistent with a 2014 study from the US Department of Health and Human Services. This study estimated the total costs of Phase I, II, III and IV trials for a drug at USD 44-115.3 million (Setkaya et al., 2014). Improved use of data and analytics can significantly reduce the costs of clinical trials.

Improving use of data and AI to increase patient recruitment and engagement

One of the most promising ways 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, especially since there are no guarantees patients will show up. To minimise this risk, companies such as Trials.ai and Vitrana have developed AI systems that can

guide site-selection decisions. The systems analyse factors such as historical site-performance data and study requirements (Kaufman, 2018; Brown, 2019).

Several companies are using AI to improve patient recruitment directly. For example, Deep 6 AI analyses structured and unstructured clinical data to better identify patients that match trial criteria, allowing trial organisers to conduct more targeted recruitment (Kaufman, 2018; Brown, 2019). London-based Antidote uses machine learning (ML) for similar purposes. Indeed, the company claims ML enabled the referral of 8 000 patients for a clinical trial relating to Alzheimer's disease in under two months. Moreover, these referrals were seven times more likely to follow through with the recruitment process than those from other sources (Sennaar, 2019).

Using machine learning to maintain patient engagement in trials

Even when a trial has enough recruits, they must participate in the full trial for it to be successful. However, failure to engage participants properly can cause them to drop out or not adhere to trial rules, thereby reducing the trial's effectiveness. Palo Alto start-up Brite Health has developed a smartphone app that uses ML to improve and maintain patent engagement to reduce this risk. The app provides users with notifications and nudges them to perform required tasks and site visits. It also uses a chatbot that can make trial information more accessible to patients, while algorithms identify and flag indicators of patient disengagement for trial organisers (Sennaar, 2019).

In some cases, patients may end their participation in a trial due to the negative side effects of a treatment. Here, too, AI can help. Researchers have developed ML algorithms that can identify the fewest and smallest doses of a chemotherapy regimen that can still shrink brain tumours, thus reducing the toxicity of the treatment (Yuaney and Shah, 2018). In a simulated trial, the researchers' ML model reduced treatment potency by between 25-50% of all doses without reducing effectiveness (Yuaney and Shah, 2018). By minimising the risk of side effects, researchers can more reliably ensure patient adherence to a clinical trial (Harrer, 2019).

New technologies also make it possible to conduct decentralised and virtual clinical trials. This can both make it easier to recruit patients from a wide area and reduce overhead costs. In October 2017, life sciences company AOBiome Therapeutics completed a 12-week clinical trial of an acne drug that proved to be safe and effective (Mantel-Undark, 2018). Unlike a traditional clinical trial, however, participants completed the trial at home. AOBiome mailed participants either the drug or a placebo, along with an iPhone that came pre-loaded with an app for participants to take and share regular selfies of their acne, as well as communicate with study organisers throughout the trial (Mantel-Undark, 2018). This approach enabled an effective clinical trial with no in-person screening or site visits, which substantially reduced both costs and barriers to participation.

Pharmaceutical companies have been actively exploring the potential to replace or augment traditional in-person trials with data technologies. For example, the French company Sanofi launched a clinical trial that had required participants to regularly visit the trial site. This allowed organisers to collect data regarding participants' weight, blood pressure and blood glucose. They then extended the trial, giving participants connected sensors and wireless technology to record and share these data from their homes (Mantel-Undark, 2018). GlaxoSmithKline sponsored a study to demonstrate the feasibility of using a smartphone and app to record survey data from rheumatoid arthritis patients. It also used the phone's accelerometer to record wrist-motion exercises. The study found the accelerometer data could be much more accurate than motion-evaluation exercises performed in-person with a physician (Mantel-Undark, 2018). Finally, Novartis has partnered with Apple to use Apple's ResearchKit, to improve clinical trial recruitment and administration. The partnership helps researchers develop apps for smart devices to collect and share medically relevant data, such as biometric sensor data and user-inputted information (McConaghie, 2018).

Site visits can cost between USD 3 000-7 000 per patient, and studies can involve dozens of visits and hundreds of patients. Thus, the potential for remote data collection could dramatically reduce the cost of clinical trials (Mantel-Undark, 2018).

New technologies such as the Internet of Things provide opportunities to collect large amounts of data outside of a traditional health-care context, known as real-world data (RWD). This might provide valuable evidence to help inform drug evaluation, known as real-world evidence (RWE) (FDA, 2018). In December 2018, the FDA published the framework for its Real-World Evidence Program. It provides guidance about how to incorporate RWD into clinical trials to create meaningful RWE (FDA, 2018).

Recommendations

Policy makers can and should play a role in accelerating data-driven innovation in drug development, both to maximise the benefits of these new technologies and to mitigate potential risks.

Expand access to institutional and non-traditional data

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.

Modernise regulatory processes

Policy makers should modernise regulatory processes, including by expanding and fully supporting programmes to evaluate and share foreign clinical trial data.

Promote equity in drug development

Racial and ethnic minorities, as well as women, have been historically underrepresented in clinical trials. This has led to evaluation of drugs based on data unrepresentative of the general population (Castro, 2014; ACC, 2018). Policy makers should invest in programmes that promote equity in drug development.

Invest in human resources

Policy makers should invest heavily in developing a workforce with the necessary AI skills to develop and implement data-driven innovations at scale.

Conclusion

Data-driven innovation promises to be even more transformative in medicine than in many other sectors. The benefits of these technologies can lead to new and safer treatments, improved patient outcomes and lower costs. The clinical research phase of drug development is particularly ripe for this kind of disruption. Already, use of AI and other data-driven technologies is transforming this field.

References

ACC (2018), “Study explores representation of women in clinical trials”, 30 April, American College of Cardiology, Washington, DC, www.acc.org/latest-in-cardiology/articles/2018/04/30/16/43/study-explores-representation-of-women-in-clinical-trials.

Brown, C. (2019), "The next step: Using AI to formulate clinical trial research questions", 8 January, Anju Life Sciences Software, Phoenix, https://anjusoftware.com/about/all-news/insights/ai-trial-research-questions.

Castro, D. (2014), "The rise of data poverty in America", 10 September, Center for Data Innovation, Washington, DC, www2.datainnovation.org/2014-data-poverty.pdf.

FDA (2018), *Framework for FDA's Real-World Evidence Program,* December, US Food and Drug Administration, Washington, DC, www.fda.gov/media/120060/download.

Harrer, S. et al. (2019), "Artificial intelligence for clinical trial design", *Trends in Pharmacological Sciences*, Vol. 40/8, pp. 577-591, https://doi.org/10.1016/j.tips.2019.05.005.

JHSPH (2018), "Cost of clinical trials for new drug FDA approval are fraction of total tab", 24 September, Press Release, John Hopkins Bloomberg School of Public Health, Baltimore, www.jhsph.edu/news/news-releases/2018/cost-of-clinical-trials-for-new-drug-FDA-approval-are-fraction-of-total-tab.html.

Kaufman, J. (2018), "The innovative startups improving clinical trial recruitment, enrollment, retention, and design", 30 November, *MobiHealthNews*, www.mobihealthnews.com/content/innovative-startups-improving-clinical-trial-recruitment-enrollment-retention-and-design.

Mantel-Undark, B. (2018), "The search for new drugs is coming to your house", 30 August, Fast Company, www.fastcompany.com/90229910/virtual-clinical-trials-are-bringing-drug-development-home.

McConaghie, A. (2018), "Novartis and Apple to scale up clinical trial collaboration", 24 January, Pharma Phorum, https://pharmaphorum.com/news/researchkit-novartis-apple-scale-clinical-trial-collaboration.

New, J. (2019), "The promise of data-driven drug development", 18 September, Center for Data Innovation, Washington, DC, www2.datainnovation.org/2019-data-driven-drug-development.pdf.

Sennaar, K. (2019), "AI and machine learning for clinical trials – examining 3 current approaches," 5 March, Emerj, Boston, https://emerj.com/ai-sector-overviews/ai-machine-learning-clinical-trials-examining-x-current-applications.

Setkaya, A. et al. (2014), "Examination of clinical trial costs and barriers for drug development", submitted to the US Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, July, https://aspe.hhs.gov/system/files/pdf/77166/rpt_erg.pdf.

Yuaney, G. and P. Shah (2018), "Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection", in *Proceedings of the 3rd Machine Learning for Health Care Conference*, Vol. 85, pp. 161-226, http://proceedings.mlr.press/v85/yauney18a.html.

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

M. Galtier, Owkin, France

D. Meadon, Owkin, United Kingdom

Introduction

Every day, millions of patient data points are collected from a wide range of sources. Harnessing the insights held within this ever-increasing volume of data in a safe, secure and ethical manner is the key to unlocking better health outcomes for everyone. Through recognising patterns at a scale unachievable by humans, machine-learning models can power the discovery of these insights. However, when it comes to health care, the journey from big data to actionable insights is fraught with challenges. Health data are sensitive and require handling with care through tight regulation. This essay explores how this challenge can be overcome through federated learning (FL), a novel machine-learning technology. FL drastically reduces privacy concerns by keeping patient data stored securely onsite during model training. The essay also discusses the salient policy implications to realise the full potential of FL technology for patients, health-care systems and life sciences companies.

The data challenges

Technology based on artificial intelligence (AI), especially machine-learning approaches, can power scientific breakthroughs at an unprecedented speed and scale. This technology can help us answer salient questions in medical research, such as: which patients should be included in clinical trials? Which molecules are the most promising targets for drug development? What is the best way to extract information from a wide variety of medical images, such as chest x-rays and scans to detect cancer or the early onset of COVID-19?

Machine-learning approaches are, however, “data-hungry”; models need access to large and diverse datasets to learn, improve accuracy and remove bias (Cahan et al., 2019). Consider a model designed to predict heart attack symptoms in the UK population. Such a model is unlikely to be widely applicable if it has only trained on the data collected at, for example, local general practices in a London suburb with a predominantly young, white population. Machine learning will not successfully transition from research

settings into everyday clinical practice without large, diverse and multimodal data (i.e. omics, digital pathology, radiology, spatial biology and clinical) (Rieke et al., 2020).

Why is it hard to gather data in health care?

Most patient data are stored at different hospitals and research centres, and distributed across servers and databases. These "silos" make it challenging for researchers and models to access enough data to train accurate and robust predictive models. Other valuable data types, such as chemical compounds used for drug discovery and development, are usually kept in-house by pharmaceutical companies. These "chemical libraries" are unlikely to be shared readily due to the competitive nature of the health-care industry.

Data scientists have to deal with heterogeneous datasets from multiple sources and in different formats. Sources include electronic health records, national or international databases, and clinical trial results. Meanwhile, formats and modalities could include medical history, laboratory test results and radiology images. This heterogeneity can make data collection and preparation for analysis tedious, lengthy and costly.

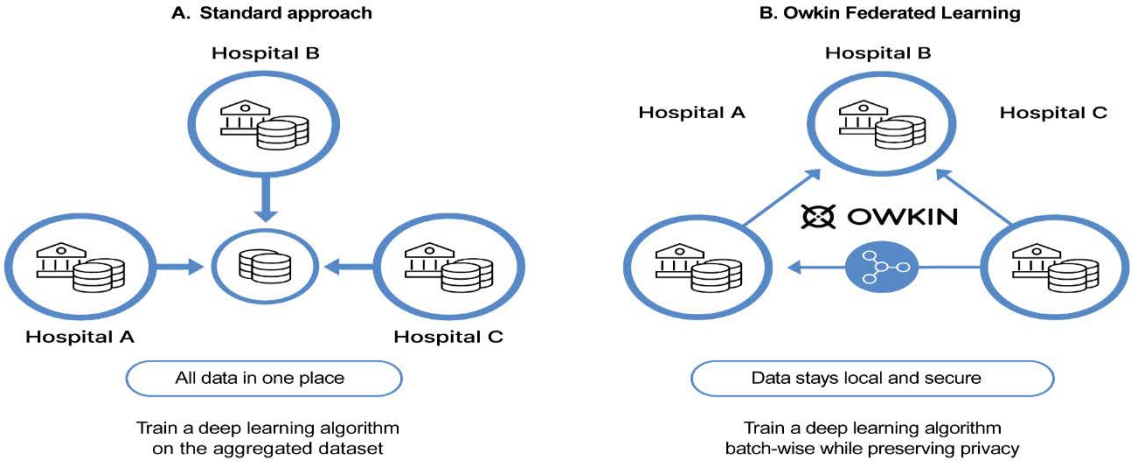
Protecting a person's identity in medical research through anonymised data is understandable but poses challenges for researchers. Completely removing identifiable information or, more commonly, using "pseudonymised" data (such as replacing a patient's name with a code) can decrease the performance of an algorithm (Rieke et al., 2020).

Consider the case of training a model to predict heart attack symptoms. This would not be clinically useful without patients' dates of birth or gender in the training dataset. Ultimately, whether anonymised or not, the data belong to patients, who may or may not consent to them being sold or transferred to third parties.

How does federated learning work?

FL emerged in 2015 to address data governance and privacy considerations by collaboratively training algorithms without exchanging the data (Figure 1). Algorithms are dispatched to different data centres, where they train locally. Once improved, they return to a central location, while the data stay local. The same algorithms are then sent to other local datasets to re-train and improve.

Figure 1. Standard machine learning versus federated learning approaches



In this way, FL enables researchers to gain insights collaboratively in the form of a consensus model without moving patient data beyond the firewalls of the institutions in which they reside. Since the machine-learning process occurs locally at each participating institution and only model characteristics are transferred, it greatly enhances patient privacy. Recent research has shown that models trained by FL can achieve performance levels comparable to those trained on centrally hosted datasets and superior to those models that only see isolated single-institution data (Huang et al., 2019).

Why is federated learning the future of health care?

The successful implementation of FL holds significant potential for enabling precision medicine at scale. Training algorithms on unprecedented quantities of heterogeneous datasets across multiple sources leads to models that yield less-biased decisions. These are sensitive to an individual's physiology while respecting governance and privacy concerns. FL still requires rigorous technical consideration for the algorithm to proceed optimally without compromising safety or patient privacy. However, it is a promising approach to creating powerful, accurate, safe, robust and unbiased models.

FL offers a solution for how to maximise the benefits of AI in health research while overcoming security and privacy issues. It is poised, therefore, to turbocharge medical research over the next few decades. In so doing, it will increase the pace and decrease the cost of developing treatments and diagnostics. At the same time, it will improve targeting and personalisation of existing treatments.

In a few years, FL will solve the problem of siloed datasets. It will allow diverse datasets from across the health sector, including from health-care settings and clinical trials, to be rapidly, safely and securely analysed. Instead of developing machine learning from scratch, foundation models for each medical field will emerge, enabling an ontology of tools and models to solve more specific problems. The impact of AlphaFold on the understanding of the protein structure will be replicated across different areas – from treatments to proteins to virus genomes.

From a patient's perspective, these FL-enabled foundation models will lead to clinical models that can be systematically checked by the US Food and Drug Administration (FDA) and other agencies. It will self-update, continually integrating and verifying clinical models that can be implemented in any medical device or clinical trial.

As patients become increasingly empowered to control their data, regulators are also implementing more rigorous data and privacy controls. In the coming decades, a collaborative process will likely be needed to publish clinical machine-learning models. The FDA would screen and verify datasets, and systematically remove any biases. Researchers will benefit from a decentralised but single source of truth for model building using FL.

What is happening in the federated learning space?

Recent years have seen a surge in real-world applications of FL in health care. The COVID-19 pandemic required scientific, academic and medical collaboration at an unprecedented scale. This accelerated the data sharing necessary to expedite critical research with direct impact on patient care. In one study, scientists leveraged the NVIDIA Clara FL platform to train the EXAM (electronic medical record chest X-ray) FL model across heterogenous, unharmonised datasets from 20 institutes. This allowed them to predict the future oxygen requirements of symptomatic patients with COVID-19. This, in turn, can aid physicians in determining the appropriate level of care (Dayan et al., 2021). NVIDIA Clara has also been used for a broad range of AI applications in medical imaging, genetic analysis and oncology. This includes segmenting pancreatic tumours or training a mammography classification model across five institutions (Wang et al., 2020).

New technology is emerging as the global research community looks to unlock the potential of FL. Large corporates including IBM have launched their own FL frameworks to build adverse drug reaction prediction models using electronic health data, among other applications (Choudhury et al., 2019). A proliferation of start-ups in the space offer their own potential platforms and solutions, but few have managed to apply these in real-world settings at scale.

The public sector has also become increasingly active. The UK Government, for example, outlined a plan to set up a federated infrastructure for management of UK genomics data. This would bring them together with other routinely collected health data to support work on cancer and rare diseases (UK Government, 2021).

Owkin Connect is an FL software that powers collaborations between hospitals, research centres, technology partners and life science companies (Owkin, 2022). It is used for clinical research in the HealthChain consortium and for drug discovery in the MELLODDY consortium (MELLODDY, 2022). The MELLODDY project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking. This, in turn, receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations. The HealthChain project has received funding from the French public investment bank (Banque Publique d'Investissement).

Life sciences

Enabled by an ever-expanding arsenal of model systems, analysis methods, libraries of chemical compounds and other agents (like biologics), the amount of data generated during drug discovery programmes has never been greater. However, the biological complexity of many diseases still defies pharmaceutical treatment. Coupled with the rise in regulatory expectations, this growing complexity has inflated the research intensity and associated cost of the average discovery project. It is, therefore, imperative to maximise learning from these data investments.

Owkin is the project co-ordinator for MELLODDY, a consortium of ten pharmaceutical and seven technical partners funded by the Innovative Medicines Initiative (IMI, 2022). The partners include well-known firms such as Janssen, AstraZeneca, Novartis, GSK, Bayer, Amgen, Astellas, Kubermatic, NVIDIA and KU Leuven. This landmark project demonstrated that collaborating in AI for drug discovery is possible at industrial scale.

Launched in 2019, MELLODDY announced the successful development and operation of a secure platform for FL without sharing each partner's proprietary data and models or compromising their security and confidentiality in 2020. In 2021, MELLODDY shared the first-ever demonstration of the predictive performance benefits of FL in drug discovery. The final results published in 2022 show that its operation at scale yields improvements across all pharmaceutical partners in the predictive performance of collaboratively trained models over single partner models. Models that more accurately predict the pharmacological and toxicological activities of molecules better support the decision-making process of which candidate drug molecules to make and test.

Clinical research

Through collaborative projects like HealthChain, Owkin has successfully trained machine-learning models on histology images, siloed at different clinical centres, to predict treatment responses in breast cancer. A model trained with Owkin Connect can now help oncologists select the most effective breast cancer treatment for each patient based on a single biopsy and identify high-risk patients for clinical drug trials.

Federated learning and data privacy

Ensuring the privacy and security of personal data has become a significant challenge posed by the rise of AI and machine learning. The standard approach of centralising data from multiple centres requires datasets to reside on a single server. This often means uploading private data to the cloud, sharing data with other parties and transferring vast amounts of data to a data centre for processing. Because FL is decentralised by design, and privacy preserving, it enables compliance with the General Data Protection Regulation (GDPR). The table below highlights the key difference in the impacts on data privacy between centralised machine learning and FL.

Table 1. Centralised machine learning versus federated learning impacts on data privacy

Centralised machine learning	Federated learning
With data merging, a personal data violation by a third-party intrusion can impact the whole dataset and paralyse research until the violation terminates.	Federated learning prevents suspension of the research, paralysing research only on the violated site. The violation concerns a limited set of data only.
The ability to centralise data in one dataset can impact personal data processing as it requires systematic production of a data impact assessment.	Data stay onsite, and interconnection is not systematic. In this situation, sites do not always require a data impact assessment.
Sites are identified as data co-controllers: a co-controller contract needs to be drafted and negotiated between the different sites.	The different sites are not identified as data co-controllers as the data processing is done between the site and the data processor (Owkin). As such, data processing is facilitated and there is no need to draft a co-controller contract.

Note: Machine-learning model parameters exchanged between parties in an FL system still conceal sensitive information, which privacy attacks can exploit. It is therefore crucial that federated learning is bolstered with efforts to ensure gold-standard privacy preservation. Owkin's software stack includes secure aggregation. This distributed cryptographic method, a combination of multi-party computation and homomorphic encryption (a form of encryption that permits users to perform computations on its encrypted data without first decrypting it), enables encrypted models to be averaged without being decrypted. As a consequence, the models are far more protected because only aggregated models are shared across network partners and the platform operator (Owkin) cannot decrypt the models.

Solving policy challenges using federated learning

FL can square the circle of harnessing the value of health-care data, while preserving patient privacy. This improved data security is critical in Europe. The GDPR has put individual privacy at the heart of the continent's approach to data governance. Moreover, the European Union will further strengthen its data protection rules through ePrivacy Regulation. This regulation will introduce strict obligations around data confidentiality and require user consent before metadata can be processed.

However, this is not just about Europe. Across the world, jurisdictions are following the European Union's lead in introducing sweeping new privacy frameworks. In 2022, the People's Republic of China (hereafter "China") enacted the Personal Information Protection Law. This comprehensive GDPR-like regulation imposes restrictions on data collection, transfer and analysis (Xu et al., 2021). Companies inside and outside of China will need to comply. In the United States, numerous individual states have passed their own privacy laws. Meanwhile, the introduction of federal regulation is increasingly a matter of "when" rather than "if". In other words, data innovation and respect for personal privacy will increasingly go hand in hand wherever one operates. By adopting FL, organisations can put themselves in a solid position to comply with present and future privacy regulations.

FL also aligns with a growing priority among policy makers: ensuring that data are stored and processed locally – otherwise known as data localisation or "data sovereignty". While data localisation requirements are more common in partner economies such as China and the Russian Federation, data localisation demands are becoming increasingly prominent among European and US policy makers (Svantesson, 2020).

The motivations behind these demands vary. Some want to safeguard national security. Others lack trust in the data protection standards of jurisdictions other than their own. Still others believe that storing and processing data locally is – or will soon be – required for economic competitiveness.

FL responds to these issues. It enables data to be analysed by algorithms where they are stored – rather than requiring them to be transferred and pooled elsewhere. It thus solves "poli-analysed" concerns, while safeguarding innovation.

FL also has a crucial role in helping policy makers unlock the full economic and social value of data. The experience of the COVID-19 pandemic has illustrated, like never before, the ability of data to help humanity solve complex collective challenges. These could range from tracking the emergence and spread of new variants to monitoring vaccine side effects in billions of individuals.

To bolster these capabilities, governments can take steps to harness the power of data across various areas – from treating disease and reducing emissions to combating financial fraud. For example, in May 2022, the European Commission presented its much-anticipated Health Data Space (HDS) (European Health Data Space, 2022). This health-specific ecosystem is comprised of rules, common standards and practices, infrastructures and a governance framework. It aims to empower individuals through increased digital access to, and control over, their electronic personal health data. At the same time, it fosters a genuine single market for electronic health records systems, relevant medical devices and high-risk AI systems.

The HDS also aims to provide a consistent, trustworthy and efficient set-up for the use of health data for research, innovation, policy making and regulatory activities (secondary use of data). The intention is to promote greater data sharing between businesses, researchers and public institutions to drive innovation and economic growth. FL provides the mechanism for achieving these aspirations. To that end, it enables insights generated from multiple datasets, while avoiding the technical challenges, confidentiality concerns and privacy risks that inevitably arise when pooling datasets.

Perhaps most important of all, through its emphasis on data privacy and security, FL could contribute to rebuilding the general public's damaged trust in the societal benefits of technology. This trust eroded in recent years due to the widespread sense that technology companies have abused the access to data that consumers have granted them. They believe these firms either failed to protect the confidentiality of that data or monetised them without meaningful user consent. The widespread adoption of FL could address these legitimate concerns and increase the public's willingness to share data that can drive innovation in health care and many other areas.

Conclusion

By enabling multiple parties to train collaboratively without the need to exchange or centralise datasets, FL addresses issues related to the transfer of sensitive medical data. Consequently, it opens novel research and business avenues and can improve patient care globally. FL is already having a significant impact on the health-care ecosystem. It is offering better diagnostic tools to clinicians by improving the analysis of medical images and other clinical data. It is driving precision medicine by helping identify patient subgroups to accelerate clinical trials. This acceleration decreases the cost and time-to-market for pharmaceutical companies, ensuring more patients receive the right treatment faster.

As a relatively new technology, FL will undoubtedly be an active research area throughout the next decade. It already offers exciting new opportunities for research breakthroughs by overcoming fundamental privacy concerns. With valuable use cases in health care expected to continue to emerge, more health-care partners will decide to use FL to collaborate safely and securely. This will lead to a true paradigm shift to make precision medicine a reality and ultimately improve patient care. By safely and ethically unlocking

the invaluable insights stored within patient data, FL will increasingly play a crucial role in ensuring every patient gets the right treatment.

However, this scenario may require some appetite from the public sphere. In geographies with stringent approaches to patient data usage, FL might be critical in facilitating access to data for secondary use, as well as for sensitive data like genomics. As an alternative to the dominant "hub approach" in which data are aggregated in a single place, FL's decentralised approach and cryptographic features offer political decision makers and public authorities a compelling solution to cybersecurity issues.

This public appetite may first come through public financing, especially in helping research centres to adopt a decentralised approach and to create shared infrastructure. It may also come through regulation, especially at the EU level. In that perspective, the EU HDS constitutes a strong opportunity to further increase policy makers' awareness on FL and to adopt a decentralised strategy.

References

Cahan, E.M. et al. (2019), "Putting the data before the algorithm in big data addressing personalized healthcare", *npj Digital Medicine*, Vol. 2/78, https://doi.org/10.1038/s41746-019-0157-2.

Choudhury, O. et al. (2019), "Predicting adverse drug reactions on distributed health data using federated learning", presentation to AMIA Symposium, https://research.ibm.com/publications/predicting-adverse-drug-reactions-on-distributed-health-data-using-federated-learning.

Dayan, I. et al. (2021), "Federated learning for predicting clinical outcomes in patients with COVID-19", *Nature Medicine*, Vol. 27, pp. 1735-1743, https://doi.org/10.1038/s41591-021-01506-3.

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).

Huang, L. et al. (2019), "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records", *Journal of Biomedical Informatics*, Vol. 99/103291, https://doi.org/10.1016/j.jbi.2019.103291.

IMI (2022), Innovative Medicines Initiative website, www.imi.europa.eu/ (accessed 25 November 2022).

MELLODDY (2022), MELLODDY website, www.melloddy.eu/ (accessed 25 November 2022).

Owkin (2022), "Owkin Connect", webpage, www.owkin.com/connect (accessed 25 November 2022).

Rieke, N. et al. (2020), "The future of digital health with federated learning", *npj Digital Medicine,* Vol. 3/119, https://doi.org/10.1038/s41746-020-00323-1.

Svantesson, D. (2020), "Data localisation trends and challenges: Considerations for the review of the Privacy Guidelines", OECD Digital Economy Papers, No. 301, OECD Publishing, Paris, https://doi.org/10.1787/7fbaed62-en.

UK Government (2021), "Genome UK: 2021 to 2022 Implementation Plan", www.gov.uk/government/publications/genome-uk-2021-to-2022-implementation-plan/genome-uk-2021-to-2022-implementation-plan.

Wang, P. et al. (2020), "Automated pancreas segmentation using multi-institutional collaborative deep learning", *arXiv,* arXiv:2009.13148 [eess.IV], https://arxiv.org/abs/2009.13148.

Xu, K. et al. (2021), "Analysing China's PIPL and how it compares to the EU's GDPR", International Association of Privacy Professionals", 24 August, https://iapp.org/news/a/analyzing-chinas-pipl-and-how-it-compares-to-the-eus-gdpr.

Part III The near future: challenges and ways forward

Artificial intelligence in scientific discovery: Challenges and opportunities

R. King, Cambridge University, United Kingdom

H. Zenil, Cambridge University, United Kingdom

Introduction

There have been cycles of hype surrounding the contributions of artificial intelligence (AI) to science (scientific discovery). However, progress has accelerated over the last decade, with machine learning (ML) now arguably one of the most exciting technologies. Indeed, the largest companies in the world have ML at the core of their technology, including Google, Facebook, Microsoft and Amazon. This essay explores challenges and opportunities associated with various forms of ML.

There are two main forms of ML: statistical and model-driven. Statistical ML, the most commonly used and successful form, is based upon complex pattern learning. It finds regularities in data, whose meaning can then be interpreted or studied further.

Statistical ML, including deep learning, is still dominant (deep learning is a type of statistical ML based on neural networks with many layers). This dominance occurs even in cases where statistical ML is ill-equipped to deal with basic symbol manipulation such as algebra and causality.

Despite the continued dominance of statistical ML, there is a trend towards approaches that construct an abstract model or representation, as humans do, rather than the statistical fitting of high-dimensional data. Such approaches are referred to as “causal” and “model-driven”.

Model-driven approaches generate mechanistic models from the data consistent with the data themselves that can be tested against newly generated data. “Mechanistic” means they can be followed state by state, as in a dynamic system, through a chain of cause and effect.

The distinction between model-driven and statistical ML is not always clear in the literature. Indeed, some statistical ML models are called “causal” or “model-driven”. This essay distinguishes between them based on the abstraction and generalisation capabilities of methods, and their ability to build mechanistic models from first principles, as scientists do.

Limitations and challenges

One way to grasp the promise of AI in science is to understand its current limitations and challenges.

Scalability

The limitation of scale is especially relevant to science. Current statistical ML approaches require large amounts of data, which are often unavailable in science. This is especially the case in areas remote from the social and economic sciences, in theoretical areas or in areas with a strong descriptive component (e.g. astrophysics or genetics).

Annotation and labels

As another challenge associated with scale, many data sources must be annotated and labelled to be useful. This is difficult for several reasons. First, it takes time and resources to label large databases by hand. Second, variation in the data in some areas of science may not allow generalisations and translation across fields.

The wide range of sizes of stars in the galaxy, for example, requires a large dataset before analyses can yield results with statistical significance. The same is true in health care, where data set-ups may involve only healthy individuals (e.g. from fitness or wellness applications) or only unhealthy individuals (e.g. in a hospital) but rarely both. This makes translating findings from one to another more difficult. By contrast, applications of ML in industry usually work with much less variable data; think, for instance, of data coming from sensors on an assembly line.

Representation of data

For specialised scientific databases, one of the main challenges is how to capture the data using a symbolic representation that can also help with calculations. Much of the mathematics of ML is based on operations on data arrays such as matrices. Consequently, symbols such as words, images and sounds can be recorded as computable matrices or vectors that computers can manipulate.

Representing data with symbols matters because they have “meaning” for computers. Symbols can be manipulated and dealt with in predictable ways. For example, they can be used to calculate distances between data features to define a similarity metric. Likewise, symbols can help modify an image to produce a larger training set showing the ways in which an object in an image could be coloured under different lighting.

While the Internet has provided businesses with millions of pictures of everyday objects and faces, scientific data are much rarer. Take the challenge of protein folding. This requires data, of course, but also the use of models to process vast matrices of data that could represent, for example, distances between molecules.

It is highly inefficient to learn about molecule distance in proteins simply by learning basic patterns in data. This is because proteins are subject to external forces, such as the laws of thermodynamics, that can add “noise” and make patterns hard to find. What is required, in such cases, is symbolic representation and understanding of causation, which is a struggle for current approaches. This is where model-driven approaches could prove more useful and powerful, eventually replacing the work of scientists.

Model-driven methods can explain more observations with less training data just as human scientists do when deriving models from sparse data (Zenil et al., 2019). For instance, Newton and others derived the classical theory of gravitation from relatively few observations. ML approaches in scientific discovery have often combined statistical and symbolic approaches in a hybrid manner. The symbolic approach mostly comes from a human intervention, particularly to add symbols that represent generalisations and abstractions.

The need for a model-driven approach to AI in scientific discovery

The ability of human scientists to reason rationally, to do abstract modelling and to make logical inferences (deduction and abduction) are central to science. However, these abilities are handled poorly by the most popular approaches to AI (statistical ML and deep learning). Current AI involves mostly "black-box" techniques: methods that successfully accomplish a task but provide little to no insight or explanation of how they do so.

The black-box dilemma

Most neural network approaches have this black-box character. Their inner workings reveal a mass of data correlations but no evident relationship to any abstract or physical states of a dynamic system in the real world (such as a weather pattern, for example). The internal parameters of a neural network model do not directly correspond to any independent variables of the phenomenon they intend to model (e.g. the features that describe a cat). The underlying mathematical and computational representation of the neural network after training does not directly correspond to any physical state-to-state behaviour of the objects learnt. This is also why simulation-based approaches (Kulkarni et al., 2020; Piprek, 2021; Lavin et al., 2021) are garnering greater attention: they force AI to be model- and state-to-state-relevant.

For science, the black-box nature of current AI is a major challenge. Scientific textbooks and literature typically pertain to first principles (foundational knowledge, axioms, scientific laws, etc.) and step-by-step models executable by, for example, a computer or a mechanical process. These are the quintessence of scientific explanation (King et al., 2018).

Current approaches to AI in science rely on domain experts to generate understanding sometimes after statistical ML has helped shed light on the phenomena being investigated (Pinheiro, Santos and Ventura, 2021). For instance, after applying AI to the field of drug discovery by traditional means, domain experts may need to interpret the results to understand the mechanisms of a drug's effectiveness.

Science is more about cumulating knowledge and finding causal explanations than seeking to classify. Classification tasks, for example, are useful in a field such as industry (e.g. for movie or song recommendations, for example), where current AI algorithms excel. While classification is an important first step in science, detecting meaningful regularities or irregularities has been foreign to statistical ML approaches, including deep learning.

Bias

The problem of bias (in the everyday non-technical sense) also affects human science. Indeed, bias in AI is a legacy of human science because AI is traditionally trained on a set of examples labelled by humans. For example, in using ML to categorise different types of astronomical images, humans might need to feed the system with a series of images they have already categorised and labelled. This would allow the system to learn the differences between the images. However, those doing the labelling might have different levels of competence, make mistakes and so on. AI could be used to detect and to some extent redress such biases.

Classification

One of the most trivial types of weakness of ML revolves around classification. ML or deep learning can correctly classify a large set of images. However, after just a single pixel is changed, a large number of those images is classified both wrongly and, at the same time, with a high degree of confidence (Wang et al., 2021).

Among other uses, Generative Adversarial Network (GAN) (Cai et al., 2021) has been used to mitigate this classification problem. GANs apply a type of neural network to reduce dependence on statistical weaknesses in ML. For example, GANs can generate new examples of an image that could plausibly have been drawn from an original dataset (of pixels).

However, GANs are limited. Producing too many modified examples of an image for training data can indeed make the image-identification system work. However, this is only because the system has been fed with so many possible examples of relevant images in the training data. In other words, GANs can be used to generate new training data – new images in this case – but do not yield a model immune to the same problematic pixel-flipping effect.

GANs can help produce quite realistic but fake image data (i.e. with a similar pattern distribution of pixels). This can enlarge the training set to make the network better at classification. However, they lead to a combinatorial explosion of images producible by changes in all possible combinations of pixels (not only one, but also two and their combinations, and then three and so on).

Current statistical AI is different from human intelligence

Statistical ML operates differently from the human mind. GANs are not easily scalable because they follow a brute force approach to a problem that humans can address more effectively. Humans do not need to be fed with all sorts of fake images with insignificant small changes to minimise margins of error. First, humans would be unable to think of all possible images produced by flipping all possible combinations of pixels as a GAN may do. Second, humans appear to operate in a fundamentally different way. Humans build abstract models of the world, which 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 their driving tests or to witness millions of counter examples to know that hitting people on the road is a bad idea.

Think of a school bus. Humans know both its shape and its function. They are also more resilient than machines at identifying a school bus by its abstract properties. For example, they know it is a transportation system for children independent of its colour, shape or picture angle, the things that a statistical ML will focus on.

Arithmetic operations

Arithmetic operations provide another example of why statistical ML falls short of human reasoning. Learning to add two numbers does not work if the arithmetic operation and the concept of a number system is not “understood”. This is because one cannot feed a purely statistical network with enough examples of all possible sums between any two numbers.

It has recently been claimed that some systems, such as GPT-3, can synthesise the processes of addition, subtraction and multiplication by learning from examples (Brown, 2020). GPT-3 is a type of neural network that operates over vectors of words. It has been tested to see if learning from unstructured text could lead to some sort of deeper learning of basic arithmetic. Most tests were performed on only two- to three-digit numbers, with positive results.

However, it soon became apparent that GPT-3 was relying on previously seen examples of those exact operations as any young child would do. There was no deeper “understanding” or generalisation of arithmetic. Thus, neural networks, of which GANs are one example, are an important step forward for ML. However, they also illustrate the fundamental limitation and challenges of ML in modelling the world, learning and generalising as humans do. Systems based on GPT-3 such as ChatGPT, are giant lookup tables crawled from and combined with what has already been written on the Internet. They match inputs and outputs in the form of ever-growing vectors of words (sentences), Their often remarkable capabilities can give a false impression of intelligence.

Overfitting

GPT-3 was also trained on 175 billion parameters, which suggests possible overfitting (i.e. solving examples only because the AI had seen them all). This is different from how human intelligence works. There is no obvious reason why a natural language, unstructured model like GPT-3 should be good at a symbolic task such as learning arithmetic.

Symbolic systems

Statistical ML is limited in its capacity for symbolic systems. A convolutional network – a popular type of neural network used to classify images – comes up against two challenges. First, it must learn to recognise numbers (or perform any simple arithmetic problem) between any two numbers. Second, it must deal with a numerical positioning system such as the decimal.

Both challenges require training sets of large numbers of examples of, say, digits, numbers, or additions and subtractions. However, there is an infinite number of such examples (e.g. all possible arithmetical operations). No countable finite training set will ever cover this infinite universe of numbers and arithmetic operations.

No neural network can thus be trained over all possible arithmetic operations. Therefore, it cannot learn to add just from being shown a large set of examples of addition. It needs to be able to identify numbers and tell them apart from the relevant arithmetic symbols. In other words, there can be no successful attempt to train a neural network with a traditional statistical architecture to learn symbolic operations from numerical examples.

No matter how much data have been supplied, a neural network needs a “symbolic engine” – such as those in calculators that can deal with basic arithmetic. This illustrates the danger of the big data dogma – the belief that enlarging the training set will solve all learning challenges of a neural network.

For computer scientists, “loss function” refers to the distance between the prediction of an AI and the factual truth. Zenil, Kiani and Tegnér (2017) showed that the limitations of loss functions based on statistical measures, such as ones widely used in deep learning, can always be exploited. This results from the lack of “invariance results”.

Invariance here means that the representation of an object has no bearing on the ability of a system to recognise and identify it by its most salient properties. For example, in geometry, any object remains the same under certain “linear” transformations such as rotation, translation or reflection. Thus, an ML system should recognise objects regardless of how they are depicted. However, as noted earlier, statistical ML is highly sensitive to small changes, even in just a few pixels.

As one of its main achievements, ML has enlarged the set of transformations that it achieves when facing new data. For example, it recognises a dog as a dog even if all previously seen images of dogs have been of other breeds. However, that set of transformations remains small and rigid.

Neural networks have often been credited for some part of invariance. However, their invariance is often not robust or the result of features alien to the object of invariance. This is what GANs and changes in single pixels show. Changing a single pixel (often referred to as an “attack”, even if not deliberate) can undermine a network's ability to recognise objects. The reflection of light of any part of a school bus, for example, may make a neural network classify it as a firefighter truck.

Beyond data size

AI for science should go beyond a focus on the size of data (or big data). It should also devote more resources to developing the methodological framework most relevant to the AI needed for the specific domain in the process of scientific discovery (Zenil, 2017). Two examples are explored below.

Alphafold 2, the Google DeepMind approach to the scientific problem of protein folding, has greatly advanced prediction in the field. There is a debate, however, over how much AI is responsible for this accomplishment. Human designers, for example, decided how to represent the problem and the ML's major processing steps. Meanwhile, the domain-expert team deployed their knowledge of protein structure and statistical ML to accomplish the task.

In the domain of self-driving cars, companies compete over how many millions of miles their cars have driven unaided. However, the right measure should be how few miles they need to drive in order to infer and understand the basic rules of driving (e.g. not to hit a pedestrian).

This problem of attribution and misaligned metrics (with autonomous agency) is not exclusive to Alphafold 2 or self-driving cars. There is generally a close relationship between the choice of statistical model and ML engineers' pre-existing knowledge of the underlying structure of the data, and their own biases regarding how to deal with such data. Consequently, the contribution of domain-expert teams and statistical ML is generally intertwined.

Symbolic regression

Some model-driven approaches involve a technique known as "symbolic regression". Simply stated, this means the capability to manipulate symbols (unlike simple statistical regression). For instance, Udrescu and Tegmark (2020) use a library of equations so the AI system will find an equation that fits the observational data. While this approach has generated interesting results, it may suggest the underlying method is actually symbolic; in fact, it still has a strong classification component.

The most interesting system of this kind would be one where the library does not exist; with an existing library of equations, the problem becomes largely one of matching and classifying. Some research groups are trying to combine the worlds of statistical ML and symbolic regression. Statistical ML is best at representing and classifying data numerically, while symbolic computation excels at inference and rule-based reasoning.

Conclusion

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 and deep learning. This is a consequence of the dominant role of some academic actors and corporate AI research and development that are now almost one and the same.

A return to first principles is needed to work out how to generate some sort of equivalence to a mental model of the subjects that science is exploring. Indeed, a distinctive feature of human intelligence is the ability to take just a small fraction of a potentially infinite number of cases of something and comprehend that thing with a mental model. AI systems for science need similar abilities.

References

Brown, T.B. et al. (2020), "Language models are few-shot learners", *Advances in Neural Information Processing Systems*, Vol. 33/159, pp. 1877-1901, https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cai, Z. et al. (2021), "Generative adversarial networks: A survey toward private and secure applications", *ACM Computing Surveys (CSUR)*, Vol. 54/6, pp. 1-38, https://doi.org/10.1145/3459992.

King, R.D. et al. (2018), "Automating sciences: Philosophical and social dimensions", *IEEE Technology and Society Magazine*, Vol. 37/1, pp. 40-46, https://doi.org/10.1109/MTS.2018.2795097.

Kulkarni, S. et al. (2020), "Accelerating simulation-based inference with emerging AI hardware", in *2020 International Conference on Rebooting Computing (ICRC)*, pp. 126-132, https://doi.org/10.1109/ICRC2020.2020.00003.

Lavin, A. et al. (2021), "Simulation intelligence: Towards a new generation of scientific methods", *arXiv*, arXiv:2112.03235 [cs.AI], https://arxiv.org/abs/2112.03235.

Pinheiro, F., J. Santos and S. Ventura (2021), "Alphafold and the amyloid landscape", *Journal of Molecular Biology*, Vol. 433/20:167059, https://doi.org/10.1016/j.jmb.2021.167059.

Piprek, J. (2021), "Simulation-based machine learning for optoelectronic device design: Perspectives, problems, and prospects", *Optical and Quantum Electronics*, Vol. 53/4, pp. 1-9, https://doi.org/10.1007/s11082-021-02837-8.

Udrescu, S.M. and M. Tegmark (2020), "AI Feynman: A physics-inspired method for symbolic regression", *Science Advances*, Vol. 6/16, p. 4, https://doi.org/10.1126/sciadv.aay2631.

Wang, P. et al. (2021), "Detection mechanisms of one-pixel attack", *Wireless Communications and Mobile Computing*, Vol. 2021, https://doi.org/10.1155/2021/8891204.

Zenil, H. (2020), "A review of methods for estimating algorithmic complexity: Options, challenges, and new directions", *Entropy*, Vol. 22/6, p. 612, https://doi.org/10.3390/e22060612.

Zenil, H. (2017), "Algorithmic data analytics, small data matters and correlation versus causation", in *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data (Computability of the World? Philosophy and Science in the Age of Big Data)*, Ott, M., W. Pietsch and J. Wernecke (eds.), pp. 453-475, Springer Verlag.

Zenil, H., N.A. Kiani and J. Tegnér (2017), "Low-algorithmic-complexity entropy-deceiving graphs", *Physical Review E*, Vol. 96/1:012308, https://doi.org/10.1103/PhysRevE.96.012308.

Zenil, H. et al. (2019), "Causal deconvolution by algorithmic generative models", *Nature Machine Intelligence*, Vol. 1, pp. 58-66, http://dx.doi.org/10.1038/s42256-018-0005-0.

Machine reading: Successes, challenges and implications for science

J. Dunietz, AAAS Science and Technology Policy Fellow (STPF), United States

Introduction

As the rate of scientific publication has skyrocketed, many researchers have proposed taming the literature with artificial intelligence (AI). By harnessing the tools of natural language processing (NLP), researchers hope to automate some of the paper reading. This essay lays out a variety of reading comprehension behaviours, or "tasks", that NLP systems might perform on scientific literature. The essay places these tasks on a spectrum of sophistication based on models of human reading comprehension. It argues that today's 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 essay also discusses implications for where NLP tools can fit into researchers' workflows and offers several policy-relevant suggestions.

The core insight of this essay is that "reading" is not one monolithic capability. A shallow reader – whether human or automated – can do far less with a text than one who has combed through it and comprehended it deeply. Accordingly, the plausibility of proposals to have machines read papers depends on precisely what capabilities NLP is imagined to have. Without a well-calibrated notion of what NLP systems can do after having "read", the scientific community risks either missing opportunities for discovery or pinning its hopes on technology that does not yet exist.[1]

Human reading suggests a hierarchy of reading comprehension skills

Among the many theoretical models of human reading comprehension, the "Construction-Integration" (CI) model (Kintsch, 1988) is one of the most influential (McNamara and Magliano, 2009). It posits that concepts and propositions are first "activated" – i.e. made available for easy mental retrieval – then iteratively selected and merged into a globally coherent interpretation.

For this essay, the CI model is significant not for its proposed cognitive processing mechanisms, but rather for the form it assumes for the interpretation. The model asserts that a reader's mental representation includes three inter-constrained levels of information:

Surface structure

This is raw linguistic information, such as what words and phrases are present and what syntactic structures connect them.

Textbase

The textbase is the set of explicit propositions expressed by phrases and sentences. Given one or more passages, the textbase includes all elementary propositions a reader would take away. For a scientific paper, this might include assertions like "Sample A was kept at 20°C", "$MgCl_2$ represses expression of PmrA-activated genes" and "the response curve was modelled as a sigmoid."[2]

Situation model

A situation model[3] is an integrated representation that stitches together all asserted propositions and their relationships, as well as their relationships with unstated knowledge. This level might include relationships between assertions (e.g. that an organism retains a trait even though a gene has been knocked out); background information (e.g. that samples from 2018 can be assumed negative for COVID-19); agents' implicit goals (e.g. why experimenters wanted to ensure their samples were pure); and mentally simulated counterfactuals (e.g. what it would have meant had a solution turned a different colour). In many versions of the CI framework (e.g. Zwaan and Radvansky, 1998), the situation model focuses on spatial, temporal, causal and motivational relationships.

The CI model's taxonomy was intended to describe how human readers represent information. However, it can also be viewed as defining a spectrum of comprehension-dependent tasks. Some tasks, such as determining a paper's topic, can be performed using surface structure alone (e.g. by identifying keywords). Other tasks can be successfully performed only if the reader's internal representation includes at least a textbase-level understanding of the document(s). Finally, the most sophisticated tasksrequire a full situation model.

From this standpoint, the taxonomy applies just as well to NLP systems as to human readers (Sugawara et al., 2021). Accordingly, the next section considers what behaviours might be wanted from scientific NLP systems at each level of reading comprehension, and how current technology fares at each.

What NLP can contribute to science at each level of comprehension

Examples of NLP tasks at various levels of representation are shown in Figure 1. Each task is discussed in more detail below, including how it applies to scientific texts and where the state-of-the-art stands.

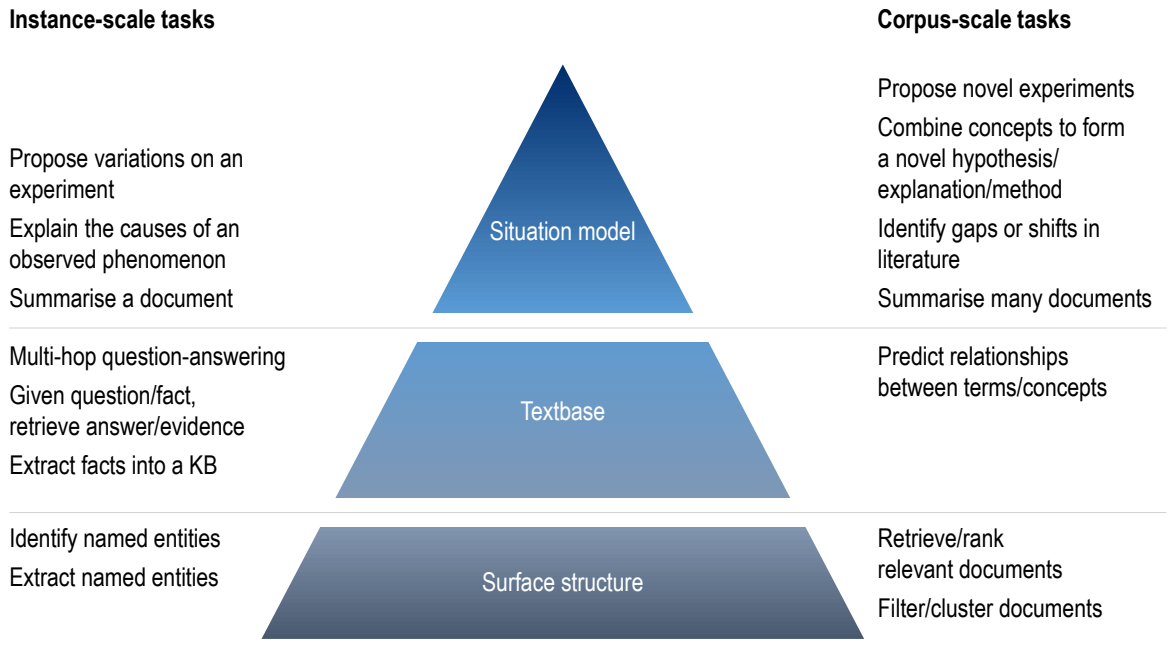
A few points should be kept in mind throughout:

- The levels of representation are best thought of not as three discrete levels but as a spectrum with three clusters. For instance, given the sentence "She didn't leave out a single one", it would be difficult to extract textbase-level propositions without background knowledge from the situation model about who "she" refers to, what she was doing and who or what she might have left out. Tasks with this property have been depicted at the textbase level but closer to the situation model boundary.[4]
- These categorisations should be seen only as rough intuitions, particularly since tasks that appear to hinge on higher-level representations may prove to be solvable using shallower techniques. If researchers were trying to answer the question, "What biomarkers indicate adenomas?", they might apply their situation model-level knowledge to home in on a paragraph about "biological markers" of "pituitary tumours". They would then extract the list of biomarkers from the textbase-level assertions in that paragraph. NLP tools, however, might succeed using surface structure alone – e.g. by looking for sentences with words that often co-occur with "biomarkers" and "adenomas".
- The CI model assumes the reader is consuming a single passage or document. In contrast, many NLP applications entail consuming an entire corpus of scientific literature. For instance, clustering

documents by topic only makes sense with multiple documents. For this essay, such tasks have been termed corpus-scale tasks. This term contrasts with instance-scale tasks that operate on single propositions, sentences or documents.[5]

- This is far from a comprehensive list of relevant NLP applications. Still, it should give the reader an intuition for what can be expected from any NLP tool.

Figure 1. Scientific NLP applications depend on a range of reading comprehension levels, from surface structure up to full situation models



Surface structure tasks

Surface-level NLP tasks mine associations between words, phrases and categories. Such tasks are far removed from anything normally considered "comprehension" (or perhaps even "reading"). They are most useful for helping a human reader locate and rapidly absorb information, particularly if researchers know exactly what terms or concepts to search for. The tools may also ease researchers' exploration and discovery processes.[6] Given that these systems leave most of the reading to humans, the occasional error matters little; the human can simply ignore irrelevant results or try a different search.

Instance-scale surface structure tasks

Two instance-scale tasks, named entity recognition and named entity identification, are described below.

Named entity recognition (NER)

This is a heavily studied NLP task, applicable to many domains, in which systems must flag "mentions" of predefined concept types. For example, a common version of the task is to scan passages for any phrase that refers to a person, a location or an organisation, and to classify each such phrase into one of these predefined categories. Intuitively, this task is about making associations between words or phrases and the categories.

Scientific text often requires specialised NER systems, both because the style differs from non-scientific text and because the categories are typically domain-specific. For instance, the CHEMDNER challenge (Krallinger et al., 2015) – the first community-wide effort to evaluate NLP methods for chemistry – included a task to automatically tag chemical names in scientific papers. Similarly, many biomedical NER systems look for phrases describing the population, intervention, comparison and outcome in papers about clinical trials (Kim et al., 2011).

With appropriate training data, scientific NER can perform quite well: recent systems achieve scores around 70-90%,[7] depending on the dataset and evaluation metric (Beltagy et al., 2019).

Named entity identification (NEI)

NER merely tags phrases such as "rabeprazole" with categories like "Intervention". Named entity identification (NEI), also sometimes called entity linking or named entity normalisation, goes a step further: it associates each tagged phrase with an entry in a structured knowledge base. For example, "rabeprazole" might be recognised as a reference to the "Rabeprazole.01" entry in a drug database.

Identifying ambiguously named entities can be easier with information from the textbase level. Still, like NER, NEI is largely about associating words and phrases with concepts – a surface structure task. Scores tend to be lower and far more varied than for NER, roughly 50-85% (Arighi et al., 2017).

NER and NEI are most often used in support of some downstream task, such as populating a knowledge base or allowing a searcher to filter papers to those that discuss a specified compound. NER and NEI can also be used to augment the reading experience, e.g. by colour-coding or hyperlinking gene names.

Corpus-scale surface structure tasks

At the corpus scale, core surface structure tasks include retrieving and ranking documents and clustering documents.

Retrieving and ranking documents

This classic task, often called information retrieval (IR), consists of returning documents that match a user query and ranking them by relevance. IR tools typically rely on some measure of alignment between the words or phrases in the query and those in a document. Examples include all competitors in the CORD-19 challenge (Roberts et al., 2021). In this challenge, systems received queries like "SARS-CoV-2 spike structure" and had to retrieve the most relevant research papers about COVID-19. Like more familiar search engines, scientific IR systems perform well, returning a satisfactory document in the top 5 results around 75-90% of the time.

Clustering documents

It can be helpful to automatically detect which documents are about similar topics, and perhaps even to organise topics into a hierarchy. A "topic" here is effectively a collection of closely related words or phrases. Recent systems that cluster scientific papers include the CORD-19 Topic Browser (MITRE, 2021) and COVID Explorer (Penn State Applied Research Laboratory, 2020). The clusters can be used either as a means of exploring the corpus or as an additional filter on search results. Though clustering is hard to evaluate objectively, it is a mature and well-studied task that generally produces respectable results.

Textbase tasks

A surprisingly large fraction of scientific NLP has focused only on the surface structure level. Still, there is plenty of research on extracting and manipulating papers' textbase-level propositions – tasks closer to

conventional "reading". In general, tools for these tasks are less reliable but still useful for investigating and generating hypotheses.

Instance-scale textbase tasks

Instance-scale textbase tasks (operating on individual phrases, papers or texts) include knowledge base construction, question answering, evidence retrieval and multi-hop question answering.

Knowledge base (KB) construction

Knowledge bases (e.g. of chemicals or genes) are widely used in science and beyond. To automatically populate a KB from one or more documents, an NLP system must turn natural language assertions into formal, machine-friendly propositions.[8] For instance, ChemDataExtractor (Swain and Cole, 2016) extracts many numerical properties of chemicals from the chemistry literature, generally with well over 95% accuracy. Kahun (2020) and COVID-KG (Wang et al., 2021) similarly extract relations such as "Condition-causes-symptom" and "Gene-chemical-interaction", albeit somewhat less reliably.

The resulting knowledge graphs can be used to look up characteristics of a gene, protein or chemical; to generate summary reports; to support question answering; or enable further machine learning on the structured relationships (e.g. predicting new drugs' possible side effects).

Question answering

There is a large NLP literature on question answering (QA) (Zhang et al., 2019; Zhu et al., 2021). The task is usually defined as responding to a user's question either with a yes/no answer or with a sentence or short phrase drawn from a body of text.[9] All of these variants have also been attempted in a scientific context (Nentidis et al., 2020). Recent examples include AWS CORD-19 Search (Bhatia et al., 2020) and covidAsk (Lee et al., 2020).

On many general-purpose QA benchmarks, NLP systems match or surpass humans. It is tempting to infer that these systems must comprehend at least some propositional content. However, systems' "comprehension" often proves brittle in the face of small changes to the question or passage (Jia and Liang, 2017) or shifts in the topics and content (Dunietz et al., 2020; Miller et al., 2020). The on-paper successes thus seem to stem largely from the benchmarks' artificial easiness: models are rewarded for exploiting ungeneralisable quirks of the data (Kaushik and Lipton, 2018).

It should not be surprising, then, that systems for scientific QA score only ~25-65% for retrieving relevant snippets and ~30-50% for retrieving relevant answer phrases (Nentidis et al., 2020).

Evidence retrieval

A similar task is to take a user-supplied assertion and look for snippets that support or refute it. This is often framed as fact-checking, particularly when systems are also asked to state whether the evidence supports or refutes the assertion. Recent systems score ~50-65% at extracting evidentiary sentences from abstracts (Wadden and Lo, 2021).

Multi-hop question answering

Many questions turn out to be answerable using little more than surface structure. In what is termed "multi-hop" QA, the questions are designed to rely on information from multiple pieces of text, theoretically requiring a higher level of comprehension and reasoning (Min et al., 2019). Several datasets have been constructed to test general-purpose multi-hop QA. Performance on these benchmarks is generally much lower than on regular QA. Among science-specific QA datasets, few (if any) target multi-hop reasoning, though at least a few of the datasets' questions likely require such reasoning.

A corpus-scale textbase task: Predicting relationships between concepts

Perhaps the most exciting textbase-level task[10] is one that only makes sense at corpus scale: predicting relationships between concepts based on large corpora. The study that launched this line of research identified materials likely to exhibit previously undiscovered thermoelectric properties (Tshitoyan et al., 2019). It used the materials science literature to train word vectors – representations that treat words as points (vectors) in a high-dimensional mathematical space of possible meanings. In such methods, word vectors are learnt from patterns of co-occurrence in the corpus. The distance[11] between two vectors corresponds to the similarity between the words' usage patterns, and hence presumably their meanings. The study authors identified chemical formulas whose vectors were close to that of "thermoelectric", thereby proposing several new possible thermoelectric materials. The work has begun to inspire similar efforts in other fields such as molecular biology (Škrlj et al., 2021).

Word vector analogies do over-generate proposed relationships; researchers must comb through to determine which hypotheses are worth testing. Still, the results can be valuable. The thermoelectrics' researchers demonstrated this by evaluating their approach retrospectively. They truncated the corpus to, say, 2009 to see what materials would have been proposed. This produced results far more likely than randomly chosen materials to have been studied later as thermoelectrics.

What emerges at the textbase level, then, is a useful but not entirely reliable suite of NLP tools. Given a well-defined question or hypothesis, these tools can help home in on relevant paper snippets. NLP tools can also help generate hypotheses, provided the researcher poses the right question (e.g. "What materials might have undiscovered thermoelectric properties?"). Systems can also extract structured data, but the resulting answers, fragments of evidence, KB entries or relationships must generally be treated with caution. If errors would be harmful, users will want to double-check the results.

Situation model tasks

Far less work has been done at the situation model level. At the instance scale – that of individual papers or passages – a NLP/AI system might help with tasks such as summarising a single document; explaining why an observation reported in a paper might have occurred; and proposing variations on an experiment described in a paper (in approximate order of increasing sophistication). Corpus-scale tasks might include summarising multiple documents (i.e. digesting an entire body of literature and summarising key takeaways); identifying gaps in the literature; combining concepts to propose a novel hypothesis, explanation or method; and proposing completely novel experiments to address knowledge gaps. To a greater or lesser degree, all of these tasks rely on manipulating a detailed, integrated representation of extensive information extracted or inferred from the text or corpus.

Of these tasks, only single- and multi-document summarisation have received significant attention. General-purpose summarisation has been widely studied in NLP (e.g. Hou et al., 2021). Approaches include a variety of extractive methods (stitching together fragments from the source text) and abstractive methods (generating an original summary). Similar techniques have been applied to scientific paper summarisation (e.g. Altmami and Menai, 2020), with some modifications for the peculiarities of scientific papers.

Results have been mixed: evaluation scores vary wildly, and it is not even clear which evaluation scores are meaningful (Kryściński et al., 2019). This is particularly true of abstractive methods, given that abstractive techniques, like many forms of natural language generation (Lin et al., 2021), struggle to ensure that output is factual (Maynez et al., 2020); they often fabricate information or misstate the facts.

Looking beyond summarisation, the level of comprehension and reasoning required for the other tasks above seems far out of reach. The fundamental problem is that current NLP techniques lack rich models of the world to which they can ground language (Bender and Koller, 2020; Bisk et al., 2020; Dunietz et al., 2020). 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.[12] A few intriguing methodological proposals have been sketched (e.g. Tamari et al., 2020), but the research trajectory towards situation models will be long indeed.[13]

Towards machines that comprehend: Possible research policy interventions

Despite the difficulties, research policies may be able to facilitate some progress towards machines that comprehend what they read – including scientific papers – at the situation-model level. Achieving that goal will likely require radical, interdisciplinary, blue-sky thinking. Yet 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 demand.

Policy makers could provide space and incentives for researchers to think more adventurously. To that end, three possible avenues are proposed below.

Reward methodological novelty

Research centres, funding streams and/or publication processes could be set up to reward methods that break with existing paradigms, even at the expense of publishing speed, performance metrics and immediate commercial applicability. Such programmes could even encourage ideas that remain half-baked or difficult to assess experimentally so long as they suggest novel, credible directions.

Seek inspiration from other disciplines

Policy makers could push NLP researchers to learn more from sociologists, philosophers and cognitive scientists, whose work on language likely holds untapped technical inspiration.[14]

Support under-studied research

Finally, policy makers can fund specific lines of under-studied research. This author is wary of some scholars' prescription of reviving symbolic methods; formal symbols tend to carve up the world too rigidly. However, perhaps NLP architectures could be explicitly designed to fluidly form, revise and apply composable concepts. In any case, funding for techniques may prove less pivotal than funding for tasks. Situation models seem likeliest to emerge from collaborative tasks where systems must communicate with humans to perform tasks in a real or simulated physical environment (e.g. Abramson et al., 2022).

Conclusion

Today's NLP provides many functionalities that can help scientists make good use of the literature. NLP can help winnow down the deluge of papers to ones relevant to a particular topic or question. It can also help researchers quickly find specific answers or pieces of evidence. It can even sometimes hypothesise previously undiscovered relationships, though humans must still do the hard work of asking the right questions and verifying systems' answers. Tools for these use cases will continue to improve, including on dimensions beyond accuracy (e.g. "few-shot training" may reduce the need for training data).

Where NLP falls short is on tasks that require deeper forms of comprehension. Distilling conclusions out of the literature remains the province of humans for the foreseeable future. This is even more true for generating creative insights about studies, though funding policies could begin to move the needle.

Of course, the history of AI contains many instances of apparently demanding tasks falling to surprisingly shallow techniques. Much NLP research amounts to finding ways of "hacking" tasks higher on the spectrum of sophistication using minimal information from the surface structure or textbase. It remains to be seen how much comprehension at the situation model level is not so difficult, after all.

References

Abramson, J. et al. (2022), "Creating multimodal interactive agents with imitation and self-supervised learning", *arXiv*, arXiv:2112.03763 [cs], http://arxiv.org/abs/2112.03763.

Aguera y Arcas, B. (2021), "Do large language models understand us?", 16 December, *Medium*, https://medium.com/@blaisea/do-large-language-models-understand-us-6f881d6d8e75.

Altmami, N.I. and M.E.B. Menai (2020), "Automatic summarization of scientific articles: A survey", *Journal of King Saud University – Computer and Information Sciences*, Vol. 34/4, pp. 1011-1028, https://doi.org/10.1016/J.JKSUCI.2020.04.020.

Arighi, C. et al. (2017), "Bio-ID track overview", in *Proceedings of BioCreative VI Workshop*, BioCreative, Bethesda, https://doi.org/10.15252/embj.201694885.

Beltagy, I. et al. (2019), "SciBERT: A pretrained language model for scientific text", in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, https://doi.org/10.18653/V1/D19-1371.

Bender, E.M. and A. Koller (2020), "Climbing towards NLU: On meaning, form, and understanding in the age of data", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, on line, https://doi.org/10.18653/V1/2020.ACL-MAIN.463.

Bhatia, P. et al. (2020), "AWS CORD-19 search: A neural search engine for COVID-19 literature", *arXiv*, arXiv:2007.09186, https://doi.org/10.48550/arXiv.2007.09186.

Bisk, Y. et al. (2020), "Experience grounds language", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistic, on line, https://doi.org/10.18653/V1/2020.EMNLP-MAIN.703.

Dunietz, J. et al. (2020), "To test machine comprehension, start by defining comprehension", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, on line, https://doi.org/10.18653/V1/2020.ACL-MAIN.701.

Gärdenfors, P. (2000), *Conceptual Spaces: The Geometry of Thought*, The MIT Press, A Bradford Book, Cambridge, MA.

Hou, S.-L. et al. (2021), "A survey of text summarization approaches based on deep learning", *Journal of Computer Science and Technology*, Vol. 36/3, pp. 633-663, https://doi.org/10.1007/s11390-020-0207-x.

Jia, R. and P. Liang (2017), "Adversarial examples for evaluating reading comprehension systems", in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Association for Computational Linguistics, Copenhagen, https://doi.org/10.18653/V1/D17-1215.

Johnson-Laird, P.N. (1980), "Mental models in cognitive science", *Cognitive Science*, Vol. 4/1, pp. 71-115, https://doi.org/10.1207/s15516709cog0401_4.

Kahneman, D. (2011), *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Kahun (2020), "Coronavirus Clinical Knowledge Search", webpage, https://coronavirus.kahun.com/ (accessed 28 October 2021).

Kaushik, D. and Z.C. Lipton (2018), "How much reading does reading comprehension require? A critical investigation of popular benchmarks", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Association for Computational Linguistics, Brussels, https://doi.org/10.18653/V1/D18-1546.

Kim, S.N. et al. (2011), "Automatic classification of sentences to support evidence based medicine", *BMC Bioinformatics,* Vol. 12/2, pp. 1-10, https://doi.org/10.1186/1471-2105-12-S2-S5.

Kintsch, W. (1988), "The role of knowledge in discourse comprehension: A construction-integration model", *Psychological Review*, Vol. 95/2, pp. 163-182, https://doi.org/10.1037/0033-295X.95.2.163.

Krallinger, M. et al. (2015), "CHEMDNER: The drugs and chemical names extraction challenge", *Journal of Cheminformatics*, Vol. 7/Suppl 1, p. S1, https://doi.org/10.1186/1758-2946-7-S1-S1.

Kryściński, W. et al. (2019), "Neural text summarization: A critical evaluation", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* Association for Computational Linguistics, Hong Kong, https://doi.org/10.18653/v1/D19-1051.

Langacker, R.W. (1987), *Foundations of Cognitive Grammar: Theoretical Prerequisites*, Stanford University Press, Stanford.

Lee, J. et al. (2020), "Answering questions on COVID-19 in real-time", in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Association for Computational Linguistics, on line, https://doi.org/10.18653/V1/2020.NLPCOVID19-2.1.

Lin, S. et al. (2021), "TruthfulQA: Measuring how models mimic human falsehoods*", arXiv*, arXiv:2109.07958 [cs], http://arxiv.org/abs/2109.07958.

Maynez, J. et al. (2020), "On faithfulness and factuality in abstractive summarization", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Association for Computational Linguistics, on line, https://doi.org/10.18653/v1/2020.acl-main.173.

McNamara, D.S. and J. Magliano (2009), "Toward a comprehensive model of comprehension", in *Psychology of Learning and Motivation – Advances in Research and Theory*, Academic Press, Cambridge, MA, https://doi.org/10.1016/S0079-7421(09)51009-2.

Michael, J. (23 July 2020), "To dissect an octopus: Making sense of the form/meaning debate", Julian Michael's blog, https://julianmichael.org/blog/2020/07/23/to-dissect-an-octopus.html#antithesis-an-ai-perspective.

Miller, J. et al. (2020), "The effect of natural distribution shift on question answering models", in *Proceedings of the 37th International Conference on Machine Learning (PMLR)*, Association for Computational Linguistics, on line, https://proceedings.mlr.press/v119/miller20a.html.

Miller, T. (2021), "Contrastive explanation: A structural-model approach", *The Knowledge Engineering Review*, Vol. 36, https://doi.org/10.1017/S0269888921000102.

Min, S. et al. (2019), "Compositional questions do not necessitate multi-hop reasoning", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, https://doi.org/10.18653/V1/P19-1416.

MITRE (2021), "MITRE CORD-19 Topic Browser", webpage, http://kde.mitre.org (accessed 28 October 2021).

Nentidis, A. et al. (2020), "Overview of BioASQ 2020: The Eighth BioASQ challenge on large-scale biomedical semantic indexing and question answering", in Arampatzis, A. et al. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer International Publishing, https://link.springer.com/chapter/10.1007/978-3-030-58219-7_16.

Nye, M. et al. (2021), "Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning", a*rXiv*, abs/2107.02794, https://openreview.net/pdf?id=P7GUAXxS3ym.

Penn State Applied Research Laboratory (2020), *COVID Explorer* (database), https://coronavirus-ai.psu.edu/database (accessed 28 October 2021).

Roberts, K. et al. (2021), "Searching for scientific evidence in a pandemic: An overview of TREC-COVID", *Journal of Biomedical Informatics*, Vol. 121, p. 103865, https://doi.org/10.1016/J.JBI.2021.103865.

Schaffer, J. (2005), "Contrastive causation", *The Philosophical Review*, Vol. 114/3, pp. 327-358, https://doi.org/10.1215/00318108-114-3-327.

Škrlj, B. et al. (2021), "PubMed-scale chemical concept embeddings reconstruct physical protein interaction networks", *Frontiers in Research Metrics and Analytics*, Vol. 6, 13 April, https://doi.org/10.3389/FRMA.2021.644614.

Sugawara, S. et al. (2021), "Benchmarking machine reading comprehension: A psychological perspective", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, on line, https://doi.org/10.18653/v1/2021.eacl-main.137.

Swain, M.C. and J.M. Cole (2016), "ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature", *Journal of Chemical Information and Modeling*, Vol. 56/10, pp. 1894-1904, https://doi.org/10.1021/ACS.JCIM.6B00207.

Tamari, R. et al. (2020), "Language (Re)modelling: Towards embodied language understanding", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Association for Computational Linguistics, on line, https://doi.org/10.18653/v1/2020.acl-main.559.

Tshitoyan, V. et al. (2019), "Unsupervised word embeddings capture latent knowledge from materials science literature", *Nature*, Vol. 571/7763, pp. 95-98, https://doi.org/10.1038/s41586-019-1335-8.

Wadden, D. and K. Lo (2021), "Overview and Insights from the SCIVER shared task on scientific claim verification", in *Proceedings of the Second Workshop on Scholarly Document Processing*, Association for Computational Linguistics, on line, https://aclanthology.org/2021.sdp-1.16.

Wang, Q. et al. (2021), "COVID-19 literature knowledge graph construction and drug repurposing report generation", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, Association for Computational Linguistics, on line, https://doi.org/10.18653/V1/2021.NAACL-DEMOS.8.

Zhang, X. et al. (2019), "Machine reading comprehension: A literature review", *arXiv*, arXiv:1907.01686 [cs.CL], https://arxiv.org/abs/1907.01686v1.

Zhu, F. et al. (2021), "Retrieving and reading: A comprehensive survey on open-domain question answering", *arXiv*, arXiv:2101.00774 [cs.AI], https://arxiv.org/abs/2101.00774v3.

Zwaan, R.A. and G.A. Radvansky (1998), "Situation models in language comprehension and memory", *Psychological Bulletin*, Vol. 123/2, pp. 162-185, https://doi.org/10.1037/0033-2909.123.2.162.

Notes

[1] This article represents the views of the author, and was written before his fellowship with AAAS began. It does not necessarily represent the views of AAAS or the US government.

[2] Of course, the words expressing each of these assertions are included in the surface structure representation. What distinguishes the textbase representation is that these assertions have been abstracted into a more conceptual form in the reader's mind, e.g. modeled-as (response-curve-1, Sigmoid).

The textbase is usually assumed to consist of formal logical propositions, although other processing-friendly abstractions may sometimes be more suitable.

[3] Situation models have also sometimes been referred to as "mental models" (Johnson-Laird, 1980).

[4] The situation-model level is especially spectrum-like: even for a human, a situation model may be more or less complete depending on how many inferences and pieces of background knowledge the reader manages to incorporate.

[5] Clearly, one could repeat an instance-scale application for every instance in a larger corpus. For example, rather than extracting a few facts from a document into a knowledge base, one could attempt to extract as many facts as possible from an entire corpus of documents. This corpus-scale version of the task might even be approached differently from its instance-scale cousin. For example, computational constraints might rule out processing each instance separately, or perhaps the resultant knowledge base could be improved by considering mutually contradictory or reinforcing pieces of evidence across the corpus.

Nonetheless, these applications were listed as instance-scale because one *could* perform them instance by instance, at least in principle. The "corpus-scale" designation has been reserved for tasks where the task does not even make sense without a larger corpus.

[6] Literature-based discovery (LBD) tasks, though omitted here for space, would generally fall into the surface structure category, as well: LBD typically tries to connect two sets of terms, A and C, by finding some set B of words or phrases that are connected to both A and C.

[7] The metrics reported in this essay vary by task but typically balance precision (what fraction of instances output by the system are correct) against recall (what fraction of correct instances are output by the system). Such a metric, known as an "F1 score" or "F-measure", ensures that to score highly, systems must simultaneously notice relevant phrases or answers and ignore irrelevant ones.

[8] Strictly speaking, some knowledge base construction can be done purely at the surface level: a system needs nothing more than entity recognition and/or identification to extract co-occurrence relationships as an indication of unspecified "relatedness".

[9] There are many additional variants on question answering tasks, including answering multiple-choice questions about a passage; declining to answer when the answer is not present; generating full-sentence answers (as opposed to retrieving decontextualized phrases); and answering multiple questions in context as part of a multi-turn dialogue. In a scientific context, these more exotic variations seem unlikely to be substantially more useful than vanilla question answering. They have therefore been elided in the main text. With the exception of multiple-choice questions, which are somewhat artificially easy (Dunietz et al., 2020), these variants generally see worse system performance than more conventional QA tasks.

[10] It is debatable whether word vectors are really operating at the textbase level, given that they are trained only on associations between words and sequences thereof. However, they do seem to capture information that are conventionally thought of as propositional – e.g. relationships between concepts – albeit only at the corpus scale, not at the scale of individual training sentences.

[11] More precisely, the cosine distance, related to the angle between the vectors. Directions in the vector space are generally taken to correspond to concepts or elements of meaning (e.g. gender). The vectors for two words such as "large" and "enormous" might have identical orientations but different magnitudes.

[12] Even researchers who argue that modern large language models do "understand us" (e.g. Aguera y Arcas, 2021) typically acknowledge that these models confabulate at best and give "off-target, nonsensical or nonsequitur [responses]" at worst.

The shortfalls are most obvious when the task involves generating text. However, even non-generative tasks – e.g. multiple-choice question answering – are typically approached using the same underlying language models. Even without the opportunity to fabricate, the lack of deep comprehension becomes painfully clear with sufficiently rigorous evaluation procedures (Dunietz et al., 2020). For any given reading comprehension system, most researchers with experience in NLP would have little trouble finding questions that trip up the system even though the answers are obvious to humans.

See also Bender and Koller (2020), Bisk et al. (2020) and Michael (23 July 2020) for discussions of whether reading comprehension systems trained purely on text could learn to construct and manipulate situation models even in principle.

[13] Reading at this level might even be "AGI-complete". In other words, achieving it might be tantamount to solving every problem in artificial general intelligence ("strong" or fully human-like AI), from planning to commonsense reasoning, social interaction and perhaps even perception and object manipulation.

[14] Examples of less technical work that could suggest NLP and AI approaches include prototypes and radial categories from cognitive linguistics (e.g. Langacker, 1987); contrastive accounts of causal language from philosophy (e.g. Schaffer, 2005; Miller, 2021); conceptual spaces from cognitive science (Gärdenfors, 2000); and the System 1/System 2 distinction from psychology (Kahneman, 2011; e.g. Nye et al., 2021).

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

H.M. Cartwright, Oxford University, United Kingdom

Introduction

Few artificial intelligence (AI) applications can do more than explain to a non-expert what they have learnt or the reasoning behind their decisions. Explanations are fundamental to understanding, but not every explanation persuades. If a doctor describes a skin lesion as possibly cancerous, her patient is likely to accept the diagnosis without asking for the doctor's medical certificates. The same patient, though, might view with suspicion a mechanic's estimate of several thousand US dollars for simple car repairs. If the "explainer" is non-human, an acceptable explanation may be particularly hard to obtain. This essay touches upon some challenges in the development of "explainable AI", focusing on applications in science and medicine.

Why do we need explanation?

For some types of problems, AIs already surpass human abilities. In these cases, it is tempting to think elaboration is unnecessary. However, AIs can also make perplexing or faulty decisions.

Retrosynthesis is the process of computationally deconstructing a target molecule of interest, such as a drug or a catalyst, into a number of simpler molecules. Each of these molecules is readily available, or can be synthesised, from still more simple chemicals. In this way, a viable route for the manufacture of the target can be found. This is a critical task in the development of commercially valuable materials.

Synthesis planning involves a combinatorial explosion of paths to examine. Meanwhile, the identification of suitable synthetic routes still relies largely on human experience, intuition and guesswork. The value of a proposed route depends on a wide variety of factors: availability of suitable reagents and solvents; stability of reactants and intermediates in storage; availability and cost of suitable synthetic equipment; the ability to suppress unwanted competing reactions; toxicity of reagents and intermediates; the necessity to limit power consumption during synthesis; and many more.

If an AI makes a surprising choice of synthetic route, further investigation is warranted. For example, it might choose a path in which each low-temperature reaction is followed by a reaction at high temperature, thus increasing energy use. In this case, an interrogation of the AI to understand its reasoning would be helpful. Such an interrogation is difficult and its value may be hard to judge if the proposed route includes unusual synthetic steps, which, because they are rarely studied, provide less reliable data.

Even when an AI's deductions are correct, more information can be valuable. On average, for example, children who spend the most time at school have poorer eyesight than their less industrious colleagues. Does prolonged studying cause eye damage? Or do myopic children spend more time in school? An AI might link school attendance to myopia, but correlation is neither explanation nor causation. Uncovering and explaining why something happens is more difficult than recognising that it does happen, especially for AIs.

This link between cause and effect is fundamental to understanding. However, what if that link is so complex that scientists find it impossible to understand?

Science is becoming more difficult. Most relatively straightforward scientific areas are heavily studied; what remains are more challenging topics. The equation in Figure 1 – from string theory – illustrates how complex these (often theoretical) areas can appear to be. Parts of mathematics, physics and quantum mechanics are accessible to only a small number of practitioners. As science continues to evolve, some topics may become so intellectually demanding that no one can understand them.

Figure 1. Science becoming more difficult: The mathematics of a part of string theory

$$E^{(0)}(\eta) = \frac{\hbar c}{2\pi^2 a^4(\eta)} \int_{mca(\eta)/\hbar}^{\infty} \frac{\lambda^2 d\lambda}{e^{2\pi\lambda} - 1} \left[\lambda^2 - \frac{m^2 c^2 a^2(\eta)}{\hbar^2} \right]^{1/2} \approx \frac{(mca(\eta)/\hbar)^{5/2} \hbar c}{8\pi^3 a^4(\eta)} e^{-2\pi x ca(\eta)/n} \Rightarrow$$

$$\Rightarrow - \int d^{26}x \sqrt{g} \left[-\frac{R}{16\pi G} - \frac{1}{8} g^{\mu\nu} g^{\nu\sigma} Tr(G_{\mu\nu} G_{\rho\sigma}) f(\phi) - \frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi \right] =$$

$$= \int_0^{\infty} \frac{1}{2\kappa_{10}^2} \int d^{10}x (-G)^{1/2} e^{-2\Phi} \left[R + 4 \partial_\mu \Phi \partial^\mu \Phi - \frac{1}{2} \left| \tilde{H}_3 \right|^2 - \frac{\kappa_{10}^2}{g_{10}^2} Tr_v(|F_2|^2) \right] \Rightarrow$$

$$\Rightarrow \frac{1}{3} \frac{4 \left[antilog \frac{\int_0^{\infty} \frac{\cos \pi t x w'}{\cosh \pi x} e^{-\pi x^2 w'} dx}{e^{-\frac{\pi^2}{4} w'} \phi_{w'}(itw')} \right] \cdot \frac{\sqrt{142}}{t^2 w'}}{\log \left[\sqrt{\left(\frac{10 + 11\sqrt{2}}{4} \right)} + \sqrt{\left(\frac{10 + 7\sqrt{2}}{4} \right)} \right]}$$

Source: Nardelli and Di Noto (2020).

Once such an extreme challenge to comprehension is reached, AIs could help science to progress. To that end, it could scan scientific databases, looking for previously undiscovered relations that could be cast as new laws. Such laws might be of considerable value in science. However, what would happen if the AIs could not explain them, or even provide a mathematical representation of them? It would become impossible for scientists to independently verify conclusions of the AI. Even more seriously, the human development of science would be inhibited as the field began to fill with laws that no one could understand.

Explanations are therefore essential. Tools, such as decision trees or reverse engineering, offer some insight into AI logic. However, most scale poorly with software complexity and are of value only to experts. This essay focuses on the needs of the non-expert, for whom explanations should be appropriate in extent and complexity to avoid the need for any further detail. This requirement may be demanding. The power of AI derives from its ability to work in high-dimensionality space. Translating into human-digestible form

what has been learnt in such a space may yield dense lines of reasoning, even if individual parts of the argument are clear.

Explanations must also be accurate. Though an obvious requirement, the implications of this are far-reaching. It is not sufficient that an AI generate reliable decisions; its explanatory model to provide a commentary must be equally effective, as the retrosynthesis example above suggests.

Challenges for explainable AI

Interpretability is the capacity of a black-box predictor to provide lucid explanations. Before considering some challenges, there is a fundamental question: is comprehensive explanation even possible?

The power and complexity of an AI are closely linked. Yet, even if the working of the underlying code is open to inspection, AI reasoning may be opaque. Dyson (2019) has argued that "… [a]ny system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand."

A pessimistic view perhaps, but interpreting AI logic is unlikely to be straightforward. In both the human brain and artificial neural networks, the representation of knowledge is distributed. However, their different structures make any direct translation between them impossible (at least currently). Humans communicate using symbolic language, and so require an AI to provide information in a compatible format. Even if an AI could do this, that ability on its own might be insufficient. The AI needs to provide meaningful explanations but also have the ability to engage in logical argument; this is a lot to ask.

If, in principle, the reasoning of an AI can be described, what makes extracting explanations from it tricky? Could it not just be sliced open for some rules and deductions? Unfortunately, matters are not that simple. At the heart of an AI lie not rules, but tens or hundreds of thousands of numbers, whose interpretation is difficult even for programmers.

These numbers encode the knowledge accumulated during training. An initial challenge is to ensure these training data are of suitable quality and that their context is sound. Observable (directly measurable) training data may be biased or influenced by factors of which humans are unaware. Therefore, access must be available to appropriate, high-fidelity training data.

Raw tabular data are readily entered into AIs, but metadata – background information that can place the raw data in context – may not always be available. Humans may know that relations exist between parts of the data; for instance, that average waist size can be related to country of residence. However, the AI may not have the benefit of this prior knowledge. It must discover relationships for itself, implying that access will be required to larger datasets.

AIs may be made more robust by combining deep neural networks with rule-based methods. Such a combination may be valuable in safety-critical applications (e.g. when people work alongside AI-controlled robots on a car production line). However, the inclusion of rule-based components in an AI system does not of itself lead to an improved ability to explain.

Effective AI systems typically comprise tens of thousands of lines of code, constructed by teams of programmers, none of whom knows how the entire system works. It has been reported, perhaps apocryphally, that older versions of the Windows operating system contained large quantities of apparently redundant code that software engineers dared not remove because its purpose was unknown. AI software, though more compact than a complete operating system, may be developed over several years. During that time, the behaviour of the software and any embedded explanation systems may drift from the original specifications, possibly without the knowledge of programmers.

Even when development of the software itself is complete, AIs may continue to learn in some fields while they work. Learning about causality, for example, requires agents that can interact with their environment

and, in doing so, create their own data. In chemistry, AI-assisted robots can be used to assess and optimise new synthetic routes, generating a pool of knowledge about practicality, safety and yields that grows as the system runs new chemical reactions. As knowledge accumulates, the AI's understanding may grow, but its ability to explain this progressively more detailed world may diminish.

One might expect that a typical question posed to an AI would be "Why did you conclude this?" By contrast, an "exception analysis" wants to understand why mistakes occur. The question then becomes: "Why did you get this wrong?"

An understanding of why a decision is faulty can uncover limitations of a trained AI model. However, exception analyses are rare in science for two reasons. First, a human must recognise the AI has taken a wrong turn. Second, a sufficiently competent explanatory system is available to uncover the origin of the failure.

Web-based image recognition applications have impressive capabilities. However, it is challenging to construct an image that an AI can use to illustrate an explanation. It often yields hybrid images in which multiple portions of the training data seem blended into a single confused new image.

Preparing reliable synthetic images as part of an explanation is an ongoing challenge. Temporal (time-varying) image streams, such as electroencephalography (EEG) scans, present further difficulties. Beyond providing a textual explanation, an AI attempting to interpret and explain a series of EEG images may need to illustrate the text with unambiguous images constructed from visual data that vary with time, patient and measurement conditions.

AIs operating in different areas may use software based on similar algorithms, provided that issues such as protection of intellectual property do not create barriers. With suitable retraining (usually from scratch), an AI may be repurposed into a substantially different field. However, porting an explanation system may be problematic. An explanation mechanism designed for a board game may be poorly suited to explain how a protein folds, no matter how much it is modified. Porting AIs between applications may also generate reliability issues, with potential consequences for features such as the ability to predict software failure modes.

Ethical considerations

The ethics of AI decision making (not just the ethics of AI use) is an area of increasing interest. This interest is concentrated in areas in which AI might be used to make decisions that affect people directly, particularly in medicine.

Triage is the process of assessing patients who enter the emergency ward of a hospital to determine what treatment they require. Medical staff must make crucial decisions, up to and including "Can the life of this patient be saved?" The number of patients moving through a large emergency unit is substantial. Over time, a significant database of previous triage decisions builds up, providing a resource that could be used to train an AI to make its own decisions. Such a database will include some data that relate to the patients' medical condition, and other data that are ethical in nature. If a large group of seriously ill patients arrives together, it may overwhelm the capacity of the unit. This could lead to delayed treatment for some patients; staff must then make judgements that contain ethical elements: "Should this patient be saved?"

An AI trained on triage data might well develop the ability to make both medical decisions about the treatment of incoming patients and ethical decisions if necessary. The defining line between the two may be blurred. Therefore, unambiguous and sympathetic justification of AI decisions would be crucial to convince medical staff, relatives and patients that decisions were appropriate.

An AI making such ethical decisions has been trained on data that include numerous examples drawn from the operation of a real emergency department. Consequently, one might expect these decisions would

simply replicate those of a human in comparable circumstances. However, AI tools may find unexpected solutions that humans may not have spotted. Humans might draw a clear line between medical and ethical decisions, for example. For an AI, information about the patient's symptoms, treatment, prognosis and eventual outcome are all just data points. The AI's assessment, based on the entirety of what it has learnt need not be in accord with human ethics.

Transparency is particularly important for ethical decisions. Privacy issues may cloud seemingly straightforward decisions. For example, should confidential data from a drug trial be released to the public before the trial is complete? If a drug trial has a small number of patients, should those patients receiving the placebo continue taking it to give the trial more statistical strength, even when data show the drug taken by other participants is effective?

Limited transparency may facilitate deception. A fair AI can create a decision tree employing a set of "if-then" rules to support reasoning that suggests a selection process is gender-neutral. Yet fair algorithms can be applied discriminately, a technique known as "fairwashing".

In fairwashing, an unfair procedure (in the selection of candidates for promotion, perhaps) is presented in a manner that suggests it is fair. The AI is used knowingly to hide evidence that critical factors, such as gender and ethnicity, influenced the decision. An opportunity to interrogate the AI would provide some degree of protection against the practice of fairwashing.

Use of AI by non-experts may also give rise to undesirable side effects. Millions of people self-diagnose using online medical chatbots, some of which employ AI. Users may feel as if, or even believe, they are chatting with another human. If the advice is faulty or misunderstood, where does the blame lie?

Without any way to assess the expertise of the software, AI medical recommendations must be taken on trust. Yet the AI may be provided by a commercial concern, which influences its recommendations (e.g. which drugs to take). The ready availability of impersonal diagnosis over the Internet can reduce visits to human doctors. This could lead to poorer outcomes since interactions with a computer may be less illuminating than those with a human doctor.

Discussion

AI is at its best when users have confidence that its decisions are reasonable and justifiable. Incomplete explanations may create suspicion that its operation is being deliberately or unintentionally obfuscated. They might also suggest the benefits of using AI accrue to the commercial or government entity rather than the user.

In science, incomplete explanations may be unavoidable when release of data is restricted by commercial interests, such as a patent application. However, science flourishes through the rapid and comprehensive dissemination of information. Therefore, deliberate or inadvertent limitation of explanations from AI-based systems can inhibit scientific progress.

The General Data Protection Regulation became law in the European Union in May 2018. It requires information about the "logic involved" in an AI to guard against discriminatory or unfair practices. While satisfying the "right to an explanation" is admirable, that aim to date is not fully achieved. Indeed, the AI's logic could be interpreted as a reference only to its algorithmic processes. Explanations of how the software operates might interest a programmer. However, they would be of little value to most people who are looking to understand how the AI reached its decision.

Explanatory systems must be developed and enhanced to match the power of the systems within which they are contained. However, even as they are, the decision-making side of AI will keep evolving. As a report from the UK House of Lords (2018) puts it: "... if we could only make use of those mechanisms that we understand, we would reduce the benefits of artificial intelligence enormously."

The implication is clear: the government will not pause development of AI decision making to allow explanatory systems to catch up. However, developers should not argue that what the AI system can do is so much more important than how it does it. They should not put work on explanatory systems to one side.

Finally, as AI becomes more powerful, the ability to predict future ramifications of its decisions will become more important. However, the use of AI in one field may affect activities in another apparently unrelated field. To anticipate such effects, the limitations and capabilities of AI operating systems must be open to interrogation.

Conclusion

As AI applications become more powerful and widespread, the demand for explanation will grow. It might seem that AI is not so different from other methods of data analysis, just more efficient. However, this underestimates both the power of the methods and the degree to which the reasoning of an AI is hidden compared to conventional data analysis tools. Halting software development until exhaustive explanations are routinely available would be disruptive, and probably impractical. However, AI users must not be encouraged to believe that providing comprehensive explanations is so difficult that its decisions should be accepted without question.

"Useful AI" (i.e. commercially valuable) risks developing at a far greater rate than that of "user-friendly AI" (i.e. that can explain itself). Moreover, if explanations of any sort are not expected, software companies may think that development of explanatory systems can be quietly put to one side. If this were to happen, the opportunity to develop them will be lost and powerful – but opaque – AIs could become the norm. Furthermore, current software largely sidesteps ethical and practical challenges in the provision of explanations. Users are often unaware they are interacting with an AI, or that it is processing their data. With no knowledge of AI involvement, the user will not be expecting an explanation.

Governments can undoubtedly play a role in helping to foster research on the explanation problem, but how best to have significant impact is unclear. Any government funding would be dwarfed by the huge amounts of money already invested by the biggest commercial players (Google, Amazon, et al.) on AI in general. Governments might pour money into national agencies like the US Defense Advanced Research Projects or other public organisations. However, it might be hard to bring together a sufficiently large group of talented people to make real progress in such a demanding area. These issues require further explanation.

References

Dyson, G. (2019), "The third law", in *Possible Minds: 25 Ways of Looking at AI,* Brockman J., (ed.), Penguin, New York.

House of Lords (2018), "AI in the UK: Ready, willing and able?", Select Committee on Artificial Intelligence, Report of Session 2017-19, https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Nardelli, M. and F. Di Noto (2020), "On some equations concerning the Casimir effect between World-Branes in Heterotic M-Theory and the Casimir effect in spaces with nontrivial topology. Mathematical connections with some sectors of Number Theory", https://vixra.org/pdf/2005.0121v1.pdf.

Combining collective and machine intelligence at the knowledge frontier

E. Malliaraki, Nesta, United Kingdom

A. Berditchevskaia, Nesta, United Kingdom

Introduction

In the past decade, machine learning and deep learning have advanced significantly. They can now assist in the process of discovery, for instance, by analysing large volumes of data. On the other hand, humans have unique abilities such as creativity, intuition, contextualisation and abstraction. Moving forward, the best of both worlds must be combined. Instead of using only artificial intelligence (AI) to navigate scientific knowledge, novel AI and human collaborations could explore complexity and advance the frontiers of scientific understanding in new ways. This essay describes emerging tools and initiatives for discovering, encoding and synthesising knowledge that could help guide the way. The recommendations outline a pathway for changing scientific infrastructures, incentives and institutions to help hybrid human-AI science to flourish.

Each day, researchers publish more than 4 000 scientific papers in the field of biomedicine alone. An estimated 200 000 articles have been written to date about the COVID-19 pandemic. While the number of publications grows exponentially (Bornmann and Mutz, 2015), the number of novel scientific ideas expands only linearly (see Staša Milojević's essay in this volume). Across fields, from medicine to agriculture to computers, the effort and money required to innovate are growing.

Traditional mechanisms for getting up to speed with, and navigating, the knowledge frontier have limitations. Textbooks are updated only slowly. Meanwhile, literature reviews are often ad hoc and may overlook relevant work from other disciplines or discount disruptive ideas in favour of canonical works (Chu et al., 2021).

As fields grow, they often divide into subspecialties, each with its own literature. These cluster in discrete areas of knowledge (Foster et al., 2015). Fragmentation leads to a combinatorial explosion of "undiscovered public knowledge" (i.e. knowledge existing in the unrevealed connections between existing bodies of publicly available knowledge) (Swanson, 1989). In addition, while older or neglected findings and hypotheses from distant disciplines could become sources of new knowledge, the means are unavailable to systematically resurrect them (Swanson, 2011). This results in undiscovered correlations, undrawn conclusions and a failure to capitalise on insights from analogous problems (see the essay by Smallheiser et al. in this volume).

At the same time, science is carried out by ever-larger teams and international consortia.[1] More than ever before, science has become a collective endeavour. This is witnessed in successes ranging from mapping the human genome to pushing the frontier of neuroscience research with the Human Brain Project and proving Einstein's theory of gravitational waves. From optimising the size (Wu et al., 2019) and diversity of

teams (Nielsen et al., 2018) to leveraging emerging methods like crowdsourcing and crowd forecasting (Sell et al., 2021), an understanding of how to make the most of collective intelligence for scientific discovery is just beginning to emerge.

Reimagining current methods and tools to better harness collective and machine intelligence will help scientists better assess the state of scientific knowledge and prioritise research at the knowledge frontier (Berditchevskaia and Baeck, 2020).

Mapping the knowledge frontier

Encoding and discovering knowledge

Scientific knowledge consists of concepts and relationships represented in research papers, patents, software and other academic artefacts. However, the existing science communication infrastructure does not help researchers make the best use of the predominantly document-centric scholarly outputs. For example, the extracted text, graphics, bibliography and metadata from PDF files often require extensive cleaning before use. While words and sentences may be searched for, images, references, symbols and other semantics are currently mostly inaccessible to machines.

In recent years, it has become increasingly possible to represent scholarly knowledge in machine-actionable form**.** So far, the emphasis has been on representing, maintaining, and linking data and metadata about articles, people and other relevant entities. However, in metadata discovery, a combination of humans and machines allows for better detection of scientific data artefacts. For example, project RePEc uses AI to infer which datasets have been used in a publication and then requests authors to validate or reject these dataset annotations (Nathan, 2019).

Once encoded, these pieces of public knowledge must be searchable and discoverable at the right level of representation. Recent advances in natural language processing (NLP), text mining and information retrieval can now help researchers find and understand scientific information. For example, citation analyses using NLP help researchers discover intersections and emerging trends across scholarly databases, as well as lines of research that are either new or not mainstream. Moreover, improved representation learning and extreme summarisation in scholarly documents can alleviate information overload. Open Knowledge Maps is a notable example of a human-AI collaboration that clusters papers in similar subfields based on keywords, datasets and research software. It then allows users to create, edit and update their own knowledge maps (Matthews, 2021).

Several initiatives have focused on identifying and extracting more granular units of scientific information from papers. These units can be problems (Lahav et al., 2021), hypotheses (Spangler et al., 2014), methods (Fathalla et al., 2017), findings (Sebastian et al., 2017), causal relations and even automated suggestions of new hypotheses (Liekens et al., 2011). Specifically, recent advances in language models (i.e. sciBERT, bioBERT)[2] create much more accurate semantic representations of scientific concepts and allow for a more contextual search of scientific documents. However, these models are typically built on pretrained language representations by domain experts. They have limited generalisability outside the domains where they are developed.

This problem of limited generalisability can be addressed by harnessing the complementary pools of expertise from scientists and policy makers. A notable example is the TREC-COVID project, which extracts useful information from publications, Twitter conversations and library searches (Roberts et al., 2021). It then collects rankings and relevance judgements over submitted paper pools from medical domain experts and uses them to increase the accuracy of search algorithms. The potential for expanding these approaches across the sciences is considerable.

Finding analogies in distant fields often drives scientific discovery. However, the growing volume of academic publications makes it difficult to identify topics in a single discipline, let alone cross-domain analogies. Project Solvent (Chan et al., 2018) addresses this by collecting annotations of problems and findings in papers contributed by large groups of people. The annotated dataset is then used to train algorithms that identify semantic analogies between research papers. Similar data-driven methods for learning abstract relationships between concepts could be used to identify sub-problems and constraints and suggest novel R&D pathways based on analogy (Hope et al., 2021).

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 semantically-rich knowledge graphs (Auer et al., 2018). Knowledge graphs are a way to organise the world's structured knowledge by mapping the connections between different concepts and integrating information extracted from multiple data sources (Chaudhri, Chittar and Genesereth, 10 May 2021). Current automatic approaches to create these graphs only achieve moderate accuracy and have limited coverage. Where they exist, they tend to describe scholarly knowledge in specific fields, such as mathematics, chemistry and the life sciences.

A knowledge network for science must enable parallel and synchronised encoding and augmentation of knowledge. In the near future, knowledge synthesis system(s) will need to continuously navigate the development of knowledge and integrate new submissions both from theory and empirical evidence, including qualitative and quantitative studies. Intelligent interfaces could help experts find connections between concepts and theories cloaked in archaic academic language and emerging datasets and computational models (Chen and Hitt, 2021). NLP can already automate certain elements of mapping and translation between concepts or constructs to enable a more dynamic grouping of similar or complementary ideas. In the future, the development of causal inference methods within NLP may start to estimate cause-and-effect relationships (Feder et al., 2021). This is likely to be a particularly fruitful touchpoint for human-machine collaboration, as domain experts use the resulting knowledge graphs to contextualise, test and explore emerging relationships between various knowledge entities such as methods and experimental data.

To do this complex task, ontologies would have to be developed to give a stable frame to knowledge. An ontology can be thought of as a schema for organising information. For example, an ontology might be developed to define fields and subfields of research. One such schema – developed by the Web of Science – defines approximately 250 subject areas in science, social sciences, and the arts and humanities. Many other types of ontology can also be used to formally characterise academic literature. For example, ontologies might be created to specify who is considered a research contributor or to tag parts of research papers.

The Open Research Knowledge Graph (OKRG) is a good example of combined collective and machine intelligence to give structure to academic content. It organises and connects scholarly knowledge using crowdsourced contributions from researchers who acquire, curate, publish and process this knowledge (Karras et al., 2021). Enhanced machine-interpretable semantic content would help support more systems like the ORKG, making it easier for human experts to find connections and map connections based on their domain knowledge and understanding. In the meantime, collective intelligence is being used to enrich documents and annotate articles. The Dokie.li project[3] collects decentralised semantic annotations from scientists and MicroPublications, which is itself a new model to represent scientific arguments semantically.

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. During the COVID-19 pandemic, communities of academics came together to curate resources relevant to the crisis response. SciBeh, created in the early days of the pandemic, is one such community. SciBeh is a network of behavioural scientists aiming to create a crisis knowledge management infrastructure fit for rapid response, while maintaining the rigour of the scientific process.[4] Collective processes of quality assurance are increasingly important, as many published findings in social sciences have been difficult to replicate (Camerer et al., 2018). The reasons are well established, ranging from poor experiment design, small sample sizes, naive data analysis practices and a publication bias towards positive findings. Investing in new tools that automatically check scientific papers for limitations,[5] automatically categorise scientific uncertainty or predict the likelihood of replication in scholarly communication[6] could help address part of the problem. However, such systems are unlikely to be completely foolproof. They will require augmentation by highly distributed peer review or crowdsourced intelligence from multiple experts able to detect and point to the evidence-containing parts of the publications.

Introducing these new tools into established scientific practice will take time**.** In the context of knowledge discovery, AI tools are more likely to be used if they are developed in response to community needs and can be adapted based on ongoing community oversight and feedback (Halfacker and Geiger, 2020). Otherwise, they risk neglect or rejection from the users they are intended to support. If such tools are not properly adapted to user needs, they might produce results of lower quality or impact than more traditional approaches.

Experiments supported through Nesta's Collective Intelligence Grants Programme over the last two years have generated early findings about some obstacles to successful human-machine collaboration.[7] In one experiment, a serendipity-inducing recommendation algorithm, developed to improve the search for novel ideas and information, confused rather than helped groups to design new solutions to societal challenges (Gill, Peach and Steadman, 2021). These experiments highlight the importance of developing tools together with established communities of users to ensure their successful integration into existing workflows.

How to integrate combined AI-human systems into mainstream science

Scholarly communication and scientific progress have much to gain from capitalising on the emerging approaches outlined here. Some future directions on how to accelerate the integration of combined AI-human systems into mainstream science are offered below for consideration by the academic community, scientific institutions and science funders.

Invest in new infrastructure to navigate the knowledge frontier

A finer-grained and machine-actionable representation of scholarly knowledge is needed, along with the infrastructure to support knowledge curation, publishing and synthesis by humans and machines. This infrastructure should be able to support the storage of documents and datasets, as well as link this content to people and institutions.

Scholarly outputs present unique challenges for processing that necessitate the development of NLP methods optimised for this domain. This requires a more extensive research and development programme to develop processes and workflows for linking AI and human-based processing components in knowledge discovery and synthesis. Promising future directions may include new workflows that organise a pipeline of combined human-machine approaches to uploading and annotating content, followed by knowledge

graph development. Lastly, knowledge from non-traditional data sources, like tweets, drug labels, news articles and web content might be integrated into and augment literature-based discovery.

Develop AI tools that focus explicitly on co-operation and enhancing collective intelligence

Another innovation opportunity lies in creating tools to optimise the collective intelligence potential of scientific teams or wider groups through mass collaboration. This will require more research into co-operative AI systems that can enhance group problem solving and decision making for collective benefit.

Co-operative human-AI systems will be inherently complex. They will have to learn 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 (Dafoe et al., 2021). Currently, this area of research has lagged behind other topics in AI when it comes to investment (Littman et al., 2021).[8]

Make use of existing social networks to experiment with human-AI collaboration

Multiple social platforms support knowledge exchange between academics and provide an infrastructure for discovering literature. Examples include ResearchGate, Academia.edu and the Loop community from the Frontiers journals group. Some of these platforms already use AI-enabled recommendation systems to tailor content for users (Matthews, 2021). In the future, such platforms should become testbeds for experimenting with new forms of combined human-AI knowledge discovery, idea generation and synthesis. Experimentation along such lines could help these platforms develop distinctive services in comparison to those of existing social networks like Twitter and LinkedIn, which researchers increasingly use to communicate about their work.

Re-think incentives for knowledge mapping and synthesis

Several institutional, educational and social conditions inhibit knowledge integration. Existing measures of publishability reward incremental advances in depth of understanding. They also motivate discoveries built on individual disciplines rather than knowledge synthesis. Editors, reviewers and academic institutions value theoretical coherence within a specific domain with their associated traditional analyses over others. Such foci are all essential to progress in science, but an exclusive focus on these approaches to knowledge creation will miss other available opportunities. To help widen perspectives, some have argued for a new market for knowledge synthesis workers (Chen and Hitt, 2021), integrative PhD programmes[9] and/or industry research programmes to innovate based on knowledge synthesis.

Research councils and academic institutions should experiment with these proposals and support new roles and career paths. They could support what might be termed “applied metascientists” – experts in curating and maintaining information infrastructure, who can also build crucial bridges between the public, academia and industry. Other initiatives could include prize or competition mechanisms, such as Science4Cast,[10] which incentivises mapping and forecasting the future of scientific research.

Strengthen collaborative institutions

Academia, government and industry must work together on a national and international scale to create the tools for better human and machine understanding of scientific knowledge. Such collaborations can also help democratise access to the latest science-of-science algorithms and maintain a common codebase for researchers and non-technical practitioners to navigate the knowledge frontier together. The United States has begun revitalising its research infrastructure. There have also been calls for the United Kingdom to establish new institutions such as the Atlas Institute[11] to fully map the world of scientific knowledge and identify gaps. The successful uptake of research infrastructures such as repositories for preprints (e.g. arXiv) and datasets (e.g. Zenodo) has demonstrated the research community can integrate new

infrastructures. However, new infrastructures may need to reach a critical level of uptake before being widely accepted. This takes time and endorsement by the wider research community, including funders and other institutions.

Conclusion

Some of the biggest scientific challenges today are cross-disciplinary in nature and will require interdisciplinary collaboration to solve. The persistence of isolated islands of knowledge risks slowing scientific progress. Scientific and policy networks require the tools, incentives and institutional structures to track new knowledge across fields, prioritise challenges and work across disciplines to solve them. A wide spectrum of health, economic, social and corporate challenges has already benefited from human-machine partnerships. In the current trajectory, only AI will be able to keep up with and make sense of the ever-growing scientific literature. Productive human-AI collaborations can harness the best of both worlds, the human and the machine, helping scientists to identify scientific priorities and the solutions to our most pressing problems.

References

Auer, S. et al. (2018), "Towards a knowledge graph for science", in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pp. 1-6, https://doi.org/10.1145/3227609.3227689.

Berditchevskaia, A. and P. Baeck (2020), "The future of minds and machines: How AI can scale and enhance collective intelligence", 10 February, Nesta, London, www.nesta.org.uk/mindsmachines.

Bornmann, L. and R. Mutz (2015), "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references", *Journal of the Association for Information Science and Technology*, Vol. 66/11, pp. 2215-2222, https://doi.org/10.1002/asi.23329.

Camerer, C. et al. (2018), "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015", *Nature Human Behaviour*, Vol. 2/9, pp. 637-644, https://doi.org/10.1038/s41562-018-0399-z.

Chan, J. et al. (2018), "Solvent: A mixed initiative system for finding analogies between research papers", in *Proceedings of the ACM on Human-Computer Interaction*, 2 (CSCW), pp.1-21, https://doi.org/10.1145/3274300.

Chaudhri, V.K., N. Chittar and M. Genesereth (10 May 2021), "An introduction to knowledge graphs", The Stanford AI Lab Blog, http://ai.stanford.edu/blog/introduction-toknowledge-graphs.

Chen, V.Z. and M.A. Hitt (2021), "Knowledge synthesis for scientific management: Practical integration for complexity versus scientific fragmentation for simplicity", *Journal of Management Inquiry*, Vol. 30/2, pp.177-192, https://doi.org/10.1177/1056492619862051.

Chu, J. et al. (2021), "Slowed canonical progress in large fields of science", in *Proceedings of the National Academy of Sciences*, Vol 118/41, pp. e2021636118, https://doi.org/10.1073/pnas.2021636118.

Dafoe, A. et al. (2021), "Cooperative AI: Machines must learn to find common ground", *Nature*, Vol. 593/7857, pp. 33-36, https://doi.org/10.1038/d41586-021-01170-0.

Fathalla, S. et al. (2017), "Towards a knowledge graph representing research findings by semantifying survey articles" in *International Conference on Theory and Practice of Digital Libraries*, pp. 315-327, Springer, Cham, https://doi.org/10.1007/978-3-319-67008-9_25.

Feder, A. et al. (2021), "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond", *arXiv*, arXiv:2109.00725 [cs.CL], http://arxiv.org/abs/2109.00725.

Foster, J.G. et al., 2015. "Tradition and innovation in scientists research strategies", *American Sociological Review*, Vol. 80/5, pp. 875-908, https://doi.org/10.1177/0003122415601618.

Gill, I., K. Peach and I. Steadman (2021), "Collective intelligence grants programme: Experiments in collective intelligence design for social impact", 14 October, Nesta, London, www.nesta.org.uk/report/experiments-collective-intelligence-design-20/.

Halfaker, A. and R.S. Geiger (2020), "ORES: Lowering barriers with participatory machine learning in Wikipedia", *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4/CSCW2, Article 148, pp. 1-37, https://doi.org/10.1145/3415219.

Hope, T. et al. (2021), "Scaling creative inspiration with fine-grained functional facets of product ideas", *arXiv*, arXiv:2102.09761 [cs.HC], https://arxiv.org/abs/2102.09761.

Karras, O. et al. (2021), "Researcher or crowd member? Why not both! The Open Research Knowledge Graph for Applying and Communicating CrowdRE Research", *arXiv*, arXiv:2108.05085 [cs.DL], https://arxiv.org/abs/2108.05085.

Lahav, D. et al. (2021), "A search engine for discovery of scientific challenges and directions", *arXiv,* arXiv:2108.13751 [cs.CL], https://arxiv.org/abs/2108.13751.

Liekens, A.M. et al. (2011), "BioGraph: Unsupervised biomedical knowledge discovery via automated hypothesis generation", *Genome Biology*, Vol. 12/6, pp.1-12, https://doi.org/10.1186/gb-2011-12-6-r57.

Littman, M.L. et al. (2021), *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*, Stanford University, Stanford, http://ai100.stanford.edu/2021-report.

Matthews, D. (2021), "Drowning in the literature? These smart software tools can help", *Nature*, Vol. 597/7874), pp.141-142, https://doi.org/10.1038/d41586-021-02346-4.

Nathan, P. (2019), "Human-in-the-loop AI for scholarly infrastructure", 14 September, *Derwen*, https://medium.com/derwen/dataset-discovery-and-human-in-the-loop-ai-for-scholarly-infrastructure-e65d38cb0f8f.

Nielsen, M.W. et al. (2018), "Making gender diversity work for scientific discovery and innovation", *Nature Human Behaviour*, Vol. 2, pp. 726-734, https://doi.org/10.1038/s41562-018-0433-1.

Roberts, K. et al. (2021), "Searching for scientific evidence in a pandemic: An overview of TREC-COVID", *arXiv*, arXiv:2104.09632 [cs.IR], https://arxiv.org/abs/2104.09632.

Sebastian, Y. et al. (2017), "Emerging approaches in literature-based discovery: Techniques and performance review", *The Knowledge Engineering Review*, Vol. 32, p. e12, https://doi.org/10.1017/S0269888917000042.

Sell, T.K. et al. (2021), "Using prediction polling to harness collective intelligence for disease forecasting", *BMC Public Health*, Vol. 21, pp. 2132, https://doi.org/10.1186/s12889-021-12083-y.

Spangler, S. et al. (2014), "Automated hypothesis generation based on mining scientific literature", in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1877-1886, https://doi.org/10.1145/2623330.2623667.

Swanson, D.R. (2011), "Literature-based resurrection of neglected medical discoveries", *DISCO: Journal of Biomedical Discovery and Collaboration,* Vol. 6, pp. 34-47, https://doi.org/10.5210/disco.v6i0.3515.

Swanson, D.R. (1989), "Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy", *Journal of the American Society for Information Science*, Vol. 40/5, pp.356-358, https://dblp.uni-trier.de/rec/journals/jasis/Swanson89.html.

Wu, L. et al. (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.

Notes

[1] The science policy community increasingly recognises the importance of large-scale international consortia. See, for example, the *Bold Ambition: International Large-Scale Science* report from the American Academy of Arts and Sciences initiative on Challenges for International Scientific Partnerships. Full text available at www.amacad.org/publication/international-large-scale-science.

[2] In particular, transformer-based language models have been leading the field. These are deep-learning models that use the method of attention to boost their training speed. The Transformer approach was first described in https://arxiv.org/abs/1706.03762.

[3] See website for more information: https://dokie.li/.

[4] The SciBeh project was established in 2020 to manage the new information, data and resources being produced by behavioural scientists during the COVID-19 pandemic, www.scibeh.org/ (accessed 24 February 2022).

[5] SciFact, a project from the AllenAI institute, hosts a competition to develop AI models that check the veracity of scientific claims, https://leaderboard.allenai.org/scifact/submissions/public (accessed 24 February 2022).

[6] The SCORE programme from DARPA aims to develop AI-enabled tools to assign confidence metrics to results and studies across behavioural and social sciences. www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence (accessed 24 February 2022).

[7] Collective Intelligence Grants 2.0 was a GBP 500 000 fund to support experiments in collective intelligence design, focusing on the interaction between humans and machines. www.nesta.org.uk/project/collective-intelligence-grants/, (accessed 24 February 2022).

[8] There is evidence that some funders and researchers in the AI community are taking note. The Cooperative AI Foundation was established in 2021, with an initial endowment of USD 15 million. See www.cooperativeai.com/foundation.

[9] The Venture Science Doctorate from Day One Project is training graduates to combine research and entrepreneurship, thereby incentivising commercialisation of academic research, www.dayoneproject.org/post/forging-1-000-venture-scientists-to-transform-the-innovation-economy (accessed 24 February 2022).

[10] This is an open competition that aims to develop machine-learning models capable of capturing the evolution of scientific concepts and predict which research topics will emerge. See www.iarai.ac.at/science4cast/ (accessed 24 February 2022).

[11] The Atlas Institute concept was proposed in a blog by the Tony Blair Institute. www.tenentrepreneurs.org/the-way-of-the-future (accessed 24 February 2022).

Elicit: Language models as research tools

J. Byun, Ought, United States

A. Stuhlmüller, Ought, United States

Introduction

How will machine learning change research within the next decade? Large language models are a machine-learning technology that has shown promise for many reasoning tasks, including question answering, summarisation and programming. This essay outlines the experience of building Elicit, an artificial intelligence (AI) research assistant that uses language models to help researchers search, summarise and understand the scientific literature.

On 11 June 2020, OpenAI released GPT-3, a language model trained on hundreds of billions of words on the Internet. Without task-specific training, the model completed many tasks, including translation, question answering, using a novel word in a sentence and performing three-digit arithmetic. It was the largest model released at that time, and the world exploded with hobbyists' demos of GPT-3 writing code, essays and more. Since then, over 300 applications have been built on top of GPT-3, using the pretrained model for customer support, storytelling, software engineering and ad copywriting.

It is still early, but language models may become among the most transformative technologies of our time, for the following reasons:

- Language models promise to automate simple "intuitive" natural language tasks, including tasks that require knowledge of the world and basic reasoning. They would do this in the same way that early computers automated simple rule-based information processing tasks. See, for example, Austin et al. (2021) and Alex et al. (2021).
- Most improvements have come from scaling up existing models by increasing dataset sizes, model parameters and training compute, not architectural innovations. This makes it easier to predict that they will continue to improve (Henighan et al., 2020; Kaplan et al., 2020).
- In just a year, multiple providers of pretrained language models have emerged, including Cohere (2022), AI21 (2022) and the open-source effort EleutherAI (2022). This suggests that pretrained language models may become commoditised.

As of early 2022, the impacts of language models on society are unclear. There are no guarantees that language models will help substantially with research, which requires deep domain expertise and careful assessment of arguments and evidence.

This essay shares what today's models can do, how Ought has built Elicit using these models and sets out its vision of how researchers might use AI as assistant in the future.

What are language models?

The language models discussed here (generative language models) 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.

The largest models are trained on web crawl data, typically Common Crawl, a corpus of more than a trillion characters. Training proceeds a few characters at a time. Given one segment of text from the dataset, the model predicts the next one. If it predicts incorrectly, the model updates its parameters to make the correct characters more likely next time.

In one of the most surprising lessons from language models, many tasks can be framed as text prediction, including summarisation, question answering, writing computer code and text-based classification. Consider the task of recalling a word given a description of that word. In the example below, a system is shown two phrases and a meaningful completion for each, such as the following:

A quotient of two quantities: Ratio

Freely exchangeable or replaceable: Fungible

The language model picks up that the words coming after each colon ("Ratio" and "Fungible") are words defined by the phrases preceding the colon ("A quotient of two quantities" and "Freely exchangeable or replaceable"). Thus, the language model should suggest a completion such as "Catalyst" when next shown a phrase such as the following:

A person or thing that precipitates an event or change.

The previous generation of language models (GPT-2) has 1.5 billion learnt parameters. For the example above, GPT-2 does not pick up on the definition-word pattern. It does not complete the third sentence successfully. Instead, GPT-2 predicts nonsensical completions to the prompt text, such as "A firestorm later". The GPT-3 generation (175 billion parameters) correctly completes the text with the word "Catalyst". This example illustrates how the behaviour of language models changes qualitatively as models get larger.

So far, language model performance (measured by error on a test set) has improved smoothly as a function of the computational power used, dataset size and number of parameters. In each case, it assumes the other two resources are not the bottleneck (Kaplan et al., 2020). This scaling law, as it is known, together with the observed qualitative changes as model performance improves, suggests that language model capabilities will continue to improve.

Language models as research assistants today

What is Elicit?

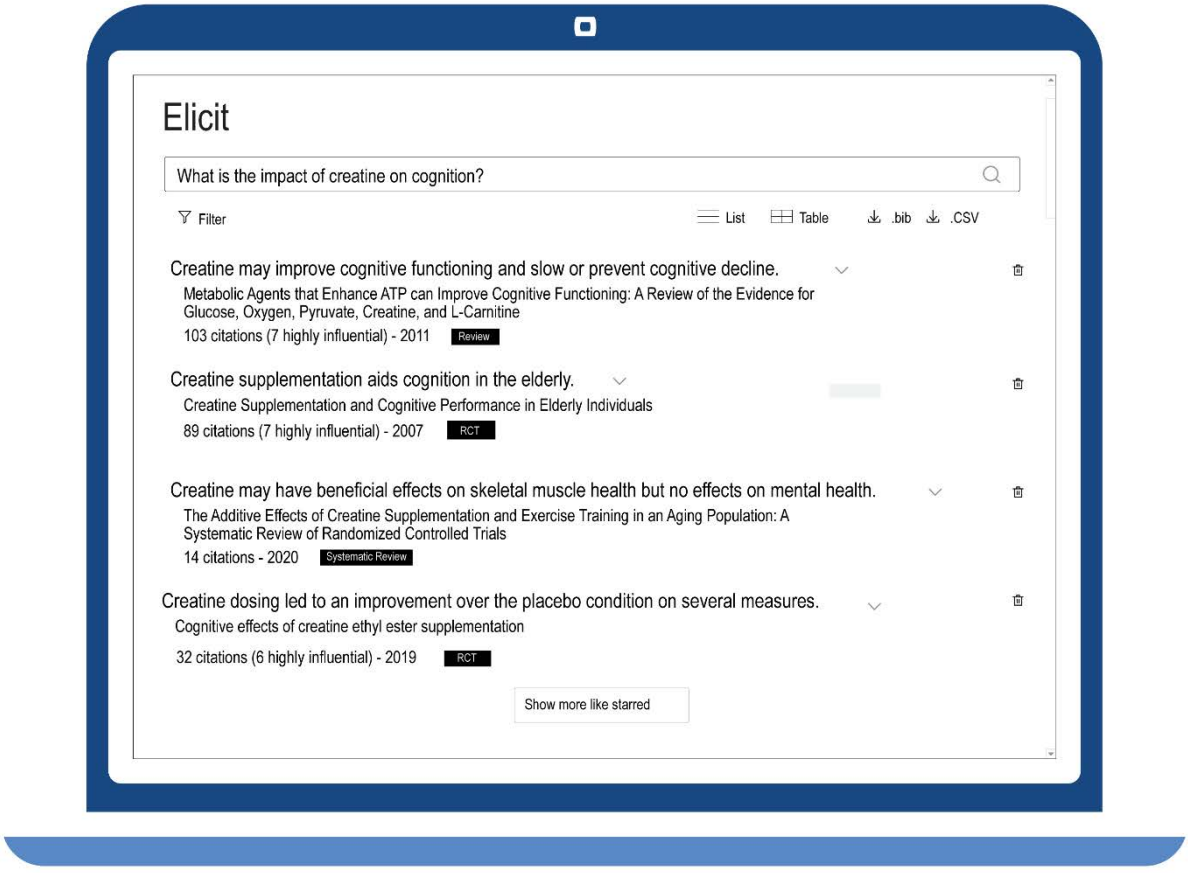
Elicit is a research assistant that uses language models – including GPT-3 – to automate research workflows. As of this writing, it is the only research assistant using large pretrained language models like GPT-3, and the only research assistant that can flexibly perform many research tasks. Researchers today primarily use Elicit for literature review (Figure 1).

Researchers can ask Elicit a question, such as "What is the impact of creatine on cognition?" Elicit returns answers and relevant academic literature. Elicit then helps researchers explore the results by surfacing key information from the papers. For example, Elicit identifies whether a paper is a randomised controlled trial, review or systematic review. Elicit can extract information about the population, intervention and outcome studied. Researchers can even ask their own questions about the returned papers, for real-time

extraction and text processing. Researchers can easily expand answers to see details about a paper and which parts of the paper Elicit used to generate its answers.

The researcher can also select particular results from a paper and Elicit will then show more papers like the selected results. Elicit accomplishes this by traversing the citation graph of the selected papers, both forwards and backwards. It looks at all the references of the selected papers, and all later papers that cited the selected papers, to find additional results. This allows the user to guide Elicit with feedback, demonstrating how AI research assistants become more effective with human feedback. Many researchers effectively run a manual and time-intensive version of this process today when they search for literature. They might start with a query in Google Scholar, open the first few papers, skim them, find interesting references and follow the citation trail. This approach quickly leads to a ballooning of possible papers to read and research directions to follow. Elicit replicates this manual process but runs it faster and more systematically.

Figure 1. An example of an Elicit literature review task



How Elicit works

The Elicit literature review workflow described above demonstrates how language models can be used for much more than text generation. It also shows how AI systems can be designed compositionally to give users more control and oversight over the AI system's work.

When a researcher conducts a literature review, their process involves subtasks such as searching, summarising or rephrasing, classifying, sorting, extracting and clustering information. Elicit trains language models to perform each of these subtasks, then builds infrastructure to string them together and automate a more complex end-to-end workflow.

Search

Elicit applies language models' understanding of semantic associations to find papers relevant to the user's query in scholarly databases like Semantic Scholar (2022). Semantic search enables researchers to find publications that help answer their questions even if they do not use the researcher's exact words.

Summarisation and rephrasing

Elicit reviews the abstracts of the papers it has found and does its best to answer the researcher's original question in a one-sentence summary. Often, this summary will be more concise and relevant than any one sentence in the abstract.

Classification

Elicit uses GPT-3 to identify whether the abstract answers "Yes" or "No" to a user's question (if the question is a yes/no question). Elicit uses another model to identify which papers are randomised controlled trials (Robot Reviewer, 2022).

Extraction

Elicit automatically extracts key information from the abstract, such as sample population, study location, intervention tested and outcome measured.

Search, summarisation, rephrasing and classification are also available as separate, individual tasks in Elicit. In addition, users can run these narrower stand-alone tasks and create their own tasks. The literature review workflow is the most advanced capability Elicit has, as it joins these tasks together to produce overall, research-backed answers to the researcher's question.

Why we need tools like Elicit

Given results from Elicit, existing research tools are clearly not designed to direct the researcher quickly and systematically to research-backed answers (Figure 2).

Google Scholar

Searches using Google Scholar (2022) often return snippets and sentence fragments that are difficult to understand. Google Scholar focuses on returning papers based on relevant keywords instead of answers. The results also require the researcher to review multiple pages and abstracts before knowing whether the papers even address their questions (Figure 3).

Semantic Scholar

Semantic Scholar is similarly designed to look up papers – the search engine is built around identifying papers given the words in a title. It is not designed to answer questions (Figure 3).

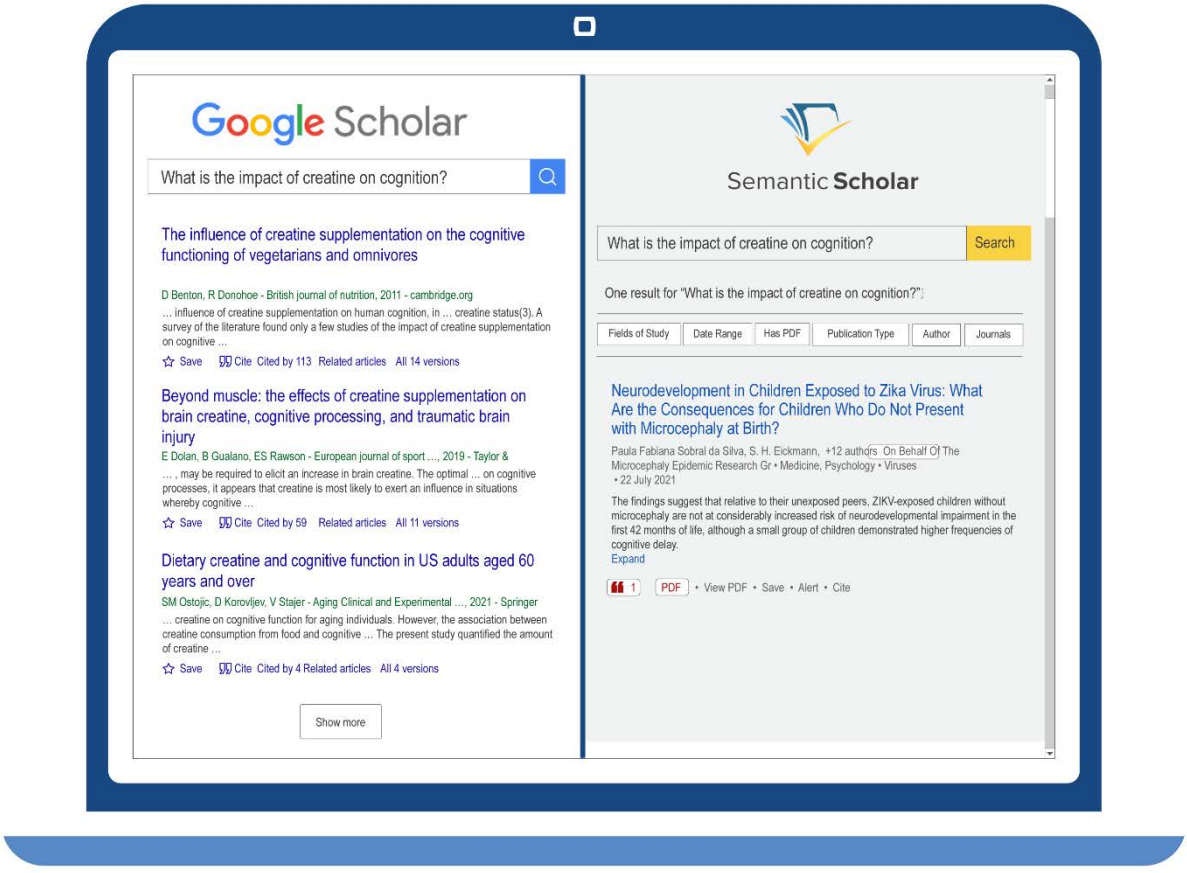
Google

Google tries harder to answer the user's question but sometimes uses less credible sources. The results answer something closer to "What does the Internet (or advertiser) think?" rather than "What does science know about this?"

GPT-3

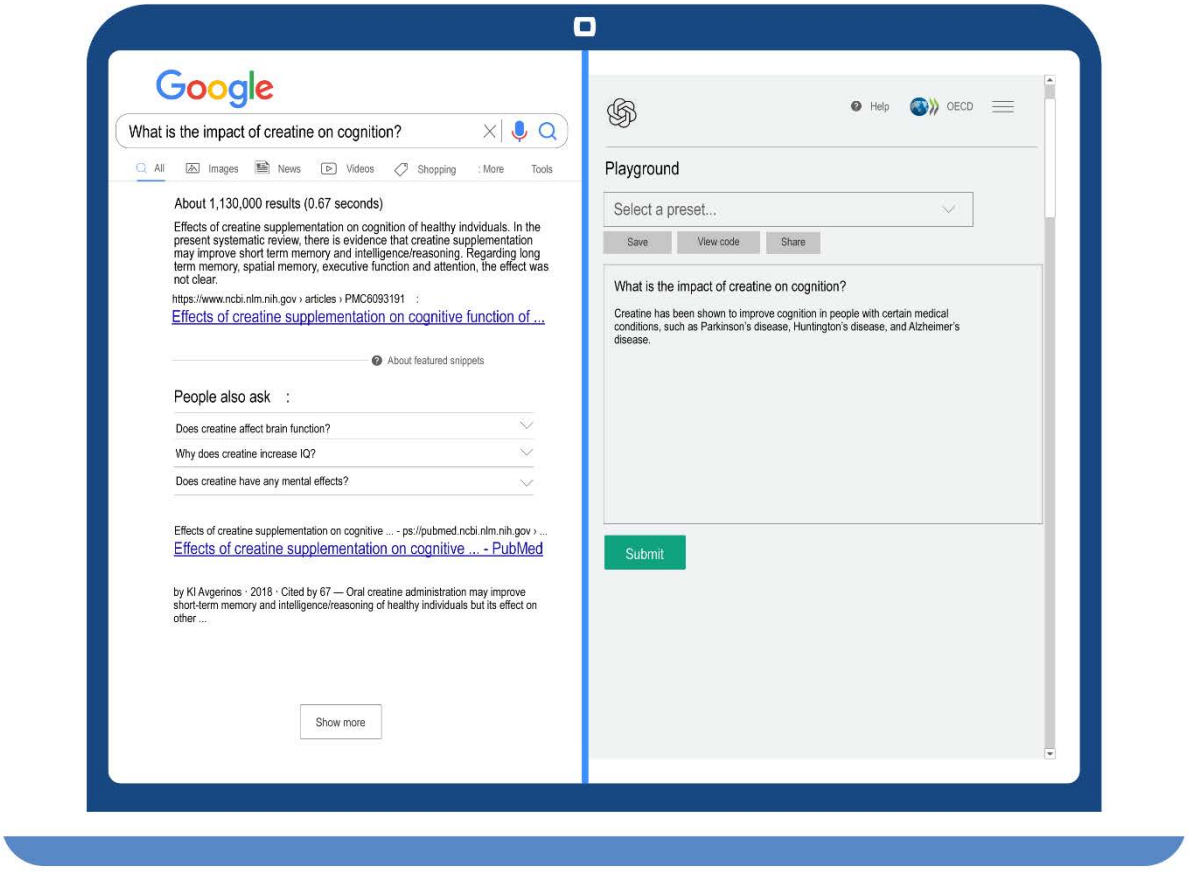
Directly prompting GPT-3 returns a coherent answer, but there is no way for the researcher to gauge its legitimacy. GPT-3 sometimes makes up information, which is a serious problem for language models (Lin, Hilton and Evans, 2021).

Figure 2. Examples of returns using search engines for the academic literature, Google Scholar and Semantic Scholar



In sum, current tools either make researchers do too much work (Google Scholar, Semantic Scholar) or are on track to generate answers without helping the researcher understand, trust or contextualise the answer. This is because either there are no sources (GPT-3) or because the results are generated in a relatively unsystematic fashion (Google). Tools are needed somewhere in the middle. Such tools would give researchers what they want as quickly as possible but also provide responses customisable enough to let users have more control over the evaluation of the results.

Figure 3. Examples of returns using the general-purpose knowledge engines Google and GPT-3



Language models as research assistants in the medium-term future

Language models today are far from automating research. However, as discussed earlier, based on trends in scaling compute, their performance is expected to continue improving. This section discusses what language models might look like on a ten-year horizon and what this may enable researchers to do. It lays out possible benefits and risks so that policy makers can help direct developments towards the benefits and away from the risks.

AI assistants in the future

In the future, researchers might spin up a "laboratory" of their own AI research assistants, each specialising in different tasks. Some of these research assistants will represent the researcher and the researcher's specific preferences about things like which questions to work on and how to phrase conclusions. Already, researchers are fine-tuning language models on their notes (Kirchner, 2021). Contrary to its portrayals in Hollywood, AI may not be a discrete entity with an independent identity (like "Samantha" from the movie *Her*) but rather highly bespoke, amplified extensions of ourselves.

Some of these assistants will use less expertise than the researcher. They will do work that researchers today might delegate to contractors or interns, like extracting references and metadata from papers (as shown in Figure 1), scraping information from websites or labelling text-based datasets.

Some assistants will use more expertise than the researcher. They might recursively simplify an explanation of the limitations of superconducting electronics, for example, as a professor might to a student. They might help a researcher evaluate the trustworthiness of findings by aggregating the heuristics of many experts and applying them across all papers. Or they might review more arguments and pieces of evidence than researchers could on their own.

Some assistants will 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. At each step, the assistants will incorporate feedback from the researcher on process and outcomes.

This compositional sub-delegation infrastructure would allow the researcher to zoom into any sub-task and troubleshoot, using assistants for help if needed. These interactions could look like workflow management tools, unstructured chat-like interactions or hybrids. Regardless of the exact interface, researchers would ideally stay in the architect's seat, overseeing the work to ensure it is aligned with their intent.

Language models and the benefits for future research

Language models can transform research in three ways: increasing productivity through time savings, enabling qualitatively new work and making research accessible to non-experts.

Increasing researcher productivity

First, language models can save researchers time. Staying on top of academic literature is already difficult. It will only get harder over the next ten years without AI tools to support researchers. The rate of new publications per year is growing exponentially in some disciplines. Some studies suggest that researchers have already surpassed the human limit for reading publications (Tenopir et al., 2015).

Many researchers have horror stories about finding the most important work they needed a year into their research. The literature review process today depends on using the right keywords. It may take hours or days before a researcher finds the exact phrase used in another domain that unlocks the most relevant literature.

In the future, language model research assistants may help researchers do the same amount of work in less time by:

- suggesting what to search for given the researcher's background
- changing search from being keyword-based to semantic, making relevant literature with different wording easy to find
- decomposing papers into units that are easier to parse (e.g. claims, evidence), and searching over those units (Chan, 2021)
- summarising parts of papers given a researcher's background, making search results easier to understand.

Saving time lets researchers do more, pushing out the frontier of science. For the same project, researchers can canvas more research. This expanded view will allow them to integrate perspectives from different disciplines and ensure they have been comprehensive.

As researchers do more, new subfields of science can emerge. Thinking about these technologies in the context of a possible decline in research productivity, it is essential to remember how much more scientific knowledge awaits discovery. Society has progressed from praying for rain to predicting rain likelihoods and quantities in 60-minute increments worldwide. What similar transformations remain in behavioural economics, neuroscience and many other domains?

Enabling qualitatively new research

The ability to apply high-quality automated reasoning to large amounts of text, and more generally at large scale, will likely catalyse fundamentally different research. This will be similar to how computers have given rise to new research fields (e.g. computer science, machine learning and biological modelling).

Bibliometric analysis may get easier over the next few years until text is as easy to analyse as numbers are today. Answering questions about research impact or productivity may not be limited to publication count analysis or well-resourced natural language processing teams. Instead, it may be done by armchair researchers and in much more depth, e.g. by conducting semi-automated reviews of research quality or impact.

Survey and interview methodology might fundamentally change. Instead of sending static questionnaires to survey participants, language models will enable dynamic question generation customised to the individual recipient and the answers already received.

Making research accessible for non-experts

When research becomes easier for researchers, it also becomes easier for research stakeholders. Better tools lower the barrier to being informed about high-quality research insight, enabling the public, industry leaders and policy makers to incorporate more research insights into their work and lives. In a future world, consuming high-quality insights could be not much harder than consuming clickbait and disinformation. It could take policy makers only minutes to comprehend the expert research they need to make mission-critical decisions.

Language models and the future of research: Possible risks

Transformative technology is necessarily unpredictable. Language models may transform the world for the better, but they could also bring risks. This section explores some possibilities to help policy makers prepare.

Making shallow work easier

Experts have mixed opinions on whether, when and to what extent language models will go beyond shallow association-based text completion and succeed at tasks that require substantial reasoning. 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 would create a disadvantage for researchers who take more time to publish higher quality research. More broadly, language models might favour certain types of research over others. The scientific community will need to carefully monitor and respond to such dynamics.

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 (May et al., 2019; Nangia et al., 2020). They also know more about famous topics and people. 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.

Misuse

In a world where language models become powerful, there will be (and already are) concerns about misuse. For examples, language models might make it easier to generate and spread false information. Language model-based tools for researchers may accelerate research on topics that come with risks, such as bioengineering and cybersecurity. Such concerns are not specific to language models but relate to progress and science broadly.[1] The best way to mitigate this risk is to direct these technologies towards reliably beneficial applications and to use them to assist people tasked with monitoring misuse and managing spurious information.

Conclusion

JCR Licklider, one of the fathers of the Internet, was one of the first people to bemoan declining research productivity. In the spring and summer of 1957, he was struggling with literature review but unable to find a rigorous study on the subject, despite copious research on potentially related topics. He cast himself as a subject and, to his dismay, found the following:

> *85 per cent of my "thinking" time was spent getting into a position to think, to make a decision, to learn something I needed to know. Much more time went into finding or obtaining information than into digesting it.... My "thinking" time was devoted mainly to activities that were essentially clerical or mechanical: searching, calculating, plotting, transforming, determining the logical or dynamic consequences of a set of assumptions or hypotheses, preparing the way for a decision or an insight. Moreover, my choices of what to attempt and what not to attempt were determined to an embarrassingly great extent by considerations of clerical feasibility, not intellectual capability (Licklider, 1960).*

In his essay “Man-computer symbiosis” (Licklider, 1960), he imagined a future where novel technologies would allow us to “think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today.”

In the 64 years since, Licklider’s vision of networked computers that transform libraries has been realised. Software and digital tools make it easier to search, calculate, plot and transform data. However, the process of preparing to think has also become harder as the amount of research required to work at the knowledge frontier has exploded. Computers have yet to help figure out what tasks to solve. In Licklider’s vision:

> *They will help not just with foreseen problems but enable people to think through unforeseen problems through an intuitively guided trial-and-error procedure in which the computer cooperated, turning up flaws in the reasoning or revealing unexpected turns in the solution...*
>
> *The human-computer interface will not be a set of precisely defined steps to take but, as we do with other people, identify an incentive or motivation and supply a criterion by which the human executor of the instructions will know when he has accomplished his task (Licklider, 1960).*

Perhaps it is fitting now, on the eve of another period of transformative technological change, to dust off these visions of human-computer symbiosis.

References

AI21 Labs (2022), “Announcing AI21 Studio and Jurassic-1 language models”, AI21 Labs, www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1 (accessed 25 November 2022).

Alex, N. et al. (2021), “RAFT: A real-world few-shot text classification benchmark”, *arXiv*, arXiv:2109.14076 [cs.CL], https://arxiv.org/abs/2109.14076v1.

Austin, J. et al. (2021), "Program synthesis with large language models", *arXiv*, arXiv:2108.07732 [cs.PL], https://doi.org/10.48550/arXiv.2108.07732.

Bommasani, R. et al. (2021), "On the opportunities and risks of foundation models", *arXiv*, arXiv:2108.07258 [cs.LG], http://arxiv.org/abs/2108.07258.

Chan, J. (2021), "Sustainable authorship models for a discourse-based scholarly communication infrastructure", *Commonplace*, Vol. 1/1, p. 8, http://dx.doi.org/10.21428/6ffd8432.a7503356.

Cohere (2022), Cohere website, https://cohere.ai (accessed 23 November 2022).

Elicit (2022), Elicit website, https://elicit.org (accessed 23 November 2022).

EleutherAI (2022), EleutherAI website, www.eleuther.ai (accessed 25 November 2022).

Google Scholar (2022), Google Scholar website, https://scholar.google.com (accessed 25 November 2022).

Henighan, T. et al. (2020), "Scaling laws for autoregressive generative modeling", *arXiv*, arXiv:2010.14701 [cs.LG], https://doi.org/10.48550/arXiv.2010.14701.

Kaplan, J. et al. (2020), "Scaling laws for neural language models", *arXiv*, arXiv:2001.08361 [cs.LG], https://doi.org/10.48550/arXiv.2001.08361.

Kirchner, J.H. (2021), "Making of #IAN", 29 August, *Substack*, https://universalprior.substack.com/p/making-of-ian.

Licklider, J.C.R. (1960), "Man-computer symbiosis", *IRE Transactions on Human Factors in Electronics*, Volume HFE-1, March, pp. 4-11, https://groups.csail.mit.edu/medg/people/psz/Licklider.html.

Lin, S., J. Hilton and O. Evans (2021), "TruthfulQA: Measuring how models mimic human falsehoods", *arXiV*, arXiv:2109.07958 [cs.CL], https://doi.org/10.48550/arXiv.2109.07958.

May, C. et al. (2019), "On measuring social biases in sentence encoders", *arXiv*, arXiv:1903.10561 [cs.CL], http://arxiv.org/abs/1903.10561.

Nangia, N. et al. (2020), "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models", *arXiv*, arXiv:2010.00133 [cs.CL], http://arxiv.org/abs/2010.00133.

Robot Reviewer (2022), "About Robot Reviewer", webpage, www.robotreviewer.net/about (accessed 25 November 2022).

Semantic Scholar (2022), Semantic Scholar website, www.semanticscholar.org/ (accessed 25 November 2022).

Tenopir, C. et al. (2015), "Scholarly article seeking, reading, and use: A continuing evolution from print to electronic in the sciences and social sciences", *Learned Publishing*, Vol. 28/2, pp. 93-105, https://doi.org/10.1087/20150203.

Note

[1] For an in-depth discussion, see chapter 5.2 of Bommasani et al. (2021).

Democratising artificial intelligence to accelerate scientific discovery

J. Vanschoren, Eindhoven University of Technology, Netherlands

Introduction

In recent years, artificial intelligence (AI) has gone from strength to strength. This has led to major scientific advances such as models that predict how proteins fold, how DNA determines gene expression, how to control plasma in nuclear fusion reactors, and many more. These advances depend on AI models – programs trained on data to recognise certain types of patterns and make predictions. The development of such models often requires large interdisciplinary teams of excellent scientists and engineers, large datasets and significant computational resources. Creating these conditions is hard, which hinders a more widespread acceleration of AI-enabled scientific discovery. This essay explores how automating a key task – the design of machine-learning models – can help democratise AI and allow many more and smaller teams to use it effectively in breakthrough scientific research.

AI models usually need to be complex to solve real-world scientific problems. They require a large amount of design and tuning based on thorough insight and intuition from both AI experts and scientists working in the domain in question. For instance, models such as AlphaFold (Jumper et al., 2021) combine deep learning (one of a broad family of AI models centred around neural networks) with built-in constraints derived from knowledge of biological and physical systems. In this way, they generate a hybrid model.

Deep-learning models have proven to be well suited for scientific problems with a massive combinatorial search space (e.g. there are 10^{300} possible conformations for an average protein), a clear metric to optimise against (e.g. how well the predicted protein structure matches the experimental observations) and lots of data to learn from (Institute for Ethics in AI Oxford YouTube Channel, 13 July 2022).

However, somewhat ironically, designing these models also requires navigating a tremendously vast search space of possible conformations of neural network architectures. For instance, some of these architectures (the structure of layers of artificial neurons, their configurations and the connections between them) can easily have 10^{18} possible configurations (Tu et al., 2022).

Discovering well-performing models is a science in itself, requiring non-trivial insight and technical expertise. A large team of research engineers can solve this problem by manual trial-and-error. However, in light of today's endemic shortage of (and intense competition for) highly trained AI experts, it is hard to scale this approach to thousands of other labs.

Democratising AI through automation

Imagine if such AI expertise could be harnessed and made universally available through easy-to-use tools that largely automate the design of machine-learning systems. This would empower all of society to apply

machine learning much more easily, in smaller teams and with fewer resources. In so doing, it would accelerate science more widely and effectively.

As Figure 1 depicts, automated machine learning (AutoML) can help democratise AI, enabling more and smaller teams to solve hard scientific problems. AutoML tools could augment effective intelligence, creating mixed teams of human scientists and AI assistants. Human scientists can make hypotheses, gather the right data and define goals. For their part, AI assistants can automatically optimise models and explore various ideas that humans could then analyse and quickly improve upon. These AI assistants could also learn across tasks and across teams, rapidly spreading effective solutions and best practices.

Figure 1. AutoML can lead to hybrid teams of humans and AI assistants to tackle major problems in science

Moreover, automation is only one aspect of democratising AI. Efficiency and safety are equally important in enabling widespread access (Talwalkar, 2018). Training many deep-learning models from scratch can be prohibitively expensive for many scientists, or require too much training data. Further advances are needed to train models much more efficiently (e.g. using continual or transfer learning where, in the latter case, knowledge gained while solving one problem would be stored and applied to different but related problems).

Widespread access to specialised computational hardware and optimised AI software is also important. In addition, it is critical to understand and audit the behaviour of AI models to verify whether they are scientifically plausible and safe to use. This may include providing interpretable explanations for predictions and evaluating any ethical ramifications.

While some requirements may be automated, human experts must often be brought into the loop and establish effective collaboration between humans and automated tools. For instance, interpretable AutoML methods have emerged that explain what health-care models have learnt (AI Pursuit by TAIR YouTube Channel, 2021) as semantically meaningful formulas. These are generally complex, such as how 20 different medical factors influence the risk of breast cancer.

Building a collective AI memory

AutoML systems often depend on hard-coded assumptions about what the models should look like. These assumptions make the search for good models faster, as long as they work well for the data at hand. As such, there is always a continuum between AutoML systems that are very general but slow, and those that are extremely specific but efficient.

AutoML-Zero (Real et al., 2020) aims to create deep-learning algorithms from scratch, evolving basic mathematical operations into entire algorithms. While it can (re)discover several known modern machine-

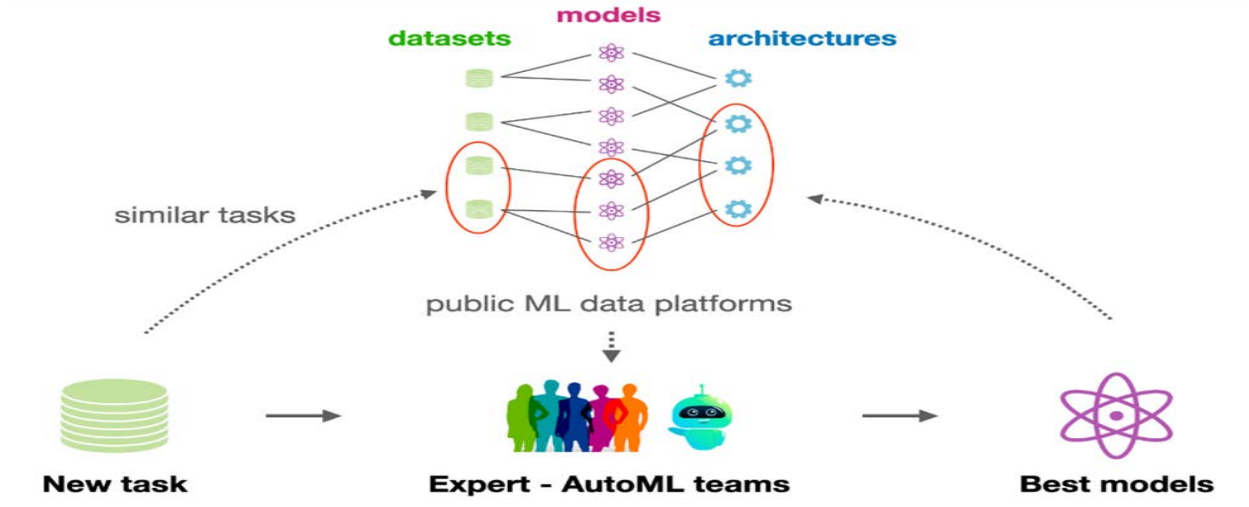
learning techniques, the process is expensive because it takes a long time to evolve such algorithms. On the other hand, by encoding elements of human knowledge, AutoML systems can become very efficient (Liu, Simonyan and Yang, 2019) but also less able to think outside the box of expert knowledge. Consequently, they are less likely to generalise to new scientific problems. As such, AutoML systems need to be imbued with assumptions that are right for each scientific problem. This can be done by embedding prior knowledge expressed by human experts (Souza et al., 2021). However, the systems could also learn directly from empirical data on the performance of AI models gathered across many scientific problems, as discussed next.

Such advances in self-learning AutoML are accelerated by the emergence of open AI data platforms, such as OpenML (Vanschoren et al., 2013). As illustrated in Figure 2, 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 neural architecture or best ways to pre-process the data they use.

Such platforms also make these data easily available through graphical, as well as programmatic, interfaces. In this way, both AutoML systems and human experts alike can use the information. For instance, researchers can look up which models work well on similar tasks and use them as a starting point.

When new models are found for new tasks, they can also be shared on the platform, thus creating a collective AI memory. Much like global databases of genetic sequences or astronomical observations, information can (and should) be collected on how to build AI models, placed on line and put through tools that help structure it to accelerate AI-driven science (Nielsen, 2012).

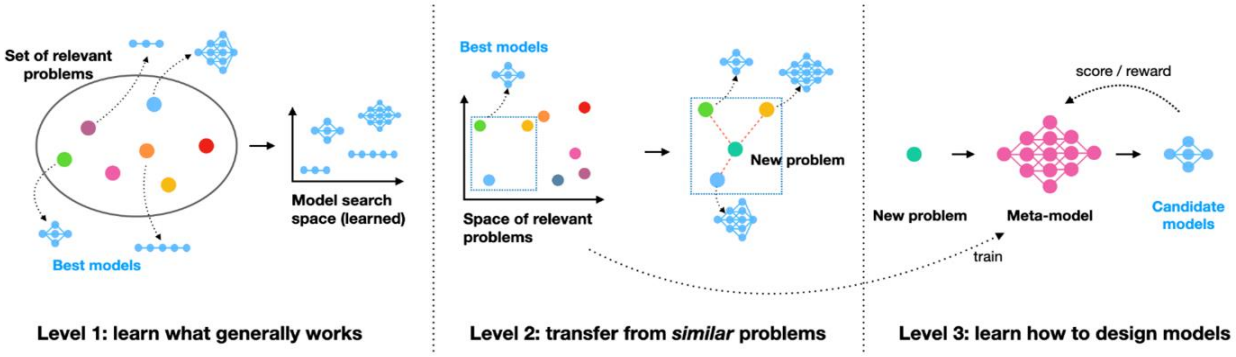
Figure 2. Collecting and organising empirical AI data across many scientific problems creates a global memory that can be leveraged to solve new scientific challenges



Learning across scientific problems

How can this global memory be leveraged effectively to find the best AI models for that problem? This is called “learning how to learn” or “meta-learning”. Meta-learning allows learning across scientific problems, transferring that learning to similar problems and using all this to design models for new problems (semi-automatically). As illustrated in Figure 3, different scenarios call for distinct approaches.

Figure 3. Meta-learning



First, the left panel shows a set of related scientific problems (e.g. various medical image analysis problems, here simply shown as dots). From these problems, models that work best for each of them can be identified. This, in turn, leads to deductions about which variety of models should be primarily considered for future problems. Learning which neural network models work best for various medical image segmentation problems, for example, would enable solving future problems faster (He et al., 2021). This can be automated further by parameterising some aspect of the model (e.g. which neural layers to use) and then learning what works well across all problems (Elsken et al., 2020).

Second, as shown in the centre panel, examining the problems themselves and extracting their key properties can help construct a (metric) space in which similar problems are close to each other and dissimilar problems are far away from each other. For instance, for problems concerning image data, the similarity of the images in two tasks could be measured (Alvarez-Melis and Fusi, 2020). Problems with similar kinds of images will then also be deemed similar. Given a new problem, information can be transferred, e.g. by recommending the best models known for the most similar prior problems (Feurer et al., 2015). Figure 3 shows how, when given a new problem, the most similar prior problems (i.e. the green, yellow and light blue dots) can be identified. The best models for old problems are likely to work well on the new problem as well.

Finally, as shown in the rightmost panel, a meta-model can be trained to predict which models to try on a given new problem. Most such meta-models go through multiple cycles, iteratively refining the model architecture to work optimally for the new problem (Robles and Vanschoren, 2019; Chen et al., 2022). Other meta-models learn how to transform the data to make them easier to model (Olier et al., 2021).

The road ahead

Automating AI has significant potential to accelerate scientific progress, but so far it 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.

Encourage more collaboration

The AutoML community and scientific communities should work closer together. While it is generally known what family of models work well for certain types of data, redesigning and tuning models to solve new scientific problems still require massive human resources. AutoML can help reduce this effort significantly. However, most AutoML researchers only evaluate their methods on specific performance benchmarks (Gijsbers et al., 2022), instead of on scientific problems where they could have much more impact. To address this issue, challenges around AutoML for science could be organised, or research that directly applies AutoML research in AI-driven sciences could be funded.

Support open AI data platforms

On a larger scale, support should be given for the development of open AI data platforms that track which AI models – such as OpenML – 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 ensure 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. Moreover, AutoML could help find these models, and even train AutoML systems on all these data to obtain even better models that help solve new scientific problems. In the past, agreements around rapid public sharing of genome data – the Bermuda principles – have led to the creation of global genome databases that now play a critical role in research. Doing the same for AI models, building databases of the best AI models for all kinds of scientific problems, could dramatically facilitate their use to accelerate science.

Moreover, to create new incentives for scientists, such platforms could track dataset and model re-use, much like existing paper citation tracking services. That way, people would get proper credit for sharing datasets and AI models. Setting this up requires public funding. The investment entailed would be both quite small overall and well worthwhile.

Create more holistic AutoML methods

AutoML methods need to become more holistic. To become true AI assistants, they need to be better at verifying and explaining the models they find to scientists. They also need to interact efficiently with domain scientists. For instance, it should be easy to define multiobjective metrics, add scientifically inspired constraints, perform safety checks and generally allow scientists to track what kind of models the system is coming up with. This would allow scientists to adjust the AutoML system's trajectory at any time.

Offer more incentives

Better incentives are needed for brilliant AI scientists and engineers to focus on solving large scientific challenges. Talent is scarce, and much of it is focused on problems that bring little long-term societal benefit. High-profile AI-driven labs could be created or supported to offer better career perspectives and sufficient compute resources. At the same time, the open release of datasets, models and infrastructure would surely help accelerate AI-driven scientific research. They may play a pivotal role in democratising AI itself.

Conclusion

AI clearly benefits science, but its true potential has not yet been reached. Since AI still relies largely on manually designed AI models, it requires extensive expertise and resources that many labs cannot easily obtain. Employing AI itself to solve this bottleneck can truly accelerate scientific discovery. This requires a data-driven approach to AI model discovery. Such an approach would collect data on which models work best for a large range of scientific problems. It would organise data in online platforms that make them easily accessible. Finally, it would leverage ALM techniques that learn from this experience and help scientists discover better models more quickly. Novel incentives for collecting data and sharing AI models can truly democratise AI and solve problems that benefit society, with machines and humans working together.

References

AI Pursuit by TAIR YouTube channel (19 October 2021), "Interpretable AutoML – powering the machine learning revolution in healthcare in the era of COVID-19", www.youtube.com/watch?v=f8JRruCRzil.

Alvarez-Melis, D. and N. Fusi (2020), "Geometric dataset distances via optimal transport", *arXiv*, arXiv:2002.02923 [cs.LG], https://doi.org/10.48550/arXiv.2002.02923.

Chen, Y. et al. (2022), "Towards learning universal hyperparameter optimizers with transformers", *arXiv*, arXiv:2205.13320 [cs.LG], https://doi.org/10.48550/arXiv.2205.13320.

Elsken, T. et al. (2020), "Meta-learning of neural architectures for few-shot learning", *arXiv*, arXiv:1911.11090 [cs.LG], https://doi.org/10.48550/arXiv.1911.11090.

Feurer, M. et al. (2015), "Efficient and robust automated machine learning", in *Advances in Neural Information Processing Systems 28*, pp. 2962-2970, http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.

Gijsbers, P. et al. (2022), "Amlb: An automl benchmark", *arXiv*, arXiv:2207.12560 [cs.LG], https://doi.org/10.48550/arXiv.2207.12560.

He, Y. et al (2021), "Dints: Differentiable neural network topology search for 3d medical image segmentation", *arXiv*, arXiv:2103.15954 [cs.CV], https://doi.org/10.48550/arXiv.2103.15954.

Institute for Ethics in AI Oxford YouTube channel (13 July 2022), "Using AI to accelerate scientific discovery", https://www.youtube.com/watch?v=AU6HuhrC65k.

Jumper, J.M. et al. (2021), "Highly accurate protein structure prediction with AlphaFold", *Nature*, Vol. 596, pp. 583-589, https://doi.org/10.1038/s41586-021-03819-2.

Liu, H., K. Simonyan and Y. Yang (2019), "DARTS: Differentiable Architecture Search", *arXiv*, arXiv:1806.09055 [cs.LG], https://doi.org/10.48550/arXiv.1806.09055.

Nielsen, M. (2012), *Reinventing Discovery: The New Era of Networked Science*, Princeton University Press.

Olier, I. et al. (2021), "Transformational machine learning: Learning how to learn from many related scientific problems", in *Proceedings of the National Academy of Sciences*, Vol. 118/49, https://doi.org/10.1073/pnas.2108013118.

Real, E. et al. (2020), "Automl-zero: Evolving machine learning algorithms from scratch", *arXiv*, arXiv:2003.03384 [cs.LG], https://doi.org/10.48550/arXiv.2003.03384.

Robles, J.G. and J. Vanschoren (2019), "Learning to reinforcement learn for neural architecture search", *arXiv*, arXiv:1911.03769, https://doi.org/10.48550/arXiv.1911.03769.

Souza, A.L.F. et al. (2021), "Bayesian optimization with a prior for the optimum", *arXiv*, arXiv:2006.14608 [cs.LG], https://doi.org/10.48550/arXiv.2006.14608.

Talwalkar, A. (2018), "Toward the jet age of machine learning", 25 April, O'Reilly Media, www.oreilly.com/content/toward-the-jet-age-of-machine-learning.

Tu, R. et al. (2022), "NAS-Bench-360: Benchmarking neural architecture search on diverse tasks", *arXiv*, arXiv:2110.05668 [cs.CV], https://arxiv.org/abs/https://arxiv.org/abs/2110.05668.05668.

Vanschoren, J. et al. (2013), "OpenML: Networked science in machine learning", *SIGKDD Explorations*, Vol. 15/2, pp. 49-60, https://doi.org/10.1145/2641190.2641198.

Is there a narrowing of AI research?

J. Mateos-Garcia, National Endowment for Science, Technology and the Arts, United Kingdom

J. Klinger, National Endowment for Science, Technology and the Arts, United Kingdom

Introduction

Large technology companies have largely driven recent advances in artificial intelligence (AI) by developing or deploying deep-learning techniques at scale. This essay examines the current state of play in AI research, including rationales for preserving technological diversity. It subsequently identifies two mutually reinforcing economic processes that may work to reduce this diversity: economies of scale and scope, and collective choice problems. The role of the private sector in further narrowing diversity is also explored. The essay ends with suggestions on how policy makers could promote technology diversity in AI research from both supply and demand sides.

Recent advances in AI have in great part been driven by deep-learning techniques developed and/or deployed at scale by large technology companies. DeepMind, a research lab owned by Alphabet, has produced important breakthroughs in game-playing and protein structure prediction. Google has developed key techniques for language modelling such as word2vec and BERT. Microsoft built the first speech recognition system to reach human-level performance. OpenAI, a not-for-profit institute backed by Microsoft, has created and commercialised GPT-3, a large language model with powerful text generation capabilities (see the essay in this report by Jungwon Byun and Andreas Stuhlmüller). The most popular software frameworks for AI research and development (R&D) – TensorFlow and Pytorch – are maintained by Google and Facebook, respectively.

Many of the ideas underpinning these advances originated in academia and public research labs, suggesting an effective flow of knowledge from the public sector to industry. Commercial demand for AI graduates is also at an all-time high (Jurowetzki et al., 2021). Meanwhile, researchers in universities and the public sector are increasingly adopting powerful open-source software tools and models developed in industry.

However, 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, a growing number of scientists and technologists has expressed concerns about the possible downsides of data and compute-intensive deep-learning methods that have come to dominate AI research. They also point to an excessive influence of corporate interests in the trajectory of such research (Marcus, 2018; Bender et al., 2021; Whittaker, 2021). Thus, the question arises: is AI research becoming too focused on a narrow set of ideas and methods aligned with the interests of influential corporate players?

In “A narrowing of AI research?”, Klinger et al. (2020) address this question by measuring the thematic diversity of the topics studied by AI researchers. They look at how topics have evolved over time, compare the diversity of AI research in academia and industry, and explore the influence of private sector research (proxied via citations) in the evolution of the field. Results suggest that technological diversity in AI research has stagnated in recent years. In addition, leading, highly influential private sector companies tend to focus

on a narrower set of state-of-the-art methods and techniques than universities. This could provide a rationale for policy interventions to preserve diversity in AI research.

The current state of play

Rationales for preserving technological diversity in AI research

It is often possible to achieve the same practical goal through different technological designs. For instance, an automobile can be powered by a combustion engine or an electric motor. Similarly, an AI system can base its decisions on a collection of logical rules or on a machine-learning model. A plurality of methods can be deployed to produce scientific knowledge in the same domain.

There are several reasons why it may be desirable to preserve such technological diversity: creativity, inclusiveness and resilience.

Creativity

Innovation involves the creative recombination of ideas, and unusual mixes are often an important source of radical and transformative innovations (Arthur, 2009). For example, today's deep-learning methods emerged at the intersection of computer science and neuroscience. Some recent advances in AI such as the AlphaGo program that defeated Go world champion Lee Sedol in 2016 also brought together state-of-the-art deep reinforcement learning techniques and traditional tree-search algorithms (Pumperla and Ferguson, 2019). Many researchers believe it is possible to overcome some limitations in deep learning-based AI systems by combining techniques from other AI traditions such as symbolic logic, causal inference or intelligence augmentation (Pearl, 2018; Marcus and Davis, 2019). A more homogenised, less diversified landscape of AI research will contain a less varied set of ideas that could be recombined in this way, potentially decreasing innovation and hindering attempts to overcome the limitations of today's dominant deep-learning designs.

Inclusiveness

Only rarely will a single technology be equally suitable for all applications, sectors and communities. In the case of AI, some sectors such as advertising and media are awash with user data that can be used to train and target deep-learning models. However, other sectors, such as education, are less data-intensive.[1] The high-stakes nature of decisions in the health sector renders deep learning less suitable than in social media or search applications. This is due to the "black-box" nature of deep-learning systems, which makes the algorithmic processes driving their results less open to direct inspection (Miotto et al., 2018; Marcus and Davis, 2019). This means that a loss of technological diversity in AI could lead to some sectors or communities lacking AI systems adapted to their needs and contexts.[2]

Resilience

Homogeneous technological ecosystems (monocultures) are more vulnerable to changes in circumstances, including the discovery of unexpected defects or limitations in a dominant design. In those cases, problems could arise from not having preserved alternative technologies that could readily be adopted. For example, the depletion of global oil reserves and the recognition of the environmental impact of CO_2 emissions called into question the reliance on combustion engines. In the case of deep learning, there is increasing evidence that AI systems based on these techniques have various weaknesses. They tend to be brittle with limited ability to generalise outside of the datasets they were initially trained on. They are also vulnerable to gaming by malicious users. In addition, they could have substantial environmental impacts due to their reliance on energy-intensive computation. Systems based on deep learning can also

be unfair. Since they have to be trained on large datasets, it is sometimes uneconomical to carefully filter out biased, prejudiced and/or inflammatory input data that could skew their outputs (Strubell, Ganesh and McCallum, 2019; D'Amour et al., 2020; Raji et al., 2021).

Reasons to expect a narrowing in AI research

There are two mutually reinforcing economic processes that may work to reduce technological diversity in AI research: economies of scale and scope, and collective choice problems.

Economies of scale and scope

In these situations, an increase in the supply or adoption of a technology makes it more attractive than its competitors. One important example of scale economies is network effects, where the size of a technology's user base increases its value to subsequent users. Network effects can potentially lead to situations where random fluctuations in adoption levels of a technology might tip the market in its favour independently from the technology's objective quality (Arthur, 1994). In two-sided markets, a technology's attractiveness depends on the presence of complementary assets such as data, computational infrastructure and/or skills. Such markets are especially important for ICT systems such as AI (Rochet and Tirole, 2006). This is demonstrated by several episodes in the history of AI where a technique benefited from independent improvements in complementary technologies that made it easier to adopt (Hooker, 2020). For example, the arrival of graphics processing units (GPUs) that could render computer game graphics efficiently lowered the barriers to deploying compute-intensive deep-learning techniques. Without the arrival of GPUs, other AI techniques may have prevailed. Once a technological design gains an edge over its competitors, this creates incentives to invest in complementary resources that strengthen that technology's advantages.

Collective choice problems

The uncoordinated behaviours of individual actors such as research teams or firms could limit technological diversity. For example, innovators could have fewer incentives to invest in developing a second-tier technology against a dominant one (Acemoglu, 2012). They may also choose to focus on technologies that generate more short-term returns, even when they know alternatives would be more beneficial in the longer term (Bryan and Lemus, 2017). In the case of AI, there is a growing sense that publication, commercial and geopolitical races could be encouraging such short-term behaviours (Armstrong, Bostrom and Shulman, 2016). Research teams and countries competing fiercely to advance the state-of-the-art in AI benchmarks, launch new products and become global AI leaders are more likely to focus their efforts on advancing the dominant paradigm (i.e. deep learning). They are less likely to explore "second-tier" techniques that preserve AI's technological diversity but have uncertain benefits.

The role of the private sector

Private sector participation in the development of a technology could intensify the pressures that narrow it. After all, commercial actors have strong incentives to invest in technologies that can be more readily deployed and to leverage their investments in technology across more markets. The results can include behaviours that drive product life cycles in certain ways. The exploration of alternative technologies, for example, could be followed by exploitation of a dominant design (Utterback and Abernathy, 1975); a competition to establish technical standards that harness network effects to dominate the market (Shapiro and Varian, 1998); and homogenisation of industries as organisations become more similar to each other to facilitate flows of knowledge and talent (Beckert, 2010). Businesses might also steer the trajectory of a technology in directions aligned with their particular interests, potentially neglecting negative externalities, unintended consequences and societal preferences.

These concerns are visible in AI research as large technology companies become more influential. These firms are making vast investments to develop deep-learning techniques that complement their assets (big data and computational infrastructure) and applications (e.g. information search, content filtering and ad-targeting). Some evidence suggests these investments are draining researchers from academia. Similarly, evidence points to skewed research priorities of public research labs that receive private funding from and/or need to collaborate with industry to access the large datasets and infrastructures required for cutting-edge research (Jurowetzki et al., 2021; Whittaker, 2021). Meanwhile, technology companies might have incentives to downplay the limitations and risks of deep-learning techniques that increasingly sit at the core of their products and services (Bender et al., 2021). All of this could lead to what some researchers have referred to as a "de-democratisation" of AI research. In such an environment, AI research focuses on a narrow set of compute-intensive techniques mainly developed and deployed by a small number of private research labs and their collaborators in elite universities (Ahmed and Wahed, 2020).

A further narrowing of AI research?

The discussion above has provided theoretical reasons and supporting evidence for three points:

1. It would be desirable to preserve AI's technological diversity.
2. Economies of scale and scope and collective choice problems could make AI research narrower.
3. Increasing private sector participation and influence in AI research may intensify this process.

The paper, "A narrowing of AI research?" (Klinger et al. (2020), sought to improve the evidence base about points (2) and (3). It conducted a quantitative analysis of 1.8 million articles from *arXiv*, a preprint repository widely used by the AI research community to disseminate its work. Having identified around 100 000 AI papers in this corpus, the authors analysed their abstracts to measure thematic concentration and heterogeneity and construct several indicators of technological diversity.[3] They then analysed the evolution of these metrics over time, thus addressing point (2).

They also extracted information about the institutional affiliation of each article's authors. This aimed to measure private sector participation in AI research. It also aimed to compare the thematic diversity and influence of "public" and "private" sector AI research overall, thus addressing point (3). Three key findings are summarised below.

There is evidence of a recent stagnation and even decline in the diversity of AI research

All metrics of diversity show that technological diversity in AI research has expanded since the late 2000s. However, this growth has stagnated and even started to decline from the mid-2010s. This is despite a substantial increase in the number of AI publications in recent years (60% of the AI articles in the corpus studied were published after 2018). Such an increase might have been expected to expand the range of AI techniques and applications explored. Analysis of the factors behind the stagnation of technological diversity in AI research shows increasing concentration of research in a small number of influential topics related to deep learning.

Private AI research is thematically narrower and more influential than academic research, and it focuses on computationally intensive deep-learning techniques

Private companies are ten times more likely to participate in AI research than in other research contained in *arXiv*. In 2020, 20% of AI papers involved at least one researcher affiliated with a private company, with large US technology companies such as Google, Microsoft, IBM, Facebook and Amazon ranking highest.

This private body of AI research is narrower than the public body according to all the metrics used, even after adjusting for differences in the number of papers produced. The analysis also shows that private companies tend to have narrower research profiles than universities after controlling for their volume of AI research, the year a paper was published and unobservable organisation-specific factors. Private

companies tend to specialise in state-of-the-art deep-learning topics in computer vision and computer language; infrastructure to scale up computationally intensive AI methods; and applications in online search, social media and ad-targeting. By contrast, they tend to be less focused on health applications of AI and analyses of the societal implications of AI.

The authors also found that AI research involving companies tends to be more highly cited even after controlling for its topic. This is consistent with the idea that the private sector might be shaping the evolution of the field directly through the research it publishes, and indirectly by providing a foundation that other researchers build on.

Elite academic institutions have similar research profiles to private sector institutions

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. These institutions include MIT; University of California, Berkeley; Carnegie Mellon; and Stanford University. Such influential universities tend to be the top collaborators of private companies, suggesting some homogenisation at the top of AI research.

Conclusion

The analysis has some limitations given the authors only consider published AI research. They were not able to make any causal statements about the direct impact of private sector participation in AI research. Nor could they make strong inferences about why companies are less thematically diverse than other institutions.

Perhaps most importantly, the analysis says little about the impacts of a loss of technological diversity in AI research. The concentration of efforts in a dominant design evidenced here might simply be making research more efficient by reducing the dissipation of efforts down unproductive dead-ends. Some theoretical and qualitative arguments were made for why this perspective might be excessively optimistic, but more evidence is required to bolster this case. This would require quantification of the loss of resilience, creativity and inclusiveness brought about by a narrowing of AI research – all of which are important questions for future work. Lacking that, portfolio theory suggests likely disadvantages in focusing all (or most) AI research on a single family of (deep-learning) techniques that, as a growing number of voices in the field argue, also have important limitations and risks. This provides a rationale for policy makers to consider how they might spur technological diversity in AI research.

What can policy makers do?

Technological diversity and the supply side

The analysis shows that universities tend to produce more diverse AI research than the private sector. Accordingly, bolstering public R&D capabilities might make the field more diverse. This could be done through increases in the levels of research funding and the supply of talent, computational infrastructure and data for publicly oriented AI research. A larger talent pool would reduce the impact of a brain drain 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 (Ho et al., 2021).

Recognising the propensity towards “research bandwagons” in academia, research and 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. Initiatives to increase the sociodemographic diversity and inclusiveness of the AI research talent pool would broaden the range of perspectives and preferences brought into the design and evaluation of AI technologies. This might make them more thematically diverse (Acemoglu, 2012).

AI researchers in the public sector have much to learn from industry teams that regularly share their code and data and build robust tools to make their findings easier to reproduce and their methods easier to deploy. Policy makers should strengthen incentives for adoption of these open science and open-source methods in academia.

Demand side of technological diversity

New datasets, benchmarks and metrics could reflect the limitations of deep-learning techniques (for example in terms of energy consumption) 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 and increase adoption of AI methods in underserved sectors. This, in turn, could spur development of new techniques more relevant for domains where deep learning is less suitable. Finally, policy makers could consider regulatory interventions, possibly focused on specific use cases, that penalise the negative externalities of deep-learning methods. One case, for example, is the impact of environmental costs and risks for minority communities. Such interventions might encourage their developers and adopters to explore alternative techniques.

Designing and implementing these initiatives will require policy makers to overcome three substantial barriers:

1. There are strong incentives for policy makers to retain a short-sighted focus on exploiting dominant AI technologies. Instead, they need to explore alternatives relevant to their countries, societal challenges, and scientific and technological capabilities.
2. Policy makers needs more expertise and know-how to help them decide what sort of technology initiatives to support.
3. Policy makers need to countervail the massive investments by the private sector on AI R&D. Their ability to do so could help prevent a premature lock-in to powerful yet limited deep-learning techniques. This could provide the foundations for future AI revolutions with fewer risks and more widely shared benefits.

References

Acemoglu, D. (2012), *Diversity and Technological Progress,* University of Chicago Press.

Ahmed, N. and M. Wahed (2020), "The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research", *arXiv,* preprint arXiv:2010.15581, https://doi.org/10.48550/arXiv.2010.15581.

Armstrong, S., N. Bostrom and C. Shulman (2016), "Racing to the precipice: A model of artificial intelligence development", *AI & Society*, Vol. 31/2, pp. 201-206.

Arthur, W.B. (2009), *The Nature of Technology: What It Is and How It Evolves,* Simon & Schuster, New York.

Arthur, W.B. (1994), *Increasing Returns and Path Dependence in the Economy,* University of Michigan Press.

Beckert, J. (2010), "Institutional isomorphism revisited: Convergence and divergence in institutional change", *Sociological Theory*, Vol. 28/2, pp. 150-166, www.jstor.org/stable/25746221.

Bender, E.M, et al. (2021), "On the dangers of stochastic parrots: Can language models be too big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, https://doi.org/10.1145/3442188.3445922.

Bryan, K.A. and J. Lemus (2017), "The direction of innovation", *Journal of Economic Theory*, Vol. 172, pp. 24772, https://doi.org/10.1016/j.jet.2017.09.005.

D'Amour, A. et al. (2020), "Underspecification presents challenges for credibility in modern machine learning", *arXiv*, preprint arXiv:2011.03395, https://doi.org/10.48550/arXiv.2011.03395.

Ho, D.E. et al. (2021), *Building a National AI Research Resource: A Blueprint for the National Research Cloud,* The Stanford Institute for Human-Centered Artificial Intelligence, Stanford, https://hai.stanford.edu/sites/default/files/2021-10/HAI_NRCR_2021_0.pdf.

Hooker, S. (2020), "The hardware lottery," *arXiv*, preprint arXiv:2009.06489, https://doi.org/10.48550/arXiv.2009.06489.

Jurowetzki, R. et al. (2021), "The privatization of AI research (-ers): Causes and potential consequences – from university-industry interaction to public research brain-drain?" *arXiv*, preprint arXiv:2102.01648, https://doi.org/10.48550/arXiv.2102.01648.

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

Marcus, G. (2018), "Deep learning: A critical appraisal", *arXiv*, preprint arXiv:1801.00631, https://doi.org/10.48550/arXiv.1801.00631.

Marcus, G. and E. Davis (2019), *Rebooting AI: Building Artificial Intelligence We Can Trust,* Vintage, New York.

Miotto, R. et al. (2018), "Deep learning for healthcare: Review, opportunities and challenges", *Briefings in Bioinformatics*, Vol. 19/6, pp. 1236-1246, https://doi.org/10.1093/bib/bbx044.

Pearl, J. (2018), "Theoretical impediments to machine learning with seven sparks from the causal revolution", *arXiv*, preprint arXiv:1801.04016, https://doi.org/10.48550/arXiv.1801.04016.

Raji, I.D et al. (2021), "AI and the everything in the whole wide world benchmark", *arXiv*, preprint arXiv:2111.15366, https://doi.org/10.48550/arXiv.2111.15366.

Rochet, J.-C. and J. Tirole (2006), "Two-sided markets: A progress report", *The RAND Journal of Economics*, Vol. 37/3, pp. 645-667, https://www.jstor.org/stable/25046265.

Shapiro, C. and H.R. Varian (1998), *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press, Boston.

Strubell, E., A. Ganesh and A. McCallum (2019), "Energy and policy considerations for deep learning in NLP", *arXiv*, preprint arXiv:1906.02243, https://doi.org/10.48550/arXiv.1906.02243.

Utterback, J.M. and W.J. Abernathy (1975), "A dynamic model of process and product innovation", *Omega*, Vol. 3/6, pp. 639-656, https://doi.org/10.1016/0305-0483(75)90068-7.

Whittaker, M. (2021), "The steep cost of capture", *Interactions*, Vol. 28/6, pp. 50-55.

Wooldridge, M. (2020), *The Road to Conscious Machines: The Story of AI*, Penguin, London.

Notes

[1] While it is likely that education will become increasingly "datafied", concerns about privacy and the challenges of measuring educational outcomes will tend to hinder the deployment of AI systems at the scale seen in the web and media sectors.

[2] Arguably, sufficient generalisability in a single (dominant) AI design could make it repurposable for all use-cases. However, such generalisability still seems far away, again providing reasons for preserving research diversity around such topics as AI techniques suitable for low-data, fast-changing and high-stakes contexts where deep learning techniques currently under-perform.

[3] To identify AI papers, salient terms for papers have been extracted that have been classified by their authors in machine learning/neural network categories. Papers with high frequency of those terms have been sought outside of those categories.

Lessons from shortcomings in machine learning for medical imaging

G. Varoquaux, Institut national de recherche en sciences et technologies du numérique (INRIA), France

V. Cheplygina, IT University of Copenhagen, Denmark

Introduction

The application of machine learning (ML) to medical imaging has attracted a lot of attention in recent years. Yet, for various reasons, progress remains slow. This essay builds upon earlier work by the authors which explores how larger datasets and more deep-learning algorithms have not yet provided practical improvements in addressing clinical problems. It recommends how researchers and policy makers can improve the situation.

Many opportunities exist to improve patients' health by applying ML to medical imaging. Through computer-aided diagnosis, for example, an algorithm is trained on existing images such as brain scans of people with and without dementia. It is later applied to unseen images to predict which group they likely belong to. There are now numerous reports of ML algorithms recognising medical images more accurately than human experts (for an overview see Liu et al., 2019).

Despite this potential, incentives in (ML) research are slowing progress in the field. For example, the impact on clinical practice has not been proportional to claims. Roberts et al. (2021) found that none of the 62 published studies on ML for COVID-19 had potential for clinical use. Studies for other clinical applications of ML have also failed to find reliable published prediction models. Two examples are for prognosis after aneurysmal subarachnoid haemorrhage (Jaja et al., 2013) and stroke (Thompson et al., 2014).

Table 1 describes key concepts, some of which might differ in their use depending on the community. The following sections summarise examples of lack of progress. The essay then provides recommendations for researchers and policy makers on how to move forward.

Table 1. Terms frequently used in machine learning in the context of medical imaging

Dataset	A dataset is a collection of (image, label) pairs where the label is either a category (such as disease or healthy), or another image (such as a segmentation map showing the locations of tumours).
Algorithm, classifier, model	This is either a general concept (such as a neural network) or a model trained on specific data.
Training	"Training" means fitting a model to a specific dataset by having it learn parameters to transform the image into the label as well as possible.
Testing, predicting	This means running a trained model on images to output their predicted labels. Note that prediction does not imply forecasting, as the data are already available.
Test set	This is part of the dataset reserved for evaluating the trained model. Ideally, these data should be previously unseen, but in practice are often already available to the researcher.

Is artificial intelligence research missing its target?

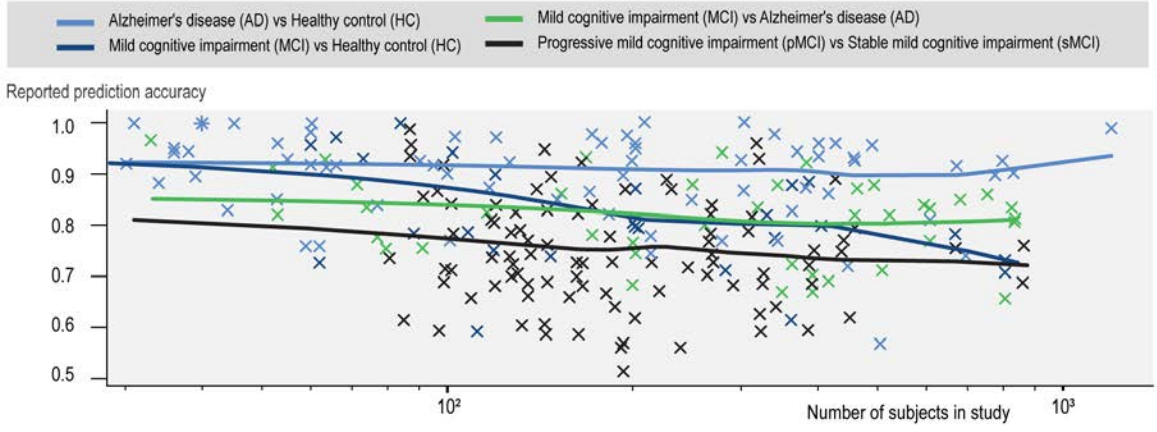
The increased popularity of ML in recent years is often explained by two developments. First, larger datasets are available. Second, deep-learning techniques permit development of algorithms without specialised domain knowledge, allowing more researchers into a field. However, the state of ML in medical imaging is not as positive as many believe for the three reasons noted below.

Large datasets are not a panacea

There is a tendency to expect that a clinical task can be "solved" if the dataset is large enough. After all, from prior research, large and diverse datasets help an algorithm generalise better to previously unseen data. There are several problems here. First, not all clinical tasks translate neatly into ML tasks. Second, creating larger and larger datasets often relies on automatic methods that may introduce errors and bias into the data (Oakden-Rayner, 2020). For example, a machine might label x-rays as showing the presence or non-presence of pneumonia based on words appearing in the associated radiology reports. In such a case, a phrase like "no history of pneumonia" might result in an x-ray wrongly labelled as showing the presence of pneumonia.

Finally, while large datasets improve algorithm training and generalisation, they also allow for better evaluation of algorithms. This is because more data are available for creating an estimate of performance on previously unseen future data. Analysis of predictions of Alzheimer's disease across six surveys and more than 500 publications in Figure 1 shows that studies with larger sample sizes tend to report worse prediction accuracy. This is worrying since these studies are closer to real-life settings.

Figure 1. Larger brain-imaging datasets do not yield better machine-learning diagnosis of Alzheimer's disease



Note: From a clinical standpoint, two important medical imaging problems using machine learning are to 1) distinguish Alzheimer's disease (AD) both from a healthy control (HC), and from mild cognitive impairment (MCI), which can signal the onset of AD; and 2) distinguish progressive mild cognitive impairment (pMCI) from stable mild cognitive impairment (sMCI).
Source: Varoquaux and Cheplygina (2022).

Algorithm research may hit diminishing returns

A lot of research within medical imaging focuses on algorithm development, but the practical benefits of the reported accuracy gains are not always clear. For this essay, the authors studied eight medical imaging competitions on Kaggle, a platform where algorithm developers can compete to solve classification tasks, and where winning can involve significant incentives. Indeed, the most famous competition on lung cancer

prediction had a prize of USD 1 million. The analysis compared two quantities: 1) the gap between the performance of the top algorithms; and 2) the expected variability in performance if a different subset of the data was used for evaluation. In other words, it tried to quantify the meaning of the final ranking. Would ranking of the winners alter if other images were used from the same or a different subset of the data? In most cases, the performance of the top algorithms is within the expected variability, and algorithms are thus not practically better or worse than one another (Varoquaux and Cheplygina, 2022).

Lack of representation from underdeveloped regions

Deep-learning studies are computationally intensive, and several ML studies have noted how this affects who gets to do research. A method may win just because more computational resources were available (Hooker, 2020). Meanwhile, the representation of prestigious labs and tech companies at conferences is increasing (Ahmed and Wahed, 2020). At a large medical imaging conference –MICCAI 2020 – only 2% of accepted papers were from underrepresented regions (Africa, Latin America, South/South-East Asia and the Middle East) (MICCAI Society, 2021). However, the need for medical AI might be even greater in these regions.

Recommendations for research communities

There are a number of things that researchers concerned with these questions can do already, especially those organising conferences, and/or editing or reviewing papers.

Build awareness of data limitations

It may not always be feasible to collect more data. However, it is important to understand the limitations of the data that are available, such as the sample size and characteristics of different patient groups. On this note, datasets should include a report of the data characteristics, as well as the potential implications for models trained on the data. Such a practice would be similar to providing “model cards”, a short document that accompanies a trained ML model and details benchmarked model performance under different conditions (Mitchell et al., 2019).

Reinvent benchmarking

Benchmarking the performance of algorithms alone is not sufficient to advance the field. Papers focusing on understanding, replication of earlier results and so forth are also valuable. If benchmarking the performance of algorithms is deemed essential in a publication, comparisons need to include both recent-and-competitive and traditional-yet-effective methods.

Furthermore, comparisons need to consider the range (rather than a single estimate) of each method’s performance. Ideally, they should use multiple, well-motivated metrics and statistical procedures (Bouthillier et al., 2021). More real-life effects of an algorithm might also be considered. This might include, for example, its carbon footprint, or how it affects the people it was designed to help (Thomas and Uminsky, 2020).

Improve publication norms

Many want to believe that publishing a novel algorithm with state-of-the-art results is the only way to create impact, but such results may be overly optimistic. In the practice of psychology, registered reports are prepared. In this approach, a planned study is reviewed and published before any experiments are done. More widespread adoption of this practice could reduce publication bias since “negative results” would also be published. From an institutional perspective, one could support different types of papers focusing

on different forms of insight. These could include replications or retrospective analyses of methods, incentivising and rewarding (e.g. through research funding, hiring decisions) such practices.

Recommendations for research policy makers: Setting incentives

As research positions and funding are often tied to the output of publications, researchers have strong incentives to optimise for publication-related metrics. With the additional focus on achieving novelty and state-of-the-art results, the publication of papers using methods that are over-engineered but under-validated is perhaps not surprising. While some researchers might choose to opt out of this dynamic and/or try to change things, many in less secure positions may pursue publication-related metrics to benefit their career. It is therefore important that external incentives are created to speed up the change towards methods with greater validation.

Quality rather than quantity

Several of the current problems stem from the way researchers are evaluated when applying for academic positions or for research funding. The focus on metrics like the h-index needs to be reduced in favour of other practices, such as, for example, an evaluation of five selected publications. Such a shift could reduce the pressures that lead to publication of research with diminishing returns. The need for new approaches for evaluating research also holds when evaluating researchers based on previously acquired funding, which can entail the propagation of existing biases.

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, replication of existing studies and prospective 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.

Better recognition for open data and software

It should be more attractive to work on curated datasets and open-source software that everybody can use. It is difficult to acquire funding, and often to publish, when working on such projects. Many team members are therefore volunteers. This creates biases against groups that are already underrepresented but that might have innovative ideas that would be vital for the field. Such groups could include, for example, women who take on a greater share of household responsibilities and lower-income countries who cannot afford to take on unpaid jobs. More regular funding and consequently more secure positions would help to improve on the status quo.

Conclusion

This essay has presented insights on a number of problems that may be slowing the progress of ML in medical imaging. These insights are based on both a review of the literature and the authors' previous analysis. In summary, not everything can be solved by having larger datasets and by developing more algorithms. The focus on novelty and state-of-the-art results creates methods that often do not translate into real improvements. The essay proposes a number of strategies to address this situation, both within the research community and at the level of research policy. Given the huge efforts invested in AI research, failure to address these issues could mean significant waste.

References

Ahmed, N. and M. Wahed (2020), "The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research", *arXiv*, preprint arXiv:2010.15581, https://doi.org/10.48550/arXiv.2010.15581.

Bouthillier, X. et al. (2021), "Accounting for variance in machine learning benchmarks" in *Proceedings of Machine Learning and Systems*, Vol. 3, pp. 747-769, https://proceedings.mlsys.org/paper/2021/hash/cfecdb276f634854f3ef915e2e980c31-Abstract.html.

Hooker, S. (2020), "The hardware lottery", *arXiv*, preprint arXiv:2009.06489, https://doi.org/10.48550/arXiv.2009.06489.

Jaja, B.N. et al. (2013), "Clinical prediction models for aneurysmal subarachnoid hemorrhage: A systematic review", *Neurocritical Care*, Vol. 18/1, pp. 143-153, https://doi.org/10.1007/s12028-012-9792-z.

Liu, X. et al. (2019), "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis", *The Lancet Digital Health*, Vol. 1/6, pp. e271-e297, https://doi.org/10.1016/S2589-7500(19)30123-2.

MICCAI Society (2021), "MICCAI Society News", 18 August, MICCAI Society, www.miccai.org/news/.

Mitchell, M. et al. (2019), "Model cards for model reporting", in *FAT* '19*: *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 220-229, https://doi.org/10.1145/3287560.3287596.

Oakden-Rayner, L. (2020), "Exploring large-scale public medical image datasets", *Academic Radiology*, Vol. 27/1, pp. 106-112, https://doi.org/10.1016/j.acra.2019.10.006.

Roberts, M. et al. (2021), "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans", *Nature Machine Intelligence*, Vol. 3/3, pp. 199-217, https://doi.org/10.1038/s42256-021-00307-0.

Thomas, R. and D. Uminsky (2020), "The problem with metrics is a fundamental problem for AI", *arXiv,* preprint arXiv:2002.08512, https://doi.org/10.48550/arXiv.2002.08512.

Thompson, D. et al. (2014), "Formal and informal prediction of recurrent stroke and myocardial infarction after stroke: A systematic review and evaluation of clinical prediction models in a new cohort", *BMC Medicine*, Vol. 12/1, pp. 1-9, https://doi.org/10.1186/1741-7015-12-58.

Varoquaux, G. and V. Cheplygina (2022), "Machine learning for medical imaging: Methodological failures and recommendations for the future", *Nature Digital Medicine*, in press. https://doi.org/10.1038/s41746-022-00592-y.

Part IV Artificial intelligence in science: Implications for public policy

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

T. Hey, UK Research and Innovation, United Kingdom

Introduction

The rapid growth of scientific data generated both by scientific experiments at large national and international facilities and by model simulations on supercomputers epitomises Jim Gray's "Fourth Paradigm" of data-intensive science. The use of artificial intelligence (AI) technologies to help automate the generation and analysis of such datasets is increasingly necessary. This essay describes the great potential for the use of AI and deep learning technologies to transform many fields of science. It draws particular attention to the conclusions of Town Hall meetings organised by the US Department of Energy (DOE). These meetings explored the potential for AI to accelerate science and the need for major public research and development (R&D) funding. Such funding could enable multidisciplinary teams of academic researchers to generate comparable breakthroughs to those of commercial companies such as Google DeepMind.

Deep learning (DL) neural networks – a sub-discipline of AI – came to prominence in 2012 when a team led by Geoffrey Hinton won the ImageNet Image Recognition Challenge (Krizhevsky et al., 2012). Their entry in the competition, AlexNet, was a DL network consisting of eight layers. The learning phase was computed on graphics processing units (GPUs), a specialised form of electronic circuit frequently used in software-based games. By 2015, building on this initial research, a Microsoft Research team used a DL network with more than 150 layers trained using clusters of GPUs to achieve object recognition error rates comparable to human rates (He et al., 2016).

DL networks are now a key technology for the IT industry and used for a wide variety of commercially important applications. These include facial recognition, handwriting transcription, machine translation, speech recognition, autonomous driving and targeted advertising. More recently, Google's UK subsidiary, DeepMind, used DL neural networks to develop the world's best Go playing systems with their AlphaGo variants.

Of particular interest for science is DeepMind's AlphaFold protein-folding prediction system (Senior et al., 2020). Their latest version of AlphaFold convincingly won the most recent Critical Assessment of Protein Structure Prediction (Jumper et al., 2021). Nobel Prize winner Venki Ramakrishnan said, "This computational work represents a stunning advance on the protein folding problem, a 50-year-old grand challenge in biology. It has occurred decades before many people in the field would have predicted. It will be exciting to see the many ways in which it will fundamentally change biological research" (DeepMind, 30 November 2020).

This essay reviews the changing face of much data-driven scientific research, and the impact of DL and other AI technologies.

The four paradigms of scientific discovery

Turing Award winner Jim Gray was the first to use the term "Fourth Paradigm" to describe the next phase of data-intensive scientific discovery (Gray, 2009). In the first paradigm, which lasted over 1 000 years, science was empirical, based solely on observation. Then, in 1687, after the discoveries of Kepler and Galileo, Isaac Newton published the *Mathematical Principles of Natural Philosophy*. This established his three laws of motion that defined classical mechanics and provided the foundation for his theory of gravity. The mathematical laws of nature provided the basis for theoretical explorations of scientific phenomena, a second paradigm for scientific discovery. Nearly 200 years later, Maxwell formulated equations for his unified theory of electromagnetism, and then, in the early 20th century, Schrödinger's equation described quantum mechanics. The use of these two paradigms – experimental observation and theoretical calculation – has been the basis for scientific understanding and discovery for the last few centuries.

In 2007, working on a study of computing futures, Jim Gray and the Computer Science and Telecommunications Board realised that computational science was a third paradigm for scientific exploration. It involved a shift towards simulation based on, and generating large volumes of, scientific data created in the first instance by digital instruments.

In his talk to the Board, Gray (2009) concluded the world of science had changed:

> *The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration (Gray, 2009).*

Each paradigm has limitations. Experiments can be slow and difficult to do at scale or even at all. Moreover, large-scale instruments, such as the Large Hadron Collider (LHC) or the Square Kilometre Array radio telescope, are expensive to build and maintain. In addition, the output of each experiment is usually analysed separately, within its own silo. This limits the potential for new knowledge to analysis of the output and input parameters of individual experiments.[1]

Mathematical models can also have limitations, including the need to simplify assumptions to create the models in the first place. In addition, scientists are often unable to solve the resulting set of complex equations to produce easily explored analytical solutions. Computer simulations can be used to address both of these limitations to a certain extent. Simulating mathematical models for a wide range of different research areas – such as climate science, molecular dynamics, materials science and astrophysics – has proved successful, and supercomputers are now used routinely for such simulations. Exascale supercomputers can perform 10^{18} floating point calculations per second. However, even supercomputer simulations are limited by the mathematical models and data representations being simulated. Moreover, one simulation represents only one instance of the problem, based on a particular set of initial conditions and constraints. In addition, some simulations can take many days or weeks to complete, thus limiting the exploration of large parameter spaces.

Simulation of climate models, which needs to be urgently improved, illustrates such limitations. The US National Center for Atmospheric Research (NCAR) collaborates in a new National Science Foundation (NSF) multidisciplinary Center for Learning the Earth with Artificial Intelligence and Physics (LEAP). LEAP will use machine learning (ML) technologies to improve NCAR's Community Earth Systems Model (CESM). The CESM comprises a complex collection of component models that can simulate the interaction

of atmosphere, ocean, land, sea ice and ice sheet processes. However, CESM is limited in its ability to incorporate an accurate mathematical representation of some important physical processes that are difficult to simulate. These include the formation and evolution of clouds at such a fine scale that the model cannot resolve them, and processes to represent land ecology that are too complicated to capture in a simulation. Climate scientists have created simplified subcomponents – known as parameterisations – to approximate these physical processes into CESM. As one of its major goals, LEAP aims to improve these approximations by using ML technologies to incorporate learning from large amounts of Earth system observational data and high-resolution model simulation data.

AI for science and engineering

AI for Science and Engineering applies AI and DL technologies to the huge scientific datasets generated by both supercomputer simulations and modern experimental facilities. Huge quantities of experimental data now come from many sources – from satellites, gene sequencers, powerful telescopes, X-ray synchrotrons, neutron sources and electron microscopes. They are also generated from major international facilities such as the LHC at the European Organization for Nuclear Research (CERN) in Geneva and the European X-ray Free-Electron Laser facility in Hamburg. These facilities already generate many petabytes of data per year and their planned upgrades will create at least an order of magnitude more data. Extracting meaningful scientific insights from these ever-increasing volumes of data will be a major challenge for scientists.

Many initiatives around the globe are now applying AI technologies to manage and analyse the ever-larger and more complex scientific datasets. Commercial tools and technologies for ML provide scientists with a good starting point. However, their application to the wide range of scientific problems requires multidisciplinary collaborative teams, including both computer scientists and physical scientists. In the United States, for example, the National Science Foundation recently established 18 National AI Research Institutes with research partnerships covering 40 states (NSF, 2022). The US DOE funds both the associated large-scale experimental facilities and the supercomputers at the National Laboratories. In 2019, the DOE Laboratories organised a series of Town Hall meetings to examine opportunities and practical next steps for AI to accelerate research in fields under the domain of the DOE's Office of Science (DOE, 2020). These meetings were attended by hundreds of scientists, computer scientists, along with participants from industry, academia and government.

The DOE Town Hall meetings used the term "AI for Science" to broadly represent the next generation of methods and scientific opportunities in computing and data analysis. This included the development and application of AI methods – a combination of ML, DL, statistical methods, data analytics and automated control – to build models from data and to use these models alone or with simulation data to advance scientific research. In line with ideas expressed in many of the contributions to the current publication, the meetings concluded that AI could transform many areas of scientific research over the next decade. It envisioned that AI technologies can:

- accelerate the design, discovery and evaluation of new materials
- advance development of new hardware and software systems, instruments and simulation data streams
- identify new science and theories revealed in high-bandwidth instrument data streams
- improve experiments by inserting inference capabilities in control and analysis loops
- enable the design, evaluation, autonomous operation and optimisation of complex systems from light sources and accelerators to instrumented detectors and high-performance computing (HPC) data centres
- advance development of autonomous laboratories and scientific workflows

- dramatically increase the capabilities of exascale and future supercomputers by capitalising on AI surrogate models (i.e. models that that mimic the behaviour of the simulation models as closely as possible while being computationally much cheaper to evaluate)
- automate the large-scale creation of findable, accessible, interoperable and re-usable (FAIR) data.

The Alan Turing Institute, the national institute for data science and AI in the United Kingdom, reached a similar conclusion. Its "AI for Science and Government" initiative includes a major research effort on AI for science in collaboration with the Scientific Machine Learning Group at the Rutherford Appleton Laboratory, the UK's National Laboratory at Harwell, near Oxford (STFC, 2022).

Thoughts on directions for research (and research policy)

Google DeepMind has applied DL techniques to make significant progress in three different fields of science – protein folding, materials modelling (Kirkpatrick et al., 2021) and fusion plasma control (Degrave et al., 2022). Researchers at DeepMind assembled multidisciplinary teams of experts and used the power of Google's Cloud computing resources for training their DL solutions to make these breakthroughs.

Can academic researchers compete with such efforts? Two actions are needed to address this question:

- A broad multidisciplinary programme is needed to allow scientists, engineers and industry to collaborate with computer scientists, applied mathematicians and statisticians to solve their challenges using a range of AI and ML technologies. This needs coherent and dedicated government funding with processes that encourage such collaboration rather than continuing with stove-piped funding allocated to individual disciplines.
- Such a programme should create a shared cloud infrastructure that allows researchers to access the competitive computing resources and tools that fuel AI R&D. In the United States, the NSF and the White House Office of Science and Technology Policy are creating a roadmap to establish a National AI Research Resource (NSF and OSTP, 2022). This is intended to be a shared research infrastructure that will provide AI researchers with significantly expanded access to computational resources and high-quality data.

The DOE work described here also detailed a rich set of topics on which research breakthroughs are needed to broaden and deepen AI's uses in science and engineering. These topics could become targets of public R&D support. In particular, as DOE (2020) describes, participants highlighted the need to:

- Incorporate domain knowledge into AI methods to improve the quality and interpretability of the models. There is a need to go beyond current models driven only by data or simple algorithms, laws and constraints. Especially key would be ML techniques driven by theory and data that could better represent the underlying dynamics specific to particular phenomena.
- Automate the large-scale creation of FAIR data. AI in science requires large datasets from a diverse range of sources – from experimental facilities and computational models to environmental sensors and satellite data streams. Adding some semantic information in the form of machine-actionable metadata could allow AI technologies to automate the creation of FAIR data. This would provide the basis for new data infrastructures to allow more interoperability and re-use.
- Advance foundational topics in the science of AI itself, with a view to developing:
  - frameworks to establish that a given problem is solvable by AI/ML methods
  - frameworks and tools to establish the validity and robustness of AI techniques, indicating the limits of AI techniques, the quantification of uncertainties and the conditions (assumptions and circumstances) that give assurance of AI predictions and decisions
  - frameworks and tools to establish which AI techniques best address different sampling scenarios and enable efficient AI on different computing and sensing environments

  - techniques that help explain the behaviour of the AI model methods
  - AI models that identify causal variables and distinguish between cause and effect
  - methods for AI models to be used to identify causal variables and distinguish between cause and effect.
- Develop new hardware and software environments. Much new AI hardware is being developed in industry for data centres, autonomous driving systems and gaming, among others. Opportunities exist for the research community to work with industry to co-design heterogeneous compute systems that use the new architectures and tools. Software is also needed to enable AI capabilities to seamlessly integrate with large-scale HPC models (see the essay in this volume by Georgia Tourassi, Mallikarjun Shankar and Feiyi Wang) and to generate and operate new scientific workflows. Such AI-enabled workflows will incorporate expert knowledge to accomplish tasks, adapt to new data and results, and refine models on the basis of cost (e.g. in energy use or run-time).

Conclusion

Greatly increased data volumes are expected for the next generation of scientific experiments. This is true for the National Laboratories in the United States with their large-scale experimental facilities, for projects such as the Square Kilometre Array Radio Telescope observatory (SKA, 2022), and for the CERN LHC upgrade, among many others. AI will be needed to automate the data collection pipelines and advance the analysis phase of such experiments. For all of these reasons, major multidisciplinary programmes on AI for science and engineering should be a high priority for public R&D investment. They can greatly increase the rate of scientific discovery and catalyse new commercial developments.

References

DeepMind (30 November 2020), “AlphaFold: A solution to a 50-year-old grand challenge in biology”, DeepMind blog, www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

Degrave, J. et al. (2022), “Magnetic control of tokamak plasmas through deep reinforcement learning”, *Nature*, Vol. 602, pp. 414-419, https://doi.org/10.1038/s41586-021-04301-9.

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.

Gray, J. (2009), “Presentation at the NRC-CSTB in Mountain View, CA, 11 January 2007”, in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Hey, T., S. Tansley and K. Tolle (eds.), Microsoft Research, Redmond.

He, K. et al. (2016), “Deep residual learning for image recognition”, in *2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, https://doi.org/10.1109/CVPR.2016.90.

Jumper, J. et al. (2021), “Highly accurate protein structure prediction with AlphaFold”, *Nature*, Vol. 596, pp. 583-589, https://doi.org/10.1038/s41586-021-03819-2.

Kirkpatrick, J. et al. (2021), “Pushing the frontiers of density functionals by solving the fractional electron problem”, *Science*, Vol. 374, pp. 1385-1389, https://doi.org/10.1126/science.abj6511.

Krizhevsky, A. et al. (2012), “ImageNet classification with deep convolutional neural networks”, *Advances in Neural Information Processing Systems*, pp. 1097-1105, https://doi.org/10.1145/3065386.

NSF (2022), “NSF-Led National AI Research Institutes”, webpage, www.nsf.gov/news/ai/AI_map_interactive.pdf (accessed 25 November 2022).

NSF and OSTP (2022), National AI Research Resource Task Force website, www.ai.gov/nairrtf/ (accessed 25 November 2022).

Senior, A.W. et al. (2020), “Improved protein structure prediction using potentials from deep learning”, *Nature*, Vol. 577, pp. 706-710, https://doi.org/10.1038/s41586-019-1923-7.

SKA (2022), "The Ska Project”, webpage, www.skatelescope.org/the-ska-project/ (accessed 25 November 2022).

STFC (2022), “Scientific Machine Learning”, webpage, www.scd.stfc.ac.uk/Pages/Scientific-Machine-Learning.aspx (accessed 25 November 2022).

Note

[1] However, AI and deep learning technologies can be applied not just to a single instance of an experiment but also to the analysis of the combined total of information from many such experiments. This can help generate new scientific discoveries and insights.

The importance of knowledge bases for artificial intelligence in science

K. Forbus, Northwestern University, United States

Introduction

For artificial intelligence (AI) systems to increase the productivity of science, they need to understand both the domains of science they are operating in, and the world in which that domain is embedded. In other words, they need knowledge bases that provide such information in explicit and verifiable forms to support reasoning that includes transparent explanations for their conclusions. This essay explains the idea of knowledge bases and knowledge graphs, summarising the state of the art and the improvements needed to support broader uses of AI in science. These improvements include commonsense knowledge to tie scientific concepts to the everyday world and to provide common ground for communication with human partners; expressive representations for encoding scientific knowledge; and robust reasoning techniques that go beyond simple retrieval. Research could work towards an open knowledge network to provide a community resource that supports re-use, replication and dissemination.

Knowledge is a hallmark of human intelligence, and a key goal of science is to generate replicable knowledge. AI systems with enough shared knowledge to reason with, and learn from, human partners could lead to revolutionary advances in science (Gil et al., 2018; Kitano, 2021). In AI, the term "knowledge base" is commonly used to refer to a system's knowledge.[1]

As this section explains, there are multiple kinds of knowledge. For some types, the commercial world has already deployed knowledge bases with billions of facts to support web search and simple forms of question answering. However, for several other kinds of knowledge – including some relevant for using AI to accelerate science – progress has been slow, despite the potential value. A US report arguing for the construction of an open knowledge network has this conclusion:

Artificial intelligence, machine learning, natural language technologies, and robotics are all driving innovation in information systems. Developing the knowledge bases, graphs, and networks that lie at the heart of these systems is expensive and tends to be domain-specific, and the largest currently are focused on consumer products (e.g. for web search, advertising placement, and question answering). An open and broad community effort to develop a national-scale data infrastructure – an Open Knowledge Network – would distribute the development expense, be accessible to a broad group of stakeholders, and be domain-agnostic. This infrastructure has the potential to drive innovation across medicine, science, engineering, and finance, and achieve a new round of explosive scientific and economic growth not seen since the adoption of the Internet. (NTSC, 2018)

This essay explains knowledge bases and knowledge graphs, what needs to be done and how it might be accomplished.

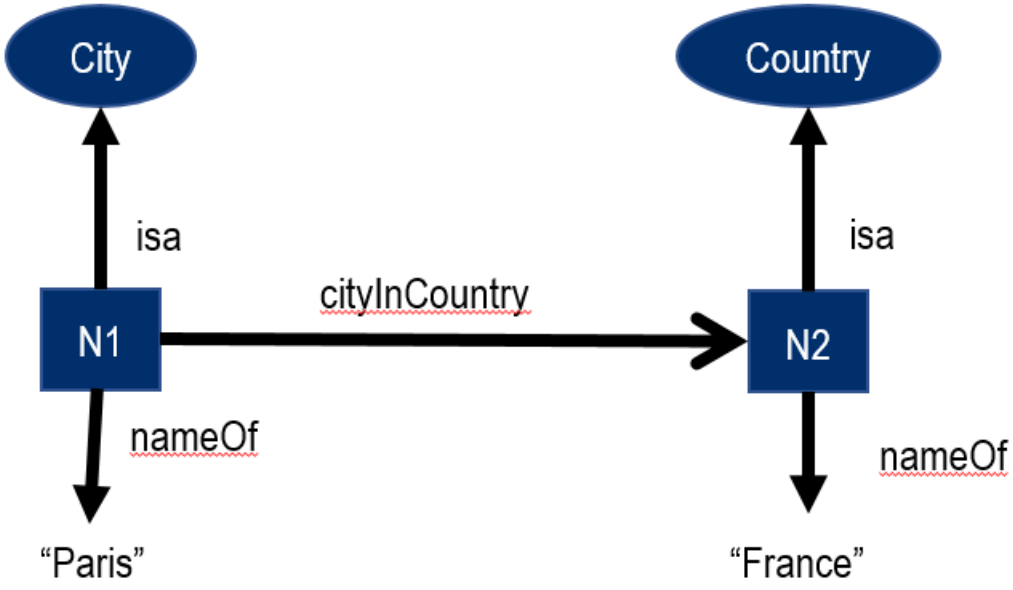
Knowledge bases and knowledge graphs

Knowledge graphs – symbolic structures that express properties of entities and relationships among them – are the most common form of knowledge bases. Entities are represented by nodes in the graph, while labelled arcs specify their properties and relationships. For example, the sentence "Paris is a city in the country of France" might be represented inside a knowledge base via the following facts:

isa(N1, City)
isa(N2, Country)
nameOf(N1,"Paris")
nameOf(N2, "France")
cityInCountry(N1,N2)

Each of these statements is a logical form expressing that a relationship (the predicate) holds between its arguments. For example, in the first statement, the predicate "isa" means that the entity provided as the first argument (e.g. N1, where "N" refers to a node) is an instance of the concept provided by the second argument (e.g. City). Similarly, the predicate "nameOf" indicates that the string given as the second argument should be used as the name for the entity given as the first argument. Meanwhile, "cityInCountry" indicates the first argument is a city geographically located within the country given as the second argument. An equivalent graphical representation of these facts is given in Figure 1:

Figure 1. Graphical representation of a knowledge base



Such symbolic representations are crucial in knowledge bases for three reasons. First, it is important to be able to vet a system's knowledge. Symbolic representations can be read by people, given the appropriate tools, as suggested by the example above. Second, such representations support reasoning. Third, they support transparent explanations, i.e. explanations that show how a system came to its conclusions.

While some researchers have tried to use "large language models" as knowledge bases (e.g. Petroni et al., 2019), such efforts have not been promising to date. Large language models fail to provide reliable ways to vet what the system knows, they do not provide accurate reasoning (Marcus and Davis, 2020), and they cannot provide transparency in their operations. Moreover, large language models have serious problems

with fairness and equity since the biases of the materials they are trained upon shine through in their application (Bender et al., 2021; Lin, Hilton and Evans, 2021). Consequently, the rest of the essay focuses on knowledge graphs.

Knowledge graphs today

Commercial knowledge graphs used in web searches (like Microsoft's Satori and Google's Knowledge Graph) contain billions of nodes, each woven into a network by even more links connecting them (Noy et al., 2019). Similarly, Amazon and other companies use knowledge graphs to represent product information to make better recommendations to customers. These large-scale knowledge graphs are constructed by a combination of manual labour (including both highly trained professionals and crowdsourcing) and automatic techniques. A great deal of engineering goes into efficient large-scale retrieval and inference using such knowledge graphs.

Most commercial knowledge graphs focus on encoding specifics about entities in the world (such as product and customer information). Similarly, some knowledge graphs have been built by extracting knowledge from textual resources, such as Wikipedia. While these knowledge graphs are useful for some kinds of factual question answering, they lack several kinds of knowledge needed to build knowledge graphs for scientific research. For example, inferential knowledge, i.e. the kinds of rules used to infer things, is missing.

Some knowledge bases have such rules as well. For example, the Cyc knowledge base (e.g. Lenat et al., 2010) is designed to support a wider variety of reasoning, via rules and more complex types of statements. For example, when asked if the planet Earth can run a marathon, it can deduce that the Earth cannot because this requires being a living thing, and the Earth, as a planet, isn't a living thing. This example illustrates an insight that Cyc researchers came to long ago: a surprising amount of commonsense knowledge isn't in explicit resources like encyclopedias. Such tacit knowledge, discussed below, also lies in the things one must know to be able to read an encyclopedia.

There have been several efforts to build knowledge bases for particular areas of science. These have been driven by the need to improve literature searches and to archive community information and expertise (e.g. genomic data, workflows). Like commercial knowledge graphs, these efforts have been facilitated by the widespread use of Semantic Web protocols. These enable the same graphs to be used with different software implementation platforms. While use of AI in science is promising, research efforts lack the breadth of expressiveness and support for inference that will be needed to fully realise this potential.

What is missing?

The impressive scale and utility of commercial knowledge graphs suggest that large-scale knowledge graphs for science are possible. However, additional research is needed in at least three areas: commonsense knowledge, professional knowledge and complex reasoning at scale. Each is discussed in turn.

Commonsense knowledge

Why commonsense knowledge? Scientific theories rest on tacit knowledge shared by all scientists due to their experience as people in the physical, social and mental worlds. Examples include the billiard ball model of gases and the lava lamp model of convection. To understand their human partners, AI systems for science need to share this common ground to some reasonable degree. Some of this tacit knowledge can be captured by qualitative representations, which provide human-like descriptions of quantities, space, causality and processes (Forbus, 2019). Other aspects require multimodal grounding, e.g. what particular

objects and systems look like. The role of experiential knowledge in commonsense remains an open research question: is most commonsense knowledge encoded via general rules, or do we mostly reason by analogy from experience?

Professional knowledge

Professional knowledge also raises important challenges. Highly expressive representations are needed to encode scientific theories. For instance, in addition to the theories themselves, how they are operationalised (by connecting the professional concepts to the everyday world) must be represented as well. In other words, knowledge must be represented to support the process of model formulation (Forbus, 2019).

Today's AI systems for science and engineering factor out model formulation by focusing on tasks and/or domains. The broader the scientific reasoning, the more tacit knowledge needs to be incorporated into the system. Representation schemes need to support explicit contexts, e.g. to represent competing theories and reason about the range of applicability for a theory. For example, when does one need to use classical, relativistic mechanics, versus quantum mechanics?

Complex reasoning at scale

No AI system comes close to the flexibility of human reasoning. Consider, for example, constructing or even just understanding thought experiments. When Einstein imagined travelling on a beam of light, it required reasoning through novel (and sometimes impossible) situations. AI systems cannot yet do this in a general way.

In specific domains, such as software verification, reasoning systems can far outstrip human capabilities. However, such systems cannot be taught to operate in a new domain without reprogramming. For specific scientific tasks and domains, special-purpose high-performance reasoning systems could likely provide important benefits. In the longer term, improved AI reasoning flexibility could bring it closer to that of people. This will enable AI to move towards being a collaborator in science, as well as a tool.

Towards knowledge bases for science: What might be done?

The commercial world is unlikely to build broad commonsense knowledge bases. Firms mostly do not care about scientific knowledge. While a large-scale high-quality commonsense knowledge graph would benefit everyone, the effort needed to build one is beyond the usual research horizons of the private sector.[2]

Partly for this reason, the idea of an open knowledge network is gaining traction in the research community. Open licensing, such as Creative Commons Attribution Only, matters. For example, the Suggested Upper Merged Ontology – intended as a foundation for a variety of computer information processing systems – provides some of the most abstract layers of an ontology (SUMO, 2022). However, it uses a licence antithetical to most commercial applications. Moreover, modules that extend it include a hodgepodge of licences that make it difficult to build on top of.[3] To maximise utility to the scientific community, in terms of reusability, replicability and dissemination, funding is needed for the construction of open knowledge graphs.

Aiming for generality is hard, and it is tempting to go straight to applications. To date, for example, the US National Science Foundation has only funded a handful of projects in open knowledge networks. All of them focused on specific domains and tasks (e.g. finding relevant legal documents, reasoning about flooding) (NSF, 2022).

A mixed approach might yield even better results. Teams of AI scientists and scientists from other domains might collaborate on knowledge graphs for fields that include both professional knowledge and relevant

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 (including simulated worlds) for some kinds of commonsense knowledge and story understanding for others.

Each effort would be required to draw upon the growing shared knowledge graph. There will be competing and complementary approaches to concepts, but that is fine if the underlying graph infrastructure supports multiple contexts.[4] The result will be a federation of knowledge graphs, ideally continually updated as research progresses and eventually encompassing all scientific knowledge.

Cognitive science and social sciences should be included, along with physical and biological sciences, in building an open knowledge network. This is important because progress in these areas will help guide AI towards more trustable, human-like algorithms and conceptual structures.

Conclusion

Progress in using AI to accelerate science will need to include efforts to build large-scale knowledge graphs of commonsense knowledge, professional knowledge and the bridges between them. This will not be done by the commercial world because it is not directly related to their everyday concerns. However, they too would benefit from aspects of it. Building an open knowledge network can improve the re-use, replicability and dissemination of scientific knowledge. This will require a long-term, large-scale community effort. However, with the right mix of projects, incremental results of value along the way will also serve to guide further efforts.

References

Bender, E. et al. (2021), “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610-623, https://doi.org/10.1145/3442188.3445922.

Carroll, J.J. et al. (2005), “Named graphs”, *Journal of Web Semantics*, Vol. 3/4, pp. 247-267, https://doi.org/10.1016/j.websem.2005.09.001.

Forbus, K. (2019), *Qualitative Representations: How People Reason and Learn about the Continuous World,* MIT Press, Cambridge, MA.

Gil, Y. et al. (2018), “Intelligent systems for geosciences: An essential research agenda”, Communications *of the ACM,* Vol. 62/1, pp. 76-84, https://doi.org/10.1145/3192335.

Kitano, H. (2021), “Nobel Turing challenge: Creating the engine for scientific discovery”, *Nature Systems Biology and Applications*, Vol. 7/29, https://doi.org/10.1038/s41540-021-00189-3.

Lenat, D. et al. (2010), “Harnessing cyc to answer clinical researchers’ ad hoc queries”, *AI Magazine*, Vol. 31/3, pp. 13-32, http://dx.doi.org/10.1609/aimag.v31i3.2299.

Lin, S., J. Hilton and O. Evans (2021), “TruthfulQA: Measuring how models mimic human falsehoods”, *arXiv*, arXiv:2109.07958v1, https://doi.org/10.48550/arXiv.2109.07958.

Marcus, G. and E. Davis (2020), “GPT-3, Bloviator: OpenAI’s language generator has no idea what it is talking about”, 22 August, *MIT Technology Review*, https://www.technologyreview.com.

Noy, N. et al. (2019), “Industry-scale knowledge graphs: Lessons and challenges”, *ACM Queue*, Vol. 12/2, https://queue.acm.org/detail.cfm?id=3332266.

NSF (2022), “Convergence Accelerator Portfolio”, webpage, https://beta.nsf.gov/funding/initiatives/convergence-accelerator/portfolio (accessed 28 January 2022).

NSTC (2018), Open Knowledge Network: Summary of the Big Data IWG Workshop, October 4-5, 2017, National Science and Technology Council, Washington, DC.

Petroni, F. et al. (2019), "Language models as knowledge bases?", *arXiv,* arXiv:1909.01066v2, https://doi.org/10.48550/arXiv.1909.01066.

SUMO (2022), Suggested Upper Merged Ontology website, https://www.ontologyportal.org (accessed 23 November 2022).

Notes

[1] This is quite different from a common informal use of the term knowledge base to mean a collection of documents for some purpose.

[2] Cycorp, founded to continue building the Cyc knowledge base, is an exception. It is funded by a mixture of commercial and government projects. Unfortunately, its proprietary nature makes it less useful for some purposes, e.g. researchers cannot freely distribute it along with their source code to ensure the replicability of their research.

[3] Personal communication. By contrast, the OpenCyc ontology, also by Cycorp, is available under a CC-Attribution Only licence. Thus, others can build up on it (e.g. www.qrg.northwestern.edu/nextkb/index.html).

[4] Any large-scale knowledge base needs to use contexts in any case, to handle culture-specific knowledge, works of fiction, and alternate explanations and theories. The OpenCyc ontology uses "microtheories" for this purpose, and a similar mechanism is found in the Resource Description Framework (RDF) semantic web representation in the form of "named graphs" (Carroll et al., 2005).

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

G. Tourassi, Oak Ridge National Laboratory, United States

M. Shankar, Oak Ridge National Laboratory, United States

F. Wang, Oak Ridge National Laboratory, United States

Introduction

The past three decades have witnessed the widespread adoption of high-performance computing (HPC) as an essential tool in the advancement of science. From accelerating critical research in the wake of a global pandemic to climate modelling and national security, HPC has become integral to the most cutting-edge scientific research around the world and across application domains. The global competition to debut the next fastest supercomputer keeps pushing the field forward. Meanwhile, increasing capabilities are bringing hallmarks of science fiction, such as artificial intelligence (AI), into daily life. However, the tremendous power of new computing systems comes with an increased concern about equitable access to these resources and their impact on supporting a thriving workforce.

The US Department of Energy advanced scientific computing research facilities

The mission of the Department of Energy (DOE) Office of Science (SC) is "to deliver scientific discoveries and major scientific tools to transform our understanding of nature and advance the energy, economic and national security of the United States." From its beginnings, the SC budget supported over 25 000 researchers at more than 300 institutions and across 17 DOE national laboratories, in addition to 27 open-access experimental and computing user facilities (DOE, 2022).

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 computing systems. These high-end computing systems are among the most advanced in the world. They are operated and available for use by researchers in industry, institutions of higher education, national laboratories and other federal agencies.

As one of DOE's leadership computing facilities, the Oak Ridge Leadership Computing Facility (OLCF) (OLFC, 2022) has consistently operated some of the nation's top supercomputers. OLCF has cemented

computation as the third pillar of scientific discovery by enabling breakthroughs in basic and applied sciences. This includes many notable contributions in energy efficiency, climate change and medical research. OLCF has provided cutting-edge supercomputing capability and pioneering ideas, operating a Top 10 supercomputer on the Top 500 list every year since its establishment in 2005. The leading OLCF systems in the last decade – Titan and Summit – debuted as the world's fastest computers.

At 200 petaflops, the IBM Summit supercomputer is the OLCF's flagship system. Launched in 2018, Summit delivers eight times the computational performance of the OLCF's previous Cray XK7 Titan supercomputer. To achieve this performance, it uses only 4 608 nodes, a fraction of Titan's 18 688 nodes. With the debut of the new HPE Cray EX Frontier supercomputer, OLCF will house the nation's first exascale system. It can perform more than 1.5 exaflops and solve calculations up to 50 times faster than today's fastest supercomputers.

As a Leadership Computing Facility, OLCF aims to provide world-class computational resources and specialised services to researchers around the world for the most computationally intensive global challenges in science and engineering. Time on DOE's Leadership Computing Systems is managed through the two competitive allocation programmes: INCITE (Innovative and Novel Computational Impact on Theory and Experiment) (ALCF, 2022) and ALCC (ASCR Leadership Computing Challenge) (ALCC, 2022). The requests typically exceed the available resources by a factor of three to five times. Therefore, selection is competitive and based on a peer-reviewed process. Allocations of computing cycles are typically 100 times greater than routinely available for university, laboratory, and industrial scientific and engineering environments.

With the rapid explosion of data-intensive science, OLCF has experienced increasing demand to support advanced scientific workflows; incorporate AI; and run rich analytics coupled with simulation software to derive valuable insights from extreme-scale experimental and observational data.

The AI compute ecosystem: Gaps and opportunities

Both the National Strategic Computing Initiative (White House, 2015) and the American AI Initiative (Parker, 11 June 2020) called for a cohesive, multi-agency, strategic vision to empower and maintain scientific leadership in the United States, as well as fuel innovations in all sectors of its economy. Since then, the landscape of HPC and AI has been changing rapidly. For example, in addition to hundreds of millions of US dollars already invested by DOE in HPC and leadership computing, more than a dozen AI institutes have been established in more than 40 states. Each has unique strengths and focus areas spanning all aspects related to AI – from fundamental methods to human-AI interaction, augmentation and collaboration.

In the European Union, with strong political endorsement, member nations have taken a concerted approach towards AI, emphasising use of AI for good and for all (European Commission, 2018). In addition to research and development (R&D) excellence, they are paying particular attention to trustworthy AI in its recent proposal for AI regulations (European Commission, 2021). Furthermore, in combination with the European Processor Initiative project, the European Union has built its own roadmap to support the convergence of extreme-scale computing, big data and AI (Kovač et al., 2022).

The infusion of enormous capital is enabling and driving R&D innovations, unleashing resources and removing barriers and training a new generation of an AI-ready workforce. However, it is becoming apparent that both AI resources and talents are highly concentrated. This could put disadvantaged groups at risk, especially in developing countries and resource-strapped universities.

The jury is still out as to whether AI is a transformational force for developing nations or a disruptive force that widens the gap between rich and poor countries (Alonso, Kothari and Rehman, 2 December 2020). One thing is clear though: there is an insatiable demand for the compute power and data storage. These

are increasingly intertwined and becoming an integral part of making ground-breaking scientific discoveries.

As part of their business strategy, cloud vendors such as Google Colab and Microsoft Azure both offer free allocations of computing resources. This service partially enables AI access to go from nothing to something. However, these offerings present notable limitations. For example, to maintain maximal resource schedule flexibility, Colab resources are not guaranteed and not unlimited. Even with a paid platform such as Colab Pro, access to the graphics processing unit (GPU) – a workhorse for AI-supporting computations – may be limited to a relatively older generation GPU and 24-hour running time. Although this is common practice, such policies are limiting for even moderate scientific and technical R&D. These limitations highlight key gaps but also opportunities for technology and policy advances.

The AI compute ecosystem: Technology and policy directions

Interest in AI and the economic potential of incorporating AI tools and methods in the commercial sector has led major corporations to develop software and purpose-built hardware for AI. Tools such as TensorFlow (originating in Google) and PyTorch (originating in Facebook) have been distributed into the open-source community. This, in turn, has led to accelerated growth in the use and adoption of AI methods in a variety of industries, academic settings and major science laboratories.

Dissemination of these tools has been accompanied by a dramatic growth in technical publications in the field of AI. It has also led to a vast array of educational materials available on line all around the world. This adoption and growth, however, is constrained by the availability of computing resources and high-quality datasets that are the basis for AI.

There are two main areas where systematic approaches led by nations at the forefront of this field can help in alleviating computing and data availability constraints: the technology and policy spheres.

Technology sphere

In the technology sphere, computing infrastructure and software availability could be stewarded and shepherded so they 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. The ways in which applications must be scaled up – crucial to serious AI campaigns – cannot be the purview of the few major commercial entities.

Nationally funded laboratories and their computing infrastructures, in collaboration with industry and academia, could nurture and support the AI ecosystems for tertiary educational entities and partner countries. This is especially useful for those entitles and countries that may lack resources or 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 would then advance to larger cloud resources or institutional-scale resources, and then on to national-scale resources. These tools and capabilities, if shared with the broader community, will enable a broader community of countries to gain from national investments.

Policy sphere

The policy sphere is associated with sharing resources, training, outcomes and guidelines. Countries at the forefront of the field, including the United States and EU leaders, 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 can provide a stepping-stone for nascent and growing initiatives. At the same time, it can also prevent reinvention and provide secondary benefits such as workforce development and rapid knowledge dissemination. The field itself will benefit from common offerings enabling reproducibility, ethical use and environment-conscious AI deployments.

Conclusion

AI has emerged as a central enabler to many existing and emerging scientific efforts. Furthermore, its rapid adoption has shown great promise across a wide array of domains – from health care and transportation to manufacturing and cybersecurity. Since their inception, the Leadership Computing Facilities have served as a strategic reserve to support open science. During the recent COVID-19 pandemic, DOE computing facilities played a central role in advancing the biomedical foundations needed for an accelerated response. The systems supported computationally intensive activities, including large AI-driven scientific campaigns (HPC Consortium, 2022). Leadership computing facilities dedicated to open science proved to be a unique asset. They leveraged their deep expertise in deploying and efficiently managing computing resources. At the same time, they built interdisciplinary teams to address some of the most critical data and computing problems associated with emerging scientific needs.

In the ever-expanding computing ecosystem, HPC will remain a critical building block. This is especially true for large-scale scientific campaigns that depend on interleaving large-scale modelling and simulation with AI. Still, since AI is a data-hungry endeavour, access to high-quality data will be as critical as access to compute resources. New capabilities and policies are needed to integrate leadership-class computing systems into distributed data ecosystems. This process will help accelerate scientific advances and ensure equity and democratisation of the resources.

References

ALCC (2022), “ASCR Leadership Computing Challenge”, webpage, https://science.osti.gov/ascr/Facilities/Accessing-ASCR-Facilities/ALCC (accessed 23 November 2022).

ALCF (2022), “INCITE Program”, webpage, www.alcf.anl.gov/science/incite-allocation-program (accessed 23 November 2022).

Alonso, C., S. Kothari and S. Rehman (2 December 2020), “How artificial intelligence could widen the gap between rich and poor nations”, IMF blog, https://blogs.imf.org/2020/12/02/how-artificial-intelligence-could-widen-the-gap-between-rich-and-poor-nations.

DOE (2022), “National Laboratories”, webpage, www.energy.gov/national-laboratories (accessed 23 November 2022).

European Commission (2021), “Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts”, 24 April, SEC(2021) 167 final, SWD(2021) 84 final, SWD(2021) 85 final, European Commission, Brussels, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.

European Commission (2018), “Artificial Intelligence for Europe”, Communication from the Commission, Brussels, 25 April, SWD(2018) 137 final, European Commission, Brussels, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0237&from=EN.

HPC Consortium (2022), “Who We Are”, webpage, https://covid19-hpc-consortium.org (accessed 23 November 2022).

Kovač, M. et al. (2022), “European processor initiative: Europe's approach to exascale computing”, in *HPC, Big Data, and AI Convergence Towards Exascale*, CRC Press, Boca Raton, FL.

OLCF (2022), Oak Ridge National Laboratory website, www.olcf.ornl.gov (accessed 23 November 2022).

Parker, L. (11 June 2020), “The American AI Initiative: The U.S. Strategy for leadership in artificial intelligence”, OECD.AI Policy Observatory blog.

White House (2015), “Executive Order – ‘Creating a National Strategic Computing Initiative’”, 29 July, Press Release, White House, Washington, DC, https://obamawhitehouse.archives.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative.

Improving reproducibility of artificial intelligence research to increase trust and productivity

O.E. Gundersen, Norwegian University of Science and Technology, Norway

Introduction

Several recent studies have shown that many scientific results cannot be trusted. While the "reproducibility crisis" was first recognised in psychology, the problem affects most if not all branches of science. This essay analyses the underlying issues causing research to be irreproducible – with a focus on artificial intelligence (AI) – so that mitigating policies can be formulated.

Studies presented at leading conferences and published in high-impact journals have shown that AI research has not escaped the reproducibility problem. Ioannidis (2022) suggested that 70% of AI research was irreproducible. He pointed to the immaturity of the field compared to more mature sciences such as physics. This accords with two findings of Gundersen and Kjensmo (2018): only 6% of research published at top AI conferences explicitly stated which research questions were being answered, while only 5% stated which hypotheses were tested.

Problems of reproducibility have been documented in image recognition, natural language processing, time-series forecasting, reinforcement learning, recommender systems and generative adversarial neural networks (Henderson et al., 2018; Lucic et al., 2018; Melis, Dyer and Blunsom, 2018; Bouthillier, Laurent and Vincent, 2019; Dacrema, Cremonesi and Jannach, 2019; Belz et al., 2021). Application domains of AI have not been spared: problems have been documented in medicine and social sciences.

Many investigations have sought to identify what causes these irreproducible results. A proper understanding of the concept of reproducibility is required to capture the causes of irreproducibility. Yet, although reproducibility is a cornerstone of science, it has no commonly agreed definition. Plesser (2018) even holds that "reproducibility" is a confused term.

Without an agreed definition, the crisis will not be mitigated and many irreproducible findings will be published. This will reduce trust in science, which is already in decline. Conversely, increasing the rate of published reproducible findings will increase the productivity of science, and more importantly, increase trust in it.

Criteria for a definition of reproducibility

Surprisingly, the most prevalent definitions of reproducibility are not helpful when designing, conducting and evaluating results of a reproducibility experiment. Here, the term "reproducibility experiment" refers to

an independent experiment that seeks to validate the results of a previous study, here called the "original study". The prevailing definitions neither specify what a reproducibility experiment entails nor what it means to reproduce results. Below, the essay presents several criteria for a more compelling definition of reproducibility.

Similarities and differences

A definition of reproducibility must help scientists specify the similarities and differences between the original and reproducibility experiments. It should provide insights into what independent researchers used from the original experiment to reproduce it and what they can and cannot change. More concretely, the definition should help answer several questions. Is an AI reproducibility experiment different enough from an original experiment if the same code is executed on a different computer but uses the same data? Did the reproducibility experiment in AI use different code or different data?

Degree of reproducibility

The definition of reproducibility should help inform when the results of an experiment have been reproduced and to what degree. In computer science, including AI, the output of the computational execution of a reproducibility experiment can be identical to the original experiment. This is due to the inherent determinism of some computational experiments. In contrast, producing identical results is highly unlikely in such domains as medicine, biology and psychology. In these domains, experiments involve humans and living material, and are far from deterministic.

Sorting out how conclusions are inferred

A definition should help show if experimental results are reproducible in one of three ways: either the same analysis has yielded the same conclusions from a different set of outputs; the same conclusion was drawn from a different analysis; or the reproducibility experiments produced an identical output. Today, the most prevalent definitions do not help researchers sort out such issues.

Generalisable to all disciplines

A good definition of reproducibility should also be generalisable to all scientific disciplines. This would be achieved if the definition is intimately related to a definition of science. While the scientific literature agrees that reproducibility is a cornerstone of science, few if any previous definitions make this relationship explicit.

Defining reproducibility

Reproducibility has no meaning outside the context of empirical studies. Reproducibility is different from repeatability, which simply means doing the same thing again. Consequently, a reproducibility experiment should be similar to, but different from, the original experiment. In addition to being generalisable, a definition of reproducibility should also help scientists pinpoint what was different between an original experiment and an experiment that confirms reproducibility (and why the researchers are justified in concluding that previous results have been reproduced). Over several years and in several publications, the author arrived at the following definition of reproducibility:

> *Reproducibility is the ability of independent researchers to draw the same conclusions from an experiment by relying on the documentation shared by the original researchers when following the scientific method. The documentation relied on by the independent researchers specifies the type of reproducibility study, and the way the independent researchers reached their conclusion specifies to which degree the reproducibility study validated the conclusion (Gundersen, 2021).*

The above definition of reproducibility differs from others in the literature in several ways. First, it emphasises that reproducibility requires independent researchers to redo a study. Second, it emphasises that an experiment from which conclusions are drawn must be described in some form of documentation shared with third parties. Third, it defines "documentation" and "drawing the same conclusions" concisely and in relation to the scientific method. Finally, it distinguishes between the type of reproducibility study and the degree to which such a study validated the original results.

For non-computational experiments, experimental documentation is written and shared in the form of reports. Analyses are typically done using statistical software such as SPSS or Excel or written as code in languages like R, Matlab and Python, which can also be shared with third parties. As reports only write about the experiments themselves, they often leave out details that could affect results.

Computational experiments, such as those mostly reported in research using AI and machine learning, have a clear advantage over non-computational experiments. Computational experiments, and their complete workflow, can often be fully captured and documented in code. This removes any ambiguity about which steps were performed in which sequence and which parameters and thresholds were used.

Computational experiments are still not fully described by code, even if all steps of an experiment are implemented in code. This is because they also depend on ancillary software, such as libraries, frameworks and operating systems, as well as hardware to run on. A computational experiment is not completely documented unless ancillary software, hardware and data are specified in the documentation, in addition to the code describing the experiment. A complete documentation, which can be supported through technical solutions, must include these descriptions as well. This documentation can support packaging all ancillary software so they can be shared with independent researchers and help capture the hardware used in experiments.

Documenting computational experiments

Many computational experiments rely on observations in the form of digitised data or fully digitised simulations. Images used for training machine-learning algorithms to recognise objects, such as handwritten digits, are one example of digitised data. Meanwhile, self-play in games, such as used when training AlphaZero, is an example of a digitised simulation (Schrittwieser et al., 2020).

In both cases, the experiments are fully executed on a computer and can be reproduced with relative ease. If the analyses and their interpretation also exist as code, the complete experiment is computational, and conclusions can be drawn without human intervention. Computers are still unable to formulate interesting research questions, design proper experiments, and understand and describe their limitations. However, efforts to fully automate the scientific process are underway.

In non-computational sciences, the importance of specifying the equipment used when performing the experiment is widely understood. It is the first thing chemistry and physics students learn as part of their laboratory assignments. Doing the same is equally important for computational experiments because the choice of hardware and software can introduce biases (Gundersen, Shamsaliei and Isdahl, 2022; Zhuang et al., 2022).

In non-computational experiments, results can also be influenced by who did the experiment. By contrast, computational experiments do not depend on the person executing the code if the experiment is fully automated as code. In other words, whether one person or another presses a button that executes a computational experiment is irrelevant to the outcome. The datasets, on the other hand, can be biased (Torralba and Efros, 2011). If another dataset is used without the same bias, it would render the experiment irreproducible.

The degree to which a result has been reproduced

The degree to which an experimental result has been reproduced depends on how much the reproducibility study is generalisable. If the same code is executed on the same ancillary software using the same data, but on a different computer, to produce the exact same output, the results are not highly generalisable. In this case, the reproducibility experiment has only shown that the results of the original experiment are generalisable to different computers.

This contrasts with reproducibility experiments that only rely on written documentation. In these cases, the independent researchers must write all code themselves, collect new data and execute the experiment on a different computer. If the results are the same, such a reproducibility experiment is much more generalisable, and the hypotheses can be trusted more deeply. In such a situation, the outcome of the re-implemented experiment should not be expected to be identical to that of the original experiment.

While reproducing an experiment by reimplementing all code and collecting new data makes the results more generalisable, it is more work for the independent researchers (Gundersen, 2019). Transparent research is easier to trust, as the researchers have nothing to hide.

Outcome reproducible

A reproducibility study is outcome reproducible if the reproducibility and original experiments have the same outcome, while the analysis and interpretation are the same as for the original experiment. This could be exemplified by an image classification experiment where a machine-learning algorithm must classify a set of images of cats and dogs. The reproducibility experiment is outcome reproducible if it produces the exact same classes for each image in the test set as the original experiment. For all practical purposes, non-computational studies will not be outcome reproducible. For example, there is a low probability that survey respondents will give the same answer if they redo the survey.

Analysis reproducible

An experiment is analysis reproducible if the outcome differs between the original and reproducibility experiments but uses the same analysis and leads to the same conclusion. Again, the reproducibility experiment could produce a different set of classes for the set of test images of cats and dogs. However, if the same analysis gives the same result, the reproducibility study is analysis reproducible. For example, the same result could be that a given algorithm performs significantly better than another one when relying on the same statistical test used in the original study.

Inferentially reproducible

Finally, an experiment is inferentially reproducible if a different analysis of the same or different outcome leads to the same conclusion in both the original and reproducibility studies. Again, using the example of image classification, the reproducibility study is inferentially reproducible if one statistical test is changed for another one when analysing the classes that the machine-learning algorithms have produced for the images in the test set. Even when the analysis is different, the result is reproducible if the same conclusion is reached.

Inferentially reproducible studies are more generalisable than analysis-reproducible studies, which in turn, are more generalisable than outcome-reproducible studies. Outcome reproducibility is a narrow interpretation of reproducibility. It cannot be achieved unless the reproducibility experiment uses the exact same data. A more robust conclusion and thus more generalisable result is made if the same analysis is done on an outcome produced by running the experiment on a different dataset. If the conclusion is still valid despite using a different analysis on the outcome, the result is even more general.

Still, while a better understanding of reproducibility is a good start, it alone will not mitigate the reproducibility crisis. One must understand what causes irreproducibility.

The sources of irreproducibility

Experiments have many sources of irreproducibility that could invalidate the conclusions drawn from an experiment. While some sources of irreproducibility are the same between sciences, others only apply to specific sciences. This essay is most interested in those that affect AI and machine learning.

Figure 1 shows graphically the entire scientific workflow, and the forms in which, and points at which, AI is brought to bear. Broadly speaking, for such experiments, the sources of irreproducibility can be divided into six types (Gundersen, Coakley and Kirkpatrick, 2022).

Study design factors

Study design factors capture the high-level plan for how to conduct and analyse an experiment to answer the stated hypothesis and research questions. For example, baselines could be poorly chosen (comparing a state-of-the-art deep-learning algorithm for a given task to one that is not state of the art). This would provide a false proof of improving state of the art.

Algorithmic factors

Algorithmic factors are design choices of machine-learning algorithms and training processes that exploit randomness in different ways. Examples include data randomised in different ways during training, learning algorithms that rely on random initialisation of features, features that are selected randomly and algorithms that are optimised by using optimisation techniques that rely on randomness. The randomness introduced by design choices leads to differences in performance. Researchers can choose results that best suit them over those that better reflect the true performance of an algorithm. This makes the study irreproducible for researchers that do not choose to “cherry pick” results.

Implementation factors

Implementation factors are consequences of choices related to the software and hardware used in the experiment. Examples include which software (e.g. operating systems, software libraries and frameworks) are used in the computational experiment, and whether computations are executed in parallel. All of these can affect the outcome to such a degree that opposite conclusions can be drawn.

Observation factors

Observation factors are related to the data or the environment of an intelligent agent. This includes how data are generated, processed and augmented, but also properties of the environments, such as the physical laws present in the benchmarking environment of the agent. The agent learns from patterns in the data. If the data represent the physical world, which in most cases they do, the agent will bring this idea of the world with it when deployed.

These issues have been most discussed when considering sex and race. However, problems can arise from even innocuous features of a dataset, as when a system fails to recognise images of coffee mugs simply because some have handles pointing in different directions than others (Torralba and Efros, 2011). Such issues can be reduced but also emphasised during data pre-processing and data augmentation (where a dataset is increased by slightly modifying samples and adding theming them to the dataset).

Other issues relate to different distributions of classes in the training and test set. For example, the test set could have more images of dogs than the training set when the algorithm has a lower error rate on dogs. The annotation quality of target values in the training data is another consideration. Different human annotators, for example, can label the same instance differently.

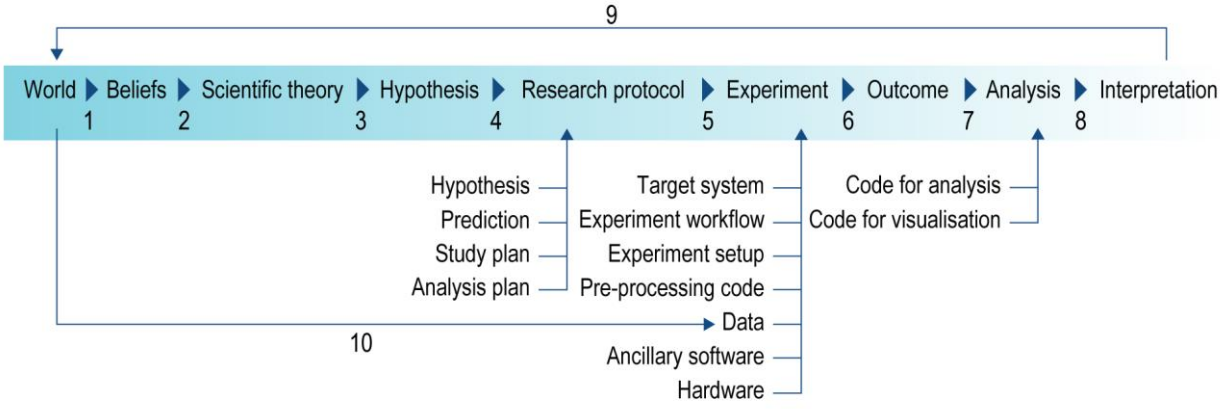
Evaluation factors

Evaluation factors relate to how investigators reach their conclusion. Examples include selective reporting of results where only datasets that show the wanted results are used in the study, over-claiming of results where conclusions go beyond the evidence, poor estimation of error and misuse of statistics when analysing results. Evaluation factors can be uncovered by reading scientific reports and thoroughly understanding the presented research and knowing the state of the art deeply.

Documentation factors

Documentation factors capture how well the documentation reflects the actual experiment. For an experiment to be documented perfectly, all choices that are made and could be a source of irreproducibility should be documented and the motivations for the choices explained. These could include up to 42 different types of choices (Gundersen, Coakley and Kirkpatrick, 2022). The readability of the documentation is of course important, as is the detail provided on the experiment's design, implementation and workflow. For computational experiments, publishing code and data will in many cases sort out the ambiguities.

Figure 1. The scientific method represented in a process diagram



Source: Gundersen and Kjensmo (2018).

Implications of sources on irreproducibility

The different sources of irreproducibility affect the conclusion in different ways (Gundersen, Coakley and Kirkpatrick, 2022). Some identified sources affect the outcome of an experiment, such as whether an algorithm uses randomness in the training process. This means that every time a model is trained on a dataset, the outcome will be slightly different when run on the same test set. Some decisions, related to how the model is evaluated, could change the analysis, leading to a different conclusion. For example, some error metrics will emphasise some characteristics of a model over others. This can be illustrated by using the mean value or the median value for comparing two sets of numbers. A couple of extreme outliers will affect the mean value but not the median value. The choice of metric can affect the conclusion. Finally,

some factors affect the inference, such as leaving out some results that do not support a researcher's desired conclusion.

Implications for the research ecosystem

False results are expected in science. Indeed, AI has a high rate of false results (Ioannidis, 2022). An achievable goal is to reduce the number of irreproducible studies to be on par with physics, which has the lowest rate of false findings. This could be done through increasing methodological rigour, such as explicitly considering the sources of irreproducibility. Individual researchers cannot do this alone. All actors in the research system must share the responsibility. Broadly, this system includes researchers, their institutions, the publishers of the research and funding agencies. The following provides some insights into the responsibilities for each actor.

Individual researchers

Individual researchers should ensure they understand and describe the limitations of their studies, taking the sources of irreproducibility into account. They must design studies that genuinely test their hypotheses and treat all algorithms that are investigated equally. They should also discuss the limitations of the experiments related to algorithmic, implementation and observation factors that can affect the conclusion. The choice of evaluation should be clearly reasoned, demonstrating clearly why it will provide trustworthy evidence for the conclusion. Finally, the researchers must document the research properly and share as much information as possible about the experiment, including code and data in addition to good descriptions.

Research institutions

Research institutions should ensure that best practices for AI research are followed. This includes training employees and providing quality assurance processes for the research under their responsibility. They should also ensure that research projects set aside enough time for quality assurance of the experiments. Finally, they should emphasise quality and transparent research practices as part of the process of hiring researchers.

Publishers

Publishers must assure the quality of the research they publish, a job often outsourced to third-party researchers. Few publishers standardise the review process and provide instructions that reviewers should follow. The peer review that occurs as part of AI and machine-learning conferences is an exception. Here, reviews are provided as checklists and forms that reviewers must answer. This contrasts with journals, where reviews are typically written in free form. Guidelines are provided but not enforced in any way. This can be improved using forms that cover different sources of irreproducibility. Furthermore, they should encourage publishing code and data as part of a scientific article. That said, publishers should not be expected to enforce the sharing of code and data.

Funding agencies

Funding agencies obviously evaluate the quality of project proposals selected for funding. While they cannot actively avert or control many sources of irreproducibility, they can significantly influence some of them.

First, funding agencies can select evaluators with a good track record of open and transparent research. As such research is easier to audit and check for reproducibility, researchers that publish open and transparent research would seemingly set a high bar for themselves.

Second, for the same reason, funding agencies can require the research they fund to 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. Governmental funding agencies in particular should require sharing of publicly funded research so it is available to the public. In January 2021, OECD countries adopted an updated *Recommendation of the Council concerning Access to Research Data from Public Funding* (OECD, 2021). This legal instrument, in force since 2006, now addresses new technologies and policy developments, and provides comprehensive policy guidance. The revision expands the scope of the earlier Recommendation to go beyond coverage of research data. It now covers related metadata, as well as bespoke algorithms, workflows, models and software (including code), which are essential for their interpretation.

Making research infrastructure available for third parties could be an option. However, it may be less important as computational experiments should be reproducible regardless of the hardware and ancillary software involved. Requiring that code and data are available publicly for third parties would allow them to run experiments on different hardware. However, this will not solve all issues with reproducibility. Producing the same outcomes does not ensure that datasets have not been chosen based on how well certain methods perform on them; other datasets could be left out for the same reason. Making code and data available for third parties will enable third parties to check the validity of published research with less effort.

Conclusion

If science involves standing on the shoulders of previous generations of scientific giants, as Newton put it, then reducing the number of false results will help scientists to see even further. This means that the productivity of science will increase. AI research needs to continue focusing on reproducibility, openness and transparency. Most high-impact research conferences care about this and have started using a reproducibility checklist as part of the submission and review process. Community-driven publications, such as the Journal of AI Research, have adopted this focus. However, funding agencies can also require researchers to share code and data as a condition of their funding. In addition, they should require funded researchers to publish in open-access journals and conferences that have clear guidelines and forms for evaluating research.

References

Belz, A. et al. (2021), “A systematic review of reproducibility research in natural language processing”, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 381-393, https://aclanthology.org/2021.eacl-main.29.

Bouthillier, X., C. Laurent and P. Vincent (2019), “Unreproducible research is reproducible”, in *Proceedings of Machine Learning Research 97*, pp. 725-734, http://proceedings.mlr.press/v97/bouthillier19a/bouthillier19a.pdf.

Dacrema, M.F, P. Cremonesi and D. Jannach (2019), “Are we really making much progress? A worrying analysis of recent neural recommendation approaches”, in *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101-109, https://doi.org/10.1145/3298689.3347058.

Gundersen, O.E. (2021), “The fundamental principles of reproducibility”, *Philosophical Transactions of the Royal Society A*, Vol. 379/2197, https://doi.org/10.1098/rsta.2020.0210.

Gundersen, O.E. (2019), "Standing on the feet of giants – Reproducibility in AI", *AI Magazine*, Vol. 40/4, pp. 9-23, https://doi.org/10.1609/aimag.v40i4.5185.

Gundersen, O.E., K. Coakley and C. Kirkpatrick (2022), "Sources of irreproducibility in machine learning: A review", *arXiv*, preprint, arXiv:2204.07610, https://doi.org/10.48550/arXiv.2204.07610.

Gundersen, O.E., S. Shamsaliei and R.J. Isdahl (2022), "Do machine learning platforms provide out-of-the-box reproducibility?", *Future Generation Computer Systems*, Vol. 126, pp. 34-47, https://doi.org/10.1016/j.future.2021.06.014.

Gundersen, O.E., Y. Gil and D.W. Aha (2018), "On reproducible AI: Towards reproducible research, open science and digital scholarship in AI publications", *AI Magazine*, Vol. 39/3, pp. 56-68, https://doi.org/10.1609/aimag.v39i3.2816.

Gundersen, O.E. and S. Kjensmo (2018), "State of the art: Reproducibility in artificial intelligence", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32/1, https://doi.org/10.1609/aaai.v32i1.11503.

Henderson, P. et al. (2018), "Deep reinforcement learning that matters", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No 1, https://dl.acm.org/doi/abs/10.5555/3504035.3504427.

Ioannidis, J.P. (2022), "Why most published research findings are false", *PLOS Medicine*, Vol. 2/8, p. e124, https://doi.org/10.1371/journal.pmed.0020124.

Lucic, M. et al. (2018), "Are GANS created equal? A large-scale study", *Advances in Neural Information Processing Systems*, Vol. 31, https://dl.acm.org/doi/10.5555/3326943.3327008.

Melis, G., C. Dyer and P. Blunsom (2018), "On the state of the art of evaluation in neural language models", in *Proceedings of the International Conference on Learning Representations 2018*, https://openreview.net/pdf?id=ByJHuTgA-.

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.

Plesser, H.E. (2018), "Reproducibility vs. replicability: A brief history of a confused terminology", *Frontiers in Neuroinformatics*, Vol. 11/76, https://doi.org/10.3389%2Ffninf.2017.00076.

Schrittwieser, J. et al. (2020), "Mastering Atari, Go, chess and shogi by planning with a learned model", *Nature*, Vol. 588/7839, pp. 604-609, https://doi.org/10.1038/s41586-020-03051-4.

Torralba, A. and A.A. Efros (2011), "Unbiased look at dataset bias", in *CVPR 2011*, pp. 1521-1528, https://doi.org/10.1109/CVPR.2011.5995347.

Zhuang, D. et al. (2022), "Randomness in neural network training: Characterizing the impact of tooling", in *Proceedings of Machine Learning and Systems,* Vol. 4, pp. 316-336, https://proceedings.mlsys.org/paper/2022/file/757b505cfd34c64c85ca5b5690ee5293-Paper.pdf.

AI and scientific productivity: Considering policy and governance challenges

K. Flanagan, The University of Manchester, United Kingdom
B. Ribeiro, Université Côte D'Azur, France
P. Ferri, The University of Manchester, United Kingdom

Introduction

Increased application of artificial intelligence (AI) is often touted as the solution to the problem of scientific productivity. This essay explores the science policy and governance implications of AI within the broader debate about scientific productivity. It reviews lessons from previous waves of automation in science and their impact on the practice of science. Since the public sector science base is also the environment in which advanced skills in science and technology are developed, the paper considers possible implications of AI use on scientific human capital. It then examines a range of policy and governance implications, including how AI tools might be used in funding and governance practices.

Scientific productivity vs. research productivity

Scientific productivity is not necessarily related to research because not all science is research. The *Frascati Manual* (OECD, 2015) defines research as "creative and systematic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture and society – and to devise new applications of available knowledge." Many scientific workers, as defined in the *Canberra Manual* (OECD, 1995), work in monitoring and testing roles, whether in the private or public sector. In addition, most research is not investigator-driven science (or "basic research" in the Frascati definition). Finally, most research is carried out in the private sector (almost three-quarters of all research and development for OECD countries). Distinguishing between public and private and applied and basic research is necessary since they have different underpinnings, dynamics and motivations.

Assuming that routine scientific work, such as testing water, air or food quality, is not relevant to the debate, research productivity could be conceived in a number of ways: the efficiency with which scientists generate outputs (and the most appropriate outputs to consider e.g. publications and research grants); the rate of important and possibly high-impact discoveries; or the generation of successful innovations, or perhaps only radical or transformational ones.

Recent debates about scientific productivity seem to switch back and forth between these different understandings of "science" and of "productivity", and between an interest in the quantity vs. quality of outputs or impacts. This is not one but several distinct debates (EC, 2022). The relationships between investigator-driven basic science and industrial (i.e. corporate) innovation are non-linear and indirect (e.g. Salter and Martin, 2001). Consequently, a change in the quality or quantity of basic science will not

necessarily drive a change in the quality or quantity of industrial innovation. In a recent expert survey, only 15% of respondents felt that research productivity had declined in the past decade, while more than half felt it had increased (EC, 2022). This lack of consensus illustrates the subjective nature of research productivity as a concept.

A recurring framing in debates about scientific productivity is that science is a process in which inputs are turned into outputs i.e. a production process. A different way to make sense of scientific productivity is to consider the policy goals behind publicly funded science, and the kind(s) of outputs and outcomes that governments are looking to realise by funding research.

The relationship between scientific productivity and the science base

The conventional wisdom that science policy is about funding research to generate knowledge that (hopefully) has positive societal and economic impacts retains a powerful hold in academic and policy circles. It is a key component of rhetoric around the public value of science and innovation (Ribeiro and Shapira, 2020). However, research funding also plays an important role in developing and maintaining the scientific workforce, as can be observed in the past century of science policy in Western countries.

The supply of "advanced scientific human capital" was a prime concern behind the emergence of science policy in the 20th century. In the United Kingdom, increasing the supply of skilled researchers in the national interest, for example, was a key aim behind the 1916 creation of the Department of Scientific and Industrial Research, the precursor of UK research councils (Clarke, 2019). In the United States, the issue found expression through President Roosevelt's letter tasking Vannevar Bush with his 1945 report *Science: The Endless Frontier* (Zachary, 1997; Dennis, 2006).

This recurring focus on the human and institutional capacity to do research, and the economic and social roles played by this capacity, is reflected in the more modern conception of the "science base". One could argue that governments fund investigator-driven science to build and sustain the research ecosystem as a key social and economic resource. If so, then the impacts of AI should be considered not just on scientific output but on the science base more broadly.

To map an agenda around these broader impacts on the science base, this paper focuses on everyday scientific labour (as opposed to an idealised, heroic view focused on moments of discovery; or proxies to assess scientific work, such as publications), on careers, and on governance issues in relation to publicly funded science. This will involve looking at previous waves of automation in science. It will consider how AI research tools might be funded in the public science base, and how that might impact on research processes and practices. It will also briefly reflect on the uses of AI in improving research governance and integrity.

The evolution of scientific research

The history and sociology of science show clearly that scientific research is not a homogenous, unchanging activity. Pickstone (2000) gives the example of how embryology moved from description through analysis to experimentation. At a more micro level, specific practices also change over time. The early embryologists observed with very different instruments and techniques from those of contemporary scientists.

Practices around the recording, communication and analysis of data have also varied from discipline to discipline and over time, as have standards of evidence. Such changes are often entangled with the adoption of new technologies. High-throughput automated sequencing technologies developed during and after the Human Genome Project created a demand for new scientific skills. They even created new disciplines such as bioinformatics (Bartlett, Lewis and Williams, 2016).

The “data deluge” in high energy physics and astronomy, but also in biomedical research, has led those scientific communities to develop new practices for managing, sharing and analysing data. In extreme cases, new practices have also emerged for verifying observations. These include, for example, multiple detectors based on different concepts and managed by different multinational teams at the Large Hadron Collider; see Junk and Lyons (2020) for a detailed discussion of replication efforts in particle physics. Funding and governance processes must often adapt to the adoption of new scientific tools. This occurred, for example, with the introduction of the genetically modified ‘knockout mouse’ in the biomedical sciences (Flanagan et al., 2003; Valli et al., 2007).

Scientific careers and scientific work

As with any other profession, scientific research careers and everyday work practices are affected by labour market dynamics, and workplace and organisational cultures. They are also affected by disciplinary cultures and national funding and evaluation practices. The content of scientific work is varied. Most researchers in the public sector science base are not full-time researchers but also play roles as teachers or managers. Research itself involves an unusual mixture of routine, often mundane, work and exploratory creative work. This can occur in conditions of high uncertainty and, often, competitive pressure, as well as sometimes extensive collaboration and sharing. Practices of collaboration, co-ordination and sometimes competition will depend on and/or be affected by automated systems.

Few scholars have studied the impact of automation on the content of scientific work. However, many studies have examined the impacts of automation on the content of other kinds of work. Studies in other domains show that automation can have a wide range of impacts on everyday work routines and interactions, depending on context. Unexpected effects may even include reduced productivity, for instance, through information overload (Azad and King, 2008).

How the adoption of new AI tools will affect science is heavily entangled with the features of scientific work. Some labour-intensive, routine and mundane practices may be replaceable by automated tools, as others have been replaced in the past. Think of how statistical analysis by hand was replaced by computers. However, the adoption of new tools can also introduce a demand for new routine and mundane tasks, which have to be incorporated into the practice of science.

In their study of scientific labour in the field of synthetic biology, Ribeiro et al. (2023) show how automation and digitalisation can lead to the amplification and diversification of tasks. New protocols and methodologies demand new skills from researchers. They also require a good deal of translational labour for interdisciplinary collaborations (e.g. computer sciences and biology).

A digitalisation paradox emerges. Robotics and advanced data analytics are aimed at simplifying scientific work by automating some repetitive tasks such as pipetting. Yet they also contribute to increasing the complexity of scientific work in terms of the number and diversity of tasks that cannot be automated (Ribeiro et al., 2023).

These tasks often involve mundane work with laboratory robots and with large volumes of data – from preparing and supervising robots to checking and standardising data. This is because automated and “intelligent” systems affect the number and types of hypothesis and scientific experiments that can be tested and performed.

Importantly, the tasks created by adoption of new AI tools are likely to be the preserve of early career researchers lower down the scientific hierarchy. This is because the application of AI tools often involves labour-intensive, time-consuming activities. Data curation, cleaning and labelling, for example, are usually performed by these researchers.

Therefore, further automation and digitalisation of scientific work – from robotics to AI models focused on discovery – might pose employment-related risks to scientific workers occupying lower positions in the

scientific hierarchy. Mundane work with data and robots has little value for promotion or job applications in scientific organisations, which value publication in prestigious journals. As argued by Ribeiro et al. (2023), the performance of scientists dealing with data and machines is intertwined with the performance of these technical systems. Should an experiment fail or be delayed due to equipment problems, the burden would mostly fall on those scientists.

The research context is also the training context

Maintaining and enhancing the human and infrastructural capacity of the science base is a key, if often implicit, aim of research policy. In other words, the research environment is also the research training environment. Graduate students and post-docs are learning by doing and learning by observing. They 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. This research training experience is a key public good motivating the public funding of research.

Because automation will change the content of scientific work in the ways outlined above, it can affect the quantity and quality of those training opportunities in the research base. If fewer research post-docs are required, or where such roles become primarily focused on work with automated systems, it could limit exposure to a wider set of scientific practices.

When automating manual or cognitive practices, there is always a risk that understanding of, and skills relating to, key procedures may be lost. Mindell (2015) notes it is critical that pilots periodically practise flying in manual mode to understand what is required should the autopilot system fail. In the scientific context, as critical techniques and processes become "black-boxed" in this way, students, as well as early career and other researchers, may not get the opportunity to fully learn or understand them.

The earlier black-boxing of statistical analysis in software packages has been argued to have contributed to the misapplication of statistical tests. This could be a contributory factor in a crisis of reproducibility in research (Nickerson, 2000). Future generations of scientists will be able to accomplish their tasks yet be unable to perform an experiment without the automated system support or fully understand the outcomes produced by algorithms. Paradoxically, the adoption of new tools could leave scientists less well equipped to understand and critique their application.

Funding and research governance implications of AI in science

As new techniques and technologies emerge in scientific research, they become fashionable, creating new demands for funding. A scientific community's shared understanding of what constitutes "leading-edge" research can change. This happened in the biomedical sciences, for example, with the increasing popularity of new technologies like high-throughput sequencing.

Researchers may change trajectories, select topics and problems amenable to apply a new tool to remain at the leading edge, publish in the most prestigious outlets and attract funding. Survey evidence suggests that the cost of meeting the performance level demanded at the leading edge of scientific research tends to grow faster than the rate at which technological innovation lowers cost (Georghiou and Halfpenny, 1996). This inevitably creates pressures on funding. These pressures have the potential to strengthen or create new structural inequalities that discriminate against less well-resourced groups or researchers in lower-income countries. Helmy, Awad and Mosa (2016), for example, examine challenges faced by developing countries in establishing themselves in genomics research. Wagner (2008) provides a more general discussion of entry barriers to the leading edge of global science.

There are also questions about how future automation in the public research base will be funded. Some commentators argue that AI tools will transform the productivity of research at little or no cost. However, AI tools have to be embedded in wider systems of data collection, curation, storage and validation. In particular, automated systems involving both AI and robotics tools are unlikely to come cheaply.

It is helpful here to consider how other items of scientific research infrastructure are funded. Small items of equipment may be funded through competitive grants. However, bigger ticket items are more likely to be funded through capital spending streams.

The cost effects of the adoption of new tools may be difficult to predict. Adoption of proven models from libraries may involve no direct costs and may reduce the direct labour costs of doing research. However, in other cases, especially at the research frontier, both capital and labour costs may rise.

Major items of research equipment typically require complementary assets. This can include refurbished or purpose-built accommodation, skilled technical and user support staff, preparation and analysis facilities. It may also require additional items of generic, supporting equipment.

There is some evidence 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. They may also lack the necessary ongoing technical support and maintenance to make that equipment productive (see Flanagan et al., 2003).

This struggle might provide a challenge to adoption of new automated tools involving AI and other forms of automation. At the very least, it could affect situations where funding is primarily competitive and research organisations lack their own private resources to complement competitively won grants. It may also be an issue for the introduction of novel, unproven AI tools. Thus, research policies need to consider not only how to fund new tools but also how to ensure support for complementary assets.

Current AI tools automate routine experimental, observational and classificational tasks in scientific research, typically within laboratories and offices in research organisations (Royal Society, 2018; Raghu and Schmidt, 2020). Future AI tools that aim to verify or even identify causal relationships might not necessarily be treated as “equipment” by funders. Instead, they could instead come to be considered like a “member” of the research team. In competitive funding systems, researchers are evaluated based on their track record of publications, an indicator of high performance. Research funders may need to consider how future AI tools that automate or partially automate the identification of causal relationships will be evaluated in the competition for funding.

AI in research governance processes

There has already been some experimentation with the application of machine-learning tools to funding body processes. These include identification of appropriate peer reviewers for grant proposals (e.g. the National Natural Sciences Foundation in China, see Cyranoski, 2019). Such tools hold the promise of speeding up the slow processes of matching reviewers with applications. They have also been lauded as a means of avoiding old boy networks or lobbying.

However, these uses of AI have also been criticised for their potential to introduce new biases into review processes. For example, they might select reviewers who have conflicts of interest or are not appropriately qualified to assess the proposal (Cyranoski, 2019).

There has also been much interest in tools to partially automate aspects of the funding or journal peer review process (Heaven, 2018; Checco et al., 2021). This has raised similar concerns about the unintended consequences of hidden biases within black-boxed processes.

A question raised less often concerns acceptance of the use of such tools by scientists themselves. Many researchers have resisted the use of metrics and proposals to replace peer review of grant applications with funding lotteries. This gives a sense of the possible response to adoption of automated processes by funders (Wilsdon, 2021).

Potential applications of AI have also been touted in tackling fraud, plagiarism and poor practice in research. These aim to improve replicability and weed out poor quality or fraudulent findings from the scientific literature. At the same time, application of AI techniques may increase, rather than resolve,

problems of reproducibility. Various issues related to data leakage such as duplicates and sampling bias, for example, have been found in machine-learning methods (Kapoor and Narayanan, 2022).

Conclusion

Scholars have emphasised the role of expectations in legitimating actions, justifying decisions, guiding activities and attracting the interest of governments, industries and research communities towards emerging technologies to make desired technological futures a reality (e.g. Borup et al., 2006). Proponents of specific nascent technologies tend to identify pressing problems that can only be solved by that technology. Many narratives of AI in science portray it as radically transforming scientific research and heralding a new era of scientific productivity. Some narratives argue that AI will free up time for researchers, augment scientific reasoning, boost diversity and promote the decentralisation of science. While often speculative, these narratives might affect the development, use and impacts of AI (Ferri, 2022). This dynamic is an ever-present aspect of the uptake of new technologies, and AI is no different. However, rhetoric around emerging technologies always risks drawing attention from alternative possible trajectories of development, and from negative or unintended consequences.

AI technologies will reconfigure the organisation and conditions of the science base. This will have many positive consequences and potentially negative and unintended ones, such as the deskilling of researchers or problems arising from the black-boxing of key processes and practices. Such negative and unintended consequences should be considered ahead of promissory statements about AI and scientific productivity. When proponents of AI in science speak, they should be asked: are they speaking about routine monitoring or leading-edge research? When they talk about productivity, what is the product produced, and why is it important? Of course, the terms “science” and “productivity” must be used consistently.

A key implicit goal of science policy is the generation of the human, organisational and infrastructural capacity to do research. This underpins problem-driven (applied) research, scientific entrepreneurship and industrial innovation. Statements about AI in science also need to be judged in light of their effects on this research capacity. Their effects on everyday scientific tasks, as well as the opportunities for (and content of) research training, should also be assessed.

It may well be that adoption of AI tools can remove some boring routine tasks from research work, giving more space for exploratory, creative and social elements of research practice. However, as discussed above, it might equally be the case that demands for new routine tasks stemming from the adoption of AI tools will create new pressures on everyday scientific work.

Our understanding of how AI might change science cannot be detached from the human capital dimension of scientific practice. Different kinds of scientists and engineers will emerge from a scientific enterprise in which AI tools are widely used. They will have different sets of skills and be accustomed to different routines from their predecessors. Fields such as synthetic biology rely heavily on interdisciplinary collaborations (e.g. computer scientists and biologists). These scientists are slowly becoming more equipped to conduct technical work when troubleshooting machines. They are also getting used to interacting with technicians from supplier companies. Their relationship with the lab is also changing as large robotics platforms take the space of the bench. More experimental work is done in office spaces away from lab benches.

These new configurations affect the way scientists collaborate and how they co-ordinate their everyday tasks. The consequences may vary according to position in the scientific hierarchy. Innovation and adoption of new AI tools in science will also create demands on research funding and governance practices. Questions will arise about how such tools will be funded and evaluated, and how they will be used *in* funding and governance practices.

AI tools have been framed as an answer not just to the problem of scientific productivity but also to problems of reproducibility and poor practice in science. However, some argue the crisis of replicability

and related problems stem from the ever-more intense "publish or perish" nature of modern scientific competition. If anything, these critical voices say, science needs to slow down (Stengers, 2018; Frith, 2020).

Perhaps the key question is not how AI tools can accelerate scientific productivity in terms of the quanta of discoveries or their direct social and economic impacts. After all, individual parcels of new knowledge are not the key mechanism through which science has such impacts. Instead, perhaps the question should be: how can AI tools help build a slower but more sustainable, more responsible and socially productive science base?

References

Azad, B. and N. King (2008), "Enacting computer workaround practices within a medication dispensing system", in *European Journal of Information Systems*, Vol. 17/3, pp. 264-278, https://doi.org/10.1057/ejis.2008.14.

Bartlett, A., J. Lewis and M. Williams (2016), "Generations of interdisciplinarity in bioinformatics", *New Genetics and Society*, Vol. 35/2, pp. 186-209, https://doi.org/10.1080/14636778.2016.1184965.

Borup, M. et al. (2006), "The sociology of expectations in science and technology", *Technology Analysis & Strategic Management*, Vol. 18/3-4, pp. 285-298, https://doi.org/10.1080/09537320600777002.

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.

Clarke, S. (2019), "What can be learned from government industrial development and research policy in the United Kingdom, 1914-1965", in *Lessons from the History of UK Science Policy,* The British Academy, London, www.thebritishacademy.ac.uk/documents/243/Lessons-History-UK-science-policy.pdf.

Cyranoski, D. (2019), "Artificial intelligence is selecting grant reviewers in China", *Nature*, Vol. 569, pp. 316-317, https://doi.org/10.1038/d41586-019-01517-8.

Dennis, M.A. (2006), "Reconstructing sociotechnical order: Vannevar Bush and US science policy" in Jasanoff, S. (ed.), *States of Knowledge: The Co-production of Science and Social Order,* Routledge, New York.

EC (2022), "Study on factors impeding the productivity of research and the prospects for open science policies to improve the ability of the research and innovation system – final report", European Commission, Brussels, https://data.europa.eu/doi/10.2777/58887.

Ferri, P. (2022), "The impact of artificial intelligence on scientific collaboration: Setting the scene for a future research agenda", presented at Eu-SPRI 2022, 1-3 June, Utrecht, Netherlands.

Flanagan, K. et al. (2003), "Chasing the leading edge: Some lessons for research infrastructure policy", presented at ASEAT Conference, Manchester, www.research.manchester.ac.uk/portal/en/publications/chasing-the-leading-edge-from-research-infrastructure-policy-to-policy-for-infrastructureintensive-research(46322065-f9d8-40aa-ba09-f9a9dfb7318b).html.

Frith, U. (2020), "Fast lane to slow science", *Trends in Cognitive Sciences*, Vol. 24/1, pp 1-2, https://doi.org/10.1016/j.tics.2019.10.007.

Georghiou, L. and P. Halfpenny (1996), "Equipping researchers for the future", *Nature,* Vol. 383, pp. 663-664, https://doi.org/10.1038/383663a0.

Heaven, D. (2018), "AI peer reviewers unleashed to ease publishing grind", *Nature*, Vol. 563, pp. 609-610, https://doi.org/10.1038/d41586-018-07245-9.

Helmy, M., M. Awad and K.A. Mosa (2016), "Limited resources of genome sequencing in developing countries: Challenges and solutions", *Applied & Translational Genomics*, Vol. 9, pp. 15-19, https://doi.org/10.1016/j.atg.2016.03.003.

Junk, T.R. and L. Lyons (2020), "Reproducibility and replication of experimental particle physics results", *Harvard Data Science Review*, Vol. 2/4, https://doi.org/10.1162/99608f92.250f995b.

Kapoor, S. and A. Narayanan (2022), "Leakage and the reproducibility crisis in ML-based science", *arXiv*, preprint arXiv:2207.07048, https://doi.org/10.48550/arXiv.2207.07048.

Mindell, D.A. (2015), *Our robots, Ourselves: Robotics and the Myths of Autonomy*, Viking, New York.

Nickerson, R.S. (2000), "Null hypothesis significance testing: A review of an old and continuing controversy", *Psychological Methods*, Vol. 5/2, p. 241, https://doi.org/10.1037/1082-989x.5.2.241.

OECD (2015), *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development,* The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264239012-en.

OECD (1995), *Measurement of Scientific and Technological Activities: Manual on the Measurement of Human Resources Devoted to S&T – Canberra Manual,* The Measurement of Scientific and Technological Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264065581-en.

Pickstone, J.V. (2000), *Ways of Knowing: A New History of Science, Technology and Medicine*, University of Chicago Press.

Raghu, M. and E. Schmidt (2020), "A survey of deep learning for scientific discovery", *arXiv*, preprint arXiv:2003.11755, https://doi.org/10.48550/arXiv.2003.11755.

Ribeiro, B. and P. Shapira (2020), "Private and public values of innovation: A patent analysis of synthetic biology", *Research Policy*, Vol. 49/1, p. 103875, https://doi.org/10.1016/j.respol.2019.103875.

Ribeiro, B. et al. (2023), "The digitalisation paradox of everyday scientific labour: How mundane knowledge work is amplified and diversified in the biosciences", *Research Policy,* Vol. 52/1, p. 104607, https://doi.org/10.1016/j.respol.2022.104607.

Royal Society (2018), "The AI revolution in scientific research", Royal Society/Alan Turing Institute, London, https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf?la=en-GB&hash=5240F21B56364A00053538A0BC29FF5F.

Salter, A.J. and B. Martin (2001), "The economic benefits of publicly funded basic research: A critical review", *Research Policy*, Vol. 30/3, pp. 509-532, https://doi.org/10.1016/S0048-7333(00)00091-3.

Stengers, I. (2018), *Another Science is Possible: A Manifesto for Slow Science,* Polity Press, New York.

Valli, T. et al. (2007), "Over 60% of NIH extramural funding involves animal-related research", *Veterinary Pathology,* Vol. 44/6, pp. 962-963, https://doi.org/10.1354/vp.44-6-962.

Vicsek, L. (2021), "Artificial intelligence and the future of work – Lessons from the sociology of expectations", *International Journal of Sociology and Social Policy*, Vol. 41/7/8, pp. 842-861, https://doi.org/10.1108/IJSSP-05-2020-0174.

Wagner, C. (2008), *The New Invisible College: Science for Development*, Brookings, Washington, DC.

Wilsdon, J. (2021), "AI & machine learning in research assessment: Can we draw lessons from debates over responsible metrics?", presentation to Research on Research Institute & Research Council of Norway workshop, January, https://figshare.shef.ac.uk/articles/presentation/AI_machine_learning_in_research_assessment_can_we_draw_lessons_from_debates_over_responsible_metrics_/14258495/1.

Zachary, G.P. (1997), Endless Frontier: Vannevar Bush, Engineer of the American Century, MIT Press, Cambridge, MA.

Part V Artificial intelligence, science and developing countries

Artificial intelligence and development projects: A case study in funding mechanisms to optimise research excellence in sub-Saharan Africa

J. Shawe-Taylor, University College London, United Kingdom

D. Orlič, Knowledge 4 All Foundation, United Kingdom

Introduction

Artificial intelligence (AI) has been attracting increased attention from researchers, entrepreneurs, investors and policy makers on all continents. Innovative national and international development collaboration mechanisms, such as micro-funding and social impact bonds, are being tested, refined and implemented to assist AI researchers, contribute to scientific excellence and scale to market. This essay looks at emerging networks of excellence in the Global South, particularly AI4D Africa. It examines how bottom-approach, small-scale investments resulted in significant research on different scientific and non-scientific, engineering and educational topics, including a profile of African languages.

Since sustainable development is a challenge for all countries and in different ways, many are developing their own approaches to using AI to help address their specific needs. Some of these approaches span different continents and regions, and others are location-specific. For example, a research group in one region might use satellite imagery to understand the quality of water resources in lakes. A group in another region might analyse news items in a number of different languages monitoring the same issue. Yet both groups are addressing water quality management, one of the Sustainable Development Goals (SDGs) (UN, 2022).

An opportunity exists to connect research groups and tap into different solutions, case studies, skills and competences via co-operation mechanisms such as networks and centres of excellence in AI and sustainability (NAIXUS, 2022). Most of these mechanisms used to boost AI in science in developing countries are having a positive impact, while being flexible and cost effective.

The introduction of AI applications in the Global South makes possible innovative, data-driven, technical innovations to help address pressing socio-economic problems and improve policy actions. AI can facilitate scientific breakthroughs, improve medical diagnoses, increase agricultural productivity, optimise supply chains and help equalise development of skills through highly personalised learning. However, AI could also widen the breach between developed and developing countries' capabilities in science.

Most AI experts work in North America, Europe and Asia, with sub-Saharan Africa barely represented in the global pool of experts. In many new AI initiatives in Africa, the expertise involved is barely visible in

developed-country technology hubs. Nevertheless, a significant AI community has grown up in Africa in recent years, with initiatives such as Deep Learning Indaba2022, Masakhane Foundation2022, Data Science Africa (DSA, 2022) and Data Science Nigeria (DSN, 2022) with many more establishing legal entities to formalise their work. Such bottom-up approaches can bypass burdensome and bureaucratic top-down university co-operation systems.

These self-mobilising and unique emerging expert communities in Africa have created novel opportunities for co-operation and innovation by introducing funding for a range of micro-scale research projects. This funding model assumes that large projects can be cumbersome and present significant bureaucratic bottlenecks; a dynamic framework of targeted and quickly disbursed financial support is more effective overall. The question is whether this approach can accelerate innovation at scale.

Background

Networks of Excellence as vehicles of success

Networks of excellence in science aim to strengthen particular areas of science and technology through collaboration. Some networks operate at European level and aim to marshal the resources and expertise needed for Europe to be a world force in a given field. PASCAL2 was the most ambitious of a number of European-funded networks of excellence in the fields of pattern analysis, statistical modelling and computational learning. It ran from 2008 until 2013 (CORDIS 2022).

The PASCAL2 organisational and financial model helped inspire more recent networks of European AI researchers. These include the European Learning and Intelligent Systems Excellence initiative (ELISE, 2022). and the European Network of Human-Centered Artificial Intelligence (Humane AI, 2022). It also includes networks outside Europe, such as AI for Development Africa, a network of researchers and practitioners in sub-Saharan Africa (AI4D, 2022).

These networks aim to encourage researchers to collaborate, to think outside of their particular research interests and help take machine learning into other fields. Early experiences with these networks offered important lessons. They had remarkable success, for example, in empowering and trusting people with the freedom to do research, with little or no funding up front and without constant pressure for results. PASCAL2 established the Knowledge 4 All Foundation (K4A) as a UK charity (K4A, 2022). Through this legacy organisation, it would make success stories, incentive models and methodologies explored in PASCAL permanently available in Europe and beyond.

K4A supports specialised scientific communities and the general public across the world with capacity building, open educational resources and education technologies. In collaboration with the Jožef Stefan Institute, for example, the Foundation has used VideoLectures (2022) to improve access to content in all subcategories of computer science. An award-winning open educational resource, VideoLectures has a tail of free machine-learning lectures dating to 2003. It is an example of a major enabler of AI uptake that can make AI tools and education accessible globally through the Internet.

African projects

In 2018, the Knowledge 4 All Foundation worked with UNESCO and Canada's International Development Research Centre (IDRC) to map the landscape of AI in emerging economies. The mapping covered 33 countries and 617 institutions spread across Asia, Latin America and the Caribbean, the Middle East and North Africa, and sub-Saharan Africa. The result – the Emerging Economies Artificial Intelligence Ecosystem Directory (K4A, 2022) – was one of the first bottom-up mappings of AI entities in the Global South. It provided a basis for types of capacity building – from policies to support AI in science and funding bodies to research methods, dissemination of information on use cases, deployment, exploration,

exploitation and operability. These could be applied to topics relevant to the SDGs. The results helped bootstrap a series of AI-related research and development initiatives in sub-Saharan Africa.

As a result of this work, IDRC provided funding to K4A in 2019 to help establish the AI4D network. This second project aimed to strengthen and develop a community of scientific and technological excellence in a range of AI-related areas. AI4D Africa developed a network of institutions and individuals working on and researching AI from across sub-Saharan Africa, via workshops and consultations. It delivered an AI-related research agenda (Gwagwa et al., 2021) with a focus on ethical, legal and social issues underpinning scientific quality in AI research. It also generated an AI capacity building agenda (Butcher et al., 2021) via a survey of universities. Further, it issued a call for multidisciplinary innovation projects within and outside the network, exploring local frontiers of research in AI.

K4A has helped co-ordinate initiatives that are already having a significant impact in the region. These include COVID-19 data challenges; a fellowship yielding 30 African language datasets covering 22 countries with 300 million speakers; a text-to-speech platform for African languages; and a registry of AI hot spots in Africa and engagement across many researchers and research institutions.

Use-case 1: Funding micro-projects

K4A designed two calls for applications for funding of micro-projects in 2019 and 2020. Projects were required to: 1) create a dataset; 2) have a novel and motivated goal; 3) involve a challenging yet manageable task with a scalable long-term vision; and 4) be accessible to the general public and researchers. The successful projects came from nine countries and encompassed AI applications in a diverse range of scientific and social objectives.[1] Elected projects were awarded between USD 5 000-8 000 each. The call for micro-projects also generated the first African Grand Challenge in AI. It focused on curing leishmaniasis, a neglected disease that affects the region.

Use-case 2: Funding the development of datasets for African languages

Through open dialogue and a collaborative methodology inspired by the PASCAL2 model, low-resourced African languages were identified as a major blind spot. A fellowship was initiated and a set of language dataset challenges established, all to incentivise the creation, collation and uncovering of African language datasets. This five-month process saw submission of 35 datasets from a variety of African languages/dialects with more than 190 data scientists enrolled to solve the challenges. The resulting datasets were released to the African crowdsourcing machine-learning challenge platform Zindi (2022). They were also published on a dedicated channel in Zenodo (2022). The awards given to the community to solve these challenges ranged from USD 500-3 000 each.

Use-case 3: Funding challenges to predict the global spread of COVID-19

Held during the first COVID-19 lockdown in 2020, this data challenge attempted to incentivise the African AI community to engage in the global response to the pandemic. Data scientists were asked on Zindi to predict the spread of COVID-19 around the world over the following few months (Zindi, 2022). Solutions were evaluated against subsequently collected data. This challenge contributed to the global body of knowledge helping stem the impact of pandemics of different sorts. The top three solutions were made available on GitHub (2022). In all, 773 data scientists enrolled in the challenge, leading to 777 submissions. For the winning entry, the average estimate of daily cumulative deaths per country was within 208 of the actual number. The selected projects were awarded between USD 500-1 000 each.

Impact of the projects

The cumulative efforts of these tailored small-scale investments resulted in significant research initiatives on different scientific and non-scientific, engineering and educational topics. They had particular successes

in profiling African languages. Evidence of success has helped unlock several other major funding initiatives. IDRC, for example, extended the AI4D initiative and created the Lacuna Fund, which mobilises funding for labelled datasets that solve urgent problems in low- and middle-income countries (Lacuna, 2022).

The outcomes described below illustrate that small-scale initiatives can be effective in helping build AI capacity, for science and other purposes, in low-income countries:

- Empowerment of grantees by creating a level playing field where researchers and data practitioners in developing countries were trusted and treated equally to those in developed countries. This considered their financial and operational constraints, and involved them in shaping the funding agenda. The principle of intentionally avoiding any biases and instead creating a machine-learner to machine-learner relationship and working environment proved highly beneficial.
- International recognition of the creation of the African languages' datasets in 2021 when two of the outcomes received the Wikimedia Foundation Research Award of the Year (Wikimedia, 2022). Neketo et al. (2020) were awarded for the paper "Participatory research for low-resourced machine translation: A case study in African languages". The development of the Masakhane online community of AI practitioners was also awarded. This community has attempted to fundamentally change how to approach the problem of under-resourced languages in Africa. Both the work of the authors and the community have been recognised across Africa.
- Direct support from K4A to establish the Masakhane Research Foundation in Nairobi and the Tanzania AI Community in 2022. This support will potentially empower a bottom-up community of researchers and practitioners to transform itself from an informal group into a registered legal entity.
- The readiness of a number of donors to create AI-specific funding schemes. For example, IDRC and the Swedish International Development Cooperation Agency launched a four-year CAD 20 million partnership, beginning in 2020, to address a range of AI challenges in Africa.
- Google.org, the Rockefeller Foundation, IDRC and Germany's Development Cooperation agency making combined contributions to the Lacuna Fund of several million US dollars. This support came to fruition thanks to the portfolio of tangible micro-projects making the case for such a large investment.
- A gender-diverse distribution of participants. A mixed gender managerial model, along with joint male and female project leads, helped create a research environment more favourable to female principal investigators.

Conclusion

The budget for all the described micro-activities was a modest CAD 500 000 over three years. This modest funding has nevertheless created significant results. This outcome is in some ways comparable, but in a different context, to the large and long-term collaborations (around 20 years) in the European networks of PASCAL, PASCAL2, ELISE and Humane AI. These networks were made possible by relatively modest funding of approximately EUR 50 million. In all cases, the micro-project model has been effective in fostering long-term innovation and impact through small-scale direct funding with minimal bureaucratic overhead.

Financial and managerial tensions arise when creating large-scale projects with ambitious, exploratory or ground-breaking visions that lack the assurance of near-term profitability or benefit, or a clear strategy on how to achieve those goals. In such cases, it is difficult to maintain research cohesion and keep a focus on creative ideas in the ways that smaller initiatives can. The much-criticised Human Brain Project (Enserink and Kupperschmidt, 2014) is perhaps an example of the former. By contrast, an example of the latter is the PASCAL Visual Object Classes challenge (PASCAL, 2022). This challenge is still relevant after

17 years, providing the vision and machine-learning communities with a standard dataset of images and annotations, and standard evaluation procedures.

Micro-projects can create dispersed rather than centralised impacts. However, co-ordinating micro-projects as part of a larger coherent programme might deliver the best of both worlds. PASCAL2 used a bottom-up and small-scale agile funding structure but around a co-ordinated research and collaborative theme of pattern analysis and machine learning. The Humane AI network is undertaking a similar experiment with its micro-project funding programme. Again, it is co-ordinated around a series of themes and "grand challenges". This network has yielded some promising initial results, delivering almost 60 micro-projects in two years. The answer appears to be "yes, we can". However, the key is good funding management rather than funding scale or a central intellectual authority that directs the research.

On first impression, independently of the funding mechanism, there is a case for sub-Saharan Africa to receive much greater funding than that available to K4A. However, funding should be disbursed in a way that enables researchers maximum opportunity to unleash their potential to innovate. This can accelerate innovation, the production of cutting-edge research accepted at top AI conferences and establishment of trust with international research institutions and donors. Such developments could go some way to closing the divide between developed and developing countries in terms of scientific achievements, enabled in part by AI tools and access to AI-related education through the Internet.

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

AI4D (2022), Artificial Intelligence for Development Africa website, https://africa.ai4d.ai (accessed 10 August 2022).

Butcher, N. et al. (2021), "Artificial intelligence in sub-Saharan Africa, 2021", International Research Centre in Artificial Intelligence under the auspices of UNESCO, https://ircai.org/project/ai4d-ai-in-sub-saharan-africa.

CORDIS (2022), "Pattern analysis, statistical modelling and computational learning 2" (fact sheet), CORDIS, European Commission, https://cordis.europa.eu/project/id/216886.

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

DSN (2022), Data Science Nigeria website, www.datasciencenigeria.org (accessed 11 September 2022).

Deep Learning Indaba (2022), "Deep Learning Indaba 2022", webpage, https://deeplearningindaba.com/2022 (accessed 16 August 2022).

ELISE (2022), European Network of AI Excellence Centres website, www.elise-ai.eu (accessed 5 July 2022).

Enserink, M. and K. Kupperschmidt (2014), "Updated: European neuroscientists revolt against the E.U.'s Human Brain Project", 11 July, *Science*, www.science.org/content/article/updated-european-neuroscientists-revolt-against-eus-human-brain-project.

GitHub (2022), "Zindi wins AI4D Predict the Global Spread of COVID-19 Insights", webpage, https://GitHub.com/Dr-Fad1/Zindi-wins-AI4D-Predict-the-Global-Spread-of-COVID-19-insights (accessed 14 August 2022).

Gwagwa, A. et al. (2021), *Responsible Artificial Intelligence in sub-Saharan Africa: Landscape and General State of Play*, International Research Centre in Artificial Intelligence under the auspices of UNESCO, https://ircai.org/project/ai4d-responsible-ai-in-sub-saharan-africa.

Humane AI Net (2022), European Network of Human-centered artificial intelligence website www.humane-ai.eu (accessed 20 September 2022).

K4A (2022a), Knowledge for All website, www.k4all.org (accessed 20 September 2022).

K4A (2022b), "Emerging Economies Artificial Intelligence Ecosystem Directory", webpage, www.k4all.org/ai-ecosystem (accessed 21 September 2022).

Lacuna Fund (2022), Lacuna Fund website, https://lacunafund.org (accessed 5 July 2022).

Masakhane (2022), Masakhane website, www.masakhane.io (accessed 16 June 2022).

Naixus (2022), Naixus website, http://naixus.net (accessed 11 November 2022).

Neketo, W. et al. (2020), "Participatory research for low-resourced machine translation: A case study in African languages", *arXiv*, arXiv:2010.02353 [cs.CL], https://arxiv.org/abs/2010.02353.

PASCAL (2022), The PASCAL Visual Object Classes Homepage website, http://host.robots.ox.ac.uk/pascal/VOC (accessed 20 August 2022).

UN (2022), "Sustainable Development Goal 6", webpage, www.un.org/sustainabledevelopment/water-and-sanitation (accessed 2 June 2022).

VideoLectures (2022), VideoLectures website, http://videolectures.net (accessed 15 July 2022).

Wikimedia (2022), "Wikimedia Foundation Research Award of the Year", webpage, https://research.wikimedia.org/awards.html (accessed 18 August 2022).

Zenodo (2022), "African Natural Language Processing (AfricaNLP)", webpage, https://zenodo.org/communities/africanlp/search?page=1&size=20 (accessed 18 August 2022).

Zindi Africa (2022a), "GIZ AI4D Africa Language Challenge – Round 2", webpage https://zindi.africa/competitions/ai4d-african-language-dataset-challenge (accessed 22 June 2022).

Zindi Africa (2022b), "AI4D Predict the Global Spread of COVID-19", webpage, https://zindi.africa/competitions/predict-the-global-spread-of-covid-19 (accessed 22 June 2022).

Note

[1] Burkina Faso (Building a Medicinal Plant Database for Preserving Ethnopharmacological Knowledge in the Sahel), Burkina Faso (Preservation of Indigenous Languages), Kenya (A Public Dataset on Poaching Trends in Kenya and a Study on the Predictive Modeling of Poaching Attacks), Kenya (Early detection of preclampsia using ambulatory blood pressure monitoring using wearable devices and Long Short Term Memory Networks (LSTM-NN) on the edge), Malawi (A Semi-Automatic Tool for Meta-data extraction from Malawi Court Judgments), Morocco (Arabic Speech-to-MSL Translator: Learning for Deaf), Nigeria (Using Artificial Intelligence to Digitize Parliamentary Bills in Sub-Saharan Africa), Tanzania (A Computer vision Tomato Pest Assessment and Prediction tool), Tanzania (Effective Creation of Ground Truth Data-set for Malaria Diagnosis Using Deep Learning), Tanzania (Improving the Pharmacovigilance system using Natural Language Processing on Electronic Medical Records).

Artificial intelligence for science in Africa

G. Barrett, Cirrus AI, South Africa

Introduction

Academic and other research institutions in Africa carry out much fundamental and applied scientific research, but few as yet use artificial intelligence (AI). African science needs to take up AI methods. In the absence of such methods, an increasing number of scientific disciplines at African institutions will border on irrelevance. A greater use of AI in scientific research in Africa will bring numerous benefits, deepening African science, broadening global research agendas and incentivising the location of corporate research and development (R&D) labs. Ultimately, the use of AI in science will have spillover effects, helping to upgrade the capabilities of civil society more broadly.

Cirrus and the AI Africa Consortium are a major response to the AI deficit in African science. They aim to broaden researcher access to computing, data, engineering resources and trained students. Ultimately, this will make AI for science feasible in numerous academic institutions across Africa – not just elite academic institutions and large technology firms. In so doing, they will help commercialise research findings. With human capital central to AI, online learning can play an important role in knowledge transfer to Africa.

Prioritising AI for science in Africa

AI-enabled scientific research is not yet happening in Africa. Most of the leading corporate research operations are active in Asia, Europe and North America but not in Africa. This is an important barrier to collaborative research and commercialisation efforts at African institutions.

Data from the QS World University Rankings, since 2012, show that Fortune 500 companies collaborated six times more with the top 50 universities than with those ranked between places 301 and 500, where most African universities are situated[1] (Ahmed and Wahed, 2020). This imbalance in collaboration exacerbates disparities between academic institutions in Africa and top tier academic institutions in the rest of the world.

Furthermore, Fortune 500 technology companies and the top 50 universities publish five times as many papers annually per AI conference than universities ranked between 200 and 500. The research budgets of premier academic research institutions like Carnegie Mellon University's Robotics Institute – at USD 90 million in 2019 (Spice, 2019) – are a fraction of that of the major industrial companies. However, they are still orders of magnitude greater than that of any academic institution in Africa.

While world-class research does take place at African institutions, African researchers lack the data, computing infrastructure and engineering resources to develop and apply the more powerful and critical AI methods. Even for the world's elite academic institutions and researchers, it is increasingly difficult to work at the frontier of AI research (Sample, 2017). For example, OpenAI analysed the relationship between the availability of computational resources and 15 relatively well-known breakthroughs in AI between 2012 and

2018 (Amodei and Hernandez, 16 May 2019). Of the 15 developments examined, 11 were achieved by private companies, while only 4 came from academic institutions.

In terms of training and human capital development, universities are fortunate that the AI field involves many feasible options to rapidly upskill researchers. This is resulting in a paradigm shift for many in academia who are accustomed to building courseware.[2] Forward-thinking universities have been steadily moving towards a "flipped classroom". In this format, learners watch videos and complete in-depth assignments and online quizzes at home, then come to class for discussions. The classes generally culminate in an open-ended final project, supported by the teaching team. The university often uses previously developed high-quality Massive Open Online Courses as the core course material. It then focuses on supplementary domain-specific materials, projects and assignments. With this approach, students in developing countries can access courseware used at elite universities. The cost to both students and the university are well below the previous alternatives.

A range of new capabilities and leadership is required to deploy AI for science

New capabilities and leadership are needed if African research institutions are to harness new AI methods. Such capabilities require engineering personnel to prepare data, and configure hardware, software and machine-learning algorithms, which are absent in most of Africa. In addition, the ad hoc mix of campus computers and commercial clouds that Africa's educators and researchers rely on today are inadequate.[3]

Simply providing underserved academic and research organisations with the data, hardware, software and engineering resources is insufficient. To truly reduce barriers to AI-enhanced research, underserved institutions need access to experts who can implement best practices. Key areas include approaches to problems, learning methods, selection of tools for tasks and optimisation of workflows.

An example is the development of AI-ready datasets. In some fields of science, data are abundant. However, in many scientific areas, sufficiently large datasets either do not exist and/or are not accessible in forms that permit the use of AI methods. Substantial effort is required to create new datasets. This could include locating and cleaning the data, aligning the schemas of disparate data, ensuring machine readability and providing relevant metadata pertaining to issues such as data provenance, quality and completeness. This expensive and error-prone process must be repeated for each analysis. This becomes a barrier to using data, and also leads to problems of research reproducibility. Furthermore, privacy and security issues need to be addressed from the beginning rather than after the fact. The process must provide integrated assurances and audit capabilities to advance research in the public interest.

Data engineering is often needed to develop specific software tools to construct the dataset for AI. Most of this tool development happens without considering possible public or inter-experiment collaborations. When collaborations are eventually sought, researchers may find their work has already been duplicated.

Providing a data management platform to enable efficient AI development and sharing is a priority for Cirrus. Such a platform will enable users to store, manage, share and find data with which to develop AI systems. This includes tracking data, versioning support for various data formats and complete metadata to allow for retraining and understanding models built from the data. Such a platform will drive advances in AI by enabling researchers to experiment with existing and new methods in new contexts. It will benefit the disciplines in which the datasets are created.

For African academic and research institutions, moving forward on AI also requires a significant increase in the scientific throughput that feeds AI systems. Governments, and academic and research institutions in the region need to generate more and better quality data, and to make data accessible. The use of findable, accessible, interoperable and reusable data principles, and participation in a centralised set of standards for benchmark datasets in scientific domains, are both needed. These will help govern data storage formats, access and metadata to reduce engineering overhead and lower the barriers to training

and comparing model performance.[4] A high priority must be to identify and use existing and potential scientific data-generating programmes to produce AI-ready data repositories. The liberating of data in a privacy-preserving manner must extend across science, from Earth observation to health care. Doing so will support science and aid in using AI to address diverse pressing social problems.

Cirrus and the AI Africa Consortium

Cirrus and the AI Africa Consortium are ambitious by African standards. Cirrus emerged in 2017 from a need to use AI in a scientific collaboration at Wits University. The university leadership then decided that Cirrus should benefit all academic and research institutions in Africa.

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 machine learning for embedded devices. Full implementation will commence following the confirmed participation of the Strategic Founding Partners (SFPs).

Cirrus

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 dedicated compute infrastructure will be enormously important. Based solely on hardware costs, it is more cost effective to own infrastructure when computing demand is close to continuous. Estimates show that commercial cloud services are more expensive per compute cycle than a dedicated high-performance computing cluster (Villa and Troiano, 30 July 2020). The initial costs of subsidising cloud use might be less than building public infrastructure. However, studies show that relying on commercial cloud services will likely be much more expensive in the long term (Wang and Casado, 2021).

Through a variety of financial and other mechanisms, Cirrus is designed to help attract corporate research in AI (and associated venture capital activity), targeting multinationals not yet active in this field in Africa. Ultimately, Cirrus would be owned, through equity, by around 15-25 multinational corporations. Each of these SFPs would commit USD 7-20 million.

The diversity in ownership should bring with it a diversity of research interests. This will help avoid AI research focused on a narrow set of ideas and methods biased to the interests of any particular private sector participant. The research mission of Cirrus is also isolated from political influence, from changes in political administrations and from politically appointed administrators. Allocation of Cirrus resources to the AI Africa Consortium will occur through a mix of peer review, lottery and equitable distribution criteria. As a private sector entity, Cirrus is also not encumbered by the intellectual property constraints that ensnare research commercialisation efforts at publicly funded universities.[5] This provides Cirrus with the flexibility to support a range of commercialisation options.

Cirrus has three components. First, it will house co-operation programmes, the state-of-the-art computing and data infrastructure, engineering personnel and the open learning programmes. Second, the Cirrus FOUNDRY – a form of business incubator – is equipped with everything needed to turn insights from scientific research into start-ups and eventually larger commercial applications. Third, the Cirrus FOUNDRY Fund is an in-house fund to support start-ups in the Cirrus FOUNDRY. The Cirrus FOUNDRY Fund has a target capitalisation of USD 35 million and will undertake pre-seed and seed stage investments.

The physical infrastructure and operations for Cirrus are to be housed at Wits University in Johannesburg, South Africa. Wits University was selected as the host institution for three reasons:

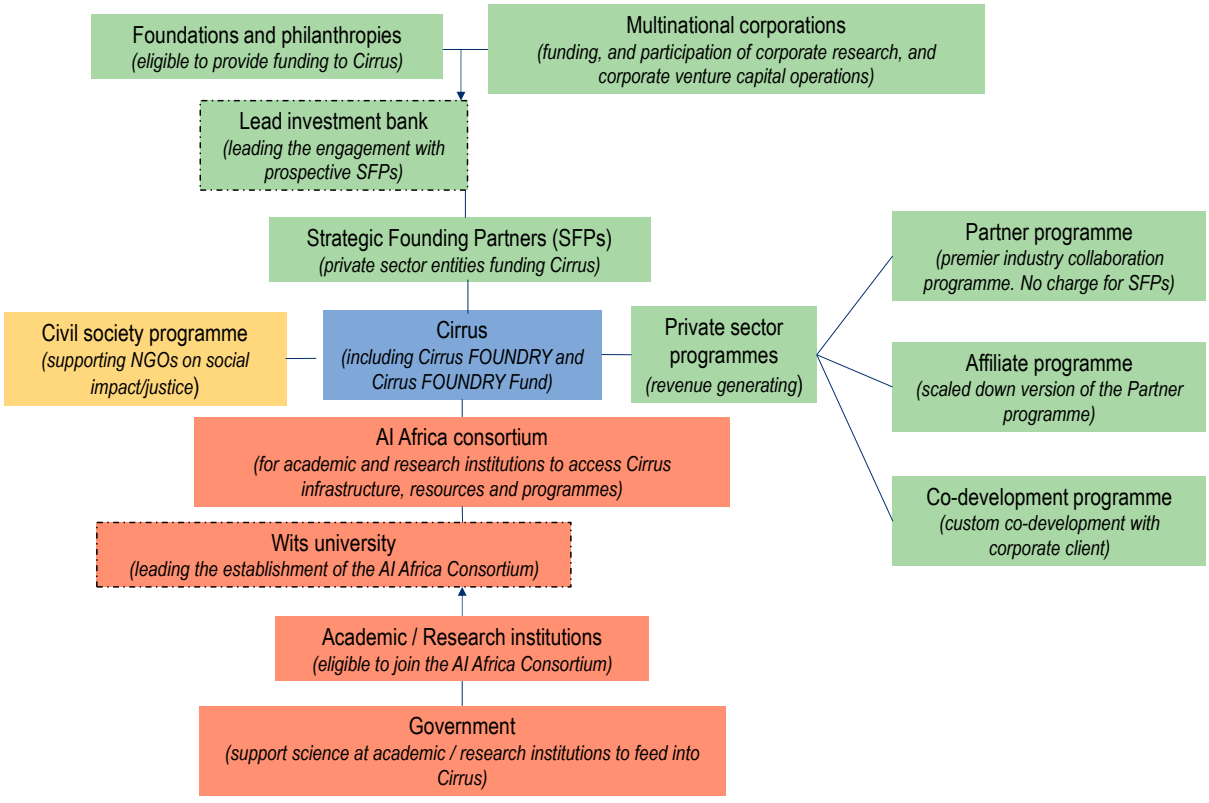
1. South Africa is the most scientifically advanced country on the African continent (Mouton et al., 2019), and Wits is one of Africa's leading academic research institutions.[6]

2. Wits is situated geographically in the highest concentration of economic, academic and research activity in Africa.
3. Wits has the land available to house the necessary infrastructure, including for energy generation and storage.

The AI Africa Consortium

The Africa AI Consortium fosters collaboration agreements with parties across the African R&D ecosystem. The agreements focus on helping identify research priorities, spreading AI research resources and engaging African research talent.[7] The Consortium aims to create significant AI research capabilities by developing skills and recruiting researchers and other skilled personnel from across Africa. It will then pair these capabilities with those provided through Cirrus.

Figure 1. The organisational layout of Cirrus and the AI Africa Consortium



Source: Cirrus AI (2022), https://aiafrica.ac.za.

The Consortium will:

- help and encourage researchers to interact and collaborate beyond disciplinary or institutional silos
- reduce redundancy of effort and cost as new research projects will not have to build capabilities or collect new data from scratch each time
- accelerate discovery and improve reproducibility through sharing of datasets, metadata, models, software, hardware and other resources
- reduce the cost for individual research programmes involved in integrating capabilities and/or comparing their work with that of others

- foster a co-design culture where teams of scientific users, engineers and instrument providers can help develop new and broadly applicable capabilities and tools
- support a research ecosystem that understands the full context for AI solutions.

Figure 1 sets out the organisational structure of the Consortium. At the time of writing, the next step is the appointment of the lead investment bank for solicitation of the SFPs. Following placement of the SFPs, the Partner, Affiliate and Co-development programmes will be rolled out.

Efforts underway within the AI Africa Consortium include:

- TinyML4D*:* The rollout of machine learning on embedded devices, targeted at developing countries. It includes the provision of free hardware kits, workshops, courseware and a network of research and collaboration opportunities.[8] TinyML4D began in 2021 and is being scaled up.
- MLCommons*:* Fostering African participation in the development of science benchmarks, particularly those relevant to African researchers.[9]
- Remote Excellence Fellowships: A remote internship system to help talented graduate students connect with leading researchers in Europe. The first cohort is planned for September 2022.

Conclusion

Fundamental and applied R&D at academic and research institutions in Africa are at risk of marginalisation. Resources essential to AI – compute, hardware, software, accessible data and machine-learning engineering – are out of reach. The growing imbalance in AI resources and innovation between Africa and the rest of the world requires an unprecedented response. The establishment of Cirrus and the AI Africa Consortium is one of Africa's responses. It aims to help spread opportunity more widely; support students and researchers at universities and research institutions across Africa; activate the talent of researchers once they have access to AI infrastructure and other resources; and create fertile ground for commercialisation through entrepreneurship.

For science in Africa, Cirrus and the AI Africa Consortium afford a major opportunity to develop and exploit AI techniques and methods. This will improve both the efficacy and efficiency of science, and also the operation and optimisation of scientific infrastructure (because system scale and complexity demand AI-assisted design, operation and optimisation).

Strengthening science in Africa by AI methods will broaden global research agendas and elevate African research. To accomplish this, Africa must also act collectively and collaborate to grow the scientific output needed to exploit opportunities presented by AI.

The goals described in this essay are challenging and the proposed solutions will require significant investment. However, the potential return on that investment is enormous: new types of data analysis; improved and even autonomous operations and performance of scientific instruments; innovative commercial products emerging from science, with even the potential for new industries; and an opportunity for Africa to become a producer of AI for science and not merely a consumer of the resulting breakthroughs.

References

Ahmed, N. and M. Wahed (2020), "The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research", *arXiv,* https://arxiv.org/pdf/2010.15581.pdf.

Amodei, D and D. Hernandez (16 May 2019), "AI and compute", OpenAI Blog, https://openai.com/blog/ai-and-compute.

Cutcher-Gershenfeld, J. et al. (2017), "Five ways consortia can catalyse open science", *Nature*, Vol. 543, pp. 615-617, https://doi.org/10.1038/543615a.

Mouton, J., et al. (2019), *The State of the South African Research Enterprise,* DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Matieland, South Africa, www0.sun.ac.za/crest/wp-content/uploads/2019/08/state-of-the-South-African-research-enterprise.pdf.

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

QS World University Rankings (2021), "QS World University Rankings" webpage, www.topuniversities.com/university-rankings/world-university-rankings/2021 (accessed 6 January 2023).

Reddi, J.V. et al. (2021), "Widening access to applied machine learning with TinyML", *arXiv*, arXiv:2106.04008v2, 9 July, https://arxiv.org/pdf/2106.04008.pdf.

Rubiera, C. (19 July 2021), "AlphaFold 2 is here: What's behind the structure prediction miracle", Oxford Protein Informatics Group blog, www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle.

Sample, I. (2017), "'We can't compete': Why universities are losing their best AI scientists", 1 November, *The Guardian*, www.theguardian.com/science/2017/nov/01/cant-compete-universities-losing-best-ai-scientists.

South African Government (2010), *Intellectual Property Rights from Publicly Financed Research and Development Act: Regulations*, www.gov.za/documents/intellectual-property-rights-publicly-financed-research-and-development-act-regulations-1.

Spice, B. (2019), "Hebert named dean of Carnegie Mellon's top-ranked School of Computer Science", 8 August, Carnegie Mellon Computer Science Department, https://csd.cmu.edu/news/hebert-named-dean-carnegie-mellons-top-ranked-school-computer-science.

The *Time*s Higher Education (2022), *Emerging Economies University Ranking*s *2022* (database), www.timeshighereducation.com/world-university-rankings/2022/emerging-economies-university-rankings (accessed 6 January 2023).

Villa, J. and D. Troiano (30 July 2020), "Choosing your deep learning infrastructure: The cloud vs. on-prem debate", Determined AI blog, https://determined.ai/blog/cloud-v-onprem.

Wang, S and M. Casado (2021), "The cost of cloud, a trillion dollar paradox", Andreessen Horowitz, 27 May, https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization.

Notes

[1] Africa's highest ranked university in 2021 was the University of Cape Town, in 220th place. For the full list of rankings, see QS World University Rankings (2021).

[2] Reddi (2021) provides an example of what it takes to build and maintain high quality courseware for machine learning.

[3] For commentary on the engineering skills that went into developing AlphaFold 2, see Rubiera (19 July 2021).

[4] For recommendations concerning access to research data from public funding, see OECD (2021).

[5] For the regulations governing intellectual property rights from publicly financed research in South Africa, see South African Government (2010).

[6] See The *Times Higher Education* Emerging Economies University Rankings (2022).

[7] For an overview of why consortia can catalyse open science, see Cutcher-Gershenfeld (2017).

[8] For information on TinyML4D, see http://tinyml.seas.harvard.edu/4D/.

[9] For information on the MLCommons Science Working Group, see https://mlcommons.org/en/groups/research-science/.

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

P.M. Addo, Agence Française de Développement (AFD), France

Introduction

COVID-19 sparked a range of uses of artificial intelligence (AI) in the search for solutions and underscored the importance of data for policy making. This essay points to the discrepancies in AI capabilities between rich and poor countries. It then considers how bilateral and multilateral development co-operation could assist, specifically in connection with AI in science.

Limited AI readiness in developing countries

The use of AI for research and development (R&D) is still out of reach for most researchers in developing countries. Europe, North America, and East and Central Asia are the world's dominant sources of AI conference publications. 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% (Zhang et al., 2021). Furthermore, researchers from developing countries often play little if any role in key international conversations on AI, especially those held in the United States, Canada and Europe.

In science and more generally, most developing countries are not yet well prepared to harness the opportunities presented by AI technologies. The overall gap in capabilities between developed and developing countries is evident in the findings of the Government AI Readiness Index 2021, which measures the capabilities and enabling factors required for a country to implement AI solutions (Oxford Insights, 2022). Billions of people are still without Internet access; basic technological and data infrastructure is often deficient; and R&D spending is limited. Meanwhile, datasets generated in developed countries are sometimes unsuited to training AI systems to meet local needs. Such deficits could exacerbate inequalities between high- and low-income countries in the productivity of science, in economic performance and in the quality of public services.

Strategic bilateral and multilateral co-operation to strengthen AI in developing-country science

COVID-19 highlighted the need for collective global partnerships: development co-operation can help. This section highlights examples of bilateral and multilateral co-operation around AI.

Identifying synergies: The Africa Regional Data Cube

Among other measures, development co-operation can provide fora for dialogue on shared challenges and relevant technological innovations, and help identify synergies between actions regionally and internationally. An example is the Africa Regional Data Cube, an initiative brought about by collaboration among many actors. These include the Committee on Earth Observation Satellites, Strathmore University in Kenya and the Global Partnership for Sustainable Development Data. Directly supporting activities in Ghana, Kenya, Senegal, Sierra Leone and Tanzania, the Data Cube has helped harness the latest Earth observation and satellite technology and data to address issues related to food security, urbanisation, deforestation and more (Global Partnership for Sustainable Development Data, 2018).

Strengthening AI readiness: The #Data4COVID19 Africa Challenge

Development co-operation can also help countries advance data protection legislation, improve data infrastructures and strengthen overall AI readiness. A good 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 (Verhulst et al., 2022). Respect for data ethics and data responsibility were a key part of the Africa Challenge. Thus, every initiative proposed complied with the European Union's General Data Protection Regulation.

Exchanging knowledge

Both bilateral and multilateral co-operation can also provide opportunities for knowledge sharing and talent attraction via mobility exchange programmes facilitated by reforms such as easing restrictions on visas.

Fostering collaboration: The IA-Biodiv Challenge

Bilateral co-operation can also help plan, finance and assist implementation of research and technological development initiatives 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 research initiative provides a space for scientists working on AI and biodiversity in France and Africa to mutually learn, share and engage.

Supporting open science, centres of excellence and networking: ARCAI and AI4D

Development co-operation can also go beyond sharing data to supporting open science initiatives. For example, most datasets on languages in Africa are not yet openly available, and local AI developers often need to use unrepresentative data from developed countries.[1]

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 (ARCAI) in the Democratic Republic of Congo (DRC). The ARCAI is the result of a collaboration between the Economic Commission for Africa and the DRC government. ARCAI will assist in AI research, collaborate with universities in Africa, participate in the creation of a network of researchers and contribute to training to help citizens actively participate in the digital transition.

Canada's International Development Research Centre in collaboration with Sweden's International Development Cooperation Agency launched the Artificial Intelligence for Development in Africa (AI4D) initiative. This partnership, with an investment of CAD 20 million over four years, aims to support African-

led research on using AI to meet local needs. AI4D, in partnership with the Human Sciences Research Council ZA in South Africa, also supports the African Observatory on Responsible AI (AORAI). In addition, AI4D works with the African Union Development Agency to develop a model African AI policy. The Observatory aims to position the African continent in global debates and policy making on responsible AI.

Encouraging private-public collaborations: The 100 Questions Initiative

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.

The selection of such questions could be informed by a dialogue between civil society, the private and public sectors, and academic and research institutions. Knowing priority questions could lead to new forms of data collaboration with the private sector to help advance the necessary science. For example, in the quest for stakeholders in Bangladesh to analyse and respond to climate extremes, a leading telecommunications provider, Grameenphone, shared its anonymous mobile call data records with three partners. Grameenphone, the United Nations University Institute for Environment and Human Security, the International Centre for Climate Change and Development, and the Telenor Group examined population movements before and after cyclone Mahasen struck Bangladesh in May 2013, an extreme climate event that affected 1.3 million people. Private-public collaborations can also stimulate investments in data infrastructures and open data sharing essential for using AI in science.

Conclusion

This essay highlights the low general level of readiness for use of AI in developing countries. It also considers how bilateral co-operation can contribute to improving the productivity of developing-country science, in particular through greater use of AI. Both bilateral and multilateral development co-operation could strengthen science in the developing world, broaden global research agendas, orient uses of AI-enabled science towards problems of particular concern to poor countries and ultimately assist global efforts to achieve the Sustainable Development Goals.

References

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 6 January 2023).

Addo, P.M. et al. (2021), "Emerging uses of technology for development: A new intelligence paradigm", *AFD Policy Paper*s, No. 6, March, Agence Française de Développement, Paris, www.afd.fr/en/ressources/emerging-uses-technology-development-new-intelligence-paradigm.

Global Partnership for Sustainable Development Data (2018), Africa Regional Data Cube Initiative website www.data4sdgs.org/initiatives/africa-regional-data-cube (accessed 6 January 2023).

Hao, K. (2021), "The race to understand the exhilarating, dangerous world of language AI", 20 May, *MIT Technology Review*, www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project.

Oxford Insights (2022), *Government AI Readiness Index 2021*, Oxford Insights, Malvern, www.oxfordinsights.com/government-ai-readiness-index2021.

The 100 Questions (n.d.), The 100 Questions website, https://the100questions.org (accessed 6 January 2023).

Verhulst, S. et al. (2022), "Building data infrastructure in development contexts: Lessons from the #Data4COVID19 Africa Challenge", *A Question of Development*, No. 56, March, Agence Française de Développement, Paris, www.afd.fr/en/ressources/building-data-infrastructure-development-contexts-lessons-data4covid19-africa-challenge.

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.

Note

[1] Currently, more than 500 researchers around the world are working together, under the BigScience project led by Huggingface, to learn more about the capabilities and limitations of large multilingual language models (Hao, 2021).

Artificial Intelligence in Science

CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH

The rapid advances of artificial intelligence (AI) in recent years have led to numerous creative applications in science. Accelerating the productivity of science could be the most economically and socially valuable of all the uses of AI. Utilising AI to accelerate scientific productivity will support the ability of OECD countries to grow, innovate and meet global challenges, from climate change to new contagions.

This publication is aimed at a broad readership, including policy makers, the public, and stakeholders in all areas of science. It is written in non-technical language and gathers the perspectives of prominent researchers and practitioners. The book examines various topics, including the current, emerging, and potential future uses of AI in science, where progress is needed to better serve scientific advancements, and changes in scientific productivity.

Additionally, it explores measures to expedite the integration of AI into research in developing countries.

A distinctive contribution is the book's examination of policies for AI in science. Policy makers and actors across research systems can do much to deepen AI's use in science, magnifying its positive effects, while adapting to the fast-changing implications of AI for research governance.

PRINT ISBN 978-92-64-44154-5
PDF ISBN 978-92-64-44621-2