Comment

MATTHEW HORWOOD/GETTY

Artificial-intelligence models could be used to sift the scientific literature and provide policymakers with the latest knowledge.

AI tools as science policy advisers? The potential and the pitfalls

Chris Tyler, K. L. Akerlof, Alessandro Allegra, Zachary Arnold, Henriette Canino, Marius A. Doornenbal, Josh A. Goldstein, David Budtz Pedersen & William J. Sutherland

Large language models and other artificial-intelligence systems could be excellent at synthesizing scientific evidence for policymakers – but only with appropriate safeguards and humans in the loop.

Recent advances in artificial intelligence (AI) have stoked febrile commentary around large language models (LLMs), such as ChatGPT and others, that can generate text in response to typed prompts. Although these tools can benefit research[1], there are widespread concerns about the technology – from loss of jobs and the effects of over-reliance on AI assistance, to AI-generated disinformation undermining democracies.

Less discussed is how such technologies might be used constructively, to create tools that sift and summarize scientific evidence for policymaking. Across the world, science advisers act as knowledge brokers providing presidents, prime ministers, civil servants and politicians with up-to-date information on how science and technology intersects with societal issues.

From solid-state batteries and antibiotic resistance to deep-sea mining, science advisers have to nimbly navigate a vast array of information. They must dig through the millions of scientific papers that are published each year, while considering reports from advocacy organizations, industry and scientific academies, each with their own take. Advisers must work fast – policy deadlines are more rigid and hastier than those in academia.

Producing policy summaries in weeks, days or sometimes hours is a daunting task. And the pressure on governments to produce and use such information is growing.

AI-based tools could increase the capacity of science advisers and help policymakers to stay afloat. But how should they be designed? Might AI undermine rigour? Could their outputs be influenced by those with an agenda? And, if AI tools are used by science advisers, what could guard against AI errors and garbled 'hallucinations' affecting public-policy decisions?

Answers to these questions are urgently needed. Powerful language models are already widely deployed in research and technological development[1], with increasingly sophisticated capabilities becoming available to more users through commercialization and open sourcing. Policymakers have already started experimenting with publicly available generative AI tools. Legislative staff members in the United States are experimenting with OpenAI's GPT-4 (see go.nature.com/3zpwhux) and, reportedly, other unapproved and potentially less reliable AI tools. This led administrators in the US House of Representatives to impose limits on chatbot use in June (see go.nature.com/3rrhm67).

Our view is that, with careful development and management, a new generation of AI-based tools could, in the near future, present an opportunity to drastically improve science advice, making it more agile, rigorous and targeted. But leveraging such tools for good will require science advisers and policy institutions to create guidelines and to carefully consider the design and responsible use of this nascent technology.

Here we explore two tasks for which generative AI tools hold promise for policy guidance — synthesizing evidence and drafting briefing papers — and highlight areas needing closer attention.

How AI can speed up evidence synthesis

Current evidence searches are time-consuming and involve a lot of judgement. Hard-pressed science advisers must take what they can get. But what if the searches could be more algorithmic?

Two main approaches are used to synthesize evidence for policy: systematic reviews and subject-wide syntheses. Both require enormous effort and take years to run. In the future, AI-based platforms should be able to make such syntheses less time-consuming, freeing subject-matter experts to focus on more complex analytical aspects.

Systematic reviews — such as Cochrane reviews in health and medicine — identify a question of interest and then systematically locate and analyse all relevant studies to find the best answer (see www.cochranelibrary.com). For example, one recent review examined evidence on whether healthy-eating initiatives were successful in young children, finding that they can be, although uncertainties remain[2].

SARAH SILBIGER/BLOOMBERG VIA GETTY

Engineer Arati Prabhakar (left) is chief science adviser to US President Joe Biden.

The alternative approach of subject-wide evidence syntheses entails reading the literature at scale[3]. For example, as part of a biodiversity project called Conservation Evidence, some 70 people spent the equivalent of around 50 person-years reading more than 1.5 million conservation papers in 17 languages, and summarized all 3,689 tested interventions. The summaries were then read by an expert panel, which assessed the effectiveness of each intervention. The synopses, on subjects ranging from bat conservation to sustainable aquaculture, are published online (https://conservationevidence.com). A parallel tool, Metadataset, allows users to tailor meta-analyses to their own needs[4].

Increasingly, machine learning can automate the search, screening and data-extraction processes that form the early stages of systematic reviews[5]. For example, LLMs such as Semantic Scholar's TLDR feature can summarize large corpuses of text — a handy feature for sifting the scientific literature. AI tools could be especially useful in making sense of emerging domains of research, in which review papers and disciplinary journals might be lacking. For instance, techniques for natural language processing can systematically classify research on AI itself[6], and graph algorithms are being used to detect emerging 'clusters' of research in the broader literature (see, for example, https://sciencemap.eto.tech). Nonetheless, assessing data quality and drawing conclusions from the amassed evidence still typically require human judgement.

Automated processes for search, screening and data extraction could also be helpful for decision-making. AI tools can create a list of possible options in a process known as solution scanning[7]. Take, for example, policies for reducing shoplifting. When prompted to list potential policy options, ChatGPT can identify topics such as employee training and store layout and design. Advisers can then collate and synthesize the relevant evidence in these areas. Such rapid assessments will inevitably miss some options, although they might also find others that conventional approaches would not. Which dimensions of credibility are most important might also differ, depending on the policy question and context.

Automation would also address another common problem: limited language skills. Science advisers who speak English have it easy, because it is the main language of science. But there is a great deal of policy-relevant literature in other languages. One analysis[8] of the biodiversity conservation literature revealed that more than one-third of papers were published in languages including Spanish, Portuguese, Chinese and French. AI tools for evidence synthesis, together with increasingly powerful LLM-based machine translation, should be able to place global information in the hands of advisers who would otherwise be constrained by a language barrier.

To realize the undoubted potential of AIs in

drawing together evidence while minimizing possible drawbacks, the following three issues must be considered.

Consistency

Many academic journals use standardized formats for reporting study results, but there is great variation across disciplines. Other sources of information, including working papers, project reports and publications from international agencies, non-governmental organizations and industry, are even more mismatched. Such diversity in presentation makes it difficult to develop fully automated methods to identify specific findings and study criteria. For example, it is usually important to know over what period an effect was measured or how large the sample was, but this information can be buried in the text. Presenting the research methodology and results in a more consistent manner could help. For instance, in medical and life-sciences research, journals published by Cell Press use a structured reporting format called STAR Methods (see go.nature.com/3ptjqcf).

Credibility

Science advisers judge whether evidence is trustworthy in five ways: the plausibility of the findings (assessed on the basis of the advisers' subject knowledge and evaluation of the research); the authors' reputations; the standing of the authors' institutions; the views of others in the field; and the perspectives of colleagues and peers. This multifaceted judgement is hard to replicate in an AI tool. Publication metrics, such as impact factors and citation counts, are found to be poor measures of research quality[9]. Which dimensions of credibility are most important might also differ, depending on the policy question and context. Experts will need to agree on standards for research quality before these can be automated in AI-based tools — a significant task, although progress is being made.

Database selection and access

Currently, conducting systematic reviews requires searching across databases — mostly proprietary ones — to identify relevant scientific literature. The choice of database matters and can have a substantial impact on the outcome. But requirements by governments to publish funded research as open access[10,11] could make it easier to retrieve study results. For research topics that governments deem as funding priorities, eliminating paywalls will enable the creation of evidence databases and ensure alignment with copyright laws.

As publishers develop other analytical tools in their databases, they might also create their own evidence-synthesis tools, but these will be limited by the scope of their coverage. Furthermore, if these tools are developed only by the private sector, this could limit access by governments in low- and middle-income countries that are least able to pay for them, but that need these services the most. Access and interoperability of databases, and government collaboration, are therefore essential foundations for large-scale automated evidence synthesis.

How AI could help to draft text for policy briefs

Advances in LLMs might make it possible for science advisers to spend less time drafting useful products for decision makers, and more time editing and crafting them. But more work is needed to test the reliability of such systems and to understand where they might err.

Policy briefs are a central part of science advice. Take, for example, the UK Parliamentary Office of Science and Technology (POST) and the US congressional science advisory body: Science, Technology Assessment, and Analytics (STAA), based in the Government Accountability Office. Both bodies write briefings (POSTnotes and Spotlights, respectively) on science and technology issues to inform a wide range of policymakers. Advisers who work for legislative committees might spend most of their time drafting questions for committee members to ask witnesses, and creating reports summarizing research evidence.

"Human reviewers and policy designers still have an essential part to play in creating policy papers."

We do not propose that policy briefs be drafted by LLM-based tools in their entirety, but AI could be used to facilitate parts of the process. Human reviewers and policy designers still have an essential part to play in creating policy papers, providing crucial quality control that ensures credibility, relevance and legitimacy. Yet, as generative AI tools improve, they could be used to provide first drafts of discrete sections, such as plain-language summaries of technical information or complex legislation.

In one experiment by the publisher Elsevier, an LLM system was constructed that referenced only published, peer-reviewed research. Although the system managed to produce a policy paper on lithium batteries, challenges remain. As others have found[12], the resulting text was bland and pitched at a high level of understanding, mirroring the language in the papers it sourced rather than an original synthesis, and far from the briefs needed. However, this system demonstrated some important design principles. For instance, forcing it to generate only text that refers to scientific sources ensured that the resulting advice credited the scientists who were cited.

Before long, AI tools might make bespoke policy briefs for different audiences. A POSTnote has to be useful for hundreds of politicians, from a wide range of political, professional and social backgrounds. But the UK Parliament holds data on politicians, including their political affiliation, voting record, and educational and professional background, as well as demographic and socio-economic information on constituencies. Next-generation AI tools could deliver automated summaries of politicians' writings and contributions to debates and committee work, as well as tailor scientific briefings for each individual. Furthermore, such tools could leverage the policymakers' previous work as a training data set to present content on science-informed issues in the voice of the policymaker to their constituents, for example (see go.nature.com/3svn2z9).

Say a POSTnote was commissioned by the UK Parliament to summarize the latest research on COVID-19 vaccines. Instead of a single publication, POST could produce a multilayered document that automatically tailored itself to different politicians. For example, a politician might receive a version that highlighted how people in their constituency made contributions to the science of COVID-19 or to vaccine manufacturing. They could be provided with targeted information on infection rates in their own region.

Another dimension might be the level of scientific explanation of how vaccines work. Science-savvy politicians could receive specialist knowledge; those with no scientific background could receive a lay version. The level of technical detail might be dialled up or down by the reader themselves.

Five issues need further thought for drafting policy notes.

Training data and model

Researchers have shown that different language models have distinct political leanings, on both social and economic fronts. Some of these biases are picked up from the data that models are trained on. These biases can then have implications for how models perform on specific tasks, such as detecting hate speech and misinformation[13]. Other forms of bias include race, religion, gender and more.

These issues highlight that AI tools for science advice cannot be black boxes — they will require transparency and participatory design processes. Advisers and policymakers should be involved in the selection of the corpus and the training process to ensure that outputs are perceived as legitimate. Researchers in evidence-based policymaking and AI bias should advise and test the systems before their widespread adoption. The training set and input corpuses need to be carefully vetted by these groups to ensure the quality of the scientific information feeding into the advice,

and thus the credibility of the output. And the algorithms used to assist in the advisory process need to be fully transparent and explainable to ensure accountability.

Governance

Such processes would be best conducted by institutions that have clear mechanisms in place to ensure robust governance, broad participation, public accountability and transparency. For example, national governments could build on current efforts, such as the US What Works Clearinghouse and the UK What Works Network. Alternatively, international bodies, such as the United Nations scientific and cultural organization UNESCO, could develop these tools in alignment with open science goals. Care should be taken to seek international collaboration between countries of all income levels. It is key to ensure not just the availability of these tools and scientific information to low-income countries, but also the consistent development of rigorous, unbiased systems for evidence synthesis that align with national and international policies and priorities.

Disinformation

Concerns have been raised that AI-drafted publications could flood the system and contaminate databases of preprints and journal submissions[14]. The same scenario could be a risk for policy information sourced from AI-based tools. Infusing political debates with biased or fabricated information — presented in a seemingly scientific manner — could create confusion and tip the perception of contested policy issues. Targeting policymakers with disinformation can be an effective strategy to divert attention and cause confusion. Disinformation attacks pose a threat to all types of online system, not just those that provide science advice, and are an increasing focus of research and policy[15]. Ensuring that systems used to produce science advice are not influenced by disinformation or 'data-poisoning' attacks might require greater oversight and understanding of the training data and process. This is a whole-of-sector issue, but in the first instance, coordinating and advisory bodies (such as the US Office of Science and Technology Policy) should work with key research funders to have a catalysing role.

Data privacy

Policy briefings often contain classified or other sensitive information, such as the details of a defence acquisition or draft findings from a public-health study, which needs to remain private until cleared for public dissemination. If advisers use publicly available tools, such as ChatGPT, they might be at risk of disclosing restricted information — a concern that has already complicated AI-model deployment elsewhere in government and in the private sector (see go.nature.com/3rrhm67). Institutions will need to establish clear guidelines about what documents and information can be fed into external LLMs and, ideally, develop their own internal models running on secure servers.

Science-advice professionals will need to be trained in AI-user skills, such as the best way to prompt LLMs to produce the required outputs. Even minor shifts in tone and context in a prompt can alter the probabilities used by the LLM to generate a response. Advisers also need to be trained to avoid inappropriate over-reliance on AI systems — such as when drafting advice on emerging topics for which information is needed rapidly. These might be areas in which LLMs perform poorly, because of a lack of relevant training data. Science advisers will require a nuanced understanding of such risks.

> **"AI literacy will be required for hiring, promotion and professional development of science advisers."**

Next steps

Science advice needs to be scientifically credible, politically legitimate and relevant to the needs of policymakers. And that must remain the case if AI tools are used.

In the short term, as policymakers begin to use available tools — for example, to write speeches or letters to constituents — governments should develop policies to prevent known risks. In the longer term, AI literacy will be required for hiring, promotion and professional development of science advisers.

Collaboration will be needed to build AI tools for science advice in a responsible way. The technical know-how is likely to come from academia and technology companies, whereas demands for robust governance, transparency and accountability can only be met by governments. These kinds of relationship between academia, business and government exist in many areas, including AI initiatives such as the US National Artificial Intelligence Research Resource Task Force.

Such issues should be discussed at forthcoming AI summits at Bletchley Park near Milton Keynes, UK, in November and in London in June 2024. Other events include the G20 Chief Science Advisers' Roundtables, and major conferences such as those hosted by the International Network for Government Science Advice and the European Parliament Technology Assessment network. Early consideration of which sector takes the lead will be important, given the concerns being voiced over regulatory capture and government competence in the broader AI domain. Consensus on this and other issues could be developed at such summits.

We still need old-school intelligence to make the most of the artificial kind.

The authors

Chris Tyler is an associate professor in the Department of Science, Technology, Engineering and Public Policy (STEaPP), University College London, UK. **K. L. Akerlof** is an assistant professor in the Department of Environmental Science & Policy, George Mason University, Fairfax, Virginia, USA. **Alessandro Allegra** is a doctoral candidate in the Department of Science and Technology Studies, University College London, UK, and a policy officer in the Directorate-General for Research & Innovation, European Commission, Brussels, Belgium. **Zachary Arnold** is an analytic lead at the Emerging Technology Observatory, Center for Security and Emerging Technology, Georgetown University, Washington DC, USA. **Henriette Canino** is a PhD candidate in the Department of Science, Technology, Engineering and Public Policy (STEaPP), University College London, UK. **Marius A. Doornenbal** is a distinguished engineer at Elsevier, Amsterdam, the Netherlands. **Josh A. Goldstein** is a research fellow at the Center for Security and Emerging Technology, Georgetown University, Washington DC, USA. **David Budtz Pedersen** is a professor of science communication in the Department of Communication and Psychology, Aalborg University, Copenhagen, Denmark. **William J. Sutherland** is a professor in the Conservation Science Group, Department of Zoology, University of Cambridge, UK.
e-mail: cptyler@ucl.ac.uk

1. Wang, H. *et al. Nature* **620**, 47–60 (2023).
2. Yoong, S. L. *et al. Cochrane Database System. Rev.* **8**, CD013862 (2023).
3. Sutherland, W. J. *et al. Biol. Conserv.* **238**, 108199 (2019).
4. Martin, P. A. *et al. Proc. Natl Acad. Sci. USA* **120**, e2221911120 (2023).
5. Marshall, I. J. & Wallace, B. C. *System. Rev.* **8**, 163 (2019).
6. Dunham, J., Melot, J. & Murdick, D. Preprint at https://arxiv.org/abs/2002.07143 (2020).
7. Sutherland, W. J. *et al. Ecol. Soc.* **19**, 3 (2014).
8. Amano, T., González-Varo, J. P. & Sutherland, W. J. *PLoS Biol.* **14**, e2000933 (2016).
9. Dougherty, M. R. & Horne, Z. *R. Soc. Open Sci.* **9**, 220334 (2022).
10. Burgelman, J. C. *et al. Front. Big Data* **2**, 43 (2019).
11. Horder, J. *Nature Hum. Behav.* **7**, 168 (2023).
12. Retkowski, F. Preprint at https://arxiv.org/abs/2305.04853 (2023).
13. Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. in *Proc. 61st Annu. Meet. Assoc. Comput. Linguist. Vol. 1* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) 11737–11762 (Association for Computational Linguistics, 2023).
14. *Nature Mach. Intell.* **5**, 1 (2023).
15. Schünemann, W. J. in *Cyber Security Politics: Socio-Technological Transformations and Political Fragmentation* (eds Dunn Cavelty, M. & Wenger, A.) 32–47 (Routledge, 2022).

The authors declare no competing interests.
Disclaimer: the views expressed are those of the authors and do not necessarily reflect the views of their employers.