Comment

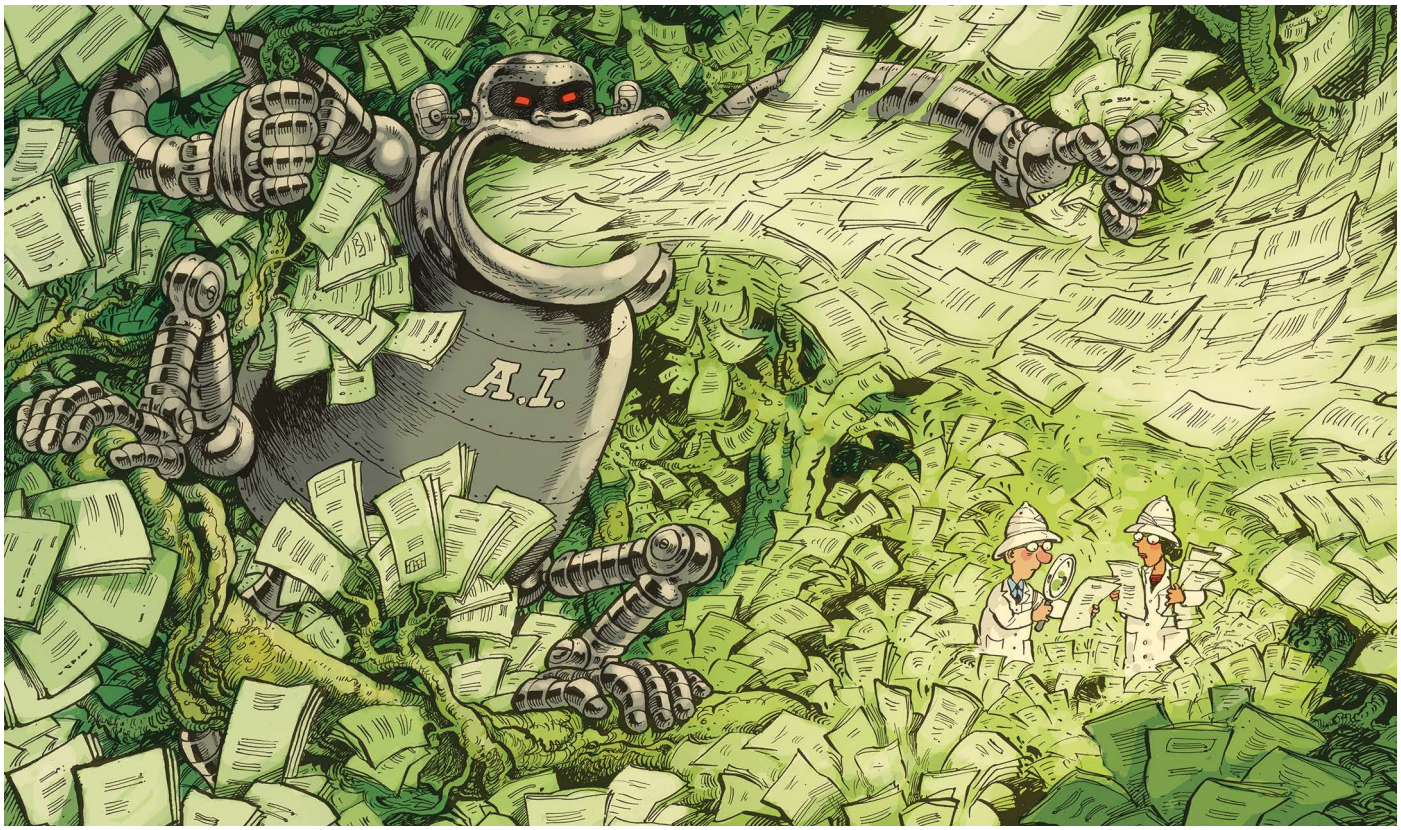


ILLUSTRATION: DAVID PARKINS

Will AI speed up literature reviews or derail them entirely?

Sam A. Reynolds, Alec P. Christie, Lynn V. Dicks, Sadiq Jaffer, Anil Madhavapeddy, Rebecca K. Smith & William J. Sutherland

The publication of ever-larger numbers of problematic papers, including fake ones generated by artificial intelligence, represents an existential crisis for the established way of doing evidence synthesis. But with a new approach, AI might also save the day.

Over the past few decades, evidence synthesis has greatly increased the effectiveness of medicine and other fields. The process of systematically combining findings from multiple studies into comprehensive reviews helps researchers and policymakers to draw insights from the global literature[1]. AI promises to speed up parts of the process, including searching and filtering. It could also help researchers to detect problematic papers[2]. But in our view, other potential uses of AI mean that many of the approaches being developed won't be sufficient to ensure that evidence syntheses remain reliable and responsive. In fact, we are concerned that the deployment of AI to generate fake papers presents an existential crisis for the field.

What's needed is a radically different approach — one that can respond to the updating and retracting of papers over time.

We propose a network of continually updated evidence databases, hosted by diverse institutions as 'living' collections. AI could be used to help build the databases. And each database would hold findings relevant to a broad theme or subject, providing a resource for an unlimited number of ultra-rapid and robust individual reviews.

Adding fuel to the fire

Currently, the gold standard for evidence synthesis is the systematic review. These are comprehensive, rigorous, transparent and objective, and aim to include as much relevant high-quality evidence as possible. They also use the best methods available for reducing bias. In part, this is achieved

by getting multiple reviewers to screen the studies; declaring whatever criteria, databases, search terms and so on are used; and detailing any conflicts of interest or potential cognitive biases.

Yet these reviews require considerable resources. Some studies suggest that Cochrane reviews — systematic reviews of specific topics in health care and health policy that meet internationally recognized criteria for the highest standards in evidence-based health care — generally cost more than US$140,000 and take more than two years to complete[3,4].

It is becoming ever harder for review authors to keep up with the rapidly expanding number of papers. The scientific literature is estimated[5] to have doubled every 14 years since 1952.

Because each reviewer tends to have access to different publications, and because databases are continually updated, systematic reviews are plagued by reproducibility issues. A study published last year concludes that only 1% of reviews report a search strategy that is fully reproducible[6]. Furthermore, many systematic reviews unwittingly cite publications that have been retracted, including those removed from the literature because of methodological or ethical issues and fraud[7].

We agree that AI could be part of the solution to these problems. It could help investigators to conduct reviews more comprehensively and more efficiently — by filtering many more papers, say, or by assessing the entire content of papers instead of just the title and abstract, as human reviewers tend to do as a first step. But one aspect seems to be underappreciated: the degree to which AI — particularly large language models (LLMs) — could exacerbate some of the problems.

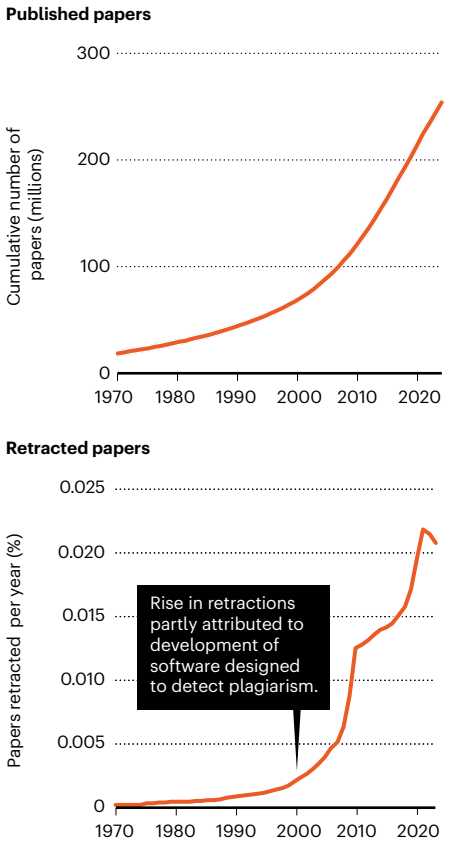
At this point, little is known about how many scientific papers generated entirely by AI are being published. As announced in March, a scientific paper[8] generated by AI Scientist (an AI tool developed by the company Sakana AI in Tokyo and its collaborators) passed peer review for inclusion in a workshop at a key AI meeting. The reviewers did not detect that an AI model had formulated the hypotheses, designed and run experiments, analysed the results, generated the figures and produced the manuscript.

And a preprint posted on arXiv estimates that at least 10% of all PubMed abstracts published in 2024 were written with the help of LLMs, on the basis that an abrupt increase in the frequency of certain words coincided with widespread access to LLMs[9]. That proportion has almost certainly gone up since.

Even if LLMs are used widely, it is difficult to separate cases in which they have been deployed to fabricate papers from those in which authors are simply using them to improve their writing[10]. Yet generative AI is likely to make the production of fake manuscripts easier, irrespective of whether those who use LLMs maliciously do so to further their careers, to manipulate the conclusions of evidence syntheses because of a specific commercial or policy objective or simply to be disruptive. The use of multiple LLMs will also make it more difficult for humans to detect textual fingerprints associated with one particular model.

MORE PAPERS, MORE RETRACTIONS

The publication of fictitious papers generated by artificial intelligence could make it even harder for researchers and policymakers to obtain an accurate picture of the science, be it in medicine, conservation biology or other fields.



SOURCE: DATA FROM OPEN ALEX (HTTPS://OPENALEX.ORG/).

In other words, the use of generative AI is likely to supercharge the already growing problem of paper mills — businesses that sell fake work and authorships to researchers seeking journal publications to boost their careers. It could even replace the paper-mill market, given that fake papers can now be generated in minutes for free.

What to do?

The Campbell Collaboration (a group of researchers and policymakers dedicated to generating evidence syntheses for economic and social policy decisions) and Cochrane already provide guidance on how to identify studies that have raised concerns or that have been retracted[11]. This includes checking studies against the Retraction Watch database, which lists retractions gathered from publisher websites, and using the CENTRAL database, a repository for clinical-trial reports that flags retracted studies[11]. Cochrane guidance also states that the authors of published reviews containing retracted studies should recalculate all results and, while doing so, flag the review with an editorial note or withdraw it and then publish the updated version[11].

Even now, this kind of reanalysis often fails to happen, presumably because the original review authors have limited resources and little incentive. In one assessment of systematic reviews of pharmaceutical compounds tested in clinical trials, retracted papers continued to be cited in 89% of the reviews one year after the review authors had been notified of the retraction[7]. With the ever-increasing production of both legitimate and spurious scientific literature, researchers' ability to maintain an accurate picture of what the data show is likely to be outstripped (see 'More papers, more retractions').

So, what system might enable the continual and rapid removal — at scale — of fraudulent or otherwise problematic papers from databases?

Although not developed with this goal in mind, our work on the Conservation Evidence project — an information resource hosted by the University of Cambridge, UK, to support decisions about how to maintain and restore global biodiversity — has convinced us that a network of AI-enabled, continually updated evidence databases is one possible solution.

As part of this project, all of the authors of this article have been involved in developing subject-wide evidence synthesis. The aim here is to identify literature containing information that is relevant to a broad theme. For the Conservation Evidence project, this is the effectiveness of management actions for biodiversity conservation.

After trawling the literature that presents evidence of an impact (or not) of a management action on some species or habitat, each paper is assessed and the outcome extracted. Newly published literature is scanned and filtered at least once each year, and the database is periodically updated. So far, more than 1.2 million papers written in 17 languages have been screened.

In an extension of this resource, called Metadataset, we have begun to extract the quantitative effect sizes of management-action impacts from the Conservation Evidence database so that users can conduct their own meta-analyses[10]. So far, we have pulled out all of the quantitative data users would need to do meta-analyses for invasive species. These data can be filtered and reanalysed depending on whether the user is interested in a particular species, a particular country or whatever else.

This kind of data set eliminates the need for the available literature to be amassed

and assessed from scratch for every research question that requires a systematic review. Thanks to Metadataset, when policymakers in China asked us in 2023 for information about the effectiveness of management strategies for the invasive grass *Spartina alterniflora*, we were able to send them the results of key meta-analyses within two hours.

An oracle for science

We propose that, with the help of AI, a global network of living evidence databases — building on our prototype — could continually capture new studies and metadata from whichever sources are available to the institutions that host the databases (see 'Making data more accessible'). High-quality systematic reviews, as well as those that are less comprehensive but faster to produce, could then be undertaken as needed, with humans kept in the loop to check that AI-mediated identification and filtering of papers is appropriate and sufficiently accurate.

Automated systems, and especially human users, could flag evidence gaps, errors in papers — including those that have not yet been corrected — and retracted studies. Each review could be associated with a unique identifier, signalling the state of the database when the review was conducted. This should enhance reproducibility and enable researchers and policymakers to determine whether papers in the review have been retracted since the initial search.

Mirroring databases across diverse institutions globally would ensure that institutional or political decisions or technical issues would not block researchers' access to them. Also, statistical confidence levels could be assigned to every decision or step in the process — especially those mediated by AI — and then tracked. This would enable researchers to identify and declare uncertainties, for instance, in cases in which answers are needed fast.

We estimate that building the Conservation Evidence database cost around $8 million over two decades. But our trials over the past 18 months suggest that AI could drastically reduce this upfront cost and enable the establishment of living evidence databases in all sorts of ways.

AI could assess whether papers meet the inclusion criteria and expedite the transfer of key data and metadata from the papers into tables. It could monitor the literature for new relevant studies and remove retracted papers from databases. It might even be able to identify fraudulent papers that have not yet been retracted. Also, multilingual LLMs mean that papers in multiple languages could be assessed and screened. (About 35% of biodiversity papers are not written in English, but few non-English papers are included in systematic reviews[11].)

"The use of generative AI is likely to supercharge the already growing problem of paper mills."

Fresh challenges could arise with the approach we are suggesting, including reproducibility, given the rate at which developers update AI models. But individual systematic reviews, even those conducted faster and more efficiently using AI, will always be constrained by the fact that they are based on narrow questions and written and maintained by a handful of authors[12]. In the longer term, we do not see how this approach can deliver the kind of ultra-fast, high-quality but flexible evidence synthesis that is now possible — and that decision-makers desire.

Conversely, a network of rigorous, transparent and dynamic evidence databases could provide an oracle that accumulates policy-relevant scientific knowledge. This goal is ever more important in a world increasingly awash with misinformation and misconceptions.

Making data more accessible

Whoever builds the 'living' evidence databases we are proposing will need access to the full text of scientific papers to enable effective and robust evidence synthesis.

Some of the literature is available through projects such as the OpenAlex catalogue of scholarly research. But most is still closed access[13].

In developing the Conservation Evidence database at the University of Cambridge, UK, we have been able to access the full texts of closed-access papers and download them in bulk thanks to various negotiations between librarians at the university and publishers. But this approach is expensive and time-consuming. Also, because the computing infrastructure used by publishers is not designed for this level of access, we have encountered endless engineering challenges.

To ensure that high-quality information from closed-access literature or other proprietary data sources is fed into evidence databases, publishers and data holders need to ensure that their publications and data are accessible through standardized formats and application programming interfaces.

A universal system would be a win–win: one that enables content to meet copyright agreements while maximizing the impact of that content in the field and in decision-making.

The authors

Sam A. Reynolds is a postdoctoral research associate in the Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK. **Alec P. Christie** is a research fellow in the Centre for Environmental Policy, Imperial College London, UK. **Lynn V. Dicks** is professor of ecology in the Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK. **Sadiq Jaffer** is a planetary computing fellow in the Department of Computer Science & Technology, University of Cambridge, Cambridge, UK. **Anil Madhavapeddy** is professor of planetary computing in the Department of Computer Science & Technology, University of Cambridge, Cambridge, UK. **Rebecca K. Smith** is a senior research associate in the Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK. **William J. Sutherland** is director of research for the Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK.
e-mail: sar87@cam.ac.uk

1. Connor, L. *et al. Worldviews Evid. Based Nurs.* **20**, 6–15 (2023).
2. Berger-Tal, O. *et al. Trends Ecol. Evol.* **39**, 548–557 (2024).
3. Michelson, M. & Reuter, K. *Contemp. Clin. Trials Commun.* **16**, 100443 (2019); erratum **16**, 100450 (2019).
4. Andersen, M. Z., Gülen, S., Fonnes, S., Andresen, K. & Rosenberg, J. *J. Clin. Epidemiol.* **124**, 85–93 (2020).
5. Bornmann, L., Haunschild, R. & Mutz, R. *Humanit. Soc. Sci. Commun.* **8**, 224 (2021).
6. Rethlefsen, M. L. *et al. J. Clin. Epidemiol.* **166**, 111229 (2024).
7. Bakker, C. *et al. BMJ Evid. Based Med.* **29**, 121–126 (2024).
8. Lu, C. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2408.06292 (2024).
9. Kobak, D., González-Márquez, R., Horvát, E.-A. & Lause, J. Preprint at https://doi.org/10.48550/arXiv.2406.07016 (2024).
10. Berdejo-Espinola, V. & Amano, T. *Science* **379**, 991 (2023).
11. Lefebvre, C. *et al.* in *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5.1 (eds Higgins, J. P. T. *et al*). Ch. 4 (Cochrane, 2025).
12. Andersen, M. Z., Zeinert, P., Rosenberg, J. & Fonnes, S. *Syst. Rev.* **13**, 120 (2024).
13. Piwowar, H. *et al. PeerJ* **6**, e4375 (2018).

The authors declare no competing interests.