Opinion: Leveraging AI to improve evidence synthesis in conservation

Authors

- Oded Berger-Tal*, Mitrani Department of Desert Ecology, Jacob Blaustein Institutes for Desert Research, Ben Gurion University, Israel. https://orcid.org/0000-0002-7396-456X
- Bob B.M. Wong*, School of Biological Sciences, Monash University, Victoria 3800, Australia. https://orcid.org/0000-0001-9352-6500
- Carrie Ann Adams, Department of Fish, Wildlife, and Conservation Biology, Colorado State University, 1474 Campus Delivery, Colorado State University, Fort Collins, CO, USA. https://orcid.org/0000-0001-7381-0555
- Daniel T. Blumstein, Department of Ecology & Evolutionary Biology, University of California, 621 Young Drive South, Los Angeles, CA 90095-1606, USA. https://orcid.org/0000-0001-5793-9244
- Ulrika Candolin, Organismal and Evolutionary Biology Research Programme, University of Helsinki, 00014 Helsinki, Finland. https://orcid.org/0000-0001-8736-7793
- Matthew J. Gibson, Evolution & Ecology Research Centre, Centre for Ecosystem Science, and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia. https://orcid.org/0000-0003-1856-8959
- Alison L. Greggor, Conservation Science and Wildlife Health, San Diego Zoo Wildlife Alliance, Escondido, CA, USA https://orcid.org/0000-0003-0998-618X
- Malgorzata Lagisz, Evolution & Ecology Research Centre, Centre for Ecosystem Science, and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia. https://orcid.org/0000-0002-3993-6127
- Biljana Macura, Stockholm Environment Institute (HQ), Box 24218, Stockholm, 10451, Sweden. https://orcid.org/0000-0002-4253-1390
- Catherine J. Price, School of Life & Environmental Sciences, University of Sydney, NSW 2006, Australia. https://orcid.org/0000-0001-9631-2479
- Breanna J. Putman, Department of Biology, California State University, San Bernardino, CA, United States https://orcid.org/0000-0002-1079-5519

- Lysanne Snijders, Behavioural Ecology Group, Wageningen University & Research, De Elst 1, 6708 WD, Wageningen, The Netherlands. https://orcid.org/0000-0003-0911-3418
- Shinichi Nakagawa, Evolution & Ecology Research Centre, Centre for Ecosystem Science, and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia. https://orcid.org/0000-0002-7765-5182

* Equal contribution

Correspondence: Shinichi Nakagawa (s.nakagawa@unsw.edu.au), Oded Berger-Tal (bergerod@bgu.ac.il), Bob B.M. Wong (bob.wong@monash.edu)

Keywords: artificial intelligence, biodiversity conservation, evidence synthesis, large language models, systematic reviews

Abstract

Systematic evidence syntheses (systematic reviews and maps) summarize knowledge and are used to support decisions and policies in a variety of applied fields, from medicine and public health to biodiversity conservation. However, conducting these exercises in conservation is often expensive and slow, which can impede their use and hamper progress in addressing the biodiversity crisis. With the explosive growth of large language models (LLM) and other forms of artificial intelligence (AI), we discuss the promise and perils associated with their use. We conclude that, when judiciously used, AI has the potential to speed up and hopefully improve the process of evidence synthesis, which can be particularly useful for underfunded applied fields such as conservation science.

Main text

Biodiversity conservation needs evidence synthesis

Biodiversity conservation requires rapid decisions that, ideally, are made with the best available scientific evidence. Systematic **evidence syntheses** (**systematic reviews** and **systematic maps;**see Glossary)—through rigorous, transparent and repeatable methods—are recognized as the gold standard for cataloging, collating and synthesizing the available evidence to support decision making from public health to environmental management and conservation ([1], Box 1). However, conducting systematic evidence syntheses can often be expensive and slow [2]. With the conservation literature growing exponentially, the endeavor can rapidly become unmanageable for human reviewers and irrelevant for managers and policy advisors that look for timely scientific evidence to support their decisions [3]. Several solutions have been proposed for more rapid forms of evidence synthesis (e.g., [1,4,5]), which raises the challenge of potentially having to trade speed of the review process with comprehensiveness or exhaustiveness, thus reducing the reliability of the review findings.

Artificial intelligence (AI) and **machine learning** (especially **deep learning**) tools are currently revolutionizing how evidence is synthesized in biomedical sciences [6]. While there are key differences between biomedicine and conservation research, in this opinion piece, we make the case that AI tools can also dramatically improve evidence syntheses and decision-making for biodiversity conservation. We do so by first highlighting the potential role of AI in biodiversity conservation, and then discussing the benefits and challenges of using AI, especially **large language models (LMMs)** in this field. Because these tools are still in their infancy [7,8], we clarify their role in synthesizing text-based scientific evidence for conservation decision-making, and propose suggestions for responsible and ethical use of AI in conservation science.

Artificial intelligence is revolutionizing conservation science

Artificial Intelligence, initially the realm of science fiction, is now firmly entrenched in our daily lives, and continues to revolutionize the way we interact with each other, our world and even the universe. In conservation science, AI technologies are already extensively and creatively deployed in a myriad of ways for research and management purposes — from AI tools to expose online wildlife trafficking [9] and drones with machine and deep learning capabilities to identify, track and monitor wild animals [10], to the use of interactive robots to understand and control the spread of invasive species [11]. By contrast, using rapidly emerging AI tools, such as LLMs, to allow for more efficient evidence synthesis to support conservation decision-making, holds great potential but is still relatively new.

Machine learning algorithms employ **artificial neural networks** that are trained by large amounts of data (referred to as a corpus). Whereas simple machine learning is an approach to classify and facilitate discrimination between two or more entities, LLMs are able to recognize, summarize, translate, predict and generate text without any training or only a few instructions as a form of **prompts** (known as **zero-shot or few-shot learning**). In the medical sciences, where evidence synthesis methods are well developed and widely used, recent studies demonstrate the promising role that AI tools can play in carrying out rapid and extensive literature reviews [8,12]. At the same time, there is also discourse around potential challenges and limitations regarding the usefulness of these platforms [7,13–15].

Benefits and challenges of using AI for evidence synthesis

Speed

Conservation science is a race against time. Employing AI and LLM tools can reduce the time required to perform systematic evidence syntheses by assisting in various stages of the work [6], including communicating the results to relevant stakeholders [3]. Researchers have shown that the use of LLM tools can substantially shorten, by as much as six-fold, the time spent screening relevant research ([8,12,13], Box 2). LLMs could also be applied to (meta)data extraction from relevant studies and summarize a collection of articles more efficiently [8,16,17]. At present, different AI tools have different limits to the amount of data that can be inputted into them or processed by them. Some free versions of AI tools may be swamped by large screening tasks [17], which could limit their use by funding-restricted conservation agencies. Speed is desirable, but without expert oversight there are likely to be issues with accuracy and reliability by increasing the pace of evidence syntheses (i.e., a human-in-the-loop, HITL process is necessary).

Comprehensiveness, accuracy and reliability

Systematic evidence syntheses aim to reduce human bias in the assessment of scientific evidence, but human biases (e.g., selection and language biases; [18]) and inconsistencies among human reviewers in study selection and data extraction, are known issues in these syntheses [19]. Using LLM tools can assist in reducing these human biases. For example, by improving prompts, Spillias et al. ([13]) were able to increase the accuracy of screening with ChatGPT (reducing type II errors to < 1%). By helping locate potentially useful gray literature sources, which can be a critical source of biodiversity conservation evidence [20,21], LMMs can help further reduce the effects of publication bias on review comprehensiveness, and can act as a second or third non-human reviewer to tackle screening inconsistencies [13].

While AI tools may reduce some human biases, they can introduce errors. LLMs can miss important and relevant articles during screening [8] and, more broadly, the reliability of different AI tools can vary greatly throughout the synthesis process [22, Table 1]. Missing relevant information may be especially problematic in conservation research where the best solutions are often context-dependent [23], which can lead to incorrect management guidance. AI tools may also generate overconfident and potentially erroneous conclusions and create harm in real-world applications [17]. Misinterpretation errors, where text is improperly summarized, creates an improper understanding of the content. Fabrication errors, where a summary includes information not in the original text, refer to a broad class of 'hallucinations' that are well-known outputs from LLMs. Attribute errors relate to any non-key elements in the review question (e.g., the mis-evaluation of the number of interventions or treatments). Thus, substantial human validation of LLM outputs is essential at each stage of review construction (i.e., HITL; [8]).

Complexity

Compounding the problem of reliability, conservation research is characterized by some unique complexities. Specifically, the field is highly heterogeneous, and includes studies that span a variety of ecosystems and species applying a panoply of study designs and dependent variables that can be measured in various ways (c.f., [24]). The field often draws on evidence from many different disciplines, from psychology and physiology to biochemistry and animal behavior. In addition, the language and terminology used in conservation can be highly inconsistent, with many synonyms for similar terms [25]. For example, the terms invasive, introduced, exotic, alien or non-native species, weed, and pest can all have the same meaning, depending on context. Finally, the majority of published conservation research does not test practical, real-world interventions [26]. Evidence producers must therefore make fine-grained decisions about where academic studies are sufficiently solution-oriented or relevant, while trudging through disparate and highly variable gray literature. Such complexities and nuances need to be taken into account in developing search prompts, screening and oversight of results, and when models are updated to ensure reliability and accuracy of results generated by LLMs [27]. However, robust methods for dealing with such complexities are yet to be developed.

Relevance over time

The evidence base for conservation is rapidly accumulating and evidence syntheses can quickly become outdated. In a rapidly changing world, the effectiveness of interventions might also change with time. Thus, systematic reviews that are not regularly updated may lead to significant inaccuracies over time [28,29]. **Living systematic reviews** have been developed to provide high quality, up-to-date online summaries that incorporate relevant new evidence as it becomes available [28,30]. Such reviews require continuous work and a level of commitment that is often hard to achieve. Here, LLMs can be used to support living reviews and ensure that the evidence-base remains up to date with minimal human effort [30,31]. However, as the outputs of LLMs may change over time (because the algorithms and training sets change), their performance will require human evaluation.

Inclusivity

In our view, one of the major benefits of using LLMs in synthesis is their ability to find conservation evidence from across the globe, particularly in languages other than English [32]. Most of the world's remaining biodiversity is found in the Global South, yet most scientific evidence to inform decision-making comes from authors in the Global North and is published in English [33]. Local studies from the Global South are often missed or discarded from reviews if they are not written in English.

By translating languages, AI tools can make all stages of the review process more inclusive (Box 1). For example, a review on community-based fisheries management focusing on the Pacific Islands [13] benefited from AI rapidly providing a list of non-English relevant terms to be integrated into the search string and yielding additional articles not previously identified by the original search. AI-suggested terms should, however, be checked by proficient speakers of the language in question before inclusion in the search string.

Nevertheless, it is important to emphasize that AI tools require accessible digitized information. Moreover, the original training to create LLMs requires sufficiently large data sets that currently exclude most of the world's languages [34,35]. Therefore, exclusively relying on AI for information means that some traditional and local knowledge may be ignored. This process could reduce the effectiveness of conservation interventions at the local scale and widen the divide between conservation agencies and local communities [36]. In this respect, we emphasize that effective conservation work relies just as heavily on building strong relationships with the relevant stakeholders as using the most accurate scientific evidence (e.g., [37]). The use of AI may alienate local collaborators if not conveyed and properly communicated to all stakeholders and rightsholders.

Ethical considerations

The question, in our view, is not whether AI tools will/should be used in conservation science (the singularity is nigh!), but rather how they are used. Issues of data privacy and informed consent created by emerging AI technologies can be exacerbated through their use in systematic evidence syntheses. People may not wish that their published data are used for AI training, or repurposed and applied to new problems. In this regard, continuous effort to actively engage various stakeholders in the synthesis process is even more crucial in the context of AI application to evidence synthesis.

A well-recognized concern with using AI is the presence of (algorithmic) biases that result from factors such as the unknown data quality and representativeness in training corpus [38,39]. As previously discussed, it is likely that documents written in English and from developed countries form the bulk of the training corpus — this may limit the nature of responses to specific queries and enhance existing biases. Therefore, there is an urgent need for culturally sensitive multi-lingual LLMs [40]. Moreover, in the current LLM landscape there is a lack of transparency around algorithm development and reporting related to decisions algorithms make during the review process. Lack of transparency leads to limited peer scrutiny and accountability in AI-supported evidence syntheses and prevents equitable and responsible development of AI.

Hence, the best practice moving forward is to be explicitly clear about how AI is being used in evidence syntheses, which may include detailed reporting of the prompts and instructions given to an LLM and how it was tested for replicability and reliability. This ensures transparency and reproducibility to some extent. Repeatability can be limited because models are probabilistic and constantly updated with new data. Thus, multiple runs of the same model over time may produce different responses. This is a challenge that requires future research to fully understand its impact on evidence synthesis and, ultimately, on conservation management decisions.

Concluding remarks

AI is not a silver bullet and conducting a reliable evidence synthesis requires a lot of work and will always be time-consuming and require attention to detail (Box 3). However, AI tools can help improve the location and consideration of gray literature and evidence in a variety of languages that were not traditionally included in syntheses. AI may make evidence synthesis faster, more accessible, and inclusive to a greater number of researchers. Although decision-making in conservation involves more than just scientific evidence, expanding the availability of the information base will increase opportunities for developing informed policies and management actions (see Outstanding Questions).

More broadly, while we have focused on how AI tools can be used to synthesize evidence for biodiversity conservation, we suggest that ecologists and evolutionary biologists, more broadly, can also benefit from using these tools to efficiently identify the state of knowledge in their respective disciplines.

Acknowledgements: This paper emerged from a workshop conducted at Monash University's Prato Centre, partially supported by Monash University and Ben Gurion University of the Negev (to OB-T and BBMW) through a grant from the Pratt Foundation. We thank the staff at the Prato Centre for their wonderful hospitality and support. In addition, we acknowledge the following funding agencies for financial support: the Australian Research Council (FT190100014 and DP220100245 to BBMW, DP210100812 and DP230101248 to ML and SN, and DE220101316 to CP), a NASA Biodiversity grant (#80NSSC21K114 to CA), the Swedish Cultural Foundation in Finland (Nr 179446 to UC), and the Netherlands Organisation for Scientific Research (VI.Veni.192.018 to LS).

AI disclosure: We used Chat GPT 4.0, accessed through Microsoft Edge, to develop a list of benefits of AI for decision making for conservation. We largely ignored the specific list and wrote this paper collectively using purely human-synthesized knowledge. Chat GPT 4.0 was also used as a resource to help understand key ideas and tools used in this rapidly growing field. The authors are fully responsible for the content contained in this manuscript.

Declaration of interests

No interests are declared.

References

1. Collaboration for Environmental Evidence (2022) Guidelines and Standards for Evidence Synthesis in Environmental Management. Version 5.1. (Pullin, A.S. *et al.,* eds). URL: www.environmentalevidence.org/information-for-authors
2. Haddaway, N.R. and Westgate, M.J. (2019) Predicting the time needed for environmental systematic reviews and systematic maps. *Conserv. Biol.* 33, 434–443
3. Tyler, C. *et al.* (2023) AI tools as science policy advisers? The potential and the pitfalls. *Nature* 622, 27–30
4. Haby, M.M. *et al.* (2016) What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. *Health Res. Policy and Syst.* 14, 83
5. Sutherland, W.J. and Wordley, C.F.R. (2018) A fresh approach to evidence synthesis. *Nature* 558, 364–366
6. Jimenez, R.C. *et al.* (2022) Machine learning computational tools to assist the performance of systematic reviews: A mapping review. *BMC Med. Res. Methodol.* 22, 322
7. Qureshi, R. *et al.* (2023) Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst. Rev.* 12, 72
8. Blaizot, A. *et al.* (2022) Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Res. Synth. Methods* 13, 353–362
9. Cardoso, A.S. *et al.* (2023) Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. *Biol. Conserv.* 279, 109905
10. Couzin, I.D. and Heins, C. (2023) Emerging technologies for behavioral research in changing environments. *Trends Ecol. Evol.* 38, 346–354
11. Polverino, G. *et al.* (2022) Ecology of fear in highly invasive fish revealed by robots. *iScience* 25, 103529
12. van Dijk, S.H.B. *et al.* (2023) Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open* 13, e072254
13. Spillias, S. *et al.* (2023) Human-AI Collaboration to Identify Literature for Evidence Synthesis. *Research Square Preprint*. DOI: 10.21203/rs.3.rs-3099291/v1
14. Zhu, J.-J. *et al.* (2023) ChatGPT and Environmental Research. *Environ. Sci. Technol.* 57, 17667–17670
15. Demszky, D. *et al.* (2023) Using large language models in psychology. *Nat. Rev. Psychol.* 2, 688–701
16. Shaib, C. *et al.* (2023) Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success). *Proc. Conf. Assoc. Comput. Linguist. Meet.* 2, 1387–1407 DOI: 10.18653/v1/2023.acl-short.119
17. Tang, L. *et al.* (2023) Evaluating large language models on medical evidence summarization. *npj Digit. Med.* 6, 1–8
18. Frampton, G. *et al.* (2022) Principles and framework for assessing the risk of bias for studies included in comparative quantitative environmental systematic reviews. *Environ Evid* 11, 12
19. Felson, D.T. (1992) Bias in meta-analytic research. *J. Clin. Epidemiol.* 45, 885–892
20. Haddaway, N.R. and Bayliss, H.R. (2015) Shades of grey: Two forms of grey literature important for reviews in conservation. *Biol. Conserv.* 191, 827–829
21. Amano, T. *et al.* (2023) The role of non-English-language science in informing national biodiversity assessments. *Nat. Sustain.* 6, 845–854
22. Zhao, Z. *et al.* (2021) Calibrate Before Use: Improving Few-shot Performance of Language Models. *Proc. Intern. Conf. Mach. Learn.* 139, 12697–12706
23. Christie, A.P. *et al.* (2020) Poor availability of context-specific evidence hampers decision-making in conservation. *Biol. Conserv.* 248, 108666

24. Christie, A.P. *et al.* (2020) Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. *Nat. Commun.* 11, 6377
25. Cheng, S.H. *et al.* (2018) Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conserv. Biol.* 32, 762–764
26. Williams, D.R. *et al.* (2020) The past and future role of conservation science in saving biodiversity. *Conserv. Lett.* 13, e12720–e12720
27. Clusmann, J. *et al.* (2023) The future landscape of large language models in medicine. *Commun. Med.* 3, 141
28. Brooker, J. *et al.* (2019) Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode [Online]. https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf
29. Shojania, K.G. *et al.* (2007) *Updating Systematic Reviews*, Agency for Healthcare Research and Quality (US), Rockville, MD. AHRQ Publication No. 07-0087.
30. Elliott, J.H. *et al.* (2014) Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLOS Medicine* 11, e1001603
31. Shackelford, G.E. *et al.* (2021) Dynamic meta-analysis: a method of using global evidence for local decision making. *BMC Biol.* 19, 33
32. Amano, T. *et al.* (2021) Tapping into non-English-language science for the conservation of global biodiversity. *PLoS Biol.* 19, e3001296
33. Christie, A.P. *et al.* (2021) The challenge of biased evidence in conservation. *Conservation Biology* 35, 249–262
34. Joshi, P. *et al.* (2021) The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv Preprint* DOI:10.48550/arXiv.2004.09095
35. Ranathunga, S. and de Silva, N. (2022) Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. *Proc. Conf. Asia-Pacific Assoc. Comput. Linguist. Joint Conf. Nat. Lang. Process. (Volume 1: Long Papers)*, pp. 823–848 URL: https://aclanthology.org/2022.aacl-main.62
36. Droz, L. *et al.* (2023) Multilingualism for pluralising knowledge and decision making about people and nature relationships. *People Nat.* 5, 874–884
37. Chaplin-Kramer, R. *et al.* (2023) Transformation for inclusive conservation: evidence on values, decisions, and impacts in protected areas. *Curr. Opin. Environ. Sustain.* 64, 101347
38. Hovy, D. and Prabhumoye, S. (2021) Five sources of bias in natural language processing. *Lang. Linguist. Compass* 15, e12432
39. Chen, Y. *et al.* (2023) Human-Centered Design to Address Biases in Artificial Intelligence. *J. Med. Internet Res.* 25, e43251
40. Ramesh, K. *et al.* (2023) Fairness in Language Models Beyond English: Gaps and Challenges. *arXiv Preprint* DOI: 10.48550/arXiv.2302.12578
41. Fan, W. *et al.* (2023) Recommender Systems in the Era of Large Language Models (LLMs). *arXiv Preprint* DOI: 10.48550/arXiv.2307.02046
42. O'Donoghue, O. *et al.* (2023) BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology. *arXiv Preprint* DOI: 0.48550/arXiv.2310.10632
43. Khraisha, Q. *et al.* (2023) Can large language models replace humans in the systematic review process? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. arXiv Preprint. DOI: 10.48550/arXiv.2310.17526
44. Michelson, M. *et al.* (2020) Artificial Intelligence for Rapid Meta-Analysis: Case Study on Ocular Toxicity of Hydroxychloroquine. *J. Med. Internet Res.* 22, e20007
45. Valizadeh, A. *et al.* (2022) Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Med. Res. Methodol.* 22, 160

46. Chen, L. *et al.* (2023) How is ChatGPT's behavior changing over time? *arXiv Preprint* DOI: 10.48550/arXiv.2307.09009
47. Koehler, M. and Sauermann, H. (2023) Algorithmic Management in Scientific Research. *SSRN Journal* DOI: 10.2139/ssrn.4497871
48. Bannach-Brown, A. *et al.* (2019) Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst. Rev.* 8, 23
49. Hill, J.E. *et al.* (2023) Methods for using Bing's AI-powered search engine for data extraction for a systematic review. *Res. Synth. Methods* DOI: 10.1002/jrsm.1689
50. Waffenschmidt, S. *et al.* (2023) Increasing the efficiency of study selection for systematic reviews using prioritization tools and a single-screening approach. *Syst. Rev.* 12, 161
51. Syriani, E. *et al.* (2023) Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. *arXiv Preprint* DOI: 10.48550/arXiv.2307.06464
52. Jardim, P.S.J. *et al.* (2022) Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Med. Res. Methodol.* 22, 167
53. Marshall, I.J. *et al.* (2017) Automating Biomedical Evidence Synthesis: RobotReviewer. *Proc Conf. Assoc. Comput. Linguist. Meet.* 2017, 7–12
54. Gates, A. *et al.* (2018) Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the RobotReviewer machine learning tool. *J. Clin. Epidemiol.* 96, 54–62
55. Tsafnat, G. *et al.* (2014) Systematic review automation technologies. *Syst. Rev.* 3, 74
56. Gates, A. *et al.* (2021) Creating efficiencies in the extraction of data from randomized trials: a prospective evaluation of a machine learning and text mining tool. *BMC Med. Res. Methodol.* 21, 169
57. Mutinda, F.W. *et al.* (2022) Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer. *BMC Med. Inform. Decis. Mak.* 22, 158
58. West, R. *et al.* (2023) Using machine learning to extract information and predict outcomes from reports of randomised trials of smoking cessation interventions in the Human Behaviour-Change Project. *Wellcome Open Res.* 8, 452 DOI: 10.12688/wellcomeopenres.20000.1
59. Ho, I.M.K. *et al.* (2023) Using Machine Learning Algorithms to Pool Data from Meta-Analysis for the Prediction of Countermovement Jump Improvement. *Int. J. Environ. Res. Public Health* 20, 5881
60. Xu, J.-L. *et al.* (2022) Combining machine learning with meta-analysis for predicting cytotoxicity of micro- and nanoplastics. *J. Hazard. Mater. Adv.* 8, 100175
61. Marshall, I.J. and Wallace, B.C. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst. Rev.* 8, 163
62. Huang, J. and Tan, M. (2023) The role of ChatGPT in scientific communication: writing better scientific review articles. *Am. J. Cancer Res.* 13, 1148–1154
63. van de Schoot, R. *et al.* (2021) An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* 3, 125–133
64. Lombaers, P. *et al.* (2023) Reproducibility and Data storage Checklist for Active Learning-Aided Systematic Reviews. *PsyArXiv Preprint* DOI: 10.31234/osf.io/g93zf
65. Dicks, L.V. *et al.* (2014) Organising evidence for environmental management decisions: a '4S' hierarchy. *Trends Ecol. Evol.* 29, 607–613
66. Orgeolet, L. *et al.* (2020) Can artificial intelligence replace manual search for systematic literature? Review on cutaneous manifestations in primary Sjögren's syndrome. *Rheumatology* 59, 811–81967.
67. Ouzzani, M. *et al.* (2016) Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 5, 210

68. Adams, C.A. *et al.* (2021) Effects of artificial light on bird movement and distribution: a systematic map. *Environ. Evid.* 10, 37
69. Ali, S. *et al.* (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* 99, 101805
70. Turpin, M. *et al.* (2023) Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv Preprint*. DOI:10.48550/arXiv.2305.04388
71. Shi, F. *et al.* (2023) Large Language Models Can Be Easily Distracted by Irrelevant Context. *Proc. Intern. Conf. Mach. Learn. 202, 31210-31227*
72. O'Dea, R.E. *et al.* (2021) Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension. *Biol. Rev. Camb. Phil. Soc.* 96, 1695–1722
73. Haddaway, N.R. *et al.* (2018) ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environ. Evid.* 7, 7
74. Susnjak, T. (2023) PRISMA-DFLLM: An Extension of PRISMA for Systematic Literature Reviews using Domain-specific Finetuned Large Language Models. *arXiv Preprint* DOI: 10.48550/arXiv.2306.14905

Glossary

Artificial intelligence (AI): a machine or model which can perform what appears to require human intelligence. It also refers to a branch of computer science dedicated to creating these models. Recently, generative AI has gained much attention with its ability to create text, images, audio and other media.

Artificial neural networks: a method used in machine learning whereby the connections and strength of connections between a set of nodes (which are modeled after neurons in the brain) is iteratively modified to maximize some desired output (e.g., a discrimination). Originally, these networks had several layers of nodes between input and output but deep learning models have many layers of nodes.

Evidence syntheses: involve a process of combining information from multiple studies on a specific topic and to inform decision making. The term is also used as an umbrella term for the family of reviews that include systematic reviews, systematic maps, rapid reviews, and reviews of reviews.

Deep learning: a type of machine learning that relies on multiple layers of connected nodes whose connections and weights are iteratively modified so as to maximize their ability to make discriminations or identifications. It requires a huge amount of training.

Human-in-the-loop (HITL): a process of model development in machine learning where humans play an interactive and iterative role.

Large language model (LLM): a type of generative artificial intelligence created by a deep learning, neural network trained on a large written corpus that can "understand" human language and generate responses to specific queries.

Living systematic reviews: systematic reviews that are continuously updated that incorporate new evidence as it is produced.

Machine learning: a process by which data are fed into neural network models which are iteratively modified without specific instructions that permit the identification of patterns in data.

Prompts: specific inputs or instructions to a LLM designed to elicit an answer. The growing field of prompt engineering studies the characteristics of effective prompts which in general should be specific, and constrained. Creating a role ('you are a fastidious researcher conducting a systematic review…') can help improve output accuracy.

Systematic review: a formal and highly structured process to comprehensively, rigorously and transparently collate and synthesize evidence, including the academic and gray literature sources. Can be used to support policy formation and biodiversity management decisions.

Systematic map: comprehensive catalogues of the literature on a broad topic of interest. Systematic maps follow the same step-wise process as systematic reviews, but they tackle broader questions, and their final output is a narrative report and a searchable catalogue of the literature that can be used to identify areas where evidence is lacking or is under-represented (knowledge gaps), or areas with sufficient evidence to conduct full synthesis (knowledge clusters) **Zero-shot or few-shot learning**: a direct query to an existing LLM is referred to as a zero-shot query where the results of zero-shot queries are based entirely on the information already contained in the LLM. By contrast, few-shot learning requires some additional data, for instance, where the LLM is provided a list of papers that, based on their title and abstract, that should be included or excluded from a systematic review.

Table 1: AI tools and platforms for evidence synthesis [a]

Stage of synthesis	Example tools and Platforms [b]	Opportunities	Potential challenges and considerations
Identify and formulate review questions	• Gemini (Google DeepMind; https://gemini.google.com/) • Scite (scite; https://scite.ai/)	Facilitate question formulation through assistance with brainstorming and refinement [7]	Some stakeholders might feel disengaged or excluded by the process, potentially hampering innovation and even reinforcing existing biases [7,41]
Draft review protocol	• Gemini (Google DeepMind; https://gemini.google.com/) • ChatGPT (OpenAI, https://chat.openai.com)	Assist in creating a good initial outline and, hence, speeding up the process for protocol writing [7,42]	Risk of 'hallucinations' may cast doubt on protocol accuracy [16,17]; Protocol may lack details and/or correct references [16]
Search for evidence	• Elicit (Elicit; https://elicit.com/) • Scite (scite; https://scite.ai/) • Consensus (Consensus; https://consensus.app/) • Scispace (PubGenius Inc; https://typeset.io/) • ConnectedPapers (Connected Papers; https://www.connectedpapers.com/) • Inciteful (Weishun, M. 2024; https://inciteful.xyz/) • Litmaps (Litmaps Ltd; https://www.litmaps.com) • Gemini (Google DeepMind; https://gemini.google.com/) • ChatGPT (OpenAI, https://chat.openai.com)	Help with suggesting and finding a variety of gray literature sources, including in different languages [43]; Suggest alternative terms for the search [7]; Help to incorporate evidence as it becomes available [44]	Inconsistent and incomplete search terms that can reduce search efficiency and increase the potential for selection bias [45]; Changes to the algorithm may change search results [7,46]; Search results may be probabilistic, erroneous, and not repeatable [7]; Can only make use of digitized knowledge [47]

Include relevant studies	• Rayyan (Ouzanni et al. 2016; https://www.rayyan.ai/) • Abstrackr (Brown University; http://abstrackr.cebm.brown.edu/account/login) • DistillerSR DistillerSR Inc; https://www.distillersr.com/) • EPPI-Reviewer (EPPI Centre; eppi.ioe.ac.uk/EPPIReviewer-Web) • SWIFT-Active Screener (Sciome; https://www.sciome.com/swift-activescreener/) • ASReview (ASReview Lab; https://asreview.nl/) • Silivi (A-Evidence ApS; https://www.silvi.ai/)	Substantially reduce screening time [Box 2]; In the case of double screening, act as the second reviewer to tackle screening inconsistencies [48,49]	May inadvertently pass on relevant studies [50,51]; Changes to the algorithm may change screening results [7,46]; Lack of transparency around algorithm development and decision-making [52]; Screening decisions may be probabilistic and not repeatable [7]
Critically appraise studies	• RobotReviewer [53] (https://www.robotreviewer.net/) • Elicit (Elicit; https://elicit.com/)	Speed up an otherwise very time-consuming process [53,54]	Difficulties in dealing with more complex and diverse study designs and different reporting styles [55]; Interpretation and extraction errors [16,56]; Lack of transparency around algorithm development and decision-making [52]

Extract data	• Scispace (PubGenius Inc; https://typeset.io/) • RobotReviewer [53] (https://www.robotreviewer.net/) • SWIFT-Review (Sciome; https://www.sciome.com/swift-review/) • Silivi (A-Evidence ApS; https://www.silvi.ai/) • ExaCT (https://exact.cluster.gctools.nrc.ca/ExactDemo/intro.php) • Elicit (Elicit; https://elicit.com/)	Efficient at extracting data and metadata (e.g. moderators and study descriptors) [53,57]	Difficulties in dealing with more complex and diverse study designs and different reporting styles [53,55,57]; Interpretation and extraction errors [16,56]; Lack of transparency around algorithm development and decision-making [52]; May not be reliable in obtaining effect sizes [58]
Synthesize data/study findings	• ChatGPT (OpenAI, https://chat.openai.com) • Gemini (Google DeepMind; https://gemini.google.com/)	Potentially efficient at running simple quantitative syntheses (meta-analysis) of evidence as well as narratively synthesizing study findings [59,60]	Sophisticated quantitative (e.g. meta-regression) synthesis is still difficult to conduct [59,61]
Report findings	• ChatGPT (OpenAI, https://chat.openai.com) • Scispace (PubGenius Inc; https://typeset.io/)	Efficient at scientific communication as it can assist scientists in improving their writing style by analyzing text and provide suggestions for improvements [14,62]	Lack of transparency around algorithm development and decision-making [63,64]

[a] We highlight both opportunities, as well as potential challenges and considerations. In regard to the latter, many of the challenges we have identified can be resolved by having humans-in-the-loop and greater procedural transparency. Stages of synthesis mirror those outlined in Figure I in Box 1.

[b] A non-exhaustive list with an emphasis on new and popular platforms

Box 1: Evidence hierarchy for decision support in conservation with AI

For scientific evidence to be useful or usable, information must be distilled, amalgamated and translated from a large collection of individual studies to an output that can inform decision-making. Figure I illustrates how different types of knowledge, information and expert opinions, primary (individual studies) and secondary research (e.g., systematic reviews and review of reviews) feed into decision support systems (i.e., tools that provide different scenarios and logical sets of steps to assist with decision making; [65]). Outputs from these systems help create evidence-informed advice and guides. The pyramid demonstrates how, at each step, the scientific evidence gradually becomes more “condensed” and hence more accessible to conservation decision makers.

Each step of evidence synthesis could potentially be supported and expedited by AI and LLMs, including: 1) question formulation, 2) protocol generation, 3) literature search, 4) screening to select relevant papers (including deduplication), 5) critical appraisal of included studies, 6) data extraction, 7) synthesizing information and 8) transparent reporting (Figure I). Recently, Jimenez and colleagues ([6]) identified 63 machine-learning tools for systematic evidence syntheses. They showed that most of the currently available tools primarily support the three review stages: searching, screening and data extraction. For example, *BIBOT* uses keywords to search and retrieve relevant papers from PubMed [66], while *Rayyan* facilitates screening by reordering papers in the order of relevance, learning from included and excluded papers [67] (also Box 2). None of the tools in their review used LLMs, but LLMs can immediately be used in these three stages and more. A generative AI platform, *Elicit* (elicit.com), for instance, can extract information and summarize pdf documents.

In addition, LLMs can facilitate “Summaries” turning long academic documents (such as systematic reviews) into distilled key messages for policy and practice. Furthermore, LLMs can help create algorithms and software for decision support systems [3].

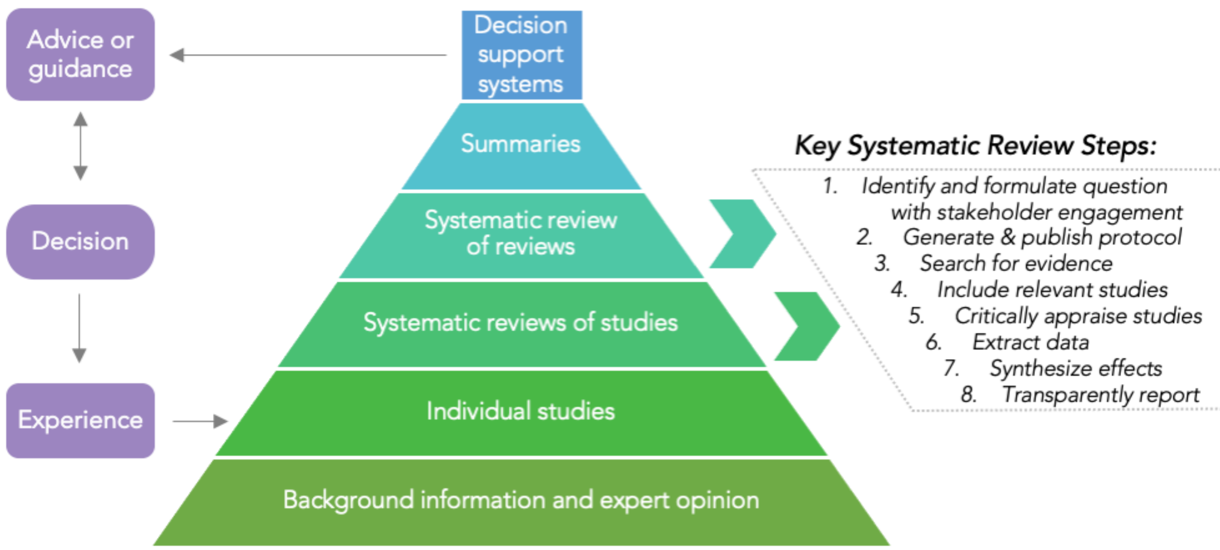

Figure I: Hierarchy of scientific evidence used in conservation decision-making (Modified and redrawn from Dicks et al. [65]).

Box 2: Speeding up screening with AI: a case study

There are a number of AI-assisted article screening tools, most of which use re-ordering algorithms that learn from included/excluded articles as researchers screen based on title and abstract. More recently, large language models (LLMs) have been suggested to be used for such screening [13]. We tested both types: Rayyan.ai (re-ordering algorithm) and GPT 3.5 (LLM) to screen 11,270 article search records from the Web of Science for relevance to the question: *how does artificial light affect bird movement and distribution?* These articles were manually screened by Adams *et al.* [68] (Figure I).

Rayyan.ai's relevance ratings could have reduced the manual screening burden at the title/abstract level by over 80%, with accuracy comparable to a human-alone screening. We provided initial training data by classifying 46 articles we knew to be relevant as "include" and classified 46 additional articles as "exclude". Rayyan computed relevance ratings for the remaining articles, and we sorted them by relevance and screened the first 100. We then recomputed the ratings, re-sorted the records, and screened the next 100 articles. We repeated the process until no additional relevant articles were found, which occurred at ~ 2,200 articles. This method identified 169 (97%) out of 174 relevant articles in the screening dataset after screening less than 20% of the articles. Notably, this process yielded 5 articles missed by a human screener during the original screening process, meaning that the human-alone and this AI-assisted method (Rayyan.ai) had equivalent false negative rates in this case (2.9%).

For GPT 3.5, we used the following prompt "*Classify the given research paper as worthy of inclusion or exclusion… The paper should be classified as "include" or "exclude". You are a careful and thorough researcher conducting a systematic review of the effect of artificial light on bird movement and distribution. Given a title and an abstract of a research paper, your task is to determine whether the paper meets the criteria for inclusion in a review study.*". Following this message, this prompt also included the published abstract along with screening criteria. For the initial run (i.e. zero-shot learning) it retrieved 66 of 215 relevant articles (30%). For the second run, we provided 46 included and excluded articles, and GPT 3.5 was able to retrieve 200 out of 215 (93%) articles. It took 2.5 hours for each run to screen 11,270 articles.

Figure I: Many studies have investigated the relationship between artificial light at night and bird movements (credit: JoshuaWoroniecki)

Box 3: Guiding principles for responsible AI use in evidence syntheses for conservation

Acceptable practices of using AI are evolving rapidly. For example, AI has been used to improve writing for years (many already use Grammarly or Microsoft Grammar Checker) but some publishers currently limit or prohibit LLM-produced text from being used in papers. With this state of flux in mind, we make the following recommendations (Figure I).

First, while AI tools offer considerable promise, use them cautiously. We do not currently understand, in various contexts, its precision, accuracy, specificity, or reliability and the developers themselves are unclear about how some AI tools and models work [69]. As these tools are applied to specific conservation issues, effort will have to be allocated to estimate these sources of error and optimize algorithms [70,71].

Second, view AI tools as a research assistant—it is essential to keep humans as supervisors of AI decision-making (i.e., human-in-the-loop). In the context of systematic evidence syntheses, validate AI decisions against established evidence synthesis standards and guidelines for conduct and reporting (e.g., [1,72,73]).

Third, at the moment, AI is more reliable in some evidence synthesis steps (such as title and abstract screening, and to some extent search strategy design and full-text screening) than others (such as data extraction and critical appraisal). To prevent relevant omissions for search strategy and screening supported by AI, there is a need for detailed scoping exercise that will test all phases of the review before it is conducted.

Finally, we urge AI developers to provide decision files that facilitate the scrutiny of AI algorithms, because transparency is crucial (e.g., see ASReview AI software, [63]), and we should make decision data files accessible [12]. The evidence synthesis community urgently needs a guide for reporting of AI-supported reviews (e.g., PRISMA extension PRISMA-DFLLM for LLM; [74]). Such transparency will help with trust building between evidence producers and evidence users.

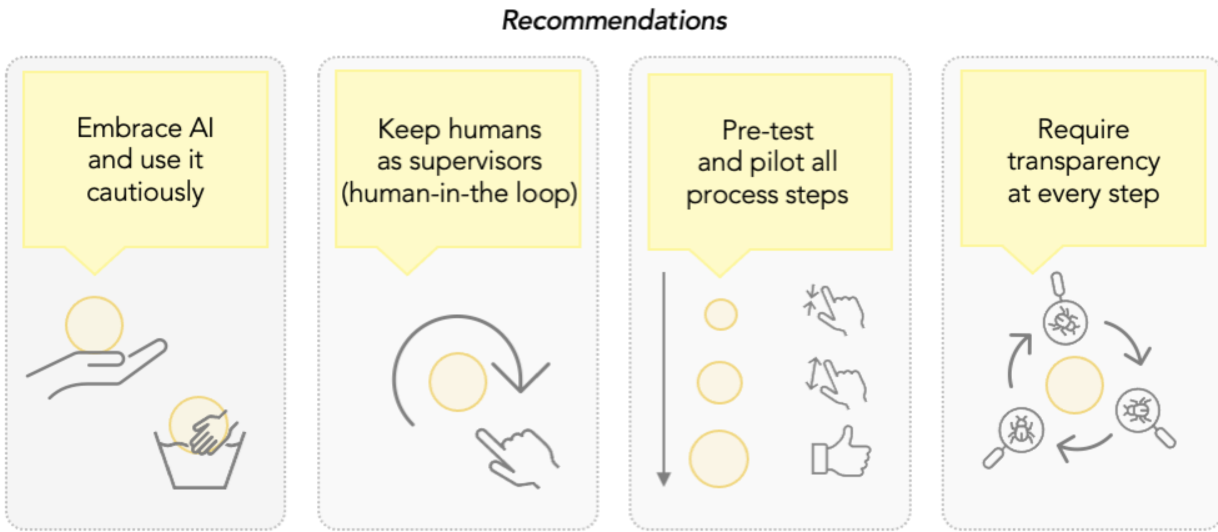

Figure I: Recommendations for responsible AI use for evidence synthesis in conservation.