The EBM of AI and the AI of EBM

EBHC, Taormina, October 2026

John P.A. Ioannidis, MD, DSc

George E and Lucy Becker Professor of Medicine
Professor of Epidemiology and Population Health
Professor (by courtesy) of Biomedical Data Science
Co-Director, Meta-Research Innovation Center at Stanford (METRICS)
Stanford University

Pre-emptive comments: The EBM of AI

- Artificial intelligence is revolutionizing both medicine and biomedical research at the level of discovery, translation, evaluation, evidence synthesis, and dissemination – along with many other related disciplines that are affected, elevated, or dismantled.
- EBM needs to rise to the challenge to appraise the evidence behind these transformations, including the benefits, harms, trade-offs of interventions, diagnostic and predictive tools, and policy where AI is involved.
- The landscape is convoluted and complex and much of the supporting evidence may often be unavailable, as start-ups and AI companies may operate outside the traditional scientific literature in trying to disrupt current medical practices and biomedical research.

Pre-emptive comments: The AI of EBM

- Concurrently, AI tools are increasingly involved in the performance of classic EBM tasks including evidence appraisal, synthesis, dissemination (including misinformation), and provision of guidance.
- The benefits and limitations of such tools need careful study and comparative assessment.
- There is clearly a lot of interesting leads and some major successes, but the greater landscape is more uncertain.
- Many different futures may be imagined and may even coexist.

Science: a current bird's eye view

- Over 50 million authors have published at least one scientific paper, ~10 million published something in the last 12 months
- Over 200 million scholarly documents published, ~8 million published in the last 12 months, most of them entirely false and/or biased
- Over 1 million evidence syntheses published, ~100,000 published in the last 12 months, most of them poorly done

Ioannidis, et al. EJCI 2013; Ioannidis et al; PLoS Biology 2019

My guess for the 8 million papers published in the last 12 months

- AI-heavy, fraudulent = 0.5-1 million
- Not AI-heavy, fraudulent = 0.2-0.5 million
- AI-heavy, of little value or heavily biased = 1-2 million
- Not AI-heavy, of little value and/or heavily biased = 5-7 million
- AI-heavy, of substantive value <0.1 million
- Not AI-heavy, of substantive value <0.1 million

Much (most?) of the scientific universe may be unpublished dark matter and not pre-specified dark energy

5% ordinary matter, 27% dark matter, and 68% dark energy

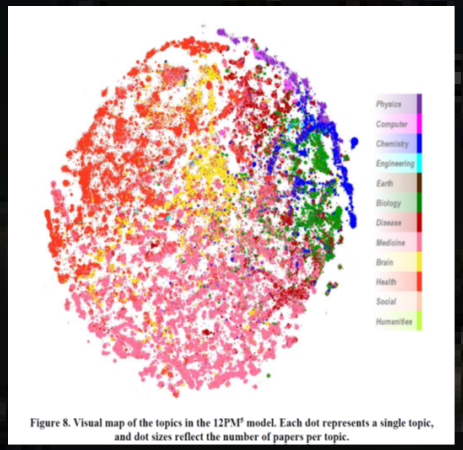


Figure 8. Visual map of the topics in the 12PM² model. Each dot represents a single topic, and dot sizes reflect the number of papers per topic.

Which part of the scientific universe will be boosted more by AI ordinary matter, dark matter, or dark energy

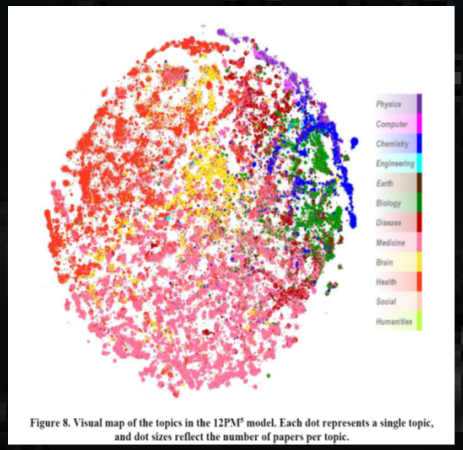
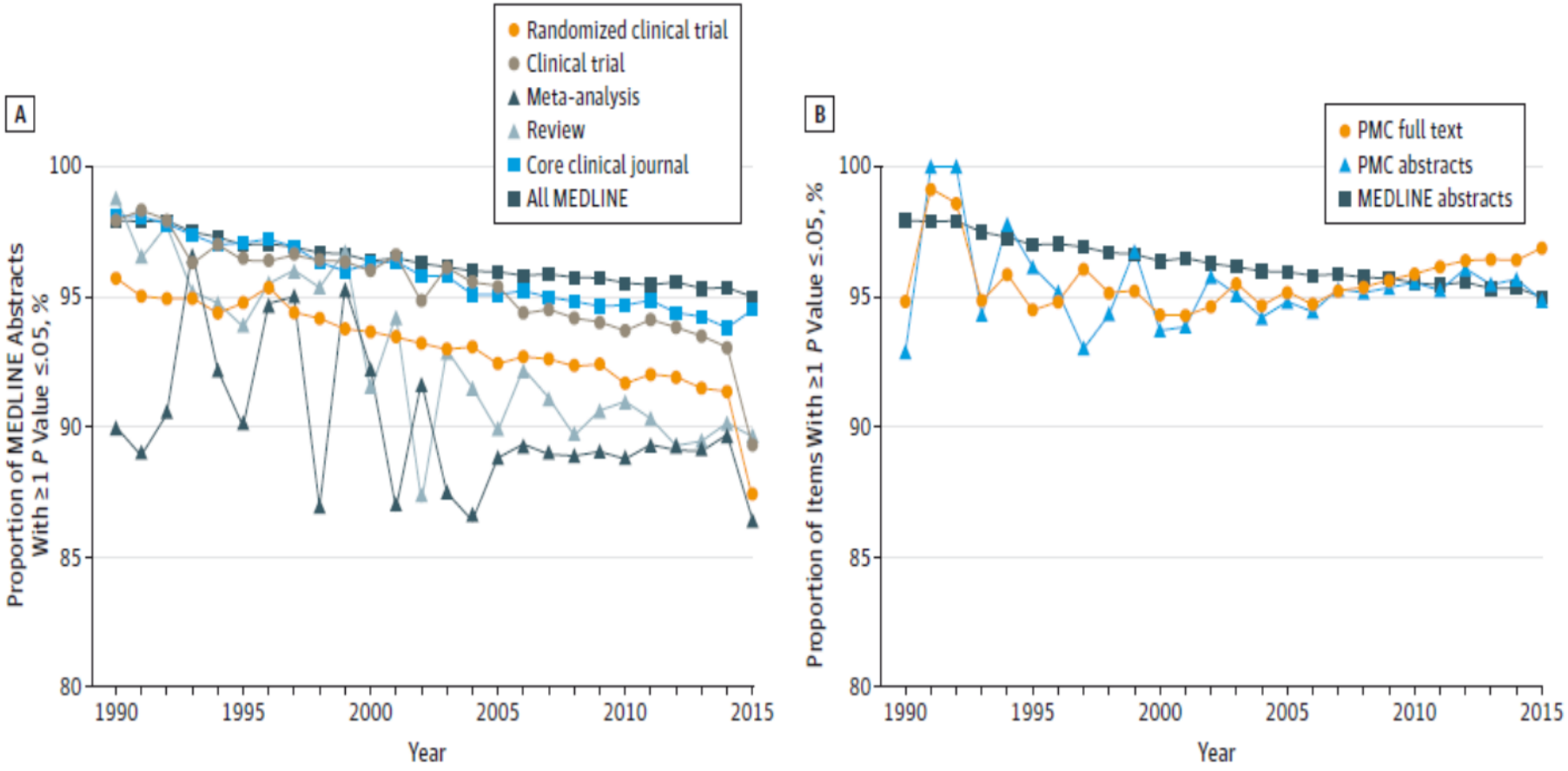


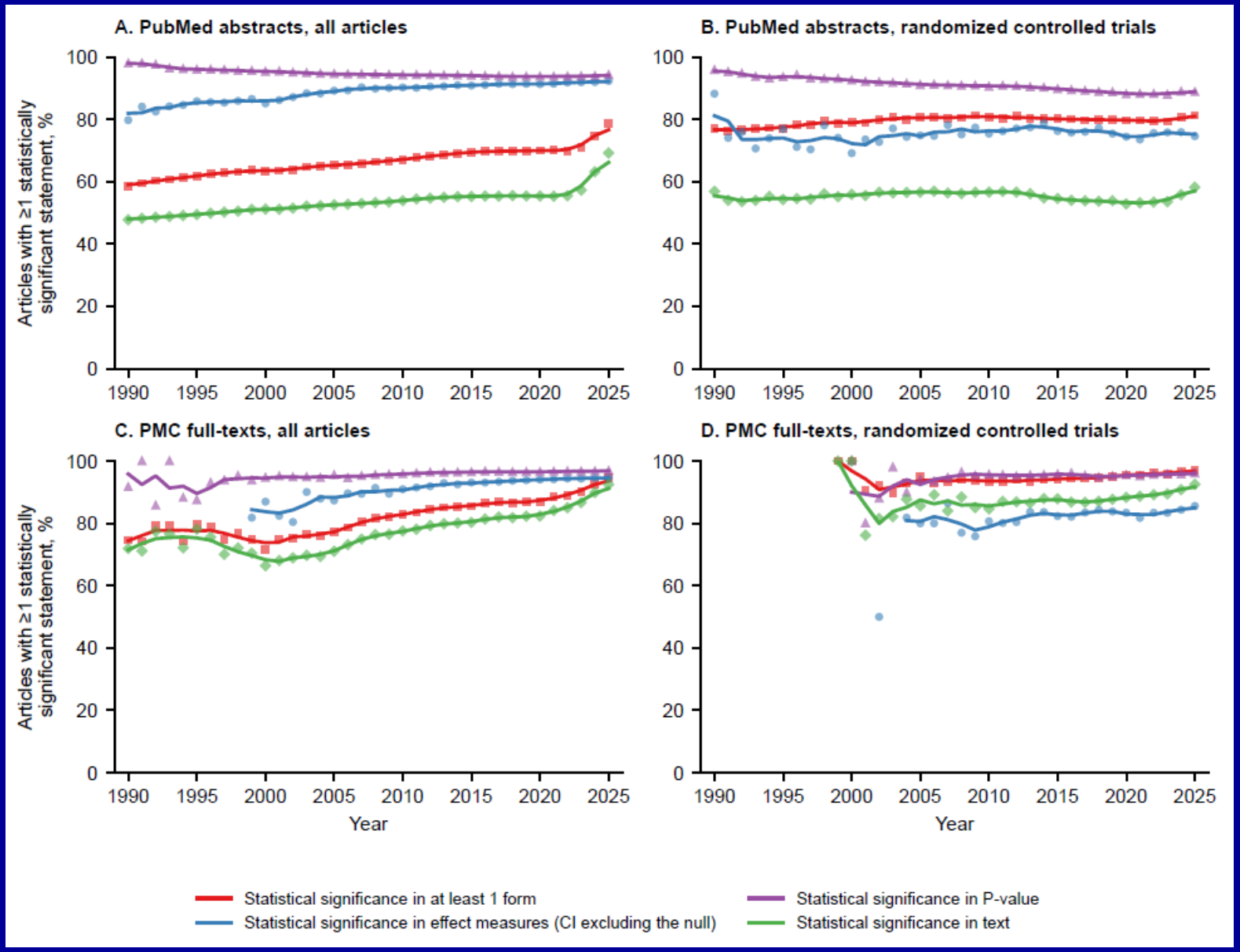
Figure 8. Visual map of the topics in the 12PM² model. Each dot represents a single topic, and dot sizes reflect the number of papers per topic.

The scientific literature has long reached the point of becoming mostly a boring nuisance: 96% of the biomedical literature claims significant results



Chavalarias, Wallach, Li, Ioannidis, JAMA 2016; and JAMA Netw Open in press

With hardly any change in the last decade



Science is getting less disruptive

Article

Papers and patents are becoming less disruptive over time

https://doi.org/10.1038/s41586-022-05543-x

Received: 14 February 2022

Accepted: 8 November 2022

Published online: 4 January 2023

Check for updates

Michael Park[1], Erin Leahey[2] & Russell J. Funk[1]✉

Theories of scientific and technological change view discovery and invention as endogenous processes[1,2], wherein previous accumulated knowledge enables future progress by allowing researchers to, in Newton's words, 'stand on the shoulders of giants'[3–7]. Recent decades have witnessed exponential growth in the volume of new scientific and technological knowledge, thereby creating conditions that should be ripe for major advances[8,9]. Yet contrary to this view, studies suggest that progress is slowing in several major fields[10,11]. Here, we analyse these claims at scale across six decades, using data on 45 million papers and 3.9 million patents from six large-scale datasets, together with a new quantitative metric—the CD index[12]—that characterizes how papers and patents change networks of citations in science and technology. We find that papers and patents are increasingly less likely to break with the past in ways that push science and technology in new directions. This pattern holds universally across fields and is robust across multiple different citation- and text-based metrics[1,13–17]. Subsequently, we link this decline in disruptiveness to a narrowing in the use of previous knowledge, allowing us to reconcile the patterns we observe with the 'shoulders of giants' view. We find that the observed declines are unlikely to be driven by changes in the quality of published science, citation practices or field-specific factors. Overall, our results suggest that slowing rates of disruption may reflect a fundamental shift in the nature of science and technology.

Public funding for research in inefficient and badly compressed and threatened

PNAS PERSPECTIVE OPEN ACCESS

The costs of competition in distributing scarce research funds

Gerald Schweiger[a,1], Adrian Barnett[b], Peter van den Besselaar[c,d], Lutz Bornmann[e], Andreas De Block[f], John P. A. Ioannidis[g,h,i,j], Ulf Sandström[k], and Stijn Conix[l]

Edited by Neil Shubin, The University of Chicago, Chicago, IL; received May 8, 2024; accepted October 18, 2024

	Recommendations		Expected Outcome
Data Collection and Accessibility	Make non-sensitive application and decision data publicly available for analysis, and facilitate responsible use of sensitive data.	• Funders should make anonymized data on rejected proposals available. • Use OpenSafely approach for sensitive data.	Facilitates research on funding processes to investigate existing systems and examine the impact of system changes, enabling transparency and accountability.
Decision Making Processes	Conduct controlled experiments to test the reliability, predictive validity, and fairness of funding decisions.	• Investigate differences in reliability across disciplines and in interdisciplinary research. • Use advanced bibliometric methods instead of citizen bibliometrics. • Use measures beyond bibliometrics: altmetrics, Overton data, connections to SDG, etc.	Improved understanding and refinement of decision-making processes in funding allocation. Demonstrates to the research community that processes are being interrogated.
Alternative Evaluation Systems	Experimentally test and compare different evaluation systems for competitive funding.	• Random audits that evaluate a small number of completed projects in great detail versus bibliometric outcomes for all completed projects that likely miss important benefits. • Field experiments with new types of reviewing and selecting processes. • Examine the time delay between funding and benefits.	A more accurate picture of the benefits of funding research which would be politically useful for science. Better data on what projects most often fail to complete the proposed work which could then inform funding calls and criteria.
Alternative Distribution Systems	Experimentally test funding models such as lotteries or base funding, or targeted funding based on peer nominations.	• Experimentally test various project-parameters such as duration or proportion of budget used on people funding.	Potentially reduces biases and administrative burdens, promoting fairness and diversity in funded research.
Economic Costs	Investigate the economic costs of competitive funding beyond the total costs for applicants.	• Investigate the division of labour in grant writing. • Investigate the costs of administration, decision-making processes, and positive effects of proposal writing. • Develop models that take into account all costs and potential positive effects of grant writing.	A detailed view on what is invested in competitive funding.
Epistemic Costs	Understand the relations between the various aims of science and different methods of funding.	• Map the relations between the various epistemic aims of science and different methods of funding science. • Develop models that can be used to design and compare different funding ecologies.	A heterogeneous view on science funding where different epistemic aims are pursued using different funding methods.
Social and Ethical Costs	Go beyond survey studies with convenience samples to map and diminish the social and ethical costs of competitive funding and other funding methods.	• Map out the potential social and ethical costs of alternative funding methods. • Randomized controlled trial of questionable research practices in competitive funding.	Creates a more supportive and ethical research environment, encouraging responsible research practices.

Reproducibility is an increasingly pervasive quest, but most science is not reproducible

A manifesto for reproducible science

Marcus R. Munafò[1,2*], Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]

Improving the reliability and efficiency of scientific research will increase the credibility of the published scientific literature and accelerate discovery. Here we argue for the adoption of measures to optimize key elements of the scientific process: methods, reporting and dissemination, reproducibility, evaluation and incentives. There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures, but their broad adoption by researchers, institutions, funders and journals will require iterative evaluation and improvement. We discuss the goals of these measures, and how they can be implemented, in the hope that this will facilitate action toward improving the transparency, reproducibility and efficiency of scientific research.

Improvements in reproducibility have happened (?)

- Reproducibility of methods: yes, in some fields, but not necessarily in those that produce many papers
- Reproducibility of results: remains unknown in many/most fields and most papers where replication is not attempted
- Reproducibility of inferences: is it even possible (worthwhile?) to improve

Interventions to improve reproducibility: how many can be facilitated by AI?

Design and conduct:

•Preregistration, Registered reports, Systematic reviews and other evidence synthesis, Electronic lab notebooks, Critical incident reporting, Multi-team/multi-center studies, Reproducibility training

Reporting:

•Reporting guidelines, Open and FAIR data, Open access, Open code, Open methods and protocols

Reviewing:

•Training, Statistical reviewing, Open peer review, Post-publication peer review

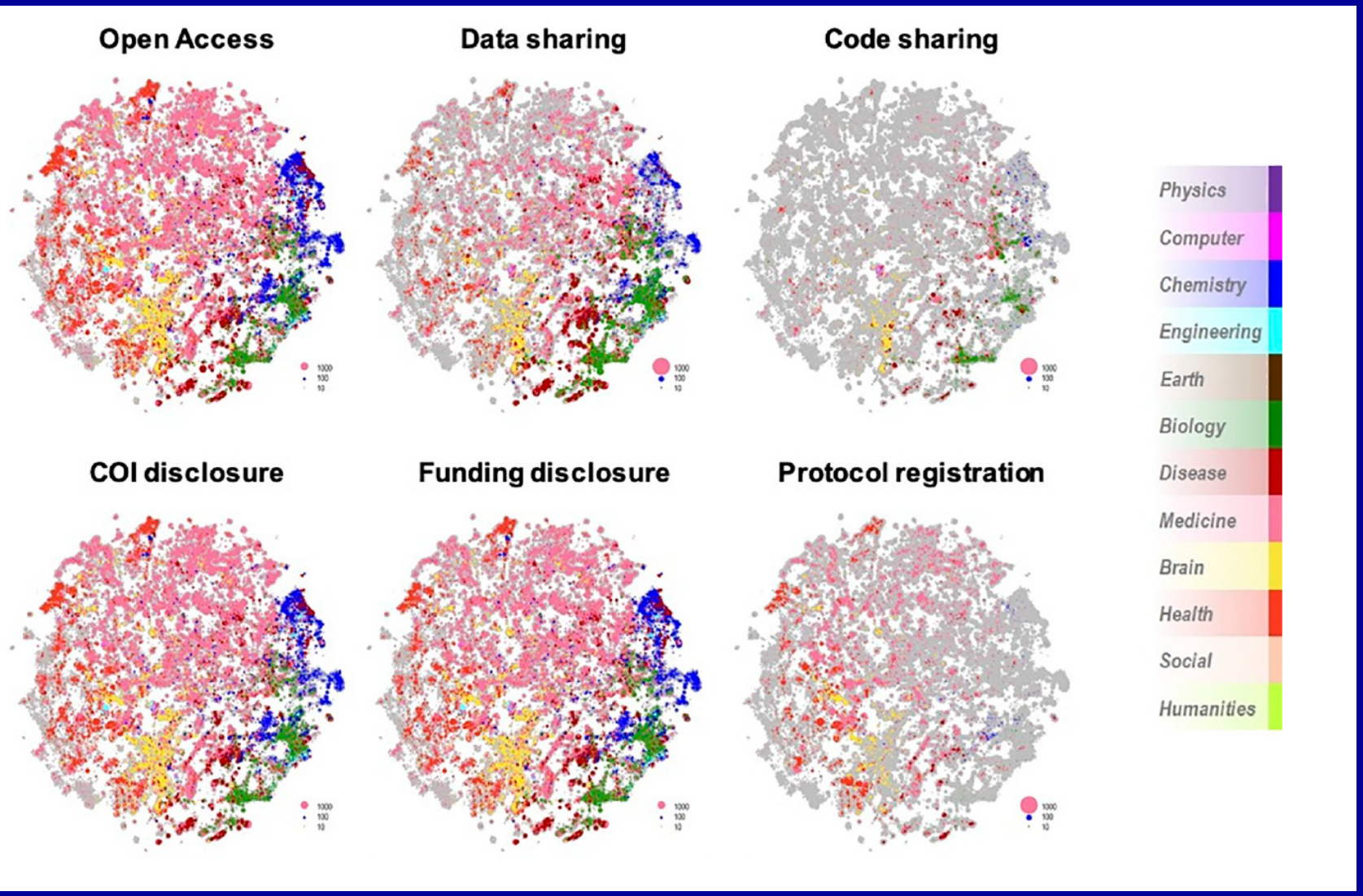
Haven and Ioannidis, Annual Review of Medicine 2025

Assessment of transparency indicators across the biomedical literature: How open is open?

Stylianos Serghiou[1,2], Despina G. Contopoulos-Ioannidis[3], Kevin W. Boyack[4], Nico Riedel[5], Joshua D. Wallach[6], John P. A. Ioannidis[1,2,7,8,9]*

Stagnation in sharing in the last 10 years – could AI help?




Open Access
Data sharing
Code sharing
COI disclosure
Funding disclosure
Protocol registration
Physics
Computer
Chemistry
Engineering
Earth
Biology
Disease
Medicine
Brain
Health
Social
Humanities

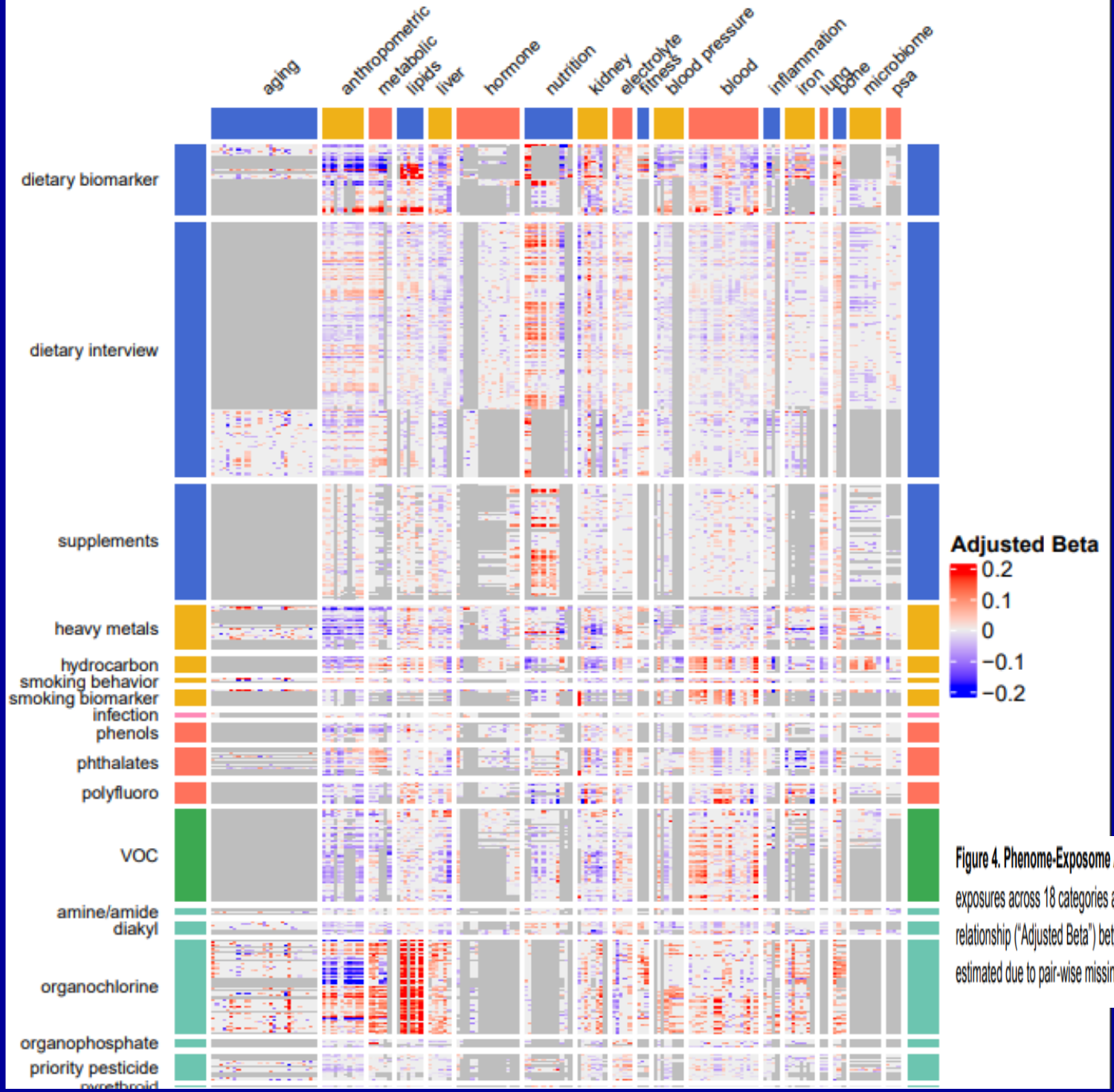


Figure 4. Phenome-Exposome Atlas. 305 phenotypes across 18 categories are depicted column wise, 625 exposures across 18 categories are depicted row wise. Each entry in the matrix is the linear association relationship ("Adjusted Beta") between exposure and phenotype. Grey denotes association that could not be estimated due to pair-wise missingness or total sample size less than 500.

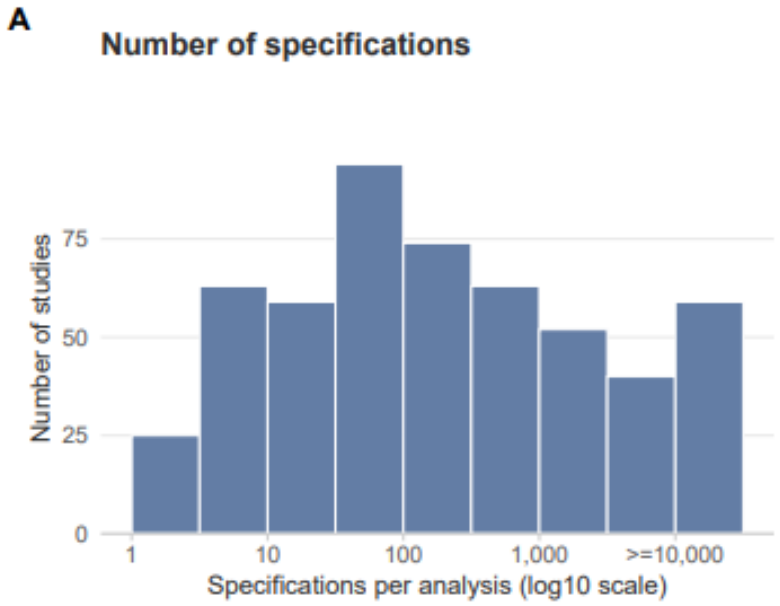
Analyze everything and report everything in an Atlas

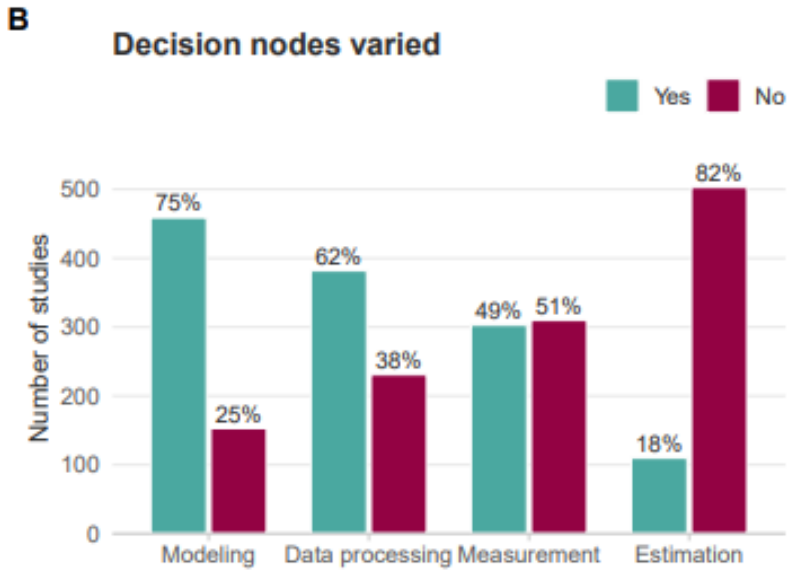
e.g. NHANHS

Patel, Ioannidis, Manrai, Nature Medicine 2026

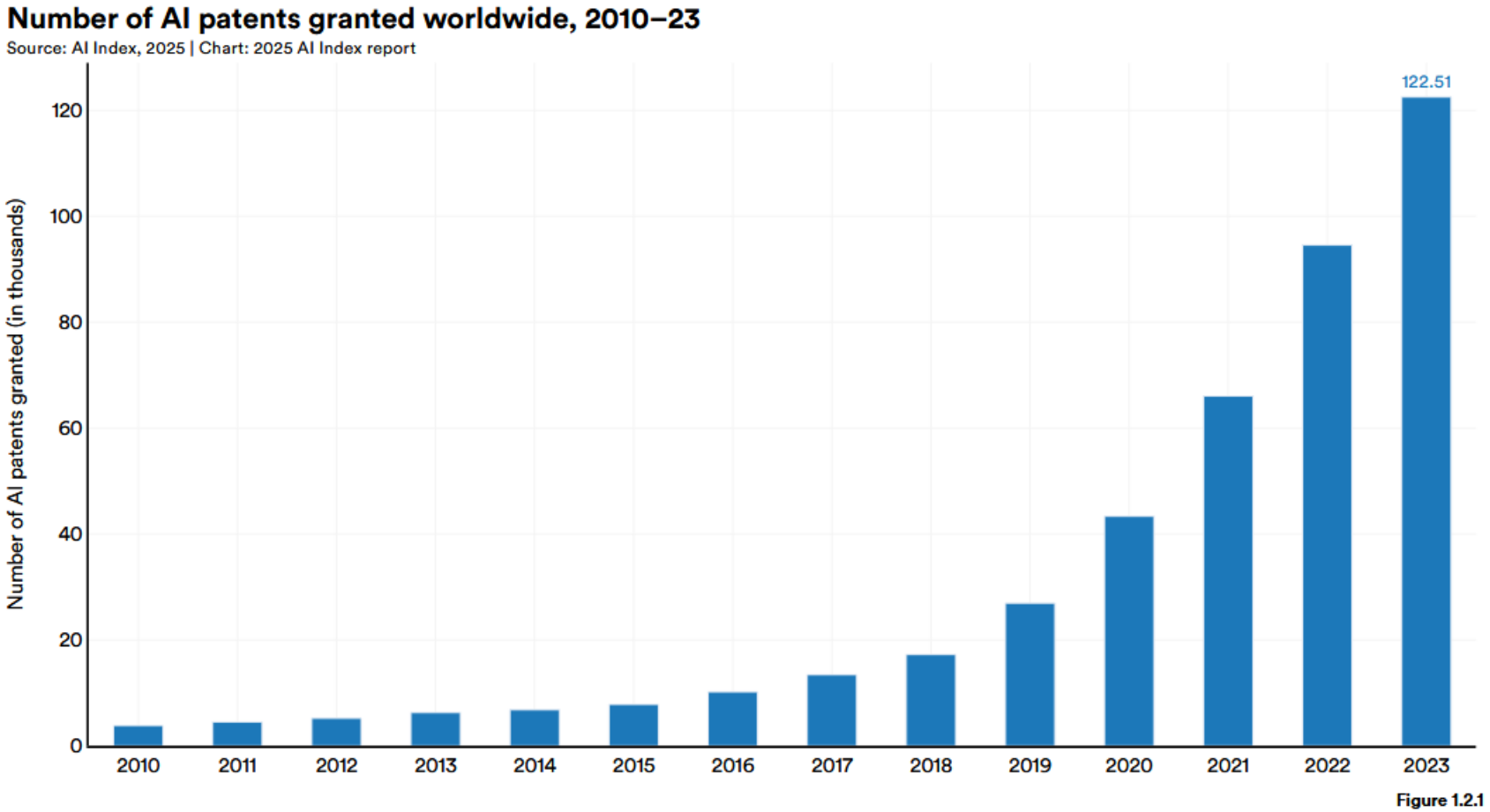
Uptake and Implementation of Multiverse-style Analyses Across 613 Studies

Andrew Nepomuceno[1] Angikar Ghosal[1]
Alejandro Sandoval-Lentisco[1*] John P. A. Ioannidis[1,2,3,4†]





AI enters the scene mostly from business, not science



Artificial Intelligence Index Report 2025

Challenges and opportunities

- AI increases computational and overall complexity
- LLMs may even defy strict reproducibility
- Agentic labs also lead to massive transformations of scientific discovery, appraisal, and misconduct
- This transformation can be both helpful and harmful in this regard

Wiping out mathematics with 722 preprints

jrl-openai Merge pull request #1 from openai/codex/update-10-7 fd4aeeb · 4 days ago 3 Commits

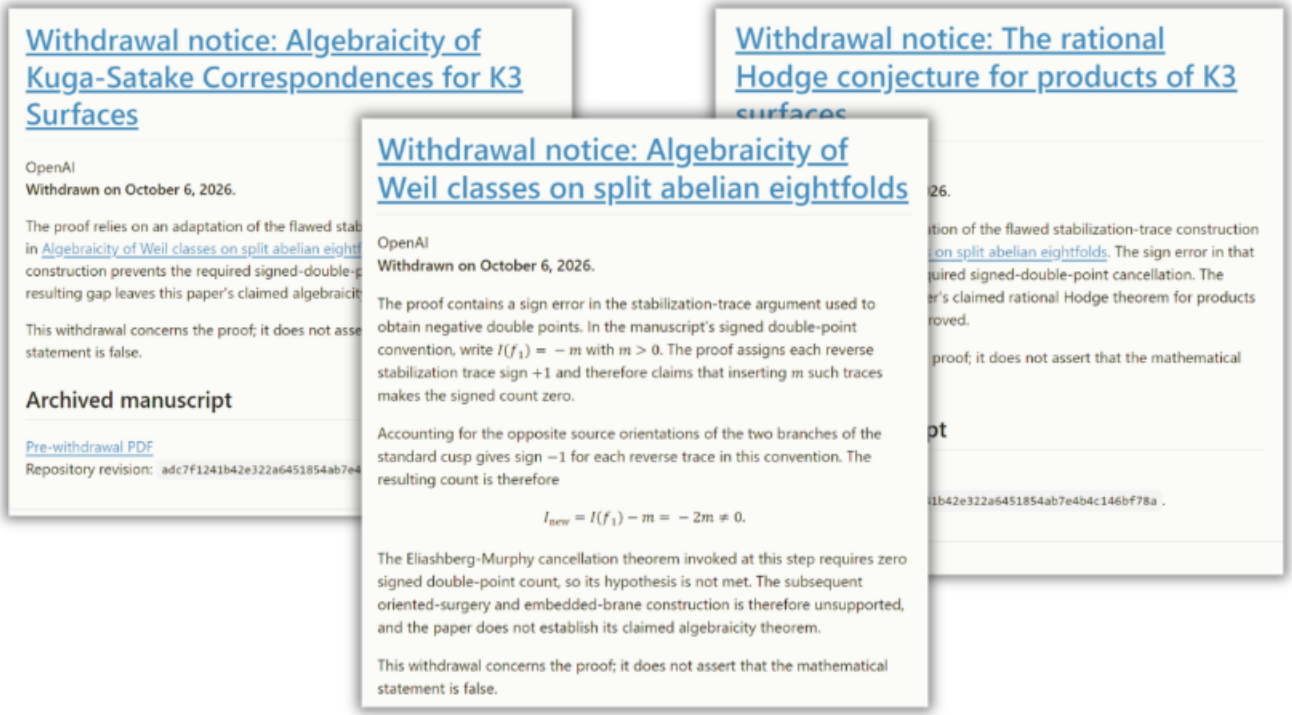
lean	Update manuscripts and Lean formalizations	4 days ago
preprints	Update manuscripts and Lean formalizations	4 days ago
reasoning_traces	Initial commit	5 days ago
CONTENTS.md	Update manuscripts and Lean formalizations	4 days ago
LICENSE	Initial commit	5 days ago
README.md	Update manuscripts and Lean formalizations	4 days ago
history.md	Update manuscripts and Lean formalizations	4 days ago
overview.pdf	Update manuscripts and Lean formalizations	4 days ago
overview.tex	Update manuscripts and Lean formalizations	4 days ago

README Apache-2.0 license

Readme

This repository contains mathematical manuscripts and supporting proof artifacts produced by an internal OpenAI model.

OpenAI withdraws three preprints a day after releasing 722 manuscripts on unsolved math problems



Withdrawal notice: Algebraicity of Kuga-Satake Correspondences for K3 Surfaces

OpenAI
Withdrawn on October 6, 2026.

The proof relies on an adaptation of the flawed stab
in Algebraicity of Weil classes on split abelian eight
construction prevents the required signed-double-p
resulting gap leaves this paper's claimed algebraicit

This withdrawal concerns the proof; it does not asse
statement is false.

Archived manuscript

Pre-withdrawal PDF
Repository revision: adc7f1241b42e322a6451854ab7e4

Withdrawal notice: Algebraicity of Weil classes on split abelian eightfolds

OpenAI
Withdrawn on October 6, 2026.

The proof contains a sign error in the stabilization-trace argument used to obtain negative double points. In the manuscript's signed double-point convention, write $I(f_1) = -m$ with $m > 0$. The proof assigns each reverse stabilization trace sign $+1$ and therefore claims that inserting m such traces makes the signed count zero.

Accounting for the opposite source orientations of the two branches of the standard cusp gives sign -1 for each reverse trace in this convention. The resulting count is therefore

$$I_{new} = I(f_1) - m = -2m \neq 0.$$

The Eliashberg-Murphy cancellation theorem invoked at this step requires zero signed double-point count, so its hypothesis is not met. The subsequent oriented-surgery and embedded-brane construction is therefore unsupported, and the paper does not establish its claimed algebraicity theorem.

This withdrawal concerns the proof; it does not assert that the mathematical statement is false.

Withdrawal notice: The rational Hodge conjecture for products of K3 surfaces

26.

ition of the flawed stabilization-trace construction
on split abelian eightfolds. The sign error in that
quired signed-double-point cancellation. The
er's claimed rational Hodge theorem for products
roved.

proof; it does not assert that the mathematical

pt

1b42e322a6451854ab7e4b4c146bf78a .

A Severe Misalignment of AI in Mathematics

Published 11 September 2026 · DOI 10.5281/zenodo.22737750 ▶ English

Over the last few months, the mathematical capabilities of LLMs have improved dramatically, to the point that they can solve major outstanding problems in many fields of mathematics. However, the push by AI companies to solve mathematical problems as a benchmark is detrimental to the science of mathematics, and to the mathematical community. The goals of the AI companies and the goals of the mathematical community are severely misaligned. We see these as part of broader alignment issues impacting other scientific and creative professions, as well as the whole of society.

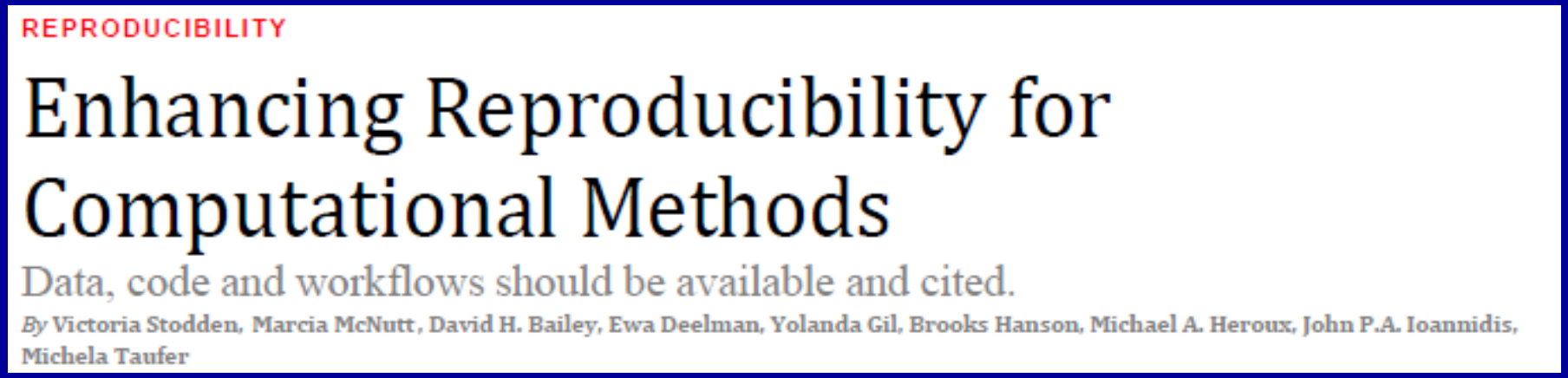
ADD YOUR NAME

If you support this declaration, you are invited to endorse it.

Endorse the Declaration ↗

Your endorsement appears right after you sign in with ORCID, or after you confirm your academic email address.

AI: a black box regarding reproducibility – or maybe not?



REPRODUCIBILITY

Enhancing Reproducibility for Computational Methods

Data, code and workflows should be available and cited.

By Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer

Science, December 2, 2016 – i.e. almost 10 years ago

Transparency and reproducibility in artificial intelligence

https://doi.org/10.1038/s41586-020-2766-y
Received: 1 February 2020
Accepted: 10 August 2020
Published online: 14 October 2020
Check for updates

Benjamin Haibe-Kains[1,2,3,4,5✉], George Alexandru Adam[2,5], Ahmed Hosny[6,7], Farnoosh Khodakarami[1,2], Massive Analysis Quality Control (MAQC) Society Board of Directors*, Levi Waldron[8], Bo Wang[2,3,5,9,10], Chris McIntosh[2,5,9], Anna Goldenberg[2,5,11,12], Anshul Kundaje[13,14], Casey S. Greene[15,16], Tamara Broderick[17], Michael M. Hoffman[1,2,3,5], Jeffrey T. Leek[18], Keegan Korthauer[19,20], Wolfgang Huber[21], Alvis Brazma[22], Joelle Pineau[23,24], Robert Tibshirani[25,26], Trevor Hastie[25,26], John P. A. Ioannidis[25,26,27,28,29], John Quackenbush[30,31,32] & Hugo J. W. L. Aerts[6,7,32,34]

ARISING FROM S. M. McKinney et al. *Nature* https://doi.org/10.1038/s41586-019-1799-6 (2020)

Table 1 | Essential hyperparameters for reproducing the study for each of the three models

	Lesion	Breast	Case
Learning rate	Missing	0.0001	Missing
Learning rate schedule	Missing	Stated	Missing
Optimizer	Stochastic gradient descent with momentum	Adam	Missing
Momentum	Missing	Not applicable	Not applicable
Batch size	4	Unclear	2
Epochs	Missing	120,000	Missing

publication, McKinney et al.[1] did not disclose the settings for the augmentation pipeline; the transformations used are stochastic and can considerably affect model performance[10]. Details of the training pipeline were also missing. Without this key information, independent reproduction of the training pipeline is not possible.

Numerous frameworks and platforms exist to make artificial intelligence research more transparent and reproducible (Table 2). For the sharing of code, these include Bitbucket, GitHub and GitLab, among others. The many software dependencies of large-scale machine learning applications require appropriate control of the software environment, which can be achieved through package managers including

Table 2 | Frameworks to share code, software dependencies and deep-learning models

Resource	URL
Code	
BitBucket	https://bitbucket.org
GitHub	https://github.com
GitLab	https://about.gitlab.com
Software dependencies	
Conda	https://conda.io
Code Ocean	https://codeocean.com
Gigantum	https://gigantum.com
Colaboratory	https://colab.research.google.com
Deep-learning models	
TensorFlow Hub	https://www.tensorflow.org/hub
ModelHub	http://modelhub.ai
ModelDepot	https://modeldepot.io
Model Zoo	https://modelzoo.co
Deep-learning frameworks	
TensorFlow	https://www.tensorflow.org/
Caffe	https://caffe.berkeleyvision.org/
PyTorch	https://pytorch.org/

AI may shift the scientific literature and move science outside the scientific literature

- Stanford has the highest computational capacity than any other university worldwide
- Still, this is only 1% (or less) of the computational capacity of Alphabet, Meta, Microsoft, Apple.
- Most AI research may grow outside the (published) scientific literature
- It may or may not be open source

Most science is happening outside the scientific literature?

317 AI unicorns: most publish nothing, all of them publish 0.1% of AI literature

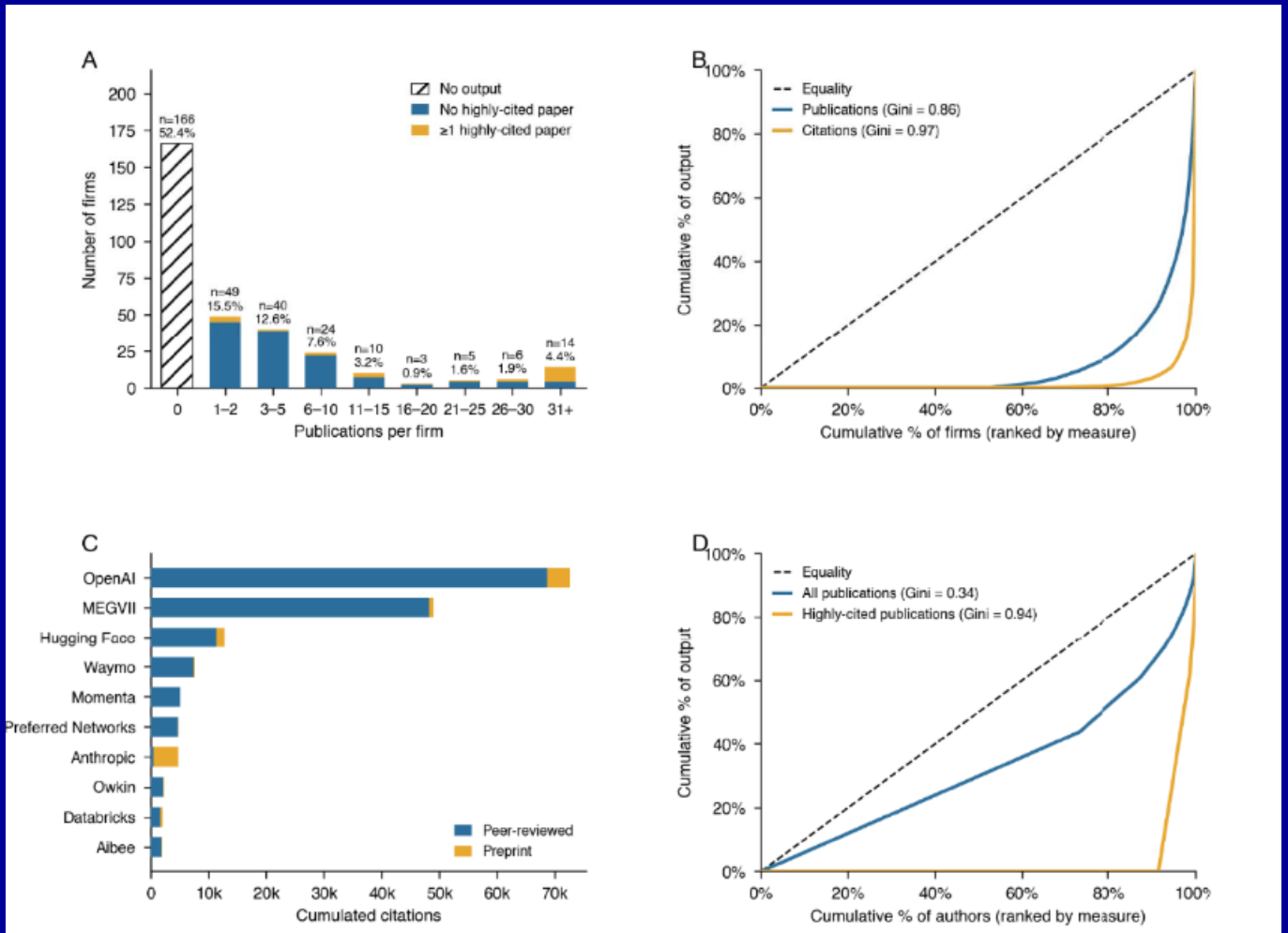


Fig. 1 – Firm distributions and concentrations of scientific output. (A) Distribution of publications per firm. (B) Lorenz curves: firm concentrations of publications and citations. (C) Top 10 firms by cumulative citations. (D) Lorenz curves: author concentrations of publications and highly cited publications.

Loisel, Sandoval-Lentisco, Ioannidis, bioRxiv 2026

Some stunning transformative successes

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1

Received: 11 May 2021

Accepted: 16 July 2021

Published online: 22 July 2021

Open access

Check for updates

Kathryn Tunyasuvunakool[1]✉, Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3]✉ & Demis Hassabis[1,3]✉

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy structure prediction will become an important tool that will allow new questions to be addressed from a structural perspective.

But mostly incremental benefits (in hybrid uses)

Machine learning augmentation reduces prediction error in collective forecasting: development and validation across prediction markets with application to COVID events

Alexander Gruen,[a] Karl R. Mattingly,[b] Ellen Monwitch,[a] Frederik Bossaerts,[b] Manning Clifford,[c] Chad Nash,[b] John P. A. Ioannidis,[d] and Anne-Louise Ponsonby[a,e,f,g,*]

Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review

Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis, Eric J Topol, Pranav Rajpurkar**

This scoping review of randomised controlled trials on artificial intelligence (AI) in clinical practice reveals an expanding interest in AI across clinical specialties and locations. The USA and China are leading in the number of

Lancet Digit Health 2024; 6: e367–73

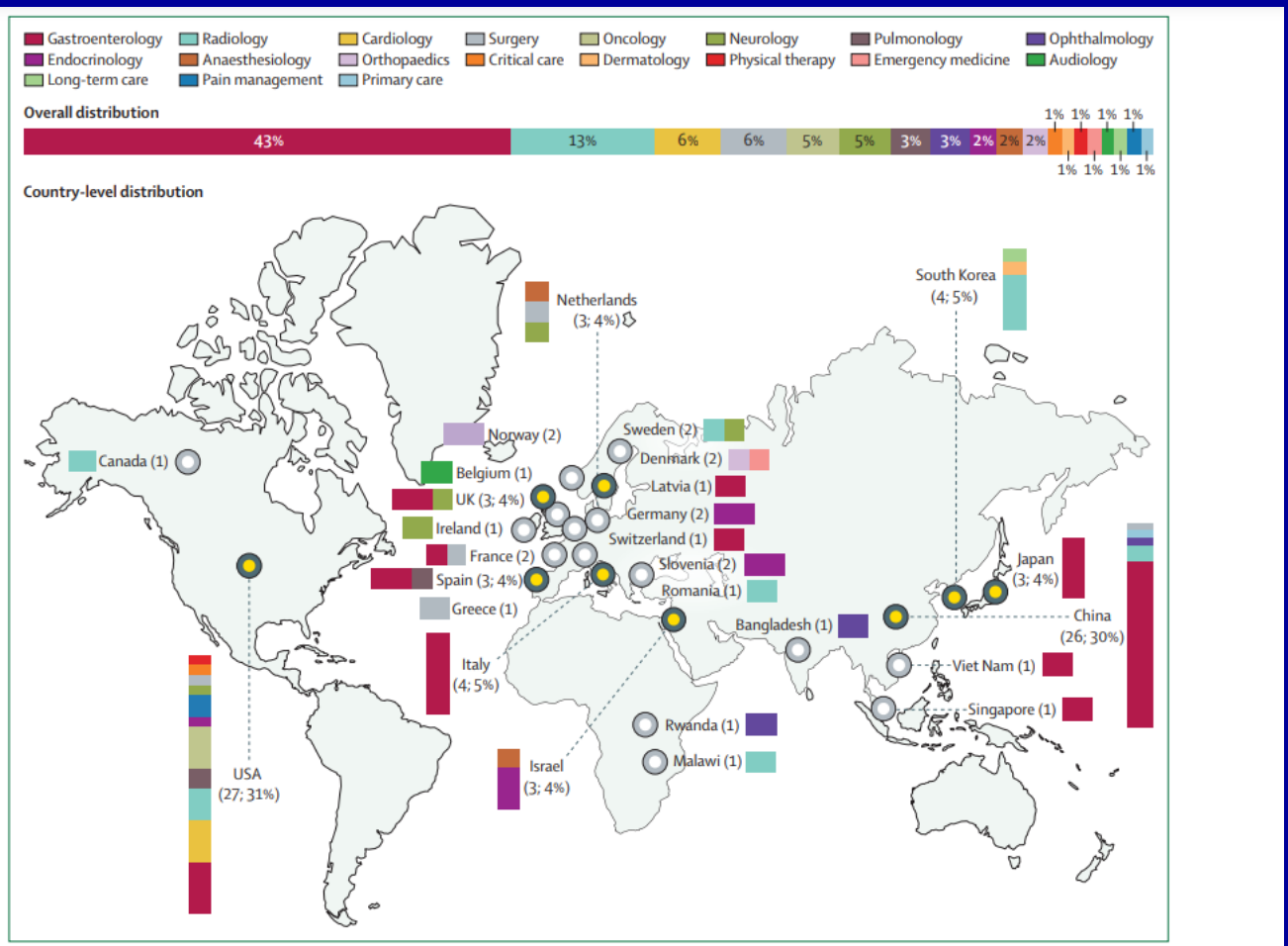


Figure 2: Randomised controlled trials of artificial intelligence in clinical practice across countries and specialties

	Statistically significant improvement	No statistically significant effect	Showed non-inferiority	Statistically significant deterioration	Total
Care management	15	1	2	..	18
Clinical decision making	6	1	..	..	7
Diagnostic yield or performance	34	10	1	1	46
Patient behaviour and symptoms	10	3	2	..	15
Total	65	15	5	1	86

Data are n.

Table 1: **Primary endpoints and types for randomised controlled trials of artificial intelligence in clinical practice**

	Statistically significant improvement	No statistically significant effect	Showed non-inferiority	Statistically significant deterioration	Total
AI *vs* clinician	3	1	3	1	8
AI *vs* routine care	16	4	..	..	20
AI-assisted clinician *vs* unassisted clinician	46	10	2	..	58
Total	65	15	5	1	86

Data are n. AI=artificial intelligence.

Table 2: **Primary endpoint results and group comparisons for randomised controlled trials of AI in clinical practice**

LLMs may help in evidence appraisal

ELSEVIER

Journal of Clinical Epidemiology 175 (2024) 111533

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Benchmarking Human–AI collaboration for common evidence appraisal tools

Tim Woelfle[a,b,c,*], Julian Hirt[a,d,e], Perrine Janiaud[a,d], Ludwig Kappos[a], John P.A. Ioannidis[f,g], Lars G. Hemkens[a,d,f,h]

[a]Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), Basel, Switzerland
[b]Department of Neurology, University Hospital Basel, Basel, Switzerland
[c]Translational Imaging in Neurology (ThINk), Department of Biomedical Engineering, University Hospital and University of Basel, Basel, Switzerland
[d]Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland
[e]Institute of Nursing Science, Department of Health, Eastern Switzerland University of Applied Sciences, St. Gallen, Basel, Switzerland
[f]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA
[g]Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, Stanford University, Stanford, CA, USA
[h]Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany

Accepted 9 September 2024; Published online 12 September 2024

Abstract

Background and Objective: It is unknown whether large language models (LLMs) may facilitate time- and resource-intensive text-related processes in evidence appraisal. The objective was to quantify the agreement of LLMs with human consensus in appraisal of scientific reporting (Preferred Reporting Items for Systematic reviews and Meta-Analyses [PRISMA]) and methodological rigor (A MeaSurement Tool to Assess systematic Reviews [AMSTAR]) of systematic reviews and design of clinical trials (PRagmatic Explanatory Continuum Indicator Summary 2 [PRECIS-2]) and to identify areas where collaboration between humans and artificial intelligence (AI) would outperform the traditional consensus process of human raters in efficiency.

Study Design and Setting: Five LLMs (Claude-3-Opus, Claude-2, GPT-4, GPT-3.5, Mixtral-8x22B) assessed 112 systematic reviews applying the PRISMA and AMSTAR criteria and 56 randomized controlled trials applying PRECIS-2. We quantified the agreement between human consensus and (1) individual human raters; (2) individual LLMs; (3) combined LLMs approach; (4) human–AI collaboration. Ratings were marked as deferred (undecided) in case of inconsistency between combined LLMs or between the human rater and the LLM.

Results: Individual human rater accuracy was 89% for PRISMA and AMSTAR, and 75% for PRECIS-2. Individual LLM accuracy was ranging from 63% (GPT-3.5) to 70% (Claude-3-Opus) for PRISMA, 53% (GPT-3.5) to 74% (Claude-3-Opus) for AMSTAR, and 38% (GPT-4) to 55% (GPT-3.5) for PRECIS-2. Combined LLM ratings led to accuracies of 75%–88% for PRISMA (4%–74% deferred), 74%–89% for AMSTAR (6%–84% deferred), and 64%–79% for PRECIS-2 (29%–88% deferred). Human–AI collaboration resulted in the best accuracies from 89% to 96% for PRISMA (25/35% deferred), 91%–95% for AMSTAR (27/30% deferred), and 80%–86% for PRECIS-2 (76/71% deferred).

Conclusion: Current LLMs alone appraised evidence worse than humans. Human–AI collaboration may reduce workload for the second human rater for the assessment of reporting (PRISMA) and methodological rigor (AMSTAR) but not for complex tasks such as PRECIS-2. © 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Review article

Artificial intelligence in systematic reviews and meta-analyses: Task-specific performance, residual error quantification, and human oversight

Jules Descamps [a,b], Nicolas Bouguennec [c], Raphael Porcher [b], Matthieu Resche-Rigon [b], Pierre-Alban Bouché [a,b,*]

Results: Across the workflow, AI maturity is heterogeneous. In a 2025 systematic review of 19 comparative studies, generative AI missed 68.0%–96.0% of relevant studies in searching, made incorrect inclusion decisions in 0.0%–29.0% and incorrect exclusions in 1.0%–83.0%, incorrect data extractions in 4.0%–31.0%, and incorrect risk-of-bias assessments in 10.0%–56.0%. A 2025 scoping review identified 37 articles: 15/37 (41%) searching, 14/37 (38%) study selection, 11/37 (30%) extraction, 33/37 (89%) GPT-based, 21/37 (57%) validation studies; 20/37 (54%) drew a promising conclusion, none as a validated implementation. Title/abstract screening was most mature (sensitivity 99.2%, specificity 83.6%) versus full-text screening (97.6% and 47.4%). For structured extraction across 30 articles, one tool reached 92.0% precision, recall, and F1, but sensitivity fell to 77.0%–80.0% for review-specific variables, with confabulations in 4/90 (4%) of data points; across 107 trials, extraction accuracy was 96.2% (97.9% with refinement) while mean time fell from 86.9 to 14.7 min per trial. Risk-of-bias assessment was most fragile (kappa 0.51, 95% CI: 0.36–0.66). For residual error, 0 errors among 60 checked items corresponds to an upper one-sided 95% bound of 4.9%.

Discussion: AI is transforming SR/MA toward hybrid human-machine workflows rather than replacing them. Performance is task-specific, and metrics alone are insufficient: residual uncertainty should be quantified. Current evidence supports AI as an assistive technology under human oversight, not autonomous use.

Level of evidence: lV; narrative review.

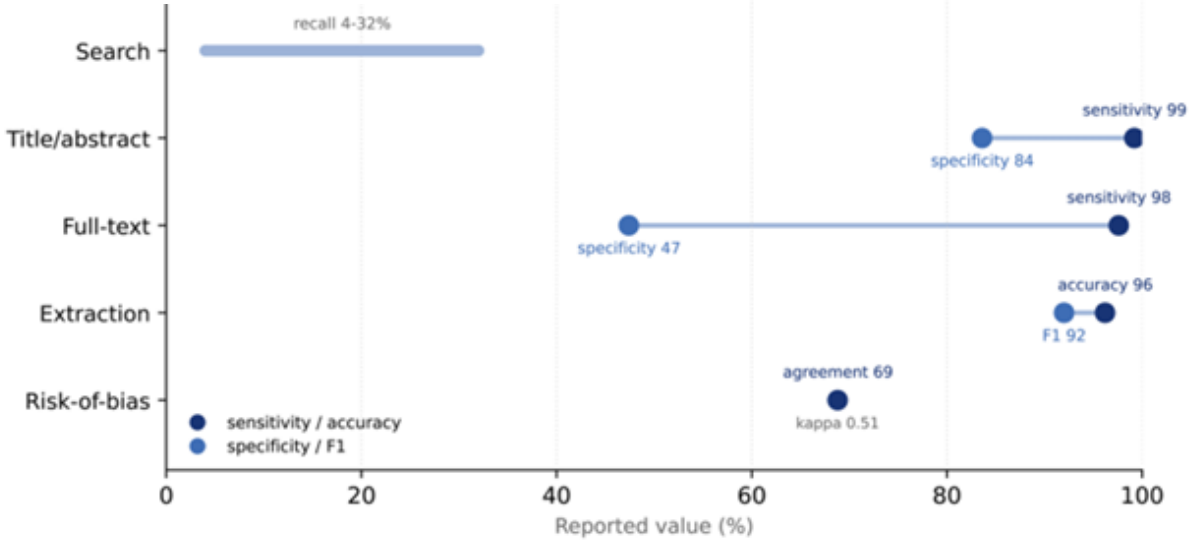


Fig. 1. Performance of artificial intelligence at different steps of the systematic review process.

Journal of the American Medical Informatics Association, 2025, **32(6)**, 1071–1086
https://doi.org/10.1093/jamia/ocaf063
Advance access publication 7 May 2025
Review

Review

The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review

Dmitry Scherbakov, PhD[1], Nina Hubig, PhD[1,2], Vinita Jansari, PhD[3], Alexander Bakumenko, MSc[3], Leslie A. Lenert, MD, MS*[,1]

[1]Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina (MUSC), Charleston, SC 29403, United States, [2]Interdisciplinary Transformation University, OG 2 A-4040 Linz, Austria, [3]School of Computing, Clemson University, Charleston, SC 29634, United States

*Corresponding author: Leslie A. Lenert, MD, MS, Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina (MUSC), 22 WestEdge Street, Suite 200, Room WG213, Charleston, SC 29403, United States (lenert@musc.edu)

D. Scherbakov and N. Hubig contributed equally to this work.

Abstract

Objectives: This study aims to summarize the usage of large language models (LLMs) in the process of creating a scientific review by looking at the methodological papers that describe the use of LLMs in review automation and the review papers that mention they were made with the support of LLMs.

Materials and Methods: The search was conducted in June 2024 in PubMed, Scopus, Dimensions, and Google Scholar by human reviewers. Screening and extraction process took place in Covidence with the help of LLM add-on based on the OpenAI GPT-4o model. ChatGPT and Scite.ai were used in cleaning the data, generating the code for figures, and drafting the manuscript.

Results: Of the 3788 articles retrieved, 172 studies were deemed eligible for the final review. ChatGPT and GPT-based LLM emerged as the most dominant architecture for review automation (n = 126, 73.2%). A significant number of review automation projects were found, but only a limited number of papers (n = 26, 15.1%) were actual reviews that acknowledged LLM usage. Most citations focused on the automation of a particular stage of review, such as Searching for publications (n = 60, 34.9%) and Data extraction (n = 54, 31.4%). When comparing the pooled performance of GPT-based and BERT-based models, the former was better in data extraction with a mean precision of 83.0% (SD = 10.4) and a recall of 86.0% (SD = 9.8).

Discussion and Conclusion: Our LLM-assisted systematic review revealed a significant number of research projects related to review automation using LLMs. Despite limitations, such as lower accuracy of extraction for numeric data, we anticipate that LLMs will soon change the way scientific reviews are conducted.

Key words: large language models; review automation; systematic review; scoping review; Covidence.

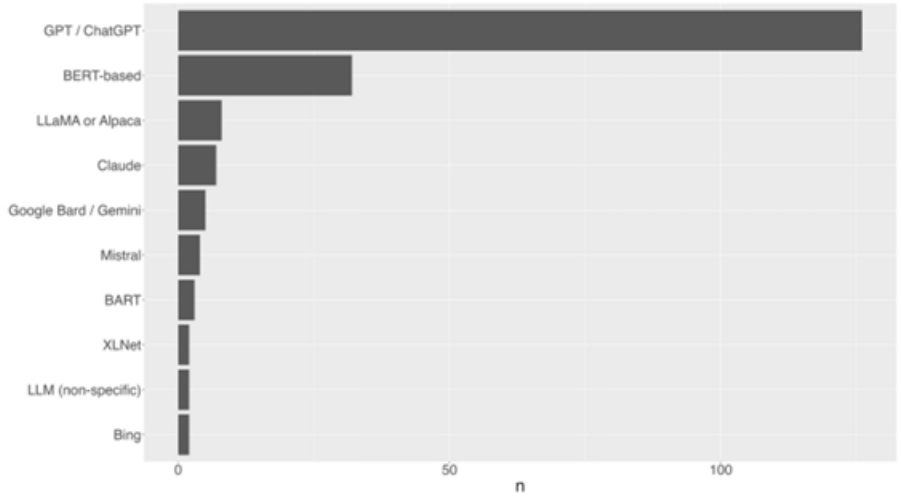


Figure 5. LLM types proposed for automation (models mentioned in 2 or more studies shown).

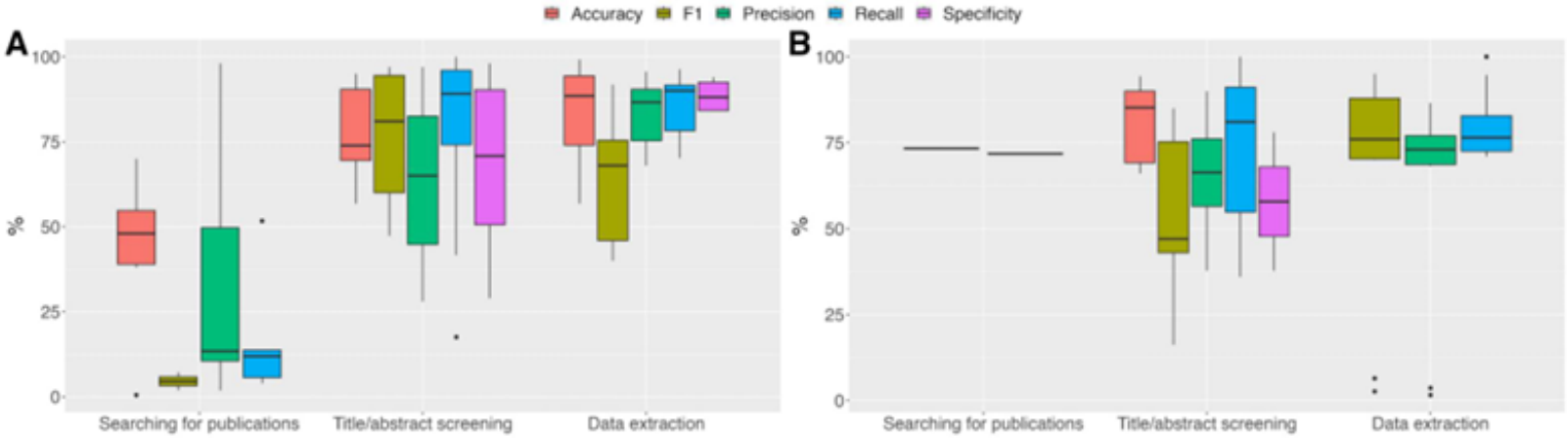


Figure 6. Performance metrics reported for the 3 most common automated stages for (A) GPT-based models and (B) BERT-based models. Boxplots display median value, while comparison is provided using means.

Automation of Systematic Reviews with Large Language Models

Christian Cao, MD[1, *], Rohit Arora, PhD[2, *], Paul Cento, BSc[3, *], Adil Budak, BSc[4], Katherine Manta, MSc, MD[1], Elina Farahani, MD[1], Matthew Cecere, MSc, MD[1], Anabel Selemon, MSc, BHSc[5], Jason Sang BSc[6], Ling Xi Gong[2], Robert Kloosterman, MSc, MD[1], Scott Jiang, BASc[7], Richard Saleh, MA, MD[1], Denis Margalik, MSc, MD[1], James Lin[8], Jane Jomy, MSc, MD[1], Jerry Xie, BSc[6], David Chen, MD[1], Jaswanth Gorla, MSc, MD[1], Sylvia Lee, MSc[9], Kelvin Zhang, BSc[6], Jennifer Kuang, BSc[6], Harriet Ware, MSc[10], Mairead Whelan, MSc[10], Bijan Teja, MD, PhD[11,12], Alexander A. Leung, MD, MPH[13], Rahul K. Arora, DPhil[10], Jennifer Pillay[14], Lisa Hartling[14], Michael Noetel, PhD[15], David B. Emerson, PhD[16], Allan S. Detsky, MD, PhD[1,17,21], Andrea C. Tricco PhD[18,19], George M. Church, PhD[2, †], David Moher, PhD[20,21,22, †], Niklas Bobrovitz, MD, DPhil[23, †]

Results: *otto-SR* showed high performance in phase 1 article screening (*otto-SR*: 96.7% sensitivity, 97.9% specificity; human: 81.7% sensitivity, 98.1% specificity) and phase 2 data extraction (*otto-SR*: 93.1% accuracy; human: 79.7% accuracy). In phase 3, *otto-SR* demonstrated high interrater reliability for risk of bias judgements (ROB2 0.98, Newcastle-Ottawa 0.95, QUADAS2 0.74; Gwet AC2). In phase 4, *otto-SR*, reproduced and updated the primary analysis from an issue of Cochrane reviews. Across Cochrane reviews, *otto-SR* incorrectly excluded a median of 0 studies (IQR 0 to 0.25), and found nearly twice as many eligible studies compared to the original authors (n= 114 vs. 64). Meta-analyses based on *otto-SR* generated screening and extraction outputs, subsequently verified through dual human review, yielded newly statistically significant effect estimates in 2 reviews and negated significance in 1 review.

Conclusions and relevance: LLMs have high performance in article screening, data extraction, and risk of bias assessments. They can rapidly reproduce and update existing systematic reviews, laying the foundation for automated, scalable, and reliable evidence synthesis.

LLMs may create a massive fraudulent scientific literature

Up to one in seven submissions to hundreds of Wiley journals flagged by new paper mill tool



Wiley, whose Hindawi subsidiary has attracted thousands of paper mill papers that later needed to be retracted, has seen widespread paper mill activity among hundreds of its journals, it announced yesterday.

Authorship and prolific authorship: a deflated currency?

Higher Education

Using AI, a professor wrote 200 papers this year. Researchers are alarmed.

An online platform removed 257 papers written or co-written by the University of Chicago's Nicholas Polson after other academics raised questions.

September 23, 2026

Article | Published: 14 January 2026

Artificial intelligence tools expand scientists' impact but contract science's focus

Qianyue Hao, Fengli Xu ✉, Yong Li ✉ & James Evans ✉

Nature **649**, 1237–1243 (2026) | Cite this article

Save article

64k Accesses | **152** Citations | **1231** Altmetric | Metrics

Abstract

Developments in artificial intelligence (AI) have accelerated scientific discovery[1]. Alongside recent AI-oriented Nobel prizes[2,3,4,5,6,7,8,9], these trends establish the role of AI tools in science[10]. This advancement raises questions about the influence of AI tools on scientists and science as a whole, and highlights a potential conflict between individual and collective benefits[11]. To evaluate these questions, we used a pretrained language model to identify AI-augmented research, with an F1-score of 0.875 in validation against expert-labelled data. Using a dataset of 41.3 million research papers across the natural sciences and covering distinct eras of AI, here we show an accelerated adoption of AI tools among scientists and consistent professional advantages associated with AI usage, but a collective narrowing of scientific focus. Scientists who engage in AI-augmented research publish 3.02 times more papers, receive 4.84 times more citations and become research project leaders 1.37 years earlier than those who do not. By contrast, AI adoption shrinks the collective volume of scientific topics studied by 4.63% and decreases scientists' engagement with one another by 22%. By consequence, adoption of AI in science presents what seems to be a paradox: an expansion of individual scientists' impact but a contraction in collective science's reach, as AI-augmented work moves collectively towards areas richest in data. With reduced follow-on engagement, AI tools seem to automate established fields rather than explore new ones, highlighting a tension between personal advancement and collective scientific progress.

ELSEVIER

Journal of Clinical Epidemiology 199 (2026) 112464

ORIGINAL RESEARCH

Paper mill or paper mine? A tentative answer to the sharp increase in research papers based on the Global Burden of Disease database

Peter Methys Degen[a,*], Andrea Riebler[b], Leonhard Held[a,c]

[a]*Center for Reproducible Science and Research Synthesis, University of Zurich, Hirschengraben 84, Zurich 8001, Zurich, Switzerland*
[b]*Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, Trondheim, Trøndelag 7034, Norway*
[c]*Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, Zurich 8001, Zurich, Switzerland*

Accepted 9 August 2026; Published online 20 August 2026

Abstract

Objectives: Recent studies have raised concerns arising from the exploitation (''mining'') of public health databases for low-quality, mass-produced papers. However, it remains challenging to disambiguate whether such papers originate from paper mills (commercial entities that sell authorships on mass-produced papers) or the uncoordinated action of individuals facilitated by artificial intelligence (AI) tools and templated workflows. Our study aims to address this question for one particular database, the Global Burden of Disease (GBD) study. We selected this database after noticing that one of our papers on Bayesian age-period-cohort models has recently been experiencing a rapid surge in geographically clustered citations from GBD papers with Chinese affiliations.

Methods: We collected bibliometric and article-level metadata from GBD papers to search for indicators of mass-produced research. Moreover, we assessed 713 full-text articles for reported R versions, availability of code and data, and declaration of generative AI use. For 180 articles, we qualitatively screened the figures for graphical similarities. Finally, we conducted an exploratory scoping investigation of online platforms (social media sites, vendor websites) dedicated to do-it-yourself workflows for secondary analyses of public health data.

Results: Although we cannot rule out paper mill involvement, our findings suggest that the geographically clustered increase in GBD publications from China is at least partially driven by independent authors. The wide variety of R versions listed in 477 articles points against centralized paper production. Moreover, despite broad graphical similarities in figure styles that suggest the use of shared visualization tools, substantial variation in ancillary details suggest independent authors finalizing figures. This is corroborated by the identification of an online ecosystem of proprietary tools and services specializing in streamlined do-it-yourself workflows for conducting, writing, and publishing secondary analyses of public health data.

Conclusion: Appropriate efforts should be directed toward evaluating the quality of the identified workflows. Stakeholders in scientific integrity should monitor online platforms dedicated to the rapid production of papers, especially in light of the increasing focus on AI-assisted workflows. Code sharing should be mandated for data-driven secondary analyses. Paywalls and proprietary software licenses hinder transparency and reusability, underscoring the importance of free and open-source software for trustworthy and reproducible research. © 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

META-RESEARCH ARTICLE

Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database

Tulsi Suchak[1], Anietie E. Aliu[1], Charlie Harrison[2], Reyer Zwiggelaar[2], Nophar Geifman[1], Matt Spick[1]*

1 School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom, **2** Department of Computer Science, Aberystwyth University, Ceredigion, United Kingdom

* matt.spick@surrey.ac.uk

Abstract

With the growth of artificial intelligence (AI)-ready datasets such as the National Health and Nutrition Examination Survey (NHANES), new opportunities for data-driven research are being created, but also generating risks of data exploitation by paper mills. In this work, we focus on two areas of potential concern for AI-supported research efforts. First, we describe the production of large numbers of formulaic single-factor analyses, relating single predictors to specific health conditions, where multifactorial approaches would be more appropriate. Employing AI-supported single-factor approaches removes context from research, fails to capture interactions, avoids false discovery correction, and is an approach that can easily be adopted by paper mills. Second, we identify risks of selective data usage, such as analyzing limited date ranges or cohort subsets without clear justification, suggestive of data dredging, and post-hoc hypothesis formation. Using a systematic literature search for single-factor analyses, we identified 341 NHANES-derived research papers published over the past decade, each proposing an association between a predictor and a health condition from the wide range contained within NHANES. We found evidence that research failed to take account of multifactorial relationships, that manuscripts did not account for the risks of false discoveries, and that researchers selectively extracted data from NHANES rather than utilizing the full range of data available. Given the explosion of AI-assisted productivity in published manuscripts (the systematic search strategy used here identified an average of 4 papers per annum from 2014 to 2021, but 190 in 2024–9 October alone), we highlight a set of best practices to address these concerns, aimed at researchers, data controllers, publishers, and peer reviewers, to encourage improved statistical practices and mitigate the risks of paper mills using AI-assisted workflows to introduce low-quality manuscripts to the scientific literature.

Scientometrics
https://doi.org/10.1007/s11192-024-05117-w

Evolving patterns of extreme publishing behavior across science

John P. A. Ioannidis[1,2,3,4,5] · Thomas A. Collins[6] · Jeroen Baas[7]



Fig 1. Signal indicators in 59 authors with ultra-precocious citation impact among top-cited scientists based on citation counts to end of 2023. For definitions of the 8 indicators, see Methods. Data are from Scopus and raw data can be found in https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/7 for country, subfield, and all citation metrics, and https://elsevier.digitalcommonsdata.com/datasets/kmyvjk3xmd/2 for extreme

RESEARCH ARTICLE

Features and signals in precocious citation impact: A meta-research study

John P. A. Ioannidis*

Artificial intelligence in scientific medical writing: Legitimate and deceptive uses and ethical concerns

Davide Ramoni [a,1], Cosimo Sgura [a,1], Luca Liberale [a,b], Fabrizio Montecucco [a,b], John P.A. Ioannidis [c], Federico Carbone [a,b,*]

Table 3
Summary of warnings/ risk factors from items related to paper mills and related unethical phenomena.

WARNING OR RISK FACTOR	Comment
At risk countries (A)	China, Iran, and Russia are most prominent producers that have been detected, but location may diversify over time; authors who buy authorship in paper mills may come from almost any country
ORCID ID lacking/ incomplete (A)	Not very specific criterion since many legitimate papers still lack ORCID
Suspicious e-mail address of authors (A, M)	E.g. not institutional, or weird institutions
Suspicious reviewers' e-mail address (A, M)	Same as above
Not consistent references (M)	Use of references/citations is very poor in the vast majority of the scientific literature, but with paper mills sometimes misuse may reach egregious levels; AI tools may not necessarily improve this, since they are trained on flawed previous papers with flawed use of references.
Plagiarism (M)	Classic plagiarism is detectable with standard software; more difficult with paraphrasing
Weird phrasing and terminology (M)	Forced rephrasing and misuse of scientific terms
Image quality (M)	Routine checking of images for signs of problems may reveal many problems that may be related to papermills or may be independent to them
Severe methodological weaknesses(M)	Not specific to paper mills, most published papers are weak methodologically
Manuscript formatting and style (M, J)	Unusual formatting and style in submission or requested by the journal
Non-provision of raw data (M, J)	Not specific, since most scientific papers unfortunately do not provide raw data, but some paper mill or other fraudulent studies may overtly seem implausible (e.g. data volume exceeds capacity of team/center)
Authorship change after acceptance (M, J)	Not specific to paper mills, but probably far more common, since authors are added even at such a late stage, some paper mills advertise that they sell authorship after acceptance
Detection of AI-generated material (M, J)	See text, methods of detection improve over time, but they are far from perfect, risk of false negatives and false positives
Too fast review time (J)	May herald a problem with the legitimacy of the journal and/or its ability to reject paper mill papers and other fraudulent or very low-quality work; extreme cases abound (e.g. highjacked sites of previously legitimate journals)
Similarities in reviewer comments (J)	Not specific, given the generally poor quality of peer review in most journals, but extreme similarity should raise concerns

COMMENT | 02 March 2026

AI agents are 'aeroplanes for the mind': five ways to ensure that scientists are responsible pilots

As artificial-intelligence systems take on more of the scientific workflow, the central goal should not be complete automation, but designing platforms that preserve creativity, responsibility and surprise.

Some concluding comments

- AI is a set of tools that can be used in both positive and negative ways in the quest for improving the reproducibility, transparency, and credibility of research and evidence as well as the synthesis, translation and dissemination thereof.
- Science and the scientific literature may be strengthened, accelerated, become boring, or obsolete.
- EBM specifically is at a crossroads to study, use, misuse, absorb or by absorbed by AI.