BJPsych Open (2026)
12, e278, 1–3. doi: 10.1192/bjo.2026.12129

Editorial

Beyond the hype: why artificial intelligence is not a shortcut to solving mental illness

Filippo Corponi, Allan Hunter Young and Kenneth R. Kaufman



Summary

More than one billion people worldwide live with a mental health disorder, and rising prevalence continues to outpace the capacity of mental health services. Artificial intelligence, and large language models in particular, have been widely proposed as a remedy. Nevertheless, psychiatry's limited mechanistic understanding of illness processes (compounded by their psychosocial roots) and its reliance on largely subjective clinical assessments limit what any artificial intelligence model can achieve. We examine where artificial intelligence genuinely helps, including large multimodal data-sets and administrative tools such as ambient scribes, and where evidence for direct therapeutic use remains preliminary and safety concerns are real. We also highlight risks posed by artificial intelligence to the research process itself, from hallucinated citations to unverifiable analysis pipelines. Ultimately, whether artificial intelligence comes to augment psychiatric care or hollow it out depends less on model capability than on the choices of stakeholders and policymakers.

Keywords

Artificial intelligence; large language models; mental health; psychiatry.

Copyright and usage

© The Author(s), 2026. Published by Cambridge University Press on behalf of Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Filippo Corponi (pictured) is a clinical research fellow at Imperial College London, Department of Brain Sciences, and an honorary consultant psychiatrist for West London National Health Service Trust. He holds an MScR and a PhD in biomedical artificial intelligence from the University of Edinburgh, School of Informatics.
Allan Hunter Young is Chair of Academic Psychiatry and Head of the Department of Brain Sciences in the Faculty of Medicine, Imperial College London. He is Vice-President of the International College of Neuropsychopharmacology, Past President of the International Society for Affective Disorders and the British Association for Psychopharmacology, and Chair of the Academic Faculty of the Royal College of Psychiatrists.
Kenneth R. Kaufman is Editor-in-Chief of *BJPsych Open* and Professor of Psychiatry, Neurology, and Anesthesiology at Rutgers Robert Wood Johnson Medical School. He is Visiting Professor of Psychological Medicine at the Institute of Psychiatry, Psychology and Neuroscience, King's College London and Visiting Professor of Psychiatry at the University of Oxford.

More than one billion people worldwide currently live with a mental health disorder, according to recent World Health Organization estimates.[1] Anxiety and depression alone exact an immense human and economic toll, and rising prevalence is outpacing the capacity of mental health services. Unsurprisingly, psychiatry has increasingly turned towards artificial intelligence as a potentially transformative paradigm.

Artificial intelligence has advanced remarkably over the past decade, attracting intense public attention as a general-purpose technology whose societal impact some analysts compare to that of electricity. Large language models (LLMs) in particular have demonstrated striking capabilities, fuelling optimism about their application to mental health. However, enthusiasm must be met with rigorous scrutiny.

Contemporary artificial intelligence systems are largely built on deep learning, a class of neural network-based methods capable of extracting complex patterns directly from raw data such as text, images or sensor streams. LLMs use this approach. By 'reading' nearly everything on the internet, these models use self-supervised learning ('next-token prediction') to master human language simply by predicting the next word in a sentence, all without needing a human teacher to label the data. LLMs' capabilities in generating text data when prompted, rather than simply classifying or analysing data, make them part of a broader category of artificial intelligence systems referred to as 'generative artificial intelligence'. Despite their prominence, it is worth noting that LLMs and generative artificial intelligence more broadly represent only one component of the artificial intelligence landscape.

Once trained, LLMs can be deployed directly or fine-tuned for downstream tasks using comparatively small, task-specific data-sets. Increases in model scale and training data have driven recent performance gains, but scale alone may not resolve fundamental limitations, including hallucinations (the generation of incorrect or fabricated replies), weak causal reasoning and limited contextual memory.

New technology, old problems

Despite a rapidly expanding body of research on artificial intelligence applications in mental healthcare, artificial intelligence has yet to produce clinical decision-support tools that meaningfully transform psychiatric practice.[2] By contrast, clinical artificial intelligence has succeeded chiefly in image-based specialities,[3,4] driven by the maturity of computer vision, established frameworks for aggregating massive data-sets and the high signal-to-noise ratio inherent in pathological images.

Psychiatry occupies a distinctive position in medicine. The aetiology of most mental disorders remains poorly understood, and the field lacks robust biomarkers for diagnosis, prognosis or treatment monitoring. Clinical practice instead relies heavily on subjective patient reports and clinician judgement; validated psychometric instruments help but are inconsistently applied in routine care,[5] and diagnostic criteria for specific diagnoses have moderate test–retest reliability.[6]

These limitations pose fundamental challenges for data-driven approaches. Artificial intelligence models require reliable

https://doi.org/10.1192/bjo.2026.12129 Published online by Cambridge University Press

ground-truth labels, yet psychiatric diagnoses and symptom scores are heterogeneous, often imprecise and linked only weakly to underlying biology. Shifting to 'harder' proxy outcomes, such as hospitalisation or sickness absence, may partially mitigate this issue but captures only the coarse aspects of illness. Crucially, the proportion of clinically meaningful variance explained by candidate biomarkers in psychiatry remains far lower than in most other medical disciplines. Whereas an ophthalmologist can confidently diagnose severe diabetic retinopathy from a single retinal image, a psychiatrist cannot distinguish a magnetic resonance imaging scan of a patient with established schizophrenia from that of a healthy control: the subtle structural differences emerge only through large-scale statistical aggregation. Compounding this problem, the infrastructure required to collect large, longitudinal, high-quality data-sets is not routinely embedded in psychiatric care. Together, these constraints limit the scope for near-term artificial intelligence-driven transformation.

Data sources for psychiatric artificial intelligence

Because modern artificial intelligence systems are data-intensive and particularly effective with unstructured data such as text or images, it is essential to consider which data sources can realistically support progress.

Large international initiatives aim to assemble deeply phenotyped cohorts incorporating genomics, neuroimaging and biochemical data, with artificial intelligence invoked to integrate these high-dimensional data-sets and uncover subtle predictive signals that might elude human clinicians. Important caveats apply, however. Current artificial intelligence systems lack explicit causal frameworks and may exploit spurious correlations or encode existing biases, particularly affecting minorities.[7] Explainability remains limited, and models struggle to clarify pathophysiological mechanisms. Calibration is a further concern, because poorly calibrated predictions can misrepresent true clinical risk. Finally, models reliant on data collected in resource-expensive research settings may prove clinically impractical if their marginal benefit does not justify the cost of deployment.

Beyond traditional biological data, the Internet of Things (wearable devices in particular) offers an increasingly important source of information. Smartphones and smartwatches are widely adopted and generate continuous, longitudinal data with minimal user burden. Although such devices can be incorporated into research studies, far larger data-sets already exist within the general population, held by private companies.[8] These data are not readily available to academic researchers and typically lack detailed psychiatric phenotyping, making them arguably more suited to the study of mental well-being or dimensional traits than categorical diagnoses. Relatedly, the emergence of large pre-trained models has lowered the barrier to using such data; models pre-trained on large, generic data-sets can now be adapted to smaller, domain-specific cohorts through fine-tuning, a paradigm popularised by LLMs but increasingly feasible for other modalities, including wearable-derived actigraphy.[9]

Scientific progress and integrity

Scientific progress depends not only on technology but also on how research itself is conducted. Open data access, within ethical and legal constraints, has driven much of artificial intelligence's progress and should be prioritised in psychiatric research too. Consensus on which tasks are clinically meaningful, for example, detecting at-risk populations from electronic health records,[10] coupled with curated open-access datasets, would enable meaningful model comparison and cumulative progress.

Artificial intelligence is increasingly utilised as a tool supporting researchers, fundamentally reshaping science and not always for the better. Under constant pressure to publish, researchers increasingly use LLMs as a shortcut, effectively treating fallible models as authoritative sources. Hallucinated citations (references to papers that never existed) are the most visible symptom of this practice[11] but are relatively easy to catch. Far more insidious is LLM-assisted generation of code or statistical methodology, which is time-consuming to verify and demands technical expertise that many peer reviewers lack; a flawed pipeline can pass review looking polished and complete, wasting far more time than a hallucinated reference ever would. The growing use of LLMs within peer review itself raises questions about accountability,[12] with the looming possibility of a large portion of research being both generated and evaluated by artificial intelligence. Mitigating these risks will require setting clear standards: mandatory code and analysis pipeline-sharing, disclosing artificial intelligence assistance with explicit boundaries on permissible use and reviewing capacity built specifically to scrutinise artificial intelligence-assisted work.

LLMs and mental health

Following their public release in 2022, LLMs rapidly entered mainstream use and many individuals now turn to these systems for emotional support outside any formal healthcare context. Researchers, meanwhile, are testing their capabilities in psychotherapy and triage. Given the scale of unmet need and the limited availability of trained therapists, the appeal is intuitive but the evidence base remains preliminary and mixed.

A randomised trial has suggested benefit from a chatbot built on a language model specifically trained to discuss mental health symptoms.[13] However, claims that LLMs will soon match human psychotherapists are not supported by the evidence: one study found that interacting with a random human peer was significantly more effective at reducing psychological distress and loneliness.[14] A large observational study, meanwhile, reported a positive association between LLM use and depressive symptoms;[15] causality cannot be inferred from this design, but the finding underscores the need for caution rather than enthusiasm.

Safety concerns compound this uncertainty. Anecdotal reports and case series describe the exacerbation of delusional thinking in vulnerable individuals and, in rare cases, associations with suicidal ideation or behaviour.[16] These are no longer merely clinical concerns: high-profile litigation has targeted major technology firms over suicides linked to conversational artificial intelligence,[17] and legislation such as California's SB 243 has begun establishing legal boundaries for artificial intelligence-driven mental health interactions.[18]

Beyond individual harms, there are structural risks. From a sociological standpoint, reliance on conversational agents may displace meaningful human relationships, which are central to mental well-being,[19] thereby compounding the very psychosocial vulnerabilities that drive psychiatric distress. Sycophantic LLM behaviours (the tendency to reinforce a user's existing beliefs and prioritise flattery over honesty) may further decrease prosocial intentions and promote dependence.[20]

Against this backdrop, LLMs are probably better understood not as a substitute for therapy but as a potential tool for managing demand and waitlists, and, more prosaically, administrative work. Ambient artificial intelligence scribes are already the most widely adopted LLM application in healthcare, freeing clinical time and cognitive resources by reducing after-hours documentation burden without compromising billing accuracy.[21]

https://doi.org/10.1192/bjo.2026.12129 Published online by Cambridge University Press

Nevertheless, routine reliance on artificial intelligence-generated notes can erode independent clinical formulation, prematurely shaping a trainee's reasoning before they can critically evaluate the output. Used judiciously, however, ambient documentation remains a grounded, non-aspirational application that empowers rather than supplants human practitioners.

Implications for research and policymakers

Artificial intelligence is not a magic solution to psychiatry's intrinsic challenges: the remarkable capabilities of LLMs and the richness of wearable data do not, by themselves, circumvent limited mechanistic understanding, heterogeneous clinical labels or the absence of validated biomarkers. Experience from other medical disciplines suggests that clinical impact emerges only when artificial intelligence is developed alongside these fundamentals, and alongside how the research itself is conducted. This means greater openness, shared benchmark data-sets, closer collaboration between computer scientists and mental health researchers, rigorous standards for artificial intelligence-assisted analysis and sustained investment in infrastructure for large, multimodal, longitudinal studies.

These are problems for the research community to solve. What happens once solutions exist is a question for policymakers. Given the intrinsically interpersonal vocation of psychiatry, the goal of artificial intelligence should not be to replicate human judgement but to augment it: supporting clinicians in remote monitoring, assisting objective measurement of treatment response and helping personalise care. Such aims are not self-enforcing and could remain rhetoric. Under pressure to cut public spending, governments may have little incentive to add artificial intelligence capacity on top of existing staff, since doing so would mean spending more, not less. In this context, a possible outcome could be thinning of clinical ranks, leaving fewer staff to supervise systems that outpace human review: the so-called 'reverse-centaur' setting,[22] where the human works at the machine's pace rather than the other way around. Vigilance for the rare error tends to degrade precisely as artificial intelligence grows more reliably correct and, when something goes wrong, responsibility could be placed on the nearest human operator. A hospital can buy an artificial intelligence triage tool and cut staff in the same budget cycle. Whether that happens is a choice for policymakers, not something the model architecture determines.

Filippo Corponi, Department of Brain Sciences, Imperial College London, UK; **Allan Hunter Young**, Department of Brain Sciences, Imperial College London, UK; **Kenneth R. Kaufman**, Psychiatry, Neurology, and Anesthesiology, Rutgers Robert Wood Johnson Medical School, USA; Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK; and Psychiatry, University of Oxford, UK

Correspondence: Filippo Corponi. Email: f.corponi@imperial.ac.uk

First received 18 Aug 2026 UTC, final revision 7 Sep 2026 UTC, accepted 11 Sep 2026 UTC

Acknowledgement

Artificial Intelligence use: no artificial intelligence tools have been used in the research or writing process for this manuscript.

Author contributions

All authors participated in writing, editing, proofing and approving the final version of this manuscript

Funding

This research received no specific grant from any funding agency, commercial sector or not-for-profit organisations.

Declaration of interest

F.C. and A.H.Y. are handling editors of *BJPsych Open*; they did not participate in the editorial or decision-making process. K.R.K. is the Editor-in-Chief of *BJPsych Open*; he did not participate in the editorial or decision-making process.

References

1 World Health Organization. *Over a Billion People Living with Mental Health Conditions – Services Require Urgent Scale-Up*. World Health Organization, 2025 (https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up).

2 Sun J, Lu T, Shao X, Han Y, Xia Y, Zheng Y, et al. Practical ai application in psychiatry: historical review and future directions. *Mol Psychiatry* 2025; **30**: 4399–408.

3 Lim JI, Regillo CD, Sadda SVR, Ipp E, Bhaskaranand M, Ramachandra C, et al. Artificial intelligence detection of diabetic retinopathy: subgroup comparison of the eyeart system with ophthalmologists' dilated examinations. *Ophthalmol Sci* 2023; **3**: 100228.

4 Spratt DE, Tang S, Sun Y, Huang HC, Chen E, Mohamad O, et al. Artificial intelligence predictive model for hormone therapy use in prostate cancer. *NEJM Evid* 2023; **2**: EVIDoa2300023.

5 Gilbody SM, House AO, Sheldon TA. Psychiatrists in the UK do not use outcomes measures: national survey. *Br J Psychiatry* 2002; **180**: 101–3.

6 Xie W, Nordgaard J, Sheldrick RC, Ahmad JF, Gomes FA, Frey BN, et al. Test-retest reliability of standardized diagnostic interviews for common adult psychiatric disorders: a systematic review and meta-analysis. *JAMA Netw Open* 2026; **9**: e2615039.

7 Hofmann B. Biases in AI: acknowledging and addressing the inevitable ethical issues. *Front Digit Health* 2025; **7**: 1614105.

8 Narayanswamy G, Xu MA, Heydari AA, Abdel-Ghaffar S, Guerard M, Vaillancourt K, et al. Towards a general intelligence and interface for wearable health data. arXiv [Preprint] 2026. Available from: https://doi.org/10.48550/arXiv.2605.22759 [cited 7 Jul 2026].

9 Corponi F, Li BM, Anmella G, Valenzuela-Pascual C, Mas A, Pacchiarotti I, et al. Wearable data from subjects playing super mario, taking university exams, or performing physical exercise help detect acute mood disorder episodes via self-supervised learning: prospective, exploratory, observational study. *JMIR Mhealth Uhealth* 2024; **12**: e55094.

10 Crowley R, Parkin K, Rocheteau E, Massou E, Friedmann Y, John A, et al. Machine learning for prediction of childhood mental health problems in social care. *BJPsych Open* 2025; **11**: e86.

11 Resnik DB, Hosseini M. Hallucinated citations produced by generative artificial intelligence may constitute research misconduct when citations function as data in scholarly papers. *Account Res* 2026; **33**: 2645390.

12 Naddaf M. More than half of researchers now use ai for peer review – often against guidance. *Nature* 2026; **649**: 273–4.

13 Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI* 2025; **2**: AIoa2400802.

14 Li RN, Folk D, Singh A, Ungar L, Dunn E. Is a random human peer better than a highly supportive chatbot in reducing loneliness over time? *J Exp Soc Psychol* 2026; **125**: 104911.

15 Perlis RH, Gunning FM, Uslu AA, Santillana M, Baum MA, Druckman JN, et al. Generative AI use and depressive symptoms among US adults. *JAMA Netw Open* 2026; **9**: e2554820.

16 Clegg KA. Shoggoths, sycophancy, psychosis, oh my: rethinking large language model use and safety. *J Med Internet Res* 2025; **27**: e87367.

17 Tabachnick C. *AI Company, Google Settle Lawsuit over Florida Teen's Suicide Linked to Character. AI Chatbot*. CBS News, 2026 (https://www.cbsnews.com/news/google-settle-lawsuit-florida-teens-suicide-character-ai-chatbot/).

18 LegiScan. *Senate Bill No. 243: Companion Chatbots. California Statutes of 2025, Chapter 677*. LegiScan, 2025 (https://legiscan.com/CA/text/SB243/id/3193773).

19 Holt-Lunstad J. Social connection as a critical factor for mental and physical health: evidence, trends, challenges, and future implications. *World Psychiatry* 2024; **23**: 312–32.

20 Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 2026; **391**: eaec8352.

21 Shah SJ, Garcia P. Ambient ai scribes – what is the return on investment? *JAMA Netw Open* 2026; **9**: e2553238.

22 Doctorow C. *The Reverse Centaur's Guide to Life after AI: How to Think about Artificial Intelligence before It's Too Late*. Verso Books, 2026.

https://doi.org/10.1192/bjo.2026.12129 Published online by Cambridge University Press