Responsible AI Integration in Mental Health Research: Issues, Guidelines, and Best Practices

Mirza Jahanzeb Beg[1] iD

In 2023, artificial intelligence (AI) captured global attention, reshaping various sectors and quickly integrating into daily life. Advanced generative models such as OpenAI's GPT-4, along with open-source innovations such as Llama 2 and Mistral 7B, expanded the capabilities of natural language processing and creative content generation, spurring an array of applications across industries. These advancements in AI not only transformed technology but also reached fields as diverse as healthcare, finance, autonomous vehicles, and education. Every day seems to bring new developments within what many are calling the "AI Bubble"—a period of accelerated AI development and investment, propelled by high expectations and speculative excitement but shadowed by concerns of a possible market correction reminiscent of past tech bubbles.[1]

The Promise and Paradox of AI's Rise: From Real Impact to Unseen Risks

AI has achieved impressive milestones, from surpassing human-level accuracy on the U.S. Medical Licensing Examination to acing the Turing's test and, most recently, exhibiting a sophisticated understanding of human mental states through the "theory of mind."[2]

In healthcare, particularly mental health services, machine learning models have identified effective psychological and behavioral interventions with a precision that sometimes rivals that of human professionals. AI promises to augment clinical reasoning, enhance decision-making, and simulate empathy, sparking discussions about AI's role in empathy training for physicians. AI-based chatbots offer scalable, personalized support, acting as virtual companions to make mental healthcare more accessible and compassionate by assisting with screenings, guiding therapy, and offering continuous support while complementing human professionals.[3] Nevertheless, despite these successes, experts need to be more cautious, warning against full reliance on AI for sensitive tasks such as answering patient queries or making medical recommendations without human oversight. Integrating AI into medical practice requires rigorous oversight due to potential inaccuracies, copyright issues, and algorithmic biases. AI should complement human expertise, guided by human judgment, to ensure patient safety and trust. Careful integration is essential to unlock AI's full potential while mitigating associated risks.

The cultural impact of AI-generated content is profound. Technologies that create art and music and simulate historical voices have sparked both wonder and ethical debates. For instance, the AI-assisted production of a posthumous Beatles song last year has recently been nominated for a Grammy, exemplifying AI's ability to merge past and present in unprecedented ways. On another front, AI-driven devices, such as the Humane AI Pin—a wearable device that integrates AI into daily interactions—signal an ever-closer relationship between humans and machines. These innovations raise questions about copyright, intellectual property, and the nature of data itself while also highlighting critical concerns regarding data privacy and the ethical use of personal information. This evolving landscape has even given rise to "machine unlearning," where the right to be forgotten mandates that

[1]Dept. of Psychology, Kumaraguru College of Liberal Arts and Science, Coimbatore, Tamil Nadu, India.

Address for correspondence: Mirza Jahanzeb Beg, Dept. of Psychology, Kumaraguru College of Liberal Arts and Science, Coimbatore, Tamil Nadu 641049, India.
E-mail: mirzajahanzebbeg@gmail.com

§ Sage

ACCESS THIS ARTICLE ONLINE
Website: journals.sagepub.com/home/szj
DOI: 10.1177/02537176241302898

machine learning applications remove specific data from a dataset and retrain the model upon user request.[4]

The rapid pace of AI innovation presents benefits but also significant concerns, prompting policymakers in the United States, European Union, and China to develop regulatory frameworks that balance fostering innovation with ethical oversight. A major concern is the rise of AI-generated fake images and videos, or "deepfakes," which have fueled scams and misinformation campaigns. For instance, a deepfake of Pope Francis went viral, highlighting AI's potential to spread false information and create confusion on a global scale.[5] Additionally, AI's ability to mimic voices and faces raises serious risks for identity theft and fraud. The prospect of superintelligent AI raises existential risks, with fears of an entity surpassing human control and altering decision-making processes. Policymakers are also addressing the surge of AI-generated emails and content designed to influence public policy and opinion, which can compromise democratic processes.[6] Ethical and philosophical questions concerning AI governance remain central to policy discussions. The misuse of AI in generating synthetic content has amplified misinformation, threatening societal stability and electoral integrity. The World Economic Forum's Global Risks Report 2024 identifies misinformation as a pressing short-term risk with the potential to disrupt civil order and erode trust in media and government institutions. Policymakers face the challenge of crafting regulations that prioritize societal safety without stifling AI's innovative potential.[7]

The AI in Scientific Writing: Accountability, Accuracy, Ethical Challenges, and Regulatory Considerations

The rapid integration of AI into academic writing, healthcare, and other domains presents complex ethical, legal, and intellectual challenges. Soon after the release of ChatGPT, there was a notable boom in the sale of AI-generated e-books on Amazon. The rise of large language models (LLMs) such as ChatGPT has significantly impacted academic publishing. Some authors have controversially cited ChatGPT as co-authors

in research, sparking debate in academic circles.[8] Numerous academic journals have published articles containing AI-generated content, where authors neglected to remove indicators of AI involvement, such as phrases like, "Certainly, here's a possible introduction for your topic" and "I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model."[9] These oversights have led to retractions and raised serious concerns about authenticity and credibility. They expose vulnerabilities in the peer-review process and pose risks to scientific integrity.

The integration of AI in scientific writing presents significant ethical challenges, particularly concerning accountability, accuracy, originality, intellectual property, data privacy, and bias, thus warranting strict regulatory oversight.[10] Unlike human authors, AI systems cannot assume responsibility for the content they generate, undermining scientific integrity. Due to limitations in their training data, AI-generated articles risk unintentional plagiarism, reinforcement of biases, and inaccuracies. Without human oversight, such content lacks the critical analysis, depth, and creativity inherent in human authorship, potentially leading to unreliable conclusions. Moreover, the inconsistent disclosure of AI involvement in academic work complicates peer review, raising concerns about transparency and intellectual honesty in scholarly publications.

The increasing use of AI tools in research writing, such as Grammarly or QuillBot, has proven beneficial for language editing and minor formatting adjustments, and many journals permit their use, provided that the use of AI is declared. However, full-scale reliance on LLMs such as ChatGPT to generate academic content introduces profound ethical and practical concerns. Chief among these issues are copyright violations and the risk of "AI hallucinations"—instances where AI fabricates or distorts information, potentially leading to inaccurate or misleading scholarly work.[11] For example, the Google Bard chatbot inaccurately claimed that the James Webb Space Telescope captured the first images of an exoplanet.[12] Such errors not only mislead the public but also highlight the need for rigorous validation and human oversight, particularly in fields such as healthcare, where misinformation can have serious consequences.

The "black box" nature of many AI models, where the decision-making processes are opaque, further complicates accountability.[13] This lack of transparency impedes users from diagnosing errors or understanding how AI systems arrive at their conclusions, making it difficult to trust and improve these technologies.

Bias in AI systems, which can arise from societal prejudices embedded in training data or algorithms, presents a significant challenge to fairness and objectivity.[14] Various forms of bias—including algorithmic, data, and implicit bias—can lead to discriminatory outcomes, such as reinforcing stereotypes or skewing results based on over- or under-representation of certain groups. These biases can affect multiple aspects of AI-generated content, from research findings to automated decision-making tools. For instance, algorithmic bias may amplify existing social inequalities, while data biases such as historical or sample bias can distort the accuracy of results across different demographics. Moreover, reporting bias, group attribution bias, and implicit bias can misrepresent or generalize traits across populations, undermining the neutrality of AI outputs. Addressing these issues requires careful curation of training data, constant monitoring, and human oversight to ensure that AI systems are used responsibly and do not perpetuate harmful stereotypes or inequities.

Privacy concerns play a critical role in the ethical use of AI, as many AI systems rely on vast amounts of personal data. This reliance raises important questions about data collection, storage, and usage. Instances of inadvertent privacy breaches underscore the need for robust data protection and transparent usage policies to ensure that individuals' privacy is respected and safeguarded. Simultaneously, the integration of AI-generated content in academic writing presents significant challenges regarding intellectual property rights. As AI systems become more sophisticated, the difficulty in determining authorship and ownership of AI-generated works intensifies. Traditional copyright laws need to be adequately equipped to address these complexities, leading to potential disputes over authorship and rightful attribution. This ambiguity impacts the protection of original ideas and the enforcement of copyright laws, underscoring the urgent need for updated legal frameworks to ensure

fair and transparent practices. Without clear guidelines and regulatory oversight, the risk of intellectual property disputes and potential exploitation of AI-generated content increases, highlighting the pressing need for new legal standards to safeguard creators' rights and maintain the integrity of intellectual property.[15]

Furthermore, AI systems have the potential to offer harmful or unsafe advice, as demonstrated by AI-driven meal planners suggesting recipes that could produce toxic gas.[16] Such risks highlight the importance of integrating safety protocols and human oversight into AI systems, particularly in applications that affect user well-being.

The overreliance on AI in academic and professional settings also poses significant risks to critical thinking and intellectual development.[17] As users become more dependent on AI-generated content, there is a growing concern that this dependency may diminish original thought and the development of independent decision-making skills. This shift can lead to a reduction in cognitive engagement, as individuals may increasingly rely on AI for answers rather than engaging in deep, analytical thinking themselves. The long-term implications of this trend are profound, potentially stunting cognitive development and undermining academic integrity. It is crucial to use AI as a tool to augment human capabilities rather than replace them, ensuring that the development of critical thinking and intellectual independence remains a priority in educational and professional environments.

Given these significant concerns, it becomes increasingly important to use AI responsibly in academic writing. The potential for privacy breaches and intellectual property disputes underscores the necessity of avoiding the use of AI to generate entire articles. Instead, AI should be employed as a supportive tool to refine and enhance writing rather than to generate content, not as a substitute for human creativity and critical thinking. Ensuring that human creativity and critical thinking remain central to the research process is essential. By adhering to ethical guidelines and maintaining rigorous oversight, researchers can ensure that AI contributes positively to academic integrity and the advancement of knowledge rather than compromising these foundational principles.

Ethical Guidelines and Best Practices for the Responsible Use of AI in Academic Writing

To address the ethical challenges surrounding AI, leading research bodies, including the Committee on Publication Ethics and the World Association of Medical Editors (WAME), have issued guidelines emphasizing its limitations in academic authorship.[18] They stress that AI should not generate substantive content, as AI cannot accept accountability, manage conflicts of interest, or navigate copyright responsibilities—functions that require human oversight. Notably, listing AI tools as co-authors or allowing them to generate primary research is discouraged within the research community. The Indian Council of Medical Research has also issued guidelines for the responsible use of AI in biomedical research and clinical practice, emphasizing transparency, fairness, security, and accountability.[19] These guidelines stress the need for rigorous oversight, particularly in high-stakes areas such as healthcare, where AI decisions directly impact human lives. Similarly, researchers in other fields must adopt comprehensive governance frameworks that prioritize transparency, accountability, and privacy to ensure ethical AI deployment. These frameworks should support AI innovation while preserving public trust and ensuring responsible technological advancement.

The AI's role in research writing can significantly enhance academic productivity and efficiency, but its use must be managed carefully to uphold academic integrity. Transparency is crucial in preventing plagiarism and intellectual dishonesty; researchers must disclose AI involvement clearly, including specifying the model used and the extent of its contribution. Additionally, the potential for AI to generate inaccuracies, or "artificial hallucinations," necessitates rigorous fact-checking to ensure the credibility of AI-assisted work. Biases in AI outputs, shaped by historical, cultural, or algorithmic factors, can further compromise the objectivity of research. Ensuring the neutrality of AI-generated content requires careful selection of training data and ongoing human oversight to prevent the reinforcement of stereotypes or the misrepresentation of certain groups.

Adhering to best practices in responsible AI use in academic writing is essential for maintaining scientific integrity and ethical standards. This includes transparency in disclosure, validation of AI-generated content, and a clear recognition of AI's limitations. While AI can be a useful tool for tasks such as literature reviews, language refinement, and idea generation, it should never replace human intellectual input, critical thinking, or ethical judgment. Researchers are ultimately responsible for the accuracy, reliability, and overall quality of their AI-assisted work, and they must take the lead in educating others about the ethical use of AI in academic settings.

To ensure scientific integrity, researchers must verify AI-generated results, disclose AI usage fully in research articles, and explain how results were validated, ensuring that tools and methods are transparent. Proper attribution of AI-generated content is crucial to avoid inaccuracies, fabricated references, or intellectual property violations. Oversight bodies, including strategic councils, should be responsible for addressing the ethical implications of AI in academic research. Public engagement and education are vital to maintaining trust in AI technologies, emphasizing responsible and ethical use of AI.

Independent verification of AI-generated results is necessary to ensure rigorous testing before they are integrated into scientific workflows, as reproducibility and accuracy are critical. These practices, along with policies that prevent AI from being listed as an author, will ensure that AI's role in research adheres to the highest academic and ethical standards.

When AI (e.g., LLMs, NLP, machine learning, chatbots) is used in research, authors should adhere to relevant reporting guidelines such as Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI), Checklist for Artificial Intelligence in Medical Imaging (CLAIM), and others, ensuring that each guideline element is reported with sufficient detail to enable reproducibility. AI-generated content must respect copyright laws, avoid including identifiable patient information, and include proper permissions for AI-generated material. The methods section should specify AI usage, including platforms, prompts, datasets, and details about machine learning models, performance metrics, and steps taken to address AI-related biases. Results should include performance

comparisons, uncertainty measures, and the reporting of any AI-related errors or missing data. The discussion section must address the potential for AI bias, inaccuracies in AI-generated content, and the generalizability of findings, especially when considering underrepresented groups, and include data-sharing statements when applicable.[20]

The ethical AI use in academia also requires researchers to verify AI outputs, safeguard against plagiarism, and protect intellectual property. Data privacy is a critical concern, particularly when AI tools rely on sensitive information that could compromise confidentiality. Researchers must adhere to data protection regulations and ensure that AI-generated content is properly cited. Furthermore, to prevent overreliance on AI, researchers should maintain their critical thinking skills and foster accountability in AI-influenced research.

Conclusion

Responsible integration of AI in mental health and academic research requires researchers to navigate complex ethical and practical considerations, upholding transparency, accountability, and rigor. As AI tools advance, they should be used as complement to human expertise, enhancing productivity and supporting, not replacing, critical thinking and intellectual engagement. Ethical guidelines must be followed to ensure accurate reporting of AI use, transparent validation of results, and responsible data handling. Researchers should openly disclose AI involvement, clarify the role of AI in their methods, and address potential biases, inaccuracies, and privacy risks associated with AI-generated content. Through interdisciplinary collaboration, adherence to established reporting standards, and public education on AI's responsible use, the research community can harness AI's benefits while safeguarding academic integrity and fostering trust in technological advancements.

ORCID iD

Mirza Jahanzeb Beg (iD) https://orcid.org/0000-0002-6308-2678

References

1. Floridi L. Why the AI hype is another tech bubble. *Philos Technol* 2024; 37. Epub ahead of print November 5, 2024. DOI: 10.1007/s13347-024-00817-w.
2. Strachan JWA, Albergo D, Borghini G, et al. Testing theory of mind in large language models and humans. *Nat Hum Behav* 2024; 8: 1285–1295.
3. Beg MJ, Verma M, Chanthar KMMV and Verma MK. Artificial intelligence for psychotherapy: A review of the current state and future directions. *Indian J Psychol Med*. Epub ahead of print July 28, 2024. DOI: 10.1177/02537176241260819.
4. Zhang H, Nakamura T, Isohara T, et al. A review on machine unlearning. *SN Comput Sci* 2023; 4. Epub ahead of print April 19, 2023. DOI: 10.1007/s42979-023-01767-4.
5. Perrigo B and Johnson V. How to spot an AI-generated image like the 'Balenciaga Pope'. *TIME*. 2023. Accessed November 10, 2024, https://time.com/6266606/how-to-spot-deepfake-pope/
6. Freedom House. The repressive power of artificial intelligence, 2023. Accessed November 10, 2024. Available from: https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence
7. World Economic Forum. Global Risks Report 2024. Accessed November 10, 2024. Available from: https://www.weforum.org/publications/global-risks-report-2024/
8. Stokel-Walker C. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature* 2023; 613: 620–621.
9. Oransky AI. Weekend reads: A paper written by ChatGPT goes viral; the Gino misconduct investigation report; superconductivity scandal. *Retraction Watch*, 2024. Accessed November 10, 2024. Available from: https://retractionwatch.com/2024/03/16/weekend-reads-a-paper-written-by-chatgpt-goes-viral-the-gino-misconduct-report-superconductivity-scandal/
10. Meskó B and Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023; 6. Epub ahead of print July 6, 2023. DOI: 10.1038/s41746-023-00873-0.
11. Alkaissi H and McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus.* Epub ahead of print February 19, 2023. DOI: 10.7759/cureus.35179.
12. Pearl M. Bard shared false information in Google's announcement. *Mashable India*, 2023. Accessed November 10, 2024, Available from: https://in.mashable.com/tech/46866/bard-shared-false-information-in-googles-announcement
13. Von Eschenbach WJ. Transparency and the black box problem: Why we do not trust AI. *Philos Technol* 2021; 34: 1607–1622.
14. Panch T, Mattie H and Atun R. Artificial intelligence and algorithmic bias: Implications for health systems. *J Glob Health* 2019; 9. Epub ahead of print November 24, 2019. DOI: 10.7189/jogh.09.020318.
15.s Heidt A. Intellectual property and data privacy: The hidden risks of AI. *Nature.* Epub ahead of print September 4, 2024. DOI: 10.1038/d41586-024-02838-z.
16. Novak M. Supermarket AI gives horrifying recipes for poison sandwiches and deadly chlorine gas. *Forbes.* 2023. Accessed November 10, 2024. Available from: https://www.forbes.com/sites/mattnovak/2023/08/12/supermarket-ai-gives-horrifying-recipes-for-poison-sandwiches-and-deadly-chlorine-gas/
17. Zhai C, Wibowo S and Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learn Environ* 2024; 11. Epub ahead of print June 18, 2024. DOI: 10.1186/s40561-024-00316-7.
18. Zielinski C, Winker MA, Aggarwal R, et al. Chatbots, generative AI, and scholarly manuscripts. *Colomb Med* 2023; 54: e1015868.
19. Indian Council of Medical Research, Government of India. Ethical guidelines for application of artificial intelligence in biomedical research and healthcare. *Indian Council of Medical Research*, 2024. Accessed November 10, 2024. Available from: https://www.icmr.gov.in/ethical-guidelines-for-application-of-artificial-intelligence-in-biomedical-research-and-healthcare
20. Flanagin A, Pirracchio R, Khera R, et al. Reporting use of AI in research and scholarly publication-JAMA Network Guidance. *JAMA* 2024; 331: 1096.