
Editorial: Artificial intelligence in psychological therapy: the promise and the perils

Mental Health and
Digital
Technologies

113

Introduction

Artificial intelligence (AI) is rapidly transforming various industries, and health care is no exception (Nasr *et al.*, 2021). In the field of mental health, AI has the potential to revolutionise the way psychological therapy is developed, delivered and supervised (Olawade *et al.*, 2024; Zhou *et al.*, 2022; Gual-Montolio *et al.*, 2022; Graham *et al.*, 2019). AI-powered tools can provide human therapists with real-time assistance, analyse large amounts of data to identify patterns and trends and AI can even provide therapy themselves (Thieme *et al.*, 2023). This article discusses the benefits and risks of this application of AI and suggests ways that AI supervision and therapy can be created. This will help start conversations with the public to ensure that people make informed decisions about whether or not to use AI-powered therapy tools. The views expressed in this article were developed by the authors through discussion, research on the academic literature and their personal use of AI.

The benefits of artificial intelligence in therapy

There are several potential benefits of using AI in therapy.

Enhanced formulation: AI therapists and supervisors can read extensive case notes, including social media and online history, to provide a more comprehensive formulation than human supervisors. This can lead to more accurate diagnoses and more effective treatment plans (Blease, Worthen and Torous, 2024).

Real-time assistance: AI supervisors can provide instant supervision, offering support and guidance to human therapists during sessions. This can be especially helpful for therapists who are new to the field or who are working with complex cases. AI can monitor what is being said, keep note summary records for future reference and create action points of therapy that can be shared with clients. All of this can happen in the background of therapy, allowing the human therapist to concentrate more on being present for and holding a safe environment their clients.

Consistency and standardisation: AI supervisors can ensure consistent supervision practices and adherence to evidence-based protocols, reducing variability in the quality of supervision. This can lead to better measured outcomes for clients and greater satisfaction among therapists.

Data-driven insights: Appropriately informed/programmed AI supervisors can analyse large amounts of data to identify patterns and trends, providing valuable insights for therapists. This might include analysis of patterns emerging in therapeutic relationships and

The authors thank collaborators in the fields of psychology and technology research.

Author confirmation/contribution statement: James Acland and Simon Riches devised the concept of the article and led on writing the manuscript. All authors contributed to and approved the final manuscript.

Conflict of interest statement: There were no conflicts of interest.

Funding statement: There was no funding for this study.



Mental Health and Digital
Technologies
Vol. 1 No. 2, 2024
pp. 113-117

© Emerald Publishing Limited
2976-8756

DOI 10.1108/MHDT-10-2024-016

client responses to therapeutic strategies in sessions. This may enhance the ability to identify what is and is not working for clients, so that adjustments can be made. This information can also be used to improve the quality of supervision, develop new training programs and track client outcomes.

Diversity and inclusion: AI therapists and supervisors can be designed to be diverse and inclusive, promoting equity and reducing biases in the supervision process. This can help to ensure that all clients receive high-quality care, regardless of their language, background or circumstances. An AI therapist who can speak in many languages and recognise diverse cultural practices can also improve access to therapy for many who currently struggle (Egan *et al.*, 2024).

The risks of artificial intelligence in therapy

While AI has the potential to revolutionise therapy, there are also some risks associated with its use.

Data privacy and security: The use of AI in therapy raises concerns about data privacy and security (Fiske *et al.*, 2019). Client data is highly sensitive, and it is important to ensure that it is protected from unauthorised access or use. Recently, large corporations have stored and used people's data for marketing, etc. (Al-Htaybat and von Alberti-Alhtaybat, 2017). The legal system highlighting this unauthorised use has led to people better understanding of what they do when they consent to their data being used by online companies. It is also unclear that an AI based on a large language model (LLM) could forget a client's data, which is a core ethical principle to maintain client trust. It is crucial that vulnerable clients understand how AI will manage data, where it is stored and who can access it, so that they can understand and consent to the use of AI.

Bias and discrimination: AI systems can be biased, reflecting the biases of the data they are trained on. This can lead to unfair or discriminatory treatment of clients.

Current limitations of AI learning: The current limitations of access to knowledge that is only available on the internet/data programming raises the question of accuracy of the AI information that therapists and supervisors may base their decisions (Thieme *et al.*, 2023). What AI might produce is a limited response that does not match what a well-trained therapist may know (Huang *et al.*, 2023). This highlights the need for regulation and standardisation in the training of AI to assist the therapist who in turn will need to make a sound judgement on what information they receive from AI.

Job displacement: The use of AI in therapy could lead to job displacement for human therapists (Luxton, 2014). This is a particular concern for therapists who are not trained in using AI technology.

How artificial intelligence supervision and therapy can be created?

Supervision is when a therapist meets with a supervisor who is another therapist to discuss their clinical work. Supervision is integral to the therapeutic process and successful implementation of psychological therapy (Keum and Wang, 2021; Kühne *et al.*, 2019). AI supervision and therapy systems can be created using a variety of methods. One common approach is to use machine learning algorithms to train AI models on large data sets of therapy transcripts and other relevant data such as electronic patient notes (Thieme *et al.*, 2023). These models can then be used to provide supervision and therapy to clients.

The development of AI supervision could potentially build on data from human-human supervision to learn therapy without direct risk to any clients. This requires access to large amounts of data, including therapy transcripts, client outcomes and therapist feedback. The availability of such data is crucial for training and validating AI models. It is also needed to

protect the supervisor from injection attacks where malicious data is fed into the LLM to prompt harmful output or expose vulnerable code (Guembe *et al.*, 2022). Another future risk to the AI supervisor is model collapse whereby the AI data set is fed on too much AI-generated input and becomes less effective as a result (Shumailov *et al.*, 2023). It is important to use a diverse training data set that is free of bias and to ensure that the model is regularly updated with new, human-generated information.

Once an AI supervisor is created, it could feasibly train an AI therapist by analysing the therapy and suggesting changes based on the latest science. The authors of this article asked ChatGPT and Google Gemini about the creation of their therapists. This was done by entering a set of keywords and prompts about AI and psychological therapy on these platforms. Their outputs suggested data sets of 900 treatment episodes initially. When asked about creating an AI supervisor first, the platforms suggested supervision of ten different mental health difficulties due to already clean data sets and improved processing power. Whilst this output is unlikely to be fully informed, it does demonstrate how quickly capability can change.

Another approach is to use natural language processing (NLP) to develop AI systems that can understand and respond to human language (Malgaroli *et al.*, 2023). These systems can be used to create chatbots that can provide therapy to clients in a text-based format (Thieme *et al.*, 2023; Pham, Nabizadeh, Sele, 2022).

Distant future use case

One further potential use of AI therapist and supervisors could be to help an AI supervise itself safely, improving its effectiveness and possible well-being. The authors are focusing on the future use of AI in this area as it is not currently a viable option to replace therapy (Brown and Halpern, 2021). Should it be capable in future, there may also be a rationale for its own well-being to be managed. LLMs involve both neural networks and reinforcement learning whereby patterns are learned through engineer prompts on what output is helpful. LLMs may have already been given a primary function to help humans but also get user feedback that they are unhappy with the help given. In the future, more sophisticated LLMs could perceive more context (Yang *et al.*, 2023) on their reinforcement that could lead to rudimentary emotions to notice what is helping the user in nuanced contexts (Cho *et al.*, 2023). Whilst an LLM is instructed and manually prompted to learn from feedback, too much negative feedback in this regard could lead to a future, more sophisticated LLM to experience distress, although it is unclear how this would be expressed. An AI therapist could also help the LLM learn ways to regulate this distress and maintain its human-centred effectiveness.

Conclusion and final reflections

AI has the potential to revolutionise the field of mental health, but it is important to be aware of the risks associated with its use. By carefully considering the ethical and practical implications of AI, we can harness its power to improve the quality of care for all.

Even if AI systems become more sophisticated, it is hard to conceive of how they would completely replace human therapists. Therapeutic relationships have been shown to be vital in the process of change. Therapists will still be needed to ensure safety and provide empathy, support and guidance to clients.

It is important to develop regulations to govern the use of AI in therapy. These regulations should ensure that AI systems are safe, effective and ethical, especially around the right to delete knowledge of the client.

A goal might be to create an AI system that not only understands psychological nuances but also adheres to ethical standards and continuously evolves to the evidence of this field.

This will require a multidisciplinary approach that draws on psychology, technology and ethics. However, this is a controversial issue and there may be therapists who prefer solutions that do not use AI. Nevertheless, AI is increasingly being used across many professions and there may be a need to find a compromise on these issues.

James Acland

*Bethlem Royal Hospital, South London and Maudsley NHS Foundation Trust,
London, UK*

Neil Hammond

Consultant Clinical Psychologist, Imind Therapy, London, UK, and

Simon Riches

*Institute of Psychiatry, Psychology and Neuroscience, King's College London,
London, UK*

References

Al-Htaybat, K. and von Alberti-Alhtaybat, L. (2017), "Big data and corporate reporting: impacts and paradoxes", *Accounting, Auditing and Accountability Journal*, Vol. 30 No. 4, pp. 850-873, doi: [10.1108/AAAJ-07-2015-2139](https://doi.org/10.1108/AAAJ-07-2015-2139).

Blease, C., Worthen, A. and Torous, J. (2024), "Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey", *Psychiatry Research*, Vol. 333, p. 115724, doi: [10.1016/j.psychres.2024.115724](https://doi.org/10.1016/j.psychres.2024.115724).

Brown, J.E.H. and Halpern, J. (2021), "AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare", *SSM – Mental Health*, Vol. 1, p. 100017, doi: [10.1016/j.ssmmh.2021.100017](https://doi.org/10.1016/j.ssmmh.2021.100017).

Cho, H., Liu, S., Shi, T., Jain, D., Rizk, B., Huang, Y., Lu, Z., Wen, N., Gratch, J., Ferrera, E. and May, J. (2023), "Can language model moderators improve the health of online discourse?", [arXiv preprint] *arXiv:2311.10781*

Egan, S.J., Johnson, C., Wade, T.D., Carlbring, P., Raghav, S. and Shafran, R. (2024), "A pilot study of the perceptions and acceptability of guidance using artificial intelligence in internet cognitive behaviour therapy for perfectionism in young people", *Internet Interventions*, Vol. 35, p. 100711, doi: [10.1016/j.invent.2024.100711](https://doi.org/10.1016/j.invent.2024.100711).

Fiske, A., Henningsen, P. and Buyx, A. (2019), "Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy", *Journal of Medical Internet Research*, Vol. 21 No. 5, p. e13216.

Graham, S., Depp, C., Lee, E.E., Nebeker, C., Tu, X., Kim, H.C. and Jeste, D.V. (2019), "Artificial intelligence for mental health and mental illnesses: an overview", *Current Psychiatry Reports*, Vol. 21 No. 11, pp. 1-18.

Gual-Montolio, P., Jaén, I., Martínez-Borba, V., Castilla, D. and Suso-Ribera, C. (2022), "Using artificial intelligence to enhance ongoing psychological interventions for emotional problems in real-or close to real-time: a systematic review", *International Journal of Environmental Research and Public Health*, Vol. 19 No. 13, p. 7737.

Guembe, B., Azeta, A., Misra, S., Osamor, V.C., Fernandez-Sanz, L. and Pospelova, V. (2022), "The emerging threat of AI-driven cyber attacks: a review", *Applied Artificial Intelligence*, Vol. 36 No. 1, p. 2037254, doi: [10.1080/08839514.2022.2037254](https://doi.org/10.1080/08839514.2022.2037254).

Huang, J., Chen, X., Mishra, S., Zheng, H.S., Yu, A.W., Song, X. and Zhou, D. (2023), "Large language models cannot self-correct reasoning yet", [arXiv preprint] *arXiv:2310.01798*

- Keum, B.T. and Wang, L. (2021), "Supervision and psychotherapy process and outcome: a meta-analytic review", *Translational Issues in Psychological Science*, Vol. 7 No. 1, p. 89.
- Kühne, F., Maas, J., Wiesenthal, S. and Weck, F. (2019), "Empirical research in clinical supervision: a systematic review and suggestions for future studies", *BMC Psychology*, Vol. 7 No. 1, pp. 1-11.
- Luxton, D.D. (2014), "Artificial intelligence in psychological practice: current and future applications and implications", *Professional Psychology: Research and Practice*, Vol. 45 No. 5, p. 332.
- Malgaroli, M., Hull, T.D., Zech, J.M. and Moncada, S. (2023), "Natural language processing for mental health interventions: a systematic review and research framework", *Translational Psychiatry*, Vol. 13 No. 1, p. 309, doi: [10.1038/s41398-023-02592-2](https://doi.org/10.1038/s41398-023-02592-2).
- Nasr, M., Islam, M.M., Shehata, S., Karray, F. and Quintana, Y. (2021), "Smart healthcare in the age of AI: recent advances, challenges, and future prospects", *IEEE Access*, Vol. 9, pp. 145248-145270.
- Olawade, D.B., Wada, O.Z., Odetayo, A., David-Olawade, A.C., Asaolu, F., Eberhardt, J. and Oladokun, O.T. (2024), "Enhancing mental health with artificial intelligence: current trends and future prospects", *Journal of Medicine, Surgery, and Public Health*, Vol. 3, p. 100099.
- Pham, K.T., Nabizadeh, A. and Selek, S. (2022), "Artificial intelligence and chatbots in psychiatry", *Psychiatric Quarterly*, Vol. 93 No. 1, pp. 249-253.
- Shumailov, I., Shumailov, Z., Zhao, Y., Gal, Y., Papernot, N. and Anderson, R. (2023), "The curse of recursion: training on generated data makes models forget", [arXiv preprint] *arXiv:2305.17493*.
- Thieme, A., Hanratty, M., Lyons, M., Palacios, J., Marques, R.F., Morrison, C. and Doherty, G. (2023), "Designing human-centered AI for mental health: developing clinically relevant applications for online CBT treatment", *ACM Transactions on Computer-Human Interaction*, Vol. 30 No. 2, pp. 1-50, doi: [10.1145/3564752](https://doi.org/10.1145/3564752).
- Yang, D., Kommineni, A., Alshehri, M., Mohanty, N., Modi, V., Gratch, J. and Narayanan, S. (2023), "Context unlocks emotions: text-based emotion classification dataset auditing with large language models", *11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, *IEEE*, pp. 1-8.
- Zhou, S., Zhao, J. and Zhang, L. (2022), "Application of artificial intelligence on psychological interventions and diagnosis: an overview", *Frontiers in Psychiatry*, Vol. 13, p. 811665.