frontiers | Frontiers in Human Dynamics

TYPE Opinion
PUBLISHED 30 May 2025
DOI 10.3389/fhumd.2025.1519872

OPEN ACCESS

EDITED BY
ZhiMin Xiao,
University of Essex, United Kingdom

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

*CORRESPONDENCE
Tita Alissa Bach
✉ tita.alissa.bach@dnv.com

†These authors have contributed equally to this work

RECEIVED 30 October 2024
ACCEPTED 14 May 2025
PUBLISHED 30 May 2025

CITATION
Bach TA and Männikkö N (2025) The importance of justified patient trust in unlocking AI's potential in mental healthcare. *Front. Hum. Dyn.* 7:1519872. doi: 10.3389/fhumd.2025.1519872

The importance of justified patient trust in unlocking AI's potential in mental healthcare

Tita Alissa Bach[1*†] and Niko Männikkö[2†]

[1]Group Research and Development, DNV, Høvik, Norway, [2]Centre for Research and Innovation, Oulu University of Applied Sciences, Oulu, Finland

KEYWORDS

trust, trustworthy AI, AI system, patient engagement, ethics

Introduction

"Mental health is a basic human right." WHO mental health (World Health Organization, 2022).

Mental health is "a state of mental wellbeing that enables people to cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community" (World Health Organization, 2022). The WHO reports that in 2019 an estimated 970 million people worldwide were affected by mental disorders, with anxiety and depression being the most prevalent. For example, 1:5 U.S. adults and 1:6 individuals in Europe live with a mental illness (WHO Team, 2021; National Institute of Mental Health, 2022). The World Economic Forum projects that these conditions will contribute to a cumulative global economic loss of $16.3 trillion between 2011 and 2030 (World Economic Forum Harvard School of Public Health, 2011). A recent study indicates rising suicide rates among individuals aged 10–24 across the UK, the USA, much of Central Latin America, and Australasia (Bertuccio et al., 2024).

As mental health challenges continue to increase in number and complexity, the shortage of mental healthcare providers has become more acute, creating gaps in care (Lee et al., 2021). Artificial Intelligence (AI)-enabled systems[1] or, AI systems, have the potential to revolutionize mental healthcare by addressing these gaps, offering solutions that range from digital diagnostics to therapeutic tools (DNV, 2024). AI systems have been used to help mental healthcare by directly interacting with patients through self-management mobile health apps to aid in the treatment of depression, anxiety, post-traumatic stress disorders, sleep disorders, and suicidal behaviors (Müller et al., 2023; Shan et al., 2022). They also assist in diagnosing behaviors or responses associated with mental health conditions, developing risk profiles, and deploying context-specific interventions (Milne-Ives et al., 2022).

However, the success of these AI-driven innovations hinges on one crucial factor: patient trust. Without trust, patients may hesitate to engage with AI systems, limiting the technology's impact. Real-world cases have already highlighted the risks of diminished trust. For instance, the National Eating Disorders Association (NEDA) recently removed

1 Any system that contains or relies on one or more AI components. AI components are distinct units of software that perform specific functions or tasks within an AI-enabled system. They consist of a set of AI models, data, and algorithms, which, through implementation, create the AI component (Bach et al., 2024c).

AI chatbot, Tessa, from a support hotline after concerns arose that it was providing harmful advice, potentially exacerbating the conditions of vulnerable users who were patients with eating disorders (Atherton, 2023). Similarly, Sonde Health's voice analysis AI, which uses vocal biomarkers to assess depression, has been criticized for overlooking the diverse speech patterns of non-typical users, such as those with disabilities or regional and non-native speech differences (Ma et al., 2023). In addition, patient concerns about data privacy and potential biases in AI systems, how patient data is used, and the potential for AI systems to perpetuate existing inequalities have been reported as key trust barriers (Lee et al., 2021). These examples highlight the fragility of trust in AI systems, particularly in the sensitive domain of mental health, where patient vulnerability is already high at baseline (Minerva and Giubilini, 2023).

Trust is delineated as the "willingness to render oneself vulnerable" to a capability, founded on an evaluation of congruence in intentions or values (Bach et al., 2024b). Trust relationships can be established among individuals and between individuals and technology (Glikson and Woolley, 2020). Trust is often described as a connection between a trustor and a trustee, with the hopeful anticipation that the trustee will meet the trustor's expectations (Kelton et al., 2008). Trust relationships usually do not have legally binding obligations and are therefore susceptible to deceit. As a result, various factors contribute to and affect the dynamic of trust relationships.

Here, we focus on the trust placed in AI systems by mental health patients who are also the direct users, highlighting the most sensitive and direct relationship between AI systems and those whose mental healthcare is impacted by them. In mental healthcare, AI systems can be used by mental health professionals (Sebri et al., 2020; Montag et al., 2024), patients, and patients' families or caregivers (Zidaru et al., 2021). However, when patients are not the direct users of these systems, their trust in them is likely to be indirect and mediated by their trust in the healthcare professionals or family members who utilize AI systems on their behalf. Patients as users have different trust needs and significantly higher risks in using these systems than do non-patient users. Trust in this context is thus delicate, as patients' emotional and cognitive states may make them particularly vulnerable to the risks of over-reliance on AI-enabled systems. Patients are usually the primary stakeholders of many AI-enabled mental health applications, irrespective of who the users may be (Müller et al., 2023). The user experience for patients is deeply personal, and their trust in these systems directly influences their engagement and, ultimately, the outcomes of their care. This is because patient commitment to actively engage in the care plan is the single most critical determinant of positive outcomes (Milne-Ives et al., 2022). Therefore, building patient trust is not only beneficial for empowering patients and giving them a sense of control over their care, it is absolutely vital for ensuring successful and meaningful care outcomes (Milne-Ives et al., 2022).

In some cases, giving patients access to use AI systems without any support from mental healthcare professionals or caregivers requires careful consideration (Tavory, 2024). Such cases arise when patients are deemed clinically incapable of making their own decisions. This complicates the trust equation, as patients may no longer be seen as users of an AI system even if they interact directly with it. As a result, their caregivers may be viewed as the users.

Fostering justified trust

While patients' lack of trust may slow the adoption of AI systems in mental healthcare, a more significant concern arises when patients "trust incorrectly" (Taddeo, 2017). Initial hesitation or skepticism is a natural and expected reaction when humans encounter new or unfamiliar technology, making it easier to anticipate and address. However, the risks associated with overtrust or blind trust in AI systems are less frequently discussed (Aroyo et al., 2021), which could lead to serious consequences. For example, the Dutch childcare benefits scandal, where thousands of low-income families were wrongly accused of fraud due to a biased algorithm, led to victims committing suicide, suffering severe mental health issues, and the removal of their children into foster care (Amnesty International, 2021). A similar case happened with the Swedish Social Insurance Agency's algorithm (Amnesty International, 2024).

Trusting correctly means that the trust is justified and based on evidence, knowledge, experiences and/or skills (Taddeo, 2017; Glomsrud and Bach, 2023; Jacovi et al., 2021). In mental healthcare, this means that while patients may place trust in an AI system, they should still engage in critical thinking while interpreting the output of the AI system. Such critical thinking can be encouraged by providing patients with skills to understand AI systems' capabilities and limitations (Lee et al., 2021), so that patients can recognize deviations of AI systems' operations and output.

Justified trust requires transparency, reliability, and appropriate human oversight rather than blind reliance on AI outputs. For example, Sonde Health's voice analysis AI claims to offer "objective" depression detection by analyzing vocal biomarkers (Ma et al., 2023). However, if users assume its outputs are definitive diagnoses rather than probabilistic assessments, this could lead to misplaced trust. To foster justified trust, these systems must clearly communicate their limitations, and patients should retain control over their data, with options to review, modify, or delete AI-collected information (Lee et al., 2021). In care treatments where the relationships between patient-professionals are fundamental to positive outcomes, such as in psychotherapy (Holohan and Fiske, 2021; Danieli et al., 2021), it is important to also provide patients with access to mental health professionals, alongside patients' use of AI systems (Danieli et al., 2021).

However, there is still too little research investigating the effectiveness of various AI systems in mental healthcare to build evidence (Milne-Ives et al., 2022). Forming justified trust thus needs to depend on users' and domain experts' experiences, knowledge and/or skills with the hope that over time they can build evidence on the positive and negative effects of an AI system on patients' mental health (Taddeo, 2017; Glomsrud and Bach, 2023; Jacovi et al., 2021). Integrating patient AI literacy can be achieved through interactive

educational modules within applications, offering insights into system capabilities, limitations, and evaluation best practices (Lee et al., 2021; Milne-Ives et al., 2022). A structured framework with updates, tailored learning, and feedback can sustain engagement and foster justified trust in AI (Milne-Ives et al., 2022).

Maintaining justified trust over time

User trust in AI systems is dynamic and can change over time (Bach et al., 2024b; Cabiddu et al., 2022). A review by Cabiddu et al. (2022) highlights that initial trust is shaped by users' propensity to trust, the presence of human-like features, and the perceived usefulness of the system (Cabiddu et al., 2022). Human-like traits enhance emotional connections, making AI interactions more familiar and trustworthy. Over time, trust is further influenced by social factors, familiarity, and system reliability (Cabiddu et al., 2022). Users assess whether AI performance aligns with initial expectations, fostering justified trust based on experience and knowledge (Glomsrud and Bach, 2023).

As users become more familiar with AI systems, especially if they have strong social support to encourage continued use and develop a positive perception of its usefulness through consistent, reliable, and predictable AI output, sustained user trust is established. In such a scenario, established user trust can still transition into user distrust or mistrust, particularly when AI systems make errors that directly impact users, or when overtrust occurs, such as when users under time pressure and/or with low cognitive capacity act upon AI output without any evaluation or judgment (Bach et al., 2024c).

To maintain justified trust, it is crucial to continually promote critical thinking so that users may base their evaluation rooted in collected evidence, if any, as well as knowledge, experiences, and/or skills. Patient education on AI's capabilities and limitations and the incorporation of patients' feedback to improve the systems are extremely valuable for maintaining justified trust. Incorporating feedback can be done by, for example, allowing users to rate AI systems' responses and flag inaccuracies (Bach et al., 2024a), which then can be used to improve the AI systems' ability to retrieve and present more relevant information (Gao et al., 2024; Shankar et al., 2024).

The downside of maintaining justified trust in this manner is that it requires a high cognitive load, and it depends on the patients' ability to think critically each time. This can become an issue for mental health patient as patients may use AI systems in their most vulnerable conditions, when their cognitive capacity is likely limited.

It is only ethical and responsible to develop, deploy, and continuously improve AI systems together with patients (Bach et al., 2024c), especially to understand what influences patients' cognitive capacity and critical thinking when using AI systems. It is crucial to match specific user populations' characteristics and needs to the design of AI systems, specifically AI interface and features where human-AI interaction happens (Bach et al., 2024b,c). For example, this can be done by identifying users' needs to determine which key aspects of AI output are to be displayed, or not, in the interface.

An AI system for patients with sensory sensitivity should use fit-for-purpose visuals and audio, avoiding bright colors, loud noises, and overstimulating displays. AI systems for PTSD or trauma can gradually introduce challenging topics as trust builds rather than overwhelming patients. Customizable trigger detection allows patients to specify distressing words, topics, or stimuli, enabling AI systems to adjust accordingly. All these examples show the importance of embedding an AI feature to personalize user interface based on users' preferences, as well as giving control to users to decide what, how much, how, and when preferred information is to be presented to them (Bach et al., 2024c). Such personalization can help patients to evaluate AI output without requiring additional workload and within their cognitive capacity at the time of use, maintaining their justified trust. Developers can use adaptive learning models (Liu et al., 2024) that adjust responses based on user interactions and multimodal AI systems that combine voice, text, and biometric inputs for tailored recommendations (Islam and Bae, 2025). For example, AI-driven therapy platform Woebot adapts to users' mood patterns (Darcy et al., 2022), ensuring more contextually relevant support.

Conclusions

Given that mental healthcare already presents unique ethical and legal challenges, the integration of AI systems demands scrutiny and fit-for-purpose regulation (Tavory, 2024). Regulators play a crucial role in ensuring that AI development and deployment adhere to responsible and ethical principles (Tavory, 2024). For instance, they are responsible for verifying that the claimed benefits of using AI systems, particularly those made by for-profit vendors, are true.

Since the use of AI systems in mental healthcare is still emerging, creating structured platforms for stakeholders to exchange insights is essential for identifying both obstacles and best practices (Hamdoun et al., 2023). Future efforts should focus on evaluating real-world effectiveness, understanding long-term impacts on patient outcomes, and mitigating biases in AI-driven decision-making.

In conclusion, ensuring that AI systems provide personalized, clinically effective care while maintaining justified user trust is fundamental. Continued interdisciplinary collaboration between researchers, clinicians, and policymakers is key to maximizing AI's benefits while safeguarding patient wellbeing.

Author contributions

TB: Writing – original draft, Writing – review & editing. NM: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors would like to thank N. Hardwick for providing valuable proofreading support.

Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could potentially create a conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

Amnesty International (2021). *Xenophobic machines: Dutch child benefit scandal*. Available online at: https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/ (accessed April 4, 2025).

Amnesty International (2024). *Sweden: authorities must discontinue discriminatory AI systems used by welfare agency*. Available online at: https://www.amnesty.org/en/latest/news/2024/11/sweden-authorities-must-discontinue-discriminatory-ai-systems-used-by-welfare-agency/ (accessed April 4, 2025).

Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., et al. (2021). Overtrusting robots: setting a research agenda to mitigate overtrust in automation. *Paladyn* 12, 423–436. doi: 10.1515/pjbr-2021-0029

Atherton, D. (2023). *AI Incident Database, Incident 545: Chatbot Tessa Gives Unauthorized Diet Advice to Users Seeking Help for Eating Disorders*. AI Incident Database.

Bach, T. A., Babic, A., Park, N., Sporsem, T., Ulfsnes, R., Skeie, T., et al. (2024a). Using LLM-generated draft replies to support human experts in responding to stakeholder inquiries in maritime industry: a real-world case study of industrial AI. a*rXiv* [Preprint]. arXiv:2412.12732. doi: 10.48550/arXiv.2412.12732

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., and Sousa, S. A. (2024b). Systematic literature review of user trust in AI-enabled systems: an HCI perspective. *Int. J. Hum. Comput. Interact*. 40, 1251–1266. doi: 10.1080/10447318.2022.2138826

Bach, T. A., Kristiansen, J. K., Babic, A., and Jacovi, A. (2024c). Unpacking human-AI interaction in safety-critical industries: a systematic literature review. *IEEE Access* 12, 106385–106414. doi: 10.1109/ACCESS.2024.3437190

Bertuccio, P., Amerio, A., Grande, E., La Vecchia, C., Costanza, A., Aguglia, A., et al. (2024). Global trends in youth suicide from 1990 to 2020: an analysis of data from the WHO mortality database. *EClinicalMed.* 70:102506. doi: 10.1016/j.eclinm.2024.102506

Cabiddu, F., Moi, L., Patriotta, G., and Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *Eur. Manag. J.* 40, 685–706. doi: 10.1016/j.emj.2022.06.001

Danieli, M., Ciulli, T., Mousavi, S. M., and Riccardi, G. (2021). A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. *JMIR Form Res.* 5:30053. doi: 10.2196/preprints.30053

Darcy, A., Beaudette, A., Chiauzzi, E., Daniels, J., Goodwin, K., Mariano, T. Y., et al. (2022). Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. *Expert Rev. Med. Devices* 19:2075726. doi: 10.1080/17434440.2022.2075726

DNV (2024). *DNV-RP-0671: Assurance of AI-Enabled Systems. Det Norske Veritas*. Available online at: https://www.dnv.com/digital-trust/recommended-practices/assurance-of-ai-enabled-systems-dnv-rp-0671/

Gao, G., Taymanov, A., Salinas, E., Mineiro, P., and Misra, D. (2024). Aligning LLM agents by learning latent preference from user edits. *arXiv* [Preprint]. arXiv:2404.15269. doi: 10.48550/arXiv.2404.15269

Glikson, E., and Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Acad. Manag. Ann.* 14, 627–660. doi: 10.5465/annals.2018.0057

Glomsrud, J. A., and Bach, T. A. (2023). The Ecosystem of Trust (EoT): enabling effective deployment of autonomous systems through collaborative and trusted ecosystems. *arXiv* [Preprint]. arXiv:2312.00629. doi: 10.48550/arXiv.2312.00629

Hamdoun, S., Monteleone, R., Bookman, T., and Michael, K. (2023). AI-based and digital mental health apps: balancing need and risk. *IEEE Technol. Soc. Mag.* 42, 25–36. doi: 10.1109/MTS.2023.3241309

Holohan, M., and Fiske, A. (2021). "Like i'm talking to a real person": exploring the meaning of transference for the use and design of AI-based applications in psychotherapy. *Front. Psychol.* 12:720476. doi: 10.3389/fpsyg.2021.720476

Islam, R., and Bae, S. W. (2025). Revolutionizing mental health support: an innovative affective mobile framework for dynamic, proactive, and context-adaptive conversational agents. a*rXiv* [Preprint]. arXiv:2406.15942. doi: 10.48550/arXiv.2406.15942

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y (2021). "Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI," *in FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery), 624–635. doi: 10.1145/3442188.3445923

Kelton, K., Fleischmann, K. R., and Wallace, W. A. (2008). Trust in digital information. *J. Am. Soc. Inform. Sci. Technol.* 59, 363–374. doi: 10.1002/asi.20722

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., et al. (2021). Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 6, 856–864. doi: 10.1016/j.bpsc.2021.02.001

Liu, Y., Tan, H., Cao, G., and Xu, Y. (2024). Enhancing user engagement through adaptive UI/UX design: a study on personalized mobile app interfaces. *World J. Innov. Mod. Technol.* 7, 1–21. doi: 10.53469/wjimt.2024.07(05).01

Ma, A., Patitsas, E., and Sterne, J. (2023). "You sound depressed a case study on sonde health's diagnostic use of voice analysis AI," in *ACM International Conference Proceeding Series* (New York, NY: Association for Computing Machinery), 639–650. doi: 10.1145/3593013.3594032

Milne-Ives, M., Selby, E., Inkster, B., Lam, C., and Meinert, E. (2022). Artificial intelligence and machine learning in mobile apps for mental health: a scoping review. *PLOS Digit. Health* 1:e0000079. doi: 10.1371/journal.pdig.0000079

Minerva, F., and Giubilini, A. (2023). Is AI the future of mental healthcare? *Topoi* 42, 809–817. doi: 10.1007/s11245-023-09932-3

Montag, C., Ali, R., Al-Thani, D., and Hall, B. J. (2024). On artificial intelligence and global mental health. *Asian J. Psychiatr.* 91:103855. doi: 10.1016/j.ajp.2023.103855

Müller, R., Primc, N., and Kuhn, E. (2023). "You have to put a lot of trust in me": autonomy, trust, and trustworthiness in the context of mobile apps for mental health. *Med. Health Care Philos.* 26, 313–324. doi: 10.1007/s11019-023-10146-y

National Institute of Mental Health (2022). *Mental Illness*. Available online at: https://www.nimh.nih.gov/health/statistics/mental-illness (accessed April 4, 2025).

Sebri, V., Pizzoli, S. F. M., Savioni, L., and Triberti, S. (2020). Artificial intelligence in mental health: professionals' attitudes towards AI as a psychotherapist. *Ann. Rev. CyberTher. Telemed.* 18, 229–233. Available online at: https://www.arctt.info/volume-18-summer-2020

Shan, Y., Ji, M., Xie, W., Lam, K. Y., and Chow, C. Y. (2022). Public trust in artificial intelligence applications in mental health care: topic modeling analysis. *JMIR Hum. Factors* 9:e38799. doi: 10.2196/38799

Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A., and Arawjo, I. (2024). "Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (New York, NY: ACM), 1–14. doi: 10.1145/3654777.3676450

Taddeo, M. (2017). Trusting digital technologies correctly. *Minds Mach.* 27, 565–568. doi: 10.1007/s11023-017-9450-5

Tavory, T. (2024). Regulating AI in mental health: ethics of care perspective. *JMIR Ment. Health* 11:e58493. doi: 10.2196/58493

WHO Team (2021). *Mental Health Atlas 2020*. WHO Publication, 1–136. Available online at: https://www.who.int/publications/i/item/9789240036703 (accessed April 3, 2025).

World Economic Forum and Harvard School of Public Health (2011). *Methodological Appendix: The Global Economic Burden of Non-Communicable Diseases*. World Economic Forum, 1–20. Available online at: http://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_MethodologicalAppendix_2011.pdf (accessed April 3, 2025).

World Health Organization (2022). *Mental Health Key Facts*. Available online at: https://www.who.int/health-topics/mental-health#tab=tab_1 (accessed April 4, 2025).

Zidaru, T., Morrow, E. M., and Stockley, R. (2021). Ensuring patient and public involvement in the transition to AI-assisted mental health care: a systematic scoping review and agenda for design justice. *Health Expect.* 24, 1072–1124. doi: 10.1111/hex.13299