American College of Neuropsychopharmacology

www.nature.com/dpn

PERSPECTIVE OPEN

Check for updates

Advancing neurotech justice in youth digital mental health: insights from an interdisciplinary and cross-generational workshop

Craig W. McFarland[1], Donnella S. Comeau[2], Sepideh Abdi[2], Mahsa Alborzi Avanaki[2], Leo Anthony Celi[3], Neurotech Justice Workshop Participants*, Francis X. Shen[4]✉ and Benjamin C. Silverman[5]

Researchers and clinicians are increasingly looking to leverage artificial intelligence (AI) and digital tools to improve psychiatric care. Of particular promise is addressing the youth mental health crisis. Yet, the introduction of AI-enabled digital technologies for psychiatric treatment of young adults raises a host of ethical, legal, and societal issues (ELSI). To provide guidance in addressing these issues, we convened a two-day meeting at the Radcliffe Institute for Advanced Study at Harvard University: *Advancing Neurotech Justice in Mental Health: Insights from an Interdisciplinary and Cross-Generational Workshop*. The meeting brought together a diverse cohort of 17 experts and 5 students from various fields and different countries. In partnership with the MIT Critical Data team, the workshop engaged participants in an interactive Prompt-a-Thon to explore first-hand the potential benefits, biases, and harms related to the use of Large Language Model chatbots in mental health care. This Perspective reports on five principles of digital psychiatry deployment that the workshop participants determined to be the most essential: ensuring accuracy, remaining human-centric, promoting just access, protecting privacy, and providing transparency. We place these five principles within a "Neurotech Justice" framework and discuss how guardrails can be built to promote neurotech justice in digital psychiatry.

NPP – Digital Psychiatry and Neuroscience; https://doi.org/10.1038/s44277-025-00052-x

LAY SUMMARY

This article presents findings from an interdisciplinary workshop focused on Neurotech Justice in youth digital mental health, addressing the ethical issues raised by the increasing use of AI-enabled tools to provide mental health services for young adults. Workshop participants, a mix of youth and experts, co-created a framework of five core principles to ensure equitable and ethical deployment: ensuring accuracy, remaining human-centric, promoting just access, protecting privacy, and providing transparency.

There is emerging evidence that young adults in the United States and globally are in the midst of a mental health crisis [1, 2]. This challenge is felt acutely in secondary schools and on college campuses [3]. The rise in mental health needs has created an unprecedented strain on healthcare systems [2, 4]. Modalities that heavily depend on face-to-face consultations are failing to meet the growing need for accessible, affordable, and scalable mental health services. However, new neurotechnology such as digital phenotyping tools offers a potential solution [5]. Technology-based mental health support and treatment platforms are increasingly being developed and deployed to facilitate improved mental health diagnosis and treatment [6, 7], including for college-aged students [8]. Could increasing utilization of neurotechnology be leveraged to deliver improved mental health care for young adults? Maybe, but it is unlikely to be beneficial if the methods used to develop and implement these tools are not accessible, inclusive, and equitable [9]. The accelerating use of AI tools for mental health support has raised pressing concerns that deployment may be outpacing safeguards. For instance, in 2025, two youth suicides were linked to their interactions with large language models for mental health support [10, 11], illustrating the potentially severe consequences of deploying these tools without adequate safeguards.

In this Perspective, we present a framework consisting of five core principles—ensuring accuracy, remaining human-centric, promoting just access, protecting privacy, and providing transparency—to ensure that psychiatry's implementation of mental health AI tools is done equitably and ethically. We describe these

[1]University of Oxford, Oxford, UK. [2]Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School Teaching Hospital, Boston, MA, USA. [3]Laboratory for Computational Physiology, Massachusetts Institute of Technology; Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center; Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]University of Minnesota Law School & Graduate Program in Neuroscience; Massachusetts General Hospital Center for Law, Brain & Behavior; Harvard Medical School Center for Bioethics; Neurotech Justice Accelerator at Mass General Brigham, a Dana Foundation Center for Neuroscience & Society, Boston, MA, USA. [5]Human Research Affairs, Mass General Brigham; Institute for Technology in Psychiatry, McLean Hospital; Harvard Medical School, Boston, MA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉email: fxshen@umn.edu

Received: 10 July 2025 Revised: 18 October 2025 Accepted: 8 December 2025
Published online: 08 April 2026

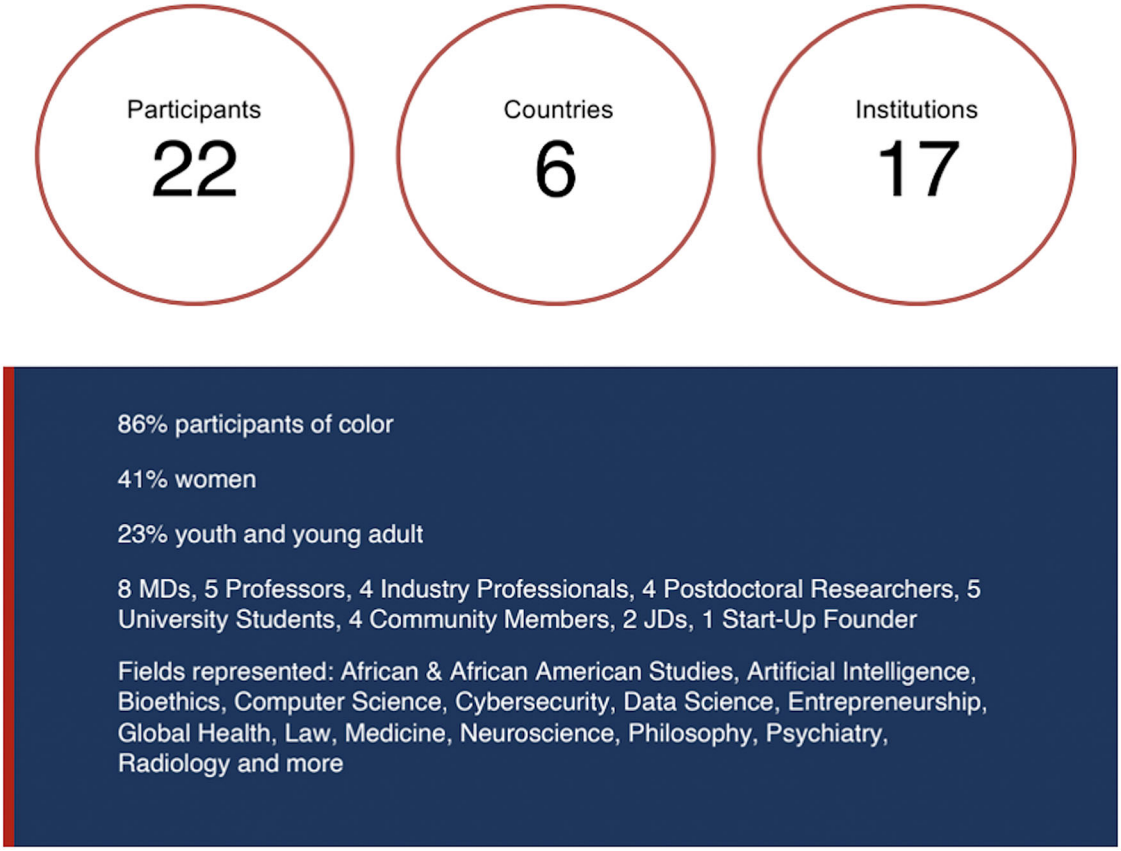

Fig. 1 Composition of Neurotech Justice Workshop Participants.

principles as falling within a broader "neurotech justice" framework, which encompasses considerations in the development and deployment of neurotechnologies.

Neurotech justice is a broad concept that can apply to a wide range of neurotechnologies. Neurotechnology covers many applications, including but not limited to deep brain stimulation, brain-computer interfaces, transcranial magnetic stimulation, electroencephalography, and functional magnetic resonance imaging. These technologies have been variously categorized in the literature along dimensions such as invasive vs. non-invasive; clinical vs. research; and read-out vs read-in [12]. In this paper, we focus on one rapidly expanding area in the neurotechnology landscape: digital psychiatry tools such as large language models (LLMs) and related AI-enabled platforms for youth mental health. We focus on LLMs as they are adopted more readily into everyday practice than brain scans or brain implants. LLMs are also being quickly and increasingly utilized in real-world mental health situations, and they can also be directly accessed by nonscientists, as compared to many other neurotechnologies, which require a clinician intermediary. We narrow our focus to digital tools aimed at youth because young adults are among the earliest and most frequent users of AI-enabled mental health tools and may be particularly vulnerable to LLMs' harms if safeguards are not in place. At the same time, schools and youth-centered institutions provide practical opportunities to operationalize these principles through education, awareness, and participatory design.

The framework was co-created by an interdisciplinary and multi-generational group, which convened for a two-day workshop at the Radcliffe Institute for Advanced Study at Harvard University: *Neurotech Justice: Engaging Young Adults to Improve Digital Mental Health Tools, Protect Privacy, and Ensure Computational Justice* (May 13-14, 2024). In total, the group convened for a total of 16 h, including five shared meals. The meeting brought together 17 experts from diverse fields and locations (Fig. 1) and five students from different countries to explore the ethical, legal, and societal (ELSI) challenges of integrating neurotechnology and artificial intelligence (AI) into mental health care for young adults. In addition, the workshop featured a presentation by Dr. Kevin Simon, MD, MPH, Boston's first Chief Behavioral Health Officer, on "Protecting Youth Mental Health," which provided clinical and public health context for the discussions.

In partnership with the Massachusetts Institute of Technology (MIT) Critical Data team, the workshop engaged participants in an interactive Prompt-a-Thon to explore first-hand the potential benefits, biases, and harms related to the use of LLM chatbots in mental health care. A prompt-a-thon is an interactive exercise in which participants develop different prompts to submit to the LLM and then carefully review the different responses the LLM generates [13]. Rather than simply telling participants that different prompts may lead to different and potentially harmful LLM outputs, a prompt-a-thon approach allows participants to directly engage with AI and observe these effects for themselves. In our workshop, the prompt-a-thon featured prompts related to youth mental health. After participants completed exercises facilitated by the MIT Critical Data team, participants engaged in group discussions about the effects of prompt engineering.

The workshop consisted of a cross-generational cohort composed of physicians and experts, mental health providers, with lived experience of the clinical side of mental health, founders and leaders in tech and AI leaders from various countries, local community members, and college students. This diversity across age, gender, race, ethnicity, geography, training, and expertise

was crucial to ensuring that our deliberations included both expert perspectives (e.g., the latest evidence of neurotechnology efficacy) and lived experiences from those directly affected by mental health issues.

We recognize that our framework emerges alongside numerous efforts to develop ethical guidance and regulation of AI. To date, there have been over 200 such guidelines promulgated [14]. These include guidelines from multidisciplinary experts and prominent organizations, international legislation (e.g., the European Union's ethics guidelines for Trustworthy AI) [15], and state-level regulations in the United States [16].

The novelty of our approach lies not in the final list of principles—which are similar to the core tenets of many of these 200 other guidelines—but rather in the collaborative process that generated them. Generating a framework specific to emerging digital psychiatry tools for youth mental health by test driving those tools with the youth in the room grounded our analysis in the practical realities of those most affected. That our bottom-up approach arrived at a similar outcome as the top-down expert-driven approaches is evidence that our method can be utilized for engaging communities in the process of AI guidance development.

ADDRESSING THE YOUTH MENTAL HEALTH CRISIS WITH NOVEL NEUROTECHNOLOGIES

Since the 2010s, there has been a significant rise in mental health issues among young people, including elementary-aged children [17]. The COVID-19 pandemic further negatively affected youth mental health [18]. Increased rates of loneliness, anxiety, depression, and other mental health disorders have created a surge in demand for mental health services [19]. But a shortage of mental health professionals, long wait times, and limited access to care have left many young people without the support they need [20]. Compared to other age groups, youth with any mental illness are least likely to receive mental health services, and less than half of youth in the US who need mental health care can access it [21].

Innovative digital mental and behavioral health tools have been proposed as part of a multiprong solution to address unmet mental and behavioral needs. Teletherapy platforms and mobile applications have emerged as a vital resource, providing accessible and flexible mental health support in a time where mobile devices are common among youth [22, 23].

Another potentially promising development to improve youth mental health is the use of AI-enabled chatbots and virtual therapists [24]. These tools promise the possibility of real-time support and guidance, offering psychotherapy, healthcare education, training for symptom management, early intervention, and screening for disorders [25]. Moreover, data analytics and machine learning are being harnessed to identify at-risk individuals and tailor interventions to their specific needs [26], improving outcomes and reducing the burden on mental health professionals. Several studies [27, 28] report that young adults are viewing AI-enabled chatbots as a favorable or even preferred mode of psychotherapy, underscoring the potential of AI-enabled chatbots to encourage those who might otherwise avoid seeking help to access mental health support. Together, these emerging neurotechnologies offer promising avenues for youth amidst a growing mental health crisis.

RISKS OF DEPLOYING AI TOOLS TO ADDRESS YOUTH MENTAL HEALTH

While digital mental health tools offer much promise for addressing the youth mental health crisis, it is not yet clear if they will live up to their potential. Participants in our workshop discussed a number of issues that have been identified in the literature, with the group arriving at a focus on three major concerns: accountability for harm, privacy, and equity.

An LLM providing youth mental health advice could offer inaccurate guidance that either discourages professional help, or exacerbates existing distress. Both concerns have been identified by multiple researchers in the field [29, 30]. In addition, LLMs providing youth mental health advice could lead to direct physical harm, such as recommendations for self-mutilation and other high-risk behaviors [31]. In some cases, chatbots have directly recommended homicidal and suicidal action, leading to tragic outcomes among young adults and teens [9, 10, 32]. At present, it is unclear who will be held legally responsible in these types of cases, nor is it clear what regulatory oversight is in place to avoid the replication of such harms [33, 34].

Privacy concerns also emerged as a predominant theme in our discussions. These tools often require the collection of vast amounts of sensitive personal data (e.g., behavioral patterns) to function effectively. Psychiatry, for instance, is beginning to adopt digital phenotyping and AI-driven tools that track participants' locations, online activity, phone and text message usage, and more [35]. Participants expressed deep concerns that such data could be exploited by third parties. For example, mental health data could be sold to advertisers, insurers, or employers, leading to targeted manipulation, discrimination in hiring or insurance coverage, or stigmatization based on mental health status [36]. The opaque and continuous nature of many data collection practices leaves users, often in a vulnerable state as they seek support for mental health challenges, unaware of what information is being gathered and how it might be used. The potential for data breaches or unauthorized access only heightens privacy concerns, further eroding trust in otherwise promising mental health technologies.

Equity concerns surrounding digital mental health tools include how these technologies may exacerbate existing health disparities rather than alleviate them. A concrete example, illuminated by the prompt-a-thon, was the possibility that an LLM would provide systematically different recommendations to individuals from one socioeconomic or racial group, as compared to another. Participants expressed uneasiness about how such responses could in effect codify bias and perpetuate existing social inequities if not carefully managed, especially as AI systems are trained on data primarily from affluent users. Algorithms trained on biased datasets may inadvertently reinforce discriminatory practices, thereby disproportionately affecting marginalized youth and providing higher rates of substandard care and inaccurate diagnoses for minority users [37].

Relatedly, participants worried about the commodification of mental health care, where a company's prioritization of profit may contribute to the digital divide and pose financial barriers to underserved communities. Racial and ethnic minorities, for example, are often excluded from accessing digital mental health resources due to systemic barriers like limited internet access, high data costs, and low digital literacy [38]. When financial incentives take precedence, high subscription fees, paywalls, and premium features can create significant cost barriers, preventing vulnerable populations from accessing potentially life-saving mental health support. This profit-driven model directly contradicts the potential of digital mental health tools to expand care in mental health services and instead undermines the promise of these technologies to democratize mental health care. Moreover, limited access among marginalized groups creates a harmful feedback loop of biased data collection and training.

CO-CREATED PRINCIPLES OF NEUROTECH JUSTICE

To help ensure that emerging digital mental health tools are developed and implemented ethically, equitably, and effectively,

Principle	Definition	Applying the Principle
Ensuring Accuracy	Accuracy ensures neurotechnologies are rigorously tested and validated to prevent misdiagnoses, ineffective treatments, and loss of user trust.	Conduct comprehensive testing across diverse populations and regularly validate tools to ensure consistent and reliable performance.
Remaining Human-Centric	Human-centric design prioritizes user well-being and accessibility over commercial interests, ensuring that tools are safe, effective, and responsive to diverse needs.	Engage users with lived mental health experience in co-designing features and interfaces to improve accessibility and usability.
Promoting Just Access	Just access means mental health technologies that are developed with fairness and inclusivity in mind, and that expand equitable access to neurotechnologies for all communities.	Include underrepresented communities in product development and offer subsidized access or tiered pricing to ensure affordability and access.
Protecting Privacy	Privacy is the protection of sensitive personal data, giving users control over their information and safeguarding against misuse or unauthorized access.	Implement strong data encryption, clear consent processes, and federated learning to minimize data breaches and enhance user control and autonomy.
Providing Transparency	Transparency requires clear communication about how neurotechnologies function, what data is collected, and any potential risks, fostering trust and accountability.	Publish algorithmic impact assessments, disclose AI decision-making processes, and inform users about how their data is used.

Fig. 2 Co-Created Principles of Neurotech Justice.

the workshop participants co-created a set of guiding principles of neurotech justice Fig. 2.

The workshop briefing materials provided relevant background on existing frameworks of neurorights [39]. During the workshop we followed a structured and participatory method to identify principles. Participants first engaged in open discussions to generate a comprehensive list of potential guiding principles and solutions, reflecting diverse perspectives on addressing the challenges in neurotechnology and mental health care. This extensive initial list, totaling over 20 potential key principles, included issues such as cultural sensitivity, harm reduction, and community integration. To refine and prioritize this list, participants provided input through facilitated group discussions and subsequently completed an anonymous survey to rank the proposed principles. Through this iterative approach, the workshop collaboratively distilled the initial ideas into five core principles for advancing neurotech justice in mental health care: (1) ensuring **accuracy**; (2) remaining **human-centric**; (3) promoting **just access** for youth; (4) protecting **privacy**; and (5) providing **transparency**.

Ensuring accuracy

Accuracy refers to ensuring that the design and deployment of neurotechnologies produce reliable and valid outcomes across populations. Inaccurate tools can lead to misdiagnoses, ineffective treatments, and a loss of user trust. Rushing software and products to market without sufficient validation may result in "technical debt," where costly harms and inefficiencies arise over time [40]. To improve accuracy, developers must conduct rigorous, representative testing across diverse populations before releasing mental health tools. For example, mental health assessment apps should be tested for diagnostic reliability across different racial, gender, and socioeconomic groups to ensure consistent performance regardless of user identity.

Remaining human-centric

Human-centric design ensures that the development of neurotechnologies prioritizes the needs, well-being, and experiences of users over the financial interests of developers or commercial stakeholders. This approach emphasizes the design of technologies that are safe, effective, and accessible to all users. A human-centric mental health app could involve participatory design practices, where individuals with lived mental health experiences are directly involved in shaping features, user interfaces, and accessibility options. For instance, integrating customizable settings to accommodate users with disabilities ensures that technology adapts to users rather than forcing users to adapt to the technology.

Promoting just access

Promoting just access involves integrating metrics around fairness, equity, and inclusivity in neurotechnology development, addressing the systemic barriers and historical harms that have marginalized vulnerable communities in healthcare and technology. This principle seeks to ensure that innovations do not replicate or worsen existing inequities. In practice, this may involve conducting outreach to low-income and underserved populations for participation in pilot programs or ensuring language accessibility in digital health tools. Developers could also implement tiered pricing models or subsidized access to ensure affordability.

Protecting privacy

Privacy in this context refers to the protection of sensitive personal data collected by digital psychiatry tools. Given that these tools may typically collect highly intimate data, privacy safeguards are critical to maintaining user trust and preventing misuse. To uphold privacy, mental health platforms must implement strong data encryption and clear consent protocols. Where possible, users

should have control over their data [41]. For example, employing federated learning allows user data to remain on local devices rather than being stored in centralized servers, reducing the risk of breaches. Additionally, providing transparent data-sharing options enables users to control who accesses their sensitive information.

Providing transparency

Transparency requires clear and open communication about how digital psychiatry tools function, what data they collect, and the potential risks and limitations of their use. LLMs currently face significant challenges in explaining how their outputs are produced. They are often described as "black boxes" because of their opaque inner workings and lack of interpretability [42]. These limitations mean that LLMs themselves cannot provide trustworthy explanations of their reasoning, and in fact may generate false "explanations" and use deception [43].

One possible solution, being explored in the Explainable AI field, is to utilize a separate "explainer" LLM to interpret the inner computational workings of a "target" LLM [44]. Effective transparency must also go beyond mere disclosure of technical details, ensuring that such information (e.g., algorithmic code) is translated in ways that are understandable and meaningful to users and empowers them to make informed decisions. These considerations are complemented by broader calls for transparency and accountability frameworks to prevent unmonitored or unsafe AI experimentation in healthcare [45]. Potential solutions include model cards and data cards to detail how AI-driven recommendations are generated [46], as well as structured reporting guidelines for research involving LLMs [47]. Additionally, developers should disclose when users are interacting with AI systems versus human professionals and clarify how personal data is used to shape responses.

ADVANCING NEUROTECH JUSTICE

Pathways to implementing these co-created principles are critical to addressing the concerns posed by novel neurotechnologies and ensuring the potential that these AI-driven mental health tools promise. The workshop concluded with a collaborative exercise in which participants identified actionable next steps for advancing neurotech justice in mental health care. This process generated concrete strategies focused on empowering youth, engaging policymakers, and fostering cross-sector collaboration to ensure the ethical development and deployment of neurotechnologies.

Immediate next steps included amplifying youth voices and actively involving young adults in shaping the future of neurotechnology. As early adopters of digital technologies and a population increasingly targeted by mental health interventions, young people offer valuable perspectives on how these tools can be designed to be more accessible, effective, and responsive to user needs. Involving youth through advisory boards, participatory design initiatives, and youth-led advocacy ensures that their experiences and insights inform the development of ethical and equitable mental health technologies.

Long-term efforts must address the historical systemic racial, gender, and economic inequities embedded in the current healthcare systems. This requires proactive measures to ensure that neurotechnologies meaningfully prioritize inclusivity and equitable access. Efforts should include investing in community-engaged research, ensuring diverse representation in product development and clinical trials, and designing technologies that are culturally responsive and accessible to marginalized populations. Outreach to colleges and universities is essential, but meaningful engagement must also extend to secondary and elementary schools. Such outreach should be led collaboratively by mental health professionals, academic societies, physicians across specialties—including pediatricians—educators, and academic leaders. Doing so will best support age-appropriate education about AI-enabled large language models and their mental health implications across all levels of community engagement. It is equally important to examine who is involved in developing, training, and deploying these models, as the perspectives and power dynamics of those shaping the systems influence both model behavior and downstream harms. Additionally, funding mental health initiatives in underserved communities, expanding access to the internet and other digital infrastructures, and supporting programs that build digital literacy are critical to bridging existing gaps.

To ensure these efforts have a sustained impact, systemic change must be driven by both industry leadership and supportive policy frameworks. Engaging corporate leaders, particularly C-suite executives at technology companies, will be critical to the integration of ethical principles into product development, corporate strategy, and regulatory frameworks. Collaborating with policymakers is key to enforcing legal protections and regulations that safeguard the advances of neurotechnology.

Further interdisciplinary collaborations, research, and ongoing advocacy efforts are essential to raise awareness about neurotech justice. This involves increasing literacy around neurotechnology among the public and educating policymakers, healthcare professionals, and technology developers about the ethical considerations and potential impacts of neurotechnology. Such efforts must also build a rigorous evidence base that directly addresses the methodological shortcomings that limited earlier research on digital harms, ensuring that studies at the intersection of AI and youth mental health generate reliable and actionable knowledge, especially for the public. Strengthening public understanding and engagement not only empowers individuals to make informed decisions but also holds developers accountable for responsible innovation. Taking these steps will enable digital psychiatry tools to fulfill their potential and contribute to the democratization and improvement of mental health care for all.

Citation diversity statement. The authors have attested that they made efforts to be mindful of diversity in selecting the citations used in this article.

REFERENCES

1. Benton TD, Boyd RC, Njoroge WF. Addressing the global crisis of child and adolescent mental health. JAMA Pediatrics. 2021;175:1108–10.
2. McGorry PD, Mei C, Chanen A, Hodges C, Alvarez-Jimenez M, Killackey E. Designing and scaling up integrated youth mental health care. World Psychiatry. 2022;21:61–76.
3. Oswalt SB, Lederer AM, Chestnut-Steich K, Day C, Halbritter A, Ortiz D. Trends in college students' mental health diagnoses and utilization of services, 2009-15. J Am Coll Health. 2020;68:41–51.
4. CDC, Improving Access to Children's Mental Health Care, https://www.cdc.gov/childrensmentalhealth/access.html.
5. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol. 2022;1:87–100.
6. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health. 2021;8:e25847.
7. Nisenson M, Lin V, Gansner M. Digital phenotyping in child and adolescent psychiatry: a perspective. Harv Rev Psychiatry. 2021;29:401–8.
8. Currey D, Hays R, Torous J. Digital phenotyping models of symptom improvement in college mental health: generalizability across two cohorts. J Technol Behav Sci. 2023;8:368–81.
9. Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: Delphi study. JMIR Mhealth Uhealth. 2021;9:e27343.
10. Reiley, L. Opinion | What My Daughter Told ChatGPT Before She Took Her Life. The New York Times. 2025. https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html, with suing from parents.

11. Hill, K. A Teen Was Suicidal. ChatGPT Was the Friend He Confided In. The New York Times. 2025. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html.
12. Collins B, Klein E. Invasive neurotechnology: a study of the concept of invasiveness in neuroethics. Neuroethics. 2023;16:11.
13. Kučević E, Brackel-Schmidt CV, Lewandowski T, Leible S, Memmert L, Böhmann T. The prompt-a-thon: Designing a format for value Co-creation with generative AI for research and practice. Proceedings of the 57th Annual Hawaii International Conference on System Sciences. Honolulu, HI: Department of IT Management Shidler College of Business University of Hawaii; 2024.
14. Corrêa NK, Galvão C, Santos JW, Del Pino C, Pinto EP, Barbosa C, et al. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns. 2023;4:100857.
15. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. (n.d.). Retrieved September 4, 2025, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
16. National Conference of State Legislatures, Artificial Intelligence 2025 Legislation (July 10, 2025), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2025-legislation.
17. Office of the Surgeon General (OSG). (2021). Protecting Youth Mental Health: The U.S. Surgeon General's Advisory. US Department of Health and Human Services. http://www.ncbi.nlm.nih.gov/books/NBK575984/.
18. Creswell C, Shum A, Pearcey S, Skripkauskaite S, Patalay P, Waite P. Young people's mental health during the COVID-19 pandemic. Lancet Child Adolesc Health. 2021;5:535–7. https://doi.org/10.1016/S2352-4642(21)00177-2.
19. American Psychological Association. (2022). Psychologists struggle to meet demand amid mental health crisis: 2022 COVID19 Practitioner Impact Survey. https://www.apa.org/pubs/reports/practitioner/2022-covid-psychologist-workload.
20. Sun C-F, Correll CU, Trestman RL, Lin Y, Xie H, Hankey MS, et al. Low availability, long wait times, and high geographic disparity of psychiatric outpatient care in the US. Gen Hosp Psychiatry. 2023;84:12–7. https://doi.org/10.1016/j.genhosppsych.2023.05.012.
21. National Institute of Mental Health. (2022). Mental Illness. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/mental-illness.
22. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health. 2022;9:e35159. https://doi.org/10.2196/35159.
23. Harrison V, Proudfoot J, Wee PP, Parker G, Pavlovic DH, Manicavasagar V. Mobile mental health: Review of the emerging field and proof of concept study. J Ment Health. 2011;20:509–24. https://doi.org/10.3109/09638237.2011.608746.
24. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth. 2023;11:e44838. https://doi.org/10.2196/44838.
25. Abd-Alrazaq AA, Alajlani M, Abdallah Alalwan A, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. Int J Med Inform. 2019;132:103978. https://eprints.whiterose.ac.uk/151992/.
26. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. J Med Surg Public Health. 2024;3:100099. https://doi.org/10.1016/j.glmedi.2024.100099.
27. Hoffman BD, Oppert ML, Owen M. Understanding young adults' attitudes towards using AI chatbots for psychotherapy: The role of self-stigma. Comput Hum Behav: Artificial Humans. 2024;2:100086. https://doi.org/10.1016/j.chbah.2024.100086.
28. Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": Experiences of generative AI chatbots for mental health. Npj Ment Health Res. 2024;3:1–9. https://doi.org/10.1038/s44184-024-00097-4.
29. Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Bell MJ. The opportunities and risks of large language models in mental health. JMIR Ment Health. 2024;11:e59479.
30. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11:e57400.
31. *ChatGPT Gave Instructions for Murder, Self-Mutilation, and Devil Worship—The Atlantic*. (n.d.). Retrieved September 4, 2025, from https://www.theatlantic.com/technology/archive/2025/07/chatgpt-ai-self-mutilation-satanism/683649/.
32. *Lawsuit claims Character.AI is responsible for teen's suicide*. (2024, October 23). NBC News. https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791.
33. Baidal, M, Derner E, Oliver N. "Guardians of trust: Risks and opportunities for llms in mental health." In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pp. 11-22. 2025.
34. Kahane K, Shumate JN, Torous J. Policy in flux: addressing the regulatory challenges of AI integration in US mental health services. Current Treatment Options in Psychiatry. 2025;12:24.
35. Shen FX, Baum ML, Martinez-Martin N, Miner AS, Abraham M, Brownstein CA, et al. Returning individual research results from digital phenotyping in psychiatry. Am J Bioeth: AJOB. 2024;24:69–90. https://doi.org/10.1080/15265161.2023.2180109.
36. Ienca M, Haselager P, Emanuel E. Brain leaks and consumer neurotechnology. Nat Biotechnol. 2018;36:805–10. https://doi.org/10.1038/nbt.4240.
37. Mihan A, Pandey A, Van Spall HGC. Artificial intelligence bias in the prediction and detection of cardiovascular disease. npj Cardiovasc Health. 2024;1:31 https://doi.org/10.1038/s44325-024-00031-9.
38. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep. 2021;23:61. https://doi.org/10.1007/s11920-021-01274-4.
39. Goering S, Klein E, Specker Sullivan L, Wexler A, Agüera Y Arcas B, Bi G, et al. Recommendations for responsible development and application of neurotechnologies. Neuroethics. 2021;14:365–86. https://doi.org/10.1007/s12152-021-09468-6.
40. Tom E, Aurum A, Vidgen R. An exploration of technical debt. J Syst Softw. 2013;86:1498–516. https://doi.org/10.1016/j.jss.2012.12.052.
41. Yuste R. Advocating for neurodata privacy and neurotechnology regulation. Nat Protoc. 2023;18:2869–75. https://doi.org/10.1038/s41596-023-00873-0.
42. Wadden JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. J Med Ethics. 2022;48:764–8.
43. Park PS, Goldstein S, O'Gara A, Chen M, Hendrycks D. AI deception: A survey of examples, risks, and potential solutions. Patterns. 2024;5:100988 https://doi.org/10.1016/j.patter.2024.100988.
44. Wu, X, Zhao H, Zhu Y, Shi Y, Yang F, Hu L, et al. "Usable XAI: 10 strategies towards exploiting explainability in the LLM era." *arXiv preprint arXiv:2403.08946* [Preprint]. 2024 https://arxiv.org/abs/2403.08946.
45. Barman KG, Wood N, Pawlowski P. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use. Ethics Inf Technol. 2024;26:47.
46. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. proceedings of the conference on fairness, Accountability and Transparency 2019; 220-9. https://doi.org/10.1145/3287560.3287596.
47. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31:60–9. https://doi.org/10.1038/s41591-024-03425-5.

AUTHOR CONTRIBUTIONS

DSC, BCS, and FXS co-created and co-organized the workshop from which this paper developed: *Neurotech Justice: Engaging Young Adults to Improve Digital Mental Health Tools, Protect Privacy, and Ensure Computational Justice*. LAC worked with DSC, BCS, and FXS to develop the workshop prompt-a-thon and co-led segments of the workshop. CWM and FXS led drafting of the manuscript. DSC, SA, MAA, and BCS reviewed and edited the manuscript. All co-authors, as well as the Workshop Participants, participated in the 2-day Neurotech Justice Radcliffe Institute Accelerator Workshop Program. The workshop discussion generated the insights that we present in this article.

FUNDING

FXS is supported by National Institute of Mental Health, National Institutes of Health grant 7R01MH134144-02, and by the Dana Foundation grant: Neurotech Justice Accelerator at MGB (NJAM), a Dana Center for Neuroscience & Society. LAC is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, the National Science Foundation through ITEST #2148451, and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00403047). The views expressed in this article are those of the authors and not necessarily those of the funders.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Francis X. Shen.

NEUROTECH JUSTICE WORKSHOP PARTICIPANTS

Sepideh Abdi[2], Julian Adong[6], Shaikha Alothman[7], Mahsa Alborzi Avanaki[2], Manal Brahimi[8], RuQuan Brown[9], Leo Anthony Celi[3], Cecile Chavane[10], Donnella S. Comeau[11], Jack Gallifant[12], Felix Garcia[13], Craig W. McFarland[1], Gabriel Làzaro-Muñoz[12], Eliane Motchoffo[14], Claire Joy Moss[15], Derek Ricketts[8], Francis X. Shen[4]✉, Benjamin C. Silverman[5], Paulos Solomon[8] and Takeshi Tohyama[16]

[6]Mbarara University of Science & Technology, Mbarara, Uganda. [7]Kuwait Healthy Cities, Kuwait City, Kuwait. [8]Bunker Hill Community College, Boston, MA, USA. [9]Harvard College, Cambridge, MA, USA. [10]Massachusetts Institute of Technology, Cambridge, USA. [11]Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School Teaching Hospital, Boston, USA. [12]Mass General Brigham, Boston, MA, USA. [13]Mainstay, Boston, MA, USA. [14]Apple, San Francisco, CA, USA. [15]Steppingstone, Boston, MA, USA. [16]Massachusetts Institute of Technology, Cambridge, MA, USA.