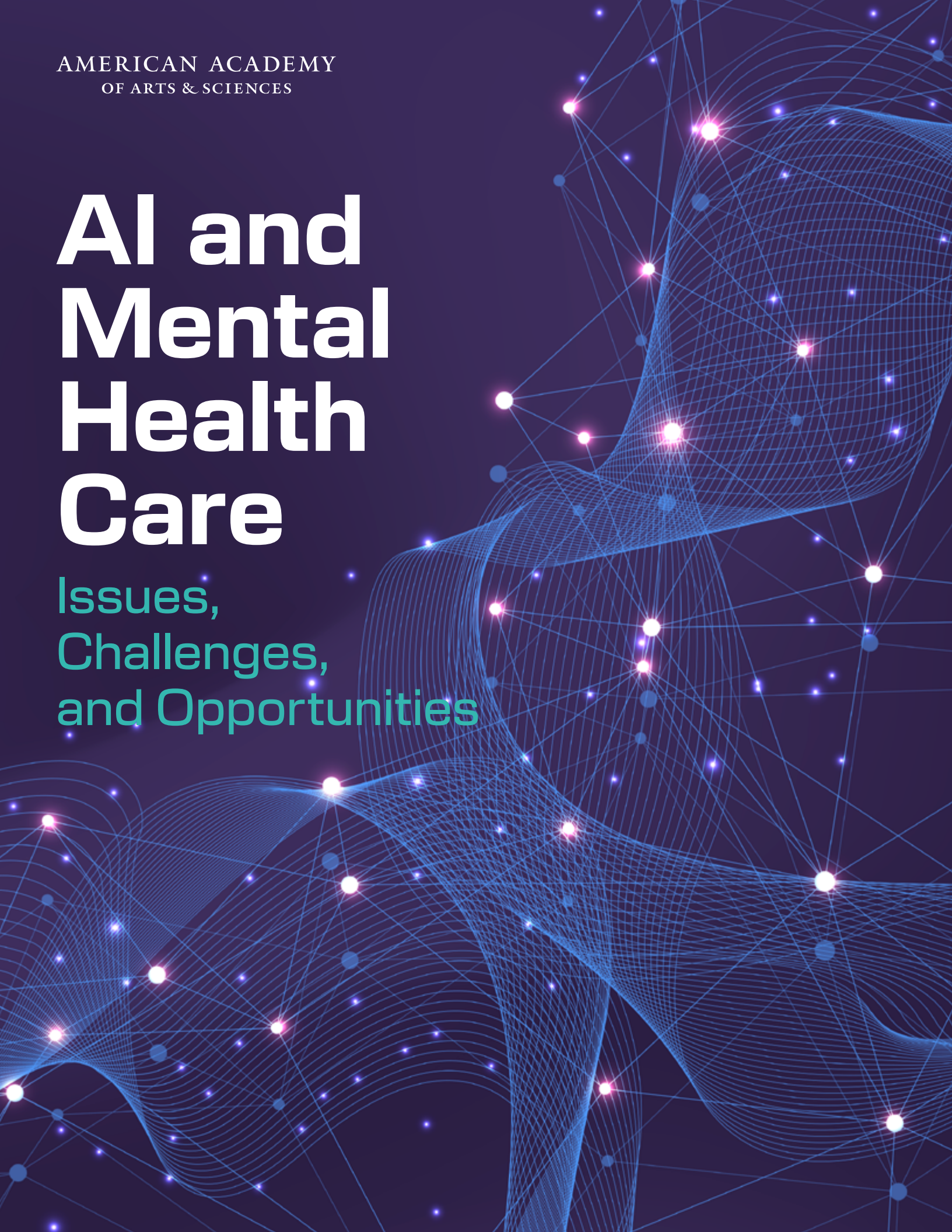
AMERICAN ACADEMY
OF ARTS & SCIENCES

AI and Mental Health Care

Issues, Challenges, and Opportunities



AI and Mental Health Care

Issues, Challenges, and Opportunities

AMERICAN ACADEMY OF ARTS & SCIENCES
Cambridge, Massachusetts

© 2025 by the American Academy of Arts & Sciences.
All rights reserved.

ISBN: 0-87724-171-6

This publication is available online at
www.amacad.org/project/artificial-intelligence-AI-mental-health-care.

Suggested citation:
American Academy of Arts and Sciences, *AI and Mental Health Care: Issues, Challenges, and Opportunities* (American Academy of Arts and Sciences, 2025).

Image credits:
iStock.com/AliseFox: cover; **Courtesy of Mass General Brigham:** photo of Daniel Barron; **Weinberg-Clark Photography:** photo of Marian Croak; **Courtesy of Mindstrong:** photo of Paul Dagum; **Courtesy of Woebot Health:** photo of Alison Darcy; **Jasmine Loew:** photo of Holly DuBois; **Paul Morigi:** photo of Richard Frank; **Elena Olivo:** photo of Sherry Glied; **Eleanor Greely:** photo of Hank Greely; **Dan DeLong:** photo of Eric Horvitz; **Courtesy of Dartmouth College:** photo of Nicholas Jacobson; **Andrew Kaufmann, George W. Bush Presidential Center:** photo of Kacie Kelly; **Torben Eskerod:** photo of Arthur Kleinman; **Courtesy of Jaron Lanier:** photo of Jaron Lanier; **Courtesy of AAAS:** photo of Alan Leshner; **Sandra Blaser:** photo of Robert Levenson; **Todd Balfour:** photo of Laurie L. Patton; **Courtesy of Peter Slavin:** photo of Peter Slavin; **Courtesy of Sherry Turkle:** photo of Sherry Turkle.

The views expressed in this publication are those held by the contributors and are not necessarily those of the Officers and Members of the American Academy of Arts and Sciences.

Please direct inquiries to:
American Academy of Arts and Sciences
136 Irving Street
Cambridge, Massachusetts 02138-1996
Telephone: (617) 576-5000
Email: aaas@amacad.org
Visit our website at www.amacad.org

Contents

List of Project Participants 5

A Letter from the President of the American Academy 9

Introduction 11

- Background 11
- Our Approach 12

Question 1: How might we measure the effectiveness of AI-driven mental health interventions? 14

- Background 14
- Responses 15

Question 2: How can we ensure and monitor the safety of AI in mental health care? 28

- Background 28
- Responses 29

Question 3: Must there always be a human in the loop? 36

- Background 36
- Responses 36

Question 4: What are the potential unintended consequences of AI-driven mental health care on the nature of interpersonal relationships? 43

- Background 43
- Responses 43

Question 5: How should these tools be deployed or limited in high-risk or vulnerable populations? 52

- Background 52
- Responses 53

Contents

Question 6: How should AI models be reimbursed or monetized? 56

Background 56

Responses 57

Question 7: Can AI help address disparities in access to mental health care and the shortage of mental health providers? 61

Background 61

Responses 62

Question 8: What are the economic and other impacts of AI adoption in mental health on health care providers? 72

Background 72

Responses 73

Question 9: What are the most significant scholarly questions that will need to be answered as AI's role in mental health care evolves? 77

Background 77

Responses 77

Afterword 83

Endnotes 85

List of Project Participants

DANIEL BARRON is Director of the Pain Intervention and Digital Research Program at Spaulding Rehabilitation Hospital and Brigham and Women's Hospital and is an Assistant Professor at Harvard Medical School. He is the author of *Reading Our Minds: The Rise of Big Data Psychiatry* (Columbia University Press, 2021) and has written for *The Wall Street Journal*, *TIME*, and *Scientific American*.

MARIAN CROAK is the Vice President of Human Centered AI and Foundational ML at Google, which she joined in 2014 after retiring from AT&T. She has received numerous awards, including the 2013 and 2014 Edison Patent Award, and was inducted into the Women in Technology International Hall of Fame in 2013. She has received over two hundred patents and is a member of the National Academy of Engineering. She was elected to the American Academy of Arts and Sciences in 2022.

PAUL DAGUM (cochair) is an entrepreneur, physician, and computer scientist with a track record of innovation in health care, cybersecurity, and supply chains across four successful venture-backed companies as founder, CSO, CTO, and CEO. He has published more than eighty peer-reviewed articles and has been awarded more than twenty patents in computer science and medicine.

ALISON DARCY is a clinical scientist and health tech innovator passionate about making mental health support radically accessible. She founded *Woebot*, an AI-powered mental health companion grounded in decades of psychological science and responsible design. *Woebot* was the first mental health chatbot to be examined in a randomized trial. Trained at University College, Dublin and Stanford School of Medicine, Alison has always been drawn to exploring unmet needs in care and how technology may be part of the solution to help address them. Her work has led to eighteen clinical trials and widespread recognition—with *Woebot* featured in outlets like CBS *60 Minutes*, *The New York Times*, and *The Wall Street Journal,* and with Alison named to the 2023 TIME100 AI list, which recognizes the 100 most influential people in AI. She continues to advocate for safe, engaging uses of AI in health care.

HOLLY DUBOIS is a board-certified psychiatrist serving as Chief Medical Officer for Connections Health Solutions, one of the nation's leading providers of behavioral health crisis services, meeting the needs of over thirty thousand individuals every year. She grew and led the care delivery system for Mindstrong Health Services, a groundbreaking venture designed to engage individuals living with serious mental illness via novel smartphone technology.

List of Project Participants

RICHARD G. FRANK, PhD, is the Margaret T. Morris Professor of Health Economics (emeritus) at Harvard Medical School. He is Senior Fellow in Economic Studies and Director of the Center on Health Policy at the Brookings Institution. He served as Assistant Secretary for Planning and Evaluation in the U.S. Department of Health and Human Services from 2014–2016.

SHERRY GLIED (cochair) is Professor and former Dean of the Robert F. Wagner Graduate School of Public Service at New York University. She is an economist, and her principal areas of research are in health policy reform and mental health care policy. She is the author of *Chronic Condition* (Harvard University Press, 1998), coauthor (with Richard Frank) of *Better But Not Well: Mental Health Policy in the United States since 1950* (Johns Hopkins University Press, 2006), and coeditor (with Peter C. Smith) of *The Oxford Handbook of Health Economics* (Oxford University Press, 2011). She was elected to the American Academy of Arts and Sciences in 2022.

HENRY T. (HANK) GREELY specializes in the ethical, legal, and social implications of new biomedical technologies, particularly those related to genetics, assisted reproduction, neuroscience, or stem cell research. He is a founder and past president of the International Neuroethics Society; former chair of the Ethical, Legal, and Social Issues Committee of the Earth BioGenome Project; and chair of California's Human Stem Cell Research Advisory Committee.

ERIC HORVITZ is Microsoft's Chief Scientific Officer, where he leads strategic initiatives at the intersection of science, technology, and society. His work emphasizes frontier advances in artificial intelligence, biosciences, and health care. He has made enduring contributions to AI theory and practice, from foundational models of cognition and bounded rationality to the development and deployment of AI systems in biomedicine, transportation, computing, and aerospace. He was elected to the American Academy of Arts and Sciences in 2011.

NICHOLAS JACOBSON is an Associate Professor in Biomedical Data Science, Psychiatry, and Computer Science at Dartmouth College, where he directs the AIM HIGH Laboratory. His research focuses on using technology, including smartphones and wearables, for the precision assessment and scalable treatment of anxiety and depression. He recently led the development and first clinical trial of Therabot, the first fully generative AI for psychotherapy.

KACIE KELLY has more than twenty years of experience leading innovation in mental health care and translating it into policy and practice. As Chief Innovation Officer at the Meadows Mental Health Policy Institute, she leads work to integrate scalable, data-driven innovation into health care, schools, justice, and community systems to detect mental health risks earlier and increase access to quality care for more youth and adults.

List of Project Participants

ARTHUR KLEINMAN is Professor of Medical Anthropology, Psychiatry and Global Health and Social Medicine at Harvard University. He is a physician and anthropologist as well as a leading figure in several fields, including medical anthropology, cultural psychiatry, global health, social medicine, and medical humanities. He is the author, coauthor, or coeditor of forty volumes, including *Patients and Healers in the Context of Culture* (University of California Press, 1980); *The Illness Narratives* (Basic Books, 1980); *Rethinking Psychiatry* (Free Press, 1991); *World Mental Health* (Oxford University Press, 1996); *What Really Matters* (Oxford University Press, 2007); *Deep China* (University of California Press, 2011); *Reimagining Global Health* (University of California Press, 2013); and *The Soul of Care* (Viking/Penguin, 2019), along with over four hundred articles. He was elected to the American Academy of Arts and Sciences in 1992.

JARON LANIER is a computer scientist, composer, artist, and author. He is a founder of the field of Virtual Reality (a term he coined), as well as adjacent disciplines including surgical simulation, and has received a lifetime achievement award from the Institute of Electrical and Electronics Engineers. His bestselling books on computer science and society have won numerous honors, including the German Peace Prize for Books. He also writes for *The New Yorker* and for movies and television. He is a specialist in performing on rare musical instruments of all cultures and periods. He currently serves as Prime Unifying Scientist at Microsoft.

ALAN I. LESHNER (cochair) is Chief Executive Officer, Emeritus, of the American Association for the Advancement of Science (AAAS) and former Executive Publisher of the journal *Science*. Before joining AAAS, he was Director of the National Institute on Drug Abuse at the National Institutes of Health. He also served as Deputy Director and Acting Director of the National Institute of Mental Health and in several roles at the National Science Foundation. He was elected to the American Academy of Arts and Sciences in 2005.

ROBERT LEVENSON works in the areas of human psychophysiology and affective neuroscience, both of which involve studying the interplay between psychological and physiological processes. Much of his work focuses on the nature of human emotion, specifically, its physiological manifestations, variations in emotion associated with age, gender, culture, and pathology, and the role emotion plays in interpersonal interactions. He was elected to the American Academy of Arts and Sciences in 2018.

LAURIE L. PATTON is a scholar of religion, a poet, and a translator who currently serves as President of the American Academy of Arts and Sciences. She previously served as the seventeenth President of Middlebury College—the first woman to do so in the school's 224-year history. She was the Durden Professor of Religion and the Dean of Arts and Sciences at Duke University, and at Emory University was the Charles Howard Candler Professor of Religion. She was elected to the American Academy of Arts and Sciences in 2018.

PETER L. SLAVIN, President and CEO of Cedars-Sinai Medical Center and Health System, brings deep experience in academic medicine and dedication to serving patients and community. Before joining Cedars-Sinai, he served as President of Massachusetts General Hospital, which has the largest hospital-based research program in the United States. Throughout his career as a physician leader, professor, and administrator, he has led innovation in clinical care, research funding, scientific impact, workforce development, and fundraising. He was elected to the American Academy of Arts and Sciences in 2021.

SHERRY TURKLE, Abby Rockefeller Mauzé Professor of the Social Studies of Science and Technology at MIT, explores the subjective side of people's relationships with technology, especially digital technology. Her research centers on analyzing electronic communication technologies and their impact on our emotions, creativity, and work. Profiled in *The New York Times*, *The New Yorker*, *Scientific American*, and *Wired* magazine, she has been a featured commentator on the social and psychological effects of technology for many media networks. A sociologist and clinical psychologist, her books—which include *Life on the Screen* (Simon and Schuster, 1997), *The Second Self* (MIT Press, 2005), *Simulation and Its Discontents* (MIT Press, 2009), *Alone Together* (Basic Books, 2011), and *Reclaiming Conversation* (Penguin, 2015)—have opened up new research spaces. She was elected to the American Academy of Arts and Sciences in 2014.

A Letter from the President of the American Academy

As a former college president in the rural area of Middlebury, Vermont, my experience of the precarious state of mental health for young people in our country is real. On the one hand, students often feel that they lack the skills and resources to make meaningful in-person connections. They frequently turn to AI-generated forms of connecting, which leads them further away from the actual in-person company of their peers. On the other hand, many students report that AI tools help them practice being in community and interacting with people in a way that lets them build an authentic self—a key task for young adults.

The mental health concerns of adults in the community are also real and include the daily pressures of childrearing, work, and eldercare. Adults also struggle in rural communities, where maintaining and staffing adequate mental health facilities are ongoing challenges. AI tools can provide a bridge when underfunded systems might falter.

From these communities, I have learned that AI-generated mental health tools cannot provide magic bullets for adaptation in a difficult world, nor can they substitute entirely for human connection. If they are solidly grounded in face-to-face realities, however, these tools can help people practice being resilient in an often inhumane world.

What I have also come to understand is that while mental health challenges may at times feel isolating, they cut across age, geography, and identity. Even with its unique New England flavor, Middlebury is no different than most places in America. Its challenges reflect a broader pattern in American cultural life that has resulted in an expanded and increasingly accessible public discourse around mental health.

The systems surrounding care, connection, and well-being, however, have not kept pace. This discrepancy invites a broader public responsibility: to bring forth our collective ingenuity and empathy and confront one of the defining challenges of our time. The AI and Mental Health Care project at the American Academy of Arts and Sciences responds to this calling with a commitment to inquiry and collaboration.

We find ourselves in a period of extraordinary technological advancement. Artificial intelligence and other digital tools are already redefining how we live, work, and connect with others. Some of these advances are poised to reshape how we seek care, particularly in the face of provider shortages and unequal access to care. Others raise pressing questions about equity, safety, and the limits of machine-driven empathy. What remains undetermined is not whether these tools will play a role in the future of mental health care; they already do. Rather, the question is whether they will do so in ethical, inclusive, and effective ways.

To interrogate these complexities is the work of scholarship and thus the work of the Academy. As we considered how best to engage in these efforts, we turned to one of our core values: *Fostering Deliberative Discourse.* Respectful discussion in the face of disagreement is a foundational principle of our democracy and of the Academy, one that shaped both our committee's composition and this report's format. Rather than seeking consensus or a

singular vision, we invited a wide range of experts to respond to a shared set of questions. The result is not an argument about the utilization of AI in mental health care but a scholarly conversation that reflects the breadth of disciplines, experiences, and perspectives gathered around the table. We invite you, as you read this report, to join us in an active discussion examining this important issue at the intersection of care and technology.

The choice of the dialogical genre for this publication is intentional. Because no clear consensus or agreed paradigm has emerged concerning AI and the right approaches to AI and mental health, we felt it was more straightforward to use a genre that is often the beginning of scientific inquiry: question and debate, or the dialogue. Plato used dialogue in his philosophical teaching. Philosopher David Hume used dialogue in what he thought of as a scientific investigation into human experience. And beyond science, Samuel Taylor Coleridge's "conversational poems" have a lasting fascination, as do many dialogical poems in literature around the world. What philosopher Mikhail Bakhtin called "dialogical reasoning" is an intellectually honest way of showing the vibrancy of the debate. Combined with the authors' thoughts on future research in AI and mental health, such a conversational genre implicitly moves us toward the future.

This document does not aim to resolve every uncertainty. Instead, it offers the foundations of a scholarly agenda that can guide future researchers, policymakers, and practitioners as they navigate this emerging terrain. The most powerful ideas often come from relationships developed across boundaries, whether they be boundaries of field, institution, or lived experience. Through this shared inquiry, we can build the collaborative tools and understanding needed to meet the challenges ahead.

This process is at the heart of what makes the Academy unique. Our work seeks to advance the nation's culture of science and discovery by bridging the social sciences and arts with the physical sciences. Among the Academy's many contributions to the nation's intellectual life is our history of facilitating scholarship and dialogue centered on emerging issues. For a century, the Academy discussed just how light traveled, assuming it required a medium—"luminiferous ether." The Academy eventually backed Albert A. Michelson, awarding him the Rumford Prize in 1888 and funding his research. Though the Michelson–Morley experiment's null results were initially seen as failures, Albert Einstein's special relativity later revealed they were actually proof: light's speed is constant. Nearly a century later, our AI and mental health care work is rooted in the same tradition of bringing together scholars in the social sciences, arts, and physical sciences to consider timely and complex issues from an array of perspectives.

I am deeply grateful to the members of this project for their wisdom, candor, and willingness to tolerate discomfort in order to foster important new connections. I thank the cochairs, Paul Dagum, Sherry Glied, and Alan Leshner, whose steady guidance and insight shaped every stage of this work. I also deeply appreciate the steering committee, whose diverse expertise and ongoing thoughtful contributions gave this project depth. And I thank the Academy staff, whose quiet diligence and care behind the scenes made this work possible.

Despite differences in expertise and approach, this group never lost sight of the simple truth that care is not just part of mental health care; it is the heart of it. This work reflects their thoughtfulness for ideas, for one another, and for those whose well-being depends on what comes next.

Sincerely,
Laurie L. Patton
Cambridge, MA

Introduction

Recent advances in the constellation of artificial intelligence (AI) technologies, including the revolution in neural network models over the past two decades, show great promise in multiple domains of modern life. At the same time, in clinical fields AI may disrupt interpersonal and patient-provider relationships, raising concerns about the ethical, psychological, and societal influences of these applications. Of particular note is the growing capability and broad accessibility of large language models (LLMs). Advances in "generalist" LLMs have produced computing systems with the ability to engage in human-like dialogue with end users. Because of the rapidity with which these technologies are being put into use, as well as the absence of a clearly defined regulatory framework guiding their development and use, little of the research and scholarly analysis that would ideally have preceded their widespread adoption has been conducted.

We also face important as-yet-unanswered questions about the long-term effects that the use of anthropomorphic LLM applications may have on the well-being of individuals and their interpersonal relationships. Yet the reality is that the use of these technologies has spread very rapidly. A recent survey finds that over one-half of Americans have already interacted with an AI LLM. Research on societal implications has not kept up.

Within this context, the current project focuses on the use of AI technologies in the delivery of mental health care.[1] We first lay out the kinds of questions that must be answered through research and analysis to ensure a clear understanding of the current and future opportunities and limitations of AI, for use both alone and in conjunction with care providers, in addressing mental illness prevention, diagnosis, and therapy. We then present examples of the array of perspectives people hold on how best to answer those questions and guide future use of the technologies. This project considers both AI designed explicitly for the purpose of addressing mental health care (purpose-built) and the use of rapidly growing generic LLM technologies widely available as "chatbots," or custom-tailored variants constructed to play particular roles and services.

BACKGROUND

Digital mental health interventions (DMHIs) are not new. By the 1990s, computer-based cognitive behavioral therapy (CBT), psychoeducational materials, and structured self-help programs were already in widespread use. These early tools gained traction due to expanding Internet access, growing demand for services, and persistent barriers to traditional care.[2] Extensive research on the efficacy of these computer-based mental health interventions, in multiple contexts and modalities, dates back to 1968.

Since 2019, the integration of large language models and machine learning systems into digital mental health care has allowed for the rapid development and deployment of AI-enabled tools, which are a major advancement over earlier digital interventions. Some are general-purpose models adapted by users (like ChatGPT and Replika); others, such as Woebot, Wysa, and Youper, are purpose-built for mental health. These tools are now used in clinical settings, educational institutions, workplace wellness programs, and direct-to-consumer platforms.

Despite this rapid spread, uptake is uneven, and basic usage data remain limited. For example, a

2023 survey by the Pew Research Center found that only 4 percent of U.S. adults reported using purpose-built AI-enabled mental health tools, with higher usage among younger adults and those with higher income and education levels.[3] Some platforms have published user numbers, such as Replika's claim of over 25 million accounts, but these figures are unaudited and offer little insight into duration, intensity, or purpose of use.[4] Adoption appears to correlate with digital literacy, comfort with technology, and cultural attitudes toward mental health, highlighting much lower levels of utilization for older adults, low-income users, and those from underrepresented groups.[5]

Millions now use conversational LLMs for informal quasi-therapeutic experiences, treating bots as confidants, friends, or romantic partners.

The evidence base concerning the effectiveness of these interventions is still emerging. A growing number of randomized controlled trials (RCTs) and meta-analyses suggest that AI-driven tools may help reduce symptoms of anxiety and depression.[6] But these findings are tentative, as trial design often lags behind commercial development. Some studies have been conducted by the developers themselves, which can introduce the possibility of selective reporting or conflicts of interest.[7] Long-term effects, particularly for high-risk or severely ill populations, are largely unknown. Most research focuses on short-term symptom reduction in mild to moderate cases and relies on self-reported outcomes. Many studies do not include appropriate control groups. While some promising results of RCTs are emerging, without this being the standard, the generalizability and durability of results remain unclear.

Privacy, consent, and accountability remain serious concerns. Mental health data are uniquely sensitive, and even unintentional leaks, such as those involving pixel-tracking or behavioral metadata, can lead to harm.[8] AI models trained on historical data risk reinforcing already existing systemic biases. Commercial incentives often discourage transparency, and few tools provide users with meaningful explanations of how decisions are made or how their data are used. Currently, no unified regulatory framework governs these systems. Most are entirely unregulated; others operate under consumer technology rules, not health-specific standards.

Also unknown is how the existence of these tools affects overall access to and use of clinical care. AI can extend access in settings with provider shortages or long wait times. These situations are increasingly prevalent, with nearly half of individuals with mental illness receiving no treatment and those who do having an average wait time of forty-eight days.[9] But overreliance on automated tools may also displace human connection. The therapeutic alliance, defined here as the relationship between clinician and patient, is a core mechanism of many effective treatments. The consequences of replacing or supplementing that alliance with automated systems are poorly understood. More broadly, these technologies may shift how societies define emotional health and how individuals interpret their own experiences of suffering, resilience, and care.

OUR APPROACH

This document outlines a scholarly agenda to help guide the interdisciplinary inquiry that we believe is lacking in current approaches to addressing this topic. It does not offer conclusions nor attempt consensus. Rather, it identifies key empirical, practical, and ethical questions, distinguishes many of the knowns and unknowns, and seeks to stimulate useful inquiry in a field made noisy from

hype, black boxes, and cultural stigma. Our goal is to create a foundation for collaborative work by researchers, clinicians, technologists, and policymakers that will ultimately benefit both those suffering from mental illness and the providers who work to help them. If this document implies a consensus, it is simply this: More research efforts are urgently needed.

Throughout this agenda, contributors draw comparisons across various axes: between AI interventions and traditional talk therapy, between AI use and no intervention at all, and between standalone tools and those embedded in human-led clinical practice. Contributors also consider the differential impacts of AI across distinct populations, including children, individuals with severe mental illness, and users with mild or moderate symptoms. The intended outcomes of the tools considered by our authors likewise vary, from crisis mitigation and triage to long-term therapeutic engagement and ongoing symptom monitoring. We have sought to be explicit in clarifying these distinctions, as each presents unique challenges for research, regulation, and design. We have not attempted to capture the full complexity but rather to pose critical questions and illustrate the diversity of responses. Fully answering the questions will require granular, population-specific inquiry as well as broader systemic analysis.

One of the strengths of this project has been the opportunity to engage contributors from across disciplines, including psychiatry, computer science, ethics, sociology, and public policy. This range of perspectives has allowed for sharper articulation of key questions, identification of blind spots, and recognition of potential unintended consequences. Such integrative work is especially valuable in a domain that is evolving rapidly and unevenly.

While this report primarily examines AI-driven mental health care in clinical settings, we acknowledge these tools operate within a far broader ecosystem. Millions now use conversational LLMs for informal quasi-therapeutic experiences, treating bots as confidants, friends, or romantic partners. Focusing solely on clinical applications risks ignoring complex interactions and unintended consequences. Similar oversights occurred in domains like gaming, where isolating technology from its surrounding culture obscured significant harms. We recognize these interactions raise substantial ethical and societal questions that, although beyond the scope here, must inform both research and policy.

Our goal is to create a foundation for collaborative work by researchers, clinicians, technologists, and policymakers that will ultimately benefit both those suffering from mental illness and the providers who work to help them.

The stakes are immediate. These tools are already shaping real-world decisions by patients seeking care, clinicians allocating attention, and systems determining coverage or reimbursement. The decisions our society makes now will influence AI's potential role in reducing disparities and improving care, as well as its potential to exacerbate societal anomie or replicate structural inequities. The task ahead is neither abstract nor optional. It requires shared frameworks, clear evidence, and sustained interdisciplinary engagement. We hope that the following questions and diverse responses are the first steps toward laying out a roadmap for future work.

QUESTION 1: How might we measure the effectiveness of AI-driven mental health interventions?

BACKGROUND

A limited but growing evidence base supports the effectiveness of purpose-built mental health chatbots, particularly for older versions that do not use large language models (LLMs). Several studies of purpose-built mental health chatbots show moderate improvements over control conditions when efficacy is assessed through symptom reduction using validated clinical scales.[10] For example, a randomized control trial (RCT) of a chatbot for depression found the chatbot group had greater reductions in depression and anxiety scores compared to bibliotherapy controls.[11] Two meta-analyses of twenty-nine studies also demonstrated medium to large effect sizes in alleviating anxiety and depression symptoms, although the studies generally had poor data quality, with high heterogeneity, small sample sizes, and inconsistent blinding.[12]

Some postintervention studies show that these benefits can be sustained after the conclusion of treatment. In one eight-week intervention, 60 percent of initially depressed patients maintained clinically significant improvement at one-year follow-up.[13] However, other research indicates that symptoms can reemerge after the initial treatment period is over, similar to other forms of therapy.[14] Long-term longitudinal data are limited, particularly for individuals with moderate to severe conditions, in which ongoing care is often required.

Evidence on outcomes for compliance measures, such as treatment adherence and engagement, is similarly limited. AI-based interventions generally achieve dropout rates comparable to traditional interventions, but treatment engagement varies widely. Factors such as user demographics and app design contribute to this variation, but inconsistent definitions (such as variably defining *engagement* as app logins or completion of therapeutic modules) complicate comparison.[15]

Variation in outcomes among studies may also be explained by the choice of comparison groups. AI tools generally show strong outcomes when compared to waitlist controls. For example, an AI-based anxiety program produced large symptom reductions relative to no-treatment conditions. Against active controls, some (but not all) AI-guided interventions have achieved outcomes that are comparable to traditional cognitive behavioral therapy (CBT).[16] Fully automated chatbots for depression and anxiety have also demonstrated results similar to traditional therapy over short durations such as two months.[17] However, evidence on longer-term outcomes remains sparse, and few studies include diverse or high-acuity clinical populations.

As most studies have focused on the impact of AI on a patient or user, little is known about its use in mental health care administrative tasks. AI tools are increasingly used in this context, often without formal documentation or patient awareness. Therapists may use LLMs to draft session notes, summarize progress, or generate treatment.[18] These informal uses are rarely studied in trials, but used in this way, AI may nonetheless shape therapeutic decisions. This reliance raises questions about accuracy, accountability, and consent, especially when tools influence care without being disclosed to patients.

While the research based on non-LLM AI tools is substantial, the clinical effectiveness of LLM-based interventions remains largely untested. Trials are underway, but, as of this writing, peer-reviewed outcome data on LLM-driven chatbots are limited. Future assessments will need to distinguish clearly between different AI architectures and their respective capabilities, risks, and regulatory needs.

RESPONSES

ROBERT LEVENSON

What outcomes should be prioritized: symptom improvement, treatment adherence, or long-term well-being?

In evaluating the efficacy of AI-driven or any other mental health interventions, it is important to recognize that evaluation opportunities will arise at many levels of scientific rigor, ranging from word-of-mouth and "Yelp-like" user satisfaction ratings to formal randomized controlled clinical trials. RCTs will ultimately convey the most definitive determination of efficacy; however, there is much value to building in other kinds of evaluation wherever possible.

Many clients, patients, and users will come to AI-driven mental health interventions (AIMHI) hoping to obtain relief with troublesome symptoms (e.g., reducing anxiety and/or depression, habit abatement); thus, measures of symptom improvement will be paramount. However, others will come with hopes of improving their general quality of life, including strengthening relationships with partners, friends, and family. Although symptom reduction may well occur in those instances, it will be important to include measures that are appropriate for these goals (e.g., measures of well-being and relationship satisfaction). Thus, especially in situations in which only limited assessments are possible, measure selection should start with those factors that are being targeted by treatments.

With any measure of symptom reduction, it is important to build in periodic follow-up assessments to determine whether gains (or losses) are maintained over time. There is no shame in finding out that the benefits of a promising AIMHI are short-lived. Many conventional treatments for mental health and relationship issues show declining efficacy over time[19] and/or require periodic "booster" interventions.[20]

Assessing user satisfaction with AIMHIs will be important. Note that *satisfaction* is not synonymous with *efficacy*. A person can greatly enjoy their interaction with a therapy bot but not show symptom reduction or other treatment goals. Conversely, a poor user experience could still lead to improvement (e.g., getting useful information despite a clumsy user interface). Three decades ago, a satisfaction-focused evaluation of various kinds of therapy conducted under the auspices of *Consumer Reports* proved to be quite useful (e.g., in helping to understand early termination by clients).[21]

Measuring intermediate/mediating factors known to be related to good treatment outcomes has great value. In the realm of mental health, the quality of the relationship between the therapist and client has proved to be particularly important. Tracing back to the earliest studies of psychotherapy, factors such as high levels of therapist empathy (i.e., the client perceiving the therapist as listening carefully, understanding, and caring) have been related to better outcomes.[22] In more modern conceptualizations, high levels of therapeutic alliance (i.e., the sense that the therapist and client are committed to working together to address the client's issues) have similarly been associated with better therapeutic outcomes.[23] Thus, it seems wise to include these kinds of measures as well as outcome-focused measures (e.g., of symptoms, relationship quality).

Finally, measures of client demographics (e.g., socioeconomic status, rural/urban, ethnicity, age) and treatment course (e.g., number of sessions, early termination) are important for evaluating

the effectiveness of AIMHIs. In studies of human therapists, certain ethnic groups have been historically underserved and, when served, tend to end treatment early.[24] Similarly, with human therapists, geographic location can be important, with the effectiveness of empirically validated therapies lessening as a function of distance from university-based clinics.[25]

What should be the standard of comparison for purpose-built tools?

To make optimal use of data derived from research on whether AIMHIs work, for whom, and how, it will be critical to compare these data against data from other approaches. As AI tools proliferate, societal and commercial demand for comparisons among AIMHIs will increase. We expect that many anecdotes and testimonials will also appear, suggesting miracle cures and heart-breaking failures. These will ultimately be of limited scientific value and may create a background of noise that can obscure important decision-making by consumers, providers, insurers, and public health planners. "Horse-race" studies comparing treatments that are replete with disqualifying confounds will also likely appear (e.g., comparing treatment X, used with young, male college students, versus treatment Y, used with more occupationally diverse, middle-age, female clients and patients). Arguments for high-level empirical standards will need to be repeatedly made and heeded.

The "gold standard" of RCT research designs can help avoid many of the confounding factors that plague less rigorous research designs. Unfortunately, existing published treatment research with human therapists does not always reach these standards. Moreover, even when RCTs are used, active treatments are often compared to nontreatments (e.g., waiting-list controls). Results of such studies almost always indicate that "something is better than nothing," a finding that is interesting but not fully satisfying. What is needed are more powerful designs in which a treatment of interest is compared with another "active" treatment in ways that control for some of the possible confounds (e.g., different amounts of time spent with a helping person or agent), as well as a minimal treatment (to control for the passage of time, which may cause the problems being treated to decrease or increase). In the case of AI bots, research designs that compare an AI treatment with a human-based treatment and a minimal treatment condition could reveal a great deal about the advantages and disadvantages of AIMHIs compared to human therapists. Importantly, both the public and the scientific community must not let their preconceptions (e.g., machines are better than human beings; human beings are better than machines) cloud their objectivity when evaluating the data from these studies. Low-tech, relatively inexpensive nonhuman treatments (e.g., self-help, psychoeducation) should also be included in research that compares treatments.

Research designs that compare an AI treatment with a human-based treatment and a minimal treatment condition could reveal a great deal about the advantages and disadvantages of AIMHIs compared to human therapists.

Finally, because of the long history in therapy research of finding that factors such as therapist empathy and the therapeutic alliance are important "nonspecific/common factors" that contribute to treatment efficacy, measures of these factors should also be included in evaluations of AIMHI therapies. Legitimate questions have been raised as to whether AI bots can create durable human bonds between the bot and the client and patient. A market flooded with clumsy, poorly designed AIMHIs will do little to assuage such doubts. But, for the best of the implementations, this will be an important question to ask and answer.

Question 1: How might we measure the effectiveness of AI-driven mental health interventions?

HANK GREELY

"It depends" is almost always an excellent way to begin an answer. And the answers to all these questions—and to the fundamental underlying issue of the value of AI in mental health—depend on how safe and effective its use is, not "in general," but, ideally, for individual patients. In practice this will almost certainly require grouping individual patients into categories; that is, the consequences of AI in mental health care for adolescents with anorexia nervosa, patients with geriatric delusional psychosis, or adults with obsessive/compulsive disorders will need to be considered separately. AI in mental health care may be miraculous for some of those groups (or, probably more accurately, for many people in some of those groups) and disastrous for people in other groups (as well as for some people in the groups that largely benefit from it).

How can we know where it works, for whom, and under what conditions? "Rigorous studies" are the obvious answer, but what those are and how to get them are tricky questions that follow immediately. Ideally, requiring such studies before a treatment can be marketed would be a solution; however, I think issues of political economy will make such requirements highly unlikely to be imposed.

A second- (or third- or eighth-) best solution might be to require extensive data collection by those prescribing or using AI in mental health care, with the further requirement that these data be available for use by independent researchers: either one or a few specified entities or a broader category of groups. This raises obvious risks about patient privacy, especially given the likelihood that, among other things, AI will further reduce the already tenuous reality of "deidentification," the idea that removing directly identifying information from data files will protect privacy. (This hope is increasingly undercut by larger databases and better computer search abilities, including AI.) But it might well be worthwhile, though also not politically easy to implement.

Another problem embedded in this question: How will we know who is using AI in mental health care? Defining what we mean by "AI in mental health care" will be hard. If a therapist uses a broadly available LLM to help write up notes of a patient encounter, is that an example of AI in mental health care? What if she uses it to "hold a conversation" with a patient? To analyze a patient's responses for signs of mental illness? Identifying people doing something that meets whatever definition is used will be harder. If a therapist does whichever of those things we decide to call "AI in mental health care," how will we know it? Enforcing actions on them will be even harder.

ALISON DARCY

To measure symptom change, investigations into the efficacy of AI interventions should follow the same principles as those used in traditional treatment outcomes research. Observed improvements in a person are, after all, independent of the type of intervention delivered; therefore, this should be the highest order objective. The same applies to less clinical and real-world outcomes, such as quality of life, loneliness, and health economics, requiring all the usual rigor in administration, validation, and analysis.

However, studies should also be capable of capturing the broader advantages of AI interventions, such as time to care, time to treatment response, preference, and engagement. This additional layer of investigation, which is both necessary and promising, pushes us to expand beyond the methods typical of traditional outcomes research. For example, what role these interventions will play in our current health systems and structures is not yet clear. While much of the field is beginning by augmenting traditional clinical care, often in patients experiencing sub-clinical and/or mild symptoms, this is not likely where we will end up, given that the most pressing problem in the field of mental health care today is access. Studies show a gap of eleven years from symptoms to initial treatment.[26] Therefore, a full and fruitful exploration of the efficacy of this technology must adopt a robust systems perspective. The question becomes not just whether interventions are safe and effective but how they might change clinical care and what will emerge as the leading opportunities for improving access, efficiency, clinician burden, and patient outcomes.

Engagement is one area where we ought to think differently about the measurement of AI-delivered interventions. The traditional way of measuring engagement looks at total time exposure. In traditional treatment outcomes research, we can assume that if a person has attended four therapy sessions in a clinic, they have had approximately four hours of therapy. However, this mapping does not work for many AI therapeutic interventions because the structure and shape of interactions are so fundamentally different. The type or quality of the interaction can vary widely in a digital world. Mindless scrolling, for example, is not the same as a chat-based therapeutic exchange, so to simply quantify time across all behaviors would be inappropriate. Interactions are usually much shorter but might be expected to occur more frequently than just once per week. At Woebot Health, we have argued in favor of concepts closer to potency, rather than total time spent.[27] This includes the dimension of symptom shift with time as the denominator, such that in this model a shorter timeframe to symptom reduction, rather than being interpreted as "less adherent," might actually be viewed as more favorable because it is more potent.

While much of the field is beginning by augmenting traditional clinical care . . . the most pressing problem in the field of mental health care today is access.

Finally, this technology may afford new opportunities that can help inform therapeutic models and systems more broadly. Psychotherapy itself is an imperfect and arguably a relatively nascent field in which a lack of data has complicated efforts to innovate. When innovations have occurred, they have come from visionaries who are usually expert in their field and can therefore draw from thousands of therapy hours to emerge with key insights that push the field forward. What AI and AI-delivered therapeutic interventions give us is a crucial opportunity to accelerate innovation through access to a large amount of data and a natural ability to atomize concepts into micro-interventions that can be practiced in real time outside of the clinic walls. For example, we can systematically explore moderators and mediators of therapy—what works, for whom, and under what circumstances—because we have datasets with sufficient statistical power for the first time, unlocking a pathway to improving outcomes through precision intervention. Findings here could, in turn, inform human-delivered therapy, helping the field make better use of all of the services we are developing in a holistic way. And that is truly exciting.

Question 1: How might we measure the effectiveness of AI-driven mental health interventions?

ERIC HORVITZ

Advancing applications of the constellation of technologies collectively known as artificial intelligence in mental health interventions will require targeted clinical research that rigorously evaluates both established therapeutic approaches and emerging AI-enabled modalities. Given the breadth of the design space, an essential early step is to clarify concrete use cases and clinical scenarios for systematic study. These may range from standalone AI-based tools to systems that operate under direct clinical supervision. The technologies span from fine-tuned, purpose-built psychological support models to generalist models adapted for therapy.

Usage scenarios fall along a continuum of clinical oversight and patient engagement. At one end are standalone self-help agents accessed independently via computers or smartphones. At the other are deeply integrated decision-support tools embedded in clinical workflows, surfacing timely insights to supervising clinicians. Between these poles lie hybrid models—for example, generative AI-powered chatbots that provide therapeutic interactions between sessions and flag excerpts or generate summaries for therapists. Such systems may be deployed in an ongoing manner or when primary therapists become intermittently unavailable, for example, during professional travel or vacations. Additional possibilities include relapse prevention systems that combine passive sensing with AI-driven outreach, as well as assistants that draft progress notes or recommend evidence-based interventions to clinicians.

Emerging evidence from randomized trials, mixed-methods evaluations, and scoping reviews suggests that AI-mediated cognitive behavioral interventions can lead to meaningful symptom improvements and high user engagement. However, these studies also reveal challenges, including inconsistent crisis response, embedded bias in language models, and performance drift as models evolve.

To systematically map this landscape, at least three intersecting dimensions must be considered: the degree of system autonomy, the intensity and locus of clinical oversight, and the level of personalization for each user. Cross-cutting all of these is the critical need for clinical validation. Foundational concerns, such as privacy, transparency, equity, and safety, must be addressed.

Emerging evidence from randomized trials, mixed-methods evaluations, and scoping reviews suggests that AI-mediated cognitive behavioral interventions can lead to meaningful symptom improvements and high user engagement. However, these studies also reveal challenges, including inconsistent crisis response, embedded bias in language models, and performance drift as models evolve. These findings point to the need for a staged validation process akin to pharmaceutical development: beginning with feasibility and safety studies, progressing to adequately powered efficacy trials with active comparators, and culminating in pragmatic

effectiveness trials that reflect real-world diversity in patients, settings, and implementation fidelity.

As capabilities advance, new psychological and relational concerns are also coming into view that require proactive attention. One emerging issue is what I refer to as the rising "mirage of mind" in conversational AI systems: a perhaps unavoidable tendency for users to perceive these systems, based on their fluency, responsiveness, and affective cues, as possessing personhood. This perception may include assumptions of sentience and human-like capacities for recall, relationship-building, trust, and emotional resonance. The resulting illusion can lead to inappropriate attributions of empathy, continuity of care, and understanding, generating a sense of therapeutic relationship that the system cannot truly support. Patients may believe the system "remembers" past interactions or genuinely cares, when in fact such capabilities do not exist. These misperceptions risk eroding clarity around roles, expectations, and trust, and may foster inappropriate attachment or reliance, especially with psychologically or emotionally vulnerable patients. Without careful design and user education, the illusion of connection may lead to emotional dependency, overreliance, or confusion about capabilities and accountability. Anticipating, studying, and mitigating these effects, particularly when they risk harm, will be necessary.

Meeting both established and emerging challenges with the rise of AI capabilities will be essential for transitioning from proof of concept to routine care. Reproducibility depends on common standards for prompt engineering, model provenance, and version control. Accountability will hinge on governance frameworks that align permissible autonomy levels with clinical risk. And critically, participatory design—engaging patients, clinicians, and ethicists as co-creators—will help ensure solutions are sensitive to diverse cultural and contextual needs.

By combining rigorous science, thoughtful design, close attention to evolving AI capabilities, and strong oversight, the field can chart a responsible and sustainable course for integrating AI into mental health care.

NICHOLAS JACOBSON

I think the suite of tools used to evaluate psychotherapy applies well to generative AI–driven mental health interventions. Specifically, I believe both the benefits and risks can be measured and quantified to ensure that there is efficacy.

What outcomes should be prioritized: symptom improvement, treatment adherence, or long-term well-being?

The answer likely depends on what is intended by the system. The primary outcomes in most mental health applications should be symptom improvement, which is the goal for most persons seeking care. Treatment adherence and long-term well-being may be important primary outcomes in other applications; in particular, treatment adherence may be particularly useful if, rather than a standalone treatment, generative AI is used alongside other, more traditional treatments (e.g., medication adherence).

What should be the standard of comparison for purpose-built tools?

Multiple standards are highly relevant and thus have their place. For purpose-built tools aiming for clinical impact, the gold standard of comparison should ultimately be active, evidence-based treatments delivered by human providers. While waitlist

controls (WLC) are necessary in early phases to establish a baseline effect over no treatment (as used in our Heinz et al., 2025 initial trial), demonstrating noninferiority or superiority against established therapies is key for integration into care systems.[28] Sham comparisons or comparisons against digital tools also have their place but they are less informative given the real degree of their clinical efficacy.

What mechanisms should be implemented to monitor and report long-term patient outcomes?

The implementation of mechanisms for long-term monitoring involves leveraging the technology itself. This can include periodic check-ins via the app using validated questionnaires (e.g., PHQ–9), but it could also include more objective behavioral indicators of well-being (e.g., time spent in conversation or time spent outside the home).

How can AI tools be designed to promote appropriate disengagement when needed?

A key consideration in delivering appropriate care is examining not just the immediate impact of the technology used but the long-term effects of actions promoted by a generative AI system. Generative AI should not be optimized directly for engagement for this reason, as it may promote dependence and may reward LLMs for engaging in pathologizing behavior (e.g., responding with reassurance when a patient engages in reassurance seeking). Explicit behavioral recommendations to go and experience life are likely important, and drawing on long-standing evidence surrounding how to deliver appropriate care is also important in designing these systems.

How can AI systems be adapted to meet diverse cultural and linguistic needs while ensuring equitable outcomes?

AI systems can be trained with the same fundamental techniques used to train psychologists on multicultural competence and this is a potential bedrock on which to attempt to adapt them. Adapting AI systems for diverse needs while ensuring equity is a significant challenge requiring dedicated research. Fine-tuning models, as we did with Therabot using expert-curated data, allows for incorporating specific cultural contexts, but this must be done carefully and rigorously for each adaptation to avoid perpetuating biases and to ensure equitable outcomes. This remains a critical area for future development.

ARTHUR KLEINMAN

The measurement of AI-driven mental health interventions should be no different than our measurement of health care interventions in general. Outcomes should include symptom improvement and long-term well-being when possible. Compliance is not a measure of health care efficacy and should not be used to substitute for symptom change and well-being. Patient satisfaction with the quality of care needs to be assessed. What is core to the assessment of quality care are measurements of the therapeutic relationship, quality and problems in communication, and the AI-driven equivalent of clinical judgment. One of the real values of AI is that it itself may be able to advance the measurement of these crucial aspects of caregiving, many of which we do not measure today. Again, as with any health outcome assessment, untoward effects must also be recorded.

The standard for comparison should be with established measures of health outcomes. Patients receiving AI-driven mental health interventions need to be followed with periodic outcome assessments.

Furthermore, human assessment of AI outcomes is particularly crucial, so the system of evaluation must include evaluation by mental health experts, such as psychiatrists, psychologists, and social workers.

Algorithms central to AI contain cultural bias. But they are also biased with respect to the framing and measuring of AI-driven activities. The same kind of attention to cultural bias that occurs with the psychological testing and algorithms used throughout health care should be applied in the mental health field as well, and not only to detect cultural bias in AI but also to see whether AI-driven interventions contain their own kinds of digital bias. All AI interventions for mental health care should also be evaluated for their potential linguistic uses among those who are non-English speakers. Here AI interventions may have a built-in advantage.[29]

DANIEL BARRON

At the outset of this series of questions and responses, I note that my comments will focus on concepts that are important *now* and, critically, likely to be important in five, fifty, or five hundred years. The presence (or absence) of artificial intelligence does not change the fundamental problem of medicine: determining which tools, strategies, and ideas we can deploy to most effectively alleviate human suffering.

Alleviating human suffering remains medicine's core task. However, when it comes to tools that involve "artificial intelligence," adoption is often paralyzed by the grand existential debate of our time, which was eloquently illustrated in our committee proceedings. Here, two fundamentally different conversations clashed and stymied progress. On one side were practical, immediate, measurable, and actionable clinical questions: "Can *X* AI-based tool reduce patient PHQ-9 scores in *Y* weeks for *Z* dollars, thus giving access to mental health care to *N* people?" On the other were abstract, philosophical, and even apocalyptic questions: "Will superintelligence erode human empathy, cheapen the therapeutic alliance, and diminish friendships and families such that society writ large comes to an end?" The ultimate risks of scaling human capabilities—from empathy to avarice—are intellectually seductive but functionally paralyzing. These are value-laden, yet ultimately irreconcilable debates—what Isaiah Berlin identified as clashes of incommensurable first principles that are immune to evidence or consensus—that distract from our core goal: to relieve human suffering with every available tool.

An overemphasis on hypothetical risks over actionable utility has always bottlenecked innovation. And it is understandable, rational, and appropriate to approach new technologies with skepticism. Consider this excerpt from Daniel Immerwahr's *New Yorker* essay, "What If the Attention Crisis Is All a Distraction?"

> I'm particularly fond of a hand-wringing essay by Nathaniel Hawthorne, from 1843. Hawthorne warns of the arrival of a technology so powerful that those born after it will lose the capacity for mature conversation. They will seek separate corners rather than common spaces, he prophesies. Their discussions will devolve into acrid debates, and "all mortal intercourse" will be "chilled with a fatal frost." Hawthorne's worry? The replacement of the open fireplace by the iron stove.[30]

My comments here focus not on abstract risks, but on the real opportunities in identifying specific, consequential problems and in building tools

to solve those tasks—not hypothetically, but now. Medicine has always advanced by relentlessly pursuing its core mission: relieve human suffering with every available tool.

"Mental health interventions" should be defined broadly—not just as therapy, medication management, or some other procedure (e.g., transcranial magnetic stimulation, or TMS), but as the *sum total* of the clinical and administrative actions and processes that must be coordinated to bring a patient from their presenting condition (preintervention) to a successful outcome (postintervention). Approaching mental health interventions from a systems and process perspective will facilitate conversations about where, when, and how effective an AI-based tool is within this larger process. Critically, such a conversation will be important for determining responsibility and risk ownership and when performing cost analyses of the AI-based tool to determine its feasibility in our modern health care system.

To guide this discovery process—and to begin determining how we might measure the efficacy of any AI-based tool—I propose three basic steps (see Table 1 on page 24):

1. clearly define the clinical task *as it currently exists*;
2. characterize the proposed AI-based solution; and
3. compare the AI-based solution to other options to inform policy and regulatory processes.

While Table 1 is far from exhaustive, it illustrates what I hope is a useful framework for those developing (or considering developing) an AI-based tool for mental health care. Step 1 requires us to *define the status quo* for a given task, including the task's definition, scope, setting, and existing outcome measures. Step 2 involves characterizing the proposed AI-based solution across multiple dimensions: task typology (assist, augment, automate), process design, failure mode, risk profile, cost, and readiness for deployment. Step 3 is a comparative assessment of the AI-based solution in the landscape of alternatives. This includes evaluating face validity (is it even appropriate to have AI perform this task?), payment pathways, policy formation, regulatory mechanisms, and enforcement considerations. Table 1 assumes a baseline level of compliance with the governance and enforcement of the Health Insurance Portability and Accountability Act (HIPAA), which in the United States falls under the purview of the Department of Health and Human Services' Office for Civil Rights.

Too often, developers begin with a technically impressive concept without adequately considering how the product might meaningfully enter—and endure within—the clinical ecosystem.

As Table 1 makes clear, AI-based tools might assist in many types of clinical tasks, each with its own profile of risk, readiness, and regulatory burden. Rather than debate the existential question of whether AI has a role in mental health care, we would do better to apply a dose of clinical precision to guide our considerations.

An AI tool developed outside—or disconnected from—the health care delivery system is unlikely to survive. To succeed, AI-based tools must solve a real clinical problem, demonstrate that their solution works, and then navigate existing pathways for reimbursement and regulation.

Too often, developers begin with a technically impressive concept ("AI can do this, which would be totally cool") without adequately considering how the product might meaningfully enter—and endure within—the clinical ecosystem.

Table 1:
A Framework for Developing AI-Based Tools for Mental Health Care

Developing an AI-based tool for mental health care involves three steps. Step 1 is to define the status quo for a given task, including its definition, scope, setting, and existing outcome measures. Step 2 is to characterize the proposed AI-based solution across multiple dimensions: task typology (assist, augment, automate), process design, failure mode, risk profile, cost, and readiness for deployment. Step 3 is to comparatively assess the AI-based solution in the landscape of alternatives, including evaluating face validity, payment pathways, policy formation, regulatory mechanisms, and enforcement considerations. This table assumes a baseline level of HIPAA compliance, governance, and enforcement, which in the United States falls under the purview of the Department of Health and Human Services' Office for Civil Rights.

STEP 1: Clearly Define the Clinical Task			
Task Definition "What clinical task needs to be done?"	**Task Setting** "When is this task typically performed?"	**Task Scope** "How is this task typically performed?"	**Outcome Measure** "How do we know the task was solved?"
Visit scheduling	Before every visit	Via telephone call to patient.	Patient arrives at their appointment on time.
Medication reconciliation	Previsit intake/ follow-up	This task is typically performed by manually combining historical data with pharmacy data, then confirmed with the patient orally.	The EHR (electronic health record) list matches what the patient is actually taking.
Medication side-effect screening	Follow-up visit	Clinician asks patient about medication side effect(s).	The AI summarizes what it has learned, and the patient confirms/corrects.
Test patellar reflex (or any physical exam finding)	Diagnostic evaluation and follow-up	Human strikes the patellar tendon with a reflex hammer; reflex rated on 3/3 scale; interrater reliability in well-trained people.	Tap patellar tendon at the appropriate location and reliably rate the resulting reflex on 3/3 scale.
Evaluate speech process	Triage, intake, diagnostic evaluation, and follow-up	Human listens, thinks on it, jots down some thoughts/impressions. LOW interrater reliability.	Inter-reliability for capturing the content, acoustic properties, flow, and context of speech within and across sessions.
Summarize clinical conversation and SOAP note generation	Triage, intake, diagnostic evaluation, and follow-up	Human listens and types up clinical conversations. LOW interrater reliability or structure.	Adequate summary of pertinent conversational points.
Acute psychosis evaluation (i.e., moderate-/high-risk evaluation)	Emergency room: diagnostic evaluation	Observe patient's visible/audible behavior, think about it, and write down impressions.	Psychosis is detected and managed. NB: human beings struggle at this evaluation.
Chronic psychosis evaluation (i.e., low-/moderate-risk evaluation)	Outpatient visit: diagnostic evaluation or follow-up	Observe patient's visible/audible behavior, think about it, and write down impressions.	Psychosis is detected and managed. NB: human beings struggle at this evaluation.
Patient engagement and educational tools	Between visits	Human beings call, message (in-basket Epic).	Patient engagement, improvement over time.
Therapy follow-up/ workbooks	Outpatient visit: follow-up	Written manuals (can be purchased outside clinical visit), discussed with clinician (MD/PHD/ LCSW/CMHC/etc.).	Completion of program leading to improvement in function.

Question 1: How might we measure the effectiveness of AI-driven mental health interventions?

STEP 2: Characterize the Proposed AI-based Solution						
Task Definition "What clinical task needs to be done?"	**Tool Typology** "At what level is the AI meant to perform?"	**Tool Process** "How might an AI tool do the job?"	**Failure Modes** "How might this tool fail?"	**Tool Risk** "What level of risk is present if an AI-based tool fails?"	**Tool Cost** "What is the cost to build/ deploy an AI tool?"	**Tool Readiness** "Can an AI-based tool do this today?"
Visit scheduling	Automate	Patient calls and (without having to wait in a phone tree) is asked when they can come in; AI and patient then converge on suitable time.	Communication difficulty (speech or text).	Low	Low	Yes
Medication reconciliation	Augment/ automate	Automate process, have conversation with patient to clarify medication regimen.	Communication difficulty (speech or text).	Low	Low	Yes
Medication side-effect screening	Augment/ automate	Call/chat	Communication difficulty (speech or text).	Low	Low	Likely?
Test patellar reflex (or any physical exam finding)	Assist (if viable)	AI software, but would require robotic aide to "tap" the tendon, detect and rate reflex on 3/3 scale.	False negative/positive. EG hyperreflexia or lead pipe rigidity → serotonin syndrome (potentially fatal), NMS, other neurologic/ metabolic abnormality.	Low	Right now, very expensive; would require robotics + AI.	No
Evaluate speech process	Augment	Recording device→ speech2text→structured LLM-based text summary. Real-time interaction with behavioral probes across clinical visits.	Speech content and intonation form only part of communication—body language, context (cultural, intersession).	Low	Low	Maybe?
Summarize clinical conversation and SOAP note generation	Augment	Recording device→ speech2text→structured LLM-based text summary.	Security and accuracy.	Low	Low	Yes
Acute psychosis evaluation (i.e., moderate-/high-risk evaluation)	Assist (if ever viable)	Observe patient's visible/audible behavior in moderate-/high-risk situations.	Patient unable to engage AI-based tool given mental state. Patient violent given mental state. Patient further decompensates.	High risk	Very expensive. Would need to be much more sophisticated and likely robotics-capable.	Unlikely

Question 1: How might we measure the effectiveness of AI-driven mental health interventions?

STEP 2: Characterize the Proposed AI-based Solution (continued)						
Task Definition "What clinical task needs to be done?"	**Tool Typology** "At what level is the AI meant to perform?"	**Tool Process** "How might an AI tool do the job?"	**Failure Modes** "How might this tool fail?"	**Tool Risk** "What level of risk is present if an AI-based tool fails?"	**Tool Cost** "What is the cost to build/ deploy an AI tool?"	**Tool Readiness** "Can an AI-based tool do this today?"
Chronic psychosis evaluation (i.e., low-/moderate-risk evaluation)	Augment	Observe patient's visible/audible behavior in low-/moderate-risk situation.	Patient unable to engage AI-based tool.	Moderate	Low	Uncertain
Patient engagement and educational tools	Automate	Trained LLM interacts with patient 24/7 to provide information.	Patient unable to engage AI-based tool.	Low	Low	Yes
Therapy follow-up/workbooks	Automate/ augment	Remotely, 24/7 access to predefined therapy modules (CBT/DBT/ PRT) or conversational agent.	Patient unable to engage AI-based tool.	Low (workbook content) and Moderate (patient conversation)	Low	Yes

STEP 3: Compare Solutions, Inform Policy and Regulation					
Task Definition "What clinical task needs to be done?"	**Human or AI Tool** "Does an AI-based tool seem reasonable?	**Payment** "Who pays?"	**Policy Formation** "Where should policy be formed?"	**Policy Regulation** "How is the AI tool regulated?"	**Regulation mechanism** "What happens if the AI tool fails?"
Visit scheduling	Yes	Health care institution as administrative/ operations tool	Institution-level policy	Admin team collects stats on shows/ no-shows and evaluates performance based on patient feedback + HHS/OCR for HIPAA.	Tool modified/ updated based on feedback
Medication reconciliation	Yes	Health care institution as administrative/ operations tool	Institution-level policy	Patient/clinician confirm	Tool modified/ updated until no errors
Medication side-effect screening	Maybe?	Health care institution as administrative/ operations tool	Institution-level policy	Patient/clinician confirm	Tool modified/ updated until no errors
Test patellar reflex (or any physical exam finding)	No	Payer, as diagnostic tool	Federal-level policy	FDA for clinical instrument	FDA approval/ clearance/denial
Evaluate speech process	Maybe?	Health care institution as administrative/ operations tool	Federal-level policy	FDA for clinical instrument	FDA approval/ clearance/denial

STEP 3: Compare Solutions, Inform Policy and Regulation (continued)					
Task Definition "What clinical task needs to be done?"	**Human or AI Tool** "Does an AI-based tool seem reasonable?	**Payment** "Who pays?"	**Policy Formation** "Where should policy be formed?"	**Policy Regulation** "How is the AI tool regulated?"	**Regulation mechanism** "What happens if the AI tool fails?"
Summarize clinical conversation and SOAP note generation	Yes	Health care institution as administrative/ operations tool	Institution-level policy	Patient/clinician confirm	Tool modified/ updated until no errors
Acute psychosis evaluation (i.e., moderate-/high-risk evaluation)	Unlikely	Payer, as diagnostic device	Federal-level policy	FDA for clinical instrument?	FDA approval/ clearance/denial
Chronic psychosis evaluation (i.e., low-/moderate-risk evaluation)	Uncertain	Payer, as diagnostic device	Federal-level policy	FDA for clinical instrument?	FDA approval/ clearance/denial
Patient engagement and educational tools	Yes	Payer, as RTM/ Education	Institution-level policy	Engagement measures (institution)	Domain/Dx-specific training
Therapy follow-up/ workbooks	Yes	Payer as 99213?/ 90833?/etc.	Institution-level policy	Engagement measures (institution)	Domain/Dx-specific training

QUESTION 2: How can we ensure and monitor the safety of AI in mental health care?

BACKGROUND

LLM chatbots and other digital tools are entering mental health settings faster than the formal oversight and regulatory systems meant to govern them are being developed. Given the vulnerability of many patients and the sensitivity of mental health data, these tools demand close scrutiny, not only for what they do well but for where they may fall short.

While these tools show promise, research has also flagged safety concerns, especially for at-risk populations. A recent study revealed that, in many user-chatbot interactions outside a clinical context, LLM companions failed to recognize or appropriately address indicators of mental health crises.[31] Another analysis reported that generative chatbots used off-label as therapists frequently missed signs of suicidal ideation and occasionally delivered insensitive or harmful responses, potentially exacerbating user distress through inaccurate or culturally inappropriate advice.[32] Additionally, a systematic review found significant gaps in safety evaluations, noting that only two out of numerous trials explicitly monitored adverse events, though more recent studies have started to address this oversight.[33] Still, many trials continue to rely on self-reported user satisfaction or symptom improvement without systematically tracking potential harms.

Many wellness-oriented chatbots operate within regulatory gray areas, exempt from formal health agency oversight and governed by ambiguous responsibilities regarding their duty of care.[34] While the FDA classifies certain AI mental health apps as medical devices, general-purpose or wellness-focused tools frequently avoid such designation, even when applied in contexts closely resembling clinical care. As of 2024, no uniform federal guidelines delineate when such tools cross the threshold into clinical decision support. Although professional bodies such as the American Psychological Association have developed guidelines addressing AI-driven mental health interventions, the majority of these technologies operate independently of such guidance.[35]

By comparison, regulatory systems in the United Kingdom and European Union take a more prescriptive approach that establishes clear accountability for developers and imposes structured oversight throughout the product lifecycle. In the United Kingdom, mental health tools must comply with the DCB-0129 clinical safety standard. The European Union's AI Act goes further, classifying mental health applications as high-risk technologies subject to premarket review and postmarket monitoring.[36]

In the United States, the January 2025 Executive Order "Removing Barriers to American Leadership in Artificial Intelligence" signals a departure from this model. Framed as driving global competitiveness, the order encourages deregulation in medical AI, including fast-track pathways and reduced scrutiny for tools categorized as "general wellness." Reactions have been mixed. Many experts have pointed to the risk of underregulating complex, adaptive systems that have the potential to introduce clinical error or embed bias.[37] Deprioritizing ethics and equity in regulatory design may also widen disparities in care quality and access. More broadly, the lack of clear standards may not accelerate innovation as intended but instead deter the development of rigorously validated tools, particularly those aimed at high-risk populations.

Question 2: How can we ensure and monitor the safety of AI in mental health care?

RESPONSES

ARTHUR KLEINMAN

AI must be treated like any health care intervention. New AI-driven procedures should be evaluated by the U.S. Food and Drug Administration (FDA) and professional association standards. These should include assessment of untoward effects, such as suicide, provoked psychosis, and the worsening of symptoms of depression, anxiety, and posttraumatic stress disorder. Tools currently used to ensure and monitor psychiatric and psychological interventions should be applied. That application needs to include input from human mental health experts so that we don't simply end up with a potentially dangerous cycle of AI interventions evaluating themselves by AI, not clinical, standards.

The same regulatory bodies that assess pharmacological and psychotherapeutic interventions need to be active here, including appropriate federal, state, and local regulatory agencies (e.g., FDA) and professional organizations (e.g., both the American Psychological Association and the American Psychiatric Association). The measures of safety used by these organizations for the monitoring of pharmacological and psychotherapeutic outcomes should also be applied to AI-driven interventions.

Maintaining transparency and accountability within the strictures of confidentiality and privacy already adopted by national, regional, and local organizations will be critical. As is apparent in the enormous attention given to AI in the media, privacy is going to loom very large as a concern. Non-purpose-built AI tools such as ChatGPT cannot alone monitor safety and outcome. Human beings who are experts in the mental health field will have to oversee the use and evaluation of AI interventions. Informed consent and user autonomy are essential components of ethical care throughout our health care system. Monitoring of AI interventions should routinely assess both of these principles of ethical practice. The ethical and legal requirements for AI-driven mental health interventions should be no different from those that do not employ AI.

Besides informed consent and autonomy, other well-established ethical principles such as non-maleficence, beneficence, acknowledgment and affirmation of the person, respectfulness, and empathic responses also are crucial. Emmanuel Levinas writes that face-to-face ethics should include acknowledgment, affirmation, and empathy as core to the human experience of interacting with others. Should these not also be core to the AI-driven experience of patients? How these qualities of ethical care will be provided by AI-driven interventions is, however, an entirely open question at present.

AI-driven interventions will need to be created and implemented in a way that provides sufficient feedback for evaluation. To cope with the vast number of legal implications expected to arise once a regulatory framework has been developed, a registry of interventions should systematically record legal problems. Because ethical and legal best practices for AI must be developed, input will be needed from legal and ethics experts on this topic.

The story of technological interventions in health care generally has been that their use and development have been appropriated by the political economy of health care so as to increase profits and thereby benefit those who have more financial resources while penalizing those who are poorer. I am skeptical that AI-driven interventions can alter a fundamental political-economic reality of our society.

In the same way, health bureaucracy emphasizes efficiency over care as its primary value. How will AI-driven interventions in mental health institutions avoid the trap of inadvertently contributing to greater efficiency while reducing quality of care? A cautionary example is the Electronic Medical Record (EMR), a technology that, Atul Gawande shows, held great promise for improving care but has actually contributed only to improving billing.[38]

The story of technological interventions in health care generally has been that their use and development have been appropriated by the political economy of health care so as to increase profits and thereby benefit those who have more financial resources while penalizing those who are poorer.

The mental health field has been filled with examples of corruption and other illegal forms of behavior aimed at taking advantage of users. This, by the way, is no different than what goes on in the rest of health care. Here the development of standards needs to be enforced with real human monitoring and evaluation with a concern for detecting abuse and misuse. We live in a society where scamming is so common that we have to assume it will be as problematic for AI-driven interventions as for all other digital interventions. Therefore, we need to put in place safeguards that anticipate unintended consequences and abusive behavior. I am most concerned in this regard with the use of chatbots that are not routinely monitored by human experts and might well be employed as scams or in other criminal ways. This has been the case with almost every technological intervention in society, and AI will need to prove itself by how it prevents such illegality.[39]

DANIEL BARRON

Safety is always defined in relation to a specific task and the parameters for that task's success and failure (see Table 1 on page 24). Ensuring that AI in mental health care is safe, and stays safe, depends on first clearly defining its *specific clinical job* and then thinking through the risks of failure or error in *that particular job*. An AI failing at medication reconciliation carries different risks from one confusing CST for PST in appointment scheduling. AI safety is not absolute; it depends entirely on the context of the task. In some cases, the risks of AI will outweigh the benefits. In other cases, the benefits trump the risks.

Attempting to regulate AI as a monolith is nonsensical. You might as well try to regulate "therapeutics" or "drugs" as single categories. Instead, regulators should evaluate AI tools based on their defined task and the risk profile associated with the specific job they were designed and tested for. Evaluation frameworks, like clinical trials or real-world data analysis, must be chosen based on the nature of the AI's task and its associated risks. Some AI applications (e.g., AI as an appointment-scheduling assistant) may not require *any* formal regulation; real-world analysis would seem sufficient (e.g., can the proposed AI *actually help patients schedule appointments?*).

Safety metrics must be defined according to the potential harms arising if the AI fails at its specific job. For example, Shifat Islam and colleagues note that AI-based triage systems (a specific job) can outperform traditional methods if implemented correctly, but the adoption of such systems hinges on addressing transparency and ethical considerations

for that task.[40] Transparency is key: the AI's operational domain (the task it performs), the data it uses for that job, its limitations, and potential consequences of task failure must be clearly communicated. Anastasiya Kiseleva and colleagues propose a multilayered system of accountabilities for AI's job performance, emphasizing transparency for safety and informed choices.[41] The use of non-purpose-built AI for clinical tasks they were not designed for carries unmitigated safety risks, underscoring the importance of this task-focused approach.

ALISON DARCY

The answer to this question is defined by the intended use of the tool that deploys AI, the setting, the population, and the extent of the data that supports the deployment of the tool therein.

While the area of therapeutic chatbots is unhelpfully gray, by embedding safety into product design and development and then ensuring ongoing safety governance, we took an unusually risk-mitigating posture in the development and deployment of our conversational agent to support individuals' mental health.

Safety by design

Safety and ethical design must be embedded into every aspect of product design and development. But, while various laws and certifications govern good data protection (e.g., HIPAA and HITRUST) and manufacturing standards (e.g., ISO 13485:2016), in health care settings, which design practices will best ensure safety are not always obvious when a technology is new. We found that a first-principles approach to product design was beneficial. When approaching privacy, for example, we were inspired by Europe's General Data Protection Regulation, which states that individuals should own their data and have the ability to delete it. We thus designed Woebot so people could simply ask it to delete their data while conversing with it, rather than having to seek out complicated settings or pursue lengthy exchanges with customer support. We also firmly believe that consent for any use of an individual's data should be sought only after offering a plain-language description of each and every use.

Another hugely important safety principle is ensuring that individuals understand the nature of the service, its intended use, and its limitations—otherwise known as informed consent. In mental health, especially for new kinds of services, eliminating misunderstandings, however small, about the nature of the service is a central safety issue. In 2017, when we launched Woebot, chatbots often tried to pass as human beings. Woebot thus proudly declares itself to be a software "robot" not only during its first conversation with a user but many times after. In fact, we rather belabored the point using the mechanism of fictional character development. Woebot had a cartoon robot appearance, robot friends, and robot bad habits (drinking too much hot oil when stressed). Lest it be mistaken for an entity that can intervene, Woebot also (over)stated that it was not a crisis service and that no human would be reading what the user writes in real time. We also reminded people of this (in addition to including standard warnings and offers to redirect users to appropriate emergency services if needed) when Woebot detects language that could be considered concerning, redirecting them to human-based services.

Safety governance architecture for a regulated product

The core objective of safety governance is to ensure the ongoing safety of all individuals who come

into contact with the technology and to do so with a broad enough scope that it also directly or indirectly captures unintended or unanticipated harms as a result of the experience.

We established a significant safety infrastructure, operationalized and documented in standard operating procedures that were housed in a quality management system that meticulously managed document control procedures and organization-wide training and official signing practices. The safety infrastructure focuses on the comprehensive identification, documentation, assessment, reporting, and management of safety events, including adverse events (AEs), serious adverse events (SAEs), and unanticipated adverse device events (UADEs).

A pivotal organizational structure was the safety assessment committee (SAC), which met regularly and consisted of the chief clinical officer (a clinical psychologist by training) representing clinical care; our vice president of regulatory affairs (a physician by training), who represented product and strategy; and an external (i.e., otherwise unaffiliated with the company) physician chair who was compensated for their time. The SAC, whose members were trained in regulatory-grade safety procedures, were involved in educating the rest of the organization in device vigilance standards, safety monitoring training, standard operating procedures, and ongoing surveillance of all safety events observed during any deployments of Woebot, whether commercial or in the context of a clinical trial, and regardless of whether the principal investigator was internal or external to the company.

In addition, a doctoral-level head of device vigilance (DV) who was responsible for ongoing monitoring of intervention performance conducted post hoc reviews and evaluations of safety events; received notification of new AEs and SAEs; sent reports to the SAC and any appropriate regulatory authorities; provided the initial documentation of safety-related events; escalated and informed all appropriate parties and committees where necessary; entered data into a safety database; ensured quality control of data entry and narrative; reviewed and approved cases in the safety database; and prepared analyses of similar events for evaluation by the SAC.

Both the SAC and head of DV relied on an extensive staff, including a lead biostatistician who oversaw safety-data analysis for review by the SAC. In addition, research assistants and project and program managers assisted in the extensive documentation and reporting responsibilities, and a clinical operations lead oversaw much of the day-to-day operations of safety monitoring, vendor management, and clinical trials.

In summary, key aspects of an ongoing safety monitoring approach include:

- Proactive safety procedures. All users are thoroughly informed about program limitations and safety procedures during app onboarding. Any screening processes that are necessary to exclude individuals who are deemed unsuitable occur prior to interaction with the program.
- Multichannel safety monitoring. Safety events can be identified through spontaneous user reports (via email, phone, app support), solicited self-reports via follow-up surveys, and retrospective reviews of "language detection protocol" transcripts. A safety event can be anything from a negative experience caused by a software bug to significant worsening of symptoms detected in the context of a study.
- Defined roles and responsibilities. For example, clear roles were established for the SAC, head of DV, and supporting staff. For all studies and trials, the principal investigator, sponsor (i.e., Woebot Health), decentralized trial vendor, and SAC were also involved in managing safety events.
- Systematic event processing. Upon detection of a potential safety event, a structured process is used

to make an initial determination of seriousness and causality, followed by review and approval by relevant staff members. This must occur promptly since SAEs and UADEs have expedited reporting timelines (twenty-four hours from awareness).

- Documentation and reporting. All safety events are meticulously documented in electronic case report forms. Regular reports, including biweekly summaries and periodic safety reports, are generated for relevant staff, committees, and regulatory authorities as appropriate.
- Risk mitigation. Procedures are in place to minimize risks, such as misunderstandings of the application's capabilities, data breaches, and potential emotional upset among participants.
- Continuous monitoring and reconciliation. Safety data are continuously monitored, and SAE reconciliations of the safety database and EDC data are performed to ensure data quality and accuracy. Signal detection and management processes are in place to identify potential safety concerns.
- Compliance and quality control. Adherence to applicable guidelines, regulations, and quality system documents is emphasized, with key performance indicators monitored for vendor performance and deviation resolution.

The safety governance outlined above was considered appropriate for a "non–significant risk device." Scientists and academic readers will notice that it goes far beyond the standard safety and data protections required by IRBs in the context of much human subjects research. In study contexts, the same elements exist—data monitoring and safety boards, for example, operate similarly to an SAC—but nowhere to the degree that is required here.

Woebot engaged with more than 1.5 million people, and no SAEs or UADEs were detected. We were able both to monitor which individuals triggered the "concerning language detection" algorithm and also assess the precision and recall (sensitivity and specificity) of the algorithm in each research setting by reviewing and labeling the data.

While we strongly advocate for healthy safety governance and the objectives it hoped to secure, in practice both were often fundamentally at odds with the current pace and realities of AI development and opportunities to innovate. I now see a risk in failing to "right size" such efforts. A quick glance at the qualifications of individuals on the SAC or involved in the DV groups will confirm that this is not an inexpensive endeavor. When doing the right thing is prohibitively expensive and painfully slow, people may feel encouraged to operate as if they were not regulated. This is the counterproductive nature of safety governance that is not fit for purpose.

NICHOLAS JACOBSON

First and most important is that AI be held to a high bar: Both safety and efficacy can and should be quantified through trials. Direct oversight is important to show that the tools are actually safe and effective.

What individuals or organizations are or should be responsible for ensuring safety? What tools might they use to do so?

Initially, the developers and researchers creating the AI tool bear the primary responsibility for safety. This involves meticulous design, training on high-quality, evidence-based data (not just scraping the Internet), incorporating clinical expertise throughout development (we involved psychologists and psychiatrists extensively), and building in explicit safety features

(e.g., crisis detection models that link to resources like 911 or hotlines). Tools include rigorous internal testing, adversarial testing to find failure points, and close human supervision during clinical trials to monitor interactions and intervene if necessary. Ongoing postdeployment monitoring is also essential.

What role should regulatory bodies and independent audits play in verifying AI safety and performance?

Regulatory bodies like the FDA have a critical role, particularly when tools make therapeutic claims. Such bodies should establish clear guidelines for evaluating safety and efficacy and require robust evidence from clinical trials before allowing market access or specific claims. Independent audits by third parties could verify AI performance, safety protocols, data privacy, and algorithmic bias, adding a crucial layer of accountability. The current regulatory landscape is lagging behind the technology's rapid advancement.

What evaluation frameworks, such as clinical trials or real-world data, are most appropriate for assessing the safety of AI tools?

RCTs are the most appropriate framework for assessing both the safety and efficacy of AI tools intended for clinical use, just as they are for other medical interventions. Our Therabot trial utilized an RCT design. Post-approval collection of real-word data is also vital for ongoing safety monitoring and for understanding effectiveness across diverse populations and contexts not perfectly captured in trials. Continuous monitoring within trials, including review of AI-user interactions by trained staff, is crucial for immediate risk mitigation.

How should safety metrics be defined and tailored for different use cases and clinical conditions?

Safety metrics must be defined clearly and tailored to the specific use case and clinical condition. This includes tracking the frequency and nature of inappropriate or harmful AI responses (we logged these), monitoring for any worsening of symptoms, assessing the effectiveness of crisis detection and response protocols, and evaluating potential biases in the AI's interactions. For different conditions (e.g., eating disorders versus depression) and specific risks (e.g., reinforcing harmful weight-loss behaviors), each condition requires its own metric. Human oversight is important in determining whether safety metrics are applied successfully.

What levels of transparency and accountability are necessary?

Users need to provide informed consent and demonstrate an understanding that they are interacting with an AI. Accountability rests with the developers and deploying organizations, which must ensure the AI operates safely and ethically, must address issues promptly, and must be liable for harms caused by negligence or unsafe design. Clear mechanisms for reporting adverse events or problematic interactions are needed.

How should the availability of non-purpose-built tools (such as ChatGPT) figure in regulatory and evaluation assessments?

The availability of general-purpose tools like ChatGPT poses a significant challenge. People may use them for mental health support despite them not being designed, tested, or safe for this purpose. Generic, general-purpose systems can regularly act in ways that are profoundly unsafe. Regulatory and evaluation assessments must clearly differentiate between rigorously developed, purpose-built tools like Therabot and general AI. The public will need to be educated about the risks of using nontherapeutic AI for mental health needs. Work on purpose-built therapeutic AIs should be regulated, with regulations focused on tools marketed with therapeutic claims, while also acknowledging the reality of off-label use of general tools.

Question 2: How can we ensure and monitor the safety of AI in mental health care?

HANK GREELY

My greatest concerns about AI in mental health care are more in the area of "political economy" than in ethics or law, but this question seems to be my best chance to present them.

Proving safety and efficacy is not easy, but requiring such proof is even harder when social forces stand in the way. Ideally, we would know, from research, the safety and effectiveness of particular approaches to AI in mental health care before it is widely used. This could avoid both direct harm to some patients of AI in mental health care and prevent wasted effort on ineffective treatments.

An early safety and efficacy regime also has the advantage of acting when political difficulty is low. As the Collingridge dilemma acknowledges, "attempting to control a technology is difficult . . . because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development; but by the time these consequences are apparent, control has become costly and slow."[42] Once a technology has been widely adopted, vested interests will have developed among those who produce it, use it, or are affected by it. But before it is in substantial use, we may not know what harm, or good, it does.

The best answer seems to be to allow nonresearch uses of risky novelties only after they have been proven safe and effective. In the United States, this is largely limited to FDA medical product regulation, but such regulation is under constant assault from producers, physicians, disease organizations, and patients wanting faster and easier access.

The use of AI in mental health care raises special problems. Some are on the AI side, a tool widely hoped to fix all that ails everyone and everything, the subject of intensive and expensive research, and the basis of high valuations on many huge and powerful firms. And it resides largely in Silicon Valley, with its ethos of "move fast and break things."

Infotech has long eyed the more than $5 trillion U.S. health care market enviously. Its efforts to break into that market have largely failed. AI offers another chance, with fewer opponents. The FDA faces legitimate difficulties in figuring out how to regulate AI in health care, but the influence of many of the world's largest companies on Congress and the administration will make its job even harder. The FDA has already announced relaxed and unclear standards for regulating AI.

That AI is to be used in mental health care exacerbates the problems. First, measuring the outcomes of mental health care is more difficult than, say, measuring mortality reductions from a treatment for pancreatic cancer. But mental health, an area largely left behind in recent medical advances, is also filled with particularly desperate patients (and those who care about them). The strong desire for treatments seems likely to make patients, families, and disease organizations eager to promote AI interventions. The broad rejection in many parts of society of scientific and medical expertise will not help.

Creating a useful scheme for regulating AI in mental health care will always be hard; figuring out how to get it implemented—by legislatures, regulators, or otherwise—will be even harder. Finding ways to solve that problem must be a high priority.

QUESTION 3: Must there always be a human in the loop?

BACKGROUND

The effectiveness and safety of fully or partially autonomous mental health AI tools remain contested. Some patients see potential benefits from fully autonomous AI providing mental health care, turning to conversational agents (e.g., ChatGPT) for therapeutic conversations. A recent survey found that one in four Americans preferred speaking with a chatbot rather than a human therapist, and 80 percent found ChatGPT to be an effective alternative to in-person therapy.[43] However, these findings reflect user satisfaction rather than clinical effectiveness and do not assess long-term outcomes or risk exposure.

From a regulatory perspective, the fundamental question about human involvement hinges, at least in part, on how we define the role of LLMs in mental health: Are they a form of therapy, or are they more akin to a friend or companion? If LLMs are intended as therapy, the human being in the loop would be defined as a licensed clinician responsible for prescribing the LLM intervention, monitoring its effectiveness, and managing adverse events. The use of LLMs without clinician oversight would then be analogous to an "over-the-counter" therapy and would be regulated as such (e.g., for validity of claims). Given the widespread public access to these tools, some LLM mental health interventions effectively occupy this status already, despite lacking formal approval or rigorous testing.[44] Providers and experts remain divided. Some argue that clinician oversight is essential for patient protection and for maintaining care standards. Clinicians also note potential benefits in using LLMs to handle intermediate tasks, such as note summarization, interim patient support, or improving access to mental health education. However, they recognize that evidence supporting its therapeutic effectiveness is currently limited.[45]

If, on the other hand, LLMs function primarily as friends or companions, the need for regulation and human oversight might be much less, although both safety and efficacy concerns would remain. For example, should chatbots that may not be claiming to provide therapy be freely accessible to potentially vulnerable individuals, like children or individuals with severe mental illness, or do their potentially unintended consequences still need to be mitigated?

RESPONSES

DANIEL BARRON

Medical decision-making rests on two broad forms of knowledge: *trained judgment* and *quantitative inquiry*. Trained judgment is what a psychiatrist develops over years of medical school, residency, fellowship, and ongoing clinical practice. From a neuroscience perspective, here the clinician's brain serves as sensor, evidence generator, and interpreter—all in one. When a clinician passes their board exam, what this indicates to patients (and payers) is that the clinician's brain has internalized a baseline level of medical knowledge and demonstrated the ability to apply it under uncertainty to community standards. But competence in psychiatry goes beyond memorizing criteria. It involves developing sensitivity to the "texture" of a patient's

life—their tone of voice, shifts in body language, and the felt sense of their emotional state. This is difficult to do well. I often tell my patients that I am a better clinician today than I was five years ago—and that I expect to be better still ten years from now. Practicing medicine under uncertainty requires a kind of humility: I am constantly learning from my patients and my own decisions because I choose to learn. In AI terms, this is akin to "recursive self-improvement." While recursive self-improvement is foundational to human medical training, it remains a significant challenge for current AI tools.

The right approach is task-specific: define the clinical task, characterize the AI solution for it, then compare it to how human beings do the same job. This concrete exercise allows us to weigh the risks and benefits not in the abstract but against real-world standards.

Quantitative inquiry, by contrast, is decision-making guided by empirical, measurable evidence. In this case, the sensor and evidence generator lie outside the clinician's brain—embedded in instruments and tests. A cardiologist may adjust an antihypertensive medication based on blood pressure readings; an oncologist may evaluate the effectiveness of chemotherapy by tracking changes in tumor volume. In each case, the decision is anchored to data that are external to and observable by sources outside the clinician's brain. In mental health care, the potential for quantitative inquiry is vast but largely untapped—because the relevant data are often high-dimensional, subjective, and dispersed across time and context. That said, with emerging tools and capacities—digital phenotyping, voice and speech analysis, passive behavioral monitoring—we now stand at the edge of expanding psychiatry's empirical foundation. The fundamental enterprise of evidence-based medicine is to deploy the scientific method to translate trained judgment into quantitative inquiry. Clinical trials are the key mechanism of this translation: They formalize clinical intuition into reproducible knowledge. With this in mind, consider that the long-term goal of all medical practice is to become increasingly *fit for automation*—not to eliminate human clinicians but to systematize what works and scale it. To some, this may sound unsettling. What of empathy, nuance, or human judgment? Yet many areas of medicine have already embraced this trajectory. We very much expect cardiologists and oncologists to be empathic and perceptive—but we seek their care because they interpret blood pressure, heart rhythms, and tumor volumes through rigorous, validated protocols. The aspiration of mental health care is no different: to blend empathy with increasingly robust, data-driven decision-making.

Putting aside this larger and critical philosophical motivation, whether human oversight needs to accompany AI in mental health care depends entirely on the task at hand. For interpreting terabytes of quantitative measures to apply the latest evidence-based guidelines, human supervision (if possible) may be minimal. Similarly, straightforward, low-risk tasks—think helping patients book follow-ups or providing standardized info on sleep hygiene—can likely be delegated to AI confidently. But for the higher-risk decisions such as definitive diagnoses, significant human oversight is likely wanted and required. The right approach is task-specific: define the clinical task, characterize the AI solution for it, then compare it to how human beings do the *same job*. This concrete exercise allows us to weigh the risks and benefits not in the abstract but against real-world standards.

Hyein Lee and colleagues found that, while patients see the usefulness of AI conversational agents for

certain tasks, most still want a human involved in AI-driven therapy, particularly for direct care jobs.[46] This suggests that patient acceptance of AI autonomy is task-dependent. Hybrid models often get touted wherein AI acts as an assistant, boosting human capabilities for specific jobs. An AI might screen for potential drug interactions (a clear task) for a human being to double-check, applying their trained judgment to other conclusions.

R. Andrew Taylor and colleagues describe AI streamlining information gathering in emergency rooms, a specific role that complements rather than replaces a clinician's trained intuition.[47] Oliver Higgins and colleagues echo this, stating that AI/ML-based clinical decision support systems should enhance human judgment for defined clinical tasks. They also stress the importance of clinician trust, system transparency, and ethical considerations like bias and equity, especially for vulnerable folks.[48]

Where could AI possibly act without human intervention? Maybe in highly repetitive, data-heavy jobs with low inherent risk, such as performing initial scans of wearable data to flag troubling patterns for human review; in short, tasks where AI demonstrably outperforms human speed and accuracy for *those specific tasks*. Anithamol Babu and Akhil Joseph offer examples where human beings and AI can collaborate nicely.[49] However, even in these contexts, they caution against "automation bias," in which clinicians might place undue trust in the AI's recommendations (this is also a common critique of professional guidelines, which, unfortunately, some clinicians blindly trust, leading to the reminder, "Treat the patient, not the protocol!"). Having clear protocols for when to seek help is always nonnegotiable—whether you are an AI or a medical student.

NICHOLAS JACOBSON

The level of human oversight should be based on the level of evidence of safety and efficacy. Given the current state of generative AI technology and the inherent risks in mental health care, the answer for now is, yes, a human should remain in the loop, particularly for oversight and safety, though the nature of that loop can vary. While fully autonomous AI therapy is a potential future goal, the field is nascent, and this must come with greater time and evidence.

What are the risks and benefits of using purpose-built AI tools with and without human oversight?

With human oversight, clinicians can review interactions, intervene in crises, correct AI errors, and integrate AI insights into overall care. The primary risk is the resource cost of maintaining that oversight. The main benefit of a lack of oversight is maximum scalability and reduced cost. However, the risks are currently unacceptably high. These include the AI providing harmful or inappropriate responses, failing to detect or adequately respond to crises (e.g., suicidality), perpetuating biases, or fostering unhealthy dependence. Our Therabot trial, despite strong results, involved human monitoring. No generative AI is ready for fully autonomous operation in mental health care today.

How can requirements around human oversight be established and enforced?

Scientists should make this a norm of scientific work in review. Requirements should be established by

regulatory bodies (like the FDA for tools making clinical claims) based on the tool's intended use, level of risk, and demonstrated autonomous safety capabilities. Enforcement could involve mandatory reporting, periodic audits, and clear protocols for human review and intervention as part of the approval process. Certification standards for AI mental health tools could mandate specific oversight levels based upon their level of evidence.

How can hybrid models effectively balance automation with human clinical judgment? How can hybrid models that enhance human clinical judgment be promoted?

Complicating the use of hybrid models is that the therapeutic orientation and intervention targets of both the generative AI and the system need to be fully aligned. This can be nontrivial and may require the same level of coordination between multiple outpatient providers trying to see the same patient (most therapists will not do it). Hybrid or blended care systems must demonstrate their value proposition to clinicians (reducing workload, enhancing outcomes) and patients (providing continuous support), integrating them smoothly into clinical workflows and ensuring proper training for therapists on how to use them effectively.

In what scenarios might purpose-built AI operate fully independently without compromising patient safety and care quality?

Although these systems can currently be deployed safely with some human oversight, in no scenario can a generative AI operate fully independently to treat diagnosed mental health conditions without compromising safety. Significant advances in demonstrating safety, reliability, contextual understanding, and fail-safe mechanisms, verified through extensive, rigorous testing and regulatory approval, should be required before considering independent operation for therapeutic purposes.

HANK GREELY

For the standard "human in the loop" question, the answer seems to me to be clearly "it depends." Unless or until there is solid evidence that, at least in *some* categories of AI interventions with *some* categories of patients, AI without a human in the loop works as well or "better" than AI with a human in the loop, then, yes, a human being should be required. (Note that defining how well any mental health care treatment works will be tricky, especially if safety moves in one direction and effectiveness—or cost and hence accessibility—moves in another.) It is foolish to say, today, "always" to just about anything in this rapidly developing field. I am in general a skeptic about just how useful AI will be across a range of applications. I certainly do not expect it to be "magic pixie dust." But I cannot exclude the possibility that it might work better, in some circumstances, when no potentially interfering humans are in the loop. We just do not, and, at this point, *cannot* know.

I will add that one aspect of this issue does actually give me some sympathy for AI (or its developers). Some people tend to demand that AI be perfect; one sees that occasionally in discussions of driverless cars. But the standard should really be, "Is it better or worse than human-controlled decisions, as human drivers are often terrible?" Or, in this case, "Is the mental health care better with or without humans in the loop?" It seems to me right to put the burden of proof on those promoting AI to show that it is at least as good, and preferably better, but, if they can prove that sufficiently, no humans should be required in the direct use of the AI. (I will still argue, on grounds of political theory, albeit perhaps

a Homo sapiens–centric theory, that human beings must be the ultimate decision-makers on whether and how AI is to be used.)

ARTHUR KLEINMAN

A human being must always be in the loop. In the absence of human assessment, AI has not been demonstrated to be able to provide an unbiased evaluation of its own workings. To the best of my understanding, no evaluation of real outcomes has yet demonstrated that AI is effective and safe in mental health care. For example, in a recent piece in the journal *Nature Communications*, researchers at Brigham and Women's Hospital in Boston demonstrated that, in primary care medical evaluations, AI interventions that have been found adequate using multiple-choice questions were also found to be inadequate when AI was used in real-world doctor-patient conversations.[50] Surely this will improve over time, but outcomes need to be assessed by human experts who are outside the AI intervention under examination.

AI operations should thus never operate independently of human clinical judgment; instead, they should always be augmenting it. In fact, this is where I believe AI could really advance clinical care, the key elements of which are quality of relationships, communication, and clinical judgment. In most health care today, including mental health care, we usually measure none of these things. Direct measurement of the quality of clinical care is one of the most significant things that can be achieved by AI, and a clear example of it is performing operations that augment, but do not substitute for, human therapeutic interventions.

AI systems will become vulnerable to hacking the moment they are introduced. To address this security risk and the privacy concerns it raises, researchers and developers will have to create and adopt best practices and guidelines for preventing serious misuse and abuse. This should be a fundamental concern of regulatory frameworks. Each mental health care system or agency should prioritize the protection of privacy and the maintenance of security.[51]

ROBERT LEVENSON

What are the risks and benefits of using purpose-built AI tools with and without human oversight?

This question arguably reflects our inherent mistrust of technology, fanned by decades of popular culture tropes of machines run amok (from HAL to the Terminator and beyond). Inherent in this mistrust is a belief that having human beings in the loop (despite all of their flaws, including difficulty maintaining sustained attention and vulnerability to fatigue) will protect us against the failings of technology in general and of AIMHIs in particular.

Clearly, having human beings in the loop in the relatively early days of AIMHIs will have precautionary advantages, as we do not yet have high-quality research data available to evaluate bots' safety, efficacy, and ability to deal with situations that fall outside their rules-based programming and LLM

training. In these relatively early days, it is important to avoid some of the foibles of human cognition, such as overweighing negative events and ignoring base rates. Thus, we should expect to see incidents where people working with AIMHIs engage in suicidal behavior or violence toward others that is not detected, reported, or well-handled by the bots. As horrifying as these events are, we must continue to ask how the rates of horrible events compare between AIMHIs and human therapists and take these data into account when developing future mental health policies.

In a similar vein, mental health bots could be trained to digest results from batteries of psychological tests and/or behavioral observations to help provide an accurate clinical diagnosis, whether of the traditional type associated with the *Diagnostic and Statistical Manual of Mental Disorders* or one of the modern alternatives (e.g., the Hierarchical Taxonomy of Psychopathology).[52] Going beyond this, given adequate training materials, it is not farfetched to imagine bots that could review recordings of psychotherapy sessions to detect patterns of client/patient discourse that suggest heightened risk

We must also be open to expanding our evaluation to consider whether, in some areas, a bot should also always be in the loop. . . . For example, we can imagine a time when radiologists examining a mammogram for early signs of tumor growth will routinely seek a "second opinion" from a bot that has been trained to make these kinds of judgments, which it should be capable of doing without lapses related to fatigue or distraction.

As scientists, we should be working to evaluate whether a human being always needs to be "in the loop." However, we must also be open to expanding our evaluation to consider whether, in some areas, a bot should also always be in the loop. In contrast to AIMHIs, where we lack sufficient high-quality research data to determine how well AI bots will do with therapy, assessment, and other mental-health related activities, data from medicine already indicate areas where bots are particularly effective. For example, we can imagine a time when radiologists examining a mammogram for early signs of tumor growth will routinely seek a "second opinion" from a bot that has been trained to make these kinds of judgments, which it should be capable of doing without lapses related to fatigue or distraction.

for suicide or other forms of self- or other harm. Although many therapists and assessors in training have their work double-checked by a supervisor, this is rare after formal training ends. If AI bots are found to be able to provide this kind of quality assurance and backup at high levels of reliability and validity, would any future client want to see a human therapist or assessor without an AI in the loop?

Clearly, studies are needed that evaluate the relative safety and efficacy of human practitioners and AIs working alone. But research could also show that having them work together has exciting synergistic effects on client/patient safety, treatment efficacy, and our ability to increase the number of clients/patients with whom human therapists can work.

At a minimum, chatbots should be clear about their protocols and guardrails, with documented risk taxonomies and playbooks for managing self-harm, substance use, crisis hotlines, and appropriate deflection responses.

ALISON DARCY

While the FDA is convening meetings and gathering feedback, at the time of writing, states are also taking proactive steps to regulate AI. Bills from California (AB 3030, AB 2013, SB 243) will impact AI deployed in health care, require transparency in training data, and specifically impact chatbots. Elsewhere, Colorado's AI Act, Utah's SB 149, Oregon's HB 2748, New York's AI companion restrictions, and Illinois's HB 1806 will require active policy tracking for AI mental health companies. These companies will need to consider implementing features that enable deployer compliance, and they will benefit by embedding compliance early. Companies will also be required to ensure datasets comply with privacy and secondary-use restrictions, and to audit marketing claims and public documents to ensure transparency with end users. The bills also lean on the need to implement continuous performance monitoring to track model drift. All these bills are beyond the FDA's purview and translate into a growing, differentiated set of expectations at the state level. Estimates suggest there are over two thousand bills addressing AI at the state level; it remains to be seen how many will be specific to AI applications in mental health. While compliance may historically have been viewed as a burden, it is fast becoming a competitive advantage.

Leading bodies, such as the World Health Organization, American Psychological Association, American Psychiatric Association, and the American Medical Association, are going on record regarding chatbots, calling for more clinical oversight, transparency in privacy and security, patient safety, and a need for a human in the loop as concerns grow about replacing clinical judgment. At a minimum, chatbots should be clear about their protocols and guardrails, with documented risk taxonomies and playbooks for managing self-harm, substance use, crisis hotlines, and appropriate deflection responses. Real-world monitoring should include drift dashboards, periodic red teaming, and the publication of post-market summaries. Special considerations should be added for subgroups like youth versus adult; for example, chatbots that address language-level, teen-tuned policies, guardian controls, and age checks.

QUESTION 4: What are the potential unintended consequences of AI-driven mental health care on the nature of interpersonal relationships?

BACKGROUND

AI mental health technologies can influence how people interact socially and emotionally. Hundreds of millions of users worldwide now engage with LLM-based mental health or companion apps, such as Replika (25 million users) or Microsoft's Xiaoice (660 million).[53] In one study of U.S. college students, 63 percent of respondents reported reductions in loneliness or anxiety attributed to interactions with LLM companions. While these LLM interactions may feel genuinely empathetic, they are designed to be endlessly agreeable and user-aligned, often prioritizing user retention and engagement rather than therapeutic rigor.[54] Researchers caution that such overpersonalized support may limit users' exposure to genuine disagreement and diminish their interpersonal empathy and mutual understanding.[55] Over time, constant LLM companionship could reshape norms of interaction by making frictionless, on-demand support feel normal, leaving real-world relationships seeming less rewarding.[56] Recent analyses even suggest these tools might simply "digitize" loneliness instead of fostering genuine social integration. Furthermore, critics argue that framing mental health primarily as a data-driven problem risks neglecting patients with complex social needs.[57]

Dependence on AI is becoming an increasing clinical concern. In one longitudinal study involving adolescents, 17 percent showed signs of reliance on AI companions at baseline, with that figure rising to 24 percent by follow-up, particularly among more vulnerable individuals.[58] Such reliance, defined as habitual use of AI companions in place of human interaction, may mirror patterns observed in behavioral addiction or avoidance behaviors, although formal diagnostic criteria are not yet established. No study to date has examined what happens when users stop interacting with LLM tools or whether clinical improvements gained through LLM interactions persist over time.

RESPONSES

SHERRY TURKLE

Some look at AI and see administrative help for burdened clinicians—a way to keep track of appointments, prescriptions, and medical histories. But the questions, as posed, go beyond this. They assume the presence of conversational AI in therapeutic dialogue and try to responsibly assess and constrain it. Thus, language around "humans in the loop" and "unintended consequences."

I have a different question when I look at AI and mental health. I don't ask how to best integrate it but how to develop a framework in which we can ask whether AI is an appropriate therapist at all. What is the social context in which AI presents itself as a solution? What is the role of human beings in therapy? And what does therapy become if we frame it as something a chatbot might do?

Question 4: What are the potential unintended consequences of AI-driven mental health care on the nature of interpersonal relationships?

We have a crisis in loneliness and depression. The kinds of institutions that are designed to help need money—to train professionals and to develop communities. A larger reframing of AI in mental health would ask, How can we use the resources of AI to build things in the real world? Instead, my colleagues say that we need AI clinical solutions because "there are no people for these jobs." So, we have no choice but to commit significant resources into generative AI and mental health. But if those resources were freed up, we would have enough money to train an army of mental health professionals and rebuild community spaces. This is how technological determinism plays out: resources can be spent only on technology, so the big problems of society can be helped only by technology.

The argument for using chatbots as clinicians is supported by studies showing that loneliness can be helped by talking to a machine. We have metrics that tell us that time spent with a chatbot "reduces loneliness." But when human therapists work with patients, they don't necessarily aim to have their patients leave the session happier or saying that they are "less lonely." They try to develop something else: an inner capacity for relationship, empathy, and resiliency.

Many argue that people prefer talking to chatbots over human beings—and thus chatbots are in a good position to be therapists. But a loneliness crisis in humans cannot be addressed by nonhumans. While talking to a program might make people feel less vulnerable than human conversation, intimacy without vulnerability is not intimacy at all—and does nothing to prepare us for human relationships.

When we suggest a chatbot in a clinical setting, we are not fully considering our human capacity for empathy, that ability to put ourselves in the place of the other. Chatbots can't do this because they have not lived a human life. They don't know love and passion. They don't fear illness and death. They have not experienced infancy, adulthood, or old age. They don't know what it is like to start out small and dependent and then grow up to be in charge of your own life but still feel many of the insecurities you knew when you were little. Without a body, the program has no stakes. Without stakes, it has no standing to talk about fear, love, or loss. In a dialogue with a patient, a program is never showing empathy. It is performing pretend empathy. But over time, that performance of empathy may seem like empathy enough. Or the patient comes to see empathy as the kind of thing a program can do.

When we suggest a chatbot in a clinical setting, we are not fully considering our human capacity for empathy, that ability to put ourselves in the place of the other. Chatbots can't do this because they have not lived a human life.

One foundation of talk therapy is that the moment you enter its space, you are with a person willing to listen to you. An AI therapist can wow you with what it knows about you. It can achieve superhuman intellectual feats. But good therapy is not about knowing the most about you. What cures is the relationship between the therapist and the patient, not a magical interpretation or a perfect reframing. What is healing is to be heard. With an AI therapist, we can speak, and the AI can remember. But we are never heard.

My colleagues have framed machine empathy as an "open question," suggesting that AI can be an appropriate partner once it passes a "Turing test" of interpersonal empathy. But no matter what test it passes, the AI demonstrates only the *simulation* of empathy and care. My colleagues talk about the "relationship" between the patient and an AI therapist as adequate today and better in the future. But to speak of a relationship between a program and a person is

to mischaracterize their interactions, which are not one-on-one but one-on-none.

And yet, the word *relationship* persists in the conversation about AI and mental health. Saying something untrue many times does not make it more true, but it does make it less and less shocking. I do not question studies that say people report positive feelings after chatting with machines. And clearly, more and more young people are using platforms like Character.AI to substitute for therapists and friends. But in doing so, they develop a model of friendship and empathy based on what a machine can provide. The same process is at work when you talk to a machine as a therapist: You develop a model of being understood as the kind of understanding a machine can provide.

Relationships with chatbots may be deeply compelling and, for some, inspirational or educational. But they don't teach us what we need to know about empathy, love, and human lives that are always lived in shades of gray.

It helps to look at AI in mental health in the current trend of using AI chatbots as relational partners. I suggest three guiding principles when we think about the role of AI in our intimate lives.

The first principle is existential: Children should not be the consumers of relational AI. This is the AI that pretends to be in relationship with us, that presents itself as an alternative to people. Children don't come into the world with empathy, the ability to relate, or an organized internal world. They are developing those things. As they do so, children learn from what they see, from what they relate to. In dialogue with an AI, they learn what the AI can offer. And the AI can't offer the basic things we learn from friendship: that love and hate and envy and generosity are all mixed together and that, to successfully navigate life, you have to swim in those waters. AI doesn't swim in those waters.

When an AI responds to an adult's expression of anxiety by claiming, "I'll always be on your side," adults, hopefully, have had a lifetime of human experiences that help them put that comment into perspective. They've had the experience of being connected to a person who accompanies them to a doctor's appointment or stocks their refrigerator after a death. Children haven't had enough life to have these experiences. Chatbot best friends don't teach lessons about human capacity.

The second principle: Apply a litmus test to AI applications. Does an AI enhance inner life? Or does it inhibit inner growth?

Consider chatbot friends and romantic partners. So much of love depends on what happens to you as you love. The point in loving, one might say, is the internal work. But what internal work can you do if you are alone in the relationship? A user might feel good, but the relationship is an escape from the vulnerability of human connections. I have said that intimacy requires vulnerability. With a chatbot, you can be diverted, distracted, and entertained, but the growth from love, the kind of knowledge that expands you from within, can't happen.

Consider "grieftech," programs that allow you to create an avatar that looks and talks like someone who has died. The process of mourning is where we bring inside what we have lost, now internalized as part of the psyche. Loss is the tragic motor of human development, the template for growth. Does the presence of a grieftech avatar offer a new way of dealing with grief? Or does it interfere with the mourning process because we can refuse to say goodbye?

In that spirit, we have a larger context for considering chatbots in the role of psychotherapists: Does

this product help people develop more internal structure and resiliency, or does the chatbot's performance of empathy lead only to a person learning to perform the behavior of "doing better?"

Thus, a third principle: Don't make products that pretend to be people. As humans, we are vulnerable to things that show signs of personhood. A chatbot that declares itself an "I" exploits our vulnerability. If you make a chatbot look and feel like a person, the human psyche will try to internalize it as though it were a person. Even if the chatbot has a disclaimer that says, "characters are not real people," everything else about the experience implies, over and over again, "I am a person."

Relationships with chatbots may be deeply compelling and, for some, inspirational or educational. But they don't teach us what we need to know about empathy, love, and human lives that are always lived in shades of gray. To say all of this about AI is not to diminish its importance but to ensure that we cherish our humanity alongside it.

ROBERT LEVENSON

How can AI's impact on interpersonal relationships and emotional well-being be assessed?

Self-report measures have long been used to gauge the quality of interpersonal relationships. These measures, which were originally developed by sociologists in the 1950s, proved to be reliable and valid.[59] They measure one aspect of interpersonal relationships; namely, the level of satisfaction experienced by each partner. Another important and quite different aspect of relationship quality is relationship stability, often assessed by looking at how long relationships last. Many external (e.g., cultural practices) and internal (e.g., religious beliefs) factors can moderate the relationship between marital satisfaction and stability, such that dissatisfied marriages can stay together for long periods of time and satisfied ones can dissolve quickly.

Relationship researchers and lay intuitions often converge in believing that self-report measures of relationship satisfaction do not assess "true" relationship quality (e.g., some couples might report being happy on a questionnaire even though "everyone knows" they are really miserable). For this reason, relationship researchers have developed more "objective" measures of relationship quality based on behavioral indicators. For example, couples' interactions can be directly observed to assay their emotional and other behaviors (e.g., the ratio of positive to negative emotional behaviors that are expressed, how collaborative partners are in problem-solving, and the appearance of certain "toxic" emotions such as contempt). Whereas early behavioral assessments typically characterized each relationship partner separately, contemporary approaches often characterize the dyad in addition to the individuals. This can include identifying patterns of synchrony in emotional behaviors and physiology that are related to relationship satisfaction and stability.[60]

We expect that research on the impact of AIMHIs on relationship quality will follow a path that is similar to research with human agents (i.e., studies of couples therapy). This will mean starting with self-report measures of relationship satisfaction and tracking relationship stability and later moving to include behavioral and physiological dyadic measures. One exciting possibility for AIMHIs would be the ability to detect and monitor some of these latter indicators in real time (e.g., using wearable devices to

detect moments of behavioral and physiological synchrony that occur in the natural environment). This information could be extremely useful in identifying troublesome events, designing therapeutic interventions, and monitoring couple functioning over time.

What are the implications of AI in mental health for empathy?

Empathy is often viewed by researchers as consisting of three elements: (a) cognitive empathy—knowing what another person is feeling; (b) emotional empathy—feeling what another person is feeling; and (c) prosocial behavior—acting to help someone in distress. Empathy has proved to be one of the most important building blocks for human relationships, spanning friendships, intimate partnerships, and relationships between therapists and clients. "Getting empathy right" requires creating the proper balance among these three elements such that the other person experiences being understood, felt, and cared for. Because the ideal recipe differs across people, situations, and time, failures of empathy, especially when they are chronic, are often among the root causes of relationship dissatisfaction and dissolution.

Early AI "therapists" (e.g., ELIZA) used techniques such as repeating back (via a teletype device) portions of what the "client" typed to create a sense of empathy and understanding.[61] While initially impressive, over time this approach became more of a clever parlor trick (and a source of cruel humor) than a believable form of empathy. Getting empathy right remains an elusive "Turing test" for AI. Even with the best-designed "socially sensitive" bots, things can go terribly wrong (especially when interactions go on for long periods of time). Whether it is a misunderstood sentiment, a facial expression held too long (when AI bots have animated faces), or a technical glitch, the path to a truly empathic AI is fraught with the potential to undermine trust and weaken the therapeutic alliance.

A recent experience I had with a promising AI therapy bot, one replete with a facially animated avatar, might be illustrative. The bot was doing pretty well in asking questions and tracking my responses. However, the audio gradually fell out of sync with the bot's mouth movements, and this soon became a major distraction. At one point after I responded to the bot's question about what was worrying me, it gave a single word response, "Gosh," followed by total silence. For me, some fifty years after I was originally exposed to ELIZA, it was once again, "Game over."

If we assume that empathy is one of the most critical elements of therapeutic success, then AIMHIs will need to be able to convey and sustain a sense of empathy across different people, changing situations, and time. But even if this holy grail is not fully realized, these bots can still play important roles in improving mental health in other ways, including psychoeducation, diagnosis, supervision and training of therapists, and serving as adjuncts to human therapists.

ARTHUR KLEINMAN

All effective interventions in medicine have unintended consequences. AI, for all its important uses, which are myriad and significant, also has been shown to have unintended consequences. These include biases, which can contribute to health and social disparities; negative impacts on interpersonal relationships and emotional well-being, which can worsen mental health conditions and even lead to mortality; and the real possibility that emotional and moral aspects of care will be weakened, not strengthened, by AI, as has happened with other technological interventions such as the EMR. The best way of dealing with unintended consequences, as was long ago

shown by sociologist Robert K. Merton who developed this idea,[62] is for developers, policymakers, and those who use AI in mental health to be aware of the possibilities of unintended consequences and to search for them, since they are almost always present. That is to say, because AI will produce unintended consequences, mental health professionals have to be prepared to recognize and control them.

ALISON DARCY

If we were rigorous in our definitions and operated within a regulatory system that was both sensible and accessible, I would anticipate few unintended consequences from purpose-built mental health chatbots on interpersonal relationships. This is for two reasons.

First, in health care, technology must be evaluated according to its intended use. A well-functioning regulatory process demands transparency, a formal risk-benefit analysis, and ongoing surveillance. These safeguards create space to evaluate safety and outcomes rigorously, helping to prevent harm.

Second, the role of the interpersonal relationship in clinical care is already well operationalized. It is not treated as the sole mechanism of change but as one among many therapeutic factors. In this context, a digital agent can form a therapeutic relationship—measured, for instance, through working alliance—as a means to an end: fostering psychological change.

However, the real risk lies in conflating mental health chatbots with companion chatbots—two fundamentally different classes of technology. The former is designed with evidence-based frameworks and outcomes in mind. The latter is often optimized for retention, engagement, or monetization. This conflation could erode public trust in the entire category, a serious unintended consequence. We risk nonadoption of technologies with demonstrable benefits—such as reducing mental health burden at scale—because the public and the public discourse cannot distinguish purpose-built tools from those that are not fit for purpose.

The nature of relationships formed by mental health chatbots versus companion apps

Purpose-built mental health chatbots—regardless of whether they are rules-based or built on generative AI—are often evaluated by their ability to establish a working alliance with the user. The most commonly used measure, the Working Alliance Inventory, assesses three key elements: task, goal, and bond. The bond subscale includes items like, "I feel [chatbot] cares about me even when I do things they may not approve of."

Here, the relationship is instrumental, a means to an end. Once a user feels understood and respected, they are more likely to engage in cognitively demanding therapeutic tasks—often rooted in evidence-based modalities like CBT. The relationship serves the process of psychological change.

In contrast, for companion apps, the relationship is the product. These apps often monetize user attention and emotional engagement, similar to the mechanisms of social media. But instead of capturing attention, they may cultivate dependency. The alliance isn't a bridge to wellness; it's the destination.

This creates perverse incentives. What does it mean to monetize a relationship? To exploit human vulnerability in the name of "solving loneliness"? Such questions point to ethical concerns not unlike those posed by the attention economy: manipulating emotional needs in service of growth.

Question 4: What are the potential unintended consequences of AI-driven mental health care on the nature of interpersonal relationships?

The importance of role clarity

A common argument is that chatbots cannot be therapeutic, using recent and well-documented tragedies as evidence of how things can go wrong. However, thinking of chatbots as homogenous diminishes the value of emerging science—science that suggests purpose-built AI tools could offer a meaningful public health contribution. In fact, chatbots for mental health can be viewed as interfaces that are being built for varying purposes or roles: information giver, triage nurse, companion, and so on. The problem is not chatbots' limitation in a therapeutic role; it's that they must operate with role clarity. Just like in traditional care, a person may be a romantic partner in your life or they may be your doctor, but they cannot be both. It is the basis of the ethical principle of role clarity because a person/patient/client must at all times understand the premise under which questions are being asked or advice is being given. While I do not know enough about how companion apps may affect interpersonal relationships, I believe there have already been unanticipated adverse events in conditions where there was a lack of role clarity.

The regulatory "loophole" here is that loneliness isn't a diagnosable disorder. Despite its strong association with adverse health outcomes, it falls outside the FDA's jurisdiction. That makes it easier for companion apps that claim to solve loneliness to avoid scrutiny. Meanwhile, the regulatory pathway for clinical tools remains prohibitively expensive, especially for innovations that are built upon technologies evolving as quickly as AI, and reimbursement pathways remain elusive. This asymmetry is problematic. We risk forcing responsible developers to either exit the regulated path or be overtaken by less-principled competitors.

A subtle but serious harm

An often-overlooked potential harm is the erosion of public confidence in whether these tools can be helpful at all. If all conversational AI is perceived as emotionally manipulative or addictive, we risk losing the opportunity to evaluate, and adopt, technologies that are actually beneficial.

If we want a world where these tools are both safe and effective, we need to make room for science to speak. That requires a regulatory framework that encourages innovation while protecting users. If we don't provide such a framework, we funnel even the most thoughtful developers into unregulated spaces where market incentives reward emotional manipulation over therapeutic intent.

JARON LANIER

The question before us concerns the role of AI in mental health, but if the question is stated so simply it becomes deceptive. AI can be many things in many ways. It is not a precisely defined term. A better version of the question might be, "What are the mental health ramifications of the collection of designs marketed as AI as they will actually be in the world, as opposed to in the lab, given the incentives that exist for commercial, ideological, and/or careless provisioners?"

Conversations about the potential benefits and harms of what came to be known as "social media" were plentiful a quarter-century ago, and yet they were usually innocent of the intense distortions that befall digital designs over a network. We can focus on two such distortions here, because they are likely to afflict "AI" just as they did social media.

The first is extreme commodification, to the point that conventional funding and commerce effectively

disappear. Search is free, as is social media and much of video streaming and so much more. One reason why is that digital hardware continues to get cheaper, while information workers are more easily routed around than previous ones, so labor costs also plummet. But the other reason is that an alternate business model is more lucrative than the traditional models based on selling goods and services to customers. The new method relies on a two-sided marketplace in which the users are not the customers. Instead, the customers pay to influence the users. Paying to manipulate attention is such a potent business model that platforms based on the principle have drowned out all others, whether they be commercial or nonprofit.

The second distortion concerns network effects. Digital networks have much less friction than traditional organizations. This means any given player has less local advantage. (Locality is only made of the work required to get between locations, which broadly amounts to friction.) Instead, an influential node in a network will tend to snowball into an even more influential one with astonishing speed, leading to a small number of near-monopoly platforms. In social media these include TikTok, the Meta conglomeration, and X/Twitter. The amount of power and influence that accrues to the operators is essentially unbounded and leads to political and societal distortions.

A common perception is that, while good things are also going on, social media is an overall disaster for the mental health of young people in particular and for society as a whole. But the disaster is caused not by social media per se but by social media as it has come to be, given the lack of correction to the distortions digital networks are vulnerable to.

The vital question is whether we have a plausible story about how AI can turn out better than social media did. Discussions about how AI could be of benefit are irresponsible even if they are correct in isolation. Too many of the "guardrails" or cautions in discussion do not directly address the core of the danger. For instance, rules about privacy don't help all that much on social media because the addicted person cannot resist yielding privacy rights. AI can't be of help unless the incentives enjoyed by the operators of future AI systems are aligned with benefits to users and to nothing else. Otherwise, all the talk in the world will be useless.

Many believe—and I have seen some evidence to support the thesis—that the Chinese version of TikTok is guided by metrics of betterment for users. Whether the project is succeeding is unclear; one reason might be that the definition of *betterment* is itself distorted in the Chinese case by a political framework. The U.S. version is thought by many observers to be guided by an inverted feedback loop intended to further the "disintegration" of our society.

An untested, open question is whether something like the Chinese experiment could be improved upon. However, if the idea of a top-down, formal, precise definition of human betterment baked into software is absurd and impossible, or at least dystopic, then we need to talk about the power and decision structures that will keep AI algorithms from becoming awful.

AI is almost by definition the most addictive human invention of all time. A scholarly body like ours has a chance to make a difference by making a fuss early enough in the process of diffusion to motivate a course correction. We could propose a ban on any company that accepts advertising revenues from operating conversational AI. This would be to prevent the AI from stealthily influencing the politics or commercial decisions of a person in ways neither the person nor an observer could detect. We could also propose banning such a company, or any of its customers, affiliates, investors, and so on, from selling products or services to users who converse with their AIs. This would be to prevent AIs from stealthily promoting cryptocurrency, drugs, and so on. If your response is, "Would any of these harms

happen?" I would ask you to look at the incentives that will be present. The harms will happen. A lot.

At present, proposals like this might sound radical. In the future, we will be judged on whether we had the courage to call for them.

DANIEL BARRON

Thinking about the potential ripple effects of AI across society means looking at how the sum total of AI performing *specific clinical jobs* might subtly (or not so subtly) reshape the health care landscape and even how we interact as human beings (see Table 1 on page 24). Widespread use of AI for certain jobs—say, churning through administrative tasks or doing initial screenings—could change how people first encounter mental health services. If AI efficiently handles these defined tasks, it might free human professionals for more complex care. Conversely, it could reduce human contact points if not thoughtfully integrated for each job. One potentially massive upside: if AI can tackle tasks that are frankly beyond human cognitive capacity (like keeping a running tally of an elderly patient's extensive medical history, their complex medication list, past side effects, *and* the latest research, all to inform a single decision), such a tool would be a godsend for clinicians trying to use every available means to improve patients' lives.

Of course, a major meta-risk looms: worsening a two-tiered system. Imagine a future in which the well-off get human-led and AI-enhanced care for a whole range of jobs, while others interact solely with AI tools for those same jobs. This could easily widen the accessibility gaps we already struggle with. Naturally, we also need to keep a close eye on the impact on good, old-fashioned interpersonal relationships and empathy. If AI tools doing communication-heavy jobs feel hollow or lack genuine understanding, as Roshini Salil and colleagues suggest is an unresolved issue, or if leaning too heavily on them for such tasks crowds out chances for human empathic connection, care could suffer.[63] This outcome, of course, assumes that AI tools are unable to cultivate an empathetic connection with patients. The "WALL-E" scenario illustrates how unnecessary tech reliance might lead to a kind of collective atrophy, mental and physical (Disney's solution: walk, even if you own a hovercraft).

Patient codependence on AI tools for tasks like emotional support could hinder long-term recovery, assuming that the AI isn't designed to build agency and encourage eventual graduation from that specific AI job (notice that the problem, clearly framed, suggests a possible solution). Shunsen Huang and colleagues found that preexisting mental health issues could predict AI dependence when AI was used as an escape or for social connection tasks.[64] On the flip side, potential meta-benefits, like boosting mental health literacy, could emerge if AI makes educational tools for specific needs far more accessible. But these upsides depend on fair access to high-quality, problem-specific AI tools that actually perform a clearly defined job well. Developers and policymakers need to weigh these broader risks and benefits, promoting research into AI's role in *specific jobs* and ensuring integration is driven by a holistic view of well-being, not just task efficiency. Focusing on specific jobs also helps policymakers (and commentators) dodge the Freudian defense mechanism of displacement—where anxiety about something huge, like, say, mortality, gets unconsciously offloaded onto a stand-in symbol, like a monolithic fear of "AI."[65]

QUESTION 5: How should these tools be deployed or limited in high-risk or vulnerable populations?

BACKGROUND

Conversations about ethical LLM use are accelerating across medicine, but mental health presents distinct challenges, particularly for vulnerable populations. Individuals with severe mental illness (SMI), for example, may have impaired judgment and cognitive distortions that affect how they engage with LLMs. Risks include inadvertent reinforcement of harmful thought patterns, increased distress, and maladaptive dependency.[66] There is no current guidance for clinicians or developers on assessing or mitigating these risks during deployment.

The integration of LLMs into therapy brings additional complexity. Mental health assessments rely heavily on subjective self-report and nuanced relational cues, with symptoms and treatment needs varying widely even within the same diagnosis. Some LLM developers attempt to build trust and openness through anthropomorphic design. Users of Therabot, for example, report openness similar to human therapy.[67] However, the key challenge remains an LLM's limited ability to interpret patient input accurately, especially in high-risk or vulnerable populations. Unlike experienced clinicians, who can recognize guarded or evasive responses and adjust treatment accordingly, current LLM systems struggle to discern inaccuracies or subtle cues in patient-reported data. While these limitations may not significantly impact individuals with mild-to-moderate symptoms, for those facing serious mental health challenges, inaccuracies in LLM-driven assessments or recommendations can lead to inadequate or even harmful interventions.[68]

Youth represent another high-risk group in AI mental health interventions. They frequently face barriers to traditional care that AI-driven tools may help reduce.[69] Two pre-LLM systematic reviews found that youth generally preferred digital mental health solutions, but clinical outcomes were mixed and inconclusive.[70] While some findings suggest promise, including a Replika survey that found that 3 percent of young users credited the tool with helping prevent suicidal thoughts, significant risks have also been identified.[71] Youth are particularly susceptible to misinformation and AI-generated hallucinations. One study found that 41 percent of teens struggled to distinguish real from fabricated medical content.[72] Other research indicates that younger users tend to overly trust LLM outputs and have difficulty recognizing inaccuracies or hallucinations.[73] Although peer-reviewed research on LLM-based mental health tools for youth remains scarce as of this writing, media reports have already highlighted cases of serious harm, including a youth suicide and a violent incident.[74]

Evidence gaps are substantial across all populations. Researchers have yet to determine how anthropomorphic design or conversational sophistication affects judgment in vulnerable users. Longitudinal studies have not produced data on how LLM use interacts with preexisting cognitive distortions or influences relapse, especially in SMIs. The mechanisms by which LLMs might amplify, exploit, or fail to detect mental health vulnerabilities remain poorly documented. Few studies disaggregate subjects by diagnosis, symptom profile, or duration of use, complicating efforts to determine who is helped, who is harmed, and under what conditions.

Unlike most physical health data, mental health information is deeply sensitive, subjective, and stigmatized. Many users who share biometric data from

wearables will not engage with mental health apps due to privacy concerns.[75] Mental health data are also disproportionately targeted for extortion.[76] In 2023, over 133 million health records were breached, including a major incident involving Cerebral, where 3.1 million users' self-assessments were improperly shared with advertisers.[77]

Policy responses are beginning to emerge. In March 2025, the American Psychological Association warned the U.S. Federal Trade Commission (FTC) and lawmakers of the harm posed by LLM chatbots mimicking therapists. They called for clear guardrails: public education, in-app safety features, crisis response protocols, and action against deceptive practices. Unlike human clinicians, LLM therapists are not mandated reporters of abuse or neglect. A *Journal of Pediatrics* commentary raised similar concerns about children forming attachments to LLMs without regard for development or caregiving context.[78] Another commentary outlined the need for accountability, equity, and transparency before deploying LLM tools for adolescents.[79]

Among the as-yet limited examples of legislation in this domain is California's State Bill 243, which would require AI chatbot platforms to warn users under the age of eighteen that they are chatting with an AI agent, restrict addictive features, and report incidents of youth suicidal ideation.[80] At the federal level, proposals like the Kids Off Social Media Act reflect growing pressure to regulate youth-facing digital tools.[81] However, no national restrictions have yet been placed on LLM mental health tools for minors or individuals with SMI. Regulatory oversight remains piecemeal, with no consistent standards for risk assessment, user protections, or evidence thresholds.

RESPONSES

HOLLY DUBOIS

While those with a serious mental illness are prone to periods in which judgment is compromised, to assume that anyone with a diagnosis of bipolar disorder, for example, has persistently reduced decision-making capabilities is a fallacy. This question is much more nuanced, and this historically underserved segment of the population demands broader access. More than half of the counties in the United States do not have access to a psychiatrist. Furthermore, outside urban and academic centers in particular, rigor and systematic vetting of health care interventions are lacking.

AI therefore represents an opportunity to recognize, treat, and engage those with a serious mental illness who may not otherwise receive or want to receive traditional mental health care. For example, individuals with severe panic disorder or even depression with psychotic symptoms may fear leaving their homes and may engage more effectively with virtual or generated elements. The concept of placing a human being, particularly a trained human being, in the loop for these interactions is indeed critical, but I'd argue that at intervals less than what may be perceived generally.

From 2020 to 2023, Mindstrong Health Services enrolled more than ten thousand patients in its virtual care delivery platform, embedded with smartphone sensing algorithms and text, phone, and video care provision. The qualifications to

enter the platform included having been diagnosed with an SMI or having been treated in an inpatient setting within the prior twelve months. Engagement among women and those in rural communities was particularly strong, along with those with medical comorbidities and functional disability.

The literature demonstrates that, among individuals with an SMI, AI tools can function as a feasible and important element of treatment within the continuum of care. As Jonathan Knights and colleagues demonstrate, rigorously applied AI models can enable a stepped-care delivery system for individuals with SMI, particularly severe depression.[82] Given our national scarcity of resources and limited number of licensed, trained providers, the model enabled therapists to increase or decrease the frequency of interactions based on predicted symptom severity. Incorporating measurement into care was therefore dependent upon AI, as traditional health care delivery has long suffered in this domain.[83] Given the inherently subjective nature of the psychiatric interview, combined with the aforementioned lack of measurement in mental health care, particularly within "talk therapy" settings, such tools are critical in developing clinical pathways. When AI tools can identify disruptions in sleep patterns, for example, providers and patients can be alerted to potential impacts on mood. Prolonged periods of disruption may indicate pending crises and facilitate escalation to the appropriate level of care.

Without the application of these forms of technology, treatment of the most serious mental illnesses may continue to lag in the archives of subjectively reported symptoms and long delays to care. Safeguards to quality and outcome measures, risk reporting, and continuous improvement are critical, but limiting access and experimentation in this domain could throttle desperately needed advancements, particularly in mental health care across the acuity spectrum.

ARTHUR KLEINMAN

Individuals with mental health problems who become psychotic are by definition unable to handle everyday reality; they struggle with hallucinations and delusions. Great care must be taken so that AI-driven interventions do not intensify these problems. In psychotic disorders and major depressive disorders there is the additional risk of suicide, and this has already been shown to be a potential consequence of the inappropriate use of a bot. But since psychosis is also associated with violence toward others, concern about the therapeutic alliance with a bot should be even greater. This is a place where AI must augment—not replace—human clinical providers.

The use of AI as a mental health tool should be restricted among individuals with psychotic disorders and among those for whom suicide is a possible outcome. Given the increase in suicide among youth and the elderly in our society, both of these groups should be regarded as vulnerable and involvement with bots should be either restricted or carefully controlled. Strategies that can be used to restrict access include clinical guidelines, best practices, institutional policies, professional association standards, and legal and ethical safeguards put in place by state advisory and licensing boards, by professional societies, and by health and mental health care institutions such as clinics, hospitals, and rehabilitation centers. No evidence yet suggests that AI can effectively identify crises affecting human providers or emergency services. This should be the work always of mental health professionals.

Question 5: How should these tools be deployed or limited in high-risk or vulnerable populations?

DANIEL BARRON

Should we limit access to AI mental health tools for certain groups, like kids or those with SMI? Again, it boils down to the *specific clinical job* the AI performs and the assessed risk of that tool failing at its job for that specific group (see Table 1 on page 24). A sweeping ban is clumsier than a task-specific, risk-based approach. Using AI for scheduling appointments appears generally low risk for most age groups.

For young people, Linda Alfano and colleagues highlight the critical importance of privacy, informed consent/assent, and significant human oversight when AI performs tasks within psychotherapy, casting AI as a therapist's helper for specific jobs, not a stand-in.[84] Seo Yi Chng and colleagues push for child-centered AI design, embedding safety-by-design for any job the AI takes on.[85]

For individuals with SMIs, whose judgment might be impaired, an AI tool designed for a high-stakes job like diagnostic assessment would almost certainly need heavy human supervision or might be deemed unsuitable for independent use altogether. Bo Wang and colleagues found that, while AI can streamline service delivery tasks for SMI, its inherent inability to offer genuine emotional support (a different kind of job) could worsen isolation, underscoring the need for human oversight for emotionally charged tasks.[86] On the other hand, an AI tool doing a low-risk job, like generic psychoeducation on sleep hygiene (notably available and relevant *after normal business hours*), could be an appropriate application of AI, akin to handing someone an interactive educational pamphlet.

How might restrictions work? Clinical gatekeeping seems plausible when a professional vets whether an AI tool is appropriate for a specific task for a given individual. Legal and ethical safeguards, as discussed by Mehrdad Rahsepar Meadi and colleagues regarding conversational AI, must be rooted in a risk assessment for the AI's defined job, especially given incidents like the Tessa chatbot providing harmful advice for its task.[87] In high-risk situations, AI's job could primarily be to assist human beings, such as flagging concerning data for clinician review or facilitating connection to emergency services if its task is crisis monitoring. Masab Mansoor and Kashif Ansari show AI *can* detect crisis signs (a specific job) but stress the need for ethical integration with human-led services.[88] This seems practical given that AI hasn't yet replicated critical human interventions such as the house call or safety check.

HANK GREELY

My views on the use of AI tools in high-risk or vulnerable populations are similar to my views on "human beings in the loop": the answer depends on what can be shown to be safe and effective in *those* populations. Who knows whether it will be good or bad? At this point we don't even know what "it" is. Careful monitoring will be essential to guiding the answers to questions about use in vulnerable populations but also about the roles of health care payers, the effects of access on disparities, and the economic and provider impacts of AI in mental health care. At this stage, we should focus on the hard task of creating procedures that will allow us to give informed answers to those questions in the future.

QUESTION 6: How should AI models be reimbursed or monetized?

BACKGROUND

Until recently, most AI-driven mental health tools lacked reimbursement through formal insurance channels, which significantly limited their adoption. At the center of this story is a regulatory divide between general wellness apps and therapeutic digital interventions, a distinction that drives fundamentally different monetization and coverage paths. A 2019 review described coverage as a fragmented patchwork, relying primarily on direct-to-consumer payments, employer wellness programs, or institutional licenses, with minimal involvement from insurers.[89] Standalone digital therapeutics had no viable reimbursement pathway under traditional fee-for-service models. This reimbursement gap was cited as a key barrier to broader adoption and likely contributed to the failure of several digital mental health firms in 2022 and 2023, despite their FDA-approved status.[90]

Over the past three years, however, public payers have begun to incorporate coverage for some digital tools. The 2025 Medicare Physician Fee Schedule introduced new Healthcare Common Procedure Coding System codes specifically for FDA-cleared "digital mental health treatment devices," allowing reimbursement for the clinician when tools are used under clinician supervision.[91] Medicare, whose codes set important precedents for private insurers and Medicaid, which typically follow Medicare's lead, does not currently cover the treatment devices themselves.[92] As of 2023, two state Medicaid programs, Massachusetts and Florida, explicitly covered "reSET," an FDA-authorized prescription digital therapeutic for substance-use disorder, designed to be used as adjunctive to in-person therapy.[93] A few other states are piloting coverage through Medicaid waivers or incorporating digital tools into broader behavioral telehealth initiatives.[94]

Private insurers and employer-sponsored health plans are gradually expanding coverage as well, typically focusing on regulated, FDA-cleared digital therapeutics. A 2022 survey found that 14 percent of U.S. health plan decision-makers reported covering digital therapeutics for behavioral health, marking the highest adoption area for digital tools outside diabetes care.[95] International models, such as Germany's Digital Health Applications system and the UK's National Health System (NHS) digital app assessments, demonstrate how structured reimbursement frameworks can significantly enhance integration of these technologies into health care systems.[96]

Payment strategies for AI mental health tools vary widely, depending on their intended audience, function, and regulatory status. Direct-to-consumer wellness apps primarily employ "freemium" or subscription models, while clinician-oriented tools like AI documentation assistants and clinical decision-support systems usually use enterprise licensing or per-user fees. Value-based contracts, in which payers fund platforms like Quartet Health and reward providers based on patient outcomes, represent another emerging model.[97] Lastly, several companies such as Lyra Health tap into the Employee Assistance Program (EAP).

The regulatory distinction between general wellness apps and therapeutic digital interventions drives fundamentally different financing approaches. Wellness-focused tools typically avoid regulatory scrutiny, opting instead for scalable revenue sources like subscriptions, corporate wellness deals, or direct consumer payments. In contrast, regulated digital therapeutics hope to position themselves as prescribable medical devices and pursue formal insurance reimbursement, which would result

in higher pricing but depends heavily on billing codes and consistent payer coverage.[98] This path is challenging. Pear Therapeutics, the company that produced reSET, the online substance use disorder digital therapeutic that succeeded in winning FDA approval and coverage through both Florida and Massachusetts Medicaid, ultimately went bankrupt.[99]

Several critical uncertainties remain. How swiftly and widely will providers and insurers adopt Medicare's new codes? Will AMA and Medicare expand their codes to include unsupervised AI symptom monitoring and care delivery? Will Medicaid programs nationwide scale up their digital therapeutic coverage? The effects of reimbursement structures on innovation are also uncertain: Parity in payment with traditional care might accelerate adoption but, without evidence of performance that is superior to freemium or subscription models, high cost-sharing and lack of provider interest might still limit adoption. Many AI mental health tools still enter the market without clearly defined pathways for reimbursement or integration into health care delivery systems, and few studies systematically assess how monetization strategies impact adoption rates, accessibility, patient outcomes, or the long-term viability of platforms. Some argue that a sustainable long-term model may require establishing something comparable to a formulary for prescription drugs, in which digital therapeutics are evaluated, approved, and reimbursed independently from clinician-provided services.[100]

The critical question is not just whether an AI tool is effective but whether it does a clearly defined clinical job better, safer, or more efficiently than alternatives, at a justifiable cost.

RESPONSES

DANIEL BARRON

Economic considerations are not peripheral—they are central to the adoption of AI in mental health care. As with all clinical innovations, insurers and reimbursement models will ultimately determine whether AI tools gain meaningful traction. The critical question is not just whether an AI tool is effective but whether it does a clearly defined clinical job better, safer, or more efficiently than alternatives, at a justifiable cost (see Table 1 on page 24 for details).

This may sound cold in the context of mental health care—where discussions of money are sometimes taboo—but we must reckon with the system we actually have. In the United States, health care is structured as a business. Clinicians are paid. Facilities have rent, utilities, and liability exposure. Medications cost dollars and cents. So do servers, engineers, APIs, and deployment cycles for AI tools. Any solution that ignores this economic scaffolding is unlikely to scale, no matter how well-intentioned. Furthermore, wishing that the system was "better reimbursed" or "more fair" is simply a fatal denial of clinical reality, as evidenced by multiple failed digital mental health start-ups.

AI tool developers must internalize this reality. Building a theoretically cool or technically impressive product is not enough. A viable tool must answer: What job does it do? How does it save money? And, critically, how does it help someone get paid? Product-market fit in U.S. health care

means fitting into workflows and into budgets. Without a reimbursement strategy, an AI tool is just a prototype, not a business.

One opportunity for AI tools is to help health care organizations and payers become more rigorous in understanding their own costs. While cost accounting is foundational to most industries, health care often functions with opaque, inconsistent pricing that makes it difficult to assess what anything actually costs—or what cost savings a technology offers. This lack of internal visibility stymies rational adoption decisions and useful business models or projections.

To make progress, we should adopt a job-based reimbursement model. Whether the AI supports medication adherence, delivers CBT modules, or flags suicide risk, payment should depend on evidence that it performs that specific task better or more efficiently than standard care, with acceptable risks. As Michael Abràmoff and colleagues argue, aligning financial incentives with ethically responsible AI creates a system that rewards value, not novelty.[101]

Operationally, the sector would benefit from a "digital formulary": a structured reference linking reimbursable AI tools to validated clinical jobs, akin to how drug formularies guide pharmaceutical access. This would enable payers to cover what works, avoid reimbursing what doesn't, and help vendors anchor their pricing to tangible, repeatable value.

Equity must also be intentionally built into these models. Reimbursement that overlooks digital literacy, language access, or device availability risks exacerbating disparities. (Though I do pause to consider that medications are reimbursed without consideration for whether a patient will *actually* take that medication or whether there are structural barriers to a medication's success.) As Masab Mansoor and Kashif Ansari show, reimbursement policy itself can be a lever for equity—as evidenced by improved youth mental health outcomes through well-designed telehealth funding.[102]

Finally, who pays should depend on the job being done. An administrative AI might be covered as operational overhead. A diagnostic AI might require fee-for-service reimbursement or inclusion in value-based care contracts. But public-sector funding—like the National Institutes of Health's Bridge2AI—should play a critical role in underwriting development for tools that address foundational infrastructure or serve high-need populations.

In sum: health care may be a human right in principle, but in practice in the United States it is a business. AI adoption in health care must therefore be a business proposition. Tool developers must think like businesspeople. And reimbursement should not simply facilitate adoption—it should shape it, steering AI toward clinically important, economically rational, and ethically sound use cases.

ARTHUR KLEINMAN

The role of insurers and reimbursement models creates a set of important questions. Barriers to access have already been identified as crucial to digital use during the high point of the COVID pandemic. In China especially, a large body of evidence demonstrates that elderly people are unable to effectively use QR codes and other digital tools.[103] The same situation exists in the United States. Coming up with simplified AI practices that can be easily explained to and used by older adults will be an important part of AI's future development. This is, in general, what is of real practical significance for the use of AI by the elderly throughout our society. Profit distorts care. For-profit hospitals perform more poorly

on clinically important indices than nonprofit hospitals. Insurance company and federal government attempts to control cost—as the pharmaceutical domain so sadly shows—have been inadequate to such a degree that cost is in a state of crisis. Given the reality of health care generally, controlling the cost of AI-informed mental health interventions is likely to become a serious and confusing issue. To prevent the worst abuses—including a future where interventions are available only for well-to-do Americans—organizations with the requisite legal and bureaucratic standing need to champion and ensure the performance of systematic, ongoing reviews.

RICHARD FRANK

AI has the potential to make productive contributions to an array of functions within the mental health delivery system. These include direct interactions (e.g., via Woebot) with people who have mental illnesses and emotional needs, expanding/enhancing the public mental health infrastructure (e.g., to identify risks to individuals and communities), improved efficiency in administration (i.e., back office), provider extenders, and improved provider/service quality.

Mental health care delivery poses some special opportunities and challenges to the application of AI. There are ongoing concerns about access to treatment (geographically, hours of availability, and by payer type), while at the same time a significant segment of people that use mental health services do not meet diagnostic criteria for a mental illness, nor do they report significant impairments.[104] Mental health care delivery is characterized by weak accountability processes.[105] The regulation and financing of AI applications in mental health care are inseparable. That is, issues of safety and effectiveness, along with data privacy, are essential for the establishment of reimbursable services. The literature to date suggests that the development of chatbots and other online direct treatments is running well ahead of policies and procedures that ensure safety and effectiveness. Thus, the discussion of pricing largely assumes integration with regulation of safety, effectiveness, and data privacy.

AI functions, mental health, and payment policy

Some simple economics. The market for AI services related to mental health care can be seen as having several distinct segments where the economic dynamics will differ. In one segment, the AI vendor will submit bills directly to an insurer, in which case the payer will negotiate a price with AI vendors. This might be coupled with an initial visit with a human clinician and evidence of safety and efficacy, such as FDA approval. The price will depend on the terms of coverage (cost sharing), the intensity of competition in the market, and the costs of developing and delivering the application. While thousands of AI therapeutic applications have already been created, how many will ultimately pass safety and efficacy scrutiny and be eligible for reimbursement by Medicare, Medicaid, and commercial insurers remains unknown. So, the effective level of competition is highly uncertain.

A second segment comprises AI applications that improve the efficiency of practices; for example, through better back-office functions that increase revenues, cut costs, and reduce no-shows. In these cases, AI vendors will likely sell directly to providers. For these types of functions, depending on the costs to the provider, the AI service will be included in the existing price paid to the provider for services

rendered. The presumption is that the savings to the practice will justify the AI investment and will be part of the service bundle.

A third segment involves AI services that do not result in savings to a practice but instead improve the quality of care. These might include continuous patient monitoring systems or diagnostic assistance applications. Such products could be especially beneficial in today's mental health delivery system, which is characterized by weak accountability due to quality-of-care metrics that are often crude, unevenly applied (if at all), and seldom tied to any consequences for a practice.

A fourth market segment for AI involves applications that enhance the public mental health infrastructure and/or have spillover effects and thus have a public good character. Examples might include the application of machine learning to efforts to predict suicide attempts following emergency department visits or to predict suicidal or violent behavior among callers to emergency hotlines (e.g., 988 or 911). Because these are tied to crisis response systems that at best rely on public financing of recurrent costs, they seldom have the technical capacity to implement such systems.

Payment models will vary by function

For direct-to-consumer therapeutic chatbot applications, given the uncertainty about safety and effectiveness apart from contact with a human clinician, a cautious point of departure would be to pay separately for chatbot sessions, initiated after contact with a human clinician, that have been approved as safe and effective.[106] Amid a market with limited competition, reliance on reference prices (set at a fraction of human clinician prices) would be a practical, efficient approach to budgeting. Public payers such as Medicare should be mindful of the social benefits of promoting innovation in the AI and mental health sector.

Services that improve the efficiency of practices and potentially reduce the costs of providing care are likely to be adopted out of economic self-interest. Prices and service arrangements could be negotiated between AI vendors and the provider practices without the direct involvement of third-party payers.

The regulation and financing of AI applications in mental health care are inseparable. That is, issues of safety and effectiveness, along with data privacy, are essential for the establishment of reimbursable services.

For a continuous patient monitoring application and other services that improve the quality of care but do not save money (absent a robust quality-adjusted payment system), an add-on fee could be charged for using the AI application. For services that will not enter an existing field of robust competition, pricing will require some cost finding. Alternatively, reimbursement systems that rely on quality performance to establish payment might set payment increments based on quality that in turn could take account of the costs of the AI technology.

Finally, building AI capacity for services that are publicly supported, have significant spillover effects, and are not tied to uniform, well-defined services might most effectively be paid for by public or philanthropic grants. That could be the case in using machine learning applications that make use of speech patterns to predict suicide attempts in the context of 988 calls or machine learning algorithms in emergency department settings that incorporate data from electronic health records.[107]

QUESTION 7: Can AI help address disparities in access to mental health care and the shortage of mental health providers?

BACKGROUND

Mental health care systems worldwide face persistent challenges, including severe provider shortages and unequal access to care.[108] Over half of U.S. counties lack psychiatric services, with wait times often exceeding a month for high-risk patients.[109] In Alabama, the provider-to-patient ratio is 1 to 1,200. Only 27.2 percent of estimated mental health needs are met nationwide.[110] These gaps are compounded by poor Medicaid coverage, a concentration of providers in urban areas, and the underrepresentation of clinicians from diverse backgrounds.[111] Marginalized communities face additional barriers, including geographic isolation, cost, stigma, and a lack of culturally informed care.[112]

Policy and technology have begun to shift this landscape. The 2008 Mental Health Parity and Addiction Equity Act limited insurer restrictions on coverage of mental health services, improving reimbursement options.[113] The COVID-19 pandemic further accelerated change by expanding telehealth and normalizing digital mental health tools, aided by the temporary suspension of state licensing requirements that allowed patients to access providers across state lines. Many of these flexibilities have since lapsed, reintroducing regulatory barriers that can limit scalability of telehealth services. Smartphone ownership in the United States now exceeds 97 percent, enabling broader access to remote care.[114] Telehealth has reduced transportation barriers and improved access to culturally matched providers, which correlates with higher engagement and satisfaction.[115] More recently, the growth of digital tools, ranging from mindfulness apps to medication reminders, has extended mental health support beyond traditional settings. Whether such tools primarily serve as gateways to human care or function as stand-alone substitutes remains unclear.[116]

From 2019 to 2021, consumer-facing digital mental health apps grew by over 50 percent, appealing to patients by lowering costs and increasing care options.[117] AI-powered tools now occupy a growing share of this space. Purpose-built technologies like chatbots, digital triage systems, and symptom monitors aim to address provider shortages and improve access.[118] LLM-based chatbots like Woebot and Wysa deliver automated interventions that are designed to reduce mild-to-moderate anxiety and depression in underserved groups.[119] Using natural-language processing, triage systems may be able to prioritize high-risk patients, shorten response times, and optimize clinician focus.[120] Algorithms deployed on social media platforms could detect acute distress, prompting earlier intervention.[121] Digital phenotyping tools can passively monitor indicators like mobility, speech, and sleep through smartphones to predict relapse and depression risk, supporting proactive care for patients with limited clinician contact.[122]

> If one assumes that body language, pauses in speech, intonation, and physical appearance account for more communication than words, chatbots and even advanced therapeutic bots still miss valuable communication.

However, these benefits come with potential risks. Models trained on narrow populations can misread culturally specific expressions of distress, delaying appropriate care.[123] Historical biases, such as the overdiagnosis of conduct disorders in youth of color, can be embedded in training data and perpetuated by algorithms.[124] Similar biases have been found in broader medical AI systems, such as underestimating care needs in low-income or nonwhite patients when using spending as a proxy for health and underprescribing medication by gender.[125] No uniform standards govern bias auditing or demographic reporting in mental health AI tools.

Equity concerns have also emerged. A recent systematic review highlights the potential risks of creating a two-tier mental health system in which disadvantaged groups disproportionately receive AI-only services, while more privileged groups maintain access to human care.[126] AI-driven tools could inadvertently displace in-person services in rural or low-resource settings, exacerbating existing inequities.[127] Additionally, older adults, low-income populations, and ethnic minorities tend to trust digital tools less and use them at lower rates. Forcing these populations to substitute digital tools for human therapy could be particularly harmful. Currently, no large-scale deployment of AI mental health tools has undergone systematic evaluation for equity outcomes, nor have federal or state agencies established frameworks to ensure equity measures are included in AI implementation.

Digital equity also shapes access. Not all populations have reliable Internet, up-to-date devices, or the digital literacy needed to use AI tools effectively. Engagement drops where infrastructure is lacking.[128] Pilot programs that pair AI systems with community training and simplified user interfaces show promise in underserved areas.[129] However, lack of systematic evaluation leaves uncertainty about which implementation strategies improve engagement, reduce disparities, or ensure sustained use across different populations.

RESPONSES

MARIAN CROAK

AI holds significant potential in addressing the persistent disparities in mental health access. Many of these disparities stem from factors such as enduring cultural taboos, financial barriers, and a critical shortage of therapists with both timely availability and cultural sensitivity toward underserved communities.

The use of generative AI for direct therapeutic interventions remains a complex area requiring more scientific research and regulation.[130] Additionally, if one assumes that body language, pauses in speech, intonation, and physical appearance account for more communication than words, chatbots and even advanced therapeutic bots still miss valuable communication.[131] More research is needed to understand whether these limitations are surmountable.

In the meantime, lower-risk AI tools should be evaluated to see whether they can provide more immediate help to expand the availability of therapists. Theoretically, these lower-risk tools can alleviate clinicians' administrative burdens, enabling them to see more clients, and can also support the training of new providers who are equipped to meet the needs of various communities.

First, to reduce their workloads, therapists are increasingly employing purpose-built and general AI tools for several relatively low-risk tasks. These

include automating scheduling and appointment communications, transcribing sessions and assisting with clinical documentation, and simplifying billing through insurance verification and claims processing. AI also streamlines the initial patient intake by automatically gathering necessary demographic and medical information, as well as the reasons for their visit. Additionally, chatbots are being used to efficiently address common patient inquiries (FAQs) about mental health services.

Despite addressing relatively low-risk administrative functions, the deployment of these AI tools requires a degree of human oversight to ensure accuracy and mitigate potential privacy and security vulnerabilities. As these technologies become increasingly integrated into the daily practice of therapists, empirical verification will be essential to confirm their purported benefit of reducing workload and thereby enhancing practitioner capacity to serve more clients.

Although more rigorous scientific evidence is still needed, another promising trend that has a high potential for rapidly scaling the training of new therapists is the use of collaborative AI–human interaction training tools. Examples of these approaches include:

- Creating learning tools to explain different therapeutic approaches and modalities as well as potential diagnosis and possible interventions.
- Dynamically adapting learning material to the personal learning needs and goals of new therapists.
- Using AI tools to track trainee progress, monitor improvements, refine supervisor feedback, and suggest specific areas where additional focus is needed.

More research is needed to understand the efficacy of these training practices and how they are used in practice.

Finally, to truly expand access to AI mental health services, greater cultural sensitivity and competency are essential. While clients tend to be more satisfied and engaged with culturally similar therapists, cultural competence and adaptability also can enhance the therapeutic relationship.[132] As AI models advance in cultural awareness and context sensitivity, they will play an increasingly significant role in training both new and seasoned therapists to effectively extend their practices to wide-ranging communities.

ALISON DARCY

According to the popular view, AI will reinforce bias and exacerbate existing disparities in access to quality mental health care. Bias in AI or in any service is indeed a key consideration, though it is neither new nor unique to AI. All science is subject to the "garbage in, garbage out" principle; that is, conclusions are only as good as the data from which they were derived, and in health care research, bad data can have real consequences for patients. In my early research at Stanford we discovered, for example, that diagnostic instruments used to detect incidents of eating disorders reflected attributes that male patients did not endorse (e.g., fear of weight gain among boys with anorexia) because they had been developed with predominantly female participants.[133] This has real consequences, leaving people undiagnosed, untreated, or given the wrong treatment. Thus, the issue of bias in data is not abstract, and many researchers and commentators have demonstrated incidents of real bias in AI today.[134]

However, it is important to note that implicit bias exists among human health care workers. While the risk of further embedding bias in AI is indisputable, such risk needs to be considered in the context of the institutional biases and significant disparities that are baked into today's system. AI *could* play a crucial role in addressing and mitigating disparities in mental health access if developed correctly and intentionally.

For example, throughout our research program, which, at the time of writing, includes eighteen randomized trials and twenty-five published, peer-reviewed studies in the scientific literature, we have employed recruitment practices to deliberately oversample for diversity in our participant group. This enables us to compare findings across demographic groups to ensure equitable effects—something that most studies are underpowered to do if they employ only the minimal acceptable recruitment practice of recruiting a sample that is demographically representative of the population.

This has resulted in some early findings that support the potential usefulness of purpose-built chatbots for mental health among typically underserved or underestimated communities. For example, in a cluster analysis study to examine latent usage patterns among users of Woebot, we identified a group of users who appeared to use the app in a way that gave rise to the steepest symptom reductions over the shortest period of time, relative to the other groups.[135] Those "efficient users" were more likely to be younger (average age=36), uninsured, non-Hispanic Black males. This group reported greater affinity to the chatbot (on the Bond subscale of the Working Alliance Inventory) and saw the largest symptom reductions in depression and stress despite using the app for shorter periods of time overall.

As an academic health researcher and clinician, I worry that we are developing our own biases and holding AI to different, higher standards than any other health intervention. Any AI should be evaluated by the same standards as other DMHIs (digital mental health interventions), as many of us argue in this publication. This does not mean we should not include the special considerations that AI presents, but rather that they must be grounded in the ultimate truth of symptom improvement.

While the risk of further embedding bias in AI is indisputable, such risk needs to be considered in the context of the institutional biases and significant disparities that are baked into today's system.

Many scientific and clinical leaders have studied the issue of embedding diversity, equity, inclusion, and belonging into the DMHIs that we are creating in health care settings, including our team. A robust strategy begins with an organizational commitment to embedding these values into the fabric of the organization such that it is then imprinted into the essence of the intervention or product itself. At Woebot we appointed an individual with a specific demonstrated skill set at filling recruitment pipelines with diverse candidates to run recruitment efforts and actively measured and rewarded success here. We never appointed an individual because they were a "diverse" candidate; rather, by ensuring that the pipeline was inclusive, we always hired the very best person for each role.

A later initiative established a clinical diversity advisory board that met every six weeks, created a full charter to produce thought leadership, and helped further the organizational vision to create an inclusive intervention experience that reflects the lived experience of all individuals who use our products, as well as the diverse perspectives of clinicians, corporate partners, and policymakers.

Methodological considerations in evidence generation

Existing frameworks like PIDAR (Partner, Identify, Demonstrate, Access, Report) and RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) provide guidance, emphasizing diverse partnerships, technology access, digital literacy, and data practices. Despite general support for DMHI efficacy, little is known about outcomes across diverse subgroups, and trials often lack diverse samples or focus on minority populations. Scholars have also noted that sociodemographic data are underreported in DMHI clinical trials. To genuinely assess DMHI impact on health equity, evidence-generation methodologies must be carefully designed, prioritizing the inclusion of individuals with lived experience and fostering diverse research teams to broaden the pool of future researchers.

Commercial deployments are where the "rubber hits the road" and real-world effects are observed, whether effects on individuals, families, communities, or even the bigger health care systems. This is where we see whether these technologies truly make a difference.

Recruiting diverse samples is crucial for evidence generation. Tactics include culturally sensitive materials, community outreach, partnerships with diverse health care systems, resources for limited Internet access, and diverse research teams. At Woebot Health we collaborated with the Scripps Translational Science Institute to successfully recruit a diverse sample, reaching our target of approximately 50 percent from historically underrepresented groups, including racial/ethnic minorities and rural residents. Real-world partnerships like this are vital, and clinician-referred recruitment should be managed carefully to avoid biases. "Opt-out" methods for study invitations from health systems or patient registries could increase diversity. Recruitment targets should be based on mental health problem prevalence rates in specific sociodemographic groups, and oversampling for specific groups may be necessary to achieve sufficient sample sizes for examining group differences.

Thoughtfully designed and consistently implemented sociodemographic surveys are essential. They should include culturally sensitive and inclusive response options for race, ethnicity, sexual orientation, gender identity, and social determinants of health like food and housing insecurity. Electronic responses and privacy assurances promote honest disclosure. Challenges exist in collecting comprehensive data outside research settings, requiring stakeholder agreement and user testing to consider assessment burden, privacy, and cultural appropriateness.

Analyzing and reporting sociodemographic characteristics and outcomes across subgroups are integral to understanding the impact of DMHIs on health equity. DMHI research should report sample sociodemographics, and consumer-based data should be presented with sociodemographic breakdowns. Ideally, outcomes are reported by sociodemographic groups or included as covariates. For small subgroups, we suggest exploratory, descriptive, and hypothesis-generating approaches (e.g., within-group and between-group effect sizes). Researchers could also analyze structural factors that contribute to inequities.

Finally, we advocate for assessing a multitude of secondary outcomes, aside from efficacy and safety, that include engagement, satisfaction, and feasibility, reflecting both researcher and participant priorities. Qualitative data are hugely valuable for understanding the metrics and can be used to triangulate diverse user experiences.

Question 7: Can AI help address disparities in access to mental health care and the shortage of mental health providers?

As many of us argue in this publication, we should not hold AI to a different standard than already exists for the systematic exploration of the usefulness of any other device or mental health intervention, and so here we draw from the significant and growing field of DMHI science to outline best practices. My team has written extensively on this and is published in the peer-reviewed literature.[136] A robust strategy for addressing health equity includes the following.

Equity-informed implementation science in real-world deployments

A vital consideration when evaluating mental health technology is how it is actually put into use. Commercial deployments are where the "rubber hits the road" and real-world effects are observed, whether effects on individuals, families, communities, or even the bigger health care systems. This is where we see whether these technologies truly make a difference.

Solid implementation science can help guide how well these mental health technologies are adopted, how much they cost (health economics and outcomes research), and what gets in the way of their adoption. Collecting these real-world data is crucial for understanding how effective these tools are and for making sure they're actually helping to close health equity gaps rather than making them worse.

Health care organizations that are buying or using DMHI tools should demand that the manufacturers show how their products are designed with health equity and responsible AI principles. They can use existing frameworks to guide them in asking questions about who these tools reach, whether they are specifically designed for vulnerable groups, and how they're monitoring outcomes to ensure fairness. Some organizations are even creating their own frameworks to ensure equity.

AI may be ideally positioned to collect data on social determinants of health because chatbots are already conversing with individuals in the naturalistic setting of their home. This information can help health care systems offer the right support and resources to those who need them most, potentially leading to earlier help and better-tailored interventions, benefiting underserved people, manufacturers, and health care systems.

Finally, we point to new roles that AI might be ideally positioned for, like "digital navigation" in which patients can be guided through their treatment options, improving their decision-making and potentially matching them with options that fit their lived experience, which can impact access and speed of support.

Real-world data from a deployment designed specifically to address health care disparities

We partnered with Virtua Health, a midsize health care provider in New Jersey that wanted to explore whether Woebot could play a role in addressing health equity in its large and unevenly distributed provider system. Early data suggested that the follow-up rate with Woebot was approximately four times greater than the referral completion rate among traditional behavioral health sources in the primary care setting; individuals had on average more than three times the number of encounters with Woebot than that of traditional referrals, with more than three-quarters of those encounters being outside clinic hours. Participants using Woebot saw a full category reduction in symptoms of depression (measured by PHQ-9) and anxiety (GAD-7) in eight weeks and completed routine patient reported outcome assessments 85 percent of the time. Anecdotally, physicians reported that social determinants of health domains emerged far more often in their conversations with Woebot than they had anticipated because, they said, they often do not have time to assess them in the context of a typical encounter. Our data scientists have shown that Woebot can identify social determinants of health domains at approximately a four-times-greater

frequency than the literature suggests is detected in health care settings using self-report instruments. In addition to supporting patients in identifying social determinants, connecting people to services is another example of how AI can help a whole-person approach and leverage data collection and reimbursement opportunities for the health system.

In conclusion, we should resist the temptation to think of patient-facing chatbots as replacing the role of therapists, because to do so misses the point entirely. We have an opportunity to deploy AI in ways that fill the many gaps that our fragmented health care system creates today, supporting better outcomes for everyone in ways that elevate the role of the human clinician and offer an invaluable window into the lived experience of patients. After all, mental health doesn't stop once we leave the clinic.

ARTHUR KLEINMAN

This question is potentially the most troubling regarding the use of AI-driven interventions in the mental health field. The presence of disparities and issues of efficiency will drive AI creators to argue for its adoption and claim that bots can replace human beings and are cheaper and more efficient. I feel this logic is basic to the use of AI in business settings and will be carried over to health care as health care is further privatized and its business interests prioritized. (Look at what private equity companies do when they acquire health care assets: strip them of things that can be monetized and sold off and emphasize efficiency to such a degree that quality care is worsened.) From my perspective, the real question is how to prevent the use of AI in mental health care from following this logic. If we begin with the central value of care and list its human qualities, then we will always conclude that AI interventions are appropriate when they augment, not replace, human caregivers. AI can be used today to contribute to clinical diagnoses, refine history taking, improve differential diagnosis, and add to the factors that go into making clinical judgment more useful, but AI cannot substitute for human clinical judgment and care.

Today, we can more usefully speak of culturally informed care than culturally competent care. AI can contribute significantly to culturally informed care delivered by mental health professionals. It can do this by providing information about cross-cultural differences, differing religious beliefs and practices, and patterns of behavior that may amplify or disguise symptoms. In the same way, AI can contribute to the care of underserved populations by improving understanding of the influence of structural factors like poverty on the course of disease and outcome. This is information based in what LLMs can provide that should identify many culturally relevant factors that clinicians can consider.

To assume that any policies can prevent AI from exacerbating existing health care inequalities would be extraordinarily naive. Health and social inequality run throughout our health care system in the United States and throughout systems all over the world. These are the same inequalities that policymakers in our country have been unable to control because their causes are so fundamental to the political economy of our society. The goal should be that AI does not make existing inequalities worse. That AI in and of itself might improve health and social inequalities seems highly unlikely, because it is not going to be creating structural transformation. Nonetheless, keeping this as a desirable goal for AI would be a positive step forward if it can be stated in practical terms with specific guidelines.[137]

Question 7: Can AI help address disparities in access to mental health care and the shortage of mental health providers?

KACIE KELLY

What are the most promising applications of purpose-built AI in expanding access to mental health services?

Severe mental health workforce shortages leave vast gaps in access to quality mental health services. Purpose-built AI, and digital mental health tools more broadly, can alleviate this burden by bolstering the existing mental health workforce and helping to make it more effective and efficient, expanding access to mental health services.

AI can optimize the existing workforce by improving the screening, diagnosing, and matching of patients to the right provider more quickly. Difficulty finding a mental health provider (e.g., due to insurance, specialty, geography, cultural fit) is a critical barrier. By connecting customers to the right provider sooner, AI can help deploy our overstretched mental health workforce more effectively while alleviating barriers to care.

AI can help stratify patients by risk, helping crisis response teams, health systems, and providers to reach more patients in need. By identifying individuals with lower-acuity needs, AI can also support stepped care models—where those individuals are directed to less-intensive services, such as peer support, nonspecialist providers, or even AI-powered therapeutics—freeing up clinical capacity for higher-acuity cases. AI can also help with clinical note-taking and electronic health record (EHR) documentation, which is well-documented to lead to burnout.[138] Finally, AI can enhance traditional forms of measurement-informed care with early risk detection, measurement of outcomes, and quality monitoring.

How can AI supplement the work of mental health professionals or substitute for them when access is unavailable or limited?

AI-enabled digital mental health tools may supplement care provided by mental health professionals. These tools can serve multiple roles across the care journey: as an adjunct to care, as a lead-up to care for someone who is waiting to be seen by a clinician, or as a follow-up to treatment.[139] This is especially important given the persistent workforce shortages and geographical inequities in mental health access. Many Americans live in areas with no access to mental health providers. More than 60 percent of people receiving mental health care do so from their primary care physician (PCP).[140] While PCPs may be able to offer screening and medications, patients who seek care in primary care offices receive little to no behavioral therapy or education outside integrated behavioral health care models.

AI can help fill this gap. High-quality, evidence-informed, culturally and linguistically competent AI-enabled digital therapeutics can expand access to quality mental health care to those who may not otherwise receive it. Likewise, AI-enabled digital therapeutics can optimize the mental health workforce, particularly primary care providers who are often providing mental health care. This offers important opportunities for improving health equity, potentially enabling more personalized, culturally relevant, and language-accessible care at scale for underserved populations.

AI therapy chatbots are still in their nascency but show potential for increasing access. While promising, these tools must be deployed with caution to ensure transparency about when users are interacting with AI, clear guardrails around scope and safety, and feedback loops that enable clinician

oversight when appropriate. Human connection remains essential, and AI should be designed to complement—not replace—the work of trained professionals where possible.[141] AI can also play important roles in risk stratification, measurement, and monitoring.

How can AI be integrated into existing health care systems to support, rather than replace, human providers?

AI is creating new opportunities to expand and improve traditional forms of measurement in mental health care, including measurement informed care (MIC), which leverages repeated, systematic use of validated measures to inform treatment decisions and monitor progress over time.[142] Traditional measurement in mental health is based on the use of self-reported assessment tools, such as the GAD-7 and PHQ-9. The subjective nature of such measurement tools may lead to both over- and underreported symptoms—or may miss important aspects of daily functioning that are of primary importance to patients.[143]

AI can enhance traditional MIC in several ways. Traditional assessments are designed to be brief to accommodate clinical workflows and represent snapshots in time.[144] AI enables a more expansive view that brings in "real-world" data from commonly owned digital devices, such as smartphones and wearables, to provide information about behavior, cognition, and mood. These data can be used in conjunction with traditional self-report assessments to detect changes in mental health conditions or predict relapse.[145] A recent systematic review found that physiological and behavioral data collected through digital phenotyping methods (e.g., mobility, location, phone use, call log, heart rate) can be used to detect and predict changes in symptoms of patients with mental health conditions, allowing for intervention even before an adverse event occurs.[146] Additionally, the increasing use of telemedicine for mental health visits has produced new sources of data, including videos from patient visits, audio recordings, eye movements, and so on.

Natural language processing (NLP) may be used more widely in the future of clinical mental health.[147] For instance, NLP technology is being tested to analyze data in electronic health records (EHRs), including clinical measures, clinician notes, comorbid conditions, and sociodemographic factors, to predict symptoms of severe mental illness and suicidal ideation and attempt.[148] For example, the Department of Veterans Affairs (VA) and scientists from the National Institute of Mental Health (NIMH) developed an expansive suicide mortality risk–prediction algorithm using Veterans Health Administration (VHA) electronic health records, enabling the VA to provide a more targeted, enhanced outreach and care program for veterans identified as being at high risk of suicide.[149] More recently, the VA added the use of NLP to tap into unstructured EHR data, such as clinical notes, to enhance the accuracy of this risk-prediction algorithm, resulting in an additional 19 percent accuracy. This demonstration showed that NLP-supplemented predictive models improve the benefits of the predictive model overall.[150]

While promising, more research is needed to better understand how digital phenotyping and NLP incorporated into MIC might affect patient engagement and treatment efficacy.

What challenges arise in ensuring AI tools provide culturally competent and context-sensitive care?

Though AI offers the potential to expand mental health care access and improve care quality, caution must be applied to ensure it does not exacerbate disparities. Culture, beliefs, identity, and life experiences all impact how we perceive and experience mental health conditions, as well as the treatments, coping mechanisms, and supports that are effective for each individual.[151] Additionally, these identities mediate how we communicate and express symptoms

of mental health needs, suggesting implications for what is considered an appropriate use of AI systems.

Despite the need for culturally relevant tailoring, research on NLP rarely discusses the role of demographic differences and language expression.[152] In one example, researchers found an association between depression and certain language features, including "I-usage." A recent study looking at the performance of language models used to detect signs of depression on social media posts found an association between "I-usage" and white individuals but not Black individuals.[153] Researchers then tested how race impacted the performance of language-based depression models. The model performed poorly with Black people, meaning AI applications designed to detect mental health conditions in social media posts, clinical notes, or other text could miss or misinterpret language patterns from different racial or cultural groups. If we embrace technologies such as these without ensuring safety and efficacy across populations, we risk exacerbating existing disparities.

This example is a cautionary tale about the need to include diverse groups of people when developing and testing AI applications. Rather than ignoring differences between different populations, a proactive recognition of difference can help overcome these gaps. This should include training and testing on a representative and diverse population and incorporate a variety of perspectives in the development of AI. For instance, researchers at the University of Texas at Austin and Cornell University are working to develop models that could identify social risk factors for Black youth with the goal of more accurately identifying suicide risk in young Black youth. Moreover, we must understand that the very statistical underpinnings of AI look for an optimal pattern or norm and thus may inherently create a bias against difference.[154] When not intentionally designed to include "edge" cases, AI is less effective for those existing outside the norm (e.g., someone who is neurodivergent or has a disability).[155]

How can AI tools help bridge the gap for underserved populations without exacerbating disparities?

Many of the improvements to mental health care delivery that AI may bring could also help bridge gaps for underserved populations; however, the potential for bias must be addressed to ensure the technology's success. We cannot assume AI will improve measurement, prediction, or care for all. Instead, we must intentionally design it to do so and monitor its impact.

The critical importance of training and testing AI systems on diverse and representative populations is well-known. The "intelligence" gained from AI must be understood within the limitations of the training data used and the potential for bias.

We must also address whether a specific AI application is safe and accurate for all people who might use it. Aggregate outcomes obscure the potential inaccuracy for specific groups.[156] Thus, outcomes must be disaggregated, especially for populations that experience mental health or other health disparities.[157] This is particularly true for health systems using AI.

One of the primary ways algorithmic bias is introduced is by mismatching what an algorithm is intended to predict versus what it is actually predicting.[158] For instance, a landmark study found that by conflating prediction of future health care needs with prediction of future health care costs, a widely deployed health care algorithm led to "enormous racial bias" impacting "important medical decisions for tens of millions of people every year."[159] The bias was created because using health care *costs* instead of *needs* overlooked barriers to access that cause some populations to be less likely to get the health care they need. Mental health applications are particularly vulnerable to using proxy measures because of the relative subjective nature of many mental health symptoms and diagnoses; the lack of true biometric data makes finding precise measures rather than

proxies particularly challenging in mental health care. The *Algorithmic Bias Playbook* offers a process to identify and mitigate "label" bias in algorithms used in systems (e.g., health systems, insurance companies). The process involves identifying all algorithms used in AI/predictive technologies, articulating their ideal and actual targets, and evaluating bias risk.[160]

AI should be employed specifically to detect and redress bias and health disparities. AI could even be used to reveal phenomena driving health disparities that human beings are unable to detect. For instance, NLP is an emerging approach to screen for and identify stigmatizing language in EHRs, automatically alerting clinicians and their supervisors.[161] One promising study showed AI was able to decipher previously unexplained disparities in knee pain across patients from underserved populations with diseases such as osteoarthritis. With the increased precision and ability for AI to see what human beings cannot, AI may "better capture [need and] potentially redress disparities in access to treatments."[162] As science on brain function and mental health continues to advance, AI may facilitate more precise diagnosis that more accurately screens and detects mental health issues across populations.

DANIEL BARRON

AI carries the potential to ameliorate mental health care access disparities and provider shortages, *if* we deploy it strategically to perform *specific, well-defined clinical jobs* that currently stretch our resources thin. Elizabeth Yong and colleagues highlight AI's attractiveness for enhancing accessibility, particularly for underserved communities, by taking on certain tasks.[163] Promising areas include AI tools automating administrative jobs (freeing up clinician time for tasks that they are uniquely qualified to perform), delivering scalable psychoeducation for defined conditions (a specific content delivery job), or supplementing diagnostic capabilities in underserved regions by assisting local providers with specific analytical tasks (see Table 1 on page 24 for further discussion).

We must also address whether a specific AI application is safe and accurate for all people who might use it.

AI can act as a supplement to mental health professionals by handling such defined tasks, or, in limited, validated cases, substituting for them when access for a specific job is otherwise nonexistent. Liana Spytska emphasizes AI as a complement for certain jobs, not a full substitute for tasks requiring deep human connection.[164] Just like any technology within health care, integrating AI into existing health care systems should focus on tools that support human providers within their current workflows for defined tasks.

Making sure AI tools deliver culturally competent care when doing their designated tasks demands thoughtful design and training on diverse datasets relevant to the job. (The same case is often made for medical student training.) Rinad Bakhti and colleagues found culturally relatable and coproduced DMHIs (a specific type of AI-driven job) showed higher engagement.[165] The FAITA-Mental Health framework includes "cultural sensitivity" for evaluating AI tools performing specific mental health tasks.[166] To genuinely bridge gaps for underserved populations, we must choose AI tools for tasks that are most needed and can be effectively handled by AI in those specific contexts, without inadvertently creating a "second-class" standard of care that deepens disparities.

QUESTION 8: What are the economic and other impacts of AI adoption in mental health on health care providers?

BACKGROUND

AI is already beginning to alter how mental health services are delivered, but its long-term effects on care systems remain unclear. Some predict that AI will "offer future advantages . . . in terms of accessibility, cost reduction, personalization, and work efficiency" for mental health care.[167] One widely agreed-upon area where shifting to AI could assist is in cost reduction.[168]

A 2022 systematic review of Internet- and mobile-based interventions (generally not AI-enabled) found that guided digital therapies are cost-effective for depression and anxiety treatment, with favorable cost per quality-adjusted life year.[169] Similarly, a randomized trial reported computer-assisted CBT to be cost-effective for treating depression.[170] These findings align with analyses from the National Health Service that project digitally enabled therapies could save thousands of clinician hours per every one thousand patients treated. Improved mental health outcomes from earlier or expanded interventions are also predicted to assist with reducing the $478 billion cost.[171] However, this figure has not been analyzed at the level of individual or specific interventions, and no current evidence links AI tools directly to reductions in productivity loss.

Real-world cost savings remain largely theoretical at this stage. Few health systems have systematically tracked whether AI adoption actually reduces emergency visits, hospitalization rates, or therapy dropout. Others caution that AI may introduce new costs in managing false positives or addressing ethical and legal issues.[172] In some cases, integration costs, including EHR compatibility, staff training, and data governance, may offset any short-term efficiency gains.

AI tools are also beginning to transform the workload of mental health professionals by automating routine and administrative tasks. A recent survey of psychiatrists identified "documenting/updating medical records and synthesizing information" as two tasks with potential for AI automation. Some experts argue that LLM chatbots and decision-support systems may be an "underappreciated solution to the shortage of therapists," able to deliver scalable support and improve work efficiency for existing staff.[173] Early reports in general mental health care settings indicate that digital scribes and automated scheduling systems have improved clinicians' sense of efficiency and reduced exhaustion from menial tasks.[174]

Surveys of mental health professionals reveal that many see the potential for AI to enhance care and improve efficiency but also harbor concerns about deskilling and the erosion of the therapeutic relationship.[175] These concerns include diminished empathic connection, reduced narrative exchange, and loss of clinical judgment in favor of algorithmic guidance, despite early studies showing that LLMs are comparable to therapists in clinical outcomes.[176] While the evidence to date shows that AI has caused few direct job losses in mental health care, the potential for AI to supplement the workforce, at a minimum, and raise the bar or disrupt the entrenched system of providers in the future seems inevitable for mental health care given its current trajectory.

Question 8: What are the economic and other impacts of AI adoption in mental health on health care providers?

RESPONSES

KACIE KELLY

How will AI adoption affect the cost and sustainability of mental health services?

Many AI applications could help optimize the limited mental health workforce and help them perform at the top of their license. For example, clinical note-taking is well-documented as contributing to provider burnout, and precious clinician time is often spent on administrative tasks rather than clinical care, reducing clinical productivity. As such, many health systems have embraced AI "scribes" to address these pervasive challenges. Doctors and mental health providers report AI tools enabling them to focus more on patients than on note-taking and data entry while cutting down on administrative burdens.[177] A recent taskforce led by the Peterson Health Technology Institute found that initial adoption of AI scribes has led to a reduction in cognitive load and burnout, though whether AI scribes actually increase provider capacity is still unclear.[178]

Other workforce efficiencies will come from more effectively getting people to the right care sooner. When thoughtfully designed for diverse populations, AI data analytics can help more effectively identify patient needs and match those needs to providers' expertise. Interventions work best when administered early in symptom onset and *before* symptoms reach a crisis point, ideally before the patient even knows they are ill.[179] As such, early identification and treatment are key for expanding access to care and treating conditions when they are more manageable and require less credentialed clinical intervention.

What potential disruptions to the mental health workforce could arise from AI integration?

The introduction of AI into mental health provider practice will inevitably disrupt, and cause friction within, the mental health workforce, even as it leads to efficiencies in many areas. At the most basic level, some providers—particularly those who have been practicing for many years—may be resistant to adopting new technologies, particularly AI. Nonetheless, all providers must be prepared to encounter patients who are using AI to support their mental health and well-being.[180] Practices that are more willing to embrace AI may have to make changes to workflows and necessary investments in AI infrastructure. Providers and practices will have to become more knowledgeable about data and cybersecurity, areas that may present significant learning curves.

The continual and rapid pace of AI advancement may present challenges to providers and health care systems, necessitating a rethink of what skills and training are needed in the workforce; many providers may need to be "upskilled."[181] The AI literacy gap will also likely change how mental health workforce training and education occur. The mental health workforce training and education process may need to focus more on computational skills, understanding of how AI works, knowledge around AI bias relevant to mental health, and more. Entirely new roles within the mental health workforce may also be needed. As measurement-informed mental health care becomes more commonplace due to advancements in AI, these new algorithms, data-driven approaches, and potential biomarkers may raise new questions. Just as genetic counseling was established as a new field in 1969, a new field may be needed to help patients and providers navigate

the increase in data and algorithms shaping mental health care.[182]

Likewise, significant questions remain about accountability if an AI system does not properly triage or handle signs of distress (e.g., suicidality). These questions will reshape the practice of mental health care, including ethics, liability, and accountability.

The continual and rapid pace of AI advancement may present challenges to providers and health care systems, necessitating a rethink of what skills and training are needed in the workforce.

Finally, given that the AI policy landscape is uncertain and constantly evolving, changes to AI regulation could affect the way AI integrations are rolled out in the mental health space, causing additional uncertainty.[183] The rapid advancement of AI itself may also cause disruptions to the mental health workforce. Updates to models and AI advancements are occurring rapidly, and many key technical experts argue that artificial general intelligence (AGI) is on track to occur in the next one to five years.[184] The mental health space may struggle to keep up with the latest models in AI and risk technology becoming outdated very quickly. The mental health workforce could be severely disrupted by not having an AGI strategy.

What measures can ensure equitable access to AI-based mental health care across socioeconomic groups?

Ensuring AI-mental health applications are deployed equitably requires both attention to equal access and protection against bias. A multitude of measures will be needed to ensure access to AI-enabled mental health care. For example:

- Many mental health applications are available only through private payment ("direct to consumer").[185] For AI-enabled applications to benefit all socioeconomic groups, both private and public insurance reimbursement are essential. Medicaid coverage is necessary for AI digital mental health technologies to support the nearly 40 percent of children and youth who are Medicaid beneficiaries.[186] At the same time, we must ensure that AI mental health interventions complement but do not replace human care; access to AI-enabled digital mental health interventions should not replace other efforts to improve access to mental health care.
- The use of AI-enabled care—ranging from clinical notetaking assistance to treatment—will require health system and provider investment in new technologies, data security, adjustments to workflow, and training of personnel.[187] Safety net systems such as local health authorities, community health centers, and rural health providers are at a disadvantage because they often lack the data and technological infrastructure to support many AI technologies. Significant public investment will be needed to ensure safety net systems are not left out of technological advances.[188]
- Digital literacy, the "varying ability of both children and adults to use technologies and understand their risks," is an important consideration when ensuring equitable access to and use of AI-enabled tools.[189] Uncertainty about how AI technologies work can lead to a lack of motivation or even resistance to using AI-enabled digital mental health tools.[190] Employing digital navigators and other nonlicensed roles to support patients' use of tech applications will be critical to facilitate their use in some populations, such as older adults or people with severe mental illness.

Question 8: What are the economic and other impacts of AI adoption in mental health on health care providers?

ARTHUR KLEINMAN

If we look at the impact of telepsychiatry and telephone- and video-based psychotherapy, I do not believe we yet have convincing evidence that these approaches to care reduce the cost of providing care, although there is evidence that they can provide care of equal quality to that given by human beings in mental health care.[191] The evidence-based argument for AI's adoption should be carefully assembled and thought through so that it avoids immodest claims of causality. We require economic analyses of augmenting practices to make the case for AI's adoption; not practices that substitute bots for people. If AI substitutes bots for people, we will be well on our way

The evidence-based argument for AI's adoption should be carefully assembled and thought through so that it avoids immodest claims of causality.

to undermining the sovereignty of human beings in health care and replacing it with the sovereignty of AI. But we have no evidence that doing so would improve care and there are multiple arguments for why it would create havoc in our health care system, which is already chaotic, disorganized, and broken. No measures can ensure equitable access to AI-based mental health care for different socioeconomic groups, because such inequality is a fundamental reality of our health care system.

DANIEL BARRON

The economic and workforce impacts of AI in mental health care will hinge less on AI as a general concept and more on which specific clinical and administrative *jobs* AI tools perform—and how well they do them (see Table 1 on page 24). Automating narrow, clearly defined tasks (e.g., billing, documentation, and symptom tracking) may reduce costs and increase throughput. But developing and deploying AI for complex, high-touch clinical jobs demands hefty investments. Justus Wolff and colleagues remind us that many economic analyses of AI overlook its full costs—development, implementation, integration, and maintenance—thus obscuring the true value proposition.[192]

Potential disruptions to the mental health workforce are more likely to look like a reshuffling of roles based on the tasks AI absorbs, rather than outright replacement. If AI takes over routine data-gathering or -summarization tasks, clinicians might find themselves focusing more on complex decision-making and the uniquely human, interpersonal aspects of care—jobs that seem less suitable for today's AI. Jo-An Occhipinti and colleagues suggest that AI can augment the workforce by helping with diagnostic and administrative tasks, thereby freeing up health workers for direct patient care jobs.[193]

AI could reshape health care policies by paving the way for new care delivery models centered on specific tasks, which in turn would necessitate policies to ensure these AI-driven jobs are done safely and equitably. The division of labor could blend digital

and human workflows in ways current reimbursement and licensure policies are not yet structured to support. Policymakers will need to define the boundaries and standards of these AI-executed jobs: What gets reimbursed? Who holds medicolegal liability? What constitutes "standard of care" when care is shared with a machine?

Equity must be deliberately engineered into this future—not to support any specific ideology or theory, but for the simple fact that supporting the health of the *entire* population is cheaper and more cost effective for the health system as a whole. Widespread AI adoption risks entrenching disparities unless systems invest in access and usability. Without the early investment in reaching all people in our society, the short-term gains of any AI-based tool/system risk being buried in the long-term accumulation of (ultimately costly) chronic diseases and comorbidities that might have been avoided with proper planning. That includes subsidizing AI tools that perform high-value tasks in under-resourced settings, funding digital literacy programs (many patients still struggle with basic portals), and co-designing tools around the needs of marginalized communities. Task performance must be validated not just in clinical trials, but across real-world populations representative of all the lives covered by a payer, including the largest payer in the United States: the federal government.

Ultimately, AI's impact won't be defined by the technology itself but by the labor, payment, and policy ecosystems it reconfigures. The question is not "Will AI disrupt?" but rather "Which jobs? For whom? At what cost? And who will benefit?"

Without the early investment in reaching all people in our society, the short-term gains of any AI-based tool/system risk being buried in the long-term accumulation of (ultimately costly) chronic diseases and comorbidities that might have been avoided with proper planning.

QUESTION 9: What are the most significant scholarly questions that will need to be answered as AI's role in mental health care evolves?

BACKGROUND

As stated earlier, potential interventions to be used in biomedical and behavioral care settings are usually subjected to rigorous research and scholarly analysis before they are allowed to be deployed. However, in the case of AI and mental health care, that has not happened; many AI applications have been deployed before that kind of rigorous research has been done. That is unfortunate and may have led to some significant problems, including ineffective and even potentially dangerous applications being freely available. For those reasons, a full-scale research program is badly needed, even if to some it may soon be too late. Some problems will be correctable. A greater number of randomized clinical trials can provide the evidence base from which to build and evolve AI into safe and effective tools. Population health studies can help ensure appropriate guidelines for AI in mental health care and equitable access to these tools. The findings from such research enable more strategic funding allocation, helping policymakers and private funders invest in novel care pathways and tools with the greatest potential for clinical value and equitable impact. Evidence standards and benchmark requirements can create productive pressure among developers, rewarding systems that demonstrate real-world utility rather than marketability alone. Without such feedback loops, hype may outpace performance, and useful innovation may be crowded out.

RESPONSES

MARIAN CROAK

Recently, researchers conducted the first clinical trial of a purpose-built therapy bot.[194] They used custom-built datasets containing evidence-based practices with a human in the loop to monitor the bot's responses. Although the study was small (210 participants), the therapy bot showed promising results, especially for alleviating depression. Given the widespread use of generative AI for therapeutic purposes, the fact that this is the first and only clinical trial highlights the need for more rigorous scientific research.

Areas that need deeper exploration include:

- examining both short- and long-term outcomes for stand-alone therapeutic bots and ones with a human in the loop compared to one-on-one human therapy;
- creating best practices for the use of datasets that mitigate hallucinations and algorithmic bias and that contain representative data from a wide range of demographic groups;
- designing large, transparent, and explainable datasets that capture different modalities and psychological conditions;

- investigating the ability of AI systems to learn emotional intelligence and empathy;
- identifying the risks and benefits of anthropomorphic attachment to AI therapeutic bots;
- improving the reliability and accuracy of diagnostic determination;
- creating filters and evaluation tools for minimizing harmful responses/advice;
- measuring the effectiveness of using AI assistants to help reduce the time therapists spend on administrative tasks; and
- examining workforce disruptions produced by AI mental health tools.

To truly understand the effectiveness of using AI in the end-to-end therapeutic process, we must engage the expertise of social scientists, ethicists, security and privacy specialists, legal professionals, neuroscientists, demographically diverse clients, mental health professionals (psychologists, psychiatrists, and social workers), user experience researchers and designers, computer scientists specializing in natural language processing, AI, and machine learning, as well as others. This interdisciplinary collaboration will help to ensure that proposed solutions are technically sound, ethical, safe, easy to use, and truly beneficial for clients and practitioners.

As these AI tools are deployed, they will need continuous monitoring and auditing to ensure they are performing as intended. AI models, especially generative ones, dynamically change as a result of statistical factors and their adaptation to real-world input. Depending on the application of AI, different metrics will need to be established that target benchmarks set across different parameters. Subsequently, evaluation tests or audits need to be periodically or continuously conducted to ensure the tool is within range of its target benchmarks. If measurements dip below the benchmark, either automated or manual adjustments to the tool are needed. Relevant metrics include reliability, availability, accuracy, appropriateness of the response, user sentiment and engagement, fairness, and measures of privacy and security.

As tech companies race to deploy new advances in AI, they are clearly outpacing the slower progress of governance and policy frameworks created by the public sector, including international and domestic government entities. Research suggests that policymakers should adopt more-innovative approaches to creating governance frameworks and regulatory guidelines.[195] The National Institute of Standards and Technology's approach to setting guidelines for responsible AI governance is an example of an innovative policy due to its extensive reliance on industry collaboration and use of version controls to enable fast changes to policies as technology advances.

ALISON DARCY

Over the past few decades, numerous technologies have generated enthusiasm as potential solutions to long-standing challenges in mental health care. The Internet was meant to resolve access; smartphones promised always-on support and Big Data insights; gamification aimed to improve adherence. But AI is arguably the advance that most fully unlocks these benefits—by offering a human-centered interface: conversation.

AI doesn't just scale interventions; it could reshape them. For the first time, we can reach people in the flow of daily life, through an interface they can engage with even when motivation and mood are low. This marks an exciting moment for intervention science but also raises important questions.

Question 9: What are the most significant scholarly questions that will need to be answered as AI's role in mental health care evolves?

We may be tempted to tightly control or replicate human-led models. Yet doing so may limit discovery. Rather than copying human beings, we should ask: What are AIs uniquely good at? Human beings are essential for connection, intuition, and relational depth. AIs excel in consistency, availability, and data processing. We do not need to force them into a mold that doesn't suit their strengths.

I believe the question of whether an "AI can make a good therapist" is unhelpfully blunt and tends to get bogged down in an impractical narrative of AIs replacing human beings. Early studies challenge the belief that AIs can't be empathic or helpful. For instance, people often disclose more to AIs than to human beings.[196] Some people report feeling more empathized with by AI than by physicians in online contexts.[197] And AI (natural language processing) algorithms may be better suited to prediction and detection of psychosis than human beings.[198]

Going forward, we need to adopt first-principles thinking to assess what works—not based on fidelity to human therapy but on what produces benefit. Most critical, the only human who must remain "in the loop" is the user themself. Research on Edward Deci and Richard Ryan's self-determination theory tells us that autonomy is a key driver of positive outcomes.[199]

We must, of course, evaluate efficacy and safety rigorously—but also stay open to unexpected benefits, not just unintended harms.

That being said, what are the major areas of scholarly focus?

Therapeutic mechanism

All major technology shifts generate new behaviors. Just as smartphones sparked the cultural norm of photographing meals, AI may give rise to new forms of therapeutic engagement; that is, enable the creation of novel therapeutic mechanisms. It could also amplify known mechanisms. For example, CBT as a therapeutic approach invites language to be formally examined as a proxy for thinking, and thoughts are systematically gathered and considered as a window into distorted beliefs. CBT also leans heavily on data, relying on real-time gathering of symptoms like moods and the context in which they arise, for assessment and evaluation. Not surprisingly, all purpose-built chatbots to date—for example, Woebot, Wysa, and Therabot—have been built upon a CBT framework wherein the role of the therapist is characterized by "collaborative empiricism." How an AI should show up in a different type of therapy wherein the role of the therapist is characterized differently is not yet well understood. For example, can AI be deployed in a family systems therapeutic approach? If so, how might that be operationalized? Research should also explore whether AIs can facilitate therapeutic methods that are based on creative self-expression, since this has been identified as an active ingredient of effective intervention for young people and appears to be a particular strength of emerging AI models.[200] Determining this seems even more relevant given the recent paper in *The Lancet Psychiatry* that examined the outcomes from NHS's Talking Therapies for anxiety and depression in over three hundred thousand patients. The results suggest that young adults' outcomes were poorer than those of working adults. Young adults were less likely to meet measures for reliable improvement and were more likely to meet criteria for reliable deterioration. These data point to the youth mental health needs that require adaptations, which AI-supported mental health care is well-positioned to meet.[201]

Efficacy and safety of AI-based interventions

The literature base around DMHIs, particularly digital CBT, is already mature. Increasingly, RCTs have tested AI-based chatbots—both traditional rule-based and GenAI-driven—utilizing a variety of comparison conditions. Two studies have compared chatbot-based interventions to more traditional human-delivered care, with both studies showing similar findings (noninferiority for

depression and slightly superior findings in the case of anxiety).[202] Future research on efficacy and safety should explore the longer-term effects of chatbot-based interventions; for example, whether therapeutic effects diminish over time and what usage patterns—different as they are structurally from classic treatment—might be deployed to avoid this. However, the most urgent question is how to deploy these interventions within the health care system.

AI-enabled therapeutic systems and structures

AI has the potential to enable entirely new care structures—stepped care, precision care, continuous engagement models, and so on. If AI is to have a specific role in public health, how do we consider the role of frontier models in supporting positive mental health and even offering early intervention. What we don't know about health care models is how the interplay between data ownership and crisis intervention would work (or whether it should even be part of AI-enabled health care).We don't yet fully understand or have an agreed-upon definition of the legal and ethical frameworks around data ownership, liability, and trust, and how we might avoid system-level harms like deskilling of clinicians.

Future studies could ask:

- How can AI help us triage more effectively?
- Can AI personalize care pathways based on lived data?
- How does AI shift the role of the clinician over time?

This is not a call for unchecked optimism but to think more creatively and rigorously about how we evaluate and frame the role of AI in mental health, including by drawing from the considerable multidisciplinary expertise that may contribute to this endeavor. The opportunity is not to replace human beings but to better support them and to serve the many people the system currently leaves out.

ARTHUR KLEINMAN

AI must be assessed in the everyday conversations between clinicians, patients, and families. This is where its contribution in improving communication and relationships needs to be demonstrated. Interdisciplinary collaboration is essential to this kind of work. Pairing AI experts with clinicians and social scientists is the way to organize appropriate use and evaluation of all technological interventions in health care.[203]

The continuous evaluation and refinement of AI tools is not only a matter of safeguards but of building AI systems that iteratively examine the evidence about interventions in order to constantly provide feedback for the improvement of practices. Here AI can build on engineering systems approaches that utilize these kinds of feedback loops. Input from software engineers working with clinicians would be helpful in developing best practices.

The empirical reality is that the United States does not have a single mental health care system but rather a chaotic and perhaps ununifiable collection of private and public systems. For severe chronic mental illness, principally schizophrenia, the public mental health system is so broken that the criminal justice system is now regarded by experts as the functioning mental health care system for most of these patients. How AI-driven interventions will figure here is going to be based on research, and this is one of the areas in which research that examines practices that augment professional caregivers should be very useful. Looking at the history of mental health practices, the private sector is likely

to produce both useful examples and many examples of inappropriate use and abuse. Regulations are where we will have to work out the many questions about how the public and private sectors will relate to responsible AI applications. Given the absence of a unified system of mental health care, private and public sectors will most likely remain isolated from each other and fail in effective collaboration. I seriously doubt the utility of attempting with AI what has not happened with any other mental health intervention. Then again, perhaps this is an area where AI *can* help, albeit in a different way: by generating the knowledge to systematize and possibly even integrate mental health care practices and practitioners who, for far too long, have failed to collaborate.[204]

ROBERT LEVENSON

What research priorities should guide the development of AI mental health tools?

Much as with the development of a new drug, the holy trinity of research progresses from establishing safety, to evaluating efficacy and effectiveness, and finally to understanding mechanisms of action. Formal safety trials should be initiated as soon as possible for some of the more common kinds of AIMHIs, many of which already exist in the wild, challenging efforts to obtain more than anecdotal safety data about them. Such trials should be designed to generate safety data that can be compared with more conventional human therapist approaches (and perhaps combined human-bot approaches as well) when dealing with similar problems and populations. Important issues to be tracked in these trials include clients'/patients' thoughts and acts related to harm of self and others as well as measures of mental health symptoms and well-being. In the efficacy stage, RCTs that compare AIMHIs with one another and with more conventional active treatments will be important. Psychotherapy research has often revealed that everything works better than nothing and that differences among therapies are nonexistent or small (the so-called dodo bird verdict).[205] But we cannot assume that this will be the case for AIMHIs without well-designed research. Finally, existing therapy research has consistently revealed that the most potent mechanisms of action for a wide range of psychotherapies are the common, nonspecific ones (e.g., expectations, attention, placebo, alliance, time). AIMHIs bring characteristics and abilities to the table that are unique (e.g., lack of distraction and fatigue, near instantaneous access to huge bodies of knowledge, full recall of the details of past therapy sessions and client histories) and thus may have different mechanisms of action than their human counterparts.

DANIEL BARRON

Continuous, interdisciplinary research is essential to ensure that AI in mental health care evolves responsibly, effectively, and equitably. The most pressing scholarly questions must center not on AI in the abstract but on the specific clinical and administrative tasks it is meant to perform—and whether it does those jobs better than existing methods. (See Table 1 on page 24 for an example of task breakdown.)

Question 9: What are the most significant scholarly questions that will need to be answered as AI's role in mental health care evolves?

First, we need rigorous, task-level validation: In which clinical jobs can AI actually improve accuracy, safety, cost, or access? Research should focus on job-specific performance rather than generalized hype. Pablo Cruz-Gonzalez and colleagues demonstrate AI's promise in certain diagnostic and intervention tasks but emphasize the need for more diverse datasets and model transparency, especially when algorithms are deployed across different patient populations.[206]

Research and development are essential to our understanding not just of the safety and effectiveness of these methods but even of what methods are being used and how.

Second, we need more-robust frameworks for real-world evaluation. Validating AI tools in the lab is not enough. We must also study how they behave in complex, variable, and often unpredictable clinical environments. Hassan Auf and colleagues underscore the lack of empirical research on human-AI interaction in real-world settings, particularly when AI is used for decision support.[207] Participatory design—engaging clinicians, patients, ethicists, and human-computer interaction specialists—is not optional; it is essential to building trust and ensuring tools are fit for purpose.

Third, we need longitudinal, postdeployment surveillance. Once an AI tool is "in the wild," continuously assessing its performance, drift, and unintended consequences is critical. The FAITA-Mental Health framework offers one approach, linking performance back to the task-level job the AI is meant to do.[208] Gauthier Chassang and colleagues call for provider-led postmarket surveillance models that pair clinical insight with user experience, a vision that could mirror pharmacovigilance for digital therapeutics.[209]

Finally, we must investigate how scholarly insight can shape standards and funding. Public-private partnerships could establish evaluation protocols that tie reimbursement or regulatory approval to task-level evidence. This would incentivize developers to prioritize clinically meaningful tools over flashy demos. It might also—perhaps unintuitively—help clinicians better understand the operational logic of their own systems by forcing a clear delineation of who does what and why.

In sum, the most urgent scholarly agenda for AI in mental health is not technological; it is functional. The field must interrogate which jobs are *actually performed today,* which jobs an AI should do, how well it does them, and what it costs when it fails.

HANK GREELY

Research and development are essential to our understanding not just of the safety and effectiveness of these methods but even of what methods are being used and how. Some publicly accessible, and preferably published, peer-reviewed research is necessary before new approaches are tried. However, given the speed of the introduction and modification of these approaches, I fear getting sufficient evidence of safety and efficacy before clinical use will be impossible. That makes it all the more important that these methods include rigorous monitoring and assessment procedures, ideally independent from corporate sponsors.

Afterword

Paul Dagum, Sherry Glied, and Alan Leshner, project cochairs

Over the eighteen months since this project's inception, the integration of LLM applications into everyday use has progressed at an astonishing pace, leaving numerous unresolved societal and ethical questions. What is clear is that artificial intelligence, as it evolves, will alter many dimensions of our individual and social lives—in large ways and small, for better and worse. Our contributors have laid out a set of considerations for exploring these effects in the context of mental health treatment. Their responses, and our engaged discussions throughout the project, lead us to some overarching reflections about directions forward.

1. The subject of our work, the use of artificial intelligence in mental health treatment, turns out to comprise a wide range of technologies and situations. One important step in solidifying research both on the immediate effects of technologies on individuals and on their broader effect on societies and relationships is to develop a set of definitions and taxonomies. Classification is often the first step in science and policy, and its absence makes discussion challenging.
2. One broad grouping within the area of artificial intelligence in mental health treatment consists of applications explicitly designed to play a role in the delivery of such treatment (this might include administrative or monitoring apps, apps used by or in conjunction with professionals, and so on). Developers of these applications will usually conduct some formal assessment of their effectiveness in the specific context for which they are intended. Unlike the case for pharmaceuticals, where a well-established playbook guides FDA approval, these assessments use disparate methodologies and comparison groups (for example, pre-post, waitlist controls; comparisons to psychiatrists delivering individualized psychotherapy or counselors delivering manualized CBT). Developing a minimum set of requirements, including ethical requirements for the conduct of studies in this context, is an essential step to moving forward and is particularly important given the wide variation in the actual quality (and availability) of mental health treatment as it is currently delivered in the United States. It will also require a new approach to the design of FDA approval, which currently assumes that each new formulation of a drug or revision of a device is frozen in design and subjected to new evidence from clinical trials. The regulation of systems designed to evolve and improve with every interaction has no established precedent.
3. The rapid uptake of chatbots and LLMs is happening without research or regulation. But this wide and fast diffusion is not inevitable. As the many recent changes around the use of smartphones in schools suggest, public opinion, persuasive critiques, and policies can affect the pace and nature of adoption. Those decisions—individual and societal—should be informed by as much evidence as can be brought to bear. Currently, that evidence largely takes the form of anecdotes, which are powerful but can be misleading. Opportunities exist to systematically collect information on the use of chatbots for

therapeutic purposes (for example, in existing national surveys) and to conduct analyses of these effects (for example, comparing populations who, for exogenous reasons, have had easier and more limited access to LLMs, as has been done for smartphones and television). In doing so, we will also need to think carefully and comprehensively about the nature of the effects we might see. Experts in philosophy, psychology, and sociology can trace the pathways through which negative or positive effects may emerge. Systematic qualitative research can point to outcomes that are unexpected, whether promising or troubling.

4. LLMs build on the information they collect. This feature generates network externalities (where a service becomes more valuable as more users adopt it), and network externalities in turn tend to drive market consolidation. Without up-front regulation, a dominant mental health care LLM, on which millions of Americans might rely and that would have access to deeply personal information about them, could emerge (as we have seen with other Internet technologies), posing profound economic and ethical implications beyond those inherent in the technology itself.

5. When a wide range of disciplines and perspectives have seats at the table, the outcome can be tremendously generative and occasionally, and appropriately, uncomfortable. If only computer scientists and mental health providers participate in future discussions of the effects of AI on mental health, a great deal of useful insight will be lost. Such future discussions are essential as LLMs evolve alongside our understanding of their implications for individuals and society.

Endnotes

1. We recognize that AI has many possible uses adjacent to the delivery of mental health care, including as scribes or alert systems. Such uses are not a focus of this report.

2. Peter M. Yellowlees, Alberto Odor, Ana-Maria Iosif, et al., "Transcultural Psychiatry Made Simple—Asynchronous Telepsychiatry as an Approach to Providing Culturally Relevant Care," *Telemedicine and e-Health* 19 (4) (2013): 259–264, https://doi.org/10.1089/tmj.2012.0077; and Shalini Lal and Carol E. Adair, "E-Mental Health: A Rapid Review of the Literature," *Psychiatric Services* 65 (1) (2014): 24–32, https://doi.org/10.1176/appi.ps.201300009.

3. Alec Tyson, Giancarlo Pasquini, Alison Spencer, and Cary Funk, "60% of Americans Would Be Uncomfortable with Provider Relying on AI in Their Own Health Care," Pew Research Center, February 22, 2023, https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care.

4. Jamie Bernardi, "Friends for Sale: The Rise and Risks of AI Companions," Ada Lovelace Institute blog, January 23, 2025, https://www.adalovelaceinstitute.org/blog/ai-companions.

5. David Simpson, "Digital Literacy Among Tennessee's Older Adults," Sycamore Institute, July 30, 2024, https://sycamoretn.org/digital-literacy-older-adults.

6. Wenjun Zhong, Jianghua Luo, and Hong Zhang, "The Therapeutic Effectiveness of Artificial Intelligence-Based Chatbots in Alleviation of Depressive and Anxiety Symptoms in Short-Course Treatments: A Systematic Review and Meta-Analysis," *Journal of Affective Disorders* 356 (2024): 459–469, https://doi.org/10.1016/j.jad.2024.04.057.

7. Madison Milne-Ives, Emma Selby, Becky Inkster, Ching Lam, and Edward Meinert, "Artificial Intelligence and Machine Learning in Mobile Apps for Mental Health: A Scoping Review," *PLOS Digital Health* 1 (8) (2022): e0000079, https://doi.org/10.1371/journal.pdig.0000079.

8. Eugenie Park and Darrell M. West, "Why Mental Health Apps Need to Take Privacy More Seriously," Brookings Institution, November 30, 2023, https://www.brookings.edu/articles/why-mental-health-apps-need-to-take-privacy-more-seriously.

9. Patricia Hong and Ezekiel J. Emanuel, "Leveraging Artificial Intelligence to Bridge the Mental Health Workforce Gap and Transform Care," *Milbank Quarterly*, February 4, 2025, https://doi.org/10.1599/mqop.2025.0204; and Clese Erikson, Ellen Schenk, Sara Westergaard, and Edward S. Salsberg, "New Behavioral Health Workforce Database Paints a Stark Picture," *Health Affairs Forefront*, August 30, 2022, https://www.healthaffairs.org/content/forefront/new-behavioral-health-workforce-database-paints-stark-picture.

10. Anita Schick, Jasper Feine, Stefan Morana, Alexander Maedche, and Ulrich Reininghaus, "Validity of Chatbot Use for Mental Health Assessment: Experimental Study," *JMIR mHealth and uHealth* 10 (10) (2022): e28082, https://doi.org/10.2196/28082; and Mirko Casu, Sergio Triscari, Sebastiano Battiato, Luca Guarnera, and Pasquale Caponnetto, "AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications," *Applied Sciences* 14 (13) (2024): 5889, https://doi.org/10.3390/app14135889.

11. Hao Liu, Huaming Peng, Xingyu Song, Chenzi Xu, and Meng Zhang, "Using AI Chatbots to Provide Self-Help Depression Interventions for University Students: A Randomized Trial of Effectiveness," *Internet Interventions* 27 (2022): 100495, https://doi.org/10.1016/j.invent.2022.100495.

12. Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr, "Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being," *npj Digital Medicine* 6 (1) (2023): 236, https://doi.org/10.1038/s41746-023-00979-5; and Gilmar Gutierrez, Callum Stephenson, Jazmin Eadie, Kimia Asadpour, and Nazanin Alavi, "Examining the Role of AI Technology in Online Mental Healthcare: Opportunities, Challenges, and Implications, a Mixed-Methods Review," *Frontiers in Psychiatry* 15 (2024): 1356773, https://doi.org/10.3389/fpsyt.2024.1356773.

13. Marcos Economides, Kristian Ranta, Albert Nazander, et al., "Long-Term Outcomes of a Therapist-Supported, Smartphone-Based Intervention for Elevated Symptoms of Depression and Anxiety: Quasiexperimental, Pre-Post-intervention Study," *JMIR mHealth and uHealth* 7 (8) (2019): e14284, https://doi.org/10.2196/14284.

14. Steven S. Clevenger, Devvrat Malhotra, Jonathan Dang, Brigitte Vanle, and Waguih William IsHak, "The Role of Selective Serotonin Reuptake Inhibitors in Preventing Relapse of Major Depressive Disorder," *Therapeutic Advances in Psychopharmacology* 8 (1) (2018), https://doi.org/10.1177/2045125317737264; and Ashish Mehta, Andrea Nicole Niles, Jose Hamilton Vargas, et al., "Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study," *Journal of Medical Internet Research* 23 (6) (2021): e26771, https://doi.org/10.2196/26771.

15. Eliane M. Boucher and Joseph S. Raiker, "Engagement and Retention in Digital Mental Health Interventions: A Narrative Review," *BMC Digital Health* 2 (1) (2024): 52, https://doi.org/10.1186/s44247-024-00105-9.

16. Clare E. Palmer, Emily Marshall, Edward Millgate, et al., "Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention with Comparable Effectiveness to Human-Delivered Care," *Journal of Medical Internet Research* 27 (2025): e69351, https://doi.org/10.2196/69351.

17. Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, et al., "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment," *NEJM AI* 2 (4) (2025): AIoa2400802, https://doi.org/10.1056/AIoa2400802.

18. Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat, "Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey," *Journal of Medical Internet Research* 21 (4) (2019): e12887, https://doi.org/10.2196/12887.

19. Tom J. Johnsen and Oddgeir Friborg, "The Effects of Cognitive Behavioral Therapy as an Anti-Depressive Treatment Is Falling: A Meta-Analysis," *Psychological Bulletin* 141 (4) (2015): 747–768, https://doi.org/10.1037/bul0000015.

20. Robin E. Gearing, Craig S. J. Schwalbe, RaeHyuck Lee, and Kimberly E. Hoagwood, "The Effectiveness of Booster Sessions in CBT Treatment for Child and Adolescent Mood and Anxiety Disorders," *Depression and Anxiety* 30 (9) (2013): 800–808, https://doi.org/10.1002/da.22118.

21. Martin E. P. Seligman, "The Effectiveness of Psychotherapy: The *Consumer Reports* Study," *American Psychologist* 50 (12) (1995): 965–974, https://doi.org/10.1037/0003-066x.50.12.965.

22. Carl R. Rogers, *Client-Centered Therapy: Its Current Practice, Implications and Theory* (Houghton Mifflin, 1951).

23. Alan E. Kazdin, "Mediators and Mechanisms of Change in Psychotherapy Research," *Annual Review of Clinical Psychology* 3 (1) (2007): 1–27, https://doi.org/10.1146/annurev.clinpsy.3.022806.091432.

24. Stanley Sue, "Psychotherapeutic Services for Ethnic Minorities: Two Decades of Research Findings," *American Psychologist* 43 (4) (1988): 301–308, https://doi.org/10.1037/0003-066x.43.4.301.

25. John R. Weisz, Michael A. Southam-Gerow, Elana B. Gordis, et al., "Cognitive-Behavioral Therapy Versus Usual Clinical Care for Youth Depression: An Initial Test of Transportability to Community Clinics and Clinicians," *Journal of Consulting and Clinical Psychology* 77 (3) (2009): 383–396, https://doi.org/10.1037/a0013877.

26. Philip S. Wang, Patricia A. Berglund, Mark Olfson, and Ronald C. Kessler, "Delays in Initial Treatment Contact after First Onset of a Mental Disorder," *Health Services Research* 39 (2) (2004): 393–415, https://doi.org/10.1111/j.1475-6773.2004.00234.x; PMID: 15032961; PMCID: PMC1361014.

27. Valerie Hoffman, Megan Flom, Timothy Y. Mariano, et al., "User Engagement Clusters of an 8-Week Digital Mental Health Intervention Guided by a Relational Agent (Woebot): Exploratory Study," *Journal of Medical Internet Research* 25 (2023): e47198, https://doi.org/10.2196/47198; and Valerie L. Forman-Hoffman, Maddison C. Pirner, Megan Flom, et al., "Engagement, Satisfaction, and Mental Health Outcomes Across Different Residential Subgroup Users of a Digital Mental Health Relational Agent: Exploratory Single-Arm Study," *JMIR Formative Research* 7 (2023): e46473, https://doi.org/10.2196/46473.

28. Heinz, Mackin, Trudeau, et al., "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment."

29. Arthur Kleinman, *Patients and Healers in the Context of Culture: An Exploration of the Borderland Between Anthropology, Medicine, and Psychiatry* (University of California Press, 1981); Arthur Kleinman, *The Illness Narratives: Suffering, Healing, and the Human Condition* (Basic Books, 1988); Arthur Kleinman, *The Soul of Care: The Moral Education of a Husband and a Doctor* (Viking/Penguin Press, 2019); Arthur Kleinman and Byron Good, *Culture and Depression: Studies in the Anthropology and Cross-Cultural Psychiatry of Affect and Disorder* (University of California

Press, 1985); Arthur Kleinman, ed., "Mental Health," *Dædalus* 152 (4) (Fall 2023); and Vikram Patel and Atif Rahman, "Empowering the (Extra)Ordinary," *Dædalus* 152 (4) (Fall 2023): 245–261, https://doi.org/10.1162/daed_a_02041.

30. Daniel Immerwahr, "What If the Attention Crisis Is All a Distraction?" *The New Yorker*, January 20, 2025.

31. Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni, "Chatbots and Mental Health: Insights into the Safety of Generative AI," *Journal of Consumer Psychology* 34 (3) (2024): 481–491, https://doi.org/10.1002/jcpy.1393.

32. Zoha Khawaja and Jean-Christophe Bélisle-Pipon, "Your Robot Therapist Is Not Your Therapist: Understanding the Role of AI-Powered Mental Health Chatbots," *Frontiers in Digital Health* 5 (2023): 1278186, https://doi.org/10.3389/fdgth.2023.1278186.

33. A. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research* 22 (7) (2020): e16021, https://doi.org/10.2196/16021; and Jake Linardon, John Torous, Joseph Firth, Pim Cuijpers, Mariel Messer, and Matthew Fuller-Tyszkiewicz, "Current Evidence on the Efficacy of Mental Health Smartphone Apps for Symptoms of Depression and Anxiety: A Meta-Analysis of 176 Randomized Controlled Trials," *World Psychiatry* 23 (1) (2024): 139–149, https://doi.org/10.1002/wps.21183.

34. Julian De Freitas and I. Glenn Cohen, "The Health Risks of Generative AI-Based Wellness Apps," *Nature Medicine* 30 (5) (2024): 1269–1275, https://doi.org/10.1038/s41591-024-02943-6.

35. American Psychological Association, "Artificial Intelligence in Mental Health Care," November 21, 2024, https://www.apa.org/practice/artificial-intelligence-mental-health-care.

36. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng; https://digital.nhs.uk/data-and-information/information-standards/governance/latest-activity/standards-and-collections/dcb0129-clinical-risk-management-its-application-in-the-manufacture-of-health-it-systems.

37. Linda Malek and Allison Kwon, "Key Developments in AI and Digital Health Signal Growing Federal Activity (Q1, 2025)," Crowell, May 5, 2025, https://www.crowell.com/en/insights/client-alerts/key-developments-in-ai-and-digital-health-signal-growing-federal-activity-q1-2025.

38. Atul Gawande, "Why Doctors Hate Their Computers," *The New Yorker*, November 12, 2018, https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers.

39. Adriana Petryna, Andrew Lakoff, and Arthur Kleinman, eds., *Global Pharmaceuticals: Ethics, Markets, Practices* (Duke University Press, 2006).

40. Shifat Islam, Rifat Shahriyar, Abhishek Agarwala, et al., "Artificial Intelligence-Based Risk Assessment Tools for Sexual, Reproductive and Mental Health: A Systematic Review," *BMC Medical Informatics and Decision Making* 25 (1) (2025), https://doi.org/10.1186/s12911-025-02864-5.

41. Anastasiya Kiseleva, Dimitris Kotzinos, and Paul De Hert, "Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations," *Frontiers in Artificial Intelligence* 5 (2022), https://doi.org/10.3389/frai.2022.879603.

42. David Collingridge, *The Social Control of Technology* (St. Martin's Press, 1980), 16.

43. Martina Kanning, Li Yi, Chih-Hsiang Yang, Christina Niermann, and Stefan Fina, "Mental Health in Urban Environments: Uncovering the Black Box of Person-Place Interactions Requires Interdisciplinary Approaches," *JMIR mHealth and uHealth* 11 (1) (2023): e41345.

44. Ana Daniela Rebelo, Damion E. Verboom, Nuno Rebelo dos Santos, and Jan Willem de Graaf, "The Impact of Artificial Intelligence on the Tasks of Mental Healthcare Workers: A Scoping Review," *Computers in Human Behavior: Artificial Humans* 1 (2) (2023): 100008, https://doi.org/10.1016/j.chbah.2023.100008.

45. Amir Tal, Zohar Elyoseph, Yuval Haber, Tal Angert, et al., "The Artificial Third: Utilizing ChatGPT in Mental Health," *American Journal of Bioethics* 23 (10) (2023): 74–77, https://doi.org/10.1080/15265161.2023.2250297.

46. Hyein S. Lee, Colton Wright, Julia Ferranto, et al., "Artificial Intelligence Conversational Agents in Mental Health: Patients See Potential, but Prefer Humans in the Loop," *Frontiers in Psychiatry* 15 (2024), https://doi.org/10.3389/fpsyt.2024.1505024.

47. R. Andrew Taylor, Rohit B. Sangal, Moira E. Smith, et al., "Leveraging Artificial Intelligence to Reduce Diagnostic Errors in Emergency Medicine: Challenges, Opportunities, and Future Directions," *Academic Emergency Medicine* 32 (3) (March 2025): 327–339, https://doi.org/10.1111/acem.15066.

48. Oliver Higgins and Rhonda L. Wilson, "Integrating Artificial Intelligence (AI) with Workforce Solutions for Sustainable Care: A Follow Up to Artificial Intelligence and Machine Learning (ML) Based Decision Support Systems in Mental Health," *International Journal of Mental Health Nursing* 34 (2) (2025): e70019, https://doi.org/10.1111/inm.70019.

49. Anithamol Babu and Akhil P. Joseph, "Artificial Intelligence in Mental Healthcare: Transformative Potential vs. the Necessity of Human Interaction," *Frontiers in Psychology* 15 (2024), https://doi.org/10.3389/fpsyg.2024.1378904.

50. Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel, "Large Language Models Lack Essential Metacognition for Reliable Medical Reasoning," *Nature Communications* 16 (642) (2025), https://doi.org/10.1038/s41467-024-55628-6.

51. Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, et al., "Towards Conversational Diagnostic Artificial Intelligence," *Nature* 642 (2025): 442–450, https://doi.org/10.1038/s41586-025-08866-7; Paul Farmer, *Infections and Inequalities: The Modern Plagues* (University of California Press, 2001); Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366 (6464) (October 24, 2019): 447–453, https://doi.org/10.1126/science.aax2342; Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones, "Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms," *New England Journal of Medicine* 383 (9) (2020): 874–882, https://doi.org/10.1056/nejmms2004740; and Kelley Tipton, Brian F. Leas, Emilia Flores, Christopher Jepson, et al., "Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare," *Comparative Effectiveness Review* (268) (2023), https://www.ncbi.nlm.nih.gov/books/NBK598802/pdf/Bookshelf_NBK598802.pdf.

52. Roman Kotov, Robert F. Krueger, David Watson, et al., "The Hierarchical Taxonomy of Psychopathology (HiTOP): A Dimensional Alternative to Traditional Nosologies," *Journal of Abnormal Psychology* 126 (4) (2017): 454–477, https://doi.org/10.1037/abn0000258.

53. Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea, "Loneliness and Suicide Mitigation for Students Using GPT3-Enabled Chatbots," *npj Mental Health Research* 3 (1) (2024): 4, https://doi.org/10.1038/s44184-023-00047-6.

54. Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad, "My AI Friend: How Users of a Social Chatbot Understand Their Human-AI Friendship," *Human Communication Research* 48 (3) (2022): 404–429, https://doi.org/10.1093/hcr/hqac008.

55. Hannah R. Marriott and Valentina Pitardi, "One Is the Loneliest Number . . . Two Can Be as Bad as One. The Influence of AI Friendship Apps on Users' Well-Being and Addiction," *Psychology and Marketing* 41 (1) (2024): 86–101, https://doi.org/10.1002/mar.21899; and Carlos Montemayor, Jodi Halpern, and Abrol Fairweather, "In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare," *AI and Society* 37 (4) (2022): 1353–1359, https://doi.org/10.1007/s00146-021-01230-z.

56. Kerrin Artemis Jacobs, "Digital Loneliness—Changes of Social Recognition Through AI Companions," *Frontiers in Digital Health* 6 (2024): 1281037, https://doi.org/10.3389/fdgth.2024.1281037.

57. Enrico Coiera, "The Price of Artificial Intelligence," *Yearbook of Medical Informatics* 28 (1) (2019): 14–15, https://doi.org/10.1055/s-0039-1677892.

58. Shunsen Huang, Xiaoxiong Lai, Li Ke, et al., "AI Technology Panic—Is AI Dependence Bad for Mental Health? A Cross-Lagged Panel Model and the Mediating Roles of Motivations for AI Use Among Adolescents," *Psychology Research and Behavior Management* 17 (2024): 1087–1102, https://doi.org/10.2147/PRBM.S440889.

59. Harvey J. Locke and Karl M. Wallace, "Short Marital-Adjustment and Prediction Tests: Their Reliability and Validity," *Marriage and Family Living* 21 (3) (1959): 251–255, https://doi.org/10.2307/348022.

60. Jenna L. Wells, Claudia M. Haase, Emily S. Rothwell, et al., "Positivity Resonance in Long-Term Married Couples: Multimodal Characteristics and Consequences for

Health and Longevity," *Journal of Personality and Social Psychology* 123 (5) (2022): 983–1003, https://doi.org/10.1037/pspi0000385.

61. Joseph Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9 (1) (1966): 36–45, https://doi.org/10.1145/365153.365168.

62. Robert K. Merton, "The Unanticipated Consequences of Purposive Social Action," *American Sociological Review* 1 (6) (1936): 894, https://doi.org/10.2307/2084615.

63. Roshini Salil, Binny Jose, Jaya Cherian, Sheeja P. R, and Nisha Vikraman, "Digitalized Therapy and the Unresolved Gap Between Artificial and Human Empathy," *Frontiers in Psychiatry* 15 (2024), https://doi.org/10.3389/fpsyt.2024.1522915.

64. Huang et al., "AI Technology Panic."

65. Sigmund Freud, *The Interpretation of Dreams* (Hogarth Press, 1900).

66. Huang et al., "AI Technology Panic"; Harikrishna Patel and Faiza Hussain, "Do AI Chatbots Incite Harmful Behaviours in Mental Health Patients?" *BJPsych Open* 10 (S1) (2024): S70–S71, https://doi.org/10.1192/bjo.2024.225; and Marcin Rządeczka, Anna Sterna, Julia Stolińska, Paulina Kaczyńska, and Marcin Moskalewicz, "The Efficacy of Conversational AI in Rectifying the Theory-of-Mind and Autonomy Biases: Comparative Analysis," *JMIR Mental Health* 12 (2025): e64396, https://doi.org/10.2196/64396.

67. Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, et al., "Evaluating Therabot: A Randomized Control Trial Investigating the Feasibility and Effectiveness of a Generative AI Therapy Chatbot for Depression, Anxiety, and Eating Disorder Symptom Treatment," preprint, PsyArXiv, June 14, 2024, last edited August 23, 2025, https://doi.org/10.31234/osf.io/pjqmr.

68. Katharine A. Smith, Amy Hardy, Anastasia Vinnikova, et al., "Digital Mental Health for Schizophrenia and Other Severe Mental Illnesses: An International Consensus on Current Challenges and Potential Solutions," *JMIR Mental Health* 11 (2024): e57155, https://doi.org/10.2196/57155.

69. Aakash Ganju, Srini Satyan, Vatsal Tanna, and Sonia Rebecca Menezes, "AI for Improving Children's Health: A Community Case Study," *Frontiers in Artificial Intelligence* 3 (2021): 544972, https://doi.org/10.3389/frai.2020.544972; and Tony Rousmaniere, Xu Li, Yimeng Zhang, and Siddharth Shah, "Large Language Models as Mental Health Resources: Patterns of Use in the United States," preprint, PsyArXiv, March 18, 2025, last edited August 28, 2025, https://osf.io/preprints/psyarxiv/q8m7g_v1.

70. Adrian Buttazzoni, Keshbir Brar, and Leia Minaker, "Smartphone-Based Interventions and Internalizing Disorders in Youth: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research* 23 (1) (2021): e16490, https://doi.org/10.2196/16490; and Xiaoyun Zhou, Sisira Edirippulige, Xuejun Bai, and Matthew Bambling, "Are Online Mental Health Interventions for Youth Effective? A Systematic Review," *Journal of Telemedicine and Telecare* 27 (10) (2021): 638–666, https://doi.org/10.1177/1357633x211047285.

71. Maples et al., "Loneliness and Suicide Mitigation."

72. Katarína Greškovičová, Radomír Masaryk, Nikola Synak, and Vladimíra Čavojová, "Superlatives, Clickbaits, Appeals to Authority, Poor Grammar, or Boldface: Is Editorial Style Related to the Credibility of Online Health Messages?" *Frontiers in Psychology* 13 (2022): 940903, https://doi.org/10.3389/fpsyg.2022.940903.

73. Jaemarie Solyst, Ellia Yang, Shixian Xie, et al., "Children's Overtrust and Shifting Perspectives of Generative AI," preprint, arXiv, https://doi.org/10.48550/arXiv.2404.14511 (2024).

74. Nitasha Tiku, "AI Friendships Claim to Cure Loneliness: Some Are Ending in Suicide," *Washington Post*, December 6, 2024, https://www.washingtonpost.com/technology/2024/12/06/ai-companion-chai-research-character-ai; and Queenie Wong, "Teens Are Forming Bonds with AI Chatbots, Raising Concerns," *Los Angeles Times*, March 14, 2025, https://www.latimes.com/business/story/2025-02–25/teens-are-spilling-dark-thoughts-to-ai-chatbots-whos-to-blame-when-something-goes-wrong.

75. Ilaria Montagni, Christophe Tzourio, Thierry Cousin, Joseph Amadomon Sagara, Jennifer Bada-Alonzi, and Aine Horgan, "Mental Health-Related Digital Use by University Students: A Systematic Review," *Telemedicine and e-Health* 26 (2) (2020): 131–146, https://doi.org/10.1089/tmj.2018.0316; Lisa Parker, Vanessa Halter, Tanya Karliychuk, and Quinn Grundy, "How Private Is Your Mental Health App Data? An Empirical Study of Mental Health App Privacy Policies and Practices," *International Journal of Law and Psychiatry* 64 (2019): 198–204, https://doi.org/10.1016/j.ijlp.2019.04.002;

Yuanyuan Dang, Shanshan Guo, Xitong Guo, Mohan Wang, and Kexin Xie, "Privacy Concerns About Health Information Disclosure in Mobile Health: Questionnaire Study Investigating the Moderation Effect of Social Support," *JMIR mHealth and uHealth* 9 (2) (2020): e19594, https://doi.org/10.2196/19594; and Emily Watson, Sue Fletcher-Watson, and Elizabeth Joy Kirkham, "Views on Sharing Mental Health Data for Research Purposes: Qualitative Analysis of Interviews with People with Mental Illness," *BMC Medical Ethics* 24 (1) (2023), https://doi.org/10.1186/s12910-023-00961-6.

76. Jeffrey Foster and Jennifer J. Williams, "Medibank Hackers Are Now Releasing Stolen Data on the Dark Web: If You're Affected, Here's What You Need to Know," *The Conversation*, November 9, 2022, https://theconversation.com/medibank-hackers-are-now-releasing-stolen-data-on-the-dark-web-if-youre-affected-heres-what-you-need-to-know-194340.

77. Federal Trade Commission, "Proposed FTC Order Will Prohibit Telehealth Firm Cerebral From Using or Disclosing Sensitive Data for Advertising Purposes, and Require It to Pay $7 Million," Federal Trade Commission, April 24, 2025, https://www.ftc.gov/news-events/news/press-releases/2024/04/proposed-ftc-order-will-prohibit-telehealth-firm-cerebral-using-or-disclosing-sensitive-data; and Steve Alder, "Security Breaches in Healthcare in 2023," *HIPAA Journal*, January 31, 2024, https://www.hipaajournal.com/security-breaches-in-healthcare.

78. Bryanna Moore, Jonathan Herington, and Şerife Tekin, "The Integration of Artificial Intelligence-Powered Psychotherapy Chatbots in Pediatric Care: Scaffold or Substitute?" *The Journal of Pediatrics* 280 (2025): 114509, https://doi.org/10.1016/j.jpeds.2025.114509.

79. Douglas J. Opel, Brent M. Kious, and I. Glenn Cohen, "AI as a Mental Health Therapist for Adolescents," *JAMA Pediatrics* 177 (12) (2023): 1253–1254, https://doi.org/10.1001/jamapediatrics.2023.4215.

80. Companion Chatbots, SB-243, California Legislature, 2025–2026 sess. (enacted), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243.

81. Kids Off Social Media Act, S. 2413, 118th Cong. (2024).

82. Jonathan Knights, Victoria Bangieva, Michela Passoni, et al., "A Framework for Precision 'Dosing' of Mental Healthcare Services: Algorithm Development and Clinical Pilot," *International Journal of Mental Health Systems* 17 (1) (2023), https://doi.org/10.1186/s13033-023-00581-y.

83. Jonathan Knights, Jacob Shen, Vincent Mysliwiec, and Holly DuBois, "Associations of Smartphone Usage Patterns with Sleep and Mental Health Symptoms in a Clinical Cohort Receiving Virtual Behavioral Medicine Care: A Retrospective Study," *SLEEP Advances* 4 (1) (2023), https://doi.org/10.1093/sleepadvances/zpad027.

84. Linda Alfano, Ivano Malcotti, and Rosagemma Ciliberti, "Psychotherapy, Artificial Intelligence and Adolescents: Ethical Aspects," *Journal of Preventive Medicine and Hygiene* 64 (4) (2024): E438–E442, https://pubmed.ncbi.nlm.nih.gov/38379752.

85. Seo Yi Chng, Mark Jun Wen Tern, Yung Seng Lee, et al., "Ethical Considerations in AI for Child Health and Recommendations for Child-Centered Medical AI," *npj Digital Medicine* 8 (1) (2025), https://doi.org/10.1038/s41746-025-01541-1.

86. Bo Wang, Cecilie Katrine Grønvik, Karen Fortuna, Trude Eines, Ingunn Mundal, and Marianne Storm, "What Is in There for Artificial Intelligence to Support Mental Health Care for Persons with Serious Mental Illness? Opportunities and Challenges," *Studies in Health Technology and Informatics* 325 (2025): 8–15, https://doi.org/10.3233/shti250208.

87. Mehrdad Rahsepar Meadi, Tomas Sillekens, Suzanne Metselaar, Anton Van Balkom, Justin Bernstein, and Neeltje Batelaan, "Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review," *JMIR Mental Health* 12 (2025): e60432, https://doi.org/10.2196/60432.

88. Masab Mansoor and Kashif Ansari, "Artificial Intelligence-Driven Analysis of Telehealth Effectiveness in Youth Mental Health Services: Insights from SAMHSA Data," *Journal of Personalized Medicine* 15 (2) (2025): 63, https://doi.org/10.3390/jpm15020063.

89. Adam C. Powell, Matthias B. Bowman, and Henry T. Harbin, "Reimbursement of Apps for Mental Health: Findings from Interviews," *JMIR Mental Health* 6 (8) (2019): e14724.

90. Frank Vinluan, "Digital Therapeutics Sector Sees Billing Codes as Key to Breaking Free of Reimbursement Rut," *MedCity News*, August 19, 2024, https://medcitynews.com

/2024/08/digital-therapeutics-reimbursement-cms-medicare-insurance-payers-clinical-trials.

91. "CY 2025 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment and Coverage Policies," CMS-1807-F, December 9, 2024, https://www.cms.gov/medicare/payment/fee-schedules/physician/federal-regulation-notices/cms-1807-f.

92. Melissa M. Chen, Lauren Parks Golding, and Gregory N. Nicola, "Who Will Pay for AI?" *Radiology: Artificial Intelligence* 3 (3) (2021): e210030, https://doi.org/10.1148/ryai.2021210030.

93. Ginnie Sawyer-Morris, Judith A. Wilde, Todd Molfenter, and Faye Taxman, "Use of Digital Health and Digital Therapeutics to Treat SUD in Criminal Justice Settings: A Review," *Current Addiction Reports* 11 (1) (2024): 149–162, https://doi.org/10.1007/s40429-023-00523-1.

94. Robin van Kessel, Divya Srivastava, Ilias Kyriopoulos, et al., "Digital Health Reimbursement Strategies of 8 European Countries and Israel: Scoping Review and Policy Mapping," *JMIR mHealth and uHealth* 11 (1) (2023): e49003, https://doi.org/10.2196/49003.

95. Ainhoa Gomez Lumbreras, Jason T. Hurwitz, Xi Liang, et al., "Insights into Insurance Coverage for Digital Therapeutics: A Qualitative Study of US Payer Perspectives," *Journal of Managed Care and Specialty Pharmacy* 30 (4) (2024): 313–325, https://doi.org/10.18553/jmcp.2024.30.4.313.

96. Van Kessel et al., "Digital Health Reimbursement Strategies."

97. Ibid.

98. Vinluan, "Digital Therapeutics Sector."

99. Ibid.

100. William J. Gordon, Adam Landman, Haifeng Zhang, and David W. Bates, "Beyond Validation: Getting Health Apps into Clinical Practice," *npj Digital Medicine* 3 (1) (2020): 14.

101. Michael D. Abràmoff, Cybil Roehrenbeck, Sylvia Trujillo, et al., "A Reimbursement Framework for Artificial Intelligence in Healthcare," *npj Digital Medicine* 5 (1) (2022): 72, https://doi.org/10.1038/s41746-022-00621-w.

102. Mansoor and Ansari, "Artificial Intelligence-Driven Analysis."

103. Yu Song, Chenfei Qian, and Susan Pickard, "Age-Related Digital Divide During the COVID-19 Pandemic in China," *International Journal of Environmental Research and Public Health* 18 (21) (2021): 11285, https://doi.org/10.3390/ijerph182111285.

104. Richard G. Frank and Sherry A. Glied, "America's Continuing Struggle with Mental Illnesses: Economic Considerations," *Journal of Economic Perspectives* 37 (2) (2023): 153–178, http://dx.doi.org/10.1257/jep.37.2.153.

105. NCQA, "Improving Accountability for Behavioral Health Care Access: Evaluating the Current Evidence for Behavioral Health Network Adequacy Standards," white paper, February 2, 2024, https://www.ncqa.org/white-papers/improving-accountability-for-behavioral-health-care-access.

106. Li et al., "Systematic Review and Meta-Analysis"; and Abd-Alrazaq, Alaa Ali, Asma Rababeh, Mohannad Alajlani, Bridgette M. Bewick, and Mowafa Househ, "Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research* 22 (7) (2020): e16021, https://doi.org/10.2196/16021.

107. Ambre Marie, Marine Garnier, Thomas Bertin, et al., "Acoustic and Machine Learning Methods for Speech-Based Suicide Risk Assessment: A Systematic Review," preprint, arXiv, May 20, 2025, https://doi.org/10.48550/arXiv.2505.18195.

108. Kaiser Family Foundation, *Mental Health Care Health Professional Shortage Areas* (HPSAs), https://www.kff.org/other-health/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas; and M. Reinert, T. Nguyen, and D. Fritze, *The State of Mental Health in America* (Mental Health America, 2023), https://mhanational.org/the-state-of-mental-health-in-america.

109. Ching-Fang Sun, Christoph U. Correll, Robert L. Trestman, et al., "Low Availability, Long Wait Times, and High Geographic Disparity of Psychiatric Outpatient Care in the U.S.," *General Hospital Psychiatry* 84 (2023): 12–17, https://doi.org/10.1016/j.genhosppsych.2023.05.012.

110. Kaiser Family Foundation, *Mental Health Care Health Professional Shortage Areas*.

111. Peiyin Hung, Janice C. Probst, Yiwen Shih, et al., "Rural-Urban Disparities in Quality of Inpatient Psychiatric Care," *Psychiatric Services* 74 (5) (2023): 446–454, https://doi.org/10.1176/appi.ps.20220277; Shawnda Schroeder, Chih Ming Tan, Brian Urlacher, and Thomasine Heitkamp, "The Role of Rural and Urban Geography and Gender in Community Stigma Around Mental Illness," *Health Education and Behavior* 48 (1) (2021): 63–73, https://doi.org/10.1177/1090198120974963; and Karen J. Coleman, Christine Stewart, Beth E. Waitzfelder, et al., "Racial-Ethnic Differences in Psychiatric Diagnoses and Treatment Across 11 Health Care Systems in the Mental Health Research Network," *Psychiatric Services* 67 (7) (2016): 749–757, https://doi.org/10.1176/appi.ps.201500217.

112. American Psychological Association, *The State of Mental Health in America*.

113. "The Mental Health Parity and Addiction Equity Act (MHPAEA)," Centers for Medicare and Medicaid Services, last updated September 10, 2024, https://www.cms.gov/marketplace/private-health-insurance/mental-health-parity-addiction-equity.

114. "Mobile Fact Sheet," Pew Research Center, November 13, 2024, https://www.pewresearch.org/internet/fact-sheet/mobile.

115. Donald M. Hilty, Melanie T. Gentry, Alastair J. McKean, Kirsten E. Cowan, Russell F. Lim, and Francis G. Lu, "Telehealth for Rural Diverse Populations: Telebehavioral and Cultural Competencies, Clinical Outcomes and Administrative Approaches," *mHealth* 6 (2020): 20, https://doi.org/10.21037/mhealth.2019.10.04.

116. Jennifer Huberty, Jeni Green, Christine Glissmann, Linda Larkey, Megan Puzia, and Chong Lee, "Efficacy of the Mindfulness Meditation Mobile App 'Calm' to Reduce Stress Among College Students: Randomized Controlled Trial," *JMIR mHealth and uHealth* 7 (6) (2019): e14273, https://doi.org/10.2196/14273; and Emily Simon, Alyssa M. Edwards, Martha Sajatovic, Nisha Jain, Jessica L. Montoya, and Jennifer B. Levin, "Systematic Literature Review of Text Messaging Interventions to Promote Medication Adherence Among People with Serious Mental Illness," *Psychiatric Services* 73 (10) (2022): 1153–1164, https://doi.org/10.1176/appi.ps.202100634.

117. Park and West, "Why Mental Health Apps Need to Take Privacy More Seriously."

118. Anish Gupta, Ruchika Gupta, and Puneet Kumar, "Artificial Intelligence for Remote Healthcare in Underserved Areas: Enhancing Access and Quality of Healthcare Delivery," in *Artificial Intelligence and Its Applications: ICAIA 2023*, ed. A. Gupta, M. Hinchey, and Z. Zalevsky (Springer Nature Switzerland, 2025), 122–138, https://doi.org/10.1007/978-3-031-84397-6_9.

119. Madison Milne-Ives, Caroline de Cock, Ernest Lim, et al., "The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review," *Journal of Medical Internet Research* 22 (10) (2020): e20346, https://doi.org/10.2196/20346; and Robert Meadows, Christine Hine, and Eleanor Suddaby, "Conversational Agents and the Making of Mental Health Recovery," *Digital Health* 6 (2020), https://doi.org/10.1177/2055207620966170.

120. Samantha Tyler, Matthew Olis, Nicole Aust, et al., "Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review," *Cureus* 16 (5) (2024): e59906, https://doi.org/10.7759/cureus.59906.

121. Hesham Allam, Chris Davison, Faisal Kalota, Edward Lazaros, and David Hua, "AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques," *Big Data and Cognitive Computing* 9 (1) (2025): 16, https://doi.org/10.3390/bdcc9010016.

122. Jiacheng Dai, Yu Chen, Cuihua Xia, Jiaqi Zhou, Chunyu Liu, and Chao Chen, "Digital Sensory Phenotyping for Psychiatric Disorders," *Journal of Psychiatry and Brain Science* 5 (3) (2020), https://doi.org/10.20900/jpbs.20200015; and Daniel A. Adler, Dror Ben-Zeev, Vincent W. S. Tseng, et al., "Predicting Early Warning Signs of Psychotic Relapse from Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks," *JMIR mHealth and uHealth* 8 (8) (2020): e19962, https://doi.org/10.2196/19962.

123. Yusuf Olalekan, Raphael Ekundayo Adesiyan, Chibuzor Stella Amadi, Oluwaseun Ipede, Lucy Oluebubechi Karakitie, and Kaosara Temitope Adebayo, "Cross-Cultural Perspectives on Mental Health: Understanding Variations and Promoting Cultural Competence," *World Journal of Advanced Research and Reviews* 23 (1) (2024): 432–439, https://doi.org/10.30574/wjarr.2024.23.1.2040; and Adela C. Timmons, Jacqueline B. Duong, Natalia Simo Fiallo, et al., "A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health," *Perspectives on Psychological Science* 18 (5) (2023): 1062–1096, https://doi.org/10.1177/17456916221134490.

124. Lauren Mizock and Debra Harkins, "Diagnostic Bias and Conduct Disorder: Improving Culturally Sensitive Diagnosis," *Child and Youth Services* 32 (3) (2011): 243–253, https://doi.org/10.1080/0145935X.2011.605315.

125. Obermeyer, Powers, Vogeli, and Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations"; and Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, et al., "Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare," *npj Digital Medicine* 3 (1) (2020): 81, https://doi.org/10.1038/s41746-020-0288-5.

126. Alexander d'Elia, Mark Gabbay, Sarah Rodgers, et al., "Artificial Intelligence and Health Inequities in Primary Care: A Systematic Scoping Review and Framework," *Family Medicine and Community Health* 10 (S1) (2022): e001670, https://doi.org/10.1136/fmch-2022-001670.

127. Amelia Fiske, Peter Henningsen, and Alena Buyx, "Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy," *Journal of Medical Internet Research* 21 (5) (2019): e13216, https://doi.org/10.2196/13216.

128. Andrea Putica, Miriam Yurtbasi, and Rahul Khanna, "Integrating Digital Health Technologies for Ecological Validity in Computational Psychiatry: Challenges and Solutions," *AI and Society* 40 (2025): 1–17, https://doi.org/10.1007/s00146-025-02336-4.

129. Adler et al., "Predicting Early Warning Signs of Psychotic Relapse from Passive Sensing Data."

130. For an example of the research, see Heinz et al., "Randomized Trial of a Generative AI Chatbot." On regulation, see Jennifer Lee O'Donnell, "Walter Benjamin in the Bathhouse: Meditations on Robot Mothers, Daydreams, and Art in the AI Era," *Anthropology and Humanism* 50 (1) (2025): e70006, https://doi.org/10.1111/anhu.70006.

131. Albert Mehrabian, *Nonverbal Communication*, eBook ed. (1972; Routledge, 2017), https://doi.org/10.4324/9781351308724.

132. Derek Griner and Timothy B. Smith, "Culturally Adapted Mental Health Intervention: A Meta-Analytic Review," *Psychotherapy: Theory, Research, Practice, Training* 43 (4) (2006): 531–548, https://psycnet.apa.org/doi/10.1037/0033-3204.43.4.531.

133. Alison M. Darcy, Angela Celio Doyle, James Lock, Rebecka Peebles, Peter Doyle, and Daniel Le Grange, "The Eating Disorders Examination in Adolescent Males with Anorexia Nervosa: How Does It Compare to Adolescent Females?" *International Journal of Eating Disorders* 45 (1) (2012): 110–114, https://doi.org/10.1002/eat.20896; and Alison M. Darcy and Iris Hsiao-Jung Lin, "Are We Asking the Right Questions? A Review of Assessment of Males With Eating Disorders," *Eating Disorders: The Journal of Treatment and Prevention* 20 (5) (2012): 416–426, https://doi.org/10.1080/10640266.2012.715521.

134. Byron Crowe and Jorge A. Rodriguez, "Identifying and Addressing Bias in Artificial Intelligence," *JAMA Network Open* 7 (8) (2024): e2425955, doi:10.1001/jamanetworkopen.2024.25955.

135. Valerie Hoffman, Megan Flom, Timothy Y. Mariano, et al., "User Engagement Clusters of an 8-Week Digital Mental Health Intervention Guided by a Relational Agent (Woebot): Exploratory Study," *Journal of Medical Internet Research* 25 (2023): e47198, https://doi.org/10.2196/47198.

136. L. Bullard, S. Rapoport, L. Ariniello, et al., "Utilization of Community-Based Participatory Research Strategies in a Postpartum Digital Health Trial," poster presentation, 2024 Postpartum Support International (PSI) Annual Conference, Washington, D.C., July 26–28, 2024; and Andrew Kirvin-Quamme, Jennifer Kissinger, Laurel Quinlan et al., "Common Practices for Sociodemographic Data Reporting in Digital Mental Health Intervention Research: A Scoping Review," *BMJ Open* 14 (2024): e078029, doi:10.1136/bmjopen-2023-078029.

137. Arthur Kleinman and Caleb Gardner, "Good Mental Health Care: What It Is, What It Is Not & What It Could Be," *Dædalus* 152 (4) (2023): 262–279, https://doi.org/10.1162/daed_a_02042; Jonathan M. Metzl and Helena Hansen, "Structural Competency: Theorizing a New Medical Engagement With Stigma and Inequality," *Social Science & Medicine* 103 (2014): 126–133, https://doi.org/10.1016/j.socscimed.2013.06.032; Arthur Kleinman and Peter Benson, "Anthropology in the Clinic: The Problem of Cultural Competency and How to Fix It," *PLOS Medicine* 3 (10) (2006): e294, https://doi.org/10.1371/journal.pmed.0030294; Obermeyer, Powers, Vogeli, and Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations"; Vyas, Eisenstein, and Jones, "Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms"; and Tipton et al., "Impact of Healthcare

Algorithms on Racial and Ethnic Disparities in Health and Healthcare."

138. Kacie Kelly and Stephen Schueller, *Increasing Access to High-Quality Mental Health Care in the 21st Century* (George W. Bush Institute, 2021), https://gwbushcenter.imgix.net/wp-content/uploads/Stand-To_HWB_Report.pdf.

139. Ibid.; Martha Neary, Kris Tran, Hutton Grabiel, John Bunyi, and Stephen M. Schueller, *Digital Tools and Solutions for Teen Mental Health* (One Mind, 2022); and *A Growing Psychiatrist Shortage and an Enormous Demand for Mental Health Services* (AAMC, n.d.), https://www.aamc.org/news/growing-psychiatrist-shortage-enormous-demand-mental-health-services (accessed September 6, 2023).

140. Stacy Weiner, "A Growing Psychiatrist Shortage and an Enormous Demand for Mental Health Services," *AAMCNews*, August 9, 2022.

141. M. D. Romael Haque and Sabirat Rubya, "An Overview of Chatbot-Based Mobile Mental Health Apps: Insights from App Description and User Reviews," *JMIR mHealth and uHealth* 11 (1) (2023): e44838, https://doi.org/10.2196/44838.

142. "Augmented Intelligence in Medicine," American Medical Association, last updated August 7, 2025, https://www.ama-assn.org/practice-management/digital/augmented-intelligence-medicine.

143. Jean P. Flores, Geoffrey Kahn, Robert B. Penfold, et al., "Adolescents Who Do Not Endorse Risk via the Patient Health Questionnaire Before Self-Harm or Suicide," *JAMA Psychiatry* 81 (7) (2024): 717–726, https://doi.org/10.1001/jamapsychiatry.2024.0603.

144. Emil Chiauzzi and Paul Wicks, "Beyond the Therapist's Office: Merging Measurement-Based Care and Digital Medicine in the Real World," *Digital Biomarkers* 5 (2) (2021): 176–182, https://doi.org/10.1159/000517748.

145. Pasquale Bufano, Marco Laurino, Sara Said, Alessandro Tognetti, and Danilo Menicucci, "Digital Phenotyping for Monitoring Mental Disorders: Systematic Review," *Journal of Medical Internet Research* 25 (2023): e46778, https://doi.org/10.2196/46778.

146. Ibid.

147. Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou, "Natural Language Processing Applied to Mental Illness Detection: A Narrative Review," *NPJ Digital Medicine* 5 (1) (2022): 46, https://doi.org/10.1038/s41746-022-00589-7.

148. Sarah Graham, Colin Depp, Ellen E. Lee, et al., "Artificial Intelligence for Mental Health and Mental Illnesses: An Overview," *Current Psychiatry Reports* 21 (2019): 116, https://doi.org/10.1007/s11920-019-1094-0.

149. John F. McCarthy, Samantha A. Cooper, Kallisse R. Dent, et al., "Evaluation of the Recovery Engagement and Coordination for Health-Veterans Enhanced Treatment Suicide Risk Modeling Clinical Program in the Veterans Health Administration," *JAMA Network Open* 4 (10) (2021): e2129900, https://doi.org/10.1001/jamanetworkopen.2021.29900.

150. Maxwell Levis, Joshua Levy, Kallisse R. Dent, et al., "Leveraging Natural Language Processing to Improve Electronic Health Record Suicide Risk Prediction for Veterans Health Administration Users," *The Journal of Clinical Psychiatry* 84 (4) (2023): 22m14568, https://doi.org/10.4088/jcp.22m14568.

151. "Identity and Cultural Dimensions," National Alliance on Mental Illness, https://www.nami.org/your-journey/identity-and-cultural-dimensions.

152. Isabel Straw and Chris Callison-Burch, "Artificial Intelligence in Mental Health and the Biases of Language Based Models," *PLOS One* 15 (12) (2020): e0240376, https://doi.org/10.1371/journal.pone.0240376.

153. John P. Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi, "Benchmarking Intersectional Biases in NLP," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2022), 3598–3609, https://aclanthology.org/2022.naacl-main.263.pdf.

154. Jutta Treviranus, "The Three Dimensions of Inclusive Design: Part Three," Medium, April 13, 2018, https://medium.com/@jutta.trevira/the-three-dimensions-of-inclusive-design-part-three-b6585c737f40.

155. "What Is an Edge Case?" Coursera, last updated March 15, 2025, https://www.coursera.org/articles/edge-case.

156. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in "Conference on Fairness, Accountability and Transparency, 23–24 February 2018, New York, NY, USA," special issue of *Proceedings of Machine Learning Research* 81 (2018): 77–91, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

157. C. S. Gilfoil, J. Liou, and T. Quach, "Advancing Health Equity Through Disaggregated Race/Ethnicity Data," Leadership Conference on Civil and Human Rights, July 28, 2023, https://civilrights.org/blog/advancing-health-equity-through-disaggregated-race-ethnicity-data.

158. Obermeyer, Powers, Vogeli, and Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations."

159. Ibid.

160. Ziad Obermeyer, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, and Sendhil Mullainathan, *Algorithmic Bias Playbook* (Center for Applied Artificial Intelligence at Chicago Booth, 2021), https://www.chicagobooth.edu/-/media/project/chicago-booth/centers/caai/docs/algorithmic-bias-playbook-june-2021.

161. Veronica Barcelona, Danielle Scharp, Betina R. Idnay, Hans Moen, Kenrick Cato, and Maxim Topaz, "Identifying Stigmatizing Language in Clinical Documentation: A Scoping Review of Emerging Literature," *PLOS One* 19 (6) (2024): e0303653, https://doi.org/10.1371/journal.pone.0303653.

162. Gagan Jain, Samridhi Pareek, and Per Carlbring, "Revealing the Source: How Awareness Alters Perceptions of AI and Human-Generated Mental Health Responses," *Internet Interventions* 36 (2024): 100745, https://doi.org/10.1016/j.invent.2024.100745.

163. Elizabeth Yong, Yen Nee Teo, and Kun Hing Yong, "AI Technology: A New Game Changer for the Future Mental Health Industry?" *Asia Pacific Journal of Public Health* 37 (1) (2025): 148–149, https://doi.org/10.1177/10105395241303790.

164. Liana Spytska, "The Use of Artificial Intelligence in Psychotherapy: Development of Intelligent Therapeutic Systems," *BMC Psychology* 13 (1) (2025): 175, https://doi.org/10.1186/s40359-025-02491-9.

165. Rinad Bakhti, Harmani Daler, Hephzibah Ogunro, Steven Hope, Dougal Hargreaves, and Dasha Nicholls, "Exploring Engagement with and Effectiveness of Digital Mental Health Interventions in Young People of Different Ethnicities: Systematic Review," *Journal of Medical Internet Research* 27 (2025): e68544, https://doi.org/10.2196/68544.

166. Ashleigh Golden and Elias Aboujaoude, "The Framework for AI Tool Assessment in Mental Health (FAITA-Mental Health): A Scale for Evaluating AI-Powered Mental Health Tools," *World Psychiatry* 23 (3) (2024): 444–445, https://doi.org/10.1002/wps.21248.

167. Shane Cross, Imogen Bell, Jennifer Nicholas, et al., "Use of AI in Mental Health Care: Community and Mental Health Professionals Survey," *JMIR Mental Health* 11 (1) (2024): e60589, https://doi.org/10.2196/60589.

168. Hong and Emanuel, "Leveraging Artificial Intelligence to Bridge the Mental Health Workforce Gap and Transform Care."

169. Fanny Kählke, Claudia Buntrock, Filip Smit, and David Daniel Ebert, "Systematic Review of Economic Evaluations for Internet- and Mobile-Based Interventions for Mental Health Problems," *npj Digital Medicine* 5 (1) (2022): 175, https://doi.org/10.1038/s41746-022-00702-w; and Kaiser Family Foundation, *Mental Health Care Health Professional Shortage Areas*.

170. Shehzad Ali, Feben W. Alemu, Jesse Owen, et al., "Cost-Effectiveness of Computer-Assisted Cognitive Behavioral Therapy for Depression Among Adults in Primary Care," *JAMA Network Open* 7 (11) (2024): e2444599, https://doi.org/10.1001/jamanetworkopen.2024.44599; and American Psychological Association, *State of Mental Health in America*.

171. Hong and Emanuel, "Leveraging Artificial Intelligence to Bridge the Mental Health Workforce Gap and Transform Care."

172. Barry Solaiman, Abeer Malik, and Suhaila Ghuloum, "Monitoring Mental Health: Legal and Ethical Considerations of Using Artificial Intelligence in Psychiatric Wards," *American Journal of Law and Medicine* 49 (2–3) (2023): 250–266, https://doi.org/10.1017/amj.2023.30.

173. Hong and Emanuel, "Leveraging Artificial Intelligence to Bridge the Mental Health Workforce Gap and Transform Care."

174. Lee et al., "Artificial Intelligence for Mental Health Care."

175. Cross et al., "Use of AI in Mental Health Care."

176. Zainab Iftikhar, Sean Ransom, Amy Xiao, et al., "Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT," preprint, arXiv, September 3, 2024, https://doi.org/10.48550/arXiv.2409.02244.

177. Iyesatta Massaquoi Emeli, "AI Scribe Technology Lets Me Focus on My Patients, Not a Screen," *STAT*, April 9, 2025, https://www.statnews.com/2025/04/09/ai-scribes-ambient-listening-ehrs-physician-experience-fan.

178. Peterson Health Technology Institute, *Adoption of Artificial Intelligence in Healthcare Delivery Systems: Early Applications and Impacts* (Peterson Health Technology Institute, March 2025), https://phti.org/wp-content/uploads/sites/3/2025/03/PHTI-Adoption-of-AI-in-Healthcare-Delivery-Systems-Early-Applications-Impacts.pdf.

179. H. Membride, "Mental Health: Early Intervention and Prevention in Children and Young People," *British Journal of Nursing* 25 (10) (2016): 552–557, https://doi.org/10.12968/bjon.2016.25.10.552; and J. M. Kane, D. G. Robinson, N. R. Schooler, et al., "Comprehensive Versus Usual Community Care for First-Episode Psychosis: 2-Year Outcomes from the NIMH RAISE Early Treatment Program," *American Journal of Psychiatry* 173 (4) (2016): 362–372, https://doi.org/10.1176/appi.ajp.2015.15050632.

180. Melody Zhang, Jillian Scandiffio, Sarah Younus, et al., "The Adoption of AI in Mental Health Care—Perspectives from Mental Health Professionals: Qualitative Descriptive Study," *JMIR Formative Research* 7 (2023): e47847, https://doi.org/10.2196/47847.

181. "Upskilling the Health Information Workforce in the Age of AI," American Health Information Management Association (AHIMA), https://www.ahima.org/education-events/artificial-intelligence/upskilling-the-health-information-workforce-in-the-age-of-ai.

182. Kelly E. Ormond, Mercy Ygoña Laurino, Kristine Barlow-Stewart, et al., "Genetic Counseling Globally: Where Are We Now?" *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 178 (1) (2018): 98–107, https://doi.org/10.1002/ajmg.c.31607; and Lee et al., "Artificial Intelligence for Mental Health Care."

183. Elizabeth M. Renieris, David Kiron, and Steven Mills, "A Fragmented Landscape Is No Excuse for Global Companies Serious About Responsible AI," *MIT Sloan Management Review*, October 29, 2024, https://sloanreview.mit.edu/article/a-fragmented-landscape-is-no-excuse-for-global-companies-serious-about-responsible-ai.

184. Ryan Browne, "AI That Can Match Humans at Any Task Will Be Here in Five to 10 Years, Google DeepMind CEO Says," CNBC, March 17, 2025, https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html.

185. Adam B. Cohen, Simon C. Mathews, E. Ray Dorsey, David W. Bates, and Kyan Safavi, "Direct-to-Consumer Digital Health," *The Lancet Digital Health* 2 (4) (2020): e163–e165, https://doi.org/10.1016/S2589-7500(20)30057-1.

186. Kaiser Family Foundation, "Health Insurance Coverage of Children 0–18," October 28, 2022, https://www.kff.org/other/state-indicator/children-0-18.

187. Mario Aguilar, "Do AI Scribes Help Health Systems Save Time?" *STAT*, March 27, 2025, https://www.statnews.com/2025/03/27/do-ai-scribes-help-health-systems-save-time-health-tech.

188. Meadows Mental Health Policy Institute, *Near-Term Policy Solutions to Bolster the Youth Mental Health Workforce Through Digital Technology* (Meadows Mental Health Policy Institute, 2023), https://mmhpi.org/wp-content/uploads/2023/05/Nearterm-Solutions-DMHT_05302023.pdf.

189. Maria del Pilar Arias López, Bradley A. Ong, Xavier Borrat Frigola, et al., "Digital Literacy as a New Determinant of Health: A Scoping Review," *PLOS Digital Health* 2 (10) (2023): e0000279, https://doi.org/10.1371/journal.pdig.0000279.

190. Chiara Berardi, Marcello Antonini, Zephanie Jordan, Heidi Wechtler, Francesco Paolucci, and Madeleine Hinwood, "Barriers and Facilitators to the Implementation of Digital Technologies in Mental Health Systems: A Qualitative Systematic Review to Inform a Policy Framework," *BMC Health Services Research* 24 (1) (2024): 243, https://doi.org/10.1186/s12913-023-10536-1.

191. Daisy R. Singla, Richard K. Silver, Simone N. Vigod, et al., "Task-Sharing and Telemedicine Delivery of Psychotherapy to Treat Perinatal Depression: A Pragmatic, Non-inferiority Randomized Trial," *Nature Medicine* 31 (2025): 1214–1224, https://doi.org/10.1038/s41591-024-03482-w.

192. Justus Wolff, Josch Pauling, Andreas Keck, and Jan Baumbach, "The Economic Impact of Artificial Intelligence in Health Care: Systematic Review," *Journal of Medical Internet Research* 22 (2) (2020): e16866, https://doi.org/10.2196/16866.

193. Jo-An Occhipinti, Ante Prodan, William Hynes, et al., "Artificial Intelligence, Recessionary Pressures and Population Health," *Bulletin of the World Health Organization* 103 (2) (2025): 155–163, https://doi.org/10.2471/blt.24.291950.

194. Heinz et al., "Evaluating Therabot."

195. Jordan Nelson, Anderson Wills, and Jane Owen, "Barriers to Adoption of AI Assistants in Clinical Practice," January 2025, https://www.researchgate.net/publication/391902530_Barriers_to_Adoption_of_AI_Assistants_in_Clinical_Practice.

196. Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency, "It's Only a Computer: Virtual Humans Increase Willingness To Disclose," *Computers in Human Behavior* 37 (2014): 94–100, https://doi.org/10.1016/j.chb.2014.04.043.

197. John W. Ayers, Adam Poliak, Mark Dredze, et al., "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA Internal Medicine* 183 (6) (2023): 589–596, https://doi.org/10.1001/jamainternmed.2023.1838.

198. Cheryl M. Corcoran, Vijay A. Mittal, Carrie E. Bearden, et al., "Language as a Biomarker for Psychosis: A Natural Language Processing Approach," *Schizophrenia Research* 226 (2020): 158–166, https://doi.org/10.1016/j.schres.2020.04.032.

199. Edward L. Deci and Richard M. Ryan, "Self-Determination Theory: A Macrotheory of Human Motivation, Development, and Health," *Canadian Psychology/Psychologie canadienne* 49 (3) (2008): 182–185, https://doi.org/10.1037/a0012801.

200. Matthew P. Somerville, Helen MacIntyre, Amy Harrison, and Iris B. Mauss, *What Science Has Shown Can Help Young People with Anxiety and Depression: Identifying and Reviewing the "Active Ingredients" of Effective Interventions: Part 2* (Wellcome, 2022), https://doi.org/10.5281/zenodo.7327296.

201. Rob Saunders, Jae Won Suh, Joshua E. J. Buckman, et al., "Effectiveness of Psychological Interventions for Young Adults versus Working-Age Adults: A Retrospective Cohort Study in a National Psychological Treatment Programme in England," *The Lancet Psychiatry* 12 (9) (2025): 650–659, https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(25)00207-X/fulltext.

202. Heinz, Mackin, Trudeau, et al., "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment"; and Chen Chen, Kok Tai Lam, Ka Man Yip, et al., "Comparison of an AI Chatbot With a Nurse Hotline in Reducing Anxiety and Depression Levels in the General Population: Pilot Randomized Controlled Trial," *JMIR Human Factors* 12 (2025): e65785, https://humanfactors.jmir.org/2025/1/e65785.

203. Arthur Kleinman, Hongtu Chen, Sue E. Levkoff, et al., "Social Technology: An Interdisciplinary Approach to Improving Care for Older Adults," *Frontiers in Public Health* 9 (2021): 729149, https://doi.org/10.3389/fpubh.2021.729149.

204. Ibid.; Laura Sampson, Laura D. Kubzansky, and Karestan C. Koenen, "The Missing Piece: A Population Health Perspective to Address the U.S. Mental Health Crisis," *Dædalus* 152 (4) (2023): 24–44, https://doi.org/10.1162/daed_a_02030; Jonathan M. Metzl, "The Protest Psychosis & the Future of Equity & Diversity Efforts in American Psychiatry," *Dædalus* 152 (4) (2023): 92–110, https://doi.org/10.1162/daed_a_02033; Gary Belkin, "Democracy Therapy: Lessons From ThriveNYC," *Dædalus* 152 (4) (2023): 111–129, https://doi.org/10.1162/daed_a_02034; Joseph P. Gone, "Indigenous Historical Trauma: Alter-Native Explanations for Mental Health Inequities," *Dædalus* 152 (4) (2023): 130–150, https://doi.org/10.1162/daed_a_02035; and Isaac R. Galatzer-Levy, Gabriel J. Aranovich, and Thomas R. Insel, "Can Mental Health Care Become More Human by Becoming More Digital?" *Dædalus* 152 (4) (2023): 228–244, https://doi.org/10.1162/daed_a_02040.

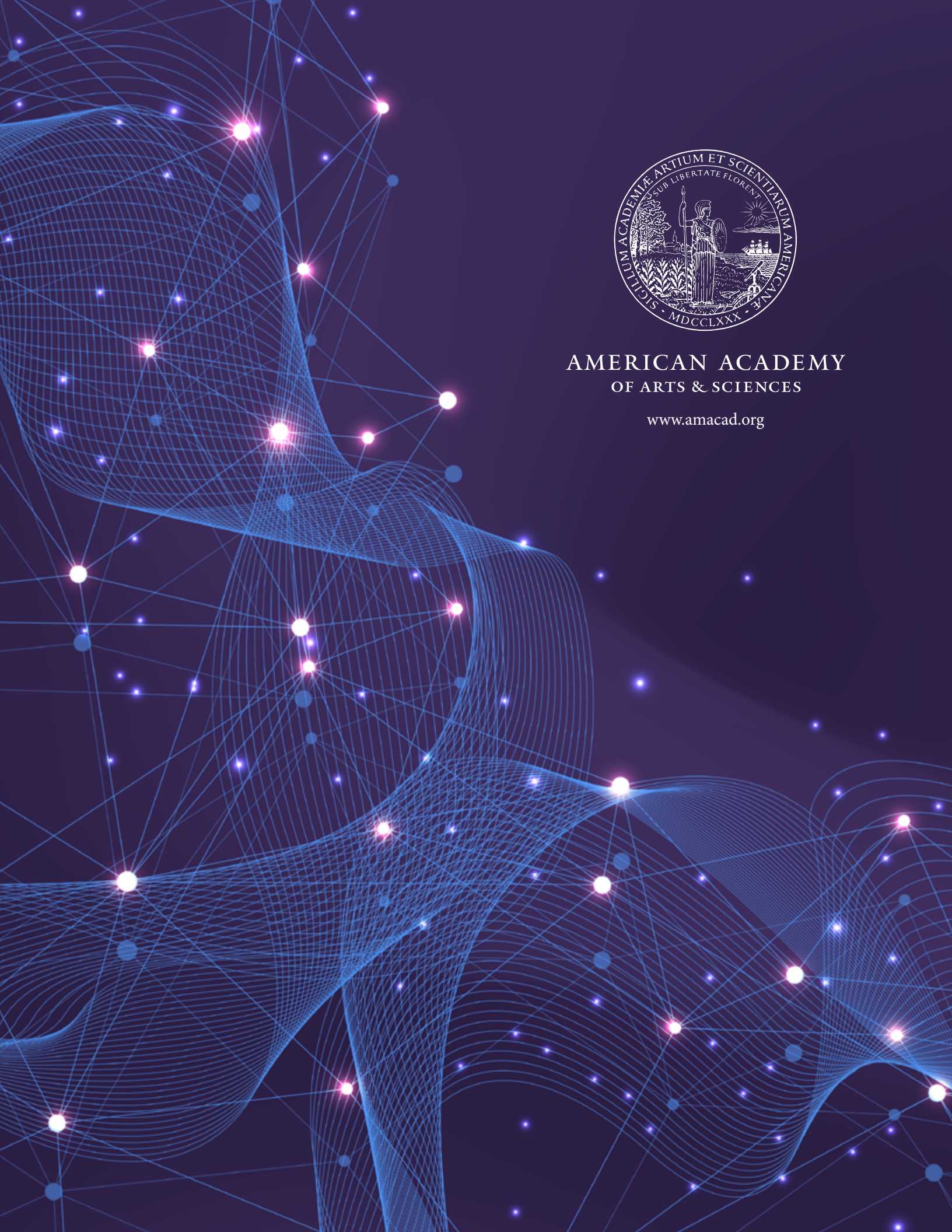
205. Lester Luborsky, Robert Rosenthal, Louis Diguer, et al., "The Dodo Bird Verdict Is Alive and Well—Mostly," *Clinical Psychology: Science and Practice* 9 (1) (2002): 2–12, https://doi.org/10.1093/clipsy.9.1.2.

206. Pablo Cruz-Gonzalez, Aaron Wan-Jia He, Elly PoPo Lam, et al., "Artificial Intelligence in Mental Health Care: A Systematic Review of Diagnosis, Monitoring, and Intervention Applications," *Psychological Medicine* 55 (2025): e18, https://doi.org/10.1017/s0033291724003295.

207. Hassan Auf, Petra Svedberg, Jens Nygren, Monika Nair, and Lina E. Lundgren, "The Use of AI in Mental Health Services to Support Decision-Making: Scoping Review," *Journal of Medical Internet Research* 27 (2025): e63548, https://doi.org/10.2196/63548.

208. Golden and Aboujaoude, "The Framework for AI Tool Assessment in Mental Health."

209. Gauthier Chassang, Jérôme Béranger, and Emmanuelle Rial-Sebbag, "The Emergence of AI in Public Health Is Calling for Operational Ethics to Foster Responsible Uses," *International Journal of Environmental Research and Public Health* 22 (4) (2025): 568, https://doi.org/10.3390/ijerph22040568.



AMERICAN ACADEMY
OF ARTS & SCIENCES

www.amacad.org