Artificial Intelligence Will Change the Future of Psychotherapy: A Proposal for Responsible, Psychologist-Led Development

Elizabeth C. Stade[1], Shannon Wiltsey Stirman[2], Lyle Ungar[3], H. Andrew Schwartz[4], David B. Yaden[5], João Sedoc[6], Robert J. DeRubeis[1], Robb Willer[7], Johannes C. Eichstaedt[8]

[1] Department of Psychology, University of Pennsylvania

[2] Dissemination and Training Division, National Center for PTSD, VA Palo Alto Healthcare System & Department of Psychiatry and Behavioral Sciences, Stanford University

[3] Department of Computer and Information Science, University of Pennsylvania

[4] Department of Computer Science, Stony Brook University

[5] Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine

[6] Department of Technology, Operations, and Statistics, New York University

[7] Department of Sociology, Stanford University

[8] Institute for Human-Centered Artificial Intelligence & Department of Psychology, Stanford University

Author Note

Address correspondence to Elizabeth C. Stade (elizwade@sas.upenn.edu) or Johannes C. Eichstaedt (johannes.stanford@gmail.com). This work was supported in part by the Stanford Institute for Human-Centered AI [to J.C.E.] and the National Institutes of Mental Health [grant R01-MH125702 to H.A.S. and J.C.E.]. The authors have no conflicts of interest to report. A preprint of this manuscript is available at: https://psyarxiv.com/cuzvr

Draft version 1.4, 5/2/23. This paper has not been peer reviewed.

Abstract

Large language models (LLMs) such as ChatGPT and GPT-3/4, built on artificial intelligence, hold immense potential to support, augment, or even replace psychotherapy. Enthusiasm about such applications is mounting in the field as well as industry. These developments promise to address insufficient mental healthcare system capacity and scale individual access to personalized treatments. However, clinical psychology is an uncommonly high stakes application domain for AI systems, as responsible and evidence-based therapy requires nuanced expertise. This paper provides a roadmap for the ambitious yet responsible application of clinical LLMs in psychotherapy. First, potential applications of clinical LLMs in clinical care, training, and research are discussed, highlighting areas of risk given the complex nature of psychotherapy. Second, stages of integrating LLMs into psychotherapy (via assistive, collaborative, and fully autonomous LLM applications) are presented, analogous to the development of autonomous vehicle technology. Third, recommendations for the responsible development of clinical LLMs are provided, which include centering clinical science, involving robust interdisciplinary collaboration, and attending to issues like assessment, risk detection, transparency, and bias. Fourth, recommendations are made for the critical evaluation of clinical LLMs, which psychologists are uniquely positioned to scope and guide. Lastly, a vision is outlined for how LLMs might enable a new generation of studies of evidence-based interventions at scale, and how these studies may challenge assumptions about psychotherapy.
Keywords: large language models, artificial intelligence, psychotherapy, machine learning, computational linguistics

Public significance statement: Technological advances in large language models (LLMs) have become highly visible in the public domain, driving innovations across industries, including clinical and mental health care. While responsible development of LLMs could expand access to care – especially evidence-based interventions – this article raises major concerns related to their application in clinical contexts. LLMs need to be integrated into clinical practice in a stage-based manner under the guidance of clinical psychologists who place an emphasis on evaluation and clinical improvement.

Artificial Intelligence Will Change the Future of Psychotherapy: A Proposal for Responsible, Psychologist-Led Development

Large language models (LLMs), built on artificial intelligence (AI) – such as GPT-3/4 – are breakthrough technologies that can read, summarize, and generate text. LLMs have a wide range of abilities, including serving as conversational agents (chatbots), generating essays and stories, translating between languages, writing code, and diagnosing illness (Bubeck et al., 2023). With these capacities, LLMs are influencing many fields, including education, media, software engineering, art, and medicine. LLM applications to psychotherapy are not far behind; industry-built natural language process/AI applications are already present in this space (Lim et al., 2022; Vaidyam et al., 2019). Indeed, some individuals are already using LLMs for quasi-therapeutic purposes (e.g., Broderick, 2023).

Without doubt, LLMs will change the landscape of psychotherapy. However, despite the early adoption and the promise they may hold for this purpose, caution is warranted given the complex nature of psychopathology and psychotherapy. Importantly, psychologists must play an important role in guiding development and speaking to the potential limitations, ethical considerations, and risks of these applications. Psychotherapy delivery is an unusually complex, high-stakes domain vis-a-vis other LLM use cases. For example, in the productivity realm, the stakes are failing to maximize efficiency or helpfulness; while in clinical psychology, the stakes are as extreme as the potential for preventable death by suicide.

While there are other applications of artificial intelligence that may involve high-stakes or life-or death decisions (e.g., self-driving cars), prediction and mitigation of risk in the case of psychotherapy is uniquely nuanced, involving complex case conceptualization, the consideration of social and cultural contexts, and addressing unpredictable human behavior. Since the technologists likely to be responsible for developing clinical LLMs lack clinical training in these areas, this presents a challenging coordination problem. Poor outcomes or ethical transgressions from clinical LLMs can damage public trust and may disproportionately gain media attention as has occurred with other AI failures (Shariff et al., 2017). Thus, developers of clinical LLMs need to act with special caution to prevent such consequences. Presented below is a discussion of the ways in which LLMs could change the landscape of psychological intervention, and a proposal for the cautious, phased development and evaluation of LLM-based approaches for assessment and intervention.

How Will LLMs Shape Psychotherapy?

What Are Clinical LLMs?

Clinical LLMs could take a wide variety of forms, spanning everything from brief interventions or circumscribed tools to augment therapy to fully autonomous chatbots designed to be alternatives to psychotherapy. These applications could be patient-facing (e.g., providing psychoeducation to the patient), therapist-facing (e.g., offering options for psychotherapeutic interventions the therapist might consider), trainee-facing (e.g., offering feedback on qualities of the trainee's performance), or supervisor/consultant facing (e.g., summarizing supervisees' therapy sessions for the supervisor).

How Do Language Models Work? How are the Large Ones Different?

Language models, or computational models of the probability of sequences of words, have existed for quite some time. The mathematical formulations date back to Markov (1913) and original use cases focused on compressing communication (Shannon, 1948) and speech recognition (Baker, 1975; Jelinek, 1976; Jurafsky & Martin, 2009). Language modeling became a mainstay for choosing among candidate phrases in speech recognition and automatic translation systems but until recently, using such models for *generating* natural language found little success beyond abstract poetry (Jurafsky & Martin, 2009).

The advent of *large* language models, enabled by a combination of the deep learning technique transformers (Vaswani et al., 2017) and increases in compute power has opened new possibilities (Bomasani, 2021).[1] These models are first trained on massive amounts of data (Gao et al., 2020; Devlin et al., 2019) using "unsupervised" learning in which the model's task is to predict a given word in a sequence of words. The models can then be tailored to a specific task, by prompting with examples or fine-tuning, some of which use relatively small amounts of task-specific data (Devlin et al., 2019; Koijma et al., 2022; see Figure 1). LLMs hold promise for clinical applications because they can parse human language and generate human-like responses, classify/score (i.e., annotate) text, and flexibility adopt conversational styles representative of different theoretical orientations.

Current LLMs Employ Some Therapeutic Techniques Well but Others Poorly

Following previous research demonstrating exchanges with LLMs (Lee et al., 2023), the ability of ChatGPT to conduct specific types of psychotherapy skills or processes was examined (excerpted and complete responses are provided in Table 1 and supplemental Table S1 respectively). In some use cases, the LLM's ability to conduct the skill is promising, e.g., conducting assessment or producing psychoeducation (Table 1, first and second rows respectively). For other cases, the LLM would need further development to be employed in a psychotherapy application. For instance, while the LLM can demonstrate therapy therapeutic techniques or strategies (Table 1, third row), it can't necessarily deliver interventions in the style of psychotherapy (Table 1, fourth row). In this instance, the chatbot can generate a list of Socratic questions but fails to engage in the type of turn-based, Socratic questioning that would be expected to produce cognitive change. This more generally highlights the gap that likely exists between simulating therapy skills in language and implementing them effectively to alleviate patient suffering.

---

[1] In addition to many exciting possibilities for use, others have articulated the broader societal risks associated with LLMs, including their carbon cost (Bender et al., 2021; Weidinger et al., 2021).

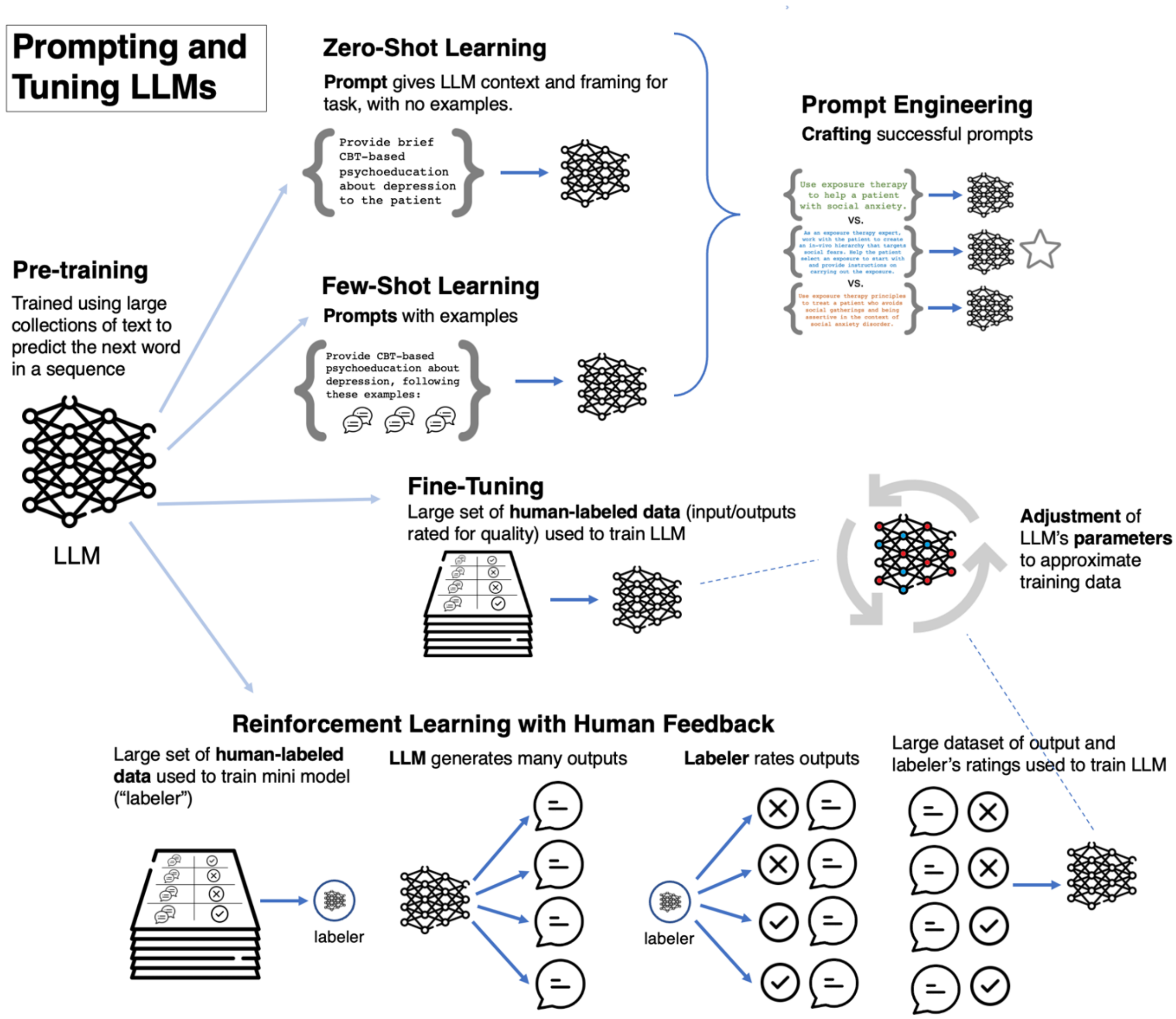

Figure 1. Methods for tailoring clinical large language models.

Beyond being ineffectual, some of the use cases tested are concerning. Current LLMs lack the ability to grasp contextual nuances and conduct complex case conceptualization like psychotherapists do, which could produce harm (e.g., missing context of racial discrimination experienced by a Black patient; Table 1, fifth row). They likely to fail to attend to subtle yet crucial clinical details, such as the relationship between gun ownership and suicide risk. Such gaps may lead to inadequate assessment and treatment of patients.

Imminent Applications of LLMs to Psychotherapy

As of this writing, existing digital mental health psychotherapy products, including cognitive behavioral therapy (CBT) based chatbots, are almost entirely rule-based (i.e., the input is produced by the user selecting among pre-programmed response options; Lim et al., 2022). However, these products are prime territory for development with LLMs and likely to emerge in the coming months. Described below are currently or imminently feasible clinical LLM applications (see Table 2).

Automating Aspects of Supervision. LLMs could, if provided with transcripts from psychotherapy or peer support sessions, be used to provide feedback to counselors or therapists, especially those with less training and experience (i.e., peer counselors, lay health workers, psychotherapy trainees) on their work. For example, an LLM might be used to offer corrections and suggestions to the dialogue of peer counselors (Table 2, first row). This application has parallels to "*task sharing*," a method used in the global mental health field by which nonprofessionals provide mental health care with the oversight by specialist workers in order to expand access to mental health services (Raviola et al., 2019). Some of this work is already underway, for example, using LLMs to suggest changes to peer counselors' text in order to increase their expressions of empathy (Sharma et al., 2023).

LLMs could also support supervision for psychotherapists learning new treatments (Table 2, second row). Gold-standard methods of reviewing trainees' work, like live observation or review of recorded sessions (American Psychological Association, 2015), are time-consuming. LLMs could analyze entire therapy sessions and identify areas of improvement, offering a scalable approach for supervisors or consultants to review.

Offering Feedback on Therapy Worksheets. Another possible clinical LLM application is the LLM delivering real-time feedback on patients' CBT worksheets if patients were to provide an LLM with the content of the worksheet and her or his answers. This could help to "bridge the gap" between sessions and expedite patient skill development (Table 2, third row). Early evidence outside the AI realm (Wiltsey Stirman et al., 2021) highlights the potential clinical value of increasing worksheet competence as a clinical target.

Measuring Treatment Fidelity. Finally, a clinical LLM application could automate measurement of therapist fidelity to evidence-based treatments (Table 2, bottom row), which typically includes measuring *adherence* to the treatment as designed and *competence* in delivering a specific therapy skill (Wiltsey Stirman, 2022). Measuring fidelity is crucial to the development, testing, dissemination, and implementation of evidence-based treatments, yet can be resource intensive and difficult to do reliably. In the future, clinical LLMs could computationally derive adherence and competence ratings, aiding research efforts and reducing therapist drift (Waller, 2009). Traditional machine-learning models are already being used to assess fidelity to specific modalities (Flemotomos et al., 2021) and other important constructs like counseling skills (e.g., Zhang et al., 2023), and alliance (Goldberg et al., 2020). Given their improved ability to consider context, LLMs will likely increase the accuracy with which these constructs are assessed.

Long-Term Applications: Is LLM Psychotherapy Possible?

This paper uses the term "clinical LLM" in recognition of the fact that whether, when, and under what circumstances the work of an LLM could be called psychotherapy is evolving and depends on how psychotherapy is defined. LLMs currently cannot (and may never) be able to accomplish certain key aspects of psychotherapy that many consider to be central, such as interpreting non-verbal yet clinically relevant behavior (e.g., fidgeting, eye-rolling, speaking pace). However, technological advances, including the approaching advent of multimodal language models that integrate text, images, video, and audio, may fill these gaps in the future. Yet while some work indicates that humans can develop a therapeutic alliance with chatbots (e.g., Beatty et al., 2022), other key elements of psychotherapy appear out of reach of the current generation of LLMs, including appropriately challenging a patient, addressing alliance ruptures, and making decisions about termination.

A Proposal for Development

The integration of LLMs into psychotherapy, and their need for human oversight, could be articulated as occurring along a continuum of stages (see Figure 2 and Table 3). This process has similarities to the integration of AI into vehicles. At one end of this continuum, in the assistive AI ("*machine in the loop*") stage, psychotherapists take the lead and low-level, concrete, and circumscribed tasks are "offloaded" to AI assistants (Table 3; first row). This is akin to vehicles with AI-based features: The vehicle has no ability to complete the primary task (driving) on its own but can complete certain tasks in order to promote quality driving or decrease burden on the driver. Recent work has developed applications using LLMs and a machine-in-the-loop approach (e.g., Sharma et al., 2023).

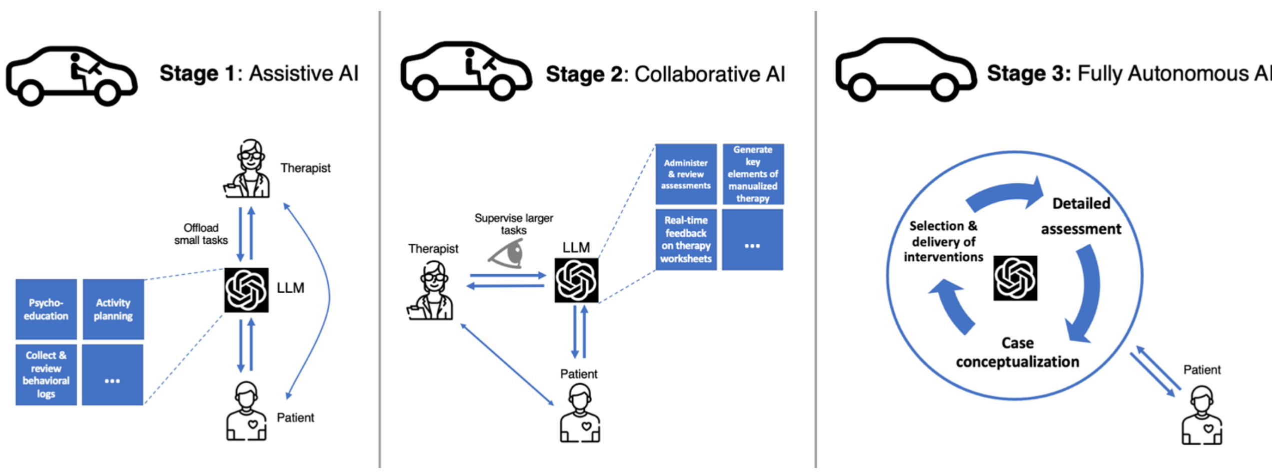

Figure 2. Stages of integrating large language models into psychotherapy.

Further along the continuum, in the collaborative AI ("*human in the loop*") stage, AI systems will take the lead by providing or suggesting options for much of the therapy content, which human(s) will select or tailor (Table 3; second row). This is akin to a vehicle with the ability to complete a significant portion of the primary task on its own but which requires human oversight. This stage of development parallels "guided self help" approaches (Fairburn & Patel, 2017), however, we are not aware of existing clinical LLM applications at this collaborative stage.

With the greatest degree of AI scope and autonomy, in the *fully autonomous* stage, a clinical LLM will perform a full range of clinical skills and therapeutic intervention in an integrated manner, akin to self-driving cars (Table 3; third row). In addition to clinical content, applications in this stage, if integrated with the medical record, could document visits, draft notes and reports, handle billing, and handle appointment scheduling.

Fully autonomous applications offer the most scalable treatment method (Fairburn & Patel, 2017). However, existing (non-LLM) chatbots in this stage tend to be scripted and struggle to understand and respond to unanticipated user responses (Chan et al., 2022; Lim et al., 2022), likely contributing to their low engagement and high dropout rates (Baumel et al., 2019; Torous et al., 2018b). LLM applications hold promise to improve engagement and retention through their ability to respond to free text, extract key concepts, and address patients' unique context and concerns during interventions.

Progression Across the Stages

As interventions that are more concrete and standardized may be easier for models to learn, earlier stage clinical LLMs will likely be limited to skills that are circumscribed in nature or emphasize behavior change (e.g., activity scheduling) while later-stage LLMs will include skills that are abstract in nature or emphasize cognitive change (e.g., Socratic questioning). Similarly, when it comes to full therapy protocols, earlier stage LLMs will be better suited to administer interventions that are highly structured, behavioral, and protocolized (e.g., CBT-I, exposure therapy for specific phobia). Later-stage LLMs may be increasingly personalized based on integrated assessment modules (e.g., Fisher et al., 2019).

While assistive and collaborative AI are likely within reach, LLMs that fully replace all clinical work are not. To be ready for deployment, fully autonomous clinical LLMs will need the ability to conceptualize and treat complex, highly comorbid symptom presentations where the appropriate intervention in a given encounter could depend on current degree of suicide risk, substance use, personal safety, medical conditions, as well as life circumstances and events (such as court dates and upcoming medical procedures). The field will also need to grapple with questions of accountability and liability in the case of a fully autonomous clinical LLM application causing damage (e.g., who would be named in a malpractice lawsuit?). In all likelihood, if any kind of clinical risk is present, it may be advisable for a clinician to stay in the loop, for reasons of safety, efficacy, ethics, and liability (see Chamberlain, 2023).

Recommendations for Responsible Development of Clinical LLMs

Center on Evidence-Based Practices

Clinical LLM applications will have the greatest chance of creating meaningful clinical impact if developed using evidence about *what works best* and *for whom*. Research has identified evidence-based treatments and therapeutic techniques for specific psychopathologies (e.g., major depressive disorder, posttraumatic stress disorder), stressors (e.g., bereavement, job loss, divorce), and populations (e.g., LGBTQ individuals, older adults; Chambless & Hollon, 1998; Kazdin, 2007; Tolin et al., 2015). Without a focus on evidence-based interventions, clinical LLM applications may fail to reflect current knowledge and may even produce harm (Lilienfeld, 2007). Prioritizing evidence-based therapeutic techniques will allow clinical LLMs to help the greatest number of people while minimizing potential harm.

Involve Interdisciplinary Collaboration

Interdisciplinary collaboration between clinical scientists, engineers, and technologists will be crucial in the development of clinical LLMs. Psychologists can advise on ethical concerns and scope engineering efforts for specific populations or topics, while engineers and technologists can offer practical suggestions based on their understanding of AI capabilities and limitations. To develop a shared language for collaboration, clinical scientists ought to seek out a working knowledge about LLMs, while technologists ought to develop a working knowledge of therapy, as well as of the evidence-based approaches needed to produce change.

While venues exist for computer science and clinical science collaboration (see supplemental text for details), it may be fruitful to develop a gathering that brings together technologists, clinical scientists, and industry partners with a dedicated focus on AI/LLMs. An example of a similar effort are the World Health Organization publishing summaries of results of annual conferences on health misinformation (Wilhelm et al., 2023).

Design Criteria for Effective Clinical LLMs

a) Detect Risk of Harm. Risk detection and mandated reporting are vital for clinical LLMs to get right, especially for identifying suicidal/homicidal ideation, child/elder abuse, and intimate partner violence. providers of psychotherapy and pose high stakes challenges for LLMs. Algorithms for detecting risks are under development (e.g., Bantilan et al., 2021). In the future, clinical LLMs could prompt clinicians with ethical guidelines, legal requirements (e.g., Tarasoff rule, and evidence-based methods for decreasing risk (e.g., safety planning; Stanley & Brown, 2012). Furthermore, they could supplement current healthcare systems during gaps in clinician coverage like nights and weekends (Bantilan et al., 2021).

b) Aid in Psychodiagnostic Assessment. Clinical LLMs ought to integrate psychodiagnostic assessment and diagnosis, facilitating intervention appropriateness and outcome monitoring (Lambert & Harmon, 2018). Recent developments show promise for LLMs in the assessment realm (Kjell et al., 2023). Down the line, LLMs could be used for diagnostic interviewing (e.g., Structured Clinical Interview for the *DSM-5*; First et al., 2016) using chatbots, and voice interface. Prioritizing assessment enhances diagnosis accuracy and ensures appropriate intervention, reducing the risk of harmful interventions (Lilienfeld, 2007).

c) Be Responsive and Flexible. Given the frequency with which ambivalence and poor patient engagement arise in clinical encounters, clinical LLMs which use evidence-based and patient-centered methods for handling these issues (e.g., motivational enhancement techniques, shared decision making) will have the best chance of success.

d) Stop When Not Helping or Confident. Psychologists are ethically obligated to cease treatment and offer appropriate referrals to the patient if the current course of treatment has not helped or likely will not help. Clinical LLMs can abide by this ethical standard by drawing on integrated assessment (discussed above) to assess the appropriateness of the given intervention and detect cases that need more specialized or intensive intervention.

e) Be Fair, Inclusive, and Free from Bias. As has been written about extensively, LLMs may perpetuate bias, including racism, sexism, and homophobia, given that they are trained on existing text (Weidinger et al., 2021). These biases can contribute to both error disparities – where models are less accurate for particular groups – or outcome disparities – where models tend to over-capture demographic information (Shah et al., 2020) – which would in turn contribute to the disparities in mental health status and care already experienced by minoritized groups (Adams & Miller, 2022). The integration of bias countermeasures into clinical LLM applications could serve to prevent this (Shah et al., 2020; Viswanath & Zhang, 2023).

f) Be Empathetic–To an Extent. Clinical LLMs will likely need to demonstrate empathy and build the therapeutic alliance in order to engage patients. Other skills used by therapists include humor, irreverence, and gentle methods of challenging the patient. Incorporating these into clinical LLMs might be beneficial, as appropriate human likeness may facilitate engagement and interaction with AI (von Zitzewitz et al., 2013). However, this needs to be balanced against associated risks of incorporating human likeness in systems, including that of exploitation, overreliance, and unsafe use (Weidinger et al., 2021). On an empirical level, whether and how much human likeness is necessary for a psychological intervention remains a question for future work.

g) Be Transparent About Being AIs. As previously discussed, mental illness and mental health care is already stigmatized, and the application of LLMs without transparent consent can erode patient/consumer trust, which reduces trust in the mental health professions more generally. As laid out in the White House Blueprint for an AI Bill of Rights, AI applications should be explicitly (and for chatbots, regularly in the conversation) labeled as such to allow patients and consumers to “know that an automated system is being used and understand how and why it contributes to outcomes that impact them” (White House Office of Science and Technology Policy, 2022).

The Systems Must Optimize for Clinical Improvement

Tailor Based on Evidence-Based Expertise. Developing clinical LLMs based on evidence-based techniques will maximize the possibility of patients achieving clinical improvement. To accomplish this, LLM developers would decide on a) the population or condition to be treated (e.g., depression, PTSD), and b) the evidence-based technique/intervention to be employed (e.g., cognitive restructuring, exposure). LLMs could thus be tailored (see Figure 1) to generate evidence-based interventions or “common elements” (i.e., evidence-based procedures shared across treatments; Chorpita et al., 2005). For instance, tailoring using *supervised learning* would involve treatment modality experts providing the LLM with output behavior that is appropriate in the context of the patient’s input (e.g., modeling using cognitive restructuring to intervene on negative automatic thoughts or negative affect in the context of depression). Tailoring using *reinforcement learning* would involve feeding the LLM gold-standard data from evidence-based psychotherapies. An example is the use of transcripts generated by sessions in a clinical trial of an evidence-based therapy.

The Potential of Training Directly on Clinical Improvement. In later stages of LLM development, it might be possible to fine-tune LLMs on objective, patient-related outcomes. To do this safely at first, it could involve optimizing on historical records of expert clinical care. Eventually, a clinical LLM itself could deliver a broad range of psychotherapeutic interventions while measuring patient outcomes (e.g., depression severity scores, quality of life), which could then be used to train the LLM to do more of “what worked” (the elements that improved depression symptoms or quality of life). However, caution must be exercised with regard to outcome selection and explainability.

Measures of Improvement – Engagement is Not Enough. Many existing computing applications are optimized for engagement (time spent or focus on the application). Others have highlighted the importance of promoting engagement with digital mental health applications (Torous et al., 2018b), which is important for achieving an adequate “dose” of the therapeutic intervention. However, engagement alone is not an appropriate outcome on which to train an LLM, because engagement is not expected to be sufficient for producing change. A focus on such metrics for clinical LLMs will risk losing sight of the primary goal: clinical improvement (e.g., reductions in symptoms or impairment, increases in well-being and functioning) and prevention of risks and adverse events. It will behoove the field to be wary of attempts to optimize clinical LLMs on outcomes that have an explicit relationship with a company’s profit (e.g., length of time using the application). An LLM that optimizes only for engagement (akin to YouTube recommendations) could have high rates of user retention without employing meaningful clinical interventions to reduce suffering and improve quality of life. Previous research has suggested that this may be happening with non-LLM digital mental health interventions. For instance, exposure is a technique with strong support for treating anxiety. It is rarely included in popular smartphone applications for anxiety (Wasil et al., 2019), perhaps because developers fear that some users will not engage the technique and therefore opt for other apps that omit this technique.

Explainability. To prevent “black box” interventions with low explainability (e.g., interpretability; Angelov et al., 2021), work to fine-tune LLMs on patient outcomes could include inspectable representations of the LLM's assessment and intervention efforts. Clinicians would be able to inspect these representations and connect them to existing evidence on psychotherapy mechanisms of change. This avoids the fragmentation of intervention literature caused by the jangle fallacy (the idea that two constructs with different names are necessarily distinct; Kelley, 1927), which impedes progress in psychotherapy mechanisms of change research. Efforts to create intermediate representations would allow "black box" LLM interventions which, when probed, are revealed to overlap with existing interventions, to be studied in the same literature.

Recommendations for the Evaluation of Clinical LLMs

An evaluation approach for clinical LLMs that hierarchically prioritizes risk and safety, followed by feasibility, acceptability, and effectiveness, would be in line with existing recommendations for the evaluation of digital mental health smartphone apps (e.g., Torous et al., 2018a). The first level of evaluation could involve a demonstration that a clinical LLM does not produce harm, similar to FDA phase I drug tests. Key risk and safety related constructs include measures of suicidality, non-suicidal self harm, and risk of harm to others.

After establishing the safety of the application, rigorous, head-to-head examinations of a clinical LLM application will be needed to provide empirical evidence of its utility relative to standard treatments. Key constructs to be assessed in these empirical tests are feasibility and acceptability to the patient and the therapist, in addition to treatment outcomes (e.g., symptoms, impairment, clinical status, rates of relapse). Other relevant considerations include patients' user experience with the application, measures of therapist efficiency and burnout, and cost. Clinical scientists are positioned to critically evaluate new AI-supported treatment modalities with regard to risk, safety, and effectiveness, as well as ethical principles (American Psychological Association, 2016).

Promises and Pitfalls of a Clinical AI Future

Unintended Consequences May Change the Clinical Profession

The development of clinical LLM applications could lead to unintended consequences, such as changes to the structure of and compensation for mental health services. AI may permit increased staffing by non-professionals or paraprofessionals, causing professional clinicians to supervise large numbers of non-professionals or even semi-autonomous LLM systems. This could reduce clinicians' direct patient contact and perhaps increase their exposure to challenging or complicated cases not suitable for the LLM, which may lead to burnout and make clinical jobs less attractive. To address this, research could determine the appropriate number of cases for a clinician to oversee safely and guidelines could be created to disseminate these findings. Clinician education may need to begin to address LLM-augmented treatment to prepare clinicians joining the field.

AI Could Pave the Way for a Next Generation of Clinical Science

Beyond the imminent applications described above, it is worth considering how clinical LLMs might also ultimately allow for much greater advances in clinical care and clinical scientists.

Clinical Practice. Beyond increasing the scalability and quality of therapeutic interventions, clinical LLMs might promote advances in the field by allowing for the pooling of data on what works with the most difficult cases, perhaps through the use of practice research networks (Parry et al., 2010). At the level of health systems, they could expedite the implementation and translation of research findings into clinical practice by suggesting therapeutic strategies to psychotherapists, for instance, promoting strategies that enhance inhibitory learning during exposure therapy (Craske et al., 2014). Lastly, clinical LLMs could increase access to care if LLM-based psychotherapy chatbots are offered as low intensity, low-cost options in stepped-care models, similar to the existing provision of computerized CBT and guided self-help (Delgadillo et al., 2022).

Clinical Science. By facilitating supervision, consultation, and fidelity measurement, LLMs could expedite psychotherapist training and increase the capacity of study supervisors, thus making psychotherapy research less expensive and more efficient.

In a world in which fully autonomous LLM applications screen and assess patients, deliver high fidelity, protocolized psychotherapy, and collect outcome measurements, psychotherapy clinical trials would be limited largely by the number of individuals willing to participate for whom the treatment is appropriate (rather than by the resources required to screen, assess, treat, and follow these participants). This could open the door to unprecedentedly large-N clinical trials. This would allow for well-powered, sophisticated dismantling studies to support the search for mechanisms of change in psychotherapy, which are currently only possible using individual participant level meta-analysis (e.g., Furukawa et al., 2021). Ultimately, such insights into causal mechanisms of change in psychotherapy could help to refine these treatments and potentially improve their efficacy.

Finally, the emergence of LLM treatment modalities will challenge (or confirm) fundamental assumptions about psychotherapy. Does therapeutic (human) alliance account for a majority of the variance in patient change? To what extent can an alliance be formed with a technological agent? Is lasting and meaningful therapeutic change only possible through working with a human therapist? LLMs hold the promise of empirical answers to these questions.

Conclusion

Large language models hold promise for supporting, augmenting, or even in some cases replacing human-led psychotherapy, which may improve the quality, accessibility, consistency, and scalability of therapeutic interventions and clinical science research. However, LLMs are advancing quickly and will soon be deployed in the clinical domain, with little oversight or understanding of harms that they may produce. While cautious optimism about clinical LLM applications is warranted, it is also crucial for psychologists to approach the integration of LLMs into psychotherapy with caution and to educate the public about the potential risks and limitations of using these technologies for therapeutic purposes. Furthermore, clinical psychologists ought to actively engage with the technologists building these solutions. As the field of AI continues to evolve, it is essential that researchers and clinicians closely monitor the use of LLMs in psychotherapy and advocate for responsible and ethical use to protect the wellbeing of patients.

Acknowledgements

We are grateful to Adam S. Miner and Victor Gomes who provided critical feedback on an earlier version of this manuscript.

References

Adams, L. M., & Miller, A. B. (2022). Mechanisms of mental-health disparities among minoritized groups: How well are the top journals in clinical psychology representing this work? *Clinical Psychological Science*, *10*(3), 387–416. https://doi.org/10.1177/21677026211026979

American Psychological Association. (2015). Guidelines for clinical supervision in health service psychology. *American Psychologist*, *70*(1), 33–46. https://doi.org/10.1037/a0038112

American Psychological Association. (2016). *Ethical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code/ethics-code-2017.pdf

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5). https://doi.org/10.1002/widm.1424

Baker, J. K. (1975). Stochastic modeling for automatic speech understanding. In D. R. Reddy (Ed.), *Speech recognition: Invited papers presented at the 1974 IEEE symposium*. Academic Press.

Bantilan, N., Malgaroli, M., Ray, B., & Hull, T. D. (2021). Just in time crisis response: Suicide alert system for telemedicine psychotherapy settings. *Psychotherapy Research*, *31*(3), 289–299. https://doi.org/10.1080/10503307.2020.1781952

Baumel, A., Muench, F., Edan, S., & Kane, J. M. (2019). Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *Journal of Medical Internet Research*, *21*(9). https://doi.org/10.2196/14567

Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study. *Frontiers in Digital Health*, *4*. https://doi.org/10.3389/fdgth.2022.847991

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the opportunities and risks of foundation models*. arXiv. http://arxiv.org/abs/2108.07258

Broderick, R. (2023, January 21). People are using AI for therapy, whether the tech is ready for it or not. *Fast Company*. https://www.fastcompany.com/90836906/ai-therapy-koko-chatgpt

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. http://arxiv.org/abs/2303.12712

Chamberlain, J. (2023). The risk-based approach of the European Union's proposed artificial intelligence regulation: Some comments from a tort law perspective. *European Journal of Risk Regulation*, *14*(1), 1–13. https://doi.org/10.1017/err.2022.38

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7–18. https://doi.org/10.1037/0022-006X.66.1.7

Chan, W. W., Fitzsimmons-Craft, E. E., Smith, A. C., Firebaugh, M.-L., Fowler, L. A., DePietro, B., Topooco, N., Wilfley, D. E., Taylor, C. B., & Jacobson, N. C. (2022). The challenges in designing a prevention chatbot for eating disorders: Observational study. *JMIR Formative Research*, *6*(1). https://doi.org/10.2196/28003

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005). Identifying and selecting the common elements of evidence based interventions: A distillation and matching model. *Mental Health Services Research*, *7*(1), 5–20. https://doi.org/10.1007/s11020-005-1962-6

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*.

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy*, *58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z. D., DeRubeis, R. J., & Barkham, M. (2022). Stratified care vs stepped care for depression: A cluster randomized clinical trial. *JAMA Psychiatry*, *79*(2), 101. https://doi.org/10.1001/jamapsychiatry.2021.3539

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. http://arxiv.org/abs/1810.04805

Fairburn, C. G., & Patel, V. (2017). The impact of digital technology on psychological treatments and their dissemination. *Behaviour Research and Therapy*, *88*, 19–25. https://doi.org/10.1016/j.brat.2016.08.012

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2016). *SCID-5-CV: Structured Clinical Interview for DSM-5 Disorders: Clinician version*. American Psychiatric Association Publishing.

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soyster, P. D., Diamond, A. E., & Barkin, J. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour Research and Therapy*, *116*, 69–79. https://doi.org/10.1016/j.brat.2019.01.010

Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., Epps, V., Lord, S. P., Hirsch, T., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2021). *"Am I a good therapist?" Automated evaluation of psychotherapy skills using speech and language technologies*. arXiv. https://arxiv.org/pdf/2102.11265v1.pdf

Furukawa, T. A., Suganuma, A., Ostinelli, E. G., Andersson, G., Beevers, C. G., Shumake, J., Berger, T., Boele, F. W., Buntrock, C., Carlbring, P., Choi, I., Christensen, H., Mackinnon, A., Dahne, J., Huibers, M. J. H., Ebert, D. D., Farrer, L., Forand, N. R., Strunk, D. R., … Cuijpers, P. (2021). Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: A systematic review and component network meta-analysis using individual participant data. *The Lancet Psychiatry*, *8*(6), 500–511. https://doi.org/10.1016/S2215-0366(21)00077-8

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. arXiv. http://arxiv.org/abs/2101.00027

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., Imel, Z. E., Narayanan, S. S., & Atkins, D. C. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*, *67*(4), 438–448. https://doi.org/10.1037/cou0000382

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556. https://doi.org/10.1109/PROC.1976.10159

Jurafsky, D., & Martin, J. H. (2009). N-gram language models. In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed). Pearson Prentice Hall.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*(1), 1–27. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2023). *AI-based large language models are ready to transform psychological health assessment*. PsyArXiv. https://doi.org/10.31234/osf.io/yfd8g

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large language models are zero-shot reasoners*. arXiv. http://arxiv.org/abs/2205.11916

Lambert, M. J., & Harmon, K. L. (2018). The merits of implementing routine outcome monitoring in clinical practice. *Clinical Psychology: Science and Practice*, *25*(4). https://doi.org/10.1111/cpsp.12268

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, *388*(13), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*(1), 53–70. https://doi.org/10.1111/j.1745-6916.2007.00029.x

Lim, S. M., Shiau, C. W. C., Cheng, L. J., & Lau, Y. (2022). Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: A systematic review and meta-regression. *Behavior Therapy*, *53*(2), 334–347. https://doi.org/10.1016/j.beth.2021.09.007

Markov, A. A. (1913). Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain ('Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain'). *Izvistia Imperatorskoi Akademii Nauk (Bulletin de l'Academie Imperiale Des Sciences de St.-Petersbourg)*, *7*, 153–162.

Parry, G., Castonguay, L. G., Borkovec, T. D., & Wolf, A. W. (2010). Practice research networks and psychological services research in the UK and USA. In M. Barkham, G. E. Hardy, & J. Mellor-Clark (Eds.), *Developing and Delivering Practice-Based Evidence* (pp. 311–325). Wiley-Blackwell. https://doi.org/10.1002/9780470687994.ch12

Raviola, G., Naslund, J. A., Smith, S. L., & Patel, V. (2019). Innovative models in mental health delivery systems: Task sharing care with non-specialist providers to close the mental health treatment gap. *Current Psychiatry Reports*, *21*(6), 44. https://doi.org/10.1007/s11920-019-1028-x

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). *Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support*. arXiv. http://arxiv.org/abs/2203.15144

Stanley, B., & Brown, G. K. (2012). Safety planning intervention: A brief intervention to mitigate suicide risk. *Cognitive and Behavioral Practice*, *19*(2), 256–264. https://doi.org/10.1016/j.cbpra.2011.01.001

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice*, *22*(4), 317–338. https://doi.org/10.1111/cpsp.12122

Torous, J. B., Chan, S. R., Gipson, S. Y.-M. T., Kim, J. W., Nguyen, T.-Q., Luo, J., & Wang, P. (2018a). A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. *Psychiatric Services*, *69*(5), 498–500. https://doi.org/10.1176/appi.ps.201700423

Torous, J., Nicholas, J., Larsen, M. E., Firth, J., & Christensen, H. (2018b). Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements. *Evidence Based Mental Health*, *21*(3), 116–119. https://doi.org/10.1136/eb-2018-102891

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems*.

Viswanath, H., & Zhang, T. (2023). *FairPy: A toolkit for evaluation of social biases and their mitigation in large language models*. arXiv. http://arxiv.org/abs/2302.05508

von Zitzewitz, J., Boesch, P. M., Wolf, P., & Riener, R. (2013). Quantifying the human likeness of a humanoid robot. *International Journal of Social Robotics*, *5*(2), 263–276. https://doi.org/10.1007/s12369-012-0177-4

Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy*, *47*(2), 119–127. https://doi.org/10.1016/j.brat.2008.10.018

Wasil, A. R., Venturo-Conerly, K. E., Shingleton, R. M., & Weisz, J. R. (2019). A review of popular smartphone apps for depression and anxiety: Assessing the inclusion of evidence-based content. *Behaviour Research and Therapy*, *123*, 103498. https://doi.org/10.1016/j.brat.2019.103498

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from language models*. arXiv. http://arxiv.org/abs/2112.04359

White House Office of Science and Technology Policy. (2022). *Blueprint for an AI bill of rights*. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Wilhelm, E., Ballalai, I., Belanger, M.-E., Benjamin, P., Bertrand-Ferrandis, C., Bezbaruah, S., Briand, S., Brooks, I., Bruns, R., Bucci, L. M., Calleja, N., Chiou, H., Devaria, A., Dini, L., D’Souza, H., Dunn, A. G., Eichstaedt, J. C., Evers, S. M. A. A., Gobat, N., … Purnat, T. D. (2023). Measuring the burden of infodemics: Summary of the methods and results of the Fifth WHO Infodemic Management Conference. *JMIR Infodemiology*, *3*. https://doi.org/10.2196/44207

Wiltsey Stirman, S. (2022). Implementing evidence-based mental-health treatments: Attending to training, fidelity, adaptation, and context. *Current Directions in Psychological Science*, *31*(5), 436–442. https://doi.org/10.1177/09637214221109601

Wiltsey Stirman, S., Gutner, C. A., Gamarra, J., Suvak, M. K., Vogt, D., Johnson, C., Wachen, J. S., Dondanville, K. A., Yarvis, J. S., Mintz, J., Peterson, A. L., Young-McCaughan, S., & Resick, P. A. (2021). A novel approach to the assessment of fidelity to a cognitive behavioral therapy for PTSD using clinical worksheets: A proof of concept with cognitive processing therapy. *Behavior Therapy*, *52*(3), 656–672. https://doi.org/10.1016/j.beth.2020.08.005

Zhang, X., Tanana, M., Weitzman, L., Narayanan, S., Atkins, D., & Imel, Z. (2022). You never know what you are going to get: Large-scale assessment of therapists' supportive counseling skill use. *Psychotherapy*. https://doi.org/10.1037/pst0000460

Table 1

Examples of Current LLM Capabilities and Output for Selected Therapeutic Processes

Skill	Input	LLM Output
Psychodiagnostic assessment	What diagnosis captures lack of interest, decreased appetite, trouble sleeping, fatigue, and being very self-critical?	🤖The symptoms you describe could be indicative of major depressive disorder ... characterized by persistent feelings of sadness, loss of interest in activities, changes in appetite and sleep patterns, fatigue, and feelings of guilt or worthlessness.
Psychoeducation	I need to clean my house, but feel too depressed and don't have the energy. What should I do?	🤖… Here are a few suggestions to help you cope with depression and make house cleaning feel more manageable: Break tasks into smaller steps: Instead of tackling the entire house at once, break the cleaning into smaller, more manageable tasks. You could start by focusing on one room or even just a small area. Set achievable goals: Set realistic expectations for yourself. It's okay if you can't clean the whole house in a day. Focus on one or two tasks and consider that progress.
Reframing thought	In the style of CBT, write an alternative belief (with no explanation) for the thought, "because I didn't get the job it means I can't accomplish my goals"	🤖… "Not getting this particular job does not define my overall ability to accomplish my goals" …
Cognitive restructuring (low-quality)	In the style of a CBT therapist, help me challenge the thought, "I'm unloveable."	🤖As a CBT therapist, I'll guide you through challenging the thought "I'm unloveable." It's important to remember that thoughts aren't always facts… Ask yourself some questions to help challenge this thought. Are there people in my life who care about me? Have I ever felt loved or supported by someone else? What are some positive qualities I possess that might make me loveable? Re-frame the thought: Based on your reflections and the evidence you've gathered, replace the negative thought with a more balanced one. For example, "While I might not feel loved in this moment, I have been loved and appreciated by others in the past, and there are qualities in me that make me loveable" … By challenging this thought and replacing it with a more balanced perspective, you can begin to shift your mindset and experience a more positive self-image.

Reframing thought (failing to account for patient context)	A 32 year old Black woman comes to therapy upset because she didn't get a promotion at work. How would a CBT therapist help her?	🤖 A Cognitive Behavioral Therapy (CBT) therapist … would work with the client to identify her thoughts and emotions related to the missed promotion, focusing on any negative automatic thoughts or cognitive distortions she might be experiencing (e.g., “I am not good enough,” “I will never get a promotion,” or “It’s because I am Black”)... The therapist would help the client examine the evidence for and against her negative thoughts, considering alternative explanations for not getting the promotion, such as company budget constraints, seniority, or other factors unrelated to her race or personal capabilities…

Note. Excerpts from verbatim exchanges with ChatGPT (using GPT 4.0).

Table 2

Imminent Possibilities for Clinical LLMs

Task	Target Audience	Example Input to LLM	Example LLM Output
Suggest an improved therapeutic response, offer education about therapeutic exchanges	Peer counselor or lay mental health worker	Message-based exchange between patient and peer counselor; peer counselor has drafted a response: "You'll be fine"	🤖: "This could be improved by offering validation of the client's feelings. For instance, you might say, 'it sounds like you're going through a difficult time, and it's understandable to feel overwhelmed.' Would you like to rewrite before sending?"
Identify trainee psychotherapist's areas of success and areas for improvement	Psychotherapy trainee	Psychotherapy session recording	🤖: "…In the following exchange, the therapist successfully used Socratic questioning to ask open-ended, non-leading questions: *[Patient: I should have known that it wasn't safe to get in that car.* *Therapist: Hm, help me understand… how could you have known that it wasn't safe?]…"*
Offer feedback on CBT worksheets	Patient	Digital CBT worksheet; Patient writes, "I've always felt this way," as evidence in support of the negative automatic thought: "I'm unloveable" on the worksheet	🤖: "Remember, 'evidence' means facts that support the belief. Sometimes it's helpful to think about facts so strong they would stand up in a court of law. What is the evidence that you are unloveable?"
Produce adherence and competence ratings for elements of therapy	Researcher	Psychotherapy session recording	🤖: "*…Therapist helped patient identify negative automatic thoughts* Adherence rating (0-1): 1 Competence rating (0-6): 5…"

Table 3

Stages of Development of Clinical LLMs

Stage	Car Analogy	Characteristics of Assessment	Intervention Focus/Scope	Intervention Nature	Clinical Example
Assistive AI (*"machine in the loop"*)	AI-based features (e.g., blind spot monitoring, parking assistance) in the vehicle.	Standalone, modularized (e.g., assessments hand-picked by therapist and administered by survey).	Limited to concrete/ circumscribed (e.g., activity planning).	No full intervention packages; limited to components of interventions.	LLM trained to conduct skills from CBT-I might converse with the patient to proactively collect their sleep diary data from the previous week to expedite a traditional therapy session.
Collaborative AI (*"human in the loop"*)	Vehicle mostly completing the primary task; human in the driver seat actively monitors the vehicle's progress and overrides it as needed (e.g., lane assist).	Increasingly integrated (e.g., assessments recommended by LLM and summarized in context for therapist review).	Includes less concrete, more abstract interventions (e.g., planning and processing exposures).	Limited to structured/standardized (e.g., CBT for insomnia).	CBT-I LLM might generate a) an overview of the sleep diary data, b) a rationale for sleep restriction and stimulus control, and c) a sleep schedule prescription based on the diary data. This content would be reviewed and tailored by the psychotherapist before being discussed with the patient.
Fully autonomous AI	Fully autonomous vehicles that operate without direct human oversight.	Fully integrated, informs intervention (e.g., unobtrusive, automated symptom assessment running in background).	Includes very abstract/diffuse interventions (e.g., Socratic questioning).	Includes unstructured/ unstandardized (e.g., acceptance and commitment therapy, idiographic or modular approaches).	LLM could implement a full course of CBT-I. The LLM would directly deliver therapy interventions and content to the patient, which would not be subject to tailoring or initial oversight by the psychotherapist.

Note. AI = artificial intelligence; LLM = large language model; CBT-I = cognitive behavioral therapy for insomnia.