The Role of Machine Learning in Shaping Mental Health Care

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Franz Charles Silva

Spring 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

William J Davis , Department of Engineering and Society

Introduction

In the 1960s, computer scientist Joseph Weizenbaum introduced the ELIZA chatbot, utilizing natural language processing to simulate an interaction with a psychotherapist (Berry, 2023, p. 2). Met with praise and promise, Weizenbaum's creation was an introduction to the impact machine learning would later have on the field of psychotherapy. Experts agree that "ML can significantly improve the detection and diagnosis of mental health conditions" while also providing access to those facing barriers with in-person therapy (Shatte et al., 2019, p. 1448). However, the ethical concerns regarding data privacy, algorithmic bias, and the autonomy of the technology raised in the 1960s are still present today.

Mental Health America revealed that 23% of adults experienced a mental illness in the past year, with 25% of adults facing frequent mental distress who were unable to seek treatment due to cost (Mental Health America, 2024). The emergence of machine learning could play a vital role in bridging the gap for those unable to currently access mental health treatment, but the nuanced and deeply personal nature of psychotherapy raises questions about the extent to which technologies can assist or potentially replace human therapists (Roth et al., 2021, p. 20).

The impressive accuracy of new Large Language Models (LLMs) and different machine learning algorithms could risk dehumanizing the process of psychotherapy by forcing patients to provide their data to machine learning models. If mental health ML technologies are forcibly adopted without clear boundaries or regulations, it may lead to a societal shift pushing for the development of new machine learning technologies and simultaneously deterring users away from seeking professional treatment (De Choudhury et al., 2016, p. 14). In this paper, I argue that the adoption of machine learning technologies within mental health care, specifically chatbots and predictive modeling, contributes to the datafication of human experiences, forcing humans to

interact with machine learning technologies to seek treatment. This paper will use a technological momentum framework to help assess how the initial control that society has over machine learning in mental health care evolves over time. I will examine how these ML-based technologies can create a feedback loop which compromises the privacy of human data and influences patient perspectives on psychotherapy, shifting power to technology developers while reducing human autonomy in mental health treatment.

What We Know and Do Not Know

The rise of machine learning within mental health care calls for research on its potential to enhance diagnostic accuracy and increase accessibility for treatment. While current research emphasizes the accuracy and effectiveness of machine learning models, significantly less attention is focused on the broader ethical and societal consequences of integrating these technologies into psychotherapy (Thieme et al., 2020, p. 34-38). The ongoing use of machine learning in mental health treatment, using technologies like chatbots and predictive models, raises concerns about data integrity and accountability, both of which remain less important in existing research (Gooding & Kariotis, 2021, p. 10).

Chatbots

Chatbots, or systems capable of communicating with human users, are currently being studied for their usability and feasibility outside of a research setting. Numerous literature reviews have been performed which highlight different studies that show there is potential for chatbots in improving mental well-being and aiding in preventative care, but also present challenges of usability and real-world integration within existing healthcare systems (Casu et al., 2024, p. 19; Abd-Alrazaq et al., 2020, p. 13).

The usage of chatbots can create a sense of distrust in mental health systems, as users prefer transparency and data privacy while interacting with these systems in fear of the hidden usage of machine learning technologies (Richards et al., 2023, p. 13). Although chatbots seemingly can help increase accessibility to those unable to attend in-person treatment, the forced interaction with a chatbot rather than a human can, according to Richards, cause distrust in the system itself and deter users from seeking treatment.

Predictive Modeling

The primary application of machine learning in mental health is predictive modeling, where algorithms are used to detect patterns in data and predict mental health outcomes. Predictive models are trained on large amounts of data, allowing them to identify correlations that might not be obvious to clinicians or therapists.

An example of this is suicide prevention, where models analyze patterns in behavior, social media data, and medical history to assess the likelihood of a person attempting suicide (Belsher et al, 2019, p. 642). Nearly half of individuals who commit suicide communicate their suicidal intentions to others (Pompili et al., 2016, p. 2240), but could this number increase if mentions of suicide within social media were also included? The effectiveness of these models ultimately depend on the quality of data collected and used within the training process. A literature review conducted on technology and risk-assessment for suicide prevention showed that “teenagers and young adults often fail to disclose risk factors to physicians, despite sharing them with the public on social media platform” (Pourmand et al., 2019, p. 880), suggesting that having access to a patient’s social media can assist in identifying suicidal acts. This focus of research on predictive modeling leads to an issue of data retrieval, as human data like emotions, personal information, and images scraped from the web can all be used to train these models.

Humans as Data Points

How much consent do humans have in preventing the datafication of their identity? Regulations regarding the retrieval of online data are sparse. There are little to no regulations that clearly state the limitations regarding using social media data, particularly when retrieving information that is publicly available (Wongkoblap et al., 2017, p. 8). As machine learning technologies grow and demand more data, how can we, as humans, preserve our autonomy and privacy online?

The datafied approach to machine learning, which reduces human emotions and humans themselves to data points used for training machine learning models, raises ethical concerns on what kinds of data should be collected and the control patients have over their data. For example, Instagram data (photos, like count, comments) can be used to extract features that can distinguish posts between depressed and nondepressed users (Reece & Danforth, 2017, p. 1). Should an image that someone posts online automatically be subject to usage for algorithms like these, which could benefit from the mass volume of available images on the internet? Or should users have control over the usage of their own images, comments, and social media?

Currently, the mining of data through the internet is not viewed as copyright infringement and "has been stretched beyond its limits to fit certain technological contexts" (Quang, 2021, p. 1435), meaning that social media data could be currently used for training in machine learning models like predictive models for mental health disorders. However, Eichstaedt et al addresses that the incorporation of social media data used for health reasons could change the ways users interact with social media in the first place (2018, p. 12205). If the user knows their social media is being monitored and input into training data for models like predictive models, they could change their social media presence on purpose to divert alerts away from their potential risk of a

disorder. As a result, predictive models that rely heavily on social media data may decrease in quality and accuracy, as they could be trained on intentionally curated or misleading posts from individuals who would have otherwise posted with genuine indicators of mental health status. This creates a dilemma of either prioritizing the privacy of data, leading to potentially less reliable predictions, or continuing the current path of using publicly available data and assuming users are aware of the implicit consent of posting online, while also raising concerns about responsibility. If a predictive model misidentifies an individual's mental health status due to self-manipulated social media activity, is the fault on the individual, the creator of the algorithm, or the algorithm itself?

Accountability

While outputs of these technologies appear accurate within research, the "black-box" nature of machine learning models leads to issues of accountability and responsibility in the event of a misprediction. Certain models, like deep neural networks, lack interpretability and make it difficult to pinpoint the specific reasons behind their predictions (Durstewitz et al, 2019, p. 1591). This lack of transparency poses a challenge when a model misdiagnoses a patient, as who is held responsible is not clearly or universally defined.

The consequences of a misprediction can be severe. Misidentifying someone as suicidal or at risk of a mental health crisis can seriously impact their self-image and potentially their employment (Thieme et al, 2020, p.34-38). The responsibility of this misprediction can be attributed to stakeholders like the algorithm, the creator of the algorithm, and the patient themselves, but since it is almost never explicitly defined, accountability becomes an issue. Without clear guidelines on accountability, mispredictions can leave affected individuals without any form of recourse and exacerbate skepticism towards machine learning in mental health.

In summary, current research focuses on the accuracy of prediction models and the potential of bringing machine learning to other aspects of mental health care. Many studies praise its ability to enhance diagnostic precision and increase accessibility to mental health resources, and while other works attempt to assess the technical positives and negatives of machine learning in mental health, rarely do they mention its potential to harm the future of the psychotherapy field. By deterring users from seeking machine learning assisted treatment, reducing personal data and images to mere data points, and removing accountability from machine learning models, the growing use of machine learning risks accelerating the (already occurring) datafication of human experiences. This shift gives power to corporations capable of developing these technologies and simultaneously threatens human autonomy and privacy as technologies demand more data for accuracy, raising urgent questions about how to effectively develop new technologies with all ethical implications involved.

Methods

Previous research on machine learning in mental health care focuses on technical accuracy rather than the broader ethical and societal consequences that result from integrating these technologies. To frame the analysis of machine learning in mental health care, this paper applies the Technological Momentum framework, which examines how technologies, once established, gain a trajectory that shapes the social and institutional structures around them and is difficult to reverse (Hughes, 1987, p.241). This approach acknowledges that social choices can shape the integration of chatbots and predictive analytics within the early stages of development, but also argues that these choices become constrained as the technologies become embedded within systems.

Guiding my analysis, I explored peer-reviewed research, systematic reviews, and meta-analyses focusing on the usage of chatbot-based treatments and predictive modeling in mental health treatment. This collection of sources offers a unique perspective into how the integration of both chatbots and predictive analytics, which provide different technical outcomes, contributes to an ongoing loop of interaction with ML-based technology, potentially worsening the relationship between patients and psychotherapy.

The research examined on chatbots explored their role in providing psychological support, raising concerns about their impact on patient trust and the challenges involved with integrating them into existing healthcare systems (Casu et al., 2024, p. 19; Abd-Alrazaq et al., 2020, p. 13). Studies on predictive modeling have focused on the accuracy of machine learning systems on assessing mental health risks, such as suicide prevention, and the ethical concerns associated with using personal data to feed into these models (Belsher et al., 2019, p. 642; Pourmand et al., 2019, p. 880), along with the influence of predictive modeling on the perspective of collection and interpretation of online data (Reece & Danforth, 2017, p. 1). Issues regarding the legality of data collection along with the change in social media usage were addressed by Wongkoblap et al. (2017, p. 8), who highlighted the sparse regulations regarding the use of publicly available data like pictures for machine learning models.

The evidence shows how early investments in machine learning technologies create expectations that drive further development, and how these systems begin to embed themselves within the infrastructure of psychotherapy, which is how technologies gain momentum. The following analysis will examine how the technological momentum behind chatbots and predictive modeling can create a feedback loop that compromises the privacy of human data and influences patient perspectives on psychotherapy. The momentum involves the ongoing

collection and analysis of personal data, which is then used to adjust the systems' outputs, potentially altering how patients experience and view mental health treatment along with changing society's perspective on what data should be private. The technological momentum framework is relevant for analyzing the implications of machine learning in mental health because it goes beyond considerations of accuracy and efficiency to explore how the design and integration of chatbots and predictive modeling creates self-reinforcing trajectories that affect human autonomy, behavior, and experiences. By looking at sources that discuss patient concerns and potential ethical consequences with ML-based treatment through the technological momentum framework, we can highlight the power that the integration of ML-based treatment places in the hands of developers and how this changes the patient-therapist relationship, alters the dynamic of power relations, and deepens the datafication of human behavior.

Datafication Loop

A "datafication loop", or a self-reinforcing cycle where the need and demand for data heavily influences human actions and technological development, operates through interconnected systems within psychotherapy. Since machine learning models within chatbots and predictive modeling require large amounts of data, what was once considered private data (social media posts, conversations with trusted therapists, patterns of digital engagement) becomes redefined as necessary for the training of these models. This normalization of harvesting personal data shifts societal expectations about privacy, reframing surveillance as a necessary trade-off for personalized care. As these technologies are deployed, they generate additional data through usage with patients. Each patient's conversation with a chatbot or predicted risk of mental illness creates a new point of data, which can then be used to train the model further. Deeply personal emotions and trauma shared with therapists now become part of

larger datasets to increase the accuracy of machine learning models, which further supports their widespread usage with the promise of expanded and improved treatment. This awareness can lead to users changing their behavior when they know their words, social media posts, and emotions are being constantly monitored, introducing bias and limiting the authenticity of and reliability of all data collected. As more and more user-generated data is collected and fed into models, the "improved" accuracy, through the addition of data, of these chatbots and predictive models can support institutional usage within psychotherapy, normalizing the collection of personal data and forcing patients to interact with machine learning systems to seek mental health treatment.

This loop represents more than just a technical process of how these systems need and generate data; it alters the societal perspective of mental health care. As machine learning systems gain momentum, they transform existing social dynamics to accommodate their own technological trajectory. The evidence on predictive modeling shows how accurate machine learning models can predict risks for certain mental health conditions, conceptualizing mental health as a problem of pattern recognition in data rather than a complex idea involving subjective experiences that may not be fully captured through data. Machine learning technologies reshape patients' expectations about therapeutic relationships and patients' understanding of their own mental health by presenting mental health as quantifiable, and something that can be predicted through data.

The evidence presented shows how chatbots initially developed to supplement human therapists gradually gain capabilities and applications that position them in society as potential replacements, masked as "increasing accessibility" to underserved areas. This aligns with Hughes' observation that technological systems can evolve beyond their initial purposes as they

develop momentum. What becomes apparent through this analysis is that technological momentum leads to a narrowing of possibilities as certain technological trajectories become institutionalized. As mental health institutions continue to adapt their practices to include machine learning technologies, alternative avenues of mental health care that do not involve the extensive use of personal data become marginalized, as the momentum behind technologies like chatbots and predictive modeling make it increasingly difficult to pursue treatment without interacting with a machine learning system.

Power

If therapy adopts widespread usage of machine learning, new stakeholders are introduced that have significant influence on patient outcomes without direct accountability to patients. Developers of these machine learning technologies gain power to shape therapeutic interactions through algorithm design. Their decisions about which data points to ignore or prioritize, which patterns to recognize as significant, and their framing of therapeutic responses become embedded into the technologies that mediate treatment. Granting developers this ability to control the responses from a chatbot or favor one pattern over another allows them to directly impact patient outcomes without facing responsibility.

This shift in power away from patients and therapists to private companies creating machine learning systems reshapes the broader landscape of mental health care. The collection of data enabled by chatbots and predictive modeling creates entities that control data that was once considered private, personal data, and the outcomes of machine learning models influence how mental health is conceptualized and treated. The technological momentum framework helps us understand that the shift in power emerges from the start of the integration of machine learning

within mental health treatment, embedding itself into institutional structures, and as these technologies gain momentum, it becomes increasingly difficult to reverse these power dynamics.

Conclusion

The technological momentum behind technologies such as chatbots and predictive analytics represents a shift in how mental health is conceptualized, diagnosed, and treated. The evidence presented demonstrates critical aspects of the momentum behind machine learning technologies. First, ML technologies create a self-reinforcing datafication loop, where technologies need data to function and use user-generated data to improve their accuracy and feasibility within existing systems. This momentum, supported by the models gaining more and more data through user interaction, shifts power from individual patients and therapists to developers and private companies creating the machine learning systems, changing how mental health is treated and how the field of mental health is understood, potentially marginalizing aspects of mental health that resist the usage of personal data.

These findings have implications for policymakers, as they emphasize an urgent need for government regulations and interventions that can guide technological momentum rather than allowing unwanted momentum to gain . Such regulations may include stricter protection of user-generated data, requirements of transparency between developers and patients, and protection for patients from any harm the technology can cause.

Limitations of this analysis should be acknowledged, as the technological momentum framework does not provide clear guidance for when momentum behind technologies should be resisted altogether or redirected instead. Moral frameworks are needed to evaluate the moral permissibility of machine learning in mental health, as technological momentum can only explain the trajectories a technology can follow, not whether the trajectories the technologies

follow are morally acceptable. Considering these limitations, the framework still provides valuable insights into the self-reinforcing dynamics driving the collection of data in mental health care and their implications for patient autonomy, privacy, and the future of psychotherapy. By acknowledging these dynamics, we can begin to work towards redirecting technological trajectories in ways that utilize the potential of machine learning while protecting patients first.

References

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *22*(7), e16021. https://doi.org/10.2196/16021

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, *76*(6), 642. https://doi.org/10.1001/jamapsychiatry.2019.0174

Berry, D. M. (2023). The limits of computation: Joseph Weizenbaum and the ELIZA chatbot. *Weizenbaum Journal of the Digital Society*, *3*(3), Article 3. https://doi.org/10.34669/WI.WJDS/3.3.2

Casu, M., Triscari, S., Battiato, S., Guarnera, L., & Caponnetto, P. (2024). AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Applied Sciences*, *14*(13), 5889. https://doi.org/10.3390/app14135889

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI Conference*, *2016*, 2098–2110. https://doi.org/10.1145/2858036.2858207

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, *24*(11), 1583–1598. https://doi.org/10.1038/s41380-019-0365-9

Gooding, P., & Kariotis, T. (2021). Ethics and law in research on algorithmic and data-driven technology in mental health care: Scoping review. *JMIR Mental Health*, *8*(6), e24668. https://doi.org/10.2196/24668

Pompili, M., Murri, M. B., Patti, S., Innamorati, M., Lester, D., Girardi, P., & Amore, M. (2016). The communication of suicidal intentions: A meta-analysis. *Psychological Medicine*, *46*(11), 2239–2253. https://doi.org/10.1017/S0033291716000696

Pourmand, A., Roberson, J., Caggiula, A., Monsalve, N., Rahimi, M., & Torres-Llenza, V. (2019). Social media and suicide: A review of technology-based epidemiology and risk assessment. *Telemedicine and E-Health*, *25*(10), 880–888. https://doi.org/10.1089/tmj.2018.0203

Quang, Jenny; (2021). *Does Training AI Violate Copyright Law?* https://doi.org/10.15779/Z38XW47X3K

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1), Article 1. https://doi.org/10.1140/epjds/s13688-017-0110-z

Richards, D., Vythilingam, R., & Formosa, P. (2023). A principlist-based study of the ethical design and acceptability of artificial social agents. *International Journal of Human-Computer Studies*, *172*, 102980. https://doi.org/10.1016/j.ijhcs.2022.102980

Roth, C. B., Papassotiropoulos, A., Brühl, A. B., Lang, U. E., & Huber, C. G. (2021). Psychiatry in the digital age: A blessing or a curse? *International Journal of Environmental Research and Public Health*, *18*(16), Article 16. https://doi.org/10.3390/ijerph18168302

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(09), 1426–1448. https://doi.org/10.1017/S0033291719000151

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput.-Hum. Interact.*, *27*(5), 34:1-34:53. https://doi.org/10.1145/3398069

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research*, *19*(6), e228. https://doi.org/10.2196/jmir.7215