Rishik Sarkar

01:355:201:A7

Elizabeth Decker

4/29/2021

Artificial Intelligence in Mental Health Treatment: A Bioethical Perspective

Introduction

Mental illness, commonly referred to as mental health disorders, is a term used to refer to a wide range of health conditions that can affect a person's mood, thinking, and behavior. According to national statistics reported by the Anxiety and Depression Association of America (ADAA), and the Depression and Bipolar Support Alliance (DBSA), two of the most common mental illnesses, anxiety, and depression, affect "40 million adults in the United States age 18 and older" (ADAA) and "approximately 17.3 million American adults" (DBSA) every year respectively. However, only a small percentage of the affected victims reach out for treatment or support owing to the negative social stigma surrounding these diseases. This lack of support for mental health patients is further aggravated by the shortage of professionals experienced in the field. A scholarly article written by Kathleen C. Thomas, et al., published by Psychiatry Online, states that "Over three-quarters (77%) of U.S. counties had a severe shortage of mental health prescribers or nonprescribers, with over half their need unmet" (Thomas 1325). However, a potential solution to the problem posed by stigma is confidentiality and anonymity of the victim seeking help, while a strategy to overcome the deficiency of trained professionals is the possibility of creating a simulation of a therapeutic environment: both of which can theoretically be achieved with the appropriate utilization of modern artificial intelligence tools and online

mental health resources. According to an article written by Sarah Graham, “AI has great potential to redefine our diagnosis and understanding of mental illnesses. An individual’s unique bio-psycho-social profile is best suited to fully explain his/her holistic mental health” (Graham 2). Thus, a virtual profile generated by an AI chatbot through conversations with a user could provide an accurate representation of their current mental state: information that could potentially allow the virtual therapist to treat the patient more effectively. That said, the involvement of uncertified techniques in a field as intricate as mental health introduces several ethical concerns from privacy risks to treatment efficacy.

In this paper, I will pose the question: is it ethical to utilize artificial intelligence and online resources to provide effective treatment and therapy methods to mental health victims? Furthermore, I will utilize the theory of the “four principles of bioethics” as a framework in my discussion of the ethics of AI. As a widely accepted theory, the four principles approach forms the cornerstone of ethicality in the medical system, and, according to Amber Johnson, a professor of Medical Ethics 101 at Stanford University: “Ideally, for a medical practice to be considered "ethical", it must respect all four of these principles: autonomy, justice, beneficence, and non-maleficence” (Johnson). I will examine each of these principles individually in the subsequent sections. In the first section, *AI, Mental Health, and Bioethics*, I will thoroughly introduce the problems posed by mental health disorders and the current limitations of traditional treatment methods that call for the need to incorporate relatively newer technologies such as the internet and artificial intelligence, relying on the theory of the “four principles of bioethics.” The second section, *Autonomy in Decision Making: Simplification or Manipulation?*, focuses on the ethical principle of autonomy, and discusses the ability of mental health patients to understand and make informed decisions about their treatment plans with the involvement of AI. The third

section, *Justice: The Future of Equity in AI Therapy*, considers the ethics behind the availability and accessibility of AI treatment methods in society. Finally, in the fourth section, *Primum Non-Nocere: The Coexistence of Beneficence and Non-Maleficence*, I highlight the ethicality of AI treatment methods in terms of their commitment to benefiting patients and their approach towards the rule of "first, do no harm." Overall, I acknowledge the various ethical concerns that accompany the introduction of AI into mental health treatment while highlighting possible solutions that can allow the technology to achieve its potential in the field.

AI, Mental Health, and Bioethics

Firstly, it is crucial to understand what a mental health disorder is, to recognize the demand for functional AI tools in its treatment, and the gravity of the ethical issues that will be discussed throughout the paper. In a very general sense, mental illnesses refer to disorders that inhibit a person's cognitive functionality by affecting their "mood, thinking and behavior," as stated in an article published by Mayo Clinic. The source also defines clinical depression as a disease that is characterized by a "persistent feeling of sadness and loss of interest" in day-to-day activities, and can lead to a "variety of emotional and physical problems." An article published by MedlinePlus states that although "Anxiety is a feeling of fear, dread, and uneasiness," and is a relatively common stress reaction, "Anxiety disorders are conditions in which you have anxiety that does not go away and can get worse over time." In this paper, I will be referring to clinical depression as "depression" and anxiety disorders as "anxiety," though it is essential to note that I am talking about long-term mental diseases and not temporary emotional states. Both these illnesses may require long-term treatment in the form of therapy and counseling: which are currently almost exclusively provided by human therapists through in-person or virtual conversation sessions.

Secondly, I will focus on defining artificial intelligence, to provide a basic overview of the technology to make the subsequent ethical conversations easier to interpret. In his article, Dalvinder Singh Grewal states that "Artificial Intelligence is the mechanical simulation system of collecting knowledge and information and processing intelligence of universe: (collating and interpreting) and disseminating it to the eligible in the form of actionable intelligence" (Grewal 13). In simpler terms, artificial intelligence algorithms analyze and "process" real-world data to make it easier to follow up on or find connections with; in the context of my paper, the information being collected would, for example, be key words in the emotions and feelings that the patient might communicate to the AI during conversations, and the output of the AI algorithms could be mental health conditions that have similar mental state symptoms. These prognoses could then be used to develop treatment plans—similar to how traditional therapy functions.

Autonomy in Decision Making: Simplification or Manipulation?

In the context of AI in mental health treatment, the ethical principle of autonomy causes the transparency of the technologies involved to become an essential feature. The concept requires that "the patient have autonomy of thought, intention, and action when making decisions regarding health care procedures" (Johnson). This essentially implies that the patient must possess the ability and knowledge to be able to select the medical treatment option that they wish to receive. Furthermore, this also necessitates the treatment technique to be transparent to the patient, and for the risk and benefits of the process to be clearly stated.

Some particular examples of AI healthcare options are the mental healthcare mobile applications, or MHapps for short, that Alyson Gamble targets in their paper. While discussing the user-centricity of MHapps, Gamble states that: "once MHapps are available on an online

marketplace, it can be difficult for users to discern what interventions are being used and which techniques have been shown to successfully address their concern" (Gamble 515). Because of the complicated design and technical ambiguity of commercially available MHapps, it becomes difficult for the user to distinguish between multiple applications, and to select a treatment option that they believe to be relevant to their health condition. Furthermore, this lack of information reduces the users' autonomy to make decisions or provide consent regarding their desired treatment option, thus questioning the ethicality of using technology that can be too complicated to be understood by a patient with limited technical knowledge. On the other hand, Trehani M. Fonseka argues against Gamble's point in her paper by stating that, "AI conversational programs have high usability scores and are well regarded for their user-adapted content and emotional responses" (Fonseka 958). Fonseka directly contradicts Gamble's theory of AI technologies being overly sophisticated for commercial use by suggesting that the software can "adapt" to its users and thus make it easier for them to understand and use the treatment processes provided by the programs. She essentially observes that although AI technology and the engineering behind it is extremely specialized, the core functionality of the programs can be presented to the users of the software in a simplified form, such that all the risks and benefits provided by the applications are thoroughly stated, and can be understood easily. This conclusion supports the principle of autonomy, since the ability to understand the purpose of individual AI therapy techniques will, in turn, allow users to decide whether or not to select them. However, one can argue that the process of "adapting" to users to make software look more uncomplicated than it is can be considered a form of deception. Additionally, the very term AI means "artificial" intelligence, which means that the software is simulating natural intelligence with enough accuracy to make it appear "human" when it is only a machine. David Luxton touches on this issue in his paper when

he discusses the basis of trust in therapeutic relationships and states: “The issue of trust raises an important philosophical question; is it unethical to simulate a human so much that people believe that the simulation is human?” (Luxton 3). He uses the term “simulate” twice to emphasize that AI is not alive, but just appears to be so—and hence brings up the argument that the ability to reflect a patient’s emotions might provide unnecessary power to the software. Because it “reflects” and “empathetically understands” (Luxton 4) a user’s emotions, AI therapists might appear to manipulate their patients—which removes the patients’ autonomy to form unbiased decisions. This presents a drawback to the point brought up by Fonseka about the user-centricity of AI software, and is a viable counterargument; however, because of the complicated nature of AI, it is unrealistic to expect patients to operate under unbiased and fully-informed consent.

Gamble, Fonseka, and Luxton all bring up essential points concerning the ethical principle of autonomy in their respective papers, and while it is true that the technical complexity of AI therapy methods, combined with the software’s replicative nature, makes it difficult for mental health patients to efficiently make decisions regarding treatment options: proper supervision and simplified communication of the systems and procedures involved can ensure that the patient is not denied autonomy. This section thus partly answers my research question about the ethicality of AI by demonstrating that the current state of the technology does not explicitly address the ethical concern of a lack of patient autonomy yet, but there exist possible solutions to this limitation.

Justice: The Future of Equity in AI Therapy

Justice is an important ethical principle in the context of AI since it ensures that the users of the technology are protected from inequity and unjust liability. The concept is defined as “The idea that the burdens and benefits of new or experimental treatments must be distributed equally

among all groups in society" (Johnson). The idea of justice considers the availability and fair use of the resources provided by medical treatment options and ensures the equitability of healthcare: while making sure that any disputes involving the treatment are resolved fairly. In the case of AI therapy, it is necessary to consider this principle in terms of traditional treatment options before we arrive at the discussion about the ethicality of current software because an understanding of the drawbacks of in-person treatment justifies the demand for online therapy.

As I mentioned earlier in the introduction to my paper, Kathleen C. Thomas states the US has a severe shortage of mental health facilities as compared to the number of victims in the country. This scarcity of resources, in addition to the general lack of awareness and negative societal stigma concerning mental illnesses, causes a large percentage of the affected population to struggle without a diagnosis, prescription, or healthcare procedure. Jimmy Tan focuses on this problem in a section of his article, "Primed for Psychiatry: The role of artificial intelligence and machine learning in the optimization of depression treatment," where he states:

> Psychotherapy can be difficult to access for uninsured patients and wait times can be lengthy. Initiating face-to-face psychotherapy may not offer the timely support required by patients at the start of treatment. An artificial intelligence (AI)-enabled, text-based conversational mobile mental well-being app can bridge this gap. (Tan 45)

Tan indirectly addresses the principle of justice by highlighting that the inadequacy of traditional mental health resources can be countered by the introduction of "conversational mobile mental well-being apps" that utilize AI. Since these are mobile applications, they can be accessed by most patients with a smartphone that meets the software requirements; and although the required specifications vary based on the application that the user wishes to install: they are generally more realistically obtainable than the resources necessary for in-person therapy. Therefore, Tan

argues that AI therapy methods are justified in their operation by being more accessible than their traditional, in-person counterparts. Additionally, the author brings up the important topic of monetary capacity by voicing the concerns of “uninsured patients” who are unable to afford the expenses of psychotherapy, and hence are incapable of seeking treatment. These points directly suggest that the inclusion of AI in therapy would mitigate the lack of mental health resources in the US, as reported by Thomas, and could provide appropriate treatment to currently undiagnosed mental health victims. When it comes to the principle of justice, most of us will readily agree that AI software and online resources are much more cost-effective than their traditional counterparts. Indeed, Michael E. Thase corroborates this viewpoint in his paper: in which he conducts a study on the cost-effectiveness of computer-assisted models of Cognitive Behavior Therapy (CCBT), all of which utilize a 9-module multimedia program called “Good Days Ahead” (GDA), versus regular Cognitive Behavior Therapy (CBT). Thase's research supports Tan's observation by indicating that the AI model of CBT "was achieved with a cost savings of USD 945 per patient" (Thase 21). Thase's discovery highlights the immense potential of migrating current therapy methods to a technological environment where they can be conveniently accessed by anyone on the internet and allows for the justification of online therapy options by ensuring the equitability of treatment expenses.

The ethical principle of justice also requires that potential conflicts be resolved fairly; to uphold the legislation, there must exist a just and objective process for holding parties accountable. In the case of AI software, however, the process of determining and assigning responsibility becomes excessively complicated owing to the complex nature of the programs themselves. David Luxton discusses this predicament in his paper, when he asks, “Who should be responsible for the actions, decisions, and recommendations that AICPs make? What if a care

seeker dies by suicide or engages in homicide after disclosing intent to an AICP? What is the appropriate responsibility of end users, developers, and others as these systems increase in autonomy?" (Luxton 5). In this case, Luxton considers a hypothetical situation in which the users of an artificial intelligent care provider (AICP) software are harmed because of its treatment recommendations. In a traditional therapeutic relationship, the therapist and medical staff involved would take full responsibility for anything that happened to the patient while in their care. Thus, it is relatively easier to determine the responsible party when only humans are involved; however, the fact that the creation of AI software involves several "end users" and "developers" makes it difficult to attribute its shortcomings to a specific individual or machine. This, in turn, makes it harder to address the issues that caused the failures in the first place. Furthermore, holding an individual or group accountable becomes even more difficult as the AI software becomes more complex and more parties become involved in its development. As a result, AICPs and other such AI therapy resources appear to be unethical in terms of justice. John P. Sullins provides a potential solution to this dilemma by stating that "the most common theoretical schema is the standard user, tool, and victim model. Here, the technology mediates the moral situation between the actor who uses the technology and the victim. In this model, we typically blame the user, not the tool, when a person using some tool or technological system causes harm" (Sullins 152). Thus, in the situation outlined by Luxton in his paper, the "user," who—in this case—is the "care seeker," would be held responsible for "suicide" or "homicide" instead of the AICP software or its developers. Although this decision might sound unjust or inhumane, it is the most effective method to distribute responsibility in such a situation and ensures that the principle of justice is consolidated in the case of AI therapy.

Thus, AI software and online resources are possibly even more ethical than their traditional counterparts when it comes to availability and cost-effectiveness; however, it seems that the process of ensuring accountability is a gray area on the morality spectrum. Overall, this analysis concludes that the current AI therapy options conform to most of the rules outlined in the principle of justice, and answers my research question by illustrating that AI software is at least partially justified from an ethical standpoint.

Primum Non-Nocere: The Coexistence of Beneficence and Non-Maleficence

Although AI software is developed to benefit its users and makes regular streamlining and renovation a priority, the relative recentness of the technology implies the existence of security issues that can potentially make it maleficent. The ethical principle of beneficence requires that "the procedure be provided with the intent of doing good for the patient involved," thus ensuring the net benefit of the patient, while that of non-maleficence demands that "a procedure does not harm the patient involved or others in society" (Johnson). I have decided to group both of these principles into a single section because I believe that they are both correlated in the context of AI.

The principle of beneficence ensures that medical processes always have the best interests of the patient in mind and that treatment methods are created through careful consideration of their circumstances. In his contemporary paper, Krešimir Ćosić states that "… one of the greatest impacts of digital psychiatry, particularly applied artificial intelligence (AI) and machine learning (ML) during the ongoing COVID-19 pandemic, is their ability of early detection and prediction of [health-care workers'] mental health deterioration, which can lead to chronic mental health disorders. Furthermore, AI-based psychiatry may help mental health practitioners redefine mental illnesses more objectively than is currently done by DSM-5" (Ćosić

279). The author mentions the "DSM-5," which is the "Diagnostic and Statistical Manual of Mental Disorders"—a list of categorized mental health disorder symptoms that are used to diagnose mental illnesses. Ćosić suggests that AI and ML systems can be used to "predict" mental health disorders using data collected from individuals—which ultimately allows more victims to promptly receive the medical attention they require; this ensures that "health care providers develop and maintain skills and knowledge" (Johnson) that they need for effective treatment. Furthermore, AI and ML models become more efficient as they analyze and are trained on larger data sets: which ties in with the principle of beneficence and its objective of ensuring continuous renovation in healthcare. Ćosić's point is corroborated by Sarah Graham in her article when she states that "Leveraging AI techniques offers the ability to develop better prediagnosis screening tools and formulate risk models to determine an individual's predisposition for, or risk of developing, mental illness" (Graham 2). In addition to mentioning the AI techniques that can predict mental illnesses, Graham's inclusion of the phrase "individual's predisposition" highlights that these technologies can consider individual circumstances and act accordingly—which further strengthens my argument for the ethicality of AI in terms of patient beneficence. Thus, by predicting the patient's risk of developing a mental illness, and either communicating crucial information about the patient's mental health to medical providers or providing supplemental therapy itself, AI appears to be developed to help its patients in mind and is therefore beneficent.

However, a complication is introduced when we consider the current state of AI technology: even though AI therapy is theoretically designed to help mental health victims, the recentness of this technological innovation gives rise to questions regarding its effectiveness in practice. Because the intended purpose is insignificant when it comes to addressing problems

caused by the practical use of the software, we must ensure that AI follows the golden medical rule of "primum non nocere," or, "first, do no harm" in its functionality. Trehani Fonseka elaborates on the current limitations of AI in this regard in her paper: "A third area of high importance is patient safety. In particular, it is essential to ensure AI programs can appropriately respond to suicidal users and not worsen their emotional state or accidentally facilitate suicide planning" (Fonseka 960). The author mentions the current limitations of the AI and ML algorithms that are being used to develop suicide-risk prediction models and states that one of the most important factors is "patient safety"—which essentially refers to the possibility that a patient's mental health might deteriorate due to the ineffective involvement of the techniques mentioned earlier. She highlights the intricate and delicate nature of mental health disorders and suggests that if not approached correctly, AI treatment techniques might be more detrimental than beneficial to the victim. Naturally, there is always the possibility of misdiagnoses even by human professionals, however, the ethical stakes are relatively higher once newer, improperly regulated technologies are involved. Hence, by pointing out one of the key drawbacks of AI usage in a field as complex as mental health, the author implies that despite its apparent advantages, the innovation still requires refinement before it can become mainstream. On a similar note, because the official regulations for AI usage are not well-enforced yet, the software poses privacy concerns such as data breaches and phishing attempts. Although this might not cause immediate physical harm to the patient, it significantly threatens their livelihood, because confidential information can serve as dangerous forms of leverage in the wrong hands. Furthermore, the disclosure of sensitive data can give rise to negative "stigmatization" against mental illnesses, as Alyson Gamble mentions in their paper, and directly reduce the "efficacy" (Gamble 517) of mental health resources. Gamble states:

> MHapp chatbots gather intimate healthcare information about a user, typically with the individual's expectation of privacy. Some MHapps track location data, permit audio recording, and may be linked to financial information. Thus, MHapp user databases can contain extremely sensitive, personal information. Yet, this information is often not governed by regulations that would normally protect health information, such as HIPAA. (Gamble 517)

Although it is essential for AI applications such as MHapps to collect user data such as "audio recordings" and "sensitive" information to properly diagnose illnesses and effectively communicate with the victim, the lack of proper regulations introduces an imminent risk to the entire process and provides AI with a negative connotation in terms of its ethicality. Organizations such as the "HIPAA" which normally preserve the confidentiality of medical and therapeutic relationships are not yet widely applicable in AI, since online resources and AI healthcare software are not yet officially recognized as medically licensed therapy providers.

Hence, the inadvertent harm that could be caused to the patient by AI software due to its current limitations makes it maleficent. However, the introduction of ethical guidelines and official organizations that enforce confidentiality in technology is a possibility; and as was stated before, AI and ML algorithms become more efficient and effective as they are trained on larger datasets: which implies that the future could theoretically see the development of software with improved efficacy. The existence of these possible solutions elucidates the idea that even though they are not yet ethical in terms of non-maleficence, AI therapy has great potential in the future: which directly corroborates my thesis.

Conclusion

In conclusion, while AI and online resources have a lot of potential in the field of mental health, it currently poses a multitude of ethical concerns, which must be acknowledged and appropriately resolved. As was presented in the research, the four bioethical principles of autonomy, justice, beneficence, and non-maleficence are not all properly satisfied by the current state of the technology. In the case of autonomy, AI software—which is currently complicated and, arguably, manipulative—can be presented to patients with a simplified user interface that clearly communicates its goals, risks, and benefits; and can be rigorously supervised to ensure the absence of bias in patient decision-making. To ensure justice, the lack of accountability in AI can be rectified by using the "standard user, tool, and victim model" proposed by Sullins. Furthermore, the ethical concern of ineffectiveness and the risk of data breaches can be alleviated through extensive training in AI and ML algorithms and the introduction of ethical guidelines and organizations that protect patient confidentiality, respectively. Overall, although AI therapy is not yet an orthodox treatment option for the general public, as compared to traditional in-person counseling sessions, and poses several ethical concerns—which makes it currently unethical in practice—it is arguably an essential asset that can mitigate contemporary challenges such as a shortage of therapeutic resources and mental health professionals, and must be optimized to be comprehensively viable from an ethical standpoint; this directly corroborates my thesis and answers my research question about whether or not it is ethical to use AI in mental health treatment. As AI programs and algorithms continue to develop in the future, many of the ethical issues discussed in this paper will likely be addressed: however, as is the case with any new technology, a lot more research must be conducted before the ethicality of AI can be thoroughly warranted.

Works Cited

"Anxiety." *MedlinePlus*, U.S. National Library of Medicine, 29 Jan. 2021.

Ćosić, Krešimir, et al. "Artificial Intelligence in Prediction of Mental Health Disorders Induced by the COVID-19 Pandemic among Health Care Workers." *Croatian Medical Journal*, Croatian Medical Schools, 5 July 2020.

"Depression Statistics." *Depression and Bipolar Support Alliance*, 12 July 2019.

"Depression (Major Depressive Disorder)." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 3 Feb. 2018.

"Facts & Statistics: Anxiety and Depression Association of America, ADAA." *Facts & Statistics | Anxiety and Depression Association of America, ADAA*.

Fonseka, Trehani M., et al. "The Utility of Artificial Intelligence in Suicide Risk Prediction and the Management of Suicidal Behaviors." *Australian & New Zealand Journal of Psychiatry*, vol. 53, no. 10, Oct. 2019, pp. 954–964. *EBSCOhost*, doi:10.1177/0004867419864428.

Gamble, Alyson. "Artificial Intelligence and Mobile Apps for Mental Healthcare: A Social Informatics Perspective." *Aslib Journal of Information Management*, vol. 72, no. 4, July 2020, pp. 509–523. *EBSCOhost*, doi:10.1108/AJIM-11-2019-0316.

Graham, S., Depp, C., Lee, E.E. *et al.* Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Curr Psychiatry Rep* 21, 116 (2019).

Grewal, Dalvinder Singh. "A critical conceptual analysis of definitions of artificial intelligence as applicable to computer engineering." *IOSR Journal of Computer Engineering* 16.2 (2014): 9-13.

Johnson, Amber. *Medical Ethics 101*, web.stanford.edu/class/siw198q/websites/reprotech/New%20Ways%20of%20Making%20Babies/EthicVoc.htm#:~:text=Bioethicists%20often%20refer%20to%20the,beneficence%2C%20and%20non%2Dmaleficence.

Luxton, David D. "Recommendations for the Ethical use and Design of Artificial Intelligent Care Providers." *Artificial Intelligence in Medicine*, vol. 62, no. 1, 2014, pp. 1-10.

M.P.H., Kathleen C. Thomas, et al. “County-Level Estimates of Mental Health Professional Shortage in the United States.” *Psychiatric Services*, 1 Oct. 2009.

Sullins, John P. "When is a robot a moral agent." *Machine ethics* 6.2001 (2011): 151-161.

Tan, Jimmy, et al. “Primed for Psychiatry: The Role of Artificial Intelligence and Machine Learning in the Optimization of Depression Treatment.” *University of Toronto Medical Journal*, vol. 96, no. 1, Jan. 2019, pp. 43–47. *EBSCOhost*.

Thase, Michael E., et al. “Improving Cost-Effectiveness and Access to Cognitive Behavior Therapy for Depression: Providing Remote-Ready, Computer-Assisted Psychotherapy in Times of Crisis and Beyond.” *Psychotherapy and Psychosomatics*, Karger Publishers, 12 May 2020.