ETHICAL CONSIDERATIONS OF CHATBOT USE FOR MENTAL HEALTH SUPPORT

by
Tiana Sepahpour

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Bioethics

Baltimore, Maryland
August 2020

© 2020 Tiana Sepahpour
All rights reserved
Abstract.

The promise of artificial intelligence thus far is greater than its actual application in healthcare, though it has been developed to be used effectively in this field. One application of technology in this setting is chatbot use for mental health support. Chatbots are conversational entities that mimic human conversations to interact with a human user as if two human agents are engaging with one another.¹ Chatbots have the potential to engage those who might otherwise be unable or unwilling to receive mental health care by circumventing many of the barriers that prevent people from receiving care from a human provider. In the mental health arena, both practical barriers (such as geographic location and financial costs) as well as societal barriers (such as self and societal stigma) prevent people from accessing treatment. Even some of those who are able to overcome some most commonly cited barriers to care cannot receive it due to the lack of providers as compared to demand for aid. Thus, there is a need for a massive enhancement of mental health services that is not currently being met, where chatbots could rise to meet these needs.

Despite lengthy discussion of the negative possibilities that I will summarize in this paper, the positive outcomes remain especially promising. This technology as a therapy, however, is investigational, as it has not yet been approved for use. Because of this, I advocate for moving forward with this technology, but cautiously. In this paper I will identify different areas of ethical concern regarding chatbot use for mental health support. I will explain why the identification of harms matter, and what steps, if any, can be taken to prevent and mitigate them. Most importantly, I will reiterate that though this technology is investigational thus far, it has the potential to be a widely successful therapeutic solution to many barriers of access, and that these

considerations are paramount for the ethical development of this technology in the mental health space.

Primary Reader and Advisor: Travis Rieder, PhD

Secondary Reader: Alan Regenberg, MBE
Acknowledgements.

This thesis would not have been possible without the invaluable contributions of many respected individuals, for whom I would like to express my deepest gratitude. I would like to thank my advisor, Dr. Travis Rieder, for his guidance in all of my academic endeavors and beyond, for his exceptional discussions in the classroom, and for his availability, enthusiasm, and diligence in navigating me through this thesis from its genesis to its final version. I would also like to thank Alan Regenberg for his thorough attention to this work, and for guiding me through my many inquiries during my time at the Berman Institute with patience and thoughtfulness. My gratitude also extends to my cohort members and Agraphia group for their constructive feedback during the most formative stages of this thesis. Further, I am proud to have gone through this program with so many curious and gifted individuals. I would also like to thank all my professors and mentors in the Berman Institute of Bioethics and in the Bloomberg School of Public Health for their time and attention to my education and academic pursuits. Finally, I cannot properly express my gratitude for the love, support, and encouragement of my family and dear friends, without whom this process would have been much less painless.
Contents.

Abstract ii

Acknowledgements iv

1. Introduction 1

2. Barriers to Mental Health Treatment 2

3. Clinical and Value Concerns 4

4. Programming and Learning Concerns 9

5. Disclosure as Upholding Public Trust 11

6. Addressing Moments of Absolute Crisis 12

7. Ethical Basis for Avoiding Harms and Bad Consequences 15

8. Moral Responsibility 17

9. Privacy and Confidentiality 19

10. FDA Approval 21

11. Injustices as a Result of Novel Nature 23

12. Conclusions 24

7.1 Discussion of Chatbots as Ethically Permissible 24

7.2 Works Cited 27

Biographic Statement 33
Introduction.

The introduction and development of artificial intelligence (AI) has lent itself as a prominent tool of healthcare in the 21st century. Technology has developed rapidly to better conform to serving clinical needs, and has been meaningfully transformative in healthcare by way of its diagnostic capabilities, and has been validated in part for its analytic strength in interpreting images, such as screening, reading x-rays, and evaluating biopsies.\(^2\),\(^3\) Cancer screening is one of these technological successes, where oncologists use screening tools, including those powered by AI, to detect approximately 1.8 million new cases of cancer annually in the United States.\(^4\),\(^5\) However, the promise of AI is far greater than the applied performance of technology in healthcare thus far, and it is to this promise that we cling. AI, if thoughtfully developed to reduce harms and mitigate bad consequences, has the potential to meet a great need in the broader field of healthcare: the mental health treatment gap.

More than half of American adults will experience a mental health illness in their lifetime, many comorbid and chronic.\(^6\),\(^7\) Further, approximately 20% of Americans will experience a mental illness in any given year.\(^8\) In the last year, only 41% of people with mental health illnesses receive mental health care in the United States, driving a demand for increased mental health services by many health care professionals to fill this unmet need.\(^9\),\(^10\) The recent evolution of AI, specifically in the development of chatbots, has brought treatment of mental health disorders to

\(^4\) Ibid.
\(^6\) From “Learn About Mental Health: Mental Health Basics”, by CDC, 2018.
\(^8\) From “Learn About Mental Health: Mental Health Basics”, by CDC, 2018.
the forefront of developing healthcare services. Chatbots are understood as conversational agents “where a computer program is designed to simulate an intelligent conversation”. In the healthcare setting, chatbots have been considered for providing some range of mental health services, such as offering Cognitive Behavioral Therapy (CBT), detecting depression, and providing broader emotional support. Employing chatbots to deliver mental health treatment by way of CBT or other forms of communicative therapy could serve as a powerful therapeutic tool, as well as relieve the burden on mental healthcare providers, who are often in high demand, yet understaffed, or are inaccessible in other ways, due to financial or physical barriers to treatment, for example. While chatbot technology is complicated and new, the increased access to mental health care, and the potential therapeutic benefits could outweigh their harms. Because of their value potential, I will argue we ought to use them despite this uncertainty regarding their risks and benefits.

**Barriers to Mental Health Treatment.**

Mental health services are expanding in availability and care options, but providers still face several practical barriers to actually delivering treatment services. The largest barrier to treatment comes by way of stigma. Societal and self-stigma expose individuals to judgement from their community, or cause them to internalize their own negative attitudes towards mental

---


Another major barrier is the limited number of mental health practitioners, where demand for mental health providers exceeds availability. Other prominent barriers include service costs and physical barriers, where healthcare providers are often inaccessible due to cost, or are sparse in more rural areas. When these barriers are compounded, treatment seeking and availability to offer care become much more difficult, further widening the mental health treatment gap. Mental health interventions built around chatbot use can be applied to each of these barriers to mitigate reported problems in seeking mental health services and encourage help-seeking behaviors.

Chatbots have many practical strengths, including ease of access. The major benefit of ease of access is eliminating the need to travel or pay high service costs for therapeutic services. Chatbots could also practically help fill in gaps of demand for those who wish to seek treatment, but have difficulty connecting with a human provider. Although, ideally, anyone who wished to seek care would be matched with a highly trained mental health provider, these conditions are unrealistic given how demand outstrips supply even in areas where these services are accessible.* Additionally, for those who do not seek treatment due to stigma, the use of chatbots can be a discreet option for offering mental health support. Those who fear judgement from human peers may feel more comfortable communicating remotely with a chatbot. Chatbots neatly fill a gap in care delivery for mental healthcare, and therefore hold great promise as a healthcare tool.

---

16 Ibid.
Clinical and Value Concerns.

While the promise of chatbots is immense, there are a number of problems with their implementation as new and unregulated technology. If chatbots are to be utilized to realistically meet the demand for mental health services, there are ethical issues that are necessary to address. Expansion of these issues helps color areas where identified issues exist, and what remedies could possibly rise to meet them. The relationship between a human provider and their patient certainly holds immense value, as seen in the success of psychotherapy historically and more recently, in telehealth.\(^{20, 21, 22}\) One concern, therefore, is that there will be diminished value in the relationship between chatbot and human user. This concern is valid, and may prove to be true. Whether it will or not, however, is an open, empirical question that we can only answer when we have greater experience with the use of chatbots in this context.

Some chatbots have been used in a recreational setting and have shown great success in terms of creating bonds that have been felt on the part of human users towards the chatbot programs they converse with.\(^ {23}\) In one instance, a senior AI designer, Steve Worswick, who has won the Loebner Prize\(^ *\) five times for his chatbot, “Mitsuku”, gave a presentation demonstrating Mitsuku’s capabilities. Worswick reported having to start his presentation with a test to prove Mitsuku was in fact non-sentient, and was responding only with phrases and reactions that he

---


\(^ {21}\) From “In pandemic, many seeing upsides to telemedicine”, by M. Van Beusekom, *Center for Infectious Disease and Research at the University of Minnesota*, 2020.


\(^ *\) The Loebner Prize is an annual competition in artificial intelligence that utilizes the Turing Test to award prizes to developers of computer programs determined by a panel of judges to be the most human-like.
had programmed her* to say, as so many people who had interacted with Mitsuku thought her to be so human-like that she must in fact be a real human woman being held captive by Worswick, and not a computer program. Worswick showed that when asked if she was a human, she would say “yes”, because that is what she was programmed to say, and not because it was true. This example is brought up to show that complex relationships, whether meaningful or not, can develop, as some chatbot programs at least have the capacity to communicate in a way that feels genuinely human, even if they are in fact non-sentient computer programs. Users perceive complex relationships between themselves and some chatbots, where in reality, their relationship is one-sided. That is to say, while they develop attachments to the chatbot, the chatbot cannot develop and cultivate these same attachments.

It is one consideration that chatbots function well in the space of entertainment or wellness, as companions rather than therapeutic treatments, where they have previously shown success. The success of chatbots as stimulating conversational partners in the field of entertainment and companionship, however, also points to their potential for success as a therapeutic technology. In one 2020 study, chatbots were found to ease the burden of too much demand for medical providers, and participants viewed chatbots even more positively than human agents, offering promising results for future chatbot use in healthcare. Other studies have also reflected success of chatbot use in more therapy-oriented settings. A 2014 study showed that human users were more likely to disclose histories and emotions in a mental health setting to

* Given that chatbots are not agents, they generally circumvent being addressed with personal pronouns that human agents choose. However, in this paper, the chatbot Mitsuku is referred to with she/her/hers pronouns to reflect the way that her creator, Steve Worswick addresses her.

24 Ibid.
25 Ibid.
27 From “Chatbots can ease medical providers' burden, offer trusted guidance to those with COVID-19 symptoms”, Indiana University: News at IU Bloomington, 2020.
a chatbot as opposed to a human provider. Further, a 2-week trial of Woebot\* also reflected positive mental health outcomes as compared to a control group. Yet, chatbots have not been used before in the application I propose in this paper. Even chatbots such as Woebot have not gained FDA approval, and do not advertise their medical capabilities. Nor do they expand and refine therapeutic interventions as a result of being on the market as medical and mental health interventions. This is the way in which the use of this technology is novel, and one way in which it will need to be evaluated, to show whether or not the value between a human provider and patient is unique and unable to be replaced or supplemented, or whether it can be replicated to some degree.

Recall that in the mental health space, demand for care exceeds the capacity to treat mental health needs. In this way, the use of chatbots in this space is operant as a means to offer resources where there were none before, rather than take over provider roles that were already meeting this need. The use of chatbots in this space can also rise to meet other needs that were previously unmet, and unable to be met by a human provider, due to lack of resources or time. One function of chatbots could be to offer supplemental support, rather than fill the role of mental health providers. Functioning in this way, chatbots might offer the solution of getting people through the week between meetings with a human mental health provider, or exist as a stepping stone to telehealth with a human provider. For those who are not yet comfortable with seeing a human provider, or experience stigma towards therapy in a more robust sense, conversation with a chatbot that could provide mental health support may aid in facilitating

---


\(29\) Woebot is a cognitive behavioral therapy (CBT) – enabled conversational agent that enables users to increase well-being by improving mental health symptoms such as those induced by depression or anxiety.

\(29\) Ibid.
acceptance of other therapeutic options, namely those that connect the user with a human provider.

Chatbots assuage not only mental barriers such as stigma towards therapeutic options, but also accommodate physical barriers, such as distance from human providers, or cost of travel to a provider’s neighborhood. Because chatbots can offer therapeutic treatments remotely, they might be attractive to those who face these types of barriers, such as inability to travel due to distance, or the wish to maintain discretion from their families. In this way, chatbots offer a solution to a unique subset of users who are unable or unwilling to receive services in more traditional ways. Though it is possible that some value is lost from seeing a human provider while the human user converses with the chatbot, it is altogether possible that conversation with the chatbot will instill confidence or comfort in the human user to reach out to a human provider for further mental health support. It is notable that many people residing in more rural areas lack the same access to care that those who live in cities or metropolitan areas enjoy.\textsuperscript{30} Though the use of telehealth is increasing rapidly and will likely persist in the future, and telehealth serves as a means to meet this healthcare gap, chatbots could also rise to meet this gap.\textsuperscript{31} Chatbots hold value in reaching a population that might be unlikely to make contact with a human provider, even remotely, for a number of other reasons, such as stigma or feelings of shame. Thus, chatbots offer aid where there was not previously aid before rather than diminishing value for those who are already receiving aid.

The success of telehealth that emerged due to the COVID-19 pandemic also reveals possible answers as well as a series of questions that might need to be considered, as similar


\textsuperscript{31} From “A Community-Based Intervention to Improve Cognitive Function in Rural Appalachian Older Adults”, by B. Katz, Innovation in Aging, 2019, 3.
therapies have been applied due to a necessity of seeing patients virtually as opposed to in an in-person setting. The exploding pandemic has shown that telemedicine can be conducted with more success than anticipated, and has revealed great benefits for some organizations who began offering mental health support virtually by necessity, and now plan to extend telehealth options into the future, even when no longer necessary. The recent need for telehealth has raised a series of important questions in the event that similar events may (and likely will) happen in the future. Among these questions are: what happens when we cannot leave our houses but need healthcare? What happens if our best options for healthcare are very far away, or are close by but do not require a physical examination? Do users feel they are receiving the same quality of care remotely as they are in-person? What are the implications if revised standards of care during the crisis lead to treatment being offered where it was not previously available?

These are interesting and important questions, though not limited to any one specific situation. As climate change continues to ravage our Earth, releasing a slew of new vector-borne and novel diseases and viruses, and as natural disasters ramp up, questions of whether telehealth as a broad form of healthcare is viable – or viable enough – to be considered and refined for current and future use materialize. Further, an onslaught of increased global disasters has been shown to have negative impacts on mental health, increasing rates of anxiety and depression, PTSD, even impacting rates of suicide. Given the likelihood of a growing need for mental

---

33 From “In pandemic, many seeing upsides to telemedicine” by M. Van Beusekom, *Center for Infectious Disease and Research at the University of Minnesota*, 2020.
health support and the recent successes of telehealth, chatbots could meet this need from a telehealth platform that has proven to be effective in delivering medical care.\textsuperscript{37}

\textbf{Programming and Learning Concerns.}

A number of conspicuous issues must first be solved before chatbots can be deployed in and be successful in healthcare settings. A distinct risk is that therapy by way of chatbot use will not be clinically effective, but it is theoretically possible that a chatbot could be effective and yet still cause harm by way of its learning arrangements. Therefore, though the question of clinical effectiveness might be resolved, there are independent concerns that are still outstanding.

Though chatbot technology has extensive conversational capacities that are designed to mimic the unstructured aspects of human-to-human conversations, they are not all designed equally.\textsuperscript{38} Chatbots learn by one of two avenues: either they are taught or fed information with a limited script, known as supervised learning (these are \textit{rule-based} systems, like early model chatbots such as ELIZA), or they are allowed access to larger datasets of human conversation by enabling continued learning, known as unsupervised learning.\textsuperscript{39, 40}

In selecting which programming style will lead to the best outcomes, it is imperative to minimize harms that may occur as a result of unintended programming consequences. Though the history of chatbot applications is short, previous noted harms of unsupervised learning reveal a clear risk in programming chatbots to learn in this way. When permitted to learn unsupervised, chatbots have turned into racist, sexist, and obscene entities in as little time as an hour.\textsuperscript{41}


\textsuperscript{39} Ibid.

\textsuperscript{40} From “Ethics and Chatbots”, by S. Worswick, 2018.

\textsuperscript{41} Ibid.
Unsupervised learning uncritically carries forward any issues related to problematic content that can and frequently does appear in these kinds of databases. These attitudes are vigorously condemned in human beings, and would certainly doom programmed technologies meant to deliver therapeutic interventions. Communicating with a human user in a disrespectful and inappropriate manner could cause harms to mental health – the very opposite goal of chatbot implementation in this space. A chatbot’s sexist, racist, or even rude behavior could decay feelings of trust and confidence that a human user has in the chatbot offering aid. Certainly, it is detrimental to the goals of chatbot use, which are to offer therapeutic interventions and support to those who are unable or unwilling to access this support by other avenues. Another consideration is that chatbots will service people of all ages; as a result, they must employ language appropriate for children, and cannot adapt to language that adult users might use freely when conversing with child users.

This offers a compelling reason to restrict chatbot learning to supervised learning only, where they cannot learn and regurgitate spiteful or inappropriate messages. It also identifies a need to develop new methods for unsupervised learning that can correct for these sorts of problems that currently exist when chatbots learn in this way. Supervised learning, though more labor-intensive for programmers, could eliminate language that could cause emotional harm or decay of trust on the part of the chatbot. There might be other venues where a cost-benefit analysis may lead to greater risk-taking in terms of allowing unsupervised learning for chatbots. One instance may be a need for rapid development. However, in the space of mental health services, this is an inappropriate gamble, and so supervised learning is required, even if this hinders speed of development or language sophistication. Therefore, given current constraints, supervised learning must be employed to uphold the very goals of this potential therapeutic solution.
Disclosure as Upholding Public Trust.

After learning programming has been determined, methods and guidelines surrounding communication between the program and the users must be established. A major ethical question stands as to whether or not chatbots need to disclose they are computer programs and not human agents to their human users. Disclosure is of particular importance as a consideration in chatbot use, as many users are unable to distinguish the chatbot as a program and not as a human agent on the other end of the screen. For this reason, a lack of disclosure can be deceptive if the human user seeking help believes they are speaking with another human agent, but in fact are not. Preventing deception is intrinsically important, but is also important as public health interventions rely largely on the maintenance of public trust. If used as a healthcare intervention, it aligns with the goals of public health to disclose that chatbots are not human beings, so as to promote public trust and maintain the autonomy of users. Autonomy in this case refers to the users’ abilities to decide whether or not to use the chatbot intervention with sufficient information as to how it works. Disclosure promotes autonomy as it offers information that might be important to users in deciding whether or not to use a chatbot intervention.

Historically, disclosure matters, and disclosure on the part of the chatbot has had impacts on the autonomy of its users. Human behavior changes when individuals become aware they are speaking with a computer program rather than another human being. Therefore, it seems ethically relevant to determine whether disclosure is necessary before supporting the use of chatbots. While it is possible that disclosing this information may depreciate the experience, it is

---

43 Ibid.
possible that not doing so may violate trust and detract from the very relationship whose sanctity must be upheld for effective delivery of aid. We protect autonomy with requirements of informed consent, where one cannot autonomously consent if they are being deceived about the very nature of the experience they are presented with. Deception itself is inherently harmful, and is only defensible with sufficient reasons.* The broader harms of deception in the space of chatbot use is damage to the endeavor at large: that news of deception could erode trust on a larger scale. In order to uphold autonomy in the healthcare setting, we do not force patients to receive care if they have the capacity to deny it. For this reason, disclosure is not only relevant and beneficial, but also required.

**Addressing Moments of Absolute Crisis.**

Other ethical worries pervade discussions of communications of conversational automated technology at large: will chatbots be able to parse human colloquialisms and detect sarcasm? What if a human user falls into crisis wherein the user presents a real threat of harm to themselves or others – will police or EMTs be called on their behalf and sent to their homes? Are police or EMTs the correct resources to offer meaningful aid to human users who might need it? The stakes are high, and the answers to these questions matter. Because chatbots are to be applied in sensitive settings, questions of reaction to crisis carry an even heavier burden than they might if they were being used in recreational settings. It is the urgency to react appropriately that questions the capabilities chatbots have currently to pass the Turing test.* The Turing Test functions to capture technology that can convince users that it expresses the intelligence and

---

* For instance, lying to Nazi officers regarding Jewish individuals hiding in your home to protect their lives.

* The Turing Test is a test developed by Alan Turing to assess artificial intelligence’s ability to exhibit intelligent behavior. In order to pass the Turing Test, a system must be indistinguishable from a human being. No system as of August 2020 has yet to pass the Turing Test.
conversational capabilities of a human agent, and in doing so can fool its user into believing they are conversing with a human agent.\textsuperscript{45} There is contention in determining whether any program has yet passed the Turing test, as some believe that no program is genuinely intelligent in that it is able to think for itself, but rather has the information necessary for human users to believe it can. In this way, programs that are colloquially named AI are actually machine learning, “an algorithmic field that blends ideas from statistics, computer science and many other disciplines to design algorithms that process data, make predictions, and help make decisions”.\textsuperscript{46} As such, technology as it stands currently to the public eye cannot be considered AI, and there is still no unanimous consensus that any technology has yet passed the Turing Test. Further, there are guidelines that determine successfully passing the Turing Test on record, which include utilizing a test panel of judges, who must believe after 5 minutes of conversation that they are speaking with a human being, and not a computer program.\textsuperscript{47, 48}

The Loebner Prize has been awarded to developers of computer programs that have come closest to passing the Turing Test. Though no Loebner Prize winner has been determined officially to have passed the Turing Test, some have come close, such as Steve Worswick with his chatbot, “Mitsuku”.\textsuperscript{49} Five time Loebner Prize winner Worswick noted that many people who had interacted with Mitsuku did believe they were speaking to a real human being, and that he was deceiving them in claiming Mitsuku was in fact a chatbot.\textsuperscript{50} This is one way in which there is still contention as to whether a program has yet passed the Turing Test. This is also where the

\textsuperscript{45} From “Turing Test”, by J. Frankenfield, 2019.
\textsuperscript{47} From “Turing Test”, by J. Frankenfield, 2019.
\textsuperscript{48} Ibid.
\textsuperscript{49} From “Trust Me, I’m a Chatbot: How Artificial Intelligence in Health Care Fails the Turing Test”, by J. Powell, \textit{Journal of Medical Internet Research}, 2019, 21.
\textsuperscript{50} From “Ethics and Chatbots”, by S. Worswick, 2018.
distinction between AI as meaningfully intelligent as opposed to programmed to respond convincingly comes into play.

Though some programs might convince some people they are meaningfully intelligent, no technology currently is truly intelligent in this way.51 As a result, in moments of absolute crisis, chatbots might respond inappropriately, without employing sensitivities that are uniquely human. They do not know how to respond – and how could they? Their design and programming capabilities only go so far. Here, it is ethically relevant to note that when they fall short, people could get hurt. These resulting harms can be addressed by first identifying moments of absolute crisis. These moments of crisis arise chiefly when a human user poses a real threat to themselves or others. Even for human clinicians, it is difficult to know when a patient’s autonomy can and should be usurped because they pose a threat of harm to themselves or others. In this way, it is a highly nuanced decision requiring them to weigh complex evidence. AI’s inability to parse human colloquialisms and sarcasm would make this difficult task even more challenging, if not impossible. For this reason, chatbots would need to do two things: firstly, recognize initially when a human user might be falling into a moment of crisis, and secondly, employing the (albeit imperfect) judgement of a human agent.

Given that chatbots are required to utilize supervised learning, they can be programmed to recognize phrases that commonly signal need in times of crisis. Common phrases lifted from psychological and therapeutic reports could be used to teach the chatbot program that when a human user uses or persistently uses these phrases, they are likely experiencing a time of extreme crisis. If these key trigger phrases are used, the chatbot could flip over to a human user. That is, allow a human to continue the conversation rather than the computer program. Doing this

allows for a type of analysis that the program cannot currently offer. Some phrases might be misinterpreted by a chatbot that a human agent might recognize as sarcastic, or not serious. In moments of absolute crisis, reacting quickly and appropriately are critical. In these moments, the chatbot would allow a human agent to continue the conversation rather than the program. Doing so will allow the human agent to assess the situation, and recommend appropriate resources, or send the appropriate resources to the human user. Further, emergency services differ based on geographic location. Because of this, while a chatbot might recommend the number for a suicide hotline, the number may not be accessible to the user in a certain location, or may not be the appropriate resource in that moment.

Although flipping to a human user might diminish one of the key assets of chatbot use – extending the reach of the inadequate supply of trained mental health providers – doing so might offer more astute and rapid responses, and may mitigate some of these harms from taking place, or offer solutions where chatbots lack the ability to reflect and react effectively. That is, the person responding to the human user in crisis could quickly evaluate what type of support the human user needs, and could send them the aid deemed necessary, be that EMTs, first responders, or connecting them to appropriate hotlines. The ability to switch to a human responder, therefore, is a protective feature primarily for the human user, and one that might operate to mitigate or prevent some foreseeable negative outcomes.

**Ethical Basis for Avoiding Harms and Bad Consequences.**

Nonmaleficence, or the moral requirement not to cause harm and to prevent harm when possible, becomes crucial to uphold when considering the necessity of preventing harms that

---

might fall on a human user as a result of inadequate software development. Nonmaleficence necessitates that possible harms are foreseen and reduced, if not eliminated altogether. In many ways, all concerns I will raise in this paper rely on the principle of nonmaleficence (even if they also rely on other concepts and principles, such as respect for autonomy) to disentangle these concerns before permitting chatbot use.

Certainly, both physical and emotional harms are commensurate in concern when navigating this complicated and relatively new field. While many concerns circle around emotional harms, what is to be done about the possibility of physical harms that might occur if a chatbot program cannot detect or intervene in a moment of crisis? Here, it is important to delineate the responsibilities of a chatbot, given it is to be used in a healthcare setting. One guideline of ethical practice in healthcare comes by way of the Hippocratic Oath. Though arguably outdated, it is still recited in medical schools across the country, and its most basic guidelines of offering aid to patients (beneficence), and protecting patients against harms (nonmaleficence) are still ethically relevant today. In this way, nonmaleficence is sensible in both medicine and AI. In medicine, the Hippocratic Oath begins, “I will do no harm”, giving pride of place to nonmaleficence. In AI, incorporating nonmaleficence into the development of AI is also prudent in protecting human users from their potentially destructive capabilities.

Reduction of human harms is woven throughout the ethical and healthcare literature at large. It is echoed in the AMA Code of Medial Ethics, ANA Code of Ethics for Nurses, and the AMA Declaration of Professional Responsibility, a few of many medical codes.

54 From “Greek Medicine”, 2012.
56 From “AMA principles of medical ethics”, American Medical Association, 2016.
Minimizing harms is even mentioned in “classic” robot literature, as in Asimov’s First Law of Robotics, a set of fictional laws that dictate a robot must not harm a human being or allow a human being to come to harm.\(^{59}\) This declaration sounds strikingly similar to the principle of nonmaleficence and aligns itself with other declarations too, such as the Hippocratic Oath. Now we return to the question of physical harms, which are predicted to be less frequently problematic, but ethically fraught nevertheless. These are questions asking what ought to happen in a time of absolute crisis? And if a human user does become a threat of harm to themselves or others, or even dies as a result of inadequate intervention, who is at fault? In order to minimize bad consequences, these questions must be addressed before chatbots are permitted to be used in the healthcare setting.

**Moral Responsibility.**

It seems misguided to assign a chatbot moral blame, in that moral blameworthiness depends on a moral agent’s ability to think and reason, and make decisions that result in bad consequences. By contrast, chatbots are not moral agents – they are programs that cannot currently meaningfully think or react for themselves, and cannot currently comprehend morality. If a chatbot cannot currently comprehend morality or distinguish actions that result in good consequences from those that result in bad consequences, they are seen as less morally responsible than those who can make those distinctions.\(^{60}\) Further, while a chatbot may present as intelligent and able to process emotions and decisions, this is currently only in appearance, and not reflective of its capabilities as a moral agent (which it is not). Therefore, all conversations had

---


with a chatbot, as authentic as they may seem, are one-sided. In this way, it cannot currently be held morally responsible if bad consequences occur because it cannot detect or respond to human concern in the way another human being can. Chatbots cannot currently think critically or make connections between a human emotion and a social service that might benefit them unless they are told to do so.

Who then is responsible, and how can we prevent bad consequences from occurring? While human agents can reason morally, no one person controls the outcomes of a chatbot conversation, and programming a chatbot involves a team of agents. As such, we must determine who is to be held responsible when harms occur. While determining legal responsibility is an important aspect of chatbot consideration that will need to be explored in this context, it is out of the purview of this paper. Therefore, when I speak to responsibility, I refer here to moral responsibility only. As discussed earlier, switching from chatbot to human agent in moments of absolute crisis can function as protective for human users. The reason for this switch is to ensure the human user is receiving the attention of a trained agent who can suggest helpful resources or facilitate an intervention to prevent the harm from occurring. However, harms might still occur. In this case, it is one option to assign moral responsibility to all those involved in the genesis and maintenance of the chatbot, or we might assign moral responsibility to the human agent who began conversing with the human user after the switch. Morally, it is more plausible to assign responsibility to the programmers or company for not coding the chatbot to be compassionate or understanding enough, or to the human agent whose function is meant to mitigate harms before they occur, or prevent them from occurring altogether. For this reason, if harms do occur, we might examine how those harms came about and whether both the chatbot program and human agent were compassionate and resourceful in their attempts to prevent the reasonably foreseeable harm from occurring.
Privacy and Confidentiality.

Foreseeable bad consequences are not only physical, but could also be emotional, as in violations of trust, privacy, and confidentiality. The use of chatbots as a medical technology falls into a grey area that has not yet been discussed or established legally, where protections that otherwise would have been drawn up in order to define the scope of protections for users do not yet exist. Because of this, it is necessary to examine where a lack of legal protections could expose harms or breaches of privacy or confidentiality. Questioning where harms may fall is particularly important in this application, where human users are people who are likely to have higher prevalence rates of mental illness, or are vulnerable in that they are seeking or need mental health support. For the purposes of this paper, as is standard, privacy refers to private personal information and confidentiality refers to private medical information. Reduction and elimination of breaches in privacy and confidentiality are not only necessary in effecting competent therapeutic solutions, but also in upholding public trust in this technology as an effective branch of support. Further, because it is expected that human users will impart personal and potentially medical information in an uncontrolled setting, the candid expectation of this behavior occurring must be met with proper protocols for protection of this information. Laws such as HIPAA (Health Insurance Portability and Accountability Act) protect patient privacy and confidentiality in the medical setting, and are followed by clinicians and all other medical staff. However, in an informal setting (or virtual setting), where a person accesses a chatbot via their phone or computer to divulge personal information, what protections are in place for the human

---

63 From “Health Information Privacy Law and Policy”, The Office of the National Coordinator for Health Information Technology, 2018.
user? Privacy and confidentiality are values that the reasonable person holds highly, therefore it is reasonable to demand high ethical standards of regulation over technology that is entrusted with sensitive and personal information such that it is not shared with other parties for any reason other than intercepting an imminent harm to the user.

Other concerns fall when lack of legal regulations and protections may offer a space for third parties to intercept sensitive information. Many third parties have a vested interest in data collection, particularly personal data collection. A lack of regulation and legal protections certainly make it easier for third parties to obtain information that might be discussed in a chatbot conversation without the consent of the human user, or might allow them to extrapolate data from these conversations about chatbot users in aggregate. These concerns could regard privacy or confidentiality, but compromising either breaches or erodes public trust in this technology, whose purpose is to be therapeutic. Breaches in public trust could occur if conversations were recorded then shared, if personal information the human user offered to the chatbot was collected and stored for distribution, or if human users become confused about speaking with a chatbot, as opposed to a human agent in times of absolute distress. Even if conversations are not recorded for the purposes of teaching the chatbot and incorporating edited conversations for the purpose of supervised learning, other violations of trust might occur. Recall that one protection against harms that might occur in moments of crisis is to switch over to a human user who can make informed and rapid decisions regarding resources to offer the human user in distress. Though this is a protective feature, if a human user shares information with a chatbot only because they feel comfortable talking with the chatbot rather than a human agent, then the flip to a human agent in a moment of absolute crisis could be violating and

---

64 From “Your Data Is Shared and Sold…What’s Being Done About It?”, Wharton, 2019.
unwelcomed. This is another way in which privacy and confidentiality might not explicitly be violated, but that human users may feel violated nevertheless. One safeguard against a violation of trust in this way is to notify users when this flip occurs, and to make them aware that the flip from chatbot to human agent is possible only in moments of absolute crisis.

As mentioned earlier as an essential goal of public health and public health research, maintaining public trust is imperative. It is particularly important to ensure human users seeking support through chatbot conversations are protected against third party interests not only to uphold their own interests of privacy and confidentiality intrinsically, but also to uphold public trust as members who have agreed to participate in a new technology. Assurance that data will not be shared with any unwanted and non-consenting parties, and transparency regarding who will be able to collect data, as well as defining what sort of data is to be collected are all ways in which protection against other interested groups may be sustained.

**FDA Approval.**

The FDA is responsible for protecting public health by ensuring safety of food, drugs, and medical devices.\(^{65}\) The FDA is responsible for the governance of medical devices, a division under which chatbots have not yet been marketed. However, if chatbots were to be marketed in this way – as they would be in order to function the way I describe in this paper – they would be under the purview of the FDA.\(^ {66}\) Some chatbots function to offer CBT, a therapeutic tool used to treat individuals with depression or those who have gone through a traumatic event.\(^ {67}\) Unlike other therapeutic options, however, even CBT chatbots are not regulated by the FDA, as they

---

\(^{65}\) From “What We Do”, by U.S. Food and Drug Administration, 2018.


are sold for their entertainment and wellness value, as opposed to a therapeutic medical device. Marketing chatbots in this way also functions to avoid FDA regulations, however their evolving function as deliberately offering mental health support necessitates ethically that they are accurately marketed and pass the appropriate regulations. This function of chatbots as applications of mental health support undoubtedly requires FDA approval. Evading labeling and recognition in this way limits the legal protections that are able to be offered to users. Because FDA oversight generally looks only at safety and efficacy/effectiveness, avoiding FDA oversight leaves us without concrete evidence of safety and efficacy of chatbot interventions. While chatbot companies may be reluctant to define their claims, therapeutic misconceptions for a CBT chatbot, for example, seems highly likely. In this way, a chatbot might be marketed as a device used for wellness and therapy, though we do not know if it is truly a therapeutic intervention until we have the evidence that it is effective, or that it is a safe intervention.

A lack of oversight also exposes human users to data security concerns, wherein private and confidential information may be sold or shared with other companies without the consent or knowledge of human users. Though this is out of the purview of FDA regulations, this is another area of concern and angle for which the regulation of chatbots is called for. Developers of such technology and their stakeholders* have a moral responsibility to protect users’ information and autonomy, insofar as private information cannot be shared or commercialized. Concerns over shared private information or shared confidential medical information are present if there are not regulations to prevent this from happening. Therefore, introducing FDA regulations in harmony with other legal regulations for chatbots to protect data and security for human users may be one way to mitigate risks associated with safety and privacy concerns.

* Such as bioethicists involved in advising the development of these technologies.
**Injustices as a Result of Novel Nature.**

Imminent harms are possible, and at least some are likely to occur. At its base, there is an aspect of experimentation when using chatbots to offer mental health services. Using chatbots for this clinical application is in a way a novel therapy, as it has not been revolutionized in the way I propose before. The most serious risks are worrisome, particularly when the greatest harms are likely to fall on a population that is already classified as vulnerable.\(^{68}\) An element of discrimination and bias emerges by default as these technologies are developed. Injustices occur as chatbot programs have been found to be biased against certain groups.\(^{69,70}\) Minority groups, and groups of lower socioeconomic status are particularly likely to encounter difficulties communicating with chatbots as they are often misunderstood, in addition to being limited to just one language option for communication.\(^{71}\) This discrimination is particularly jarring as groups that are marginalized experience higher rates of mental health illnesses, and face an additional barrier to access when chatbots are discriminatory and biased.\(^{72}\) While chatbots are validated for their ease of access, this also may exclude people without internet access, or without smart devices – already a vulnerable population. Therefore, while chatbots increase access in offering a therapeutic avenue where one did not exist previously, justice calls for equity for all users, not only the least marginalized populations who are predisposed to benefit most.\(^{73}\)

---


\(^{71}\) From “We Need to Talk About Biased AI Algorithms”, by K. Matthews, 2018.


\(^{73}\) From “Principles of biomedical ethics”, by T. Beauchamp and J. Childress, 1983.
Discussion of Chatbots as Ethically Permissible.

Though specific concerns and harms have been drawn out and discussed, what of the larger question at hand? Is the use of chatbots ethically permissible at all, and ought they to be authorized on moral grounds to fill this sensitive role? This question is legitimate. Indeed, if the use of chatbots would lead to harms that would shroud their benefits, they should not be endorsed ethically. Even if benefits would only balance the harms, and would not greatly outweigh the harms, we might say their use is not ethically justified. For this reason, we must perform an analysis of the benefits and harms at large. Many signs point to the concept of the project as being in itself suspect, as it is largely untested, and has a number of possible shortcomings.* As unapproved interventions, there are risks and harms associated with a lack of oversight – however this novel technology holds great promise. The blunt reality is that we do not have the full suite of information. This makes the project of permitting chatbots to serve as a therapeutic option for care investigational (though their scaled-up use will follow FDA approval and trials). Given that chatbots in this application are meant to serve the need for therapeutic interventions, FDA approval will be needed, necessitating trials to determine precisely how serious the risks of this intervention are and how tangible the benefits are. With proposed best practices and governance, is it likely that the actual benefits of applying chatbots clinically outweigh their possible harms. Treating chatbots in this way, with a level of skepticism and more importantly, caution, works to ensure that human users are protected from possible harms and bad consequences, and respected in this way as participating in using a novel technology.

* Though all novel interventions are, at some point, largely untested prior to approval.
This technology is indeed novel. As such, its use strikes many as alarming and raises legitimate concerns. I do not argue that these reactions are inappropriate or even misguided. They are very much justified as the use of chatbots for mental health support attempts to fill healthcare gaps in a sensitive space, and this caution is a reaction to concerns over patient safety. However, investigational therapies require evidence in order to be approved. Here, two levels of evidence would be required through this process. Firstly, preliminary evidence would be needed to show that this intervention could work and might be safe in order to justify clinical trials. Secondly, clinical trial evidence would be needed to approve the chatbot for clinical use. Currently, we are at the first level, where in this context, evidence of potential for success is data pulled from chatbot use in an informal entertainment and wellness setting. However, chatbot use as presented in this paper implies passage through both levels of this process.

Moving forward with chatbot use in this space is novel, though safeguards can be applied as they are in other settings to ensure participant safety and respect are upheld. In this way, I advocate for moving forward with chatbot use for this application, but cautiously. Further, thinking of ways to mitigate bad consequences or harms is also an important feature of research that ensures preparedness works in favor of those who commit themselves to test this technology. Testing this therapy and seeking FDA approval will either confirm or reject the hypothesis that chatbot use in this space can offer great benefits and fill an immense gap in this field. If the hypothesis fails, then it is in line with the ethical goals that inspired this experiment to terminate it. However, if the hypothesis is supported, and supported repeatedly, this area of need may see immense benefits from the application of chatbots for mental health support. The stakes are high, and the actual implementation of chatbots are more than just a promise of improved care – they are the avenue through which mental health care can be revolutionized.
Ideally, all who sought mental health support would be able to access a highly trained mental health provider, though we know that this ideal scenario is far from our reality. In order to take seriously the reality of the world, in which significant injustice and imperfection exists, we must entertain solutions to the barriers we face in this arena. Because everyone cannot receive their own mental health provider, chatbots are likely to be an improvement to our world, offering aid where we lacked aid before. Although there are uncertainties in their development, the benefits are likely to outweigh their harms, and a strong ethical consideration of these possible harms can guide their ethically sound development as effective tools used in health care.
Works Cited:


Fitzpatrick KK, Darcy A, Vierhile M, Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled TrialJMIR Mental Health 2017;4(2):e19


Johannsen, Florian; Leist, Susanne; Konadl, Daniel; and Basche, Michael, "COMPARISON OF COMMERCIAL CHATBOT SOLUTIONS FOR SUPPORTING CUSTOMER INTERACTION" (2018). *Research Papers.* 158. https://aisel.aisnet.org/ecis2018_rp/158


Powell J. Trust Me, I'm a Chatbot: How Artificial Intelligence in Health Care Fails the Turing Test Journal of Medical Internet Research: 2019;21(10):e16222


Your Data Is Shared and Sold…What’s Being Done About It?. *Knowledge@Wharton* (2019, October 28). Retrieved from https://knowledge.wharton.upenn.edu/article/data-shared-sold-whats-done/
Biographic Statement.

Tiana Y. Sepahpour was born in 1997 in the United States. Tiana completed her undergraduate studies at Fordham University in the Bronx, NY. During this time, she earned a B.A. in Environmental Studies as well as Philosophy, and became interested in public health ethics in relation to the changing global climate. Resultingly, she wrote her senior thesis, “Ethics of Population Growth and Reduction”. In 2019, Tiana began her MBE at the Berman Institute of Bioethics at Johns Hopkins University. During this time, she became involved with the IRB office at the Bloomberg School of Public Health, grant writing for non-profits, and tutoring in the Baltimore community. During her tenure at Johns Hopkins, she also became interested in the department of Mental Health and completed her practicum on a project involving virtual humans, which led her to the topic of her thesis, “Ethical Considerations of Chatbot Use for Mental Health Support”.

33