Viewpoint

A Medical Ethics Framework for Conversational Artificial Intelligence

Eleonore Fournier-Tombs[1*], PhD; Juliette McHardy[2*], LLM

[1]United Nations University Centre for Policy Research, New York, NY, United States

[2]London School of Economics, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Eleonore Fournier-Tombs, PhD
United Nations University Centre for Policy Research
767 Third Avenue, Floor 35
New York, NY, 10017
United States
Phone: 1 646 905 5225
Email: fourniertombs@unu.edu

Abstract

The launch of OpenAI's GPT-3 model in June 2020 began a new era for conversational chatbots. While there are chatbots that do not use artificial intelligence (AI), conversational chatbots integrate AI language models that allow for back-and-forth conversation between an AI system and a human user. GPT-3, since upgraded to GPT-4, harnesses a natural language processing technique called sentence embedding and allows for conversations with users that are more nuanced and realistic than before. The launch of this model came in the first few months of the COVID-19 pandemic, where increases in health care needs globally combined with social distancing measures made virtual medicine more relevant than ever. GPT-3 and other conversational models have been used for a wide variety of medical purposes, from providing basic COVID-19–related guidelines to personalized medical advice and even prescriptions. The line between medical professionals and conversational chatbots is somewhat blurred, notably in hard-to-reach communities where the chatbot replaced face-to-face health care. Considering these blurred lines and the circumstances accelerating the adoption of conversational chatbots globally, we analyze the use of these tools from an ethical perspective. Notably, we map out the many types of risks in the use of conversational chatbots in medicine to the principles of medical ethics. In doing so, we propose a framework for better understanding the effects of these chatbots on both patients and the medical field more broadly, with the hope of informing safe and appropriate future developments.

(J Med Internet Res 2023;25:e43068) doi: 10.2196/43068

KEYWORDS

chatbot; medicine; ethics; AI ethics; AI policy; conversational agent; COVID-19; risk; medical ethics; privacy; data governance; artificial intelligence

Introduction

With the launch of OpenAI's GPT-3 model in June 2020 came a new era for conversational chatbots [1]. While there are chatbots that do not use artificial intelligence (AI), conversational chatbots integrate AI language models that allow for back-and-forth conversation between an AI system and a human user. GPT-3, since upgraded to GPT-4, harnesses a natural language processing technique called sentence embedding and allows for conversations with users that are more nuanced and realistic than ever before. The launch of this model came in the first few months of the COVID-19 pandemic, where increases in health care needs globally combined with social distancing measures made virtual medicine more relevant. GPT-3 and other conversational models have been used for a wide variety of medical purposes, from providing basic COVID-19–related guidelines to personalized medical advice and even prescriptions. The line between medical professionals and conversational chatbots is somewhat blurred, notably in hard-to-reach communities where the chatbot replaced face-to-face health care [2].

Considering these blurred lines and the circumstances accelerating the adoption of conversational chatbots globally, we analyze the use of these tools from an ethical perspective. Notably, we map out the many types of risks in the use of conversational chatbots in medicine to the principles of medical

ethics. In doing so, we propose a framework for better understanding the effects of these chatbots on both patients and the medical field more broadly, with the hope of informing safe and appropriate future developments.

The Use of Conversational Chatbots During the COVID-19 Pandemic

During the COVID-19 pandemic, many different types of conversational chatbots were developed. This was both triggered by and an enabler of increased social distancing requirements that also helped support overburdened health systems. This has also allowed public health actors to respond to the "infodemic" of health-related misinformation that has co-occurred with the pandemic with evidence-based health messaging delivered on the same platforms as the misleading or false information [3]. During the pandemic, these chatbots have been useful for disseminating preventive- and vaccine-related messaging, and as tools for triaging, guiding treatment, monitoring symptoms, and providing mental health support for those social distancing or isolating at home [4]. The World Health Organization (WHO) rapidly provided access to its global alerts system via chatbot interfaces on Whatsapp, Facebook, and Viber. It later followed up these efforts by updating its tobacco use cessation virtual assistance, Florence, to provide COVID-19–related advice. The WHO European Regional Office launched, in partnership with UNICEF (United Nations Children's Fund), HealthBuddy+ to both provide information and allow users to report disinformation and give opinions on the pandemic [5].

Almalki and Azeez [6] had already at the beginning of the pandemic listed nine such uses. A later review found 61 chatbots deployed in response to COVID-19 in 30 countries across areas such as risk assessment, disease surveillance, and information dissemination [7]. Albites-Tapia et al [8] found chatbots being used for the screening and detection of COVID-19 symptoms outside of the health sector including by education providers, retailers, banks, and tourist operators—with 64 cases noted.

Ethics and Risks in Chatbots for Medicine

Several ethical risks have been documented in conversational chatbots. These include risks related to human rights, such as discrimination, stereotyping, and exclusion; risks related to data, including privacy, data governance, and stigma [9]; and technical risks, such as error tolerance, overconfidence in chatbot advice and decay of trust in health professionals, and, more broadly, technological solutionism [10].

The human rights–related risks are addressed in several recent AI standards. For example, the European Union's draft AI Act, which was published in April 2021 [11], refers to eight applications of AI that are at higher risk for discrimination: biometric identification; management and operation of critical infrastructure; education and training; employment; access to essential services; policing; migration, asylum, and border management; and administration of justice and democratic processes. It should be noted that AI in health care is not explicitly listed here but is covered elsewhere in the text and in earlier legislation on medical devices [12]. UNESCO (United Nations Educational, Scientific and Cultural Organization) has also developed its Recommendation on the Ethics of Artificial Intelligence, adopted by United Nations member states in December 2021. This document refers to AI in health care and being sensitive to human rights, which member states should monitor quite closely [13]. Going into more depth, the WHO has published guidance on the Ethics & Governance of AI for Health, in which it discusses several risks in medical chatbots, notably in relation to discrimination and privacy [3].

Conversational AI chatbots have several characteristics that could, if improperly used, increase these risks for vulnerable populations. Some of these risks apply to all initiatives collecting data, especially patient data, such as data governance and privacy [14]. Others apply to all AI models, namely, biases in training data, which could lead to the marginalization of certain groups, exclusion of groups in the development and governance of the tool, and error tolerance [15]. Other risks, finally, are unique to the type of AI used, which is natural language processing. Risks in this domain exist in both the interpretation of the input text as well as the construction of the response. Researchers have found many examples of gender and racial stereotyping in GPT-3 and other natural language processing models, which have not yet been corrected by the model owners [16,17]. Some of these risks exist also in other medical applications of AI. However, conversational AI is unique in that it also features specific risks related to large language models [18]. In Table 1, we summarize and illustrate these risks.

Table 1. Known risks in conversational chatbots.

Risk category and description	Definition
Human rights	
Discrimination	The chatbot makes different recommendations or has a higher error rate based on the patient's group (gender, ethnicity, race, religion, etc).
Stereotyping	The chatbot interprets or uses language that propagates harmful prejudices, such as the inferiority of certain groups, sexualization, or lack of credibility.
Exclusion	Development, governance, or use of the chatbot does not include certain already marginalized groups.
Data protection	
Lack of privacy	The data generated by the chatbot is not protected.
Poor data governance	The data generated by the chatbot is governed improperly or without including the patient.
Stigma	The data generated by the chatbot can lead to stereotyping or marginalizing certain individuals.
Technical	
Error tolerance	Errors, even if they are not discriminatory, cause harm to the patients.
Overconfidence and trust decay	Patients place excessive trust in chatbots resulting in overconfidence and relative decay of trust in human health professionals.
Technological solutionism	Investment in chatbot technology diverts from an actual societal problem.

A Hippocratic Oath for Chatbots

The Hippocratic Oath has undergone many versions and modifications throughout the history of the medical profession. After World War II, a more streamlined version was adopted by the World Medical Association, which was rewritten in 1964 and adopted as the current version in many medical schools globally, although not without some criticism [19]. Broadly, it contains four principles that all health practitioners must adhere to [20]. These principles, therefore, make up the backbone of accepted norms for health professions in many settings and are generally similar to alternative formulations of the leading principles to be applied in medical ethics [19]. Textbox 1 below summarizes these four principles along with their definition.

Textbox 1. Principles of medical ethics.

Beneficence

- Acting for the benefit of patients and promoting their welfare

Nonmaleficence

- Not harming the patient

Autonomy

- Respecting the patient's right to and capacity for self-determination (this includes informed consent, truth-telling, and confidentiality)

Justice

- Treating patients in a fair and equitable manner

Illustratively, these would play out in conversational artificial intelligence by mitigating the risks such that a chatbot would be able to provide appropriate medical advice without bias or stereotypes, or any of the other risks described.

As we have discussed, the ethical risks of conversational chatbots in medicine have not been mapped out to these principles of medical ethics. However, as we have seen with the recent development of GPT-3, conversational AI is becoming increasingly detailed and realistic, for example, the ability to pass the Turing test [21]. In the deployment of chatbots at scale, in particular during health emergencies, the ethical imperatives of public health focusing on population (rather than individual) health may appear more relevant from the perspective of those designing, commissioning, and delivering them. This is because they will see themselves as institutions delivering often preventative information to large groups of people [22]. From a user point of view, however, these chatbots will often appear to be individual-level interactions and, in certain cases, may substitute partially or entirely for any physician or health practitioner interaction. In analyzing the ethical implications of chatbots, it is necessary to prefer the insider perspective of intended users and the way they will likely construe the interaction [23]. Accordingly, medical ethics may provide a more appropriate framework, which will only become more applicable as AI chatbots grow increasingly realistic and capable

of assisting tasks conventionally performed only by health practitioners.

In the section below, we map out the main risks of conversational chatbots for medicine as they relate to the principles of medical ethics. We find that each risk can be related to at least one principle. For example, errors in medical chatbots can lead to harm if they make recommendations, diagnoses, or prescriptions that are wrong. The harm from incorrect diagnoses can then be compounded when chatbots are able to instill such trust in patients that they are unduly confident in the diagnosis, and human health professionals find displacing these erroneous diagnoses in the mind and actions of the patients challenging or impossible [24]. Discrimination can similarly cause harm to certain groups, as well as contravene the principle of justice, since it leads to patients not being treated in a fair and equitable way. Stereotyping similarly leads to direct harm, as discrimination does, and can lead to secondary or societal harms, which might go beyond the medical question (as does stigma). Exclusion is linked to beneficence, in that those that are not represented by the chatbot or cannot use it are not able to access its benefits. Stigma, like stereotyping, can cause harm beyond the immediate medical condition by affecting the patient's position in society. Lack of privacy and poor data governance can affect the patient's capacity for self-determination, as well as their right to confidentiality. Overconfidence in technology and trust decay can lead to a lack of adherence to physician guidelines, leading to ill health. Finally, technological solutionism can impact the patient's ability to receive good care by other means by diverting funds better used to improve in-person health services or address social determinants of health. In Table 2 below, we describe an illustrative relationship between the principles of medical ethics and the risks of conversational chatbots in medicine while also acknowledging that each risk may have a bearing on all of the principles, depending on the implementation.

Table 2. Illustration of the framework for chatbot risks to the principles of medical ethics.

Risks	Ethical principles			
	Beneficence	Nonmaleficence	Autonomy	Justice
Errors	—[a]	The chatbot makes the wrong recommendation to patients based on a bug in the system.	—	—
Discrimination	—	The chatbot has a bias that causes it not to understand requests based on women's health.	—	The chatbot provides more appropriate recommendations for men than for women.
Stereotyping	The chatbot responds to the patient in derogatory terms.	The chatbot's recommendations based on stereotypes lead to harm the patient.	—	The chatbot gives unfair and derogatory responses to patients.
Exclusion	—	The chatbot excludes certain users because of language and literary skills, withholding medical support.	—	The chatbot excludes certain patients and no alternative is provided.
Stigma	—	Use of the chatbot is not anonymous and leads to stigmatization of certain patients.	—	—
Lack of privacy	—	—	There are data leaks from the chatbot system leading to a breach of confidentiality.	—
Poor data governance	—	—	Patients do not consent to have their data collected by the chatbot and mechanisms for data governance are not clear.	—
Overconfidence and trust decay	—	The chatbot harms the relationship between the patient and their physician by providing contradictory recommendations.	—	—
Technological solutionism	A chatbot is not the best option for providing medical recommendations to certain patients.	—	—	—

[a]Not applicable.

Applications and Limitations of the Model

This paper provides a simple yet comprehensive framework for the use of conversational chatbots in the health sector. It addresses the extraordinary developments of the last few years in AI conversations and increasing reliance on them due to COVID-19 as well as the likelihood that chatbots will increasingly be used to dialogue directly with people in medical and other health contexts.

In terms of applicability, this framework could be adapted based on locally appropriate norms in medical ethics to underpin an impact assessment process. The use of this process would then be required in assessing and monitoring the deployment of chatbot technology in any circumstance comparable to that of a patient-physician relationship. Concretely, to implement the framework, Table 2 would be used as a guide for practitioners seeking to implement a conversational chatbot, allowing them to reflect on each intersection of risk and principle to consider how this might apply to their tool. This would allow them to consider risks more thoroughly and find solutions to mitigate them before deployment.

As regulatory systems covering AI develop in sophistication to match or exceed what was proposed in the EU Draft AI Act, the results of these medical ethics assessments for chatbots could be required as one component of the reporting requirements for high-risk AI. It is also conceivable that similar medical ethics assessments may be required or beneficial for other deployments of AI in the health sector. However, it should also be emphasized that in other areas, such as when AI is a tool used with a health professional's mediation, other ethical frameworks such as those of public health or professional responsibility may be more appropriate.

Conclusion

Over the last few years, conversational chatbot use has increased, driven by a general movement toward the digitalization of health care and public health considerations such as social distancing and remote accessibility. The technology behind conversational chatbots has substantially improved too, first with the Bidirectional Encoder Representations From Transformers (BERT) model developed by Google and more recently with OpenAI's GPT-3 model, which allows for extremely nuanced and realistic conversations with an AI agent.

At the same time, efforts globally have been made to understand the ethical use of conversational AI, and much research has gone into understanding possible biases, stereotypes, and other uses. Governments globally are in the process of developing regulations that will account for risks in AI technologies to mitigate them. AI regulation, however, does not happen in a vacuum. It will be inspired by existing human rights frameworks, as well as regulations in other domains, such as the European Union's regulation of medical devices.

The ethical principles of medicine highlighted here in the form of the Hippocratic Oath have informed many regulations around medicine and medical tools globally. It is therefore our hope that this paper will serve to inform the development of a stronger connection between AI ethics and an underlying medical ethics framework, to feed into stronger and more appropriate regulations, and to inform the risk assessment of individual tools.

Conflicts of Interest

None declared.

References

1. GPT-3 API. OpenAI. URL: https://openai.com/api/ [accessed 2023-06-02]
2. Azevedo Chagas B, Ferreguetti K, C Ferreira T, S Marcolino M, B Ribeiro L, S Pagano A, et al. Chatbot as a telehealth intervention strategy in the COVID-19 pandemic. Latin Am Center Inform Stud Eletronic J 2021 Dec 13;24(3):1-17 [doi: 10.19153/cleiej.24.3.6]
3. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2023-06-02]
4. Zhu Y, Janssen M, Wang R, Liu Y. It is me, Chatbot: working to address the COVID-19 outbreak-related mental health issues in China. User experience, satisfaction, and influencing factors. Int J Hum Computer Interaction 2021 Nov 01;38(12):1182-1194 [doi: 10.1080/10447318.2021.1988236]
5. HealthBuddy+: access to trusted information on COVID-19 in local languages using an interactive web- and mobile-based application. World Health Organization. 2022. URL: https://cdn.who.int/media/docs/default-source/science-translation/case-studies-1/cs12_healthbuddy.pdf?sfvrsn=369de46a_4 [accessed 2023-06-02]
6. Almalki M, Azeez F. Health chatbots for fighting COVID-19: a scoping review. Acta Inform Med 2020 Dec;28(4):241-247 [FREE Full text] [doi: 10.5455/aim.2020.28.241-247] [Medline: 33627924]
7. Amiri P, Karahanna E. Chatbot use cases in the Covid-19 public health response. J Am Med Inform Assoc 2022 Apr 13;29(5):1000-1010 [FREE Full text] [doi: 10.1093/jamia/ocac014] [Medline: 35137107]
8. Albites-Tapia A, Gamboa-Cruzado J, Almeyda-Ortiz J, Lázaro AM. Chatbots for the detection of Covid-19: a systematic review of the literature. Int J Adv Computer Sci Applications 2022;13(4):A [doi: 10.14569/IJACSA.2022.01304113]
9. Hamdoun S, Monteleone R, Bookman T, Michael K, Michael K. AI-based and digital mental health apps: balancing need and risk. IEEE Technol Soc Mag 2023 Mar;42(1):25-36 [doi: 10.1109/mts.2023.3241309]
10. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. NPJ Digit Med 2020;3:65 [doi: 10.1038/s41746-020-0280-0] [Medline: 32377576]
11. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. EUR-Lex. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 [accessed 2023-06-02]
12. Vokinger KN, Gasser U. Regulating AI in medicine in the United States and Europe. Nat Mach Intell 2021 Sep;3(9):738-739 [FREE Full text] [doi: 10.1038/s42256-021-00386-z] [Medline: 34604702]
13. Ethics of artificial intelligence. UNESCO. 2021. URL: https://en.unesco.org/artificial-intelligence/ethics [accessed 2023-06-02]
14. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [doi: 10.1038/s41746-020-00362-8] [Medline: 33398052]
15. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38 [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
16. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. 2016 Presented at: Advances in Neural Information Processing Systems 29; December 5-10, 2016; Barcelona, Spain
17. Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. 2021 Presented at: Third Workshop on Narrative Understanding; June 2021; Virtual [doi: 10.18653/v1/2021.nuse-1.5]

18. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual Event, Canada p. 610-623 [doi: 10.1145/3442188.3445922]
19. Huxtable R. For and against the four principles of biomedical ethics. Clin Ethics 2013 Sep 10;8(2-3):39-43 [doi: 10.1177/1477750913486245]
20. Gillon R. Medical ethics: four principles plus attention to scope. BMJ 1994 Jul 16;309(6948):184-188 [FREE Full text] [doi: 10.1136/bmj.309.6948.184] [Medline: 8044100]
21. Elkins K, Chun J. Can GPT-3 pass a writer's turing test? J Cultural Analytics 2020;5(2):1-16 [doi: 10.22148/001c.17212]
22. Swain G, Burns KA, Etkind P. Preparedness: medical ethics versus public health ethics. J Public Health Manag Pract 2008;14(4):354-357 [doi: 10.1097/01.PHH.0000324563.87780.67] [Medline: 18552646]
23. Kerr OS. The problem of perspective in internet law. Georgetown Law J 2003;91:357 [doi: 10.2139/ssrn.310020]
24. Parviainen J, Rantala J. Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos 2022 Mar;25(1):61-71 [FREE Full text] [doi: 10.1007/s11019-021-10049-w] [Medline: 34480711]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations From Transformers
UNESCO: United Nations Educational, Scientific and Cultural Organization
UNICEF: United Nations Children's Fund
WHO: World Health Organization

Edited by T Leung, V Arnold, H Gouda; submitted 29.09.22; peer-reviewed by J Parviainen, W Cheng; comments to author 08.12.22; revised version received 31.03.23; accepted 13.04.23; published 26.07.23

Please cite as:
Fournier-Tombs E, McHardy J
A Medical Ethics Framework for Conversational Artificial Intelligence
J Med Internet Res 2023;25:e43068
URL: https://www.jmir.org/2023/1/e43068
doi: 10.2196/43068
PMID: 37224277