Why Conversational AI won't replace Healthcare Providers

Rashmi Gangadharaiah
Amazon, AWS AI
East Palo Alto, CA, USA
rgangad@amazon.com

Taha Kass-Hout
Amazon, AWS AI
Seattle, WA, USA
tahak@amazon.com

Chaitanya Shivade
Amazon, AWS AI
East Palo Alto, CA, USA
shivadc@amazon.com

Yi Zhang
Amazon, AWS AI
Seattle, WA, USA
yizhngn@amazon.com

Parminder Bhatia
Amazon, AWS AI
Seattle, WA, USA
parmib@amazon.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Advances in Artificial Intelligence (AI) will help automate expensive and laborious tasks with ever greater accuracy and throughput. Conversational Intelligence, the ability by which humans connect, engage, problem solve and navigate with others is critical for healthcare providers to help patients reach good outcomes. The rapid rise in the availability of data in healthcare has raised the promise that research in Conversational AI can be applied to replace medical experts with artificial agents or bots. We argue that conversational AI will increase efficiency and accuracy of medical experts and will create new avenues for jobs that never existed before and will not replace healthcare providers.

Author Keywords
Conversational AI in Healthcare; Decision Support Systems; Chatbots

CCS Concepts
•Human-centered computing $\rightarrow$ Human computer interaction (HCI);

Introduction
Conversational AI encompasses a wide range of fields, including, speech recognition, computer vision, text analytics, summarization, translation and reasoning making conversational AI one of the most challenging and important tasks

in AI. We are interested here in the impact of advances in Machine Learning (ML) and AI on how healthcare is provided. In the early days of AI, interactive clinical decision support systems employed AI built on rules [28]. ELIZA[25] and PARRY[7] used Natural Language Processing (NLP) techniques to interact with humans via a text interface. Although these systems showed promising results, they were limited by the static knowledge bases powering them. Researchers have continued to invest significant effort in decision support systems, especially in diagnosis and treatment suggestions or recommendations [8, 3, 18, 12] [1]

Dialog systems [4, 30], typically contain several modules- (1) Natural Language Understanding (NLU) component is responsible for understanding the user's intent, (2) A Dialog State Tracker component is responsible for maintaining and updating the state of the dialog (3) the Dialog Policy is responsible for selecting the next best action, (4) The Natural Language Generation component then converts the action into a natural language response, optionally, using information from knowledge bases. In addition, these systems can also include a Speech (or other modalities such as vision) Recognition module and a Text-to-Speech system. Each of these components can be rule-based, deep/machine learning-based or a combination. Milabot [21], XiaoIce [29], CleverBot [2], Mitsuku and Gunrock [27] do provide human-like responses but rely on such individually trained components. More recent approaches use deep learning models to create end-end trainable systems [11, 5, 14]. These systems rely on the assumption that appropriate follow up responses have been observed in the training data. These systems are attractive since they do not require expert knowledge about the domain or specially annotated data, and replace many of the separated modules in the traditional dialog systems. Chatbots such as Meena or GPT2 [1], (although not specifically applied in Healthcare), appear to converse as humans do.

Given these advances in Conversational AI, it is now imperative to define the properties of a good conversational AI solution for healthcare applications. When a patient presents their chief complaint, a good conversational AI should collect all pertinent information, by interacting with the patient. Based on this information, the system should present a plan with next steps, including treatments. When uncertain, the system should request additional information for assessment. For example, an agent may look for lab tests to differentiate a bacterial infection from a viral one in order to triage or diagnose. If the system is still uncertain, it may need to then default back to a human expert. Hence, in addition to 'conventional' ML requirements, the properties of a good conversational AI include the need to 1) be easily extensible with advances in ML and AI, research into new diseases, outbreaks, 2) incrementally learn with a human-in-the-loop and 3) understand when there is model uncertainty and should fall back to an expert. In the following section, we list some of the challenges in achieving these requirements. We claim that despite these challenges, conversational AI can continue to help patients and healthcare providers.

Factors that inhibit Expert Replacement

Challenges in Speech Recognition

While Automatic Speech Recognition (ASR) systems have made great strides, their performance in specific domains having distinct vocabularies may not always generalize [10]. The ASR system design is further complicated since it is typically followed by a module identifying speakers associated with specific utterances, as well as detecting the role

[1] https://www.isabelhealthcare.com/products/symptom-checker, https://www.curai.com/, https://www.saykara.com

[2] 3https://www.cleverbot.com/

of the speaker (doctor or a patient in healthcare settings) [9]. ASR performance typically drops in noisy environments and results are sensitive to the quality and arrangement of microphones [24]. These factors are hard to control in settings where users interact with conversational agents, e.g. over mobile phones. Finally, ASR systems can only recognize speech. Non verbal indicators of communication are crucial in medical practice[13] and are difficult to quantify without a clinical examination.

Healthcare requires more than just detection
Most of the machine learning applications today have been narrowly focused, specializing on repetitive tasks that machines perform well with greater accuracy and speed.However, medical experts use information from different sources, and knowledge gained over several years in the field play a major role in arriving at an answer. Until AI models are capable of factoring in all scenarios of complex cases, conversational systems will be relegated to providing recommendations to a medical expert, with the ultimate decision taken by the medical expert [23].

Bias and Transparency
Adoption and success of AI technologies requires patient and provider trust in the AI algorithms. Understanding how the models arrived at a result (model interpretation), is a key challenge that machine learning researchers are faced with. Note that the definition of interpretability is itself unclear. Unless these challenges are successfully tackled, technology developers and adopting hospitals will remain anxious, limiting wide-spread adoption [20]. Another artifact of using data to train algorithms, is the tendency for algorithms to reflect and sometimes magnify biases in data. For instance, a widely used algorithm takes healthcare costs as a proxy for health conditions. This choice resulted in a system that was racially biased, falsely indicating that black patients tend to be sicker than white [19]. Since model performance completely depends on the data input, expert oversight is needed to check the results.

Lack of Representative Data
One of the key ingredients to the success of deep learning models in recent times is the availability of large corpora of *labeled* training data. Owing to privacy concerns, healthcare data cannot be hosted on crowd-sourcing platforms for collecting annotations. Often, the annotation task at hand also requires domain expertise. Thus collecting accurate clinical data is time-consuming and expensive. Another level of complexity comes from the large number of medical conditions that exist today. As a result of these challenges, clinical training data is never complete and cannot cover all possibilities of care for a problem. This often results in sparse representation for many situations. Thus, a good conversational AI system will again need to fall back on experts when presented with an edge case or if it is uncertain about a decision and collect additional data or have an expert weigh in.

Integration difficulties and Accountability for errors
Translating a technology from development to point of care involves meeting regulatory compliance [22]. This entails accurate assessment and education of users about the risks and benefits of the technology. Data collected by agents in healthcare domain is sensitive and has to follow privacy laws of respective countries. Often this results in silo-ed data present across hospitals, vendors, imaging facilities, and patients, without a central repository. Analogous to the automobile industry, where self-driving cars elicit changes in automobile regulation and insurance, changes in healthcare will also evolve through deliberation of policies [20].

Despite recent advance in deep learning leading to some claims of super-human performance in specific clinical application tasks, the nature of these models means that they are never free of error. Important decisions such as responsibility for adverse events in case of technology failure will always remain relevant mandating human involvement. In a clinical application, the consequences of such errors, even when rare, can be severe. And it is unclear who should be held accountable for such errors.

Future of Conversational AI in Healthcare

Today, deep learning allows us to build better technology than ML did before. This section lists some of the avenues where Conversational AI can aid medical experts.

Reduce medical errors

Conversational AI can provide assistance to doctors in the form of diagnosis, treatment recommendations, and safety alerts. In contrast to retrospective human chart review, automated systems are real-time, adaptive and can improve detection capabilities of safety monitoring systems significantly [17]. Medical experts spend substantial amounts of time on regulatory and administrative work such as clinical documentation. This has led to burnout and increase in medical errors [26]. Automated language understanding of the conversations can help reduce the overload on medical experts allowing them to concentrate on more superior decision making tasks that require human expertise [15].

Improve patient outcome and experience

Conversational AI that can help patients in understanding their ailments, reduces time lags that exist between when a patient becomes sick and the time at which she is actually treated. Conversational AI supported remote monitoring can eliminate unnecessary appointments. Such systems can also assist patients or the elderly at home and even go to the extent of making emergency calls under critical conditions. By using AI supported tools to collect and manage data, health care providers can spend more time caring for the patients who need them the most.

Improve efficiency and reduce costs

Conversational systems can be used for triaging purposes, where the system interacts with the patient to understand his or her ailments and use predictive models to decide the severity of the conditions. This not only helps predict the number of hospital admissions, it also reduces costs of hospital admissions, thereby improving efficiency in hospitals [6]. Nurses and other medical experts can then concentrate on critical cases that require immediate care.

Create more jobs

With the adoption of AI technologies in healthcare, new jobs that require both technology expertise and medical expertise will arise [16]. A growing number of adults in the US are using the internet to diagnose their medical concerns.[3] With appropriate conversational agents in place, these individuals will be automatically screened for medical problems and triaged to healthcare professionals for further care. There will be an increased demand for medical professionals to treat these screened patients [2].

Conclusion

In this paper, we discussed potential that Conversational AI can offer in transforming many aspects of patient care. We believe AI systems will not replace medical experts but rather augment their efforts and make them more efficient. AI will empower patients and medical practitioners with powerful tools and as a result, healthcare will no longer be confined within the walls of medical facilities.

[3]https://www.beckershospitalreview.com/healthcare-information-technology/google-receives-more-than-1-billion-health-questions-every-day.html

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. (2020).

[2] Sanjeev Arora, Karla Thornton, Glen Murata, Paulina Deming, Summers Kalishman, Denise Dion, Brooke Parish, Thomas Burke, Wesley Pak, Jeffrey Dunkelberg, and others. 2011. Outcomes of treatment for hepatitis C virus infection by primary care providers. *New England Journal of Medicine* 364, 23 (2011), 2199–2207.

[3] Hupp JA Hoffer EP. Barnett GO, Cimino JJ. 1987. DXplain. An evolving diagnostic decision-support system. *JAMA* (1987).

[4] S. Bennacef, L. Devillers, S. Rosset, and L. Lamel. 1996. Dialog in the RAILTEL Telephone-Based System. In *In Proceedings of International Conference on Spoken Language Processing, ICSLP'96*. 550–553.

[5] Antoine Bordes and Jason Weston. 2016. Learning End-to-End Goal-Oriented Dialog. *CoRR* abs/1605.07683 (2016). http://arxiv.org/abs/1605.07683

[6] Young Moon Chae, Joo Heon Lee, Seung Hee Ho, Hee Ja Kim, Ki Hong Jun, and Jong Uk Won. 2001. Patient satisfaction with telemedicine in home health services for the elderly. *International journal of medical informatics* 61, 2-3 (2001), 167–173.

[7] M.D. Kenneth Mark Colby. 1975. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Pergamon.

[8] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 3 (2010), 59–79.

[9] Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikanth Narayanan. 2018. Combined Speaker Clustering and Role Recognition in Conversational Speech.. In *Interspeech*. 1378–1382.

[10] Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of biomedical informatics* 35, 4 (2002), 222–235.

[11] Rashmi Gangadharaiah, Balakrishnan Narayanaswamy, and Charles Elkan. 2018. What we need to learn if we want to do and not just talk. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. 25–32.

[12] Shortliffe EH Greenes RA. 2009. Commentary: Cognitive informatics in biomedicine and healthcare. *Acad Med* (2009).

[13] Kathryn M Larsen and Charles Kent Smith. 1981. Assessment of nonverbal communication in the patient-physician interview. *J Fam Pract* 12, 3 (1981), 481–488.

[14] Xiujun Li, Yun-Nung Chen, Lihong Li, and Jianfeng Gao. 2017. End-to-End Task-Completion Neural Dialogue Systems. *CoRR* abs/1703.01008 (2017). http://arxiv.org/abs/1703.01008

[15] Steven Y Lin, Tait D Shanafelt, and Steven M Asch. 2018. Reimagining Clinical Documentation With Artificial Intelligence. *Mayo Clinic proceedings* 93, 5 (2018), 563–565. DOI: http://dx.doi.org/10.1016/j.mayocp.2018.02.016

[16] JP Marcin, ME Rimsza, and WB Moskowitz. 2015. The use of telemedicine to address access and physician workforce shortages. *Pediatrics* 136, 1 (2015), 202–209.

[17] Germana Montesi and Alessandro Lechi. 2009. Prevention of medication errors: detection and audit. *British journal of clinical pharmacology* 67, 6 (2009), 651.

[18] Lars Müller, Rashmi Gangadharaiah, Simone C. Klein, James Perry, Greg Bernstein, David Nurkse, Dustin Wailes, Rishi Graham, Robert E. El-Kareh, Sanjay Mehta, Staal A. Vinterbo, and Eliah Aronoff Spencer. An open access medical knowledge base for community driven diagnostic decision support system development. In *BMC Med. Inf. & Decision Making 19(1): 93:1-93:7 (2019)*.

[19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[20] W Price and II Nicholson. 2017. Regulating black-box medicine. *Mich. L. Rev.* 116 (2017), 421.

[21] Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A Deep Reinforcement Learning Chatbot. *CoRR* abs/1709.02349 (2017). http://arxiv.org/abs/1709.02349

[22] Nancy S Sung, William F Crowley Jr, Myron Genel, Patricia Salber, Lewis Sandy, Louis M Sherwood, Stephen B Johnson, Veronica Catanese, Hugh Tilson, Kenneth Getz, and others. 2003. Central challenges facing the national clinical research enterprise. *Jama* 289, 10 (2003), 1278–1287.

[23] Abraham Verghese, Nigam H Shah, and Robert A Harrington. 2018. What this computer needs is a physician: humanism and artificial intelligence. *Jama* 319, 1 (2018), 19–20.

[24] Adam P Vogel and Angela T Morgan. 2009. Factors affecting the quality of sound recording for speech and voice analysis. *International journal of speech-language pathology* 11, 6 (2009), 431–437.

[25] Joseph Weizenbaum. 1967. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman and Company.

[26] Colin P West, Mashele M Huschka, Paul J Novotny, Jeff A Sloan, Joseph C Kolars, Thomas M Habermann, and Tait D Shanafelt. 2006. Association of perceived medical errors with resident distress and empathy: a prospective longitudinal study. *Jama* 296, 9 (2006), 1071–1078.

[27] Dian Yu, Michelle Cohn, Yi Mang Yang, Chun-Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu. 2019. Gunrock: A Social Bot for Complex and Engaging Long Conversations. (2019).

[28] Victor L. Yu, Lawrence M. Fagan, Sharon M. Wraith, William J. Clancey, A. Carlisle Scott, John Hannigan, Robert L. Blum, Bruce G. Buchanan, and Stanley N. Cohen. Antimicrobial Selection by a Computer: A Blinded Evaluation by Infectious Diseases Experts. *JAMA* 242, 12 (????), 1279–1282. DOI:http://dx.doi.org/10.1001/jama.1979.03300120033020

[29] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. (2018).

[30] Victor Zue, Stephanie Seneff, James Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, and Lee Hetherington. 2000. Jupiter: A Telephone-Based Conversational Interface for Weather Information. *IEEE Trans. on Speech and Audio Processing* 8 (2000), 85–96.