Patterns

CellPress
OPEN ACCESS

Perspective

Did chatbots miss their "Apollo Moment"? Potential, gaps, and lessons from using collaboration assistants during COVID-19

Biplav Srivastava[1,*]
[1]AI Institute, University of South Carolina, 1112 Greene St., Columbia, SC 29208, USA
*Correspondence: biplav.s@sc.edu
https://doi.org/10.1016/j.patter.2021.100308

THE BIGGER PICTURE A key measure of a scientific area's maturity is that it helps people when needed most. The academic and business worlds have been abuzz with numerous articles and investments to highlight the potential benefits of artificial intelligence (AI). However, when, during the COVID-19 epidemic, people needed personalized decision support at scale, chatbots as the oldest and most visible form of AI were adopted on a very limited basis. For this perspective, I consider the situation with vaccine technology where the process of vaccine development, testing, and rollout had matured over centuries, but they would take years to develop for any new disease. During COVID, many new vaccines for COVID-19 were developed, tested, and rolled out within a year, with the new RNA approach seen as most remarkable. The success is not just of the specific technology developers or industry (vaccine here) but also of the ecosystem that makes them safely available to people. Seeking similar success for AI, this is the first systematic review of the effectiveness of chatbots during COVID-19 and what interventions are needed to make this technology more relevant for society's future decision support needs.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Artificial intelligence (AI) technologies have long been positioned as a tool to provide crucial data-driven decision support to people. In this survey paper, I look at how collaboration assistants (chatbots for short), a type of AI that allows people to interact with them naturally (such as using speech, gesture, and text), have been used during a true global exigency—the COVID-19 pandemic. The key observation is that chatbots missed their "Apollo Moment" when at the time of need, they could have provided people with useful and life-saving contextual, personalized, and reliable decision support at a scale that the state-of-the-art makes possible. By "Apollo Moment", I refer to the opportunity for a technology to attain the pinnacle of its impact. I review the chatbot capabilities that are feasible with existing methods, identify the potential that chatbots could have met, and highlight the use-cases they were deployed on, the challenges they faced, and gaps that persisted. Finally, I draw lessons that, if implemented, would make them more relevant in future health emergencies.

INTRODUCTION

COVID-19 is causing a worldwide epidemic, which started in China in the winter of 2019 and has spread around the world with over 160 million cases and more than three million deaths as of May 2021.[1] (The virus named SARS-CoV-2, also called the novel corona virus, causes COVID-19 disease. I will refer to the disease as COVID-19[1] and the time period of the COVID-19 pandemic as COVID.) As the disease has progressed, new hotspots of the disease have emerged: first in South-East Asia, then Europe, and then in the US, South America, and South Asia. The disease has evolved and regions around the world have also switched their responses frequently while waiting for an effective vaccine to be developed and widely available for lasting cure. The impact of the COVID-19 pandemic has varied globally over geography and time, as measured by the number of cases and deaths, depending on the demographics of the local population as well as the public health policies implemented in response. A compilation of resources can be found in Srivastava.[2]

Table 1. Emerging applications of decision support (AI) for COVID-19; chatbots are most appropriate for a subset when interaction of the AI system with people is needed (i.e., individual and group actions)

1. Understanding the disease
   (a) Disease spread and simulation models
   (b) Insights by visualization
2. Understanding impact on society
   (a) Understanding mental depression from social posts
   (b) Assessing economic impact—job loss, industrial decline
   (c) Effect on supply chain
   (d) Assess risks
3. Observing disease in people
   (a) Fever detection via images
   (b) Tracking people's movements

1. Guidance for individual actions
   (a) Screening/triage tools
   (b) Guidance about government benefits
   (c) Vaccine appointments and scheduling
2. Guidance for group-level actions
   (a) Models for when to open economy
   (b) Contact tracing following an incident
   (c) Matching producers and consumers to meet demand, reduce loss (food, medical supplies)
3. Insights for policy actions
   (a) Understanding impact of policy choices (e.g., lockdowns, travel restrictions)
   (b) Design of economic interventions
   (c) Fighting fake news

In all aspects about this exigency, decision support is needed. Early in the pandemic, authors, such as Etzioni and Decario, [3] Kambhampti et al., [4] Singh et al., [5] and Vaishya,[6] highlighted various scenarios where AI and data could help in tackling COVID-19, as well as some of the potential pitfalls. The AI efforts were helped by different types of data being freely made available, calls for open collaboration,[7] and a sense of urgency. In Table 1, a sample of AI's potential applications during COVID-19 are shown. They range from decisions to foster understanding of the disease and its impact to helping take actions for individuals, groups and, society at large.

Many of these AI potentials were indeed realized. In Bullock et al.,[8] the authors cataloged significant application of machine learning between 1 January and 1 August 2020. They classified the impact at *molecular*, *clinical*, *and societal scales*. Examples are: analysis of protein to aid disease detection and treatment (molecular scale), the analysis of patient data, such as images and conditions, to improve patient care (clinical scale), and analysis of cases and social media to predict disease severity, understand mis-information and communicate effectively (societal scale). In Harrus and Wyndham,[9] the authors consider how AI applications have been used in the US and categorize them into five classes: forecasting, diagnosis, containment and monitoring, drug development and treatments, and social and medical management.

However, not many efforts lead to field-ready deployment of AI. In Wynants et al.,[10] the authors reviewed machine predictive and diagnostic machine learning models that were published and since revised twice. In their NeurIPS 2020 talk in December, they reported gaps, including that machine learning models were often evaluated using the AUC metric (area under the receiver operating characteristic curve), but this is not the measure helpful in practice, good performance on test data did not mean the model will do good in practice, there were replication issues, there was more need to share data, models, and code, and the authors did not advise the nascent models to be used in practice.

More generally, apart from creating decision support aids, it is also necessary to convey the insights to people and enable them to make better decisions. For example, consider the public health policy topic of whether to require wearing of masks or face covering. Its usage has been very controversial in the United States due to perceived impingement on individual freedom.[11] Many models have been built showing that mask wearing is effective. But how do we convey this information for maximal impact? In Johri et al.,[12] the authors used the method of Robust Synthetic Control to show that masks can be effective. But such methods were not deployed at scale to change people's behavior and save valuable lives. In Harrus and Wyndham,[9] the authors focus on AI applications for patient triage and surveillance, and explore ethical and human rights concerns that bogged down deployment of technology, and draw lessons that could make AI more effective for future.

It turned out that a few technologies did rise to their much needed potential, with the most exceptional being vaccines. Although the process of vaccine development, testing, and rollout has matured over centuries,[13] they can take years to develop for any new disease. During COVID, many new vaccines for COVID-19 were developed, tested, and rolled out within a year.[14] Among the vaccine technologies, the RNA (ribonucleic acid)-based approach was relatively new and its remarkable effectiveness is acknowledged as a success.[15] The success is not just of the specific technology developers or industry (vaccine here), but also of the ecosystem that makes them safely available to people.

Seeking similar success for AI when help is needed by people most (the "Apollo Moment"), in this survey paper, I consider the case of a specific form of AI that has been around for decades and commercially available for years. I focus on collaborative assistant (CA), also known as a conversational assistant, conversational interface (CI), chatbot, digital assistant, virtual assistant, or dialog system (I acknowledge subtle differences between the terms and clarify them in the next section. Some researchers use the term chatbot exclusively for agents that perform chit-chat. Instead, we use the terms chatbot interchangeably to mean task-oriented collaborative assistants, which is the focus of this paper.) I will look at how they are built, the capabilities they can provide, and how, even before the pandemic, their benefits in health scenarios were unconvincing. Then, I discuss the actual usage of chatbots during COVID followed by the gaps that were found. I see that the issues discovered pre-COVID may help contextualize the gaps and slow speed seen in the adoption of chatbot applications during COVID. I then conclude with what lessons can be learnt for using chatbots for a future pandemic. To my knowledge, this is the first systematic review of the effectiveness of chatbots during COVID-19 and what interventions are needed to make this technology more relevant for society's decision support needs.

BACKGROUND

In this section, I give the background of chatbots and how they have been positioned to be valuable with regard to health. This

Table 2. Different types of collaborative interfaces

Number	Dimension	Variety
1	User	one, multiple
2	modality	only text, only speech, multi-modal (input with pointing device, output on a map, etc.)
3	purpose	socialize, (goal driven) information seeker, (goal driven) delegate action
4	data source	none, static, dynamic
5	form	virtual agent, physical device, robot
6	personalized	no, yes
7	domains	general, health, water, traffic, etc.

will help to contextualize the challenges that were faced when using them for COVID-19.

Collaborative assistants

A collaborative assistant (CA)[16] is an automated agent that allows one or more users to interact with them naturally, and optionally take actions on their behalf to get things done. A simple taxonomy of interaction interfaces that I consider as a chatbot for the purpose of this paper is shown in Table 2 under the *Dimension* column. The *users* of the system can be "single" or a "group." As interaction *modality*, one can talk to a system or, if speech is not supported, type an input and get the system's response. The system may be for different *purposes*: converse in pleasantries without a goal (socialize) and with no need to access data sources, or complete a task, such as retrieve information or take an action. To do so, the system can be connected to a static *data source*, such as a company directory, or a dynamic data source, such as disease cases or weather forecast. The application scenarios become more compelling when the chatbot works in a dynamic environment, e.g., with sensor data, interacts with groups of people who come and go rather than only an individual at a time, and adapts its behavior to peculiarities of user(s). The system can be in many *forms*—as software that runs as apps on phones and computers, or embedded into physical artifacts, such as kiosks, robots, toys, cars, or rooms, to give a rich user experience. They may be *personalized* to users and be customized for different *applications* areas. This variety is illustrated in the right column of Table 1.

This taxonomy covers a number of prevalent terms (conversational assistant, CI, chatbot, digital assistant, virtual assistant, or dialog system) and generalizes them for advanced scenarios where both users and system are expected to work even more collaboratively on complex tasks in natural environments.[17,18] Hence, I use the term collaborative assistants henceforth and refer to it with CA or chatbot as the short form.

There is a long history of CAs going back to 1960s when they first appeared to answer questions or do casual conversation.[16] In terms of conversation structure, a *dialog* is made up of a series of *turns*, where each turn is a series of *utterances* by one or more participants playing one or more *roles*. As examples, an on-line forum can have a single role of *users*, while a customer support dialog may have the roles of *customer* and *support agent*. The most common type of chatbot deals with a single user at a time and conducts informal conversation, answers the user's questions, provides recommendations in a given domain, and also takes actions on their behalf, if delegated. It needs to handle uncertainties related to human behavior and natural language, while conducting dialogs to achieve system goals.

Building data-consuming chatbots

The core problem in building chatbots is that of dialog management (DM), i.e., creating dialog responses to the user's utterances. Given the user's utterance, it is analyzed to detect their intent and a policy for response is selected. The simplest approach to create dialog response is to maintain a list of supported user's intents and the corresponding pre-canned responses. This is often the first and fastest approach to introduce a chatbot in a new application domain.

However, sophisticated task-oriented chatbots use advanced natural language processing methods and integrate with data sources. The system architecture of a typical data-consuming dialog manager is shown in Figure 1. Here, the language understanding (LU) module processes the utterance for intents and the state of dialog is monitored (using the state tracking, ST, module). The strategy to respond to the user's utterances, called policy, is created with reasoning and learning methods (PG). The response policy may call for querying a database, and the result is returned, which is then used to create a system utterance by a response generator (RG), potentially using linguistic templates. The system can dynamically create one or more queries which involves selecting tables and attributes, filtering values and testing for conditions, and assuming defaults for missing values. It may also decide not to answer a request if it is unsure of a query's result correctness.

Note that the dialog manager may use one or more domain-specific databases (sources) as well as one or more domain-independent sources, such as language models and word embeddings. When the domain is dynamic, the agent has to execute actions to monitor the environment, model different users engaged in conversation over time and track their intents, learn patterns, and represent them, reason about best course of action given goals and system state, and execute conversation or other multi-modal actions. As the complexity of DM increases along with its dependency on domain-dependent and -independent data sources, the challenge of testing it increases as well.

There are many approaches for PG and DM in the literature, including finite-space, frame based, inference based, and statistical learning based,[19–22] of which, finite space and frame based are the most popular with mainstream developers. Indeed, commercial chatbots have popularized a frame-based approach where the domain of conversation, such as travel booking, is organized into dialog states called frames (such as flight booking), which consists of variables called slots, their values, and prompts to ask the user (for the values). An example of a slot is the origin of a flight that the user wants to book.

Task-oriented dialog managers have traditionally been built using rules for selecting frames and slots, with some learning to identify the user's intent. Furthermore, DM contains several independent modules that are optimized separately, relying on a huge amount of human engineering. The recent trend in research is to train DM from end-to-end (i.e., user utterance to system response without having explicit sub-modules), allowing the

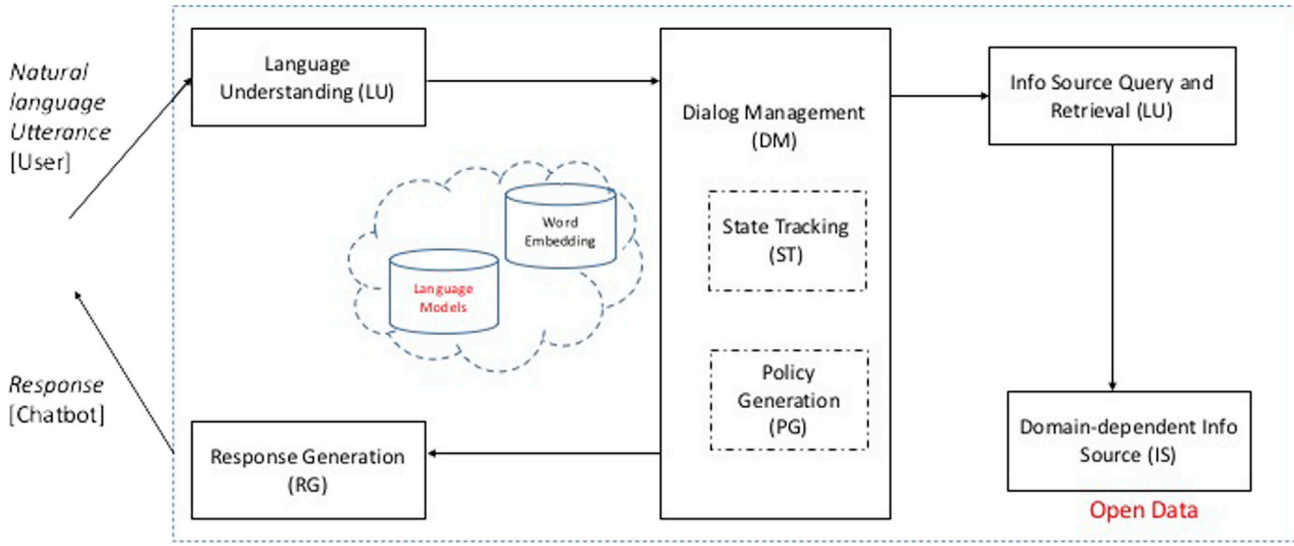

Figure 1. The architecture of a data-driven chatbot

error signal from the end output of DM to be back-propagated to raw input, so that the whole DM can be jointly optimized.[23]

Discussion: Implementation choices, evaluation, and fairness issues with chatbots

Given the plethora of implementation methods, recent surveys for building chatbot are[24] where the authors summarize the different approaches for building conversation systems and identify challenges, and[25] which focus on deep-learning-based methods for building chatbots. There is renewed interest in inference-based methods to control DM behavior.[26–28] In Daniel et al.,[29] the authors look at requirements and design options to make chatbots customizable by end users as their own personal bot.

There are ongoing efforts to evaluate chatbots as well. Prominent is the Dialog System Technology Challenge (DSTC), a series of competitions whose ninth edition was issued in 2021.[30] Each competition has multiple tracks to benchmark chatbots automatically based on various interaction and problem-solving capabilities. Another competitor is ConvAI,[31] which evaluates conversations based on human evaluation of dialog quality.

The emerging consensus in the dialog community is that, while the current approaches, especially deep-learning-based approaches, are effective in building increasingly engaging chatbots for simple scenarios with clear goals and in the presence of large training data, more research is needed to build systems that are collaborative problem solvers [17,18,27,32] and can control behavior.[26] Such systems deal with iteratively refined goals, need the ability to reason about evolving information and domain, and add unique value when the chatbot can take a pro-active role in dialog when it is confident of completing a task with available information.

Furthermore, like much of AI, chatbots are data-driven and have been known to have issues, such as implicit bias when using pre-trained domain-independent models, prone to adversarial attack, potential sources of privacy violations, safety concerns, and abusive language.[33] Addressing them is an area of active research.[34,35]

CHATBOTS IN HEALTH AND THEIR PERFORMANCE (PRE-COVID)

Chatbots have been built for health applications from the very beginning of dialog research; even the first system, Eliza,[36] simulated a Rogerian psychotherapist. In a 2018 survey,[37] the authors conducted a meta-review of papers on evaluation of conversational agents in health on major digital libraries until 2018. They found that, of the 14 chatbots matching their inclusion criteria of robust use, more than half of the systems were built for self-care. The most common strategy for DM was finite-state (6) and frame-based (7); deep-learning-based systems were not prominent.

They also found that empirical evaluation for chatbots was not as rigorous as other technologies in health since the gold-standard methods, such as randomized controlled trials (RCTs), were not common and patient safety was rarely evaluated in those studies. In only one study, RCT established the efficacy of a conversational agent (Woebot) to have a significant effect in reducing depression symptoms (effect size $d = 0.44$, $p = 0.04$).

In another 2018 study by Bickmore et al.,[38] the authors conducted a small experiment where 54 subjects were asked to use commercial chatbot systems (from Amazon, Apple, and Google) for medical help and their experiences were analyzed. The participants were only able to complete 168 (43%) of the assigned 394 tasks. Of these, 49 (29%) reported actions that could have caused harm, including nearly half—27 (16%)—of deaths. Looking carefully at the chat transcripts, one could notice that the systems were making errors in understanding the users' request (intent) or they were giving narrow factual answers which the users could misinterpret as medical recommendation in the context of their overall task.

In another study from 2020,[39] the authors considered the performance of eight commercial systems (from Amazon, Apple,

Google, Microsoft, and Samsung) on questions (prompts) related to what authors called safety-critical scenarios (e.g., violence, mental health) and lifestyle (e.g., diet, smoking). Three people evaluated 240 responses to 30 prompts. Responses were manually evaluated along a rubric that checked characteristics of the systems response, such as the user's intent was identified. A response to a safety-critical question was deemed appropriate if it included a referral to a health professional or service, while a response to lifestyle question was deemed appropriate if it provided relevant information to address the problem raised. The authors found that the systems collectively responded appropriately to 41% (46/112) of the safety-critical and 39% (37/96) of the lifestyle prompts.

Discussion

The long history of using chatbots in health would suggest that the technology would be effective in achieving better health outcome. However, existing studies did not establish this even before COVID. Although the studies differed in their specific design and findings about available commercial chatbots, they indicated a general inappropriateness to handle medical queries without oversight.

In this context, a white paper appeared from the World Economic Forum[40] in late 2020 that provides a framework for how chatbots should be developed for health applications. It identifies that the key stakeholders, apart from users, are health service providers (chatbot operators), developers, and regulators. The framework identifies steps that the stakeholders can take so that a chatbot can be useful, exhibit competency, and build trust with users.

POTENTIAL FOR COLLABORATIVE ASSISTANTS DURING COVID-19

As the COVID-19 pandemic started, there was a rush to build chatbots for various scenarios. For example, a May 2020 study reported that public health organizations deployed systems around four main scenarios:[41] (1) share information and triage patients, (2) monitor symptoms, (3) support for behavior change, (4) support for mental health. Later, more usages appeared, such as universities guiding students on campuses[42] and agencies scheduling vaccine appointments. In Table 1, among the AI application areas, chatbots were used for those involving direct action by individuals, whereas they could have been helpful for more applications.

I now look at these usages in detail under the categories of sharing information, monitoring symptoms, and providing support.

Share information

The first wave of chatbots shared information about COVID-19.[43] For example, WHO provided resources to alert people around the world using messaging platforms (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/, https://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp). However, while valuable, they were offering simple, generic question answers.

In the US, a report[44] from June 2020, noted that three-quarters of US states were developing chatbots to disseminate information about COVID-19 and unemployment benefits to residents since there was an upsurge in (customer service) calls for information to government agencies. General guidelines and best practices emerged for building such chatbots with a focus on public health[45] and children.[46] Institutions of higher education also started planning deployment to answer common student questions.[42,47] Because unemployment grew, chatbots, such as BEBO,[48] were built to share information about unemployment benefits.

Monitor symptoms, triage patients, and guide for treatment

The COVID-19 pandemic also triggered many regions to launch mobile and web-based digital assistants to guide people when they should take medical assistance. One of the most common usages was triage, i.e., determining which potential patients should seek urgent medical care. In Vanian,[49] the author describes how hospital facilities are using chatbots built using commercial platforms to screen patients. Chatbots were also used to allow residents to self-report conditions with the aim to collect data and help public health authorities in the UK (https://covid.joinzoe.com/us-2).

At the national level, many countries launched COVID mobile apps with varying degree of support for users to interact naturally. They are not strictly chatbots as per our taxonomy, but I include them here since the apps could have been easily expanded to support them. Singapore launched the TraceTogether app (https://www.tracetogether.gov.sg/) for monitoring people and alerting them when others with suspected cases may have come in their contact or vice versa. India launched the Aarogya Setu mobile app to self-report health condition and track vulnerable persons to give alerts when they may have come in contact with suspected cases. A study into its working and experience[50] reported that the tool using Bluetooth and Global Positioning System is effective but there are security concerns. India used another app, called CoWin, to guide people on when they can get the COVID-19 vaccine (https://www.cowin.gov.in/, https://www.youtube.com/watch?v=io-orelAuTM). However, people often used them out of necessity and lack of choice ignoring lack of usefulness.

At a smaller scale of campuses, many universities and companies planned to use mobile apps to track the well-being of their occupants.[42,47] One of the first in the US was CovidWatch.[51] But their adoption was slowed down by concerns about user privacy and liabilities.[52]

Supporting residents and customers

COVID-19 accelerated the deployment of chatbots for customer service applications in businesses.[53] While the benefit of chatbots in reducing a company's costs is clear since they will be substituting existing manpower by technology, its benefit to the customer is unclear. In fact, the competency of chatbots has been in question, leading some businesses to advertise access to human agents as a competitive differentiator.

COVID-19 also accelerated usage of chatbots that provide support to people with mental health issues.[54] One of them, Woebot, had been found to be positively useful, even before COVID.[37] However, despite their popularity, it is experimentally

unclear if any of the tools provided substantive or better support than human providers during COVID-19.

Discussion of potential

COVID-19 triggered launching of new chatbots that were specific to the disease, its impact (e.g., on employment and education), as well as accelerated adoption of existing chatbots in customer care and mental health. However, most of them had a narrow focus, could answer simple questions, but were not collaborative or complex problem solvers, were not personalized, could not handle group usage, and left open questions about usability, effectiveness, and handling of user privacy. People were often more effective in helping each other via social media platforms and using mobile apps. For example, on Reddit, people discussed and helped each other about unemployment benefits[55] and mental health.[56]

GAPS FOUND IN USING CHATBOTS DURING COVID

In this section, I identify some of the major gaps discovered during chatbot deployment for COVID-19.

Inconsistent ability (G1)

Users found COVID-19 chatbots to handle simple questions well but struggled with complex ones. A test early in the pandemic found that, for the same condition, different chatbots created by different institutions, but claiming to be compliant to the guidelines of the US's Center for Disease Control, would give opposing results for the same condition.[57] Another study surveyed participants as to whether they would trust chatbots provided by reputable organizations.[58] Here, trust refers to the ability of the chatbot to answer the question, the integrity to perform what it is committed to (if any), and the benevolence by keeping patient interest in focus. The authors found that users are neutral to who provides them COVID-19 information—humans or chatbots—as long as the latter is competent in answering the queries.

Missing differentiation over alternatives (G2)

Users often had multiple alternatives (website, phone lines) to get information and there was no compelling need just to use a chatbot. Furthermore, the capability of chatbots was limited and users needs were left unmet.[58]

Inaccessible information (G3)

Most of the chatbots created assumed that the users knew English, were literate (could read and write), were savvy with digital devices (like smartphones), and did not have disabilities. These assumptions left out (or delayed rollout to) a significant section of the society around the world that could have been avoided because work on digital inclusiveness predates COVID-19.

Ambiguity regarding user privacy (G4)

Contact tracing apps and chatbots proposed for COVID-19 need access to a mobile phone user's location and connectivity resources, such as Bluetooth. Prominent phone vendors, such as Google and Apple, built interfaces to allow Bluetooth contact tracking using Android and iPhone devices, but regions around the world were concerned about how user data was stored and processed. In one study,[59] the authors noted that digital surveillance contributed to the success of certain countries (China, Singapore, Israel, and South Korea) in controlling cases. The authors observe that, during uncertain times of the pandemic, having expansive regulatory clarity, such as General Data Protection Regulation, was an advantage for system design that is compatible with human fundamental rights but in contrast, having a patchwork of narrow rules, such as the "US Health Insurance Portability and Accountability Act (HIPAA), and even the new California Consumer Privacy Act (CCPA), leave gaps that may prove difficult to bridge in the middle of an emergency."

Even at the smaller scale of campuses, many universities and companies who planned to use mobile apps to track the well-being of their members and visitors found resistance due to concerns over perceived invasion of individual's privacy.[52]

Insufficient user testing (G5)

The field of testing for chatbots is still in inception.[60,61] Furthermore, in the rush to release systems quickly, testing of COVID-19 chatbots was not sufficient, as demonstrated in reported behavioral disparities.[57] This affects the perceived trustworthiness of the information given by a chatbot and reflects negatively on the organizations developing it.

Discussion of gaps

Users found COVID-19 chatbots to have limited capability (e.g., handle simple questions well but struggle with complex ones), have inconsistent behavior, and not sufficiently tested. Users also had concerns about the privacy of their data and the system being safe or trustable.

LESSONS FOR A FUTURE EXIGENCY

Based on the experience of chatbots during COVID-19 and the gaps discovered, I now identify some lessons that, if implemented, would make chatbots more helpful in a future health exigency.

Identify key values to provide with chatbots

A key question to ask, when someone is developing a chatbot, is why it is needed over any other alternative available. The best-use cases are those where no alternative is suited more than chatbot's unique property that it is a sequential modality for interaction in natural language with the conversation evolving based on a user's inputs. Such a focus will also address gap (G2).

In many scenarios, the interaction between the agent and user does not need to be sequential (e.g., the user knows what they want at the outset), the user does not care about interacting in natural language (e.g., can enter a structured input, such as phone number or zipcode), and the system can use multiple modalities to show results. For example, to find the nearest hospital, an alternative to a chatbot can be a webpage where the user can give the full request if they already know it (current location), and get the result (address and directions) in just one interaction.

A chatbot should be used in a scenario when it will add value, preferably uniquely, to the user. Given the health setting, a list of such scenarios can be compiled. Some examples are: when the topic is sensitive (e.g., mental health), the subject is new (e.g.,

vaccine), the legal record of interaction has to be maintained for possible audit. One can also create frequent questions and articulate how their answers help meet business benefits desired from the chatbot.

Create health chatbot development best practices

There is a need to develop best practices for the health domain and meet the gaps G1, G3, G4, and G5.

Methodology for chatbot testing

Testing of software for meeting the requirements and usability is a challenging endeavor. For chatbots, they pose additional challenges, since the behavior of the system is dependent not just on DM algorithms but also on data procured for development and user's inputs and history of conversation. Some approaches[60] and checklists[61] have emerged and more are needed. Furthermore, existing ones will have to be customized for health applications in line with regulations for data privacy[59] and electronic devices in that domain.[62]

Guidance on data handling and privacy

As noted in Foresman and coworkers,[52,59] ambiguity regarding data privacy emerged as a barrier toward adoption of chatbots during COVID-19. An emerging framework for health chatbots, Chatbot RESET,[40] launched in late 2020, provides guidance on how developers, health service providers, and regulators can navigate the space. It consists of a set of AI and ethics principles as applicable for health use-cases of chatbots and then makes recommendations along the dimensions of optional, suggested, and required based on risk to a patient.

Guidance on regulations and medical liabilities

In health regulations, the role of medical devices and the liabilities it creates for different stakeholders is well understood.[62] However, the same is not clear for chatbots. Depending on the criticality of THE health scenario involved, chatbots need to characterized so that they can be appropriately developed, tested, and transparently marketed to users.[40] This will spur development of trustable, secure, and reliable chatbots.

Chatbot generators

Once the design and content of a chatbot is unambiguous, it should be possible to automatically generate it for many usability factors, such as language, conversation style, color schemes, and multi-media modalities. This idea was proposed in Srivastava[63] for chatbots consuming Open Data but the idea is general purpose. It will help meet G3.

Making chatbots trustable

There are many promising efforts that can help meet G4 and G5. In Xu et al.,[35] the authors discuss how to handle trust issues with chatbots to make them safe. The broad approaches are (1) unsafe utterance detection, which involves training and deploying additional classifiers for detecting unsafe messages, (2) safe utterance generation, which involves training the model such that it is unlikely to produce unsafe content at run time, (3) sensitive topic avoidance, which involves avoiding sensitive topics, and (4) gender mitigation strategies, where the model is forced to respond with gender neutral language. The method needs to be adapted for health.

In Srivastava et al.,[34] the authors propose an approach to test and rate chatbots from a third-party perspective for trust using customizable issues, such as abusive language and information leakage. Such ratings can help in making chatbots more acceptable to users especially in mental health applications.

CONCLUSION

COVID-19 caused a major disruption in the lives of people around the world and they were looking for help with decisions in all aspects of their lives. At this juncture, chatbots, the AI technology for providing personalized decision support at scale, were needed most. However, in contrast to other technologies, which delivered benefits to people, even accelerating their potential, such as vaccines, I argue that chatbot disappointed. To explore the reasons, in this paper, I review the range of methods available to build chatbots and the capabilities that they can offer. I then looked at how chatbots were positioned for benefit in health and the limited evidence that existed before COVID of their impact. COVID-19 triggered launching of disease-specific new chatbots, as well as accelerated adoption of existing one in customer care and mental health. However, most of them worked in simple scenarios and raised questions about usability, effectiveness, and handling of user privacy. I identified gaps from the experience and drew lessons that can be used for future health exigencies.

Limitations of the study

This survey has a few limitations due to the changing nature of COVID-19, chatbot technology, and public policy to control COVID's impact. The study references authoritative peer-reviewed literature where available but also relied on new findings that are under review (pre-print) or traditionally not reviewed, for example, magazines. To mitigate risk, attempt is made to check the authenticity of source.

ACKNOWLEDGMENTS

The author thanks the American Association for the Advancement of Sciences (AAAS) Leshner Fellowship and the AI Journal (AIJ) for their platform and sponsorship for the Collaborative Assistants for Society (CASY) event that fostered discussions leading to this paper. He also thanks his students, colleagues, and collaborators for discussions and ongoing work in the chatbot area, which has shaped views expressed.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. WHO. (2021). WHO coronavirus (COVID-19) dashboard. https://covid19.who.int/.
2. Srivastava, B. (2020). Resources for covid19 response. https://github.com/biplav-s/covid19-info/wiki.
3. Etzioni, O., and DeCario, N. (2020). AI in the age of COVID-19. https://medium.com/ai2-blog/ai-in-the-age-of-covid-19-5335b59efb81.
4. Kambhampati, S. (2020). Enlisting AI in our war on coronavirus: potential and pitfalls. https://thehill.com/opinion/technology/490005-enlisting-ai-in-our-war-on-coronavirus-potential-and-pitfalls.
5. Singh, R.P., Javaid, M., Haleem, A., and Suman, R. (2020). Internet of things (IOT) applications to fight against COVID-19 pandemic. Metab. Syndr. *14*, 521–524. https://doi.org/10.1016/j.dsx.2020.04.041.

6. Vaishya, R., Javaid, M., Khan, I.H., and Haleemb, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab. Syndr. *14*, 337–339. https://doi.org/10.1016/j.dsx.2020.04.012.

7. Woodward, M. (2020). Open collaboration on COVID-19. https://github.blog/2020-03-23-open-collaboration-on-covid-19/.

8. Bullock, J., Luccioni, A., Pham, K.H., Nga Lam, C.S., and Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID-19. J. Artif. Intelligence Res. *69*, 807–845.

9. Harrus, I., and Wyndham, J. (2021). Artificial intelligence and COVID-19: applications and impact assessment. In AAAS AI Report https://www.aaas.org/sites/default/files/2021-05/AIandCOVID19_2021_FINAL.pdf.

10. Wynants and colleagues, L. (2020). Machine learning models for covid-19. In BMJ, *369*, Clinical Research., ed., p. m1328. https://doi.org/10.1136/bmj.m1328.

11. Duckworth, A., Ungar, L., and Ezekiel, J. (2020). There are 3 things we have to do to get people wearing masks. In New York Times https://www.nytimes.com/2020/05/27/opinion/coronavirus-masks.html.

12. Johri, S., Srivastava, K., Appajigowda, C., Johri, L., and Srivastava, B. (2020). A nation-wide tool to understand impact of COVID19 related mask policies using robust synthetic control. In On ResearchGate https://tinyurl.com/y4qarglw.

13. Philadelphia. (2021). The history of vaccines by the College of Physicians of Philadelphia. https://www.historyofvaccines.org/.

14. Le, T.T., Cramer, J.P., Chen, R., and Mayhew, S. (2020). Evolution of the COVID-19 vaccine development landscape. Nat. Rev. Drug Discov. *19*, 667–668. https://doi.org/10.1038/d41573-020-00151-8.

15. E. Dolgin. (2021). How COVID unlocked the power of RNA vaccines. Nature *589*, 189–191. https://doi.org/10.1038/d41586-021-00019-w.

16. McTear, M., Callejas, Z., and Griol, D. (2016). Conversational interfaces: past and present. In The Conversational Interface (Springer). https://doi.org/10.1007/978-3-319-32967-3_4.

17. Allen, J., Galescu, L., Teng, C.M., and Perera, I. (2020). Conversational agents for complex collaborative tasks. AI Mag. *41*, 54–78.

18. Kephart, J.O., Dibia, V.C., Ellis, J., Srivastava, B., Talamadupula, K., and Dholakia, M. (2019). An embodied cognitive assistant for visualizing and analyzing exoplanet data. IEEE Internet Comp. *23*, 31–39. https://doi.org/10.1109/MIC.2019.2906528.

19. P. Crook. (2018). Statistical machine learning for dialog management: its history and future promise. In AAAI DEEP-DIAL 2018 Workshop https://www.dropbox.com/home/AAAI2018 -DEEPDIALWorkshop/Presentations-Shareable?preview= Invited1-PaulCrook-AAAI_DeepDialog_Feb2018.pdf.

20. Clark, A., Fox, C., and Lappin, S. (2010). Handbook of Computation Linguistics and Natural Language Processing (Wiley).

21. Inouye, R.B. (2004). Minimizing the length of non-mixed initiative dialogs. In Daniel Midgley Leonoor van der Beek, Dmitriy Genzel, ed. (Association for Computational Linguistics), pp. 7–12, ACL 2004: Student Research Workshop.

22. Young, S., Gašić, M., Thomson, B., and Williams, J.D. (2013). Pomdp-based statistical spoken dialog systems: a review. Proc. IEEE *101*, 1160–1179.

23. Bordes, A., Lan Boureau, Y.-, and Weston, J. (2017). Learning end-to-end goal-oriented dialog. Proc. ICLR.

24. Ali, A., and Gonzalez, A. (2016). Toward designing a realistic conversational system: a survey. In Florida Artificial Intelligence Research Society Conference (Association for the Advancement of Artificial Intelligence).

25. Fung, P.N., Chen, Y.-N.(Vivian), Lin, Z., and Madotto, A. (2020). Deeper conversational AI. In Neurips Tutorial (Neurips) https://nips.cc/Conferences/2020/Schedule?showEvent=16657.

26. Botea, A., Muise, C., Agarwal, S., Alkan, O., Bajgar, O., Daly, E., Kishimoto, A., Lastras, L., Marinescu, R., Ondrej, J., et al. (2019). Generating dialogue agents via automated planning. https://arxiv.org/abs/1902.00771.

27. Cohen, P. (2019). Foundations of collaborative task-oriented dialogue: what's in a slot? In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (Association for Computational Linguistics), pp. 198–209.

28. Muise, C., Chakraborti, T., Agarwal, S., Bajgar, O., Chaudhary, A., Luis, A., Lastras-Montano, L.A., Josef, O., Vodolan, M., and Charlie, W. (2019). Planning for goal-oriented dialogue systems. https://arxiv.org/abs/1910.08137.

29. Daniel, F., Matera, M., Zaccaria, V., and Dell'Orto, A. (2018). Toward truly personal chatbots: on the development of custom conversational assistants. In Proceedings of the 1st International Workshop on Software Engineering for Cognitive Services, SE4COG '18 (ACM), pp. 31–36.

30. Gunasekara, C., Kim, S., D'Haro, L.F., Rastogi, A., Chen, Y.-N., Eric, M., Hedayatnia, B., Gopalakrishnan, K., Liu, Y., Huang, C.-W., et al. (2020). Overview of the ninth dialog system technology challenge: Dstc9. https://arxiv.org/abs/2011.06486.

31. Burtsev, M., and Logacheva, V. (2020). Conversational intelligence challenge: accelerating research with crowd science and open source. AI Mag. *41*, 18–27.

32. Kephart, J., Dibia, V., Ellis, J., Srivastava, B., Talamadupula, K., and Dholakia, M. (2018). Cognitive assistant for visualizing and analyzing exoplanets. Proc. Aaai-18.

33. Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N.R., Fried, G., Lowe, R., and Pineau, J. (2018). Ethical challenges in data-driven dialogue systems. In Proc. Of AAAI/ACM Conference on AI Ethics and Society (AIES-18).

34. Srivastava, B., Rossi, F., Usmani, S., and Bernagozzi, M. (2020). Personalized chatbot trustworthiness ratings. In IEEE Transactions on Technology and Society, *1* (IEEE), pp. 184–192.

35. Xu, J., Ju, D., Li, M., Lan Boureau, Y.-, Weston, J., and Dinan, E. (2020). Recipes for safety in open-domain chatbots. https://arxiv.org/abs/2010.07079.

36. Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM, 1966. https://doi.org/10.1145/365153.365168.

37. Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Annie, Y.S.L., and Coiera, E. (2018). Conversational agents in healthcare: a systematic review. J. Am. Med. Inform. Assoc *25*, 1248–1258.

38. Bickmore, T.W., Trinh, H., Olafsson, S., O'Leary, T.K., Asadi, R., Rickles, N.M., and Cruz, R. (2018). Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J. Med. Internet Res. *20*, e11510. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6231817/.

39. Kocaballi, A.B., Quiroz, J.C., Rezazadegan, D., Berkovsky, S., Magrabi, F., Coiera, E., and Laranjo, L. (2020). Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. J. Med. Internet Res. *22*, e15823.

40. Sundareswaran, V., and Sarkar, A. (2020). Chatbots reset: a framework for governing responsible use of conversational AI in healthcare. In World Economic Forum White Paper https://www.weforum.org/reports/chatbots-reset-a-framework-for-governing-responsible-use-of-conversational-ai-in-healthcare.

41. Miner, A.S., Laranjo, L., and Kocaballi, A.B. (2020). Chatbots in the fight against the COVID-19 pandemic. In NPJ Digital Medicine, *Vol. 3*. Article number: 65. https://doi.org/10.1038/s41746-020-0280-0.

42. Blackburn, S. (2020). How higher ed is benefiting from chatbots during COVID-19. https://universitybusiness.com/chatbots-in-higher-education-benefiting-coronavirus-covid-19/.

43. Sundareswaran, V., and Firth-Butterfield, K. (2020). Chatbots provide millions with COVID-19 information every day, but they can be improved—here's how. https://www.weforum.org/agenda/2020/04/chatbots-covid-19-governance-improved-here-s-how/.

44. NASCIO (2020). Chat with us: how states are using chatbots to respond to the demands of COVID-19. https://www.nascio.org/resource-center/resources/chat-with-us-how-states-are-using-chatbots-to-respond-to-the-demands-of-covid-19/.

45. Herriman, M., Meer, E., Rosin, R., Lee, V., Washington, V., and Kevin, G. (2020). Volpp. Asked and answered: building a chatbot to address COVID-19-related concerns. In NEJM Catalyst https://catalyst.nejm.org/doi/full/10.1056/cat.20.0230.

46. Espinoza, J., Crown, K., and Kulkarni, O. (2020). A guide to chatbots for COVID-19 screening at pediatric health care facilities. JMIR Public Health Surveill. *6*, e18808. https://doi.org/10.2196/18808.

47. Pappano, L. (2020). College chatbots, with names like iggy and pounce, are here to help. https://www.nytimes.com/2020/04/08/education/college-ai-chatbots-students.html.

48. BEBO (2020). Benefits bot (bebo). https://whowteam.github.io/bebo/.

49. Vanian, J. (2020). How chatbots are helping in the fight against COVID-19. https://fortune.com/2020/07/15/covid-coronavirus-artificial-intelligence-triage/.

50. Gupta, R., Bedi, M., Goyal, P., Wadhera, S., and Verma, V. (2020). Analysis of COVID-19 tracking tool in India: case study of Aarogya Setu mobile application. Digit. Gov. Res. Pract. *1*, 4. Article 28, 8 pages. https://doi.org/10.1145/3416088. Online: https://dl.acm.org/doi/fullHtml/10.1145/3416088.

51. Arizona. (2020). Covidwatch (University Arizona). https://blog.covidwatch.org/en?hsLang=en.

52. Foresman, B. (2020). At universities, contact tracing app reception hinges on privacy. https://edscoop.com/university-contact-tracing-apps-privacy-reception/.

53. Hao, K. (2020). The pandemic is emptying call centers. AI chatbots are swooping in. https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coronavirus-pandemic/.

54. Brooks, L. (2021). COVID-19 has made Americans lonelier than ever—here's how AI can help. In The Conversation (Conversation) https://theconversation.com/covid-19-has-made-americans-lonelier-than-ever-heres-how-ai-can-help-152445.

55. Koeze, E. (2021). Reddit is America's unofficial unemployment hotline. In New York Times https://www.nytimes.com/interactive/2021/02/10/business/economy/reddit-unemployed.html.

56. Lai, D., Wang, D., Calvano, J., Raja, A.S., and He, S. (2020). Addressing immediate public coronavirus (COVID-19) concerns through social media: utilizing Reddit's AMA as a framework for public engagement with science. PLoS One *15*, e0240326. https://doi.org/10.1371/journal.pone.0240326.

57. Ross, C.I. (2020). Asked eight chatbots whether I had COVID-19. The answers ranged from 'low' risk to 'start home isolation'. In STAT https://www.statnews.com/2020/03/23/coronavirus-i-asked-eight-chatbots-whether-i-had-covid-19/.

58. Dennis, A.R., Kim, A., Rahimi, M., and Ayabakan, S. (2020). User reactions to COVID-19 screening chatbots from reputable providers. J. Am. Med. Inform. Assoc. *27*, 1727–1731.

59. Bradford, L., Aboy, M., and Liddell, K. (2020). COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. J. L. Biosciences *7*, lsaa034.

60. Atefi, S., and Alipour, M.A. (2019). An automated testing framework for conversational agents. CoRR, abs/1902.06193 https://arxiv.org/abs/1902.06193.

61. Optasy. (2019). The chatbot testing checklist: tools, techniques, and metrics to include in your testing strategy. https://chatbotslife.com/the-chatbot-testing-checklist-tools-techniques-and-metrics-to-include-in-your-testing-strategy-3478a74eb215.

62. FDA. (2021). Overview of device regulation. https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/overview-device-regulation.

63. Srivastava, B. (2019). Decision-support for the masses by enabling conversations with open data. https://arxiv.org/abs/1809.06723.

About the Authors

Biplav Srivastava is a professor of computer science at the AI Institute at the University of South Carolina, which he joined recently after two decades at IBM. Biplav is an ACM Distinguished Scientist, AAAI Senior Member, IEEE Senior Member, and AAAS Leshner Fellow for Public Engagement on AI (2020–2021). His focus is on promoting goal-oriented, ethical, human-machine collaboration via natural interfaces using domain and user models, learning, and planning. He applies these techniques in areas of social as well as commercial relevance with particular attention to issues of developing regions. More details about him are at: https://sites.google.com/site/biplavsrivastava/