Chatbots, Humbots, and the Quest for Artificial General Intelligence

Jonathan Grudin
Education Insights & Data
Microsoft Corporation
Redmond, WA, USA
jgrudin@microsoft.com

Richard Jacques
AI + Research
Microsoft Corporation
Redmond, WA, USA
rjacques@microsoft.com

ABSTRACT

What began as a quest for artificial general intelligence branched into several pursuits, including intelligent assistants developed by tech companies and task-oriented chatbots that deliver more information or services in specific domains. Progress quickened with the spread of low-latency networking, then accelerated dramatically a few years ago. In 2016, task-focused chatbots became a centerpiece of machine intelligence, promising interfaces that are more engaging than robotic answering systems and that can accommodate our increasingly phone-based information needs. Hundreds of thousands were built. Creating successful non-trivial chatbots proved more difficult than anticipated. Some developers now design for human-chatbot (humbot) teams, with people handling difficult queries. This paper describes the conversational agent space, difficulties in meeting user expectations, potential new design approaches, uses of human-bot hybrids, and implications for the ultimate goal of creating software with general intelligence.

CCS CONCEPTS

• Human-centered computing → Ubiquitous and mobile computing → Ubiquitous and mobile devices → Personal digital assistants
• Human-centered computing → Human computer interaction (HCI) → Interaction paradigms → Natural language interfaces

KEYWORDS

Chatbot; Intelligent Assistant; Conversational Agent; Virtual Companion; Humbot; Natural Language Interfaces; Artificial Intelligence; Artificial General Intelligence (AGI).

CHI 2019, May 4-9, 2019, Glasgow, Scotland, UK.
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05...$15.00.
https://doi.org/10.1145/3290605.3300439

ACM Reference format:
Jonathan Grudin and Richard Jacques. 2019. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. 11 pages. https://doi.org/10.1145/3290605.3300439

1 INTRODUCTION

The goal of artificial intelligence left the realm of science fiction when Alan Turing wrote in the *London Times* in 1949, "I do not see why [the computer] should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms." [35] In 1956, the term 'artificial intelligence' was coined and the field coalesced. Leading researchers forecast in the 1960s that ultra-intelligent computers would appear by 1980 or 1985 [8, 11, 17].

They didn't, but early efforts such as ELIZA in 1966 and PARRY in 1972 mimicked human beings, conversing by teletype and keyboard. Ongoing interest in anthropomorphic conversational software is reflected in Alice, Cleverbot, Xiaoice, Zo, Hugging Face and others. (Descriptions of all of these are found on Wikipedia and/or product pages.)

However, most recent activity has more limited goals; the initial AI focus is now the subfield artificial general intelligence (AGI). Nevertheless, intelligent assistants such as Siri and task-focused chatbots could lay a foundation for AGI and reveal the nature and extent of the effort that will be required to realize it.

Interactive bot software had precursors in mainframe and minicomputer networks and on the ARPANET. Interest picked up as low-latency internet and web use spread. Software agents such as recommender systems appeared in the 1990s [22], although they did not use natural language.

Most bots are not conversational. They are automata used for background tasks such as web crawling and notifications; for example, monitoring Wikipedia entries and notifying people on a watchlist of changes. A user's interaction may be restricted to setting parameters such as the scope or frequency of notifications. Social media led to the spread of social bots. Twitter's telegraphic message style made simulating a human easier. A Twitterbot can access a large audience by 'following' people. Broadcasting, forwarding, and responding are simple Twitter tasks for a bot masquerading as a person.

By 2010, Twitter was flooded with social bots that broadcast, followed, and responded to or forwarded tweets generated by people or other bots. Some were 'anti-social bots' with malicious intent: fraud, denial of service attacks, trolling, and political subversion [1, 38]. Algorithms were developed to differentiate bots and humans [10]. Varol et al. [36] studied Twitterbots in 2016 and concluded that between 9% and 15% of Twitter posters were bots, which meant there were 30-45 million of these simple bots.

Over the past several years, conversational agents benefited from massive data sources and advances in machine learning. The principal functions of intelligent assistants Siri (2011), Cortana (2014), Alexa (2014), Google Assistant (2016) and Bixby (2017) are to retrieve requested information and send notifications on a broad range of topics such as the weather, your meeting schedule for the day, and music requests. Also benefiting are chatbots that have deeper understanding of a specific domain, such as customer service chatbots. Task-focused chatbots can be a good match to mobile phone use where display space and input modalities are limited to voice or text. Entrepreneurs and investors saw opportunities for interfaces that are more engaging than telephone answering systems and standard software interfaces. Chatbots could interface to FAQ's, help order pizzas, track packages, or buy products online.

Between 2015 and the present, as many as a million task-focused chatbots flooded the Web, residing on social media platforms or striving to replace apps.[1] With the Internet of Things poised to deliver ever more digital data, the potential is unlimited. However, success has been unexpectedly elusive. For the foreseeable future, non-trivial task-focused chatbots might focus on handling the simplest queries and hand off others to human partners. This reduces their appeal as a low-cost interface, yet it may succeed where unassisted chatbots can't now and enable us to gather data and better understand the requirements for interacting with people. Undertaking to create a simulated person that converses with real people engaged in meaningful activities provides a powerful motivation for understanding people. Below, we lay out the recent evolution of this technology.

2 BOT TERMINOLOGY

Chatbots.org [6] lists 161 synonyms for "humanlike conversational AI," ranging from 'artificial conversational entity' to 'virtual support agent.' They aren't all really synonyms. There are important distinctions. Wikipedia's 'bot' entry lists 10 subcategories with distinct entries, some of which link to additional subcategories. Some overlap. For example, 'Twitterbots' are considered 'social bots' which are "a particular type of 'chatterbot' [or 'chat bot']" defined by the use of natural language. A few categories do not involve natural language, such as the web crawlers that constitute most web traffic.

Agreeing on terminology will enable us to converse about conversational software and identify related work. Our goal is not to impose terminology, but we need to be clear here about what we are referring to in our use of terms, in order to communicate effectively. All too often in the chatbot literature, different species of bot are given the same label and mixed together in studies. This creates confusion and muddies the interpretation of the results.

This paper focuses most on the chatbots that recently became a topic of intense activity, those that are oriented around a particular task, but first we consider ways to think about and categorize bots more broadly.

3 BOT TAXONOMIES

Depending on your goal, bots can be usefully grouped in several ways. As noted above, most do not converse, including Twitterbots that appear to be human as they generate posts but do not respond meaningfully to

[1] There isn't a chatbot registry, but in May 2018, Facebook reported 300,000 chatbots on the Facebook platform and Microsoft reported 300,000 chatbot developers using its platform [16; 5]. The older Pandorabot platform hosts over 200,000 chatbots. IBM, LINE and Amazon also offer chatbot platforms.

replies. From this point on, we will only consider those that engage in conversation.

One distinction is physical context. Is the chatbot on a laptop, phone, stationary device, or a mobile robot? Does it interact with single users, groups or both [29]?

Bunardzic [3] considers differences in a chatbot's memory of an exchange. Cleverbot maintains no state information—it retains no memory from one statement to the next, yet it won awards for its ability to converse. For each utterance, it searches a billion responses people have made to Cleverbot to find a human response when Cleverbot said something similar. Some chatbots maintain state for the current session, flagging information to guide subsequent responses, but retain nothing when the session ends. Users might like chatbots to remember their past conversations, like the fictional Samantha in *Her* (Warner Bros., 2013), but it can be technically challenging to access and make use of data across devices and constraints on retaining personal information may inhibit this capability.

The relationship of the conversational software and the task software can differ [9, 16]. Chat can be a layer over the task software; a user goes to one corner of an application interface to converse while using the application. At the other extreme, the conversation is the backbone, central to the interaction, with other task-related software out of sight. In between is a hybrid: the conversation is tightly integrated with the application interface; for example, a restaurant menu is displayed and a diner and the chatbot discuss choices.

3.1 A Taxonomy Based on Depth and Breadth of Focus

The taxonomy in Table 1 aligns with the development efforts noted in the introduction. You may prefer terms other than those in the left column, but the important distinctions lie in their breadth and depth of focus and the duration of their exchanges with humans.

Virtual companions engage on any topic and keep a conversation going. Intelligent assistants such as Siri also take on any topic but work to keep conversations short. Task-focused chatbots have a narrow range and go deeper, yet brief conversations are also their goal.

Table 1. A taxonomy based on conversation focus.

Type	Focus	Typical sessions	Examples
Virtual companions	Broad, deep	10 to 100's of exchanges	ELIZA, Cleverbot, Tay, Xiaoice, Zo, Hugging Face
Intelligent assistants	Broad, shallow	1-3 exchanges	Siri, Cortana, Alexa, Google Assistant, Bixby
Task-focused chatbots	Narrow, shallow	3-7 exchanges	Dom the Dominos Pizza Bot, customer service bots, Russian trolls, non-player characters

Recent research has addressed intelligent assistants and task-focused bots. The former are few in number and used by millions of people. The latter are vast in number and most are used by few people, as discussed below. Although forces are bringing about a degree of convergence, their different contexts of use and properties are important. Studies that merge data from both types obscure interpretation.

In prolonging casual conversations, virtual companions are descendants of ELIZA, taking steps toward the AI dream of general intelligence. They have a lot of personality, providing casual opinions or therapeutic responses. Handling open conversation effectively is a major challenge. Most virtual companions and intelligent assistants are developed and maintained by a few large software companies.

Customer service chatbots captured the attention of investors and software platform companies. Who could resist an interface that is more engaging than a telephone answering system? Other task-focused chatbots were built to replace apps or inhabit widely used platforms such as Kik, Slack, and Facebook Messenger. Residing on a platform simplifies installation, reduces the storage needed on a developer's device, creates opportunities to interact with other bots on the platform, and enables users to access the bot without leaving the platform they are on. Task-focused bots often follow a scripted tree structure similar to those used in less friendly answering systems.

More sophisticated malicious chatbots have appeared, conversing with people or other bots. Non-player characters (NPCs) in games are task-focused chatbots. Initially NPCs were triggered by simple phrases or actions and had limited scripted responses, which was OK—in games, robotic interactions can be acceptable. Some NPCs *are* charming robots. "Warning! Warning! This does not compute!" repeated a television robot popular fifty years ago. Today, scripted branching dialogues enable less predictable responses; machine learning algorithms enable some to adapt to changing conditions in the course of play.

The three bot types in Table 1 have analogs in human conversation. Consider a restaurant where one person takes you to a table, a server takes your order, and you talk with friends. Your exchange with the greeter is very efficient—how many in your party? —as it is with an intelligent assistant. The waiter shows more personality as you discuss options, but also quickly reaches a desired outcome, like a customer service bot. With dinner companions you engage in open-ended conversation that ideally resembles that with a virtual companion.

4 THE CHATBOT TSUNAMI

By 2014, several intelligent assistants were early stages of use for simple tasks. Investment in chatbots that could provide more depth on specific topics took off. In August 2015, Facebook launched M, a Messenger chatbot that handled purchases, arranged travel and made restaurant reservations. In January 2016, tech evangelist and hashtag inventor Chris Messina proclaimed, "2016 will be the year of conversational commerce." [26] His message was picked up by media and investors. In 2016, Facebook, Microsoft, IBM and LINE launched chatbot platforms. Slack launched an investment fund for bot development. Consulting companies joined the fray. A survey of senior executives and officials in Europe and South Africa reported that 80% expected to have customer-facing chatbots by 2020 [4]. A sample of Gartner predictions:

By 2020, 80% of new enterprise applications will use Chatbots.[2] *Nov. 4, 2016*

By 2021, most enterprises will treat Chatbots as the most important platform paradigm; and "Chatbots First", will replace the meme "Cloud First, Mobile First."[3] *Nov. 4, 2016*

By 2021, more than 50% of enterprises will spend more per annum on bots and chatbot creation than traditional mobile app development. Individual apps are out. Bots are in. In the "post-app era," chatbots will become the face of AI…[3] *October 16, 2016*

Enthusiasm continued through 2017. In September, Gartner predicted, "By 2022, 85% of customer service interactions will be powered by chatbots." [24] In the last five months of 2017, *TechCrunch* ran 14 excited chatbot articles.

This activity did not go unnoticed by researchers. As study results surfaced—several are discussed below—a counter-current appeared. Facebook reportedly found that 70% of Messenger chatbots were unable to answer simple questions [34]. In January 2018, *Inc.* magazine published an article about Digit founder and strong bot enthusiast Ethan Bloch: "How this founder realized the tech trend he'd built his company on was all hype." [23] Bloch's comment, "I'm not even sure if we can say 'chatbots are dead,' because I don't even know if they were ever alive" was widely quoted. In the first nine months of 2018, TechCrunch only published 6 paeans to chatbots. In January 2018, Facebook shut down M, the chatbot that helped inspire the surge.

Thoughtful articles by people involved in bot development asked "What happened?" and offered possible answers. They ended on positive notes, some citing the Gartner Hype Cycle phases: The Innovation Trigger, The Peak of Inflated Expectations, The Trough of Disillusionment, and The Slope of Enlightenment leading to the Plateau of Productivity. Having fallen into the trough, they looked back at the peak and wrote to contribute to enlightenment. Not all product ideas make it back up to the plateau, but these reflective practitioners felt that conversational AI will get there.

5 RECENT CHATBOT RESEARCH AND MEDIA ANALYSES

In 2017, Zamora [39] published a study of 54 people who used one of six chatbots for a week. Some used Siri,

[2] http://www.chatbotinsider.ai/industry-news/gartner-predicts-enterprise-chatbots-2020/

[3] https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/

which they had on their phone; others used highly regarded publicly accessible task-focused chatbots to check finances, news, and social media, order a meal to go, and book movie tickets. As the paper title indicates— "I'm sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations"—expectations were not met. Users reported slow replies, unwanted notifications, and no gain in speed or efficiency. Apart from Siri users, few said they might continue using the chatbot (Zamora, personal communication).

Jain et al. [13] enlisted 13 relatively tech-savvy first-time users to try each of eight Facebook Messenger chatbots daily for 3 days. The authors selected Chatbottle's top-rated Messenger chatbot in each of eight domains: news, travel, shopping, social, game, utility, chit-chat and entertainment. Each chatbot had over 1000 likes. Bots and participants exchanged 9968 messages. The principal finding: "expectations of the users were not met." It was a short albeit active trial. The next study indicates that initial experiences are often decisive, at least for intelligent assistants.

In a study subtitled "The gulf between user expectation and experience," Luger and Sellen [19] studied existing users of Siri, Cortana, and Google Now. They concluded that intelligent assistants usually failed to meet expectations and users retreated to using them for menial, low-level tasks. Users reported trying new tasks a few times, then giving up when it was unsuccessful. Although not a study of task-oriented chatbots, it joined the chorus.

Luger and Sellen encountered a dilemma: All save one participant reported initially trying the software as a playful exercise. They liked occasional humorous programmed bot responses. However, the authors concluded that humor creates expectations of human-like qualities that intelligent assistants cannot deliver. They advised designers to avoid playful responses.

The opposite conclusion was reached in a 2018 study of a human resources FAQ chatbot deployed over several months to 337 new employees [18]. The authors recommended careful design that includes humor, noting that users of a previous chatbot ignored its humorless efforts at congenial conversation and resorted to typing keywords to search for FAQ answers. Humor also attracts some people to virtual companions.

A team of eight spent 18 months building, testing and refining a meeting scheduling chatbot, calendar.help [7]. Requests were processed in three tiers: those handled entirely by software (Tier 1); routine problems beyond the system's capability ('microtasks') were handled by people using simple scripts to carry them out (Tier 2); and difficult cases were passed on to experts (Tier 3). The fourth version of the system could handle 39% of scheduling requests in Tiers 1 and 2, with about 5% of all processes completed without human assistance [Cranshaw and Monroy-Hernández, personal communication]. The team analyzed the escalations to automate microtasks where possible and break difficult tasks down into sets of more easily handled microtasks.

This is striking, because meeting scheduling seems an ideal task for a bot. A person designates participants (most are two-person meetings). The bot has access to calendars and email addresses. It is not under time pressure. It can send email proposing times, coordinate responses that are generally straightforward, and send reminders when replies are overdue. Yet it required a human in the loop. We will return to this example.

In 2018, Dave Feldman, the former design manager of the Facebook Messenger chatbot platform, published "Chatbots: What happened?" [9]. Other authors extended his analyses [16, 27]. In 2017 and 2018, the authors of this paper were independently involved in some task-oriented chatbot projects that did not meet expectations. In the next section we collect and synthesize our observations and those from these studies and essays addressing the question of why most task-focused chatbots failed to meet user expectations.

6 SUCCESSES AND CHALLENGES WITH TASK-FOCUSED CHATBOTS

FAQ's. Interfacing to a list of Frequently Asked Questions is a tractable chatbot task. For questions that are frequently asked, queries and terminology are relatively predictable and responses have been written. Unexpected questions can be collected and used to improve the chatbot. For users, a short FAQ list can be visually scanned quickly but for a long FAQ list, verbal queries can be faster. Chatbots compete with keyword search which can be faster still, so the challenge is to be congenial enough to be preferred [18]. Tools are available that do much of the work of converting a FAQ to a chatbot.

Niche Successes. Although the highly rated chatbots used in studies did not meet all user expectations, many people like them. Some chatbots find niche audiences large enough to sustain the effort. Some contexts offer advantages; the chance observation that people liked ELIZA as a counselor was followed years later by studies finding that some people like and benefit from the non-judgmental nature of chatbot counselors [11]. However, not many chatbots succeed for many people. The following are obstacles faced by chatbot designers.

Tasks are complicated. The calendar.help meeting scheduler addressed a seemingly straightforward task that proved to be complex. After a year of refinement by a capable team, only 5% of meeting scheduling processes were automated. The blocker in automating workflow is often exception-handling: We focus on the standard workflow and fail to anticipate the wide range of contingencies that arise. A deviation in one step or another may occur almost every time. Lucy Suchman [33] saw this as a challenge for AI systems in general. When a system that can handle 80% of events is faced with a situation that it can't handle, the software is less effective than a person in seeing how close a solution is, whether it should be escalated or abandoned, and if escalated to whom.

Many existing apps benefit from decades of experience and refinement. Chatbots that undertake to replace apps may have a success rate even lower than applications did in the past since the chatbot must compete with mature software, and marketing to reach the right audience is not trivial.

Conversations are complicated. Unlike scripted dialogue trees, natural language conversations are not linear. They can be multi-threaded, hop back and forth, and circle around. Human conversations are more than words. Software is generally oblivious to posture, eye gaze, gestures, facial expressions, tone, shifts in conversation direction that reveal a speaker's state, the conversation history from prior sessions, and so on. Some of these are being researched, but the science has a way to go.

Natural language is complicated. Despite major advances, speech and natural language understanding fall short of human capability. Also, chatbot systems have at best primitive models of their conversational partners, so setting appropriate expectations is important. However, they can't be lowered too far: engagement is the primary attraction of chatbots, so the experience must clear a reasonably high bar.

Platforms and tools are complicated—and who is leading by example? Chatbot construction is unlike other software development. It requires different tools, guidance and skills, such as generating a sample of utterances for a given conversation intent that will maximize the ability of machine learning algorithms to match a user utterance to the right intent. In our experience, chatbot designers are often unaware of the range of available tools and select inappropriately.

Feldman takes the major platform companies to task for marketing platforms and tools but not demonstrating how to use them. Why didn't they build exemplary chatbots? Why weren't they more proactive with funds and incubators? Why didn't they make mentors available or provide design and engineering resources? Facebook tried with M, the Messenger chatbot withdrawn after two and a half years. Some of Facebook competitors engaged in fundraising and partnership efforts, but Feldman's points are well taken.

The Uncanny Cliff and Mission Creep—complications of anthropomorphism. Robots that appear almost human elicit feelings of eeriness. This phenomenon, called the uncanny valley, has been linked to the human visual system [28]. Task-oriented chatbots encounter a different problem. A bot that is knowledgeable within a narrow task focus often cannot answer a query on a related topic that any human expert could. It also can't handle simple general information questions. There is an unexpected sharp boundary between what is known and what isn't, an uncanny cliff. For example, a chatbot designed to answer a visitor's questions about the home team players and the stadium may quickly lose a visitor's confidence if it has no response to "Where I can I find a meal near the stadium?" or "How can I get from the train station to the stadium?"

When a chatbot appears dim because it lacks knowledge or information associated with its expertise, the developers' reaction is often to expand the bot's capability. This 'mission creep' can be a resource black hole. It can also negatively impact performance if the platform does not scale up gracefully.

The Demo Trap is a general phenomenon to which chatbot projects are unusually susceptible. A strong team may in a couple weeks create a working chatbot through which a real path—no smoke or mirrors—produces the desired outcome. Strong team, solid effort, successful demo. Management greenlights the project. The hard work appears to have been done. Isn't it just a matter of entering more intents (the basic user question or goal), sample queries, and responses? No, scaling up is far more challenging. The demo matched queries to the right intent—with five intents coded. With 500 or more intents—people given the chance can ask anything—matching is less accurate and the time to respond can lag. During the demo, no one tried to provoke the chatbot or get it to say something inappropriate, but they will after launch. We have seen successful demos of working prototypes that could not scale without a much larger effort.

Sharing lessons learned. Expectations generated by successful demos and the subsequent large investments in chatbot projects amplify the natural reluctance to examine unsuccessful projects closely and share what was learned. Advice from veteran chatbot developers sometimes concludes, "When your bot is released, don't consider your job over. The most important step is to get feedback from your first users to guide improvements." The hidden message is, "Listen to users, because most chatbots do not meet expectations."

Tool and platform providers are in a bind. They need developers to use them to provide feedback for improving the tools and to learn where their chatbots fare best. If low success rates are publicized and use drops, so does an opportunity to make progress. Yet with hundreds of thousands of unsuccessful efforts, sharing outcomes should be a priority.

7 HUMANS AS INVISIBLE CHATBOT PARTNERS

An ethical debate over whether people should be informed when they are conversing with a chatbot rose in volume when Google Duplex succeeded in convincing people that it was a person. [31] The human-sounding bot inserted "uh"s, "mhm"s, and other mannerisms to fool people. Another debate arose when it was discovered that humans are hired to pretend they are chatbots: "The humans hiding behind the chatbot," [12] and "The rise of 'pseudo-AI': how tech firms quietly use humans to do bots' work." [32]. Companies such as X.ai and GoButler.com offered services to people who believed they were being served by chatbots when their requests were actually handled by employees. Facebook Messenger's flagship chatbot M secretly forwarded requests it couldn't handle to human agents.[4] Had developers known this, some of the chatbots built to run on Messenger might not have been built. Similarly, calendar.help did not reveal that the respondent was usually a person.

A full account of these human bots or humbots is nuanced. People masquerade as bots for various reasons. Through half a century of natural language understanding research, researchers have played chatbots in "Wizard of Oz" experiments ("pay no attention to that man behind the curtain"). They test to see whether planned system features could handle human queries before doing the work of building them. 'GUS, a frame-driven dialog system' [2], published in 1977 by five leading AI and cognitive science researchers, described an airline reservation bot tested in a Wizard of Oz study, with an author pretending to be the chatbot [2]. The researchers found that even for this narrow task, handling natural language was very tricky. They did not build GUS and airline reservation chatbots are not in routine use forty years later.

Another use of humbots is to intrigue customers who will hire an AI servant that is available around the clock to have food delivered, book theatre tickets, buy last-minute gifts, schedule meetings, and so on, "making it look like magic." These customers might not feel as good if they knew that their requests were handled by minimum-wage night shift workers with a job "so mind-numbing that human employees said they were looking forward to being replaced by bots." [32]

Humbots can save face. "Using a human to do the job lets you skip over a load of technical and business development challenges. It doesn't scale, obviously, but it allows you to build something and skip the hard part early on." [32] Many companies were caught in demo traps in 2016 after announcing the imminent release of chatbots after demos impressed decision-makers. Humbots can preserve the aura of being on the leading edge and be thought of as Wizard of Oz experiments conducted outside the laboratory, with plans to collect data to be used to develop a working chatbot.

There is a wave of work by developers who know that their tasks cannot be fully automated. Their chatbots handle queries or comments for which there is high confidence of a correct response and pass the rest

[4] https://en.wikipedia.org/wiki/M_(virtual_assistant)

to humans. Difficult cases can be used to further refine or train the bot. This can be applied to tasks for which human partnership is affordable. The [24]7.ai platform promotes itself as “Chatbots helping Agents, Agents Helping Chatbots.”[5] When routine work is handled by a chatbot, the human tasks may not be mind-numbing.

Examples are Facebook M and calendar.help. The developers of the latter first conducted a Wizard of Oz experiment, then tested systems ‘in the wild’ with people using them in their daily work. The goal was to improve the system as it is used by analyzing difficult scheduling events. A similar service is marketed by Claralabs.com. Meeting scheduling was also among M’s human-assisted tasks before Facebook retired the bot.

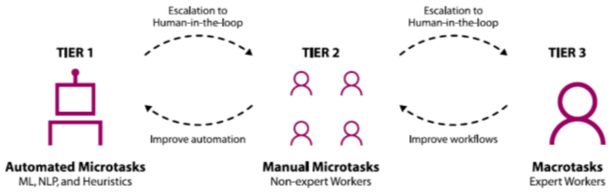

Figure 1: A three-tier architecture for handling tasks. (From [7], with permission.)

The architecture of calendar.help is shown in Figure 1, from Cranshaw et al. [7] Analysis of macrotasks that take expert humans minutes to carry out (Tier 3) is undertaken to decompose them into a workflow of microtasks that can be executed manually in seconds by non-expert workers (Tier 2). When possible, microtasks are automated (Tier 1).

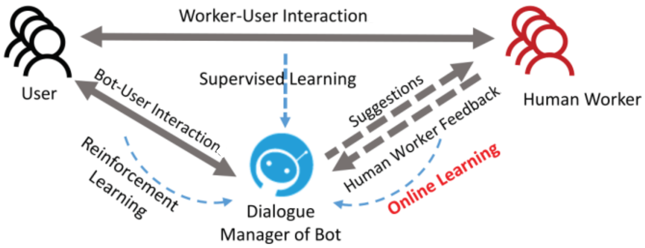

Figure 2: Co-Chat framework. Adapted from [20].

Luo et al. [20] uses a similar approach for Co-Chat, a system that can support multiple tasks (Figure 2). In ‘How to build human/bot hybrid customer service’ Kalvainen [15] provides a more detailed architecture but writes little about the development process.

Georgia Tech deployed a virtual teaching assistant for an online master’s degree course in 2016 [21]. Built on the IBM Watson platform and deployed without an indication that “Jill Watson” was a bot, it performed reasonably well, although less well the next semester when students knew the teaching team likely included a chatbot. (It actually included two.) Like other human in the loop systems, Jill Watson calculated a confidence level for responses and passed difficult queries to human teaching assistants.

It is not surprising that service-oriented chatbots would gravitate toward bot-human partnerships. Tiered response is integral to many customer service processes that rely on humans. The first tier answers the phone and tries to handle your inquiry, but if unlikely to succeed, may escalate your call to a more expert tier.

This is good news for those worried about job loss due to automation. The software may be a junior partner that shifts many tasks to people. We get some work done for us and we help the bots grow.

8 IMPLICATIONS FOR DESIGN

After plunging from the peak of inflated expectations to the trough of disillusionment, it can be difficult to discuss experiences. But lessons have been learned that can be applied before, during, and after chatbot development, one of which is to look hard to find lessons learned!

Planning. Teams should assess possible platforms and tools, which have different strengths. Inspect examples of bots built with each. Some that work well for a small chatbot, such as an FAQ, do not scale up. We have seen projects reach a point where extensions to improve the bot degraded its performance. Learn how to use the tools. For example, a common approach is to identify intents (user goals) and create responses for each, then enter examples of user queries or utterances that correspond to each intent. The optimal number and type of examples to generate can be counterintuitive.

Personality. Personality can increase engagement, which is a key goal of task-focused chatbots, but it raises expectations significantly. It is common to ask a new acquaintance questions, and a bot with personality must be ready to field personal questions, whether polite, playful, or mischievous. Robotic responses or shutdowns are undesirable but to respond meaningfully

[5] https://www.247.ai/

expands the design task dramatically. Some chatbots beat a hasty retreat from personality. For example, "Casey the UPS Bot" was released with fanfare in 2016, but after some withering reviews was replaced by a "UPS virtual assistant" that sits namelessly alongside keyword search and other customer service tools.

Personality is especially salient for virtual companions that aim for extended conversations. The lack of memory for past conversations and no persistent model of an individual over time becomes a personality trait that can annoy users. Chronic amnesiacs feature in amusing film comedies, but ultimately it can be tiresome to explain yourself repeatedly to a virtual companion who is otherwise congenial.

Task difficulty. Chatbot developers often marvel at how difficult seemingly simple tasks such as booking a dinner reservation prove to be. Determine in advance what will be involved. Do not be misled by competitors who may secretly employ humbots to carry out tasks. If an existing app does do the job, examine it closely. Shadow people carrying out the tasks, paying special attention to exceptions to the routine that are encountered and how they are addressed. Decide early whether your use scenarios and business model could work if a human must be added to handle difficult cases.

Use context. Will the bot be accessed primarily by phone? How forgiving of occasional speech recognition errors can it be? This can guide the selection of voice or typing. A display alternative that can greatly simplify an exchange by constraining options is to present a handful of cheerful text bubbles. Another approach that has been used is to engage users in a discussion around a viewed object, such as a catalog page or a restaurant menu, with the conversation on the side but items ordered via the application. This provides terminology, constrains options, and is more flexible than a sequential climb up a branching tree.

Scripted responses. Developers of any type of bot can create humorous or interesting canned responses to frequently asked questions, such as personal queries about the bot's background or tasks such as dividing a number by zero. Amusing canned replies boost engagement but as noted above can raise expectations and trigger user efforts to find more humor, which may not be forthcoming. Mission creep is a risk.

Transparency. Should chatbots and humbots reveal who they are? The Google Duplex controversy brought calls for chatbots to come out of the closet, but before saying "of course," consider the other cases. Wizard of Oz studies rely on deception until a participant is done. The "Wizard of Oz in the wild" collection of task complexity data by employing a humbot introduces nuance. Calendar.help was known to be a chatbot by the meeting scheduler but not by all meeting attendees. Some schedulers were delighted to appear important enough to appear to have a human admin working for them whereas others felt uneasy. Georgia Tech's Jill Watson humbot went incognito; students might not have as readily followed her advice otherwise, but most gave her good marks.

Plan to iterate. The most important lesson could be that the release of a chatbot is the beginning and not the end of development [5]. It won't meet expectations on every dimension and may not initially succeed on any dimension. Small-scale testing is advisable. Examine experiences that people have with your adolescent chatbot, as even with machine learning algorithms it will be unable to learn much on its own. Build adequate staffing into the plan. Like parents, those building on chatbot platforms are often dismayed at the level of ongoing support required by their offspring when they have left home and ventured into the world.

9 IMPLICATIONS FOR ARTIFICIAL GENERAL INTELLIGENCE: A CONVERGENCE OF CHATBOTS

The goal of many AI pioneers was to create one intelligence to rule them all. AI initially equated to artificial general intelligence and was considered to be within reach. The Turing Test was devised to measure progress. ELIZA, the first well-known conversational software, masqueraded as a Rogerian-style therapist that coaxed some people into reflecting and talking about themselves in ways that they found useful [37], as do virtual companions today [31].

General intelligence will require software that has general knowledge, is capable of carrying out tasks and can engage in open-ended conversations. Could there be a synthesis of the skills of today's intelligent assistants, task-focused chatbots, and virtual companions?

User expectations create pressure to bring them together. Disappointment with shallow intelligent assistant knowledge inspires developers to add skills, but comprehensive coverage is too much for one company. Perhaps an intelligent assistant could partner with many task-focused chatbots. Two illustrations of the potential for distributing the effort are Amazon Lex, which allows developers to bridge their chatbots to Alexa, and Microsoft Business Bot, which converts detailed information that a business enters into a form into a bot that appears in search engine results for that business and optionally also on the business's site.

Similarly, user expectations drive task-oriented chatbots to expand their breadth of knowledge to handle peripheral questions and appear more personable. And developers of virtual companions such as Xiaoice and Zo add to their repertoires of skills to maintain user interest [30].

The field has made progress in handling open-ended questions. Intelligent assistants have found niches that are likely to expand. Today's virtual companions are more engaging than their predecessors. We will see how many people they attract and keep engaged over time.

Tasks are the critical component. AI pioneers defined intelligent machines by describing a wide range of tasks they could carry out, most of which are much more complex than buying a movie ticket. Today's chatbots are designed to automate tasks that most people can do, perhaps assisted by a YouTube video: Cook a meal, manage finances, book airline tickets, shop, establish a diet and exercise plan, file taxes, choose a restaurant, scan the news and go deep on an item of interest, organize a meeting, plan a vacation, schedule medical appointments, buy a pet, find recipes to match the food on hand, and so on. There are a lot of tasks out there. Hundreds of thousands of task-oriented chatbots have revealed how complex those seemingly straightforward tasks actually are, the degree of problem-solving and exception-handling they demand.

After falling into the trough of disillusionment, we are seeing flashes of enlightenment. Will we move forward and upward to the plateau of productivity, and if so, how quickly?

10 STEPPING INTO A NEW ERA

The challenges may seem to paint a gloomy picture, but the state of bot art is following a path laid out in the early 1960s by JCR Licklider in prescient papers that outlined a field, human-computer interaction, that did not yet exist. Licklider was a psychologist and engineer at MIT who as ARPA director envisioned and funded the ARPANET, which evolved into the internet. He forecast three eras in the interaction of people and digital technology. The first would be devoted to improving input devices and displays, including speech recognition and language understanding—a blueprint for HCI. The final era would be governed by ultra-intelligent computers that his AI colleagues assured him would arrive by 1980. In between would be the era in which humans and computers would work as partners. In 1960 he wrote, "there are many human-machine systems. At present, however, there are no human-computer symbioses." [17]

Today there are human-computer symbioses. It is by no means an equal partnership, but software works autonomously around the clock on our behalf. Interactive bots are part of this. In building them, we discover much about ourselves and how we work, the range of our knowledge and curiosity, the subtlety of language, and the number and complexity of the tasks we routinely handle.

The capable machines that were envisioned over half a century ago must synthesize elements of virtual companions, intelligent assistants, and task-oriented chatbots. We are discovering the scope of this undertaking.

It took 50 years to complete the work of Licklider's first era: achieving the foundation for human-computer interaction. Licklider concluded that the second era of human-computer interaction, which we have just stepped into, could last 10 or 500 years and "should be intellectually the most creative and exciting in the history of humankind."

ACKNOWLEDGMENTS

The authors thank the reviewers and Clayton Lewis for useful comments, the many people we worked with and interviewed, and several authors of cited papers who provided additional information.

REFERENCES

[1] Norah Abokhodair, Daisy Yoo, and David W. McDonald. 2015. Dissecting a social botnet: Growth, content, and influence in Twitter. Proc. CSCW 2015, 839–851.

[2] Daniel G. Bobrow, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. GUS, a frame-driven dialog system. Artificial Intelligence, 8, 2, 155–173.

[3] Alex Bunardzic. 2016. Four types of bots. Chatbots Magazine, May 17. https://chatbotsmagazine.com/four-types-of-bots-432501e79a2f

[4] Business Insider. 2016. 80% of businesses want chatbots by 2020. December 14. https://www.businessinsider.com/80-of-businesses-want-chatbots-by-2020-2016-12

[5] Elaine Chang and Vishwac Sena Kannan. 2018. Conversational AI: Best practices for building bots. Build 2018. https://medius.studios.ms/Embed/Video/BRK3225

[6] Chatbots.org. 2018. 161 humanlike conversational AI synonyms. https://www.chatbots.org/synonyms/

[7] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. Proc. CHI 2017, 2382-2393. https://dl.acm.org/citation.cfm?id=3025780

[8] Brad Darrach. 1970. Meet Shaky: The first electronic person. *Life Magazine*, November 20.

[9] Dave Feldman. 2018. Chatbots: What happened? Chatbots life, April 10. https://chatbotslife.com/chatbots-what-happened-dcc3f91a512c

[10] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. Communications of the ACM, 96-104. https://arxiv. org/abs/1407.5225

[11] Jonathan Grudin. 2017. From tool to partner: The evolution of human-computer interaction. Morgan and Claypool.

[12] Ellen Huet. 2016. The humans hiding behind the chatbots. Bloomberg, April 18. https://www.bloomberg.com/news/articles/2016-04-18/the-humans-hiding-behind-the-chatbots

[13] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Schwetak N. Patel. 2018. Evaluating and informing the design of chatbots. Proc. DIS 2018, 895-906. https://dl.acm.org/citation.cfm?id=3196735

[14] Khari Johnson. 2018. Facebook Messenger passes 300,000 bots. VentureBeat, May 1. https://venturebeat.com/2018/05/01/facebook-messenger-passes-300000-bots/

[15] Erik Kalviainen. 2016. How to build human/bot hybrid customer service. Medium, July 13. https://medium.com/making-meya/how-to-build-human-bot-hybrid-customer-service-99201a7e2703

[16] Justin Lee. 2018. Chatbots were the next big thing: what happened? medium.com, June 5. https://medium.com/swlh/chatbots-were-the-next-big-thing-what-happened-5fc49dd6fa61

[17] J.C.R. Licklider. 1960. Man-computer symbiosis. IRE Transactions of Human Factors in Electronics HFE-1, 1, 4-11.

[18] Q. Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco P. Crasso, Dakuo Wang, Michael Muller, N. Sadat Shami, and Werner Geyer. 2018. All work and no play? Conversations with a question-and-answer chatbot in the wild. Proc. CHI 2018, paper 3.

[19] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. Proc. CHI 2016, 5286-5297.

[20] Xufang Luo, Zijia Lin, Yunhong Wang, and Zaiqing Nie. 2018. CoChat: Enabling bot and human collaboration for task completion. Proc. AAAI-18, 5301-5308.

[21] Jason Maderer. 2017. Jill Watson, round three. Georgia Tech news center, January 9. http://www.news.gatech/2017/01/09/jill-watson-round-three

[22] Pattie Maes. 1994. Agents that reduce work and information overload. Comm. ACM, 37, 7, 31-40. https://dl.acm.org/citation.cfm?doid=176789.176792

[23] Sonya Mann. 2018. How this founder realized the tech trend he'd built his company on was all hype. Inc., Jan. 30. https://www.inc.com/sonya-mann/digit-chatbots-are-dead.html

[24] Brian Manusama and Guneet Bharaj. 2017. Making live chat a must-have engagement channel. Gartner, September 28.

[25] Terrence McCoy. 2014. A computer just passed the Turing Test in landmark trial. Washington Post, June 9. https://www.washingtonpost.com/news/morning-mix/wp/2014/06/09/a-computer-just-passed-the-turing-test-in-landmark-trial/

[26] Chris Messina. 2016. 2016 will be the year of conversational commerce. Medium, January 19. https://medium.com/chris-messina/2016-will-be-the-year-of-conversational-commerce-1586e85e3991

[27] Casey Newton. 2018. Bots didn't flop; they just became invisible. The Verge, August 15. https://www.theverge.com/2018/8/15/17689322/bots-comeback-intercom-eoghan-mccabe-interview-converge-podcast

[28] Ayse Pinar Saygin, Thierry Chaminade, Hiroshi Ishiguro, Jon Driver, and Chris Frith. 2011. The thing that should not be: Predictive coding and the Uncanny Valley in perceiving human and humanoid robot actions. Social Cognitive and Affective Neuroscience, 7: 413–422.

[29] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K.E. Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. Proc. CHI 2018, paper 391.

[30] Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Journal of Zhejiang University Science C, 19(1), 10-26. https://arxiv.org/abs/1801.01957

[31] Olivia Solon. 2018. Google's robot assistant now makes eerily lifelike phone calls for you. Guardian, May 8. https://www.theguardian.com/technology/2018/may/08/google-duplex-assistant-phone-calls-robot-human

[32] Olivia Solon. 2018. The rise of 'pseudo-AI': how tech firms quietly use humans to do bots' work. Guardian, July 6. https://www.theguardian.com/technology/2018/jul/06/artificial-intelligence-ai-humans-bots-tech-companies

[33] Lucy Suchman. 1987. Plans and situated action. Cambridge University Press.

[34] Laurie Sullivan. 2017. Facebook chatbots hit 70% failure rate as consumers warm up to the tech, MediaPost, Feb. 22. https://www.mediapost.com/publications/article/295718/facebook-chatbots-hit-70-failure-rate-as-consumer.html

[35] Alan Turing. 1949. London Times letter to the editor, June 11.

[36] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. https://arxiv.org/abs/1703.03107

[37] Joseph Weizenbaum. 1966. ELIZA — A computer program for the study of natural language communication between man and machine. Comm. ACM, 9, 1, 36-45.

[38] Samuel C. Wooley. 2016. Automating power: Social bot interference in global politics. First Monday, 21, 4, April 1. http://firstmonday.org/ojs/index.php/fm/article/view/6161

[39] Jennifer Zamora. 2017. I'm sorry, Dave, I'm afraid I can't do that: Chatbot perceptions and expectations. Proc. HAI 2017, 253-260. https://dl.acm.org/citation.cfm?id=3125766