HCI Research Challenges of the Next Generation of Conversational Systems

Claudio S. Pinhanez
IBM Research
São Paulo, SP, Brazil
csantosp@br.ibm.com

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

Abstract
As we move past from the deployment of the first conversational systems, a new generation is shaping up with interaction patterns beyond the Q&A paradigms of today. The next wave is likely to include systems with clearly defined personalities, nuanced and emotional speech, and contexts with multiple bots and users. We propose here that a new set of design and interface challenges will be raised in the context of those upcoming systems. Among them, we address five challenges which we believe are going to become relevant for the next generation of conversation systems: handling human-machine pidgins, managing language precision, creating and conveying personality, knowing when to speak, and creating the illusion of a mind.

Author Keywords
Conversational interfaces; HCI research challenges; voice-based interaction.

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);** User studies;

Introduction
The first generation of conversational systems to reach a large spectrum of users has already been deployed during the last decade, through smart speakers, mobile phone

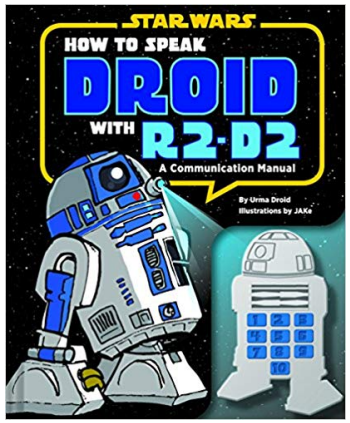

Figure 1: Cover of the 2013 book *"How to Speak Droid with R2-D2"* by Umar Droid, a satire on how human machine communications in *Star Wars* can be learned. It includes a button panel allowing the reader to produce the correct tones.

apps, and *intent-action* [20] enterprise chatbots. The results have been mixed and most of the successful systems have relied on simple question-and-answer or command-and-control frameworks. Many challenges still remain, such as building a system able to chit-chat effectively [11], .

There is a considerable amount of research and development being done to bring the next generation of conversational systems. For instance, a lot of progress has been made recently in *end2end* [4] generation of the text output, in *speech-to-intent* technology, on improving the humanness of dialogue *Duplex* [16], and on multi-party conversation [5]. However, research on the HCI aspects of the next generation of conversational systems is still limited [3].

This paper lists and discusses some key HCI research themes which are particularly relevant in the context of the likely advancement of those technologies. Although not an exhaustive list, we hope it brings light to some interesting and important HCI challenges that this new generation of machines, and their users, are likely to face.

Conversing in Human-Machine Pidgins

A *pidgin* is a language that develops between two or more groups who do not have a language in common and use a simplified language, known to both, to exchange communications [10]. It has been reported that users of conversational systems often learn, in a short amount of time, that they can get answers from those systems by using a simplified form of natural language [12]. For instance, instead of asking a full sentence like *"What is the time?"*, they simply say *"Time?"*. We refer as *human-machine pidgins* such languages constructed by users to simplify and make more effective their communication with conversational systems.

We could, in an extreme, say that human beings can converse with a machine using both natural language (such in chatbots) or in machine language (such in *ssh scripts*). An allegorical way to look into this is to contrast the two main robots of the *Star Wars* saga: *C3PO* can allegedly communicate in six million different forms, following etiquette and customs, while *R2D2* speaks only *Droidspeak* which some humans, like *Luke Skywalker*, are able to understand (see fig. 1). In practice, we could say that communication between humans and machines seems to fall somewhere in the continuum between the C3PO and R2D2 abilities, wherever users feel comfortable and machines are accurate.

One of the key challenges facing HCI researchers and designers of conversational systems is to understand and control the development of such human-machine pidgins and their effects on the user experience. The way users communicate with conversational systems seems to start in the natural language realm but eventually drift to a conversational pattern which deviates from traditional natural language, and likely to be personalized to the comprehension abilities ascribed by the user to the machine.

Simply, as current NLP technology is far from the abilities of C3PO, the tendency of each user developing his own pidgin is strong, making design and evaluation of conversational interfaces considerable more difficult. To understand better the mechanisms in which users develop pidgins, to design interfaces which evolve with them, and to devise strategies to avoid the establishment of human-machine pidgins, are all important challenges facing HCI researchers of conversational systems.

Managing Language Precision

Many practical situations of professional conversation have very strong requirements on the precision of the language used. A common reason is legal liability, which often dictates non-colloquial patterns of language such as the one

Figure 2: A commercial mug depicting the major personality elements of Apple's *Siri*: Smart, Wow (twice), Talented, Love, Loving, Good Look, Cute, Mature, Super, Super Mind, Intelligent, Nice, Cool, Funnyest, Lucky, Understanding, Rocking, You Always Make Me Happy.

call-centers use, conversations with health professionals, and management of human resources. Similarly, some domain-specific contexts, such as in some governmental situations, often have very established norms of discourse.

Although language precision issues are also present when recognizing the users' utterances, the main requirements normally happen in the process of language generation by the machine. In the case of typical *intent-action* conversational systems, where the machine utterances are composed manually by the designers and developers of the system, assuring the precision of the language is often a time-consuming task requiring a well-developed curatorial system. Changing a single utterance, in such cases, may require many levels of approval within a business process.

In the case where the machines' utterances are generated by *end2end* systems [4] trained with conversation and user log data, the assurance that the correct level of precision of the language is employed is even more difficult. In particular, it is very hard to verify whether the system generates correctly-worded utterances in all dialogue circumstances.

In practice, a conversational system moves between contexts where different levels of language precision are needed. A key challenge for HCI researchers is to support the development and evaluation of systems which output text with the right kind of precision for a given conversational context. In particular, determining what people and businesses consider the "right" language precision in a context is itself an important area for HCI research.

Designing and Conveying Personhood

In general, conversational systems of the first generation have portrayed dry and dull, impersonal personalities (like *Siri* and *Alexa*), or seem to be modeled after call-center attendants, often female. Interestingly, the latter led to criticism that giving feminine characteristics to conversations systems seems to reinforce old stereotypes of servitude [13].

Nevertheless, some key questions have remained unsolved: how to determine and design the human characteristics of a conversational system; and how to convey them effectively. Considering the former question, some early research in voice systems [14] seems to indicate that lack of personhood traits, such as gender, or ethnicity, triggers distrust in users. Similarly, non-coherent voice or conversation traits are not well received by users [15, 1].

So given a system, a context, and its potential users, what should interface designers do to determine the most effective human characteristics? In spite of some design methods proposed [17], there is a general lack of design guidelines and methods to address this question.

But even the best-defined personality for a conversational system will not succeed if it is not effective in conveying it. If an enterprise wants a customer service chatbot to be smart, funny, and warm, realizing it in terms of what it says and how it behaves is still a formidable challenge. Although the devise of AI methods and algorithms to produce language with such characteristics depends heavily in research mostly outside the HCI realm, there is a fundamental role to be played by our community in creating and defining metrics which evaluate those algorithms.

In particular, because it is necessary that those systems are perceived by the users as having such characteristic and traits, it is insufficient to evaluate only the formal quality of the generated output. The context of a machine uttering language must be considered as part of the equation: the same utterance could be perceived as funny if spoken by a person, but cynical if by a machine. All this calls for basic research on how people perceive conversational systems.

Figure 3: Majel Barrett-Roddenberry, American actress, who voiced the *USS Enterprise* computer in most of its seasons and movies, besides occasional roles as a member of the crew.

Knowing When to Speak

Most of today's voice-based systems have the annoying behavior of requiring to be called to a conversation with a vocative, often a branded name like *Siri* or *Cortana*. Such systems would likely be much more user friendly if they were able to understand the context enough to answer if and when they were needed. Moreover, they would have to do that in moments people consider appropriate to be interrupted, known in linguistic as *turn-taking points* [19].

Although there has been some recent work in this direction, notably in the context of multi-party conversation [18], just getting rid of the use of vocatives is a formidable challenge. Part of the difficulty comes from the trade-off between the need to monitor the conversation going on among people to be able to determine when action from the machine is required and the privacy people expect to have in the presence of AI systems, in particular the ones listening them.

The current design of those systems, which portrays them as "deaf" until the correct vocative is used, seems to convey a strong message of privacy preservation. However, not only in reality much more than vocatives is analyzed by the system, but also this practice forces the user into this annoying pattern, reminiscent of the 1960s' *Star Trek* TV shows (see fig. 3), of announcing with a vocative that is time for the machine to join the conversation.

We see also here an opportunity for creative research towards different and more natural mechanisms by which users can bring such systems to a conversation without use of vocatives. This may include non-verbal gestures such as looking towards the system [7], subtle references, and similar. At the same time, as multi-party conversation systems become more common, understanding what kinds of interruption and turn-taking behavior is acceptable when interacting with a machine also becomes important.

Creating the Illusion of a Mind

The AI community has a long, and still running, conversation about what a *mind* is and whether machines can have one [8, 9]. Drawing from this discussion, but not bound by it, HCI researchers must look increasingly into the issue of whether people, while interacting with a machine, perceive or not elements of an interior AI mind, and, more importantly, whether the perception of an AI mind changes elements and characteristics of the user experience.

For instance, suppose we have a situation where a user believes, from the machine's actions, that the machine has its own intentions, contrary to hers. The user may then decide to try to fool the machine, or lie, so to achieve her own goals. Employing for a moment Dennett's *intentional stance* paradigm [6], possession of its own intents is a cue for a mind. In this situation it may be advisable to downplay the elements triggering the perception of an AI mind inside the machine, while in other contexts we may have to do the opposite. This begs the question of how can we design and manage the perception of a mind in a machine by users?

To better understand the issue, we draw here a comparison with a concept created by Disney's animators called *the illusion of life* [21]. Starting in the 1930s, Disney's animators created a set of 12 animation rules which described techniques to make animated characters look like they are alive. Following ideas also used by puppet masters, the rules employ mechanisms which are not commonly seen in real live animals and humans, such as the notorius *squash and stretch* rule (fig. 4). It prescribes that, in order to see life-like movement, living things must be drawn stretched in the beginning of a movement, and squashed at the end.

Is there a similar set of rules, or guidelines, governing the perception of conversing with a machine which has a mind? When should those guidelines be used? What are the pri-

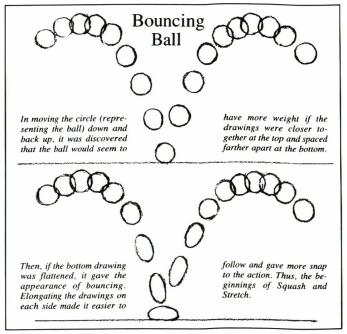

Figure 4: *Squash and stretch*, a key technique for creating the *illusion of life* in animated movies, as described by Disney animators [21].

vacy and trust implications of creating conversational systems which are perceived as mindful? Is it ethical to artificially boost the perception of a mind in a machine?

Notice tha the issue of creating an illusion of a mind is especially important for conversational systems because being able to converse in natural language, and to use it in witty or funny ways, can be easily be mistakenly perceived as an indication of a machine having a mind of their own.

Finally, it is important to consider that creating an *illusion of a mind* is an independent issue from creating a machine which actually has a mind, whatever that is. Puppets and animated characters have been very successful in making us suspending our disbelief that they are not alive, and making people engage with them intellectually and emotionally as if they were. For instance, there is considerable work on expressing emotions in *social robots* [2], which is, in many cases, independent of the existence or use of emotions in the robots internal mechanisms.

Final Remarks

It can be argued in many ways that we are, in conversational systems, in a state similar to the Internet in the end of the 1990s. That is, we have seen the first successful deployments of chatbots and voice-based Q&A systems, but like the first web-pages, we are still trying to understand what can actually be done with them, and how to tailor the interface appropriately. It is difficult, at this moment, to determine which services, interface designs, and business paradigms will be the most successful in the next generation of conversational systems. At the same time, the Internet history says that those successes will happen often guided by the best design methods and interfaces.

We tried in this paper to foresee some of the upcoming needs and translate them into specific research challenges to be addressed by the HCI community. We argued that the following are major themes for research to be pursued by the conversational user interface researchers: conversing in human-machine pidgins; managing language precision; designing and conveying personhood; knowing when to speak; and creating the illusion of a mind.

Far from being a definitive list, we want it to foster the debate of emerging key themes in the HCI community, particularly among those who are interested in the rich intersection between human-computer interfaces and natural language systems.

REFERENCES

[1] Scott Brave and Cliff Nass. 2007. Emotion in human-computer interaction. In *The human-computer interaction handbook*. CRC Press, 103–118.

[2] Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International journal of human-computer studies* 59, 1-2 (2003), 119–155.

[3] Heloisa Candello, Claudio Pinhanez, Mauro Pichiliani, Paulo Cavalin, Flavio Figueiredo, Marisa Vasconcelos, and Haylla Do Carmo. 2019. The effect of audiences on the user experience with conversational interfaces in physical spaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[5] Maira Gatti de Bayser, Paulo Cavalin, Renan Souza, Alan Braz, Heloisa Candello, Claudio Pinhanez, and Jean-Pierre Briot. 2017. A hybrid architecture for

multi-party conversational systems. *arXiv preprint arXiv:1705.01214* (2017).

[6] Daniel C Dennett. 1978. *Brainstorms: Philosophical essays on mind and psychology*. MIT press.

[7] Rahul R Divekar, Jeffrey O Kephart, Xiangyang Mou, Lisha Chen, and Hui Su. 2019. You Talkin' to Me? A Practical Attention-Aware Embodied Agent. In *IFIP Conference on Human-Computer Interaction*. Springer, 760–780.

[8] John Haugeland. 1981. *Mind Design*. MIT Press.

[9] John Haugeland. 1997. *Mind design II: philosophy, psychology, artificial intelligence*. MIT press.

[10] Dell H Hymes. 1971. *Pidginization and creolization of languages*. CUP Archive.

[11] Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Raefer Gabriel, Ashwin Ram, and Rohit Prasad. 2018. Alexa PrizeâĂŤState of the Art in Conversational AI. *AI Magazine* 39, 3 (2018), 40–55.

[12] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5286–5297.

[13] Alexander Maedche. Gender Bias in Chatbot Design. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers*. Springer, 79.

[14] Clifford Nass and Scott Brave. 2005. Wired for speech. *How voice activates and advances the human-computer relationship. Cambridge* (2005).

[15] C Nass and S Najmi. 2003. Race vs. culture in computer-based agents and users: Implications for internationalizing websites. (2003).

[16] Daniel E O'Leary. 2019. GOOGLE'S Duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management* 26, 1 (2019), 46–53.

[17] Claudio S Pinhanez. 2017. Design methods for personified interfaces. In *Proceedings of the 1st International Conference on Computer-Human Interaction Research and Applications. INSTICC, Funchal*.

[18] Claudio S Pinhanez, Heloisa Candello, Mauro C Pichiliani, Marisa Vasconcelos, Melina Guerra, Maíra G de Bayser, and Paulo Cavalin. 2018. Different but Equal: Comparing User Collaboration with Digital Personal Assistants vs. Teams of Expert Agents. *arXiv preprint arXiv:1808.08157* (2018).

[19] Emanuel A Schegloff. 1996. Turn organization: One intersection of grammar and interaction. *Studies in interactional sociolinguistics* 13 (1996), 52–133.

[20] Jetze Schuurmans and Flavius Frasincar. 2019. Intent Classification for Dialogue Utterances. *IEEE Intelligent Systems* (2019).

[21] Frank Thomas and Ollie Johnston. 1995. *The illusion of life: Disney animation*. Hyperion New York.