Moving Parts surrounding Conversational UX.

Henriette Cramer
Spotify
SF, CA & Boston, MA, USA.
Henriette@spotify.com
henriettecramer.com

Jennifer Thom
Spotify
Boston, MA USA.
jennthom@spotify.com
jennthom.com.

Position paper Conversational UX Workshop at CHI'17

Abstract
Building conversational interfaces requires more than design of the conversation alone. Services have to be consistent, but also adjust to the different conversational platforms or voice devices and their assistants. Going beyond a scripted conversation requires an infrastructure that supports learning 'how people talk', and collection of training data. Understanding people's language also means understanding the off-topic parts of conversations – and preparing for them. We here outline three types of challenges to illustrate the ecosystem surrounding conversational user experiences.

Author Keywords
Conversational UX, bots, human side of machine learning, ecosystems

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
Building successful conversational interactions requires anticipating how people will talk to your service. This involves understanding people's language, integrating into different chat platforms or voice devices, and understanding how people's social behavior will go

beyond the functionality that you have to offer. While classic HCI literature on conversational interfaces may have existed for quite a while [e.g. 6] and strides are being made in better understanding the personality aspects of living with voice assistants [4,5] many open questions remain on how to develop and design *around* the conversational functionality and personality itself. We outline 3 specific challenges that deserve further attention.

Understanding how people talk...

…with your service

Services that build conversational interfaces have to be consistent across platforms, but also have to consider the specific access point; whether mobile text chat, voice devices, or access through another assistant. When working well, voice interactions can for example take away the complexity of limited screen space. That does however require understanding how people would talk to your service. What words would they use, what intents do you need to fulfill? Designing a more limited conversation with limited, but focused, functionality is often a more successful endeavor as expectations are clear. Just as mobile platforms' app stores are the gateway to different companies' apps, voice devices and their voice assistants are the gateway to different skills. Assistants such as Alexa are integrated into different devices (Echo, tap, but now also fridges and cars), services and applications are integrating with different assistants. Similar to ways that a mobile OS affects the aesthetics and interaction models of its apps, the self-presentation of a conversational agent is limited by the chat platform on which it's built.

…with your domain

If you're developing a more complex Machine Learning-based agent, you will have to support its learning. Training data to support Machine Learning teams can be collected in different ways. The ways this data is collected, by whom, and from whom affects the types of conversations that the agent will be able to hold and the types of language it understands. Specific domains come with different needs. Different subcultures may use different language to describe similar intents or content [2], and language evolves over time. Your agent will have to keep up; yesterday's hashtag may be obscure by tomorrow. This means that ontology or folksonomy matching may be necessary, as well as constantly updating knowledge bases so the agent keeps up to date with current trends and terms.

Organizational needs & the end-users inside.

Building conversational interfaces can require a complex conversation between conversational designers, editorial and brand teams, together with the engineering, Machine Learning teams that support the natural language processing. For more advanced conversational interfaces, it's very likely that dedicated training data needs to be collected and labeled. The editorial teams [2] that monitor consumer, system and user trends require advanced tools to do their work well. These tools need to support and value the specific expertise of those who label, categorize and monitor data used in machine learning. In addition, these supporting systems should also work towards transparency to highlight the nature of the data at hand, so that tradeoffs and biases are apparent to those who build machine learning systems.

Conversations beyond designed intent

A conversation isn't simply focused on to-the-point functionality. Robots and agents invoke a certain amount of play or testing [1]. It's not necessarily obvious what a conversational agent can do – functionality is not shown on-screen and has to explicitly or implicitly explained. This invokes even more exploration. This means that designing a conversational UX doesn't only require making functionality discoverable, it also needs to consider how people will playfully test an agent.

For example, we ran a survey via Mechanical Turk and Crowdflower asking 96 users of voice assistants (Siri, Cortana, OK Google or Alexa) whether they ever tried to play with their voice assistant, tried to get it to do something they thought it shouldn't or tried to stump it. This delivered a number of examples of playful, somewhat norm-breaking behavior. This behavior varied on a couple of characteristics. For example, it included **inquisitive limit-testing,** such as "I've tried to get it to give me directions to odd places like the moon.", "I tried to get it to find Neverland and Peter Pan", and "Gave her convoluted directions to see if she could figure them out[…]" We contrast these responses with more straightforward **testing performance,** "The only thing I have tried was to speak a little faster or a little mumbled to see if they could still figure out what I was trying to say. In most cases they still were able to hear me.", or "I have tried to do a super long math problem but she always got it right.[…]"

From these types of interactions, we propose the following speculative design guideline: *expect your system to be pushed by your users, anticipate challenges, and also fail gracefully*. A conversational system should be smart enough to discern the difference between something it can't, or won't, handle or offer the opportunity for a user to try again.

More social-oriented testing however at times involved **social norm breaking** such as getting the machine to *swear at you or insult you***:** "[…] get it to curse and call me all kinds of insulting names. I was very successful getting it to call me names", *or directing insults to the machine:* "if it misunderstands me I'll mildly insult it to see what kind of response it comes up with. It usually just doesn't understand." Other social norms involve **relationship, physical relations and love** where respondents want to know how an agent would or could express these terms, such as "[…] I've asked her if she was dating anyone and she said that she dates but hasn't had a boyfriend in years." It could also involve, or potentially affect, **other humans** such as in "I asked Siri about a fight I had with my girlfriend and what I should do […]".

We propose another speculative design guideline as follows: *how the system responds to these deliberate provocations can encourage or discourage these interactions*. If the designed response of the conversational system ignores the content of these norm-breaking messages, how will users react? If the response of the conversational agent is slightly scolding, the persona of your system will change. However, reacting in a 'too accepting' or 'too playful' manner may send the implicit message that certain norm-breaking behaviors are ok, and encourage them. This can be especially problematic considering the gendered nature of these interactions.

In specific cases, this behavior highlights some potential security risks as well. Such as for example trying to trick **other devices**, " […] to get two SIRIs to talk to each other […]", and can (accidentally) involve **other services**. It's unclear whether we should see this as just banter, manifesting engagement and bonding. We for example could also see some of these as playful variants of social engineering attacks, especially when for example trying to get a device to insult friends, or when creating endless loops between devices. A playful-but-slightly-offensive conversation may be entertaining, but may not necessarily be on-brand or without consequence. However, that play will have to be anticipated – and can be both a challenge and a good opportunity for engaging interactions. It does however raise questions on what behavior a design may implicitly condone, and what the consequences are of that signal on users and their interactions with others.

In summary, we identify opportunities and challenges that we have observed in our work in designing conversational interfaces. These challenges are structural, as we need to understand the interactions between language, platform and context, and they are also social and performative, as users engage and test the boundaries of conversational interfaces.

About us.

Henriette Cramer: I'm a sr. research lead at Spotify, where I focus on the dialogue between people, data and machines; Spotify voice integrations and the human side of machine learning. Before joining Spotify, I was part of Yahoo's user engagement & metrics team where I worked on mobile and search-related projects, which included the Yahoo conversational bot platform. Prior, I led projects on human-robot interaction and location-based services at the Mobile Life Center in Stockholm, Sweden. My original academic background is in people's responses to adaptive and autonomous systems.

Jennifer Thom: I am a senior researcher at Spotify also focusing on the human side of machine learning and in particular, the social and collaborative aspects of the work conducted by those who label and collect the data that underlie these systems. Previously, I was a research scientist at Amazon where I used various crowdsourcing techniques to provide data to improve the machine learning models that power the Alexa assistant and investigated informal question-answer behavior as a research scientist at IBM Research.

References

1. Christoph Bartneck, Michel van der Hoek, Omar Mubin, and Abdullah Al Mahmud. 2007. "Daisy, Daisy, give me your answer do!": switching off a robot, HRI'07.
2. Erving Goffman, The Presentation of Self in Everyday Life, Doubleday, 1956.
3. Vindu Goel, When Yahoo Ruled the Valley: Stories of the Original 'Surfers', July 16, 2016, *New York Times*,http://www.nytimes.com/2016/07/17/technology/when-yahoo-ruled-the-valley-stories-of-the-original-surfers.html.
4. Sarah Mennicken, Oliver Zihler, Frida Juldaschewa, Veronika Molnar, David Aggeler, and Elaine May Huang. 2016. "It's like living with a friendly stranger": perceptions of personality traits in a smart home. UbiComp'16.
5. Ewa Luger, Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation

and Experience of Conversational Agents. CHI '16. New York, NY, USA, ACM, 5286–5297.

6. Clifford Nass, Scott Brave, Wired for Speech, *MIT Press*, 2005.