"Like Having a Really bad PA": The Gulf between User Expectation and Experience of Conversational Agents

Ewa Luger
Microsoft Research, UK
ewluge@microsoft.com

Abigail Sellen
Microsoft Research, UK
asellen@microsoft.com

ABSTRACT
The past four years have seen the rise of conversational agents (CAs) in everyday life. Apple, Microsoft, Amazon, Google and Facebook have all embedded proprietary CAs within their software and, increasingly, conversation is becoming a key mode of human-computer interaction. Whilst we have long been familiar with the notion of computers that speak, the investigative concern within HCI has been upon multimodality rather than dialogue alone, and there is no sense of how such interfaces are used in everyday life. This paper reports the findings of interviews with 14 users of CAs in an effort to understand the current interactional factors affecting everyday use. We find user expectations dramatically out of step with the operation of the systems, particularly in terms of known machine intelligence, system capability and goals. Using Norman's 'gulfs of execution and evaluation' [30] we consider the implications of these findings for the design of future systems.

Author Keywords
Conversational Agents; mental models; evaluation

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

INTRODUCTION
Framed as "dialogue systems often endowed with 'humanlike' behaviour" [43 p.357], conversational agents (CA) are becoming ever more common human-computer interfaces. The launch of Siri (Apple, 2011), Google Now (2012), Cortana (Microsoft, 2015), and Alexa (Amazon, 2015) indicate a spike in mainstream market commitment to this form of experience and, in a departure from their traditional services, even Facebook have thrown down the gauntlet by launching 'M'; a hybrid dialogue system that employs both artificial intelligence and human responses to task requests. Equally, such products are no longer solely tied to the handset. Both Siri and Cortana are now core components of their respective operating systems and Alexa finds its home in the form of Amazon Echo, giving us every reason to believe that spoken dialogue interfaces will become the future gateways to many key services.

Whilst the past 4 years have clearly seen a reinvigoration of such systems, this is very much a return to an old idea; that conversation is the next natural form of HCI. It has also long been argued that "when speech and language interfaces become more conversational, they will take their place along with direct manipulation in the interface" [6]. Moreover, they will have the potential to enhance both the system usability and user experience [43]. However, despite these expectations, the weight of research has veered away from such single modalities and tended towards multimodal developments, with a focus upon embodiment and anthropomorphism rather than voice alone. Indeed, our fascination with computers that converse can be traced back as far as 1964 when, seeking to create the illusion of human interaction, Joseph Weizenbaum of MIT created Eliza [10], a computer program that responded on the basis of data gleaned only from human respondents' typed input. Whilst script-based, it is considered the first convincing attempt to simulate natural human interactions between a user and a computer. This chatterbot, rudimentary by today's standards, was designed in the form of a Rogerian psychotherapist and, due to the high level of emotional involvement exhibited by users, was hailed as the beginnings of an automated form of psychotherapy [45]. Fast-forward 50 years and, whilst psychotherapy-bots for the time being remain the stuff of science fiction, HCI is again seeing moves towards serious adoption of naturalistic human-computer dialogue systems.

However, despite tech giants vying to develop the most compelling experience, the field of HCI has developed little empirical knowledge of how such agents are used in everyday settings. Whilst CA research exists, it tends towards either technical papers related to architecture [37], CAs studied in experimental settings, or systems created for specific contexts, such as guiding users around a space [24], delivering information [41], or for the support of language learning [40]. Whilst each study brings us closer to understanding effective design, without concurrent knowledge of the pragmatics of everyday use, we fail to truly understand dynamics such as how and why such systems are used and "which factors influence acceptance and success in such scenarios" [24 p.329]. In light of this deficit, our paper seeks to understand user experience of CA systems by answering two simple questions; (a) what factors currently motivate and limit the

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858288

ongoing use of CAs in everyday life, and (b) what should we consider in future design iterations? To this end we present the findings of 14 semi-structured interviews conducted with existing users of CA systems and, on the basis of our analysis, present four key areas where current systems fail to support effective user interaction.

PRIOR WORK

When considering human-machine dialogue from an HCI perspective, some of the earliest work can be traced back to vision statements and position papers such as JCR Licklider (1960), 'Man-machine symbiosis' [25] and the early work of Richard Bolt [4]. Whilst this work was visionary in nature, more recent studies have focussed around the development of CA systems for specific contexts or settings [24] with a preference towards multimodality, or embodiment, in order to more effectively simulate the nuances of human communication.

What is a conversational agent?

The notion of a humanlike virtual character has been framed and realised in many different ways. When using the term Conversational Agent, for example, one might think of a chatbot, virtual companion, "interface agent, embodied conversational agent, virtual assistant, autonomous agent, [or] avatar…often synonymously" [44 p.1641] and use these labels interchangeably, or so as to articulate particular subtle distinctions. In order to distinguish such terms more clearly, Wilks (2010) suggests a series of differentiating characteristics. From this perspective, CAs are distinct by their function; to carry out tasks. In contrast, "chatbots" have no memory or knowledge but instead mimic conversation; for example, Richard Wallace's Alice technology [1] or Microsoft's Xiaoice [13]. Digital companions on the other hand are not designed for any central or overriding tasks, have *long term* discourses and personal knowledge of the user, and are able to lay the foundations of a 'relationship' [47]. Although the emerging class of CAs may indeed have aspirations to become a user's artificial companion, their current functionality places them some way from realising this. In this paper we use the term 'Conversational Agent' to label the emergent form of dialogue system that is becoming increasingly embedded in personal technologies and devices. At a minimum, such spoken dialogue systems require "an automatic speech recognizer to perform speech to text conversion, some form of dialogue manager (controller) to control the interaction with the user, and a mechanism for conveying information to the user (e.g. text and/or speech generation" [15; 1]. These systems, whilst often accompanied by some form of graphic such as the Cortana ring or the wavy line that appears when Siri is activated, are not strictly speaking embodied. Neither is anthropomorphism their principal goal, nor is it to represent a specific person and as such they cannot be considered avatars. More accurately, these types of systems draw directly from the notion of the 'virtual/digital assistant', or 'virtual/digital butler' [32] in that their purpose is both support for real time task completion and to develop sufficient knowledge about the user in order to exert agency on their behalf. In this way, whilst they are not a 'companion', they might seek to exhibit a level of the associated characteristics in order to (a) better perform their function, (b) present a more compelling experience, or (b) mimic human-human relationships and build user trust [22]; setting behavioural realism as a principal goal.

Embodiment, multimodality and humanlike behaviour

Reflective of this, the weight of work in this area has tended towards embodiment and multimodality. Von der Pütten *et al* (2010) argue that the social influence of an autonomous agent relates strongly to the levels of behavioural realism it exhibits. From this perspective, an embodied conversation agent (ECA) will elicit significantly more sympathetic social behaviour than those agents without physical form, and even less than those exhibiting a high level of anthropomorphism, thereby framing embodiment as *the* means to natural interaction. This desire, to create the illusion of human-human dialogue, is catalogued through a range of experimental studies. Most prominent is the work of linguist/psychologist Justine Cassell (2000) who created Rea, an embodied CA that incorporated non-verbal behaviour to show that such modalities were key to sustaining convincing experiences. Unlike the emerging class of CAs, embodied agents take their own physical form with the intent of eliciting more natural human-computer communication. Embodied conversational agents are "specifically conversational in their behaviors, and specifically humanlike in the way they use their bodies in conversation" [9 p.29], though the requirements of such agents are arguably partly transferable to the non-embodied CA. Such requirements include (a) recognising/responding to verbal/non-verbal input (b) generating verbal and non-verbal output, (c) dealing with functions of conversation such as turn-taking, giving feedback, and repair mechanisms, (d) giving "signals that indicate the state of the conversation, as well as contribut[ing] to new propositions to the discourse" [10 p.29]. One example within this vein is Max, a museum guide that presents as an avatar on screen to guide visitors and answer questions [24]. Like many embodied CAs, Max is multimodal in that it employs manual gestures, gaze, facial expressions, locomotion and capability for small talk. Similar developments have been seen within a range of spheres such as gaming [19], mixed reality environments [14], tourism [29], train timetabling [41] and agents to support learning [23; 14]. Whilst it is clear that the anthropomorphic value of embodiment has driven much recent work in this field, focus upon a single modality has faded and subsequently the majority of HCI work around dialogue systems is somewhat dated.

The value of dialogue as interaction

It has been argued that the true value of dialogue interface systems over direct manipulation (GUI) can be found where task complexity is greatest [6]. Specifically, when used (a) to filter/browse or issue commands over large amounts of information/sets of things (b) to follow navigational paths (c) to allow negation and quantification (d) to distinguish individuals from kinds (e) to filter and request information in

ways not predetermined by design, (f) to build a complex query that takes more than a single turn, (g) for referring to things that you cannot see or point to, or doing anything that's not in the here and now, and (h) for delegating complex or redundant actions. However, the realities of spoken language processing are such that "success is blighted by recognition errors, unintelligible responses and hard to navigate dialogues" [28 p.123], leaving us somewhere in the void between experience and potential. According to Moore, this gulf will remain unbridged until dialogue systems better understand user drivers and motivations, and are able link these meaningfully to their communicative behaviour. In this regard, the notion of a conversational agent might be considered something of a misleading concept. In line with Moore's assertions, the act of conversing with another person implicates a raft of transactions beyond the mere vocalising of words, more akin to what Harper (2010) describes as 'structural patternings', which are dictated by judgments made in response to certain social cues. He gives the following example - the opening 'hello' is actually preceded by a glance that allows the speaker to ascertain context, judge the mood of the recipient, and assess the correct tone to take in order to know, for example, whether one is interrupting. Framed in this way, even a simple conversational act becomes highly complex, inferentially fertile, and can be said to perform multiple functions. So, one might ask, is the principle design goal to replicate human-to-human interactions?

Conversing with Computers

In 1996, Reeves & Nass claimed that individuals respond to computer-based agents in the same way that they do to other humans during social interactions. Shechtman & Horowitz (2003) however found that in fact this was not the case. The authors highlighted three kinds of conversation goals: 1) *task goals*, in which conversation is used to achieve a joint activity, or construct a plan together; 2) *communication goals*, in which the goal is to ensure the conversation itself runs smoothly, and 3) *relationship goals*, in which people are driven to achieve a certain tone of conversation and maintain certain kinds of relationships (whether friendly, hostile, professional, intimate, etc). In order to assess interaction, subjects were blinded as to whether they were interacting with a human or computer. It was found that subjects put more effort into their conversations when they thought their partner was human, used more statements relating to their relationship, were more engaged in the conversation, and reacted more assertively to assertive responses from their partner. Wilkes confirms this view by suggesting that, whilst people expect politeness in human interactions, they are actually 'repelled' by excessive politeness and repetitions when they know the interaction is with a machine [48].

Whilst research has shown that people talk differently to humans and computers, there are still common requirements. Both cases require a degree of context dependence in the conversation, and users expect connectedness across the whole sequence of conversational turns rather than a response only to a single turn [6]. In addition, CAs need to exhibit 'social smarts' that enable engagement of the human in "an interesting and relevant conversation" [9 p.30], which should be both automatic and context appropriate. Prior work has shown that designing for human-computer dialogue is a complex space. However, it is also clear that if we are seeking to make CAs successful applications, we first "need to make them capable of interacting with naïve, uninformed humans in everyday situations" [24 p.329]. Having given an overview of cognate work, the following section outlines the methodology and results of our study.

METHODOLOGY

Within this paper we seek to understand the high-level factors that motivate and limit ongoing use of CAs in everyday life. To this end we conducted a series of semi-structured interviews [7] with 14 individuals who considered themselves 'regular' users of a CA. Fourteen users were interviewed; 9 male and 5 female, aged between 25 and 60. Interview durations ranged from 19:32 to 48:50. All interviews were conducted at a distance (telephone/Skype) between November 2014 and February 2015. Participants were recruited through an open call via email and social media, and were selected on the basis of having 'regularly' used a CA for at least one month. They included; a health telecare advisor, two consultants, a magazine editor, a data analyst, a lecturer, three managers, two researchers, a strategic advisor (educational policy), a homemaker and one PhD student. Twelve participants identified as British, one as American and one as a Czech national (see Fig.1).

Name	CA type	Age	Gender	Role
Adam	Siri	40-49	M	Lecturer
Viv	Siri	20-29	F	Researcher
Dan	G Now	30-39	M	Consultant
Allan	Siri	30-39	M	PhD student
Sam	Siri	50+	M	Strategic advisor
Richard	Siri	50+	M	Project manager
Graham	Siri/G Now	40-49	M	Consultant
Sarah	Siri	30-39	F	Magazine editor
Mike	Siri/ G Now	20-29	M	Researcher
Andy	Siri	40-49	M	Telecare advisor
Rob	G Now	30-39	M	Data analyst
Emily	G Now	30-39	F	Regional manager
Laura	Cortana	30-39	F	Homemaker
Denise	Siri	40-49	F	Account manager

Figure 1: Participants by age/gender/CA type/occupation

Definitions of CA 'use' were not prescribed, in order to allow a range of perspectives, and individuals self-identified as 'regular' users. A CA was described as a 'task-driven system or application that you interact with through speech' and

examples were given. Respondents were predominately male and principally Siri users.

Transcripts were coded using a Thematic Network Analysis (TNA) approach in order to align with the interpretive epistemology of the wider research. The thematic network is a 'web-like coding structure' [2; 385] that allows for a rich understanding of the conceptual interplay occurring within the narrative. TNA organizes coded text into three types of theme; (i) basic (lowest order, coded statements or beliefs that related to organizing themes), (ii) organizing (that cluster basic themes into organizing issues) and (iii) global themes (super-ordinate themes that organize all codes into meta-groups or metaphor) [2]. Interview questions sought to elicit participants' technical knowledge, the broad frequency and duration of CA use, type of use, most/least complex tasks, use location, manner of use and emotions elicited by the experience, personal preferences, and perceived benefits/limitations in respect of their specific CA product. Participants were also prompted to describe at least one example in detail; 'describe step by step the way in which you [user-defined example]'. Transcripts were coded on an ongoing basis and new participants were sought until theoretical saturation was reached [17]. Themes were derived of the data, and drew from both the frequency of their occurrence and their perceived substantive significance. Having described the approach to recruitment, data collection, and analysis, the following section frames the findings in terms of stages of CA use.

FINDINGS

Participants were all daily users of technology (predominately smartphones, laptops and tablets) with mobile devices cited as those most frequently used. All but one owned a smartphone through which they accessed their CA; Mike was the only exception to this as he accessed Siri solely via his iPad. Only two participants made use of more than one CA; Mike and Graham, who used both Google Now and Siri. Overall, Siri was the most frequently used CA (10) followed by Google Now (5). Only one user made use of Cortana as the study was conducted prior to the UK launch date and our Cortana user had set her phone to United States settings specifically in order to access it.

Motivations and Type of Use

Whilst a third of users employed their CA only a few times a week, the majority made use of it on a daily basis, with only one user reporting the CA (Siri) as their preferred task interface. When asked about the most frequent type of CA use, the tasks reported were relatively simple; checking upcoming weather followed by reminders. When asked to describe these interactions in more detail, users reported using natural/colloquial language in the first instance (e.g. 'should I take an umbrella/my coat today'), but then simplifying language until the desired result was achieved. When describing their rationale for use, participants spoke of broadly similar motivations. Principal amongst these was the desire for the CA to enable multitasking, particularly where their hands were otherwise engaged. For example: Adam regularly cycled to work and made most use of the CA in that context; Sarah worked from home whilst looking after two small children, and wanted to be able to carry out multiple tasks; and, Rob spent a considerable amount of time driving to work and liked to use his CA for productivity in what would otherwise be dead time. Indeed, time-saving was a key related motivation *"I'm also constantly asking Siri to set appointments, reminders, alerts and alarms because on the iPhone each of those takes, you know, 4 to 7 steps, not including typing. So Siri is good for those things that would otherwise cause me to go to the keyboard"* (Graham).

Learning to Use the CA

When learning to use their CA, all participants described making use of a particular economy of language. Dropping words other than 'keywords', removing colloquial or complex words, reducing the number of words used, using more specific terms, altering enunciation, speaking more slowly/clearly and changing accent were the most commonly described tactics: *"I try to enunciate as much as I can and keep my requests to relatively simple words…..you know possibly in a very exaggerated, trying to speak in a very clear fashion….rinse and repeat as necessary until it works out what I mean"* (Rob). For over half of participants this was seen as affecting ease of use *"I think it would be easier…if you didn't have to enunciate everything or feel like you have to*" (Denise). In this way they reported 'learning' to speak to their CA as a process necessary for successful interaction. *"To a certain extent you have to change the language you use because sometimes it doesn't pick up certain words or it doesn't understand the inflection in your voice obviously…you have to say things in a particular way"* (Laura).

Several users noted that they had attempted to speak to their CA as though it were a person, though the majority were unlikely to engage in conversational-style interactions when in public. One user noted that he was more likely to speak to Siri colloquially when in private – indeed, there was a desire by several participants to be able to carry out more natural conversational interactions: *"Often, if I'm in private, I talk to it a bit more like it's a person. So, I'll say 'Siri, can you tell me what the weather's like today', whereas if I was on the street I'd just say 'directions from St Pancras to Waterloo station'"* (Allan). Despite attempting the use of more colloquial phrases, there was no sense that users spoke to the CA in the same way as one might to a human. Each interaction was uniquely framed and in the majority of cases participants commented on the lack of ability of the CA to bring to bear contextual understanding between interactions, even when they were temporally close. This was seen as limiting the tasks one might ask: *"I don't ask for more information from it. It tends not to be very good at that. You ask it to do one thing and you know that that's the one thing it knows how to do. Asking it to do sub-tasks, to follow up or to give you more information about something you've just asked it, it tends to be really bad at."* (Allan)

Effective use requires ongoing 'work' and investment

For all except three participants, Conversational Agents (CA) were considered an entertaining/gimmicky addition to their device rather than a key application; Graham and Andy were the only participants who defined themselves as 'serious' users. Both such users described a strong desire for their CA to function efficiently and were each prepared to do additional 'work' to ensure the most successful operation. Such work included researching the capability if the CA, introducing use of the CA within social/professional settings, considering and testing best strategies of use, training it to recognize key contextual variables such as locations/people, systematically testing appropriate speech syntax in order to 'speak its language' and create a dynamic mental model of its capabilities, and making time to test these capabilities: *"I found out through poking online out of curiosity about what Siri can do…I played around with it just to see how it worked… So, I tried a very simple test. I told Siri who my father and mother were and then I said, both of those entries have addresses and so, you know, after I drive from my mother's to my father's my mother wants me to let her know I arrived alive, and sometimes I forget, you know, because I'm an adult man and I forget to call my mother when I get to someplace after a perfectly normal drive. So I told Siri 'remind me, call mom when I get to my father's house'…so there was a pre-set up just to get to a point where I could ask something very basic."* (Graham)

This form of preparatory work resulted in the user being able to more easily ask complex tasks than those participants who used the CA less frequently. Such ongoing work also served to build user 'trust' in the system; a word frequently used. Andy, in particular, was prepared to invest time in discovery and set up because he wanted to know both whether the technology would be useful for his clients, and the extent to which it would be helpful for his own day-to-day life; *"I suppose ever since I got it I would probe it as much as I could so I suppose I am more serious about using it but I was serious I suppose from the beginning"* (Andy). Whilst a relatively recent (1 month) and less frequent user, Richard also sought to discover more features of his CA, though his motivation was stimulated by seeking common ground with his son, who also used Siri. Whilst the most effort was brought to bear by those who had a clear use case, all users engaged in some level of 'work' to ensure successful interaction (see 'learning to talk to computers'). As one might expect, the level of work/investment of time appeared to have a direct relationship with user satisfaction.

Those who used their CA frequently (at least daily) also found themselves becoming more successful in their use. This was reportedly due to (a) their growing accustomed to the types of commands the CA would respond to, and (b) the greater likelihood that a successful CA interaction would motivate them to search for other things their CA could do. "*I think by playing with it and understanding what it could do well and what it could do badly….through that I found that I developed a series of things that I used it for that were a quite discrete number of its functions"* (Allan). In this way, users' satisfaction with, and trust in, the product had a strong relationship to the extent to which they were prepared to invest time in both understanding what their CA could do, and practicing those interactions. Equally, the more personally compelling the described use proposition (primarily professional need), the more likely users were to frame task failure as forgivable.

'Play' as a point of entry

All users except one began their engagement through playful interactions such as finding 'Easter eggs' in the system; the inbuilt humorous responses triggered by specific phrases. This, they described, as allowing them to both better understand what Siri could do and enabling their familiarity with the interface. *"I started out in a playful thing simply because…you just don't know what you can ask it… and it was like 'can you do this'?"* (Dan). The only participant who did not start with play (Andy) approached Siri as an instrument to be used with his clients - people with physical and mental health issues. In this case, he was professionally motivated to find practical uses for Siri, which subsequently transferred to his day-to-day life. This user was also far more forgiving of errors or points of failure as his desire for the promise of a fully functioning CA raised his threshold of acceptable failure. *"It wasn't just an added feature on the phone…. I got (iPhone) especially for it and when I got it I was so excited that I could find the use for my clients"* (Andy). In the case of families with small children (3 participants), Siri and Cortana were a particular source of entertainment. *"Cortana has this function where she can sing to you… she does…if you ask her to sing you a song or tell you a joke, she'll sing to you or tell you a joke and we use that quite regularly…. I have numerous children so they play with my phone quite a lot"* (Laura). However, once the amusement had passed, levels of engagement dipped and the framing of Siri as a fallible but amusing tool pervaded much of the narrative: *"We played games with the kids when we first got it…we did ask it 'what the fox said'… and it said frakakakakaka, which is one of the lyrics from the song. Which made us all laugh. So that's my happiest story of using Siri so far"* (Adam). In most cases, these playful interactions resulted in longer-term use. This was particularly true where users felt they had time to invest in the product, or where the CA was seen as 'easier' or 'faster': *"My feeling originally with Siri was that it was a toy….you'd ask it to do stupid stuff and then you start to do certain things with it and it starts to work, you know, like putting stuff in your calendar, and then it just becomes like an easier way of doing things"* (Mike).

However, beyond the initial framing of 'playful' experiences, users became less forgiving of failure. As early successes resulted in sustained use, so early failures affected the frequency and type of on-going use, particularly where users expected the system to respond similarly to the way in which it had during play: *"I don't think maybe I was speaking clearly enough to it so it wasn't really getting it…she nailed the jokes though so I guess it was, you know, surprising"*

(Denise). When Siri failed to perform a key task more than a few times (reportedly between 2 and 6 in the majority of cases), user expectations were set. *"I gave it the benefit of the doubt…and then I thought no, you're always going to be rubbish"* (Sarah). Having failed at more complex tasks, the CA was often relegated to performing very basic tasks such as setting reminders. When asked to describe the factors that had affected use, participants most frequently cited misunderstanding of words or commands; *"I'd heard it had had difficulty recognising accents and I guess mine was amongst [those]"* (Emily). This was particularly the experience of the female users as all except one woman found the CA unable to dependably recognise their voice. Male users found this less of a consistent problem, though it was still the most commonly reported reason that CAs were perceived to fail. However, where users were aware of software updates or improvements, they reported subsequent improvements in voice recognition, though they were often uncertain as to the extent this was imagined. The two associated global themes, of a lack of awareness regarding the operation of the system, and the limitations of its 'intelligence', were recurrent across all participants.

'Hands free' as the principle use case

As one might expect, the principle use-case for the CA was 'hands free', which was tied strongly to the theme of time-saving and convenience. *"So I was walking through London and it was just more convenient to ask Siri how to get between St Pancras and Waterloo then it would be to stand in the street and type it in…it's quicker"* (Allan). Reasons for hands-free use were cited as when (a) hands were necessarily otherwise engaged, (b) hands were dirty (c) the handset couldn't be easily reached, (d) speech was felt to be faster, or (b) when attention was distributed, particularly during another primary activity: "*So, my main use case is when my hands are otherwise occupied, which is when I'm cycling into work typically"* (Adam). Time saving was also a global theme. *"In terms of economy of key presses, it's just an awful lot quicker"* (Richard). Indeed, where a desire for hands-free was not the motivating rationale, *perceived* time savings drove the preferred mode of interaction. However in such cases, when users felt they had not saved time, they fell back upon using the touchscreen, a practice they described as 'traditional', 'normal' or 'old-school'. In instances where the CA responded to task requests by defaulting to on-screen web-search results, this was universally seen as a failure.

Locus of attention and task complexity

Within the global theme of 'hands-free', the locus of user attention was a key organizing theme and use of the CA was rarely described as a primary activity. Participants made the CA their principal focus only when engaging it in playful interactions, or during preparatory or exploratory work/activities such as experimenting/teaching it/testing the limits of its capability. For occasional users, the CA was perceived to be most useful during activities that demanded high attentional resources such as cycling, driving, or looking after children. This desire by users, to employ their CA to support distribution of their cognition, was also reflected in the second most frequently cited task type, that of memory augmentation: "*I constantly have a really shitty memory…if I don't write it down I'll remember it a week later and it's like, oh that important thing I had to remember just came back to me…I travel a lot, I'm out and about a lot…so I pull out Siri and I just say 'set appointment tomorrow for 10am, call mom"* (Graham).

In line with our working definition, users viewed their CA principally as a simple task-based system and, other than Graham, they did not attempt complex tasks (particularly without visual confirmation) or tasks where they perceived a high social cost to failure. Simple tasks included asking the CA to check the weather, setting reminders, setting alarms, getting directions, making lists, adding notes, adding items to their calendar, searching the Internet, searching their address book, activating music playlists, or activating FaceTime. In contrast, more complex or socially sensitive tasks, such as launching a call or sending a long email, were mostly considered tasks with which the CA could not be trusted. In each case where use of the CA distracted the user from their primary task, or required manual intervention, it was seen as a failure, and recurrence was rarely forgiven: *"I've tried to use it when driving, I've tried to use it when I've got my hands full with the kids, I've tried to use it when I've kind of been engaged in other stuff and all that happens is that I've had to abandon what I was doing and sort of do it in the old-school way of actually typing in what I need and so I have by and large kind of given up on using it.* (Sarah). However, where users came to the CA with a clearly defined proposition for use, such as Graham or Andy, they were more willing to return to the application, consider ways in which they might improve the outcome, and set more complex tasks.

System Feedback and User Evaluation of CAs

With the exception of the more exogenous limitations of infrastructure or social context, the majority of reported issues during use related to a lack of system feedback/transparency. Half of users explicitly stated that they did not know what their CA could do. This resulted in them either feeling overwhelmed by the unknown potential, or led them to assume that the tasks they could accomplished were highly limited. *I felt let down that I didn't get any feedback from it...It has a captive audience it could have just told me…just to let you know, this is what we've done. Just a few examples of what could be done. The things it can do are so broad that I just feel lost"* (Dan). User understanding of the operation of the system was a key organizing theme within participant interviews. Insufficient feedback or visibility of both the limits and capabilities of the CA was often cited as a factor limiting users' ability to make the system work: *"I'd have put the [CA] down and never picked it up again, I know that. So there was quite a big barrier I think… I don't consciously know what I've done to change it"* (Richard). Specific factors limiting interaction included a lack of understanding of (a) what the system *could* do, (b) what it *was* doing (c) *how* it was doing it, (d) whether or not its capabilities altered over time,

and (e) the extent to which that alteration was a direct result of user interaction; i.e. whether or not the CA learned from user behaviour. Indeed, the extent to which a CA could be said to 'know' or 'learn' things was mentioned by all participants. Additionally, users who had previously defined themselves as not/less technically knowledgeable, tended to have higher initial expectations of CA capability and intelligence. Equally, when tasks failed, they were more likely to see the CA as static and *unable* to learn, resulting in them being less experimental in the tasks they asked it to perform. *"I don't use it as a functional tool as it were…it's more sort of oooh look at this, I can speak through my phone. Then when I use it…to look up recipes for sloe gin or something like that it simply directs me to a website. That's it. It seems to be a very blunt tool. It's literally just replacing my finger on a keyboard in a kind of clever smoke and mirrors kind of way"* (Sam).

In contrast, where users were more technically knowledgeable, their expectations were relatively modest and they tended to be more forgiving when the CA failed. *"I think I'm probably quite forgiving because I know how hard voice recognition is as a problem, and especially in the time frame that it's attempting to work on, just a few seconds, I know how difficult a problem that is. So it doesn't really annoy me if it doesn't work it out because I assume it's not going to because I know how difficult it is"* (Rob). Such users were also more likely to persevere, doing further research or using different types of commands or syntax, until the task was accomplished:*" I know it's got like certain words that work. You just say 'weather' and it will work, you don't have to ask it a question. You know that it's searching for keywords and it doesn't care what you say around it really"* (Viv). Equally, those with a reportedly higher level of technical knowledge were more likely to locate task failure with the technology, whereas those who identified as non-technical were more likely to see the failure as their own: *"I wasn't using the right words and it would confuse words with things that I was looking for"* (Richard). This assignation of blame often resulted in users reporting feelings of being 'stupid', 'slow', 'unaware' or lacking in 'technical savvy': *"And it's all my fault because I haven't put the numbers in under the right, you know, headings and it's all like well if you'd have done the right thing…if you'd have done the right you wouldn't have made that mistake and you know it's making me feel really stupid right now"* (Dan).

Assessing system intelligence

In terms of perceptions of CA intelligence, all except the most highly technical users (Graham, Rob, Adam and Allan) felt unsure as to whether their CA had a capacity to learn, and even for those who were highly technical, this was not always clear: *"I don't know whether it learns or not, that was the other thing, I'm not sure whether it learns or not"* (Richard). This resulted in the majority of users being unsure as to the interaction dynamic; is the computer learning to adapt to the user or visa versa. However, for those with lower levels of technical knowledge, a combination of (a) the system's failure to learn/adapt to either their accents or the ways in which their questions were posed, and (b) its tendency to resort to web search, led them to frame the CA as simply a voice-based search engine extension. Even for those with technical knowledge, the tendency for their CA to default to search was a frustration: *"I've asked it about the weather today, say, and then I want to ask it about tomorrow….and it will just come back and say 'do you want me to do a web search for what about tomorrow, not to give me tomorrow's weather"* (Allan).

Whilst all users described a process of learning to speak to their CA, for those with lower levels of technical knowledge, there continued to be mismatch between their initially high expectations, and the perceived intelligence of the system. However, even where users described the system as failing, they were at the same time able to ascribe to it a high level of contextual understanding, sarcasm, anger or situational humor. *"There was one time I was very [sarcastic)] to it, I was like 'oh thanks that's really helpful' and it just said, I swear, in an equally sarcastic tone 'that's fine it's my pleasure'"* (Sarah). Irrespective of whether they believed the system capable, a fluid understanding of context was something considered key to making interactions appear more naturalistic. Users expected the CA to be able to infer, from all previous interactions, the context of the current task. In particular, once an interaction/task was complete, the majority of users expected the CA to remember the context of the preceding interaction. Equally, more positive 'conversational' experiences were reported when the CA was perceived to have understood the context of use, for example knowing that reading a message would likely lead to the user wanting to reply; *"It's brilliant that it reads messages to you and gives you the option would you like to reply…automatically, without pressing the button again it goes into what would you like to reply and it's almost like you're having a conversation with your co-driver"* (Andy).

In the context of Siri, those who took a more instrumental view of the CA, or reported high levels of technical understanding, tended to describe less anthropomorphic behaviour and referred to Siri using gender-neutral pronouns. In contrast, those who reported less technical knowledge were found to use both gendered pronouns and greater anthropomorphism in their descriptions: "*He was like a bad boyfriend that was just never going to make the grade"* or *"like having a really bad PA"* (Sarah). In many such cases users reasoned about perceived limitations of the system in terms of the CA being 'just' a computer. Indeed, the term 'computer' was often thematically tied to pejorative language and was regularly set against human intellect; the latter framed as the superior form. Irrespective of user experience, the CAs were universally seen as a blunt instrument - limited in their capacity to learn. In all cases, even where improved performance was perceived, or failure was expected and forgiven, CAs were found to fail to live up to initial user expectations: "*So my expectations were always high, I adjusted them and the product improved, but …it's still not*

quite met with my original expectations or my adjusted expectations" (Emily).

Issues Affecting Engagement and Ongoing Use

Where the capability or operation of the CA was felt to be unknown, so the issue of trust emerged. Asides from the two most frequent users who tended to be more experimental and forgiving, all of those interviewed raised issues of trust as limiting the tasks they would ask their CA to perform. For example, after several attempts Allan had not succeeded in getting Siri to book cinema tickets and had subsequently abandoned asking for help with this task. As in the case of most users, if Siri failed after successive attempts he felt compelled to abandon that task and now uses it *"only where I know I'm going to get a reasonably reliable result…."After it's got it wrong a couple of times, oh I might as well just set it normally"* (Alan). This was particularly true where failure might result in social embarrassment. Within this theme, the most frequently-cited activity users would *not* use their CA for was to make a call on their behalf. *"I would never risk calling anyone because I know that that would not work for me and it'd end up being awkward because I'd end up calling someone I didn't want to call"* (Viv).

"*I tend to not use it for dialing and stuff like that because I'm worried that it's going to ring the wrong person….and because you're not paying attention it sort of rings that person on your contact list that you haven't spoken to for 7 or 8 years. It's like 'hello darling how are you'….what?"* (Allan). However, one participant (UK-based but whose family live in the US) noted that this may well relate to cultural context, his feeling being that people in the UK had a lower threshold of social embrassement than other cultures: *"Oooh God no, absolutely not…where I'm trying to get to here is the idea that just like ringing the wrong number, if it was a US thing, it's the opportunity to have a quick chat and that's fine. I just reached out and touched a friend. But in the UK, God no."* Irrespective of the activity, however, all users described seeking visual confirmation of complex tasks.

Precision and visual confirmation

In all cases, the issue of required precision was a factor linked to trust and reliability. Where precision was necessary, for example when drafting an email, users tended not to trust their CA to complete the task. Even amongst the two 'serious' users, 'send' would not occur without visual confirmation and intervention, and 'call' only where the respondent was well known. *"I'm not going to trust Siri where the language needs to be precise…but when it can be imprecise then I'll trust Siri"* (Graham). All users described this need, for visual confirmation that complex tasks had been carried out accurately. However for simpler tasks, such as setting an alarm, repeated successful completion was sufficient for them to trust spoken confirmation from the CA alone; the exception being when the alarm was important. Indeed where tasks did not require precision, such as setting reminders, alarms, or asking about the weather, visual confirmation was not always required. In contrast sending messages was considered a potentially more sensitive task, particularly when the audience was not family or friends. Here, despite the potential effort/time-saving of using a CA, users felt the need to return to the screen and check/amend errors before sending.*" I can sometimes make it read an incoming text, you know, a notification comes in and I can't see what it is because the phone's in my pocket, and I quite like the audio output. At the moment I certainly can't trust it to reply. You know, I can't rattle a reply off by speech and trust that it's got it right, and that whole interaction….of not being able to see the screen of what it's translating"* (Adam).

Google versus Siri

Whilst the purpose of this study was not to compare products, it was clear that there was a considerable difference in the ways that users approached use of Google Now compared to Siri or Cortana. Google Now was used predominantly as voice-activated search. Conversely, Siri was seen as having more of a personality, which was well received. Even in the case of 'serious' users or those who self-reported as technically skilled, there was recognition that it had been designed to give that impression. *"You know I tell Siri 'tell me a joke' and it has a few, you know…and my hope that Siri…I mean Siri is very much at the vanguard. I haven't asked Google like, you know, tell me a joke. Siri seems very much trying to be a personal thing whereas Google seems to me very much task-oriented"* (Graham). None of the users of Google Now referred to the product with either gendered pronouns or in any way that suggested humanlike characteristics, rather they framed it as a tool: *So with Google Now, because it's very tightly integrated into Google maps, you can ask it a place and it will show you where it is. Siri isn't as developed that way"* Mike).

DISCUSSION

In light of our findings it is clear that user expectations of CA systems remain far from the practical realities of use. This is illustrated by the level of effort users described when seeking to elicit the expected task response from their CA. Unlike human-human interactions, the nature of CA human-computer dialogue was, as one might suppose, limited. Our findings show that this is due to the manifest dissonance between user expectations and the pragmatics of CA communication, particularly regarding their assessment of system intelligence. This speaks to what Norman describes as the 'gulfs of execution and evaluation', understood as the degree to which the system representations can be perceived and decoded by the user into accurate expectations and intentions of use. The smaller the gulf, the more satisfying the user experience, and a small gulf is achieved "when the device provides information about its state in a form that is easy to get, is easy to interpret and matches the way the person thinks of the system" [30, p.39]. According to Cassell, there are three components that affect perceptions of intelligence in the design of a CA user interface; (a) how the system is represented through its interface, (b) how the system represents information (and the world) to the user, and (c) how the system's 'internal representation' impacts upon users' interactions with that system. Whilst the focus of Cassell's work has been

multimodal representations of intelligence (physical gestures in addition to voice), the central concept, of the users' need to 'locate intelligence', and thereby the need to represent intelligence to the user, is equally salient here.

Setting realistic expectations scaffold the learning process

For all our participants, expectations of how to interact with the CA, its capability and operation, were out of step with reality. In the majority of cases users were unable to make accurate judgments about system capability. This varied in accordance with the referential frame of the user; for example, where users had knowledge of computer science then the gulf of execution and to some extent evaluation were much smaller. Here, users had a more developed mental model of system capability and were less likely to abandon tasks. For those without technical knowledge, however, the model upon which they drew tended towards that of human-human dialogue, illustrated by (a) beginning their interactions with fuller and more natural sentences, which they then amended (b) a tendency towards anthropomorphism, and (c) ungrounded attribution of intelligence. During conversation-based human interactions, we use a variety of cues to communicate intelligence, what Cassell describes as 'social interactional intelligence' such as initiation, appropriately interrupting a conversation, feedback, error-handing and turn-taking. In contrast our users described each interaction with their CA as clearly bounded, which problematized their instinctive approach to task commands and led them to desire more colloquial dialogue, more akin to Harper's 'structured patternings' [21]. In the absence of expected cues, our users tended towards two overlapping approaches; (a) to endow the system with imagined anthropomorphic qualities, such as an advanced understanding of context, or (b) to employ an economy of interaction, such as avoiding complex tasks, limiting the types of language used, and gradual abandonment of the CA for activities other than those they 'trusted' it to perform. Each of these approaches were problematic in their own way and, where they overlapped, served to confuse the user experience.

(a) CA design should reveal system intelligence: Whilst such economic interaction resulted in a much more effective approach to human-agent conversation and task completion, it was ultimately mechanistic and shut down opportunities for developing meaningful CA use. In contrast, anthropomorphism set unrealistic expectations that framed user perceptions of what constituted system failure. Our results showed that technically skilled participants were better able to see beyond artificial humanlike qualities to devise their own mental models of interaction. However, those with lower levels of knowledge described little alteration in their expectations and greater levels of frustration, leading them to question the 'intelligence' of the system, indicating that user expectations of CAs should be scaffold through more considered revelation of system intelligence through design.

System feedback and representation of intelligence – recognising humour as an indicator of state

Despite users actively engaging in the process of 'learning' to speak more simply to their CA, this did not seem to affect their high expectations of system intelligence. Even where users perceived CA failure, they continued to attribute elevated levels of episodic social intelligence to the system such as sarcasm or humour. One reason for this might be the expectations set by the act of conversation and use of humour as a form of interaction. When considering the nature of human-human/face-to-face interaction, "unique informational conditions prevail" [16, p33]. Humans expect to hear intonation and "mutual evaluations will be conveyed by very minor things" [16, p.33]. In this way it is possible that the familiar 'Easter Eggs' and humorous trigger responses to particular phrases or words may have conveyed a level of 'social smarts' that belied the true system capabilities. Expecting to hear sarcasm and humour in conversation then is perhaps unsurprising, although a concern arises where this expectation frames wider perceptions of system capability, resulting in dissatisfaction. This poses something of a conundrum. Whilst users are drawn into serious uses of the system through Easter eggs and playful interactions, our finding show that these interactions also act as affordances, in that they suggest the possibility of action/interaction. This is not entirely unexpected. It has been suggested that framing systems as anthropomorphic "raises user expectations about the extent of their capabilities for intelligence, language, judgement, autonomy, and social norms" [39, p 193]. For our non-technical participants, these expectations were cemented through the playful interactions which characterised initial engagement with these systems.

(b) Reconsidering the interactional promise made by humour: Whilst being a key mechanism for drawing users into more exploratory practices, these playful/humorous interactions had the effect of reinforcing anthropomorphic qualities, thus compounding users' expectations of CA capability. For example, where users were not familiar with the 'intelligence' of the system, or had no technical frame of reference, anthropomorphic language pervaded much of their descriptions of use. It also, however, framed their descriptions of failure and frustration. Equally, users of Google Now had more accurate expectations of the system and were able to reason fairly about what they felt Google did best. Here, Google Now was clearly understood as a hands free interface to a search facility and there were no expectations of wider intelligence or anthropomorphic qualities. In light of these findings, we suggest that future iterations reconsider the interactional promises made by humorous engagement and explore how such engagements could instead support user assessment of system intelligence.

Supporting user evaluation - revealing system capabilities

One of the overarching themes throughout our findings was the inability of users to assess the intelligence of the CA. During a human-human conversation, we use a series of physical indicators and indexical verbal shorthand in order to communicate not only concepts and our identity, but also our

intelligence/ capability. Our conversational partner will seek to elicit such information through their interactions with us; what do we know, think or feel. In these instances our behaviour is tightly synchronised to our conversational partner and, as such, feels natural and immersive. Here, the listener is able to bring to bear a ready model of human interaction developed from prior experiences which allows them to either "map the speaker's behaviours onto richer underlying representations, functions, and conventions – to attribute intelligence to the other", [8 p.71] or to "reach an understanding with another person about something in the world" [20, p.215]. Our findings indicate that whilst users applied this mental model of human communication, it was revised in light of their CA experiences and reflected in their interactions. Whilst the need for this type of effort or learning is reported in other types of 'natural' user interfaces, such as those relying on gesture [31], the absence of any 'natural' means to interrogate and assess system capabilities or state meant that the gulf of evaluation remains too great to result in positive user experience. Within human-human dialogue, the 'internal nature' of the speaker is conveyed via speech.

(c) Consider new ways of conveying CA capability though interaction: Currently, the reality of CAs is such that the system response presents only task-related information to the user. In some cases this has the consequence of conveying capability, for example reverting to visual web-search as an indicator that the system is struggling, or through polite trigger responses that might tell you about limited connectivity. Where users were not able to draw from a technical frame of reference, they tended to find blame in themselves, and often abandoned particular types of task requests, a behaviour seen where systems present a gulf of execution [30]. If we are to truly reflect the expectations set by 'conversation' as interface, some thought should be given to how to convey system limitations and capabilities in instances other than the moment when the system has visibly failed in its task.

Supporting ongoing user engagement by clearly defining the goal of the system

Ideally a conversational interaction should be immersive, resulting in a 'binding hypnotic effect' (p.113) through 'joint spontaneous involvement'. However, the experience of using a CA is currently very far from this. Indeed, our research showed that the principle use-case of the CA was 'hands-free', meaning that an alternative primary task, rather than the conversation, was the focus of attention. Where cognition is split across two tasks, fluidity of user experience is critical to supporting interaction. The majority of our participants used a CA where their primary task required a high level of attention (e.g. driving, cycling, child-minding), what Preece *et al* describe as 'fast thinking' meaning that we "perceive, act, and react to events around us intuitively and effortlessly" [33, p.66]. In each use case described by participants, the activity was not only hands free but required a level of visual attention. Where participants felt they had to resort to 'old school' techniques, or where the CA reverted to screen-based response, the resulting stress and extra effort was seen as failure from their perspective. In contrast, where the user goal was successful operation of the system, then the extra work involved was acceptable and satisfaction of the system was greater. In the majority of cases, however, the primary user goal was not solely to use the CA, making the system a means to an end rather than an end in itself.

(d) Rethink system feedback and design goals in light of the dominant use case: It is clear that the majority of users engage with the system only up to the point that it ceases to provide utility. However, the way in which the system interacts, handles tasks and delivers information does not reflect the dominant use case. This begs the question, what are the design goal of current CA system and how might these be rethought to deliver a more compelling user experience.

Limitations of the study: This study is based upon a UK-centric sample of mostly male participants. Whilst every effort was made to balance gender, the greater number of users responding to the call were male. Though we found no meaningful differences between genders, and theoretical saturation was reached, the imbalance is acknowledged. Beyond this, Siri was the most commonly used system, reflective of the market at the time. Equally, our analysis focused upon interactional themes rather than specific capabilities. Whilst system capability influenced user experience, our study represents the state of the art at the time.

CONCLUSIONS

Overall, in the majority of instances, the operation of the CA systems failed to bridge the gap between user expectation and system operation. Our study showed that users had poor mental models of how their CA worked and that these were reinforced through a lack of meaningful feedback regarding system capability and intelligence. Equally, where playful aspects and trigger responses were programmed into the systems, these served to act as engagement mechanisms, whilst concurrently setting unrealistic expectations that framed the ongoing user experience. A deep 'gulf of evaluation' was also found. This was demonstrated through the extent to which users were consistently unable to ascertain the level of system intelligence, their need to visually confirm all but the simplest tasks, and their reluctance to use the CA for complex or sensitive activities. Finally, whilst the key use case for CAs was found to be 'hands free', system error handling made this largely untenable but for the most economic interaction, calling into question the design goals of such systems. Whilst CAs offer the promise of an engaging and natural user interface, much design and interaction work is required before this potential is realised. Without the humanlike cues and affordances relied upon by multimodal systems, CAs have a particular challenge. We suggest considering (a) ways to reveal system intelligence (b) reconsidering the interactional promise made by humorous engagement, (c) considering how best to indicate capability though interaction, and (d) rethinking system feedback and design goals in light of the dominant use case, as areas for future investigation and development.

REFERENCES

1. ALICE http://www.alicebot.org/bios/richardwallace.html Artificial Intelligence Foundation
2. Jennifer Attride-Stirling. 2001. Thematic networks: an analytic tool for qualitative research. Qualitative Research, 1(3), 385–405
3. Margaret A Boden. 2007. Conversationalists and Confidants. In *Proceedings of Artificial Companions in Society: Perspectives on the Present and Future*. Oxford Internet Institute. Retrieved from http://www.academia.edu/2669520/Artificial_Companions_in_Society_Perspectives_on_the_Present_and_Future
4. Richard, A Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface, *In the proceedings of SIGGRAGH '80* Proceedings, 14, 3 (July 1980), 262-270. doi: 10.1145/800250.807503
5. Timothy Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*. 12, 2: 293-327. doi: 10.1145/1067860.1067867
6. Susan Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. In *The Art of Human-Computer Interface Design*. B.K. Laurel (Ed.), Reading, MA: Addison-Wesley.
7. Alan Bryman. 2004. Social Research Methods (2nd Ed). New York: Oxford University Press
8. Justine Cassell. 2001. Embodied Conversational Agent: Representation and Intelligence in User Interfaces. *AI Magazine.* 22, 4: 67-83. doi: http://dx.doi.org/10.1609/aimag.v22i4.1593
9. Justine Cassell. Tim Bickmore, Lee Campbell, Hannes Vihjalmsson & Hao Yan. 2000. Human Conversation as a System Framework: Designing Embodied Conversational Agents. *In Embodied Conversational Agents.* Justine Cassell. (ed) MIT Press: 29-63
10. Martin Davis. 2000. *Engines of Logic: Mathematicians and the Origin of the Computer*. W W Norton & Company
11. Anind Dey and Jennifer Mankoff. 2005. Designing mediation for context-aware applications. *ACM Transactions on Computer-Human Interaction* 12, 1: 53-80. doi 10.1145/1057237.1057241
12. Abbe Don, Susan Brennen., Brenda Laurel and Ben Shneiderman. 1992 Anthropomorphism: from Eliza to Terminator 2. Panel In Proc. CHI '92, Bauersfeld, P., Bennett, J., and Lynch, G. (eds.). ACM, 67-70
13. Jilian D'Onfro. 2015. *Microsoft Created a Chatbot in China that has Millions of Loyal Followers who talk to it like in the Movie 'Her'*. Business Insider UK. http://uk.businessinsider.com/microsoft-chatbot-xiaoice-2015-8?r=US&IR=T
14. Mauro Dragone, Thomas Holz, Brian R. Duffy, Greogory M.P. O'Hare. 2005. Social Situated Agents in Virtual, Real and Mixed Reality Environments. In *Intelligent Virtual Agents.* Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier & Thomas Rist (Eds.) Proceedings of 5th International Working Conference, IVA 2005 Kos, Greece, September 12-14, 2005. Springer: 166-177. doi: 10.1007/11550617_15
15. James R. Glass, 1999. Challenges for Spoken Dialogue Systems. *In the proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. (ASRU), Colorado, USA
16. Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behaviour.* Pathenon
17. Aurthur Graesser, Haiying Li & Carol Forsyth. 2014. Learning by Communicating in Natural Language with Conversational Agents. Current Directions in Psychological Science. 23, 5: 374-380 doi: 10.1177/0963721414540680
18. Greg Guest, Arwen Bunce & Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, 18, 1: 59-82. doi: 10.1177/1525822X05279903
19. Joakim Gustafson, Johan Boye, Morgan Fredriksson, Lasse Johanneson, Jürgen Königsmann. 2005. Providing Computer Game Characters with Conversational Abilities. In *Intelligent Virtual Agents.* Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier & Thomas Rist (Eds.) Proceedings of 5th International Working Conference, IVA 2005 Kos, Greece, September 12-14, 2005. Springer: 37-51 doi: 10.1007/11550617_4
20. Jürgen Habermas. 1998. *On the Pragmatics of Communication*. Cambridge: Polity Press.
21. Richard H. R. Harper. 2010. *Texture: Human Expression in the Age of Communications Overload*. MIT Press
22. Kerstin Heuwinkel. 2012. Framing the Invisible – The Social Background of Trust. In *Your Virtual Butler: The Making-of*. Robert Trappl (ed). Springer, 16-26. doi: 10.1007/978-3-642-37346-6_3
23. Ido A. Iurgel & Manuel Ziegler. 2005. Ask & Answer: An Educational Game Where It Pays to Endear Your Capricious Virtual Companion. In *Intelligent Virtual Agents.* Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier & Thomas Rist (Eds.) Proceedings of 5th International Working Conference, IVA 2005 Kos, Greece, September 12-14, 2005. Springer: 15-24 doi: 10.1007/11550617_2
24. Stephan Kopp, Lars Gesellensetter, Nicole, C. Krämer & Ipke Wachsmuth. 2005. A Conversational Agent as Museum Guide: Design and Evaluation of a Real-World Application. *Intelligent Virtual Agents*. 3661: 329-343 doi: 10.1007/11550617_28
25. Joseph Carl Robnett Licklider. Man-Computer Symbiosis . IRE Transactions on Human Factors in Electronics, HFE-1,(March, 1960), 4-11

26. Paul P. Maglio and Christopher S. Campbell. 2003. Attentive agents. *Communications of the ACM* 46, 3: 47-51 doi: 10.1145/636772.636797
27. Cade Metz. 2015. Get a Peek at Using Facebook's New Assistant, 'M'. *Wired*. Retrieved from http://www.wired.com/2015/09/get-peek-someone-using-facebooks-new-assistant-m/ (accessed 07.09.15)
28. Roger Moore. 2012. Spoken Language Processing: Where do we go from Here? In *Your Virtual Butler: The Making-of*. Robert Trappl (ed). Springer, 119-133 doi: 10.1007/978-3-642-37346-6_10
29. Andreea I Niculescu, Kheng Hui Yeo, Luis F D'Haro, Seokhwan Kim, Ridong Jiang & Rafael E Banchs. 2014. Design and evaluation of a conversational agent for the touristic domain. *In Proceedings of APSIPA'14*: 1–10. doi:10.1109/APSIPA.2014.7041744
30. Don Norman. 2013. *The Design of Everyday Thing*s. Basic Books.
31. Kenton O'Hara, Richard Harper, Helena Mentis. Abigail Sellen and Alex Taylor. 2013. On the naturalness of touchless: putting the "interaction" back into NUI. *ACM Transactions on Computer-Human Interaction*. 20, 1 doi: 10.1145/2442106.2442111
32. Sabine Payr. 2012. Virtual Butlers and Real People: Styles and Practices in Long Term Use of a Companion. In *Your Virtual Butler: The Making-of*. Robert Trappl (ed). Springer, 134-178 doi: 10.1007/978-3-642-37346-6_11
33. Jenny Preece, Yvonne Rogers & Helen Sharp. 2015. *Interaction Design: Beyond Huma- Computer Interaction.* Wiley& Sons Ltd
34. Stephen Pulman. Johan Boye, Marc Cavazza, Cameron Smith & Raúl Santos de la Cámara. 2010. How was your Day? *In Proceedings of the 2010 Workshop on Companionable Dialogue Systems. Assoc. for Computational Linguistics*, Stroudsburg, PA, USA: 37-42
35. Byron Reeves & Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge University Press.
36. Deborah Richards. 2012. Agent-based museum and tour guides: applying the state of the art. In *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System (IE '12).* ACM (2012) doi: 10.1145/2336727.2336742
37. Lina Maria Rojas-Barahona and Christophe Cerisara. 2014. Bayesian Inverse Reinforcement Learning for Modelling Conversational Agents in a Virtual Environment. *Computational Linguistics and Intelligent Text Processing*. 8403: 503-514 doi: 10.1007/978-3-642-54906-9_41
38. Nicole Shechtman, and Leonard M Horowitz. 2003 Media Inequality in Conversation: How People Behave Differently when Interacting with Computers and People. *In Proceedings of the. SIGCHI Conference on Human Factors in Computing Systems (CHI '03).* ACM, (2003), 281-288 doi: 10.1145/642659.642661
39. Nathan Shedroff & Christopher Noessel. 2012. *Make it so: Interaction Design Lessons from Science Fiction*. Brooklyn, New York: Rosenfeld
40. Stergios Tegos, Stavros Demetriadis, Thrasyvoulos Tsiatsos. 2012. Using a Conversational Agent for Promoting Collaborative Language Learning. In *Proceedings of Fourth International Conference on Intelligent Networking and Collaborative Systems*: 162-165 doi: 10.1109/iNCoS.2012.105
41. Janienke Sturm, Else den Os, Lou Boves. 1999. Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System. *In the proceedings of ESCA Workshop on Interactive Dialogue in Multi-Modal Systems (UDS-99)*, Kolster Irsee, Germany.
42. Robert Trappl. 2012. From Jeeves Jeannie to Siri, and Then? In *Your Virtual Butler: The Making-of*. Robert Trappl (ed). Springer, 1-8 doi: 10.1007/978-3-642-37346-6_1
43. Giorgio Vassallo, Giovanni Pilato, Agnese Augello & Salvatore Gaglio. 2010. Phrase Coherence in Conceptual Spaces for Conversational Agents. In *Semantic Computing.* Sheu, Yu, Ramamoorthy, Joshi & Zadeh.(eds). IEEE: 357-371 doi: 10.1002/9780470588222.ch18
44. Astrid M. Von der Pütten, Nicloe C. Krämer, Jonathan Gratch & Sin-Hwa Kang. 2010. "It doesn't matter what you are!" Explaining Social Effects of Agents and Avatars. In *Computers in Human Behaviour*, 26: 1641-1650 doi: 10.1016/j.chb.2010.06.012
45. Joseph Weizenbaum. 1976. *Computer Power and Human Reason: From Judgement to Calculation*. W. H. Freeman
46. Joseph Weizenbaum. 1965. Eliza: A Computer Program for the Study of Natural Language Communication between Man and Machine. In *Communications of the ACM*. 9,1:36-45 doi: 10.1145/365153.365168
47. Yorick Wilks. 2010. Is a companion a distinctive kind of relationship with a machine? .*In Proceedings of the 2010 Workshop on Companionable Dialogue Systems (CDS '10).* Association for Computational Linguistics, Stroudsburg, PA, USA, 13-18
48. Yorick Wilks (ed.). 2010. *Close Engagements with Artificial Companions. Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins.