Inferring Capabilities of Intelligent Agents from Their External Traits

BART P. KNIJNENBURG, Clemson University
MARTIJN C. WILLEMSEN, Eindhoven University of Technology

We investigate the usability of humanlike agent-based interfaces for interactive advice-giving systems. In an experiment with a travel advisory system, we manipulate the "humanlikeness" of the agent interface. We demonstrate that users of the more humanlike agents try to exploit capabilities that were not signaled by the system. This severely reduces the usability of systems that look human but lack humanlikehumanlike capabilities (overestimation effect). We explain this effect by showing that users of humanlike agents form anthropomorphic beliefs (a user's "mental model") about the system: They act humanlike towards the system and try to exploit typical humanlike capabilities they believe the system possesses. Furthermore, we demonstrate that the mental model users form of an agent-based system is inherently integrated (as opposed to the compositional mental model they form of conventional interfaces): Cues provided by the system do not instill user responses in a one-to-one matter but are instead integrated into a single mental model.

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems**]: Human Factors; H.5.2 [**User Interfaces**]: Interaction Styles; H.5.2 [**User Interfaces**]: Natural Language; H.5.2 [**User Interfaces**]: Evaluation/methodology

General Terms: Design, Human Factors, Measurement, Theory

Additional Key Words and Phrases: Agent-based interaction, anthropomorphism, usability, feedforward and feedback, mental model

ACM Reference Format:
Bart P. Knijnenburg and Martijn C. Willemsen. 2016. Inferring capabilities of intelligent agents from their external traits. ACM Trans. Interact. Intell. Syst. 6, 4, Article 28 (November 2016), 25 pages.
DOI: http://dx.doi.org/10.1145/2963106

1. INTRODUCTION

Advice-giving systems have been around for several decades in the form of intelligent tutoring systems [Sleeman and Brown 1982], expert systems [Carroll and McKendree 1987], and recommender systems [Schafer et al. 1999]. Recent advances in machine learning and artificial intelligence have given advice-giving systems the capability to assist users with a wide variety of tasks. As the flagship commercial advice-giving systems—Apple's Siri, Microsoft's Cortana, and Google Now—are tremendously complex, their designers have come to realize that a standard Graphical User Interface (GUI) is not sufficient to harness their power. Instead, these systems have a renewed interest in *agent-based interaction* (an interaction paradigm that has been a topic of human-computer interaction research for several decades [Behrend and Thompson

The reviewing of this article was managed by special issue associate editors Nava Tintarev, John O'Donovan, and Alexander Felfernig.
Author's addresses: B. P. Knijnenburg, Human-Centered Computing Division, School of Computing, Clemson University, 215 McAdams Hall, Clemson, SC 29634; email: bart@clemson.edu; M. C. Willemsen, Human-Technology Interaction Group, School of Innovation Sciences, Eindhoven University of Technology, IPO 0.17, P.O. Box 513, 5600 MB Eindhoven; email: m.c.willemsen@tue.nl.

DOI: http://dx.doi.org/10.1145/2963106

2011; Benbasat et al. 2010; Qiu and Benbasat 2009; Yoo and Gretzel 2009; Nowak and Rauh 2005; Al-Natour et al. 2006; Hess et al. 2006; Cowell and Stanney 2005; Bickmore and Cassell 2001; Dehn and van Mulken 2000; Bradshaw 1997; Walker et al. 1994; Quintanar et al. 1987; Nickerson 1976]). In these systems, the user interacts with a virtual entity using natural language.

Because agent-based interaction is richer, finer grained, and more natural than our interaction with more tool-like interfaces, it should be more suitable for the increasingly common situation where computers give advice and decision support [Negroponte 1997; Laurel 1990; Williges et al. 1987]. Moreover, research in highly controlled environments shows that people tend to find agent-based interaction more enjoyable and more natural than interacting with a standard GUI [Kang et al. 2012; Yoo and Gretzel 2009; Serenko 2008; Hess et al. 2006; Cowell and Stanney 2005]. At the same time, though, agent-based interfaces developed in academia seem to be unable to live up to these promises of enjoyable and natural interaction [Nowak 2004; Andersen and Andersen 2002; Bickmore and Cassell 2001; Dehn and van Mulken 2000; Shneiderman 1997], and even the available commercial systems seem to fail us on many occasions. Many of them remind us of the late MS Office agent "Clippy," which caused annoyance rather than fluent interaction [Trott 1998].

Why do these systems fail to provide the more flexible and enjoyable interaction they promise? In this article, we argue that the key to this failure lies in the fact that designers are attempting to give these agent-based advice-giving systems a *humanlike appearance*: They have a human name, speak with a human voice, use full sentences, employ a varied sentence structure and wording, and sometimes even have a humanlike avatar. In response to this, users will assume that the system has *humanlike capabilities*, and they may *overestimate* the system, which breaks the interaction. For example, try to ask Siri the following question: "Do I have a meeting on February 15?" and then follow up that question with the following: "At what time does the sun rise on that day?" You will find that Siri is unable to understand your reference to the previous question and thus answers the latter question by giving you the time of sunrise for the *current* day rather than February 15.[1]

In this article, we investigate the cognitive principles behind this overestimation phenomenon. While existing research on humanlike agents focuses on the social psychological effects of agent interfaces [Behrend and Thompson 2011; Benbasat et al. 2010; Qiu and Benbasat 2009; Nowak and Rauh 2005; Al-Natour et al. 2006], we instead focus on how the usability of agent-based interaction differs from traditional GUIs. In Section 2, we develop a theory that pinpoints the fundamental difference between humanlike agents and GUIs. This theory revolves around users' understanding of the way the system works (i.e., the "user's model" [Norman 1986]). Specifically, the usability of GUIs is determined by the formation of a compositional mental model, in which cues of the system relate to user reactions in a one-to-one manner (cf. Brinkman 2003]). Our theory argues that agent-based interfaces, on the other hand, have an *integrated user's model*, in which users anthropomorphize the system and infer a broader set of humanlike capabilities that are not necessarily related to the specific cues displayed by the system. We argue that this integrated mental model is the cause of both the theoretical advantage of agent-based interfaces (i.e., by providing an instant schema of humanlike intelligence that provides guidelines for how to interact with the system), as well as the disappointing nature of current applications (i.e., because of the almost-inevitable overestimation of the system's humanlike capabilities).

In Section 3, we develop a series of hypotheses that allow us to (1) demonstrate that the user's mental model of agent-based interfaces is indeed more likely to be integrated

[1]Tested with iOS 9.2.

rather than compositional, (2) test our expectations about the consequences of an integrated mental model on the users' humanlike and capability-exploiting responses, and (3) evaluate the effect that these instilled responses have on the usability of the interaction (i.e., a potential overestimation effect). In Section 4 we outline a Wizard-of-Oz experiment to test these hypotheses, and in Section 5 we provide the results of this experiment. In Section 6, we discuss these results, and in the conclusion (Section 7), we highlight the pitfalls of using an agent-based interaction paradigm and suggest ways to improve the usability of agent-based interaction.

2. RELATED WORK AND THEORY DEVELOPMENT

In this section, we extend Norman's theory of human-computer interaction [Norman 1986] to agent-based interfaces. Based on related work on agent-based interaction, we posit a number of conjectures to argue that agent-based interfaces have an *integrated user's model*. These conjectures form the basis for the experiment we conducted in order to empirically validate this theory of human-agent interaction.

2.1. The Compositionality of Traditional User Interfaces

First, let us consider how users interact with traditional GUIs using Norman's theory of human-computer interaction [Norman 1986], which remains one of the most prominent theories in human-computer interaction today [Norman 2013]. To explain why some systems are more usable than others, Norman argued that there are two gulfs between the user and the system: the gulf of execution and the gulf of evaluation. The gulf of execution manifests itself when the user has to discover how to manipulate the system to accomplish a task. The gulf of evaluation emerges after the user has provided some input and now has to interpret what the system has done and whether this is in line with what he/she wanted to happen.

Norman states that a system is usable if users are able to easily overcome (bridge) these two gulfs. Users do this by forming a "user's model," a mental representation of the way the system works. A user's model may contain some gaps and inconsistencies, and it rarely matches the actual internal workings of the system, but an appropriate mental model assists users to infer which interface actions fulfill their goal (bridging the gulf of execution) and what the output of the system means (bridging the gulf of evaluation). According to Norman, the formation of an adequate mental model is greatly facilitated by providing appropriate feedback and feedforward. For instance, salient cues in the system interface such as carefully worded labels on buttons (feedforward) let users infer what the system can do, and understandable output (feedback) allows them to see if the system actually did what they wanted.

An operationalization of Norman's "user's model" is the Layered Protocol Theory (LPT) [Taylor 1988]. LPT decomposes user-system interaction into a set of layers that each have a different level of abstraction. On each successive layer, users' intentions are broken down into smaller components. Brinkman [2003; Brinkman et al. 2007] argued that this compositional character of the interaction is reflected in the users' mental model and that usability is therefore *compositional*. This implies that the user's mental model (and thus the usability) of a graphical user interface is simply the aggregate of the mental models of its widgets (e.g., levers, buttons, text fields, scrollbars). Brinkman showed that this assumption held for various interfaces.

The idea of a compositional mental model has been assumed by many other usability researchers and designers. Evaluation techniques like Heuristic Evaluation [Nielsen 1994] explicitly evaluate the usability of the separate parts of the interface, so the effectiveness of such techniques partially depends on the legitimacy of the compositional character of user interfaces.

2.2. Agent-Based Mental Models

Norman's usability theory is applicable to "real-life" interfaces (e.g., doors and phones) as well as our current software interfaces [Norman 1988]. However, agent-based interfaces typically lack the common levers, buttons, text fields, and scrollbars. How do users form a mental model of an agent-based interface? Cook and Salvendy [1989] argued that users infer the model of an agent-based system from the way it "looks" and "talks" and the apparent intelligence of its responses, just like they would do when interacting with other human beings. In light of Norman's theory, the agent's cues in terms of appearance and language provide feedforward, and the actual system capabilities provide feedback. This notion is in line with Laurel [1990], who stated that a system shows signs of humanlike intelligence when it shows humanlike responsiveness (meaning that it is able to respond flexibly to incomplete requests) and when it shows humanlike accessibility (meaning that it looks like a human being and uses grammatically correct sentences).

In terms of feedforward, we will distinguish between humanlike appearance and capability cues. *Human-like appearance cues* are cues that make the system look human, such as full sentences, a varied sentence structure and wording, and a humanlike avatar. *Human-like capability cues* are cues that signal that the system has capabilities similar to humans. In this article, we focus on linguistic capabilities, specifically, the capability to understand *references* to the context (time, place, previous sentences) of the conversation [Levinson 1983; Halliday and Hasan 1976]. An agent that displays humanlike capability cues makes extensive use of such implicit references, using words "you" and "I" (references to persons), "there" and "here" (references to place), "now" and "then" (references to time), and "that" (references to things mentioned in previous sentences).

Previous studies have shown that users attribute common human intelligence to systems that provide more humanlike appearance and capability cues. For example, users of a system with a cartoon character that "talks" in full sentences and personifies itself believe that it shows some form of common human intelligence as well, while these users do not show a similar belief when using a system without such a cartoon character that talks "computerese" [De Laere et al. 1998; Quintanar et al. 1987]. We thus argue that:

Conjecture 1: An agent-based user's model is one of "believed humanlike intelligence": The more humanlike the system looks (appearance cues) and the more capabilities it displays (capability cues), the more intelligent users believe the system to be.

This conjecture is presented on the left side of Figure 1. Note that the instilled use image may not hold foot in the long run: *Actual* capabilities might not necessarily co-occur with capability *cues*; the system might exhibit specific linguistic capabilities (e.g., using the word "here" to refer to the current location) without actually being able to understand them in the user dialog (e.g., it may not be able to infer the current location when the user uses the word "here"). In effect, cues of humanlike appearance and capabilities can underplay or overplay the agent's actual capabilities. Over time, the actual capabilities of the system provide feedback about the accuracy of the use image, but this process is believed to be much slower, because it requires users to find the boundaries of the system's capabilities by trial and error (see Figure 1, bottom).

2.3. Anthropomorphism

What psychological mechanisms could underlie the user's model of believed intelligence as postulated in Conjecture 1? Thompson [1980] found that users of a natural language-based system showed a tendency to anthropomorphize the behavior of the system:

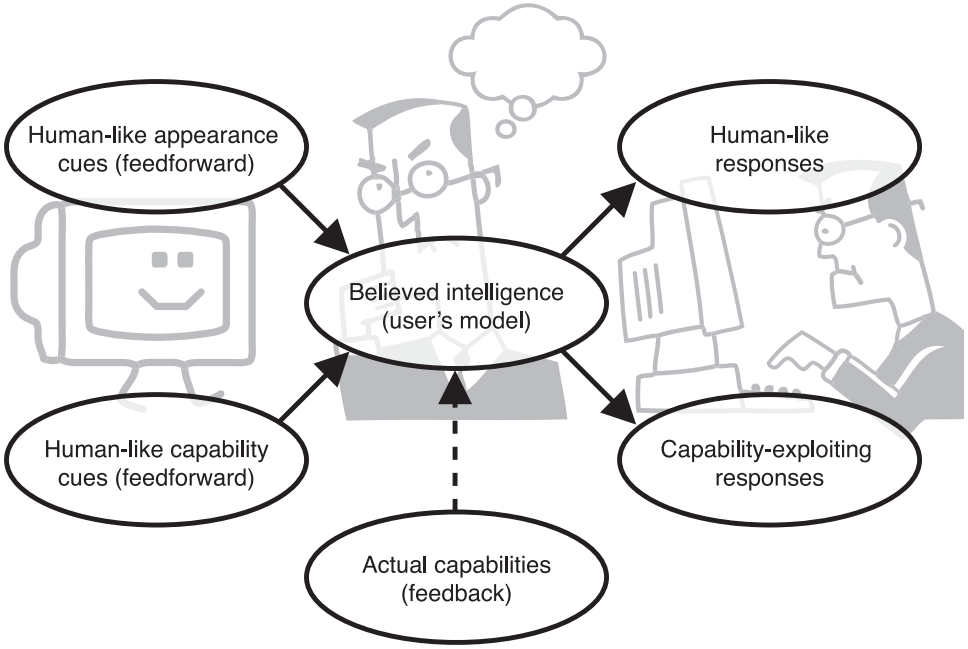


Fig. 1. Feedforward (cues) and feedback (actual capabilities) constitute the user's model (believed intelligence), which leads users to give humanlike responses and exploit the capabilities they believe the system has.

Users treated the system as if it were a human being. Other research confirms this tendency to anthropomorphize systems that display humanlike cues: Researchers have found that the use of personalization, conversational tone, affective responses, and diversified wording leads users to perceive agents as being more "human" [Quintanar et al. 1987; De Laere et al. 1998].

Not only agent-based systems are subject to anthropomorphism. Researchers have shown that users of any computer system occasionally engage in negative anthropomorphism (e.g., shouting [Chin, et al. 2005]) and unconsciously adhere to social principles that normally apply between humans [Reeves and Nass 1996]. Reeves and Nass [1996] show that computer users may, for example, show a "politeness effect": They evaluate a computer more positively when asked by that same computer than when asked by another computer. Why would human beings try to be polite towards computers or shout at them? Bradshaw [1997] offers an explanation for this phenomenon: When a system's behavior is too complex to understand, users are inclined to take the "intentional stance" [Dennett 1987] when reasoning about these systems: They attribute intentional behavior to systems as a convenient shortcut towards explaining complicated behavioral patterns, and this also leads them to adhere to human social principles. Although the intentional stance holds for any type of system, agents seem to instill stronger anthropomorphic reactions [Nowak 2004]. In sum:

Conjecture 2: In agent-based systems, the intentional stance is at the heart of the construction of the user's mental model. The user's model is an anthropomorphic construct, instilled by humanlike cues.

2.4. Consequences of the User's Mental Model

As the user's model is a mental construct, one cannot directly observe whether it is anthropomorphic. However, observable reactions to the user's model can provide

evidence of the nature of the latent mental model. For example, if the user's model is anthropomorphic, users will interact with the agent in a way that is in accordance with human-human interaction. Such "humanlike responses" (the top-right ellipse in Figure 1) are reactions that previous research has found to occur in human-human interaction but not in human-computer interaction. In this article, we focus on linguistic indicators of believed human intelligence, such as the use of long and grammatically correct sentences. Indeed, existing research has found that the use of a humanlike avatar and personalized feedback (humanlike cues that may lead to an anthropomorphic mental model) leads users to be more verbose and grammatically correct to computer systems in their responses [Brennan 1991; Rosé and Torrey 2005; Walker et al. 1994; Richards and Underwood 1984]. Hence, we argue:

Conjecture 3: Since the user's mental model of an agent is anthropomorphic, users will act in a more humanlike way towards a system they believe to be more intelligent.

Moreover, if the system looks and behaves human, then users will believe it has typical human capabilities and will try to exploit these capabilities (the bottom right ellipse in Figure 1). As mentioned earlier, in this article we focus on the linguistic capability to understand implicit references to the context of the conversation [Levinson 1983; Halliday and Hasan 1976]. References to the textual context ("anaphora") or the situational context ("deixis") are often used in human conversation to speed up the interaction. However, computers are notoriously bad at understanding such references [Winograd 1972; Dey 2001; Moratz and Tenbrink 2006; Smith 2013; Scheutz et al. 2011]. In the case of agent-based interaction, users may believe that a humanlike system, like a human being, can resolve such implicit references [cf. Reichel et al. 2014]. In this article, we therefore define *capability-exploiting responses* as the use of words referring to the current or previously mentioned location, like "here" and "there" (spatial anaphora/place deixis); the current or previously mentioned time, like "now" or "then" (temporal anaphora/time deixis); or a previously mentioned object, like "that trip" or "that ticket" (nominal anaphora/discourse deixis). In sum, we argue that:

Conjecture 4: Users will assume that systems they believe to be more intelligent also have more advanced linguistic capabilities, and they will try to exploit these capabilities.

2.5. An Integrated Mental Model?

If an agent-based system would be just like a traditional GUI, then its user's model would be compositional: There would be a one-to-one mapping from its cues to a related functionality. Each cue would then instill its own mental model and induce a corresponding response. In other words, users would "mirror" the agent's behavior and try to exploit a certain humanlike capability if and only if the system would provide a cue of this capability; for example, they would use contextual references ("here" and "now") if and only if the system gave them an explicit cue that this is possible (by using contextual references in its responses). Brennan [1991] found support for such a one-to-one mapping. In her experiments with both human-human and natural language-based human-computer interactions, she found that participants were likely to show *syntactic entrainment*, a direct reflection of the conversation partner's responses. According to Brennan's findings, one could evoke a certain behavior in the users' response by expressing the same behavior in the agent. Let us take, for example, a number of *cues* that an agent-based system could display: cue A, a human-looking avatar; cue B, using grammatically correct sentences; and cue C, referring to the current location as "here." Furthermore, let us consider the following *responses* that the user can give: response B, using grammatically correct sentences; response C, referring to the current location as

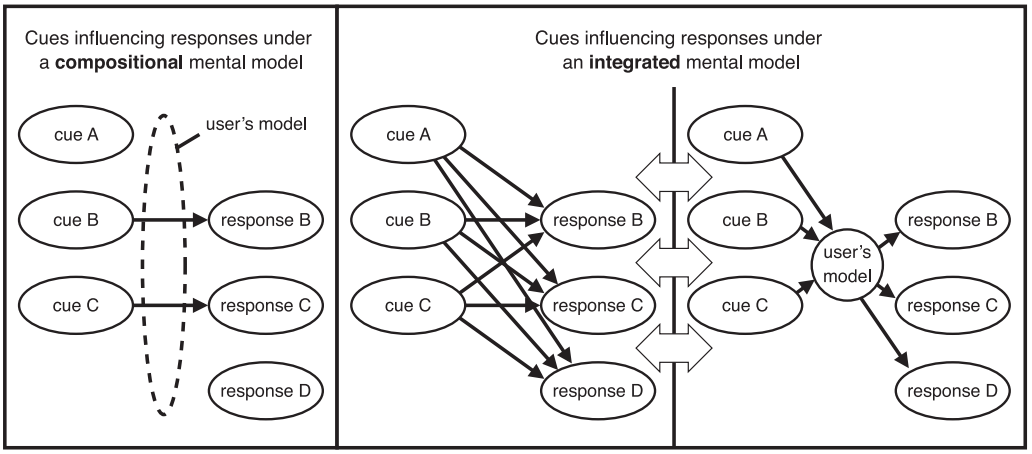


Fig. 2. The effects of system cues on user responses differ for compositional mental models and integrated mental models.

"here; and response D, asking multiple questions at once. According to Brennan's theory of *syntactic entrainment*, cue B will instill response B and cue C will instill response C, but cues would never be able to instill any responses that are not directly related (e.g., cue B cannot instill response C). Consequently, cue A (which has no corresponding response) does not instill any humanlike responses, and response D (which has no corresponding cue) should not occur at all. The left panel of Figure 2 shows a graphical representation of a compositional mental model.

However, the intentional stance [Dennett 1987] should allow users to create an *integrated* mental model based on the behavior of the system *as a whole*. If the system is sufficiently humanlike, then it will be attributed intentional behavior, and this attribution is based on the "humanlikeness" of the agent as a whole, not on a specific part of its behavior. This gives users a "shortcut" to the functionality of the system, because they can predict what the system can and cannot do based on what they know their fellow humans can and cannot do. For example, users may infer that humanlike systems are able to handle complex sentences with implicit references. In terms of Norman's usability theory, it is stated as follows: The fact that the "system" is "human" provides them *instantaneously* and *effortlessly* with a detailed mental model of what it can do and how to interact with it. In the words of Laurel [1990]: "[An agent-based interface] makes optimal use of our ability to make accurate inferences about how a character is likely to think, decide and act on the basis of its external traits. This marvelous cognitive shorthand is what makes plays and movies work [. . .] With interface agents, users can employ the same shorthand—with the same likelihood of success—to predict, and therefore control the actions of their agents" (pp. 358–359).

If users *integrate* the system cues they receive into a mental model of believed intelligence, then there would be a more complex relation between system cues and user responses than the one-to-one mapping of a compositional mental model. As can be seen in the middle panel of Figure 2, in such a case, cue B not only instills response B but also can instill response C (e.g., if the system uses grammatically correct sentences, then users are not only more likely to use grammatically correct sentences as well but also more likely to refer to the current location as "here"). Furthermore, a cue that does not have a directly related response (e.g., cue A) may still instill certain responses (e.g., a human-looking avatar may instill grammatically correct sentences and implicit references to place or time). Also, responses not directly connected to a related cue

(e.g., response D) can still be evoked by other cues (e.g., the user might think that the agent can handle multiple requests at the same time, even if it does not display any feedforward cue that suggests that it has this capability). In this case, the relation between system cues and user responses is mediated by the formation of a single, integrated mental model (Figure 2, right panel). In effect, we argue that:

CONJECTURE 5: All cues about the intelligence of the system will be integrated into a single mental model and instill a series of possibly unrelated responses.

Taken together, Conjectures 1–5 present the cognitive principles of human-agent interaction: If a system employs humanlike appearance and capability cues, then users will believe it to be intelligent (C1). This anthropomorphic mental model (C2) will cause users to employ humanlike (C3) and capability-exploiting (C4) responses towards the system. Most importantly the user's mental model is *integrated* (C5), which means that any kind of cue can instill any kind of response. This integrated mental model has profound effects on the usability of agent-based systems; the following subsection discusses those effects.

2.6. The Effects of an Integrated Mental Model on Usability

An integrated mental model of agent-based interfaces can have both positive and negative effects on their usability (for an overview of selected studies in this field, see Dehn et al. [2000] and, more recently, Qiu and Benbasat [2009]). On the positive side, the integrated mental model makes agent-based interfaces especially suitable for performing complex tasks such as giving advice to support decisions. Typical GUIs have a compositional mental model with a one-to-one mapping from the layout of the interface to the functions of the system. The more complex the system gets, the more interface elements it requires, and the more difficult it becomes for users to retrieve the adequate mental model for each widget. Therefore, it might be impossible to create a really usable GUI for a complex system. If agent-based interfaces are subject to an integrated use image, then they would instantly provide users with a heuristic to determine what they can and cannot do and with an easy way to access this functionality. So instead of having to form a mini-mental model for every widget in the interface, the users will identify the humanness of the interface, instantly infer its integrated mental model, and act accordingly. In effect, the more humanlike the system, the more usable it will be. In line with this, the experiments of Dharna et al. [2001] and Quintanar et al. [1987] found that more humanlike interaction increased usability.

On the negative side, though, an integrated mental model poses severe problems when it is incorrect. This occurs when the system cannot perform a certain function—or a number of functions—that the user expects the system to be able to perform based on the formed mental model. If the system looks more capable than it actually is, then users might *overestimate* the system's capabilities, which would then result in confusion and reduced usability. This would be especially true for users' initial interactions with the system, when they have not yet learned the actual capabilities of the system from its feedback. In other words, among the dimensions of usability, the learnability will arguably suffer the most. Several studies have suggested that such overestimation might occur [cf. Richards et al. 1984; Brennan 1991; Erickson 1997; Forlizzi et al. 2007; Walker et al. 1994; Shneiderman 1981], and some have demonstrated it in qualitative accounts [cf. Serenko 2006], but it has to our best knowledge never been tested in a controlled experiment.

By definition, the usability of interaction with a system is good when users try to use the capabilities that the system provides, and do not try to use any capabilities the system does not provide. Since users base their interaction decisions on their mental model of the system, this means that this mental model has to match the actual system

capabilities [Norman 1986]. If the user's model of an agent were compositional, then it would be fairly easy to "manage" this mental model so that it matches the actual system capabilities: the system would simply have to provide a capability cue for each actual capability. However, an integrated mental model implies that it is much harder to control the user's model, as it suggests that there is more than just a one-to-one relation between cues and responses. In effect, even humanlike *appearance* cues may instill *capability*-exploiting responses: Merely "looking human" may be enough to make users believe that the system has certain typically humanlike capabilities (even if they are not actually present).

In sum, the presumed integrated mental model is responsible for the greatest advantage but at the same time also the most significant drawback of agent-based interaction: Due to our natural tendency to anthropomorphize, it is very easy to *instantly* create a complex, integrated mental model from which users can effortlessly infer a myriad of complex functions to perform with the system, along with possible ways to exploit them. However, since these functions are not directly coupled to a specific underlying cue, an overestimation effect can easily occur in which one or more of these functions are actually not available in the system, and it will be rather difficult to tweak the user's mental model such that it perfectly matches the actual system capabilities.

3. HYPOTHESIS DEVELOPMENT

The five conjectures outlined in the previous section present a general theory of human-agent interaction and usability. In this section, we will develop hypotheses to empirically validate this theory. Specifically, we will test whether humanlike agents indeed instill humanlike (C3) and capability-exploiting (C4) responses, whether the user's mental model of agent-based interfaces is more likely to be integrated rather than compositional (C5), and what kind of an effect this has on the usability of the interaction (i.e., a potential overestimation effect if the system cannot live up to its promises). Together, these hypotheses provide support for our conjectures that the user's mental model of an agent-based system is an anthropomorphic construct (C2) of "believed human intelligence" (C1).

To test these hypotheses in our experiment, we built an advice-giving system and independently varied its feedforward cues and its actual capabilities. The feedforward provided by the system is varied in three different "cues" conditions: "computer-like cues" (the agent looks like a computer and talks "computerese"), "humanlike appearance cues" condition (the agent looks and talks like a human being), and "humanlike appearance and capability cues" (the agent uses deictic and anaphoric references (e.g., the word "here" to refer to a place); this signals its capability to understand such references). Furthermore, the actual capabilities of the system are varied in two conditions: The system with "low capabilities" can only process simple, complete requests, while the "high capabilities" system can process complex requests with implicit (deictic or anaphoric) references, just like a human being would be able to handle. We then asked users to perform a number of tasks with the system and measured their humanlike and capability-exploiting responses, as well as the usability of the interaction.

In terms of usability, we argue that a system with high capabilities should generally be more efficient (i.e., time and number of requests per task) and satisfying to use:

H1. Efficiency and user satisfaction in the "high capabilities" condition will be higher than in the "low capabilities" condition.

One might argue that it is obvious that a system with more capabilities would be easier to use. However, if the user's mental model is *not* integrated, then more capabilities mean that the user needs to form more (separate) mini-mental models of the system, which leads to more cognitive overhead. A formal test of this hypothesis is therefore still a useful endeavor.

Within the "low capabilities" conditions, there is an opportunity for overestimation to occur, that is, when the system appears humanlike and displays humanlike capabilities but is not able to understand complex requests like a human being would (i.e., the system with humanlike cues but low capabilities). In this situation, users may take much longer to learn the actual capabilities of the system, or they may even give up if entirely if they feel unable to learn the actual system capabilities. In other words, in the "low capabilities" conditions, the "humanlike" systems may have a lower learnability (i.e., in terms of the reduced time per task from the first to the last task) and a lower effectiveness (i.e., in terms of the proportion of participants finishing all tasks) than the "computer-like" systems:

H2. Within the "low capabilities" conditions, the occurrence of "humanlike appearance cues" and "humanlike appearance and capability cues" will lead to lower learnability and effectiveness than "computer-like cues."

To test the existence of an anthropomorphic mental model, we inspect users' humanlike and capability-exploiting responses in the "high capabilities" conditions in hypotheses 3 and 4. We can only use the "high capabilities" conditions for these tests, because in the "low capabilities" conditions the system provides feedback in the form of errors when the user tries to exploit nonexistent capabilities. Consequently, if the system has low capabilities, users will eventually learn to restrict their vocabularies in order to "make the system work." So to accurately measure how these different cues conditions instill different types of user behavior, we restrict these hypotheses to the conditions where users' behavior is not influenced by the actual capabilities of the system (i.e., the "high capabilities" conditions).

In terms of humanlike responses, we predict that users are expected to give more of such responses when more humanlike cues are provided. This suggests conducting planned contrast comparisons between the two "humanlike" conditions and the "computer-like" condition:

H3. Within the "high capabilities" conditions, users in the "humanlike appearance cues" and the "humanlike appearance and capability cues" conditions exhibit more humanlike responses than in the "computer-like cues" condition.

Finally, we can test whether this anthropomorphic mental model is compositional or integrated. If and only if the user's mental model is compositional, then there would be a one-to-one mapping between cues and responses. Specifically, a humanlike *appearance cue* given by the agent (e.g., a humanlike avatar) cannot evoke capability-exploiting responses from the user (e.g., using context of time and place). In that case, users will try to exploit humanlike capabilities in the "humanlike appearance and capability cues" condition only, and users in the "humanlike appearance cues" condition use the same (low) amount of capability-exploiting responses as users in the "computer-like cues" condition. In other words, this suggests a planned contrast comparison between the "humanlike appearance and capability cues" condition on the one hand and the "humanlike appearance cues" and "computer-like cues" conditions on the other hand:

H4a. If and only if the use image is compositional, then—within the "high capabilities" conditions—users exhibit more capability-exploiting responses in the "humanlike appearance and capabilities cues" condition than in the "humanlike appearance cues" condition or the "computer-like cues" condition.

On the other hand, if and only if the user's mental model is integrated, then humanlike appearance cues *can* evoke capability-exploiting responses. In that case, users will try to exploit humanlike capabilities in both the "humanlike appearance cues" *and* the

Table I. A Description of the System Capabilities Manipulation

Low capabilities *The low capabilities system lacks several typically human capabilities, specifically:*	**High capabilities** *The high capabilities system has several typically human capabilities, specifically:*
The system cannot infer information that is implicitly stated, for example, if no departure time is given, it will not process the request.	The system can infer implicitly stated information: for example, if no departure station is given, it assumes the user's current location (the 登rigo •of deixis).
The system does not understand spatial or temporal references, for example, it will not understand 塗ere •or 渡ow •	The system can determine the meaning of deictic references, for example, it will understand 塗ere •and 渡ow •
The system treats every request as a new entity, for example, a request like: 努hat is the price for that trip? •will not be processed.	The system can infer anaphoric references to times, places or trips mentioned in previous requests, for example, it will understand 鍍hen •and 鍍hat trip •
The system can only handle one request at a time.	The system can handle multiple connected requests.
The system cannot handle convoluted sentences, that is, it will only interpret the first 10 words of a request.	The system can handle requests of any length.

"humanlike appearance and capability cues" conditions. In other words, if capability-exploiting responses occur even in the "humanlike appearance cues" condition, then this would rule out the strictly one-to-one mapping between cues and responses that the compositional mental model would predict and thereby provide evidence for an integrated mental model. This suggests conducting a different planned contrast comparison, namely between the two "humanlike" conditions and the "computer-like" condition:

H4b. If and only if the user's mental model is integrated, then—within the "high capabilities" conditions—users exhibit more capability-exploiting responses in the "humanlike appearance cues" condition and the "humanlike appearance and capabilities cues" condition than in the "computer-like cues" condition.

4. EXPERIMENTAL SETUP

For our experiment, we created an agent-based system for requesting travel information for the Netherlands Railways (NS). The specific capabilities and cues of the system in the 2×3 between-subjects design are described in Table I and Table II, respectively. The experiment was conducted online with university students from all over the Netherlands. Ninety-two participants (35 male; age $M = 21.8$, SD = 3.55) took part in the experiment, which was conducted entirely over the Internet, with participants logging in from their home computers. Since only the high-capabilities condition allows us to test H3 and H4, 59 participants were assigned to the "high capabilities" condition and only 33 to the "low capabilities" condition. Table III shows the full factorial design and the number of participants in each of the six experimental conditions.

A Wizard-of-Oz technique was used to provide the functionality of the system: users were led to believe that they were interacting with a real system, but actually the experimenter read their inputs and provided the system responses. Participants were asked to sign up for a specific time slot to interact with the system.[2] This prevented the experimenter from becoming overwhelmed or having to stand by 24/7 to wait for participants. All users performed the following predefined tasks using the system[3]:

[2]Participants were recruited from a database that was also used for lab experiments. Participants were thus familiar with the "sign up for a time slot" experience; it did not seem to raise suspicion.

[3]We included four tasks so users would have multiple interactions with the system. Task order was not randomized, because the task did not represent any experimental manipulation. Moreover, all tasks were about equally complex and provided an equal opportunity to give humanlike and capability-exploiting responses.

Table II. A Description of the System Cues Manipulation

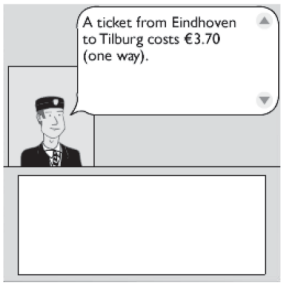
Computer-like cues	Human-like appearance cues	Human-like appearance and capability cues
		
These systems are used as a baseline. They present a logo and a textbox that displays system responses. They display 澱roken ・sentences and a strict sentence structure, which provides users with the feel of computer-style dialogue.	These systems introduce some humanlike, representational cues. They present a humanlike character that 鍍alks ・through a speech bubble. The character uses full sentences, and a more varied sentence structure and wording. These systems do not use capability-implying cues; their appearance and interaction style are just more humanlike.	These systems have the same cues as the 塗umanlike appearance cues・systems, but they also signal certain humanlike capabilities. In their responses, they make extensive use of implicit references, in the form of personal deixis (you, I), temporal and spatial reference (there, then) and deixis (now, here), and nominal references (that trip).

Table III. Number of Participants Per Condition

	Computer-like cues	Humanlike appearance cues	Humanlike appearance and capability cues
Low capabilities	10	11	12
High capabilities	20	19	20

TASK 1: It is Monday morning 11am and you are at Eindhoven station. Your first trip is to Tilburg. Try to find out from which platform the train leaves, and whether you have to switch trains somewhere.

TASK 2: Monday, 11:30am: You arrive in Tilburg, and decide to walk around town for a bit. Before you leave the station, you first look up at what time your train to Leiden leaves. You want to arrive in Leiden at 5pm. What time do you have to leave? What does this trip cost?

TASK 3: You are experiencing some delay due to the evening rush hour, and you arrive in Leiden at 5:30pm. You decide to stay there for the night. Tomorrow you want to be in Roosendaal at 9:45am. What time do you have to leave? What does this trip cost?

TASK 4: Tuesday 9pm. You are walking back to the Roosendaal station. Find out when the next train to Eindhoven leaves, and what the trip will cost.

4.1. Measures

Our hypotheses require us to measure the humanlike and capability-exploiting responses that our participants provide during the interaction, as well as the overall usability of the interaction. Humanlike responses are behaviors that typically occur in interaction between two humans, and that do not occur when a human is interacting with a computer. Brennan [1991] found that participants used more first-person references ("I" and "me") when talking to humans compared to when they were talking

Table IV. Differences in Usability Metrics for Low versus High Capabilities System

	Low capabilities	High capabilities	Significance*	Effect size*
# of requests per task	3.05	1.66	$p < .001$	$r = .78$
Time per task (seconds)	181	106	$p < .001$	$r = .71$
Satisfaction (min. 9, max. 45)	24.48	31.47	$p < .001$	$r = .53$

* Tests are based on linear mixed-effects models, except for satisfaction, which is based on a t-test.

to computers. Researchers have found that participants also used significantly longer sentences when talking to other humans compared to when they were talking to computers [Rosé et al. 2005; Shechtman and Horowitz 2003; Richards et al. 1984]. Thompson [1980] found that participants used more grammatically correct sentences when talking to humans compared to when they were talking to computers. We therefore use the number of personal references, the number of words per request, and the grammatical correctness of the requests (correct sentences versus command-style language) as humanlike responses.

We define capability-exploiting responses as behaviors that exploit the typical capabilities of a human conversation partner; specifically, the linguistic capability to understand implicit references, which is commonly not available in computers. This capability includes understanding "anaphora" (a reference to an earlier question/sentence), "time deixis" (a reference to the current date or time), "space deixis" (a reference to the current place), and "multiple connected questions" (asking for multiple things within one request). Other humans can exploit these capabilities by using "anaphora" (referring to earlier questions), "time deixis" (using words like "now" or "tomorrow" or not indicating a time, thereby implicitly meaning "now"), "space deixis" (using words like "here" or not indicating a place, thereby implicitly meaning "here"), and asking multiple questions at once (asking "how do I get to Amsterdam, and what does it cost?"). We used the presence of such user behavior to measure capability-exploiting responses.

Usability is typically conceived as a multi-dimensional concept, including effectiveness (whether users are able to perform the task with the system or not), efficiency (the amount of time or number of actions users needed to accomplish a task), and satisfaction (a self-reported reflection of the users' feelings with regard to using the system) (International Standards Organization [ISO] 9241-11). As participants performed multiple similar tasks with the system, we added learnability (the time it takes to learn to do a task efficiently or the amount of efficiency that can be gained over time) to this list. We measured usability as follows:

—Discontinuation of the experiment (proxy of ineffectiveness)
—Number of requests a user needs to make and time per task (proxy of efficiency)
—Satisfaction, as measured by the "overall reactions to the software" Section of the Questionnaire for User Interface Satisfaction [Chin et al. 1988]
—Difference in time per task between the first and the last task (proxy of learnability)

5. RESULTS

5.1. Low versus High Capabilities

We first confirm that the system with high capabilities is actually more usable (in terms of efficiency and user satisfaction) than the system with low capabilities (H1). Table IV shows that users needed significantly fewer requests and less time per task in the high compared to the low-capabilities conditions and were also significantly more satisfied, indicating that the highly capable system was indeed more usable (in terms of efficiency and satisfaction) than the less capable system.

Table V. Decrease in Time (Seconds) Needed to Perform the Task between Task 4 and Task 1

	Computer-like cues	Humanlike appearance cues	Humanlike appearance and capability cues
Low capabilities	108.56	40.12	25.89
High capabilities	−6.50	20.42	13.30

5.2. Overestimation

H2 suggests that in the "low capabilities, humanlike appearance cues" and the "low capabilities, humanlike appearance and capability cues" conditions users overestimate the capabilities of the system, resulting in a lower learnability and effectiveness than the "low capabilities, computer-like cues" condition.

Strong evidence for overestimation was found in terms of system effectiveness: Five of the 23 participants interacting with a system with low capabilities but humanlike cues (and none for computer-like cues) prematurely quit the experiment—something that, in our experience with this participant database, rarely happens. Two participants in the "low capabilities, humanlike appearance cues" condition and three participants in the "low capabilities, humanlike appearance and capability cues" condition were not able to adapt their questions to the limited capabilities of the system. Probably frustrated with the numerous error messages they encountered, they either closed their browsers or skipped all remaining tasks to end the experiment. Typically, these users completed none or just a single task before giving up.

Additional evidence for the overestimation effect was found in terms of learnability. Table V shows that within the low-capabilities condition, users of the computer-like interface showed a higher time decrease (learned faster) than users of the humanlike systems.[4] A Factorial ANOVA on the decrease in time used to perform the first task and the last task[5] was performed with capabilities and cues as independent variables. First, a main effect of capabilities showed that across cue conditions, users in the low capabilities conditions exhibited a stronger decrease in time used per task than users in the high capabilities conditions ($M_{low} = 58.2$ seconds versus $M_{high} = 9.3$s, $F(1,83) = 8.56$, $p < .005$, *partial* $\eta^2 = 0.098$). This result indicates that there was less need for learning in the high capabilities system. More importantly, the interaction between capabilities and cues was significant ($F(2,82) = 3.95$, $p < .05$, *partial* $\eta^2 = 0.091$), which indicates that the strongest decrease in time (our measure for learnability) occurred in the "low capabilities, computer-like cues" condition. Indeed, the time decrease in this condition is marginally higher than in the "low capabilities, humanlike appearance cues" and the "low capabilities humanlike appearance and capability cues" conditions (contrast: $F(1,24) = 2.66$, $p = 0.11$, *partial* $\eta^2 = 0.10$), and there is no significant difference between the two "humanlike" conditions (contrast: $F(1,15) = 0.073$, $p = 0.79$). In the "high capabilities" conditions, the time decrease in the "computer-like cues" condition is actually *lower* than in the other two conditions (contrast: $F(1,57) = 4.283$, $p = 0.043$, *partial* $\eta^2 = 0.070$), and there is again no significant difference between the two "humanlike" conditions (contrast: $F(1,37) = 0.325$, $p = 0.57$).

Note that, according to the findings above, overestimation occurred not only in the "humanlike appearance and capabilities cues" condition but also in the "humanlike appearance cues" condition. This suggests that participants tried to exploit inexistent capabilities even if the system gave no explicit (one-to-one) cues of these capabilities.

[4]Participants who quit the experiment are excluded from this analysis.
[5]Since all tasks were about equally difficult, this is an unbiased indicator of learnability.

This already provides evidence for the existence of an integrated mental model, which we will test in more detail in the next sections.

5.3. Existence of an Anthropomorphic Mental Model

To demonstrate the existence of an anthropomorphic mental model, we compare measurements of humanlike and capability-exploiting responses in participants' interaction between the different cue conditions. We can perform these tests only on the high capabilities systems, because the direct feedback (system errors) in the low capabilities systems mitigates these behavioral differences.

H3 suggests that the occurrence of humanlike responses in the "high capabilities" conditions is higher for systems with "humanlike appearance cues" and "humanlike appearance and capability cues" than for systems with "computer-like cues." As dependent measures of humanlike responses, we tested the number of first-person references, total number of words per request, and the grammatical correctness of the requests. See Figure 3 for overall and task-specific numbers of humanlike responses for these three dependent measures. Each of these measures differed in their underlying scale (nominal, frequency, and continuous). To analyze the data, we therefore used mixed (Poisson, linear, and nominal) regressions with random intercepts. As independent factors, we used cue-level and task number. Figure 3 reveals substantial effects of task on most measures, so we used a linear and quadratic task term in the regression to account for these variations. Below we concentrate our discussion on the effects of cue-level, though. Further details of the regressions can be found in the appendix.

Figure 3(a) shows that the number of first-person references used is significantly higher in the "humanlike appearance cues" and "humanlike appearance and capability cues" conditions than in the "computer-like cues" condition (contrast "CuesHum" in the appendix; $\beta = 1.86$, $p = 0.021$). Figure 3(b) shows the same for the number of words per chat request ($\beta = 4.22$, $p = 0.005$), and Figure 3(c) shows the same for grammatical correctness ($\beta = 10.88$, $p < 0.001$). Concluding, we found evidence for the existence of an anthropomorphic mental model (H3), as several humanlike responses were significantly higher when the system had humanlike cues.[6]

H3 also suggests that the occurrence of capability-exploiting responses in the "high capabilities" conditions increases with cue level. A sum measure[7] of the five capability-exploiting responses (see Section 4.1) was taken for each task. Figure 4 shows that the number of capability-exploiting responses is dependent on the cue level and varies by task. Notice that even in the "computer-like cues" condition, the number of capability-exploiting responses observed is unexpectedly high. This is the case because our interface allowed users to type anything into the text field of our system, which for some of the more adventurous participants seemed to be a prompt in itself to try to exploit as many of the system capabilities as possible. More importantly, however, the number of capability-exploiting responses in the humanlike conditions is significantly higher. A linear mixed regression with cue level and task number as independent factors shows that the number of capability-exploiting responses is significantly higher

[6]Note that the difference between "humanlike appearance cues" and "humanlike appearance and capability cues" (contrast "CuesCap" in the appendix") is not significant for any of these humanlike responses.

[7]When a number of items are all indicators of the same behavior, and measured on the same measurement scale, taking a sum measure to create an "index variable" increases the robustness of the measurement, as well as the power of subsequent statistical analyses. In our case, the sparsity of capability-exploiting behaviors is another reason to create an index variable: This variable is more normally distributed than its individual indicators. Running multilevel logistic regressions with the individual indicators provides the following results: anaphora: no effect; date deixis: no effect, time deixis: $p = 0.14$, place deixis: $p = 0.034$, asking multiple questions: $p = 0.009$.

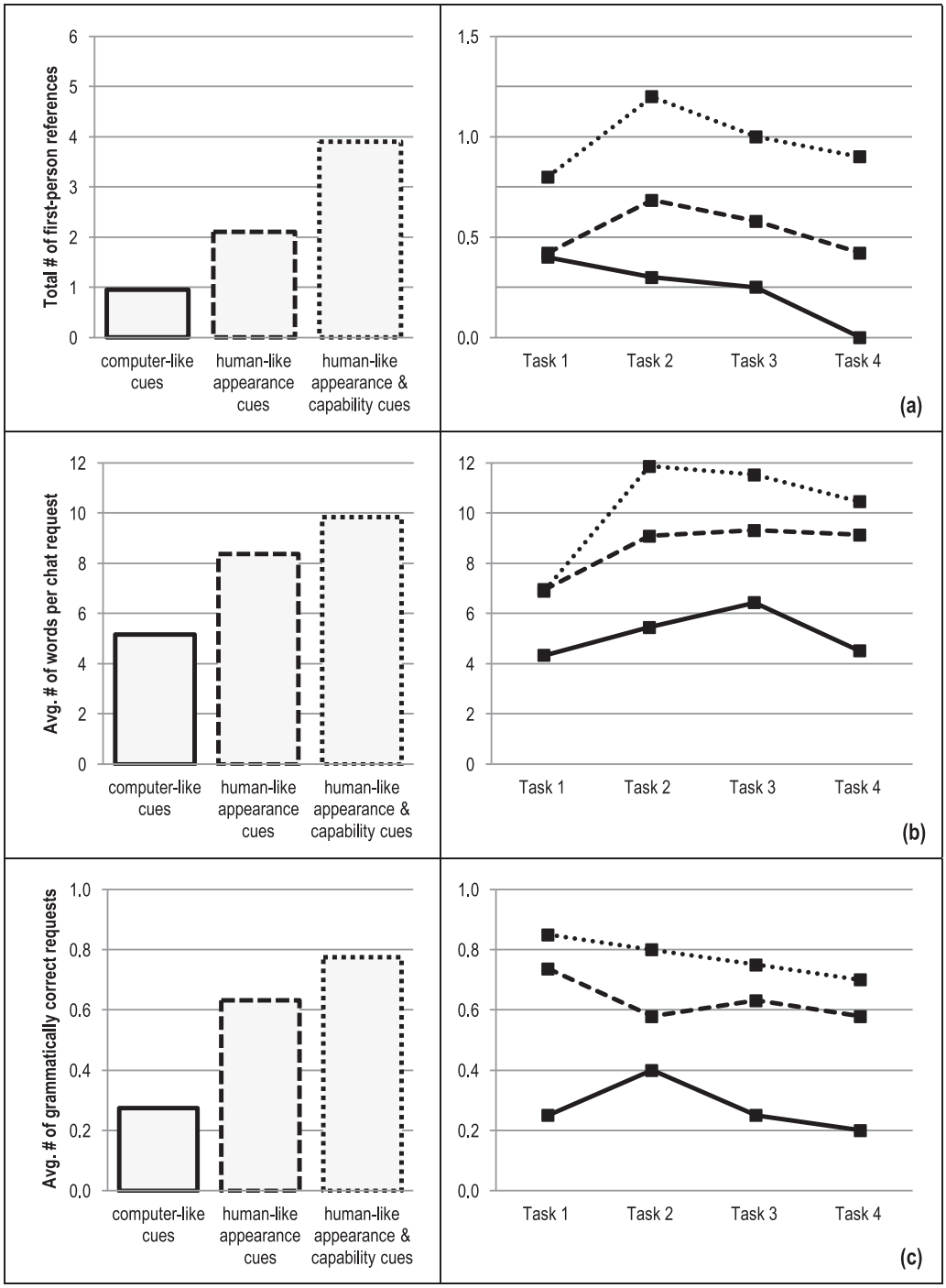


Fig. 3. People use more first-person references (a), longer sentences (b), and better grammatical correctness (c) towards systems with more humanlike cues. The left-side graphs show the humanlike response measures for each cue level. The right-side graphs show how these measures vary per task for each cue level; their vertical axis is measured in the same units as the left-side graphs.

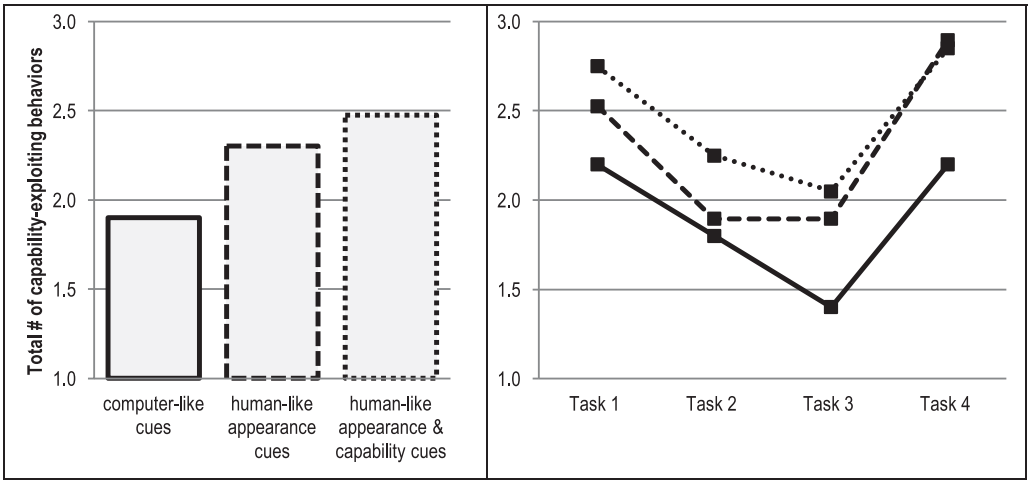


Fig. 4. Capability-exploiting responses (i.e., anaphora, date deixis, time deixis, place deixis, and asking multiple questions) are higher for systems with more humanlike cues. The left-side graph shows the number of capability-exploiting responses for each cue level. The right-side graph shows how this measure varies per task for each cue level.

in the "humanlike appearance cues" and "humanlike appearance and capability cues" conditions than in the "computer-like cues" condition (contrast "CuesHum" in the appendix; $\beta = 0.49$, $p = 0.019$). These results provide further evidence for the existence of an anthropomorphic mental model (H3), as the total number of capability-exploiting responses was significantly higher when the system had humanlike cues.

5.4. Integrated Mental Model

If users have an integrated mental model, then they should show capability-exploiting responses even when the system does not give humanlike capability cues, that is, when it gives humanlike appearance cues only. In such a case, a compositional mental model is ruled out, since it predicts that only capability cues can induce capability-exploiting responses. In other words, the compositional mental model hypothesis (H4a) predicts that the capability-exploiting responses in the "humanlike appearance cues" condition are as low as in the "computer-like cues" condition and only higher in the "humanlike appearance and capability cues" condition (i.e., {low, low, high}); while the integrated mental model hypothesis (H4b) predicts that the capability-exploiting responses in the "humanlike appearance cues" condition are significantly higher than in the "computer-like cues" condition, and equally high as in the "humanlike appearance and capability cues" condition (i.e., {low, high, high}).

As can be seen in Figure 4, the capability-exploiting responses are indeed higher in the "humanlike appearance cues" systems than in the "computer-like cues" condition and almost at the same level as in the "humanlike appearance and capability cues" condition. In line with this, the planned contrast between "computer-like cues" and the other two conditions provided significant results (contrast "CuesHum" in the appendix; $\beta = 1.86$, $p = 0.021$), supporting H4b. Moreover, the contrast between "humanlike appearance cues" and "humanlike appearance and capability cues" was not significant (contrast "CuesCap" in the appendix; $p > 0.05$). H4a would suggest that a gap exist between the first two cue conditions and the last one: "computer-like cues" + "humanlike appearance cues" versus "humanlike appearance and capability cues." This contrast

was not significant ($p > 0.05$). Thus, in line with H4b, there is a significant increase in capability-exploiting responses between the system with "computer-like cues" and the other two conditions, meaning that even with humanlike appearance cues alone, people show more capability-exploiting responses than for computer-like cues. Apparently, these cues can elicit cue-unrelated responses. This suggests that the cues do not elicit a specific response but that they are integrated into a complex mental model that can trigger any kind of response that is in correspondence with this mental model, even outside the range of the provided cues. In other words, this provides evidence that the user's model is integrated rather than compositional.

6. DISCUSSION

We conducted our experiment with the goal of empirically supporting our theory of human-agent interaction, which argues that the user's mental model of an agent-based system is an anthropomorphic construct (C2) of "believed human intelligence" (C1). This theory argues that a system that employs humanlike appearance and capability cues can instill humanlike (C3) and capability-exploiting (C4) responses from the user. Most importantly, it argues that agent-based mental models are *integrated* (C5), which means that any kind of cue can instill any kind of response.

Results from the experiment indeed show that users of humanlike systems show more humanlike and capability-exploiting responses than users of computer-like systems (H3). This results in more first-person references, longer sentences,[8] and more grammatically correct sentences. Users of humanlike systems also (try to) make more use of humanlike capabilities, like implicit references to context and earlier parts of the conversation, and asking multiple questions at the same time.

We argued that some systems may not be able to handle such humanlike and capability-exploiting responses. In these cases, the actual system capabilities would be overestimated by the user, which would result in a reduction in system usability. Results from the experiment show that while low capability systems are in general less usable (H1), *overestimation* occurs when such low capability systems employ humanlike cues, which causes reductions in effectiveness and learnability (H2). Effectiveness is dramatically reduced when cues indicate humanlike intelligence while the system has low capabilities; some participants even gave up entirely when they overestimated the capabilities of the system. Learnability is also better when cues match system capabilities. If a system has limited capabilities, then only with computer-like cues will users eventually learn to use the system in a significantly faster way.

Finally, we hypothesized that the constructed mental model would not be compositional like in "normal" human-computer interaction [Brinkman 2003], but that it would instead be integrated. We set out to provide empirical evidence for this "integratedness" by showing how certain cues could elicit unrelated responses (H4b) rather than just related responses (H4a). Results from the experiment show that "humanlike appearance cues" (i.e., cues not directly related to capabilities) alone increase the number of capability-exploiting responses. This suggests that there is not a one-to-one mapping from cues to responses, but that these responses are instead mediated by an underlying integrated mental model.

[8]One might conjecture that using longer sentences may reduce the efficiency of agent-based interaction. However, participants in the humanlike conditions were not slower than in the computer-like condition. In fact, they were marginally faster ($\beta = -15.26$, $p = 0.054$).

Our results can be explained by the realization that humanlike agents are a *metaphor*. Metaphors are an easy way to instill a complex, integrated mental model, because capabilities are not directly related to cues but can be inferred. This anthropomorphic mental model may, however, cause the user to overestimate the system's capabilities, which reduces the usability of the system. The results of our experiment confirm these statements. In a system using an agent-based interface, the usability is higher when the system is more capable. In our system with low capabilities, however, the usability decreases even further when the agent looks more humanlike. We argue that this effect may be caused by an anthropomorphic mental model: Arguably, the user thinks such a humanlike agent possesses some form of humanlike intelligence and thereby overestimates the actual system capabilities. This presumed mental model instilled by the agent leads the users to respond in a humanlike fashion and to exploit the humanlike capabilities that the user presumably believes the system to have. For usable agent-based interaction, each cue must therefore be delicately tuned to instill the right beliefs; otherwise, it will inadvertently lead users to overestimate the capabilities of the system.

Moreover, the provided cues are integrated into a single mental model that results in a set of responses that do not necessarily need to be directly related to the provided cues. Specifically, capability-exploiting responses can be induced not only by capability cues but even by appearance cues alone. This last finding is most important, since it explains that it is very hard to fix overestimation problems in agents *exactly because* the capabilities are inferred instead of directly related to cues. The integrated mental model makes finding the right set of cues a matter of trial and error, and it might well be the case that there is no possible configuration of cues that will *not* lead to overestimation.

7. CONCLUSION

In this article we have demonstrated that users of humanlike systems anthropomorphize the system and thereby instantly infer a mental model of humanlike intelligence. Our results suggest that agent-based interaction provides system designers with a powerful metaphor, but that this metaphor is too powerful for most of our current systems. These systems cannot live up to the expectations elicited by the agent-based metaphor and thereby suffer from bad usability.

A limitation of our study is that the agent we built somewhat differs from the agents that are in commercial use today. Unlike our agent, most of today's agents use spoken rather than written language to interact with the user. Moreover, unlike today's agents, our agent has a humanlike avatar. Both of these aspects arguably increase the humanlikeness of the agent, and we argue that it is a good thing that today's agents do not employ a humanlike avatar, lest they suffer even more from the effects of overestimation.

Another limitation is that we tested our agent on a culturally uniform sample of university students. While we took care to avoid having engineering students only, our sample is arguably younger and more tech-savvy than the general population. That said, we argue that in the general population the overestimation effect might be stronger: a less tech-savvy audience will be less likely to understand the agent and therefore more likely to take the intentional stance [Bradshaw 1997] and make (unwarranted) inferences about the humanlike capabilities of the system. Future work could further investigate potential cultural and demographic differences in the effect of integrated mental models on human-agent interaction.

Our findings have specific implications for designers of advice-giving systems. As advice-giving systems become more powerful, designers may be tempted to employ a

humanlike agent-based system as an interaction metaphor. This is especially true for advice-giving systems that are inherently *conversational* (cf. [Andersen and Andersen 2002; Holzwarth et al. 2006; McBreen and Jack 2001; Pazzani and Billsus 2002; Semeraro et al. 2008; Spiekermann and Paraschiv 2002]). Our results suggest, though, that although agent-based interaction may seem intuitive for advice-giving systems, designers have to be very careful when introducing a humanlike agent into their systems. While an agent-based system can instantly create a complex mental model (voiding the need for numerous buttons and labels), it is really hard to control this mental model [Keeling et al. 2004]. Because the user's mental model is integrated, it is hard to switch a tiny part of it "on" or "off." Therefore, to prevent usability issues, the system should provide every bit of functionality a user could possibly induce from the agent's appearance. As this is often impossible due to technological constraints, overestimation is likely to occur, and usability may be significantly reduced. It is likely that the other benefits of using agents (attractiveness, fun, social facilitation) do not outweigh these negative effects on usability. Users will arguably more satisfied with a simple but elegant GUI-based interaction method (e.g., "example critiquing" [Chen and Pu 2012]). For systems that already use an agent-based paradigm, it would be advisable to continuously remind users of their limited capabilities. In this sense, designers should follow existing systems like Siri, Cortana, and Google Now, and refrain from giving the agent an ostensive human face. Moreover, a somewhat "robotic" voice could also dampen users' high expectations.

Moreover, system designers should realize that their conventional theories and methods for usability testing may not work on agent-based systems, because they do not adhere to the conventional (compositional) "user's model" as postulated by Norman [1986]. For instance, modular usability tests of agent-based interfaces cannot be integrated, since it is impossible to present part of the functionality without affecting the other parts. Methods like Heuristic Evaluation [Nielsen 1994] that inspect each interface widget and reason about its usability are less suitable for agent-based interfaces, since it is not the widgets that determine the user's mental model.

Given that agent-based mental models are often misaligned with reality, what could the designers of such systems do to better manage users' expectations? As human-agent interactions are similar to human-human interactions, *communication research* can arguably also be applied to agent-based interfaces. For example, established strategies for self-presentation (see Kenrick et al. [2007], chapter 4 for a review) may apply to agents as well [Bailenson and Yee 2005]. Research for instance suggests that when you want to help someone, you should focus on appearing likable rather than competent [Casciaro and Lobo 2005]. One way to do this is to make the agent similar to the user in terms of gender, ethnicity, or personality [Behrend and Thompson 2011; Benbasat et al. 2010; Qiu and Benbasat 2010; Nowak and Rauh 2005; Al-Natour et al. 2006].

Recent advances in speech processing, as demonstrated by Siri, Cortana, and Google Now [Sateli et al. 2012; Lieberman et al. 2014], show that agents definitely have potential. A usable agent-based interface, however, calls for very sophisticated mental model fine-tuning. From an organizational perspective, such fine-tuning projects call for computer scientists and artificial intelligence specialists who can develop smarter systems, social psychologists who know all kinds of self-presentation techniques, designers who can build these techniques into their characters, and usability researchers who can test the correctness of the formed mental model with users. Arguably, only such a multidisciplinary team can bring about a paradigm shift from graphical user interfaces to agent-based interfaces.

APPENDIX

The following tables display the results of the mixed regressions used to measure the effect of cues on humanlike and capability-exploiting responses. Contrast "CuesHum" tests the "humanlike appearance cues" and "humanlike appearance and capability cues" conditions against the "computer-like cues" condition. Contrast "CuesCap" tests the "humanlike appearance and capability cues" condition against the "humanlike appearance cues" condition. All variables are centered, so main effects are interpretable regardless of interaction effects. Moreover, note that all tests are performed in the "high capabilities" conditions only.

Regression A: First Person References

A mixed Poisson regression on the number of first-person references, with task number (both linear and squared) and cues as predictors.

Variable	Estimate	Stand. Error	z-value	p-value
Intercept	−2.37	0.492	−4.81	<0.001
CuesHum	1.86	0.806	2.31	0.021
CuesCap	0.90	0.795	1.13	0.258
Task	−0.14	0.070	−2.05	0.040
CuesHum * Task	0.42	0.193	2.19	0.029
CuesCap * Task	0.02	0.094	0.17	0.863
$Task^2$	0.26	0.122	2.12	0.034
CuesHum * $Task^2$	−0.28	0.324	−0.87	0.385
CuesCap * $Task^2$	−0.08	0.198	−0.39	0.696
Random effect (standard deviation):				
Intercept	2.07		$\chi^2(1) = 115.39$	<.001

Regression B: Words per Chat Request

A mixed Linear regression on the number of words per chat request, with task number (both linear and squared) and cues as predictors.

Variable	Estimate	Stand. Error	t-value	p-value
Intercept	8.00	0.681	11.74	<0.001
CuesHum	4.22	1.439	2.93	0.005
CuesCap	1.56	1.676	0.93	0.355
Task	0.31	0.088	3.52	<0.001
CuesHum * Task	0.35	0.186	1.89	0.060
CuesCap * Task	0.18	0.217	0.82	0.412
$Task^2$	0.95	0.197	4.83	<0.001
CuesHum * $Task^2$	0.29	0.416	0.69	0.491
CuesCap * $Task^2$	0.93	0.484	1.92	0.056
Random-effect (standard deviation):				
Intercept	5.010		$\chi^2(1) = 162.87$	<0.001

Regression C: Grammatical Correctness

A mixed Nominal regression on the grammatical correctness, with task number (both linear and squared) and cues as predictors.

Variable	Estimate	Stand. Error	z-value	p-value
Intercept	1.34	1.194	1.12	0.263
CuesHum	10.88	4.327	2.51	0.012
CuesCap	0.22	2.413	0.09	0.926
Task	−0.43	0.168	−2.55	0.011
CuesHum * Task	−0.22	0.299	−0.74	0.458
CuesCap * Task	−0.08	0.377	−0.21	0.835
Task^2	−0.05	0.295	−0.17	0.866
CuesHum * task^2	−1.09	0.651	−1.68	0.093
CuesCap * task^2	0.45	0.738	−0.61	0.544
Random effect (standard deviation):				
Intercept	6.15		$\chi^2(1) = 120.01$	<0.001

Regression D: Capability-exploiting Responses

A mixed Linear regression on the number of capability-exploiting responses used (connectedness + context of date + context of time + context of place + multiple questions), with task number (both linear and squared) and cues as predictors.

Variable	Estimate	Stand. Error	t-value	p-value
Intercept	2.23	0.095	23.71	<0.001
CuesHum	0.49	0.202	2.42	0.019
CuesCap	0.17	0.235	0.73	0.466
Task	0.01	0.022	0.61	0.539
CuesHum * Task	0.05	0.046	1.09	0.278
CuesCap * Task	−0.05	0.054	−0.94	0.351
Task^2	−0.34	0.048	−7.06	<0.001
CuesHum * Task^2	−0.07	0.103	−0.64	0.520
CuesCap * Task^2	0.08	0.120	0.69	0.491
Random-effect variance and covariance term(s):				
Intercept	0.62		$\chi^2(1) = 46.95$	<0.001

ACKNOWLEDGMENTS

The authors thank Iris van Rooij for her feedback on an early draft of this article.

REFERENCES

Verner Andersen and Hans H. K. Andersen. 2002. *Evaluation of the COGITO System*. Technical Report Risø-R-1363 (EN). Risø National Laboratory, Roskilde, Denmark.

Sameh Al-Natour, Izak Benbasat, and Ronald T. Cenfetelli. 2006. The role of design characteristics in shaping perceptions of similarity: The case of online shopping assistants. *J. Assoc. Inform. Syst.* 7, 12 (Dec. 2006), 821–861.

Jeremy N. Bailenson and Nick Yee. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychol. Sci.* 16, 10 (Oct. 2005), 814–819. DOI: http://dx.doi.org/10.1111/j.1467-9280.2005.01619.x

Tara. S. Behrend and Lori Foster Thompson. 2011. Similarity effects in online training: Effects with computerized trainer agents. *Comput. Hum. Behav.* 27, 3 (May 2011), 1201–1206. DOI: http://dx.doi.org/10.1016/j.chb.2010.12.016

Izak Benbasat, Angelika Dimoka, Paul A. Pavlou, and Lingyun Qiu. 2010. Incorporating social presence in the design of the anthropomorphic interface of recommendation agents: Insights from an fMRI study. In *Proceedings of the 31st International Conference on Information Systems (ICIS'10).* Paper 228.

Timothy Bickmore and Justine Cassell. 2001. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI'01)*. ACM Press, New York, NY, 396–403. DOI: http://dx.doi.org/10.1145/365024.365304

Jeffrey M. Bradshaw. 1997. An introduction to software agents. In *Software Agents*, Jeffrey M. Bradshaw (Ed.). AAAI Press, London, 3–46.

Susan E. Brennan. 1991. Conversation with and through computers. *User Model. User-Adapt. Interact.* 1, 1 (Mar. 1991), 67–86. DOI: http://dx.doi.org/10.1007/BF00158952

Willem-Paul Brinkman. 2003. *Is Usability Compositional?* Ph.D. Dissertation. Technische Universiteit Eindhoven, Eindhoven, The Netherlands. DOI: http://dx.doi.org/10.6100/IR562468

Willem-Paul Brinkman, Reinder Haakma, and Don G. Bouwhuis. 2007. Towards an empirical method of efficiency testing of system parts: A methodological study. *Interact. Comput.* 19, 3 (May 2007), 342–356. DOI: http://dx.doi.org/10.1016/j.intcom.2007.01.002

John M. Carroll and Jean McKendree. 1987. Interface design issues for advice-giving expert systems. *Commun. ACM* 30, 1 (Jan. 1987), 14–32. DOI: http://dx.doi.org/10.1145/7885.7886

Tiziana Casciaro and Miguel Souse Lobo. 2005. Competent jerks, lovable fools, and the formation of social networks. *Harvard Bus. Rev.* 83, 6 (Jun. 2005), 92–99.

Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Model. User-Adapt. Interact.* 22, 1–2 (Apr. 2012), 125–150. DOI: http://dx.doi.org/10.1007/s11257-011-9108-6

John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'88)*. ACM, New York, NY, 213–218. DOI: http://dx.doi.org/10.1145/57167.57203

Matthew G. Chin, Valerie K. Sims, Linda Upham Ellis, Ryan E. Yordon, Bryan R. Clark, Tatiana Ballion, Michael J. Dolezal, Randall Shumaker, and Neal Finkelstein. 2005. Developing an Anthropomorphic Tendencies Scale. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES'05)*, 1266–1268. DOI: http://dx.doi.org/10.1177/154193120504901311

John Cook and Gavriel Salvendy. 1989. Perception of computer dialogue personality: An exploratory study. *Int. J. Man-Mach. Stud.* 31, 6 (Dec. 1989), 717–728. DOI: http://dx.doi.org/10.1016/0020-7373(89)90023-0

Andrew J. Cowell and Kay M. Stanney. 2005. Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *Int. J. Hum.-Comput. Stud.* 62, 2 (Feb. 2005), 281–306. DOI: http://dx.doi.org/10.1016/j.ijhcs.2004.11.008

Kevin H. De Laere, David C. Lundgren, and Steven R. Howe. 1998. The electronic mirror: Human-computer interaction and change in self-appraisals. *Comput. Hum. Behav.* 14, 1 (Jan. 1998), 43–59. DOI: http://dx.doi.org/10.1016/S0747-5632(97)00031-9

Doris M. Dehn and Susanne van Mulken. 2000. The impact of animated interface agents: A review of empirical research. *Int. J. Hum.-Comput. Stud.* 52, 1 (Jan. 2000), 1–22. DOI: http://dx.doi.org/10.1006/ijhc.1999.0325

Daniel C. Dennett. 1987. *The Intentional Stance.* MIT Press, Cambridge, MA.

Anind K. Dey. 2001. Understanding and using context. *Pers. Ubiq. Comput.* 5, 1 (Feb. 2001), 4–7. DOI: http://dx.doi.org/10.1007/s007790170019

Nisha Dharna, Susan L. Thomas, and Jerry B. Weinberg. 2001. The effects of animated characters with human traits on interface usability and user's perception of social interactions. In *Proceedings of the Confrerence on Group Processes in Computer Supported Interaction: Technological and Social Determinism.* Miami University, Oxford, OH.

T. Erickson. 1997. Designing agents as if people mattered. In *Software Agents*, Jeffrey M. Bradshaw (Ed.). AAAI Press, London, 79–96.

Jodi Forlizzi, John Zimmerman, Vince Mancuso, and Sonya Kwak. 2007. How interface agents affect interaction between humans and computers. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces (DPPI'07)*, ACM, New York, NY, 209–221. DOI: http://dx.doi.org/10.1145/1314161.1314180

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English.* Longman Group Ltd., London, UK.

Traci J. Hess, Mark A. Fuller, and John Mathew. 2006. Involvement and decision-making performance with a decision aid: The influence of social multimedia, gender, and playfulness. *J. Manag. Inform. Syst.* 22, 3 (Winter 2005-6), 15–54. DOI: http://dx.doi.org/10.2753/MIS0742-1222220302

Martin Holzwarth, Chris Janiszewski, and Marcus M. Neumann. 2006. The influence of avatars on online Consumer Shopping Behavior. *J. Market.* 70, 4 (Oct. 2006), 19–36. DOI: http://dx.doi.org/10.1509/jmkg.70.4.19

International Standards Organization. 1998. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. Retreived July 2016 from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en.

Yi-Lin Kang, Fiona Nah, and Ah-Hwee Tan. 2012. Investigating intelligent agents in a 3D virtual world. In *Proceedings of 33rd International Conference on Information Systems (ICIS'12)*. Paper 81.

Kathy Keeling, Susan Beatty, Peter McGoldrick, and Linda Macaulay. 2004. Face value? Customer views of appropriate formats for embodied conversational agents (ECAs) in online retailing. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04)*, IEEE Computer Society, Washington, DC, 178–187.

Douglas T. Kenrick, Steven L. Neuberg, and Robert B. Cialdini. 2007. *Social Psychology: Goals in Interaction* (4th ed.). Allyn & Bacon, Boston, MA.

Brenda Laurel. 1990. Interface agents: Metaphors with character. In *The Art of Human-Computer Interface Design*, Brenda Laurel (Ed.). Addison-Wesley, Reading, MA, 67–77.

Steven C. Levinson. 1983. *Pragmatics.* Cambridge University Press, Cambridge, UK.

Henry Lieberman, Christopher Fry, and Elizabeth Rosenzweig. 2014. The New Era of High-Functionality Interfaces. In *Revised Selected Papers of the International Conference on Agents and Artificial Intelligence (ICAART'14)*, Springer International Publishing, Switzerland, 3–10. DOI:http://dx.doi.org/10.1007/978-3-319-25210-0_1

Helen M. McBreen and Mervyn A. Jack. 2001. Evaluating humanoid synthetic agents in e-retail applications. *IEEE Trans. Syst. Man Cybernet. A: Syst. Hum.* 31, 5 (Sep. 2001), 394–405. DOI:http://dx.doi.org/10.1109/3468.952714

Reinhard Moratz and Thora Tenbrink. 2006. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cogn. Comput.* 6, 1, 63–107. DOI:http://dx.doi.org/10.1207/s15427633scc0601_3

Nicolas Negroponte. 1997. Agents: From direct manipulation to delegation. In *Software Agents*, Jeffrey M. Bradshaw (Ed.). AAAI Press, London, 57–66.

R. S. Nickerson. 1976. On conversational interaction with computers. In *Proceedings of the ACM / SIGGRAPH Workshop on User-oriented Design of Interactive Graphics Systems (UODIGS'76)*. ACM Press, New York, NY, 101–113. DOI:http://dx.doi.org/10.1145/1024273.1024286

Jakob Nielsen. 1994. *Usability Engineering.* Morgan Kaufmann, San Francisco, CA.

Donald A. Norman, 1986. Cognitive engineering. In *User Centered System Design: New Perspectives on Human-Computer Interaction*, Donald A. Norman, and S. W. Draper (Eds.), Lawrence Erlbaum, Hillsdale, NJ, 31–61.

Donald A. Norman. 1988. *The Design Of Everyday Things.* Basic Books, New York, NY.

Donald A. Norman. 2013. *The Design Of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, NY.

Kristine L. Nowak. 2004. The influence of anthropomorphism and agency on social judgment in virtual environments. *J. Comput.-Med. Commun.* 9, 2 (Jan. 2004). DOI:http://dx.doi.org/10.1111/j.1083–6101.2004.tb00284.x

Kristine L. Nowak and Christian Rauh. 2005. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *J. Comput.-Med. Commun.* 11, 1 (Nov. 2005), 153–178. DOI:http://dx.doi.org/10.1111/j.1083-6101.2006.tb00308.x

Michael J. Pazzani, and Daniel Billsus. 2002. Adaptive Web Site Agents. *Auton. Agents Multi-Agent Syst.* 5, 2 (Jun. 2002), 205–218. DOI:http://dx.doi.org/10.1023/A:1014849311433

Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *J. Manag. Inform. Syst.* 25, 4 (Spring 2009), 145–182, DOI:http://dx.doi.org/10.2753/MIS0742-1222250405.

Lingyun Qiu and Izak Benbasat. 2010. A study of demographic embodiments of product recommendation agents in electronic commerce. *Int. J. Hum.-Comput. Stud.* 68, 10 (Oct. 2010), 669–688. DOI:http://dx.doi.org/10.1016/j.ijhcs.2010.05.005

Leo R. Quintanar, Charles Crowell, and Patrick J. Moskal. 1987. The interactive computer as a social stimulus in human-computer interactions. In *Social, Ergonomic, and Stress Aspects of Work with Computers*, G. Salvendy, S. Sauter, and J. L. Hurrell (Eds.). Elsevier Science Publishers, Amsterdam.

Byron Reeves and Clifford Nass. 1996. *The Media Equation; How People Treat Computer, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, MA.

Sven Reichel, Ute Ehrlich, André Berton, and Michael Weber. 2014. In-car multi-domain spoken dialogs: A Wizard of Oz study. In *Proceedings of the EACL 2014 Workshop on Dialogue in Motion (DM'14)*. Association for Computational Linguistics, Stroudsburg, PA, 1–9.

M. Richards, and K. Underwood. 1984. How should people and computers speak to one another? *Proceedings of the 1st IFIP Human-Computer Interaction (INTERACT'84).* Elsevier Science, New York, NY, 33–36.

Carolyn Penstein Rosé and Cristen Torrey. 2005. Interactivity and expectation: Eliciting learning oriented behavior with tutorial dialogue systems. In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, Springer-Verlag, Berlin, 323–336. DOI: http://dx.doi.org/10.1007/11555261_28

Bahar Sateli, Gina Cook, and René Witte. 2013. Smarter mobile apps through integrated natural language processing services. In *Proceedings of the 10th International Conference on Mobile Web Information Systems (MobiWIS'13)*, Springer-Verlag, Berlin, 187–202. DOI: http://dx.doi.org/10.1007/978-3-642-40276-0_15

J. Ben Schafer, Joseph Konstan, John Riedl. 1999. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce (EC'99)*. ACM Press, New York, NY, 158–166. DOI: http://dx.doi.org/10.1145/336992.337035

Matthias Scheutz, Rehj Cantrell, and Paul Schermerhorn. 2011. Toward human-like task-based dialogue processing for HRI. *AI Mag.* 32, 4 (Winter 2011), 77–84. DOI: http://dx.doi.org/10.1609/aimag.v32i4.2381

Giovanni Semeraro, Verner Andersen, Hans H. K. Andersen, Marco de Gemmis, and Pasquale Lops. 2008. User profiling and virtual agents: A case study on e-commerce services. *Univers. Access Inform. Soc.* 7, 3 (Sep. 2008), 179–194. DOI: http://dx.doi.org/10.1007/s10209-008-0116-1

Alexander Serenko. 2006. The use of interface agents for email notification in critical incidents. *International J. Hum.-Comput. Stud.* 64, 11 (Nov. 2006), 1084–1098. DOI: http://dx.doi.org/10.1016/j.ijhcs.2006.06.004

Alexander Serenko. 2008. User perceptions and employment of interface agents for email notification: An inductive approach. In *Proceedings of the 14th Americas Conference on Information Systems (AMCIS'08)* Paper 267.

Nicole Shechtman and Leonard 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, New York, NY, 281–288. DOI: http://dx.doi.org/10.1145/642611.642661

Ben Shneiderman. 1981. A note on human factors issues of natural language interaction with database systems. *Inform. Syst.* 6, 2, 125–129. DOI: http://dx.doi.org/10.1016/0306-4379(81)90034-X

Ben Shneiderman. 1997. Direct manipulation versus agents: Paths to predictable, controllable, and comprehensible interfaces. In *Software Agents*, Jeffrey M. Bradshaw (Ed.). AAAI Press, London, 97–106.

Derek Sleeman and John Seely Brown (Eds.) 1982. *Intelligent Tutoring Systems*. Academic Press, London.

Dustin Arthur Smith. 2013. *Generating and Interpreting Referring Expressions in Context*. Ph. D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA. Local System Number: 002302584.

Sarah Spiekermann and Corina Paraschiv. 2002. Motivating human–agent interaction: Transferring insights from behavioral marketing to interface design. *Electron. Commerce Res.* 2, 3 (July 2002), 255–285. DOI: http://dx.doi.org/10.1023/A:1016062632182

M. M. Taylor. 1988. Layered protocol for computer-human dialogue. *Int. J. Man-Mach. Stud.* 28, 23 (Feb.–Mar. 1988), 175–257. DOI: http://dx.doi.org/10.1016/S0020-7373(88)80036-1

Bozena Henisz Thompson, 1980. Linguistic analysis of natural language communication with computers. In *Proceedings of the 8th International Conference on Computational Linguistics (COLING'80)*. Association for Computational Linguistics Stroudsburg, PA, 190–201. DOI: http://dx.doi.org/10.3115/990174.990206

Bob Trott. 1998. Microsoft's paper-clip assistant killed in Denver. *CNN.com*, October 16, 1998. Retrieved July 4, 2016, from http://www.cnn.com/TECH/computing/9810/16/clipdeath.idg/.

Janet H. Walker, Lee Sproull, and R. Subramani. 1994. Using a human face in an interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'94)*. ACM Press, New York, NY, 85–91. DOI: http://dx.doi.org/10.1145/191666.191708

R. C. Williges, B. H. Williges, and J. Elkerton. 1987. Software interface design. In *Handbook of Human Factors*, G. Salvendy (Ed.). John Wiley, New York, NY, 1414–1449.

Terry Winograd. 1972. *Understanding Natural Language.* Academic Press, Orlando, FL.

Kyung-Hyan Yoo and Ulrike Gretzel. 2009. The influence of virtual representatives on recommender system evaluation. In *Proceedings of the 15th Americas Conference on Information Systems (AMCIS'09)*. Paper 533.

Received June 2015; revised May 2016; accepted June 2016