Anthropomorphic Agents as a UI Paradigm: Experimental Findings and a Framework for Research

Technical Report GIT-GVU-02-10
June 2002

Richard Catrambone[1], John Stasko[2], and Jun Xiao[2]
[1]School of Psychology, [2]College of Computing / GVU Center
Georgia Institute of Technology
Atlanta, GA 30332 USA
rc7@prism.gatech.edu, stasko@cc.gatech.edu, junxiao@cc.gatech.edu

Abstract

Research on anthropomorphic agent interfaces has produced widely divergent results. We suggest that this is due to insufficient consideration of key factors that influence the perception and effectiveness of agent-based interfaces. Thus, we propose a framework for studying anthropomorphic agents that can systematize the research. The framework emphasizes features of the agent, the user, and the task the user is performing. Our initial experiment within this framework manipulated the agent's appearance (lifelike versus iconic) and the nature of the user's task (carrying out procedures versus providing opinions). We found that the perception of the agent was strongly influenced by the task while features of the agent that we manipulated had little effect.

Keywords

Anthropomorphic user interfaces, empirical study, agent-based interfaces, evaluation, help systems

1. INTRODUCTION

If you could ask for assistance from a smart, spoken natural language help system, would that be an improvement over an on-line reference manual? Presumably the answer, in most cases, is yes for two reasons. First, the spoken natural language aspect would allow you to speak your questions rather than having to type them. Generally this is a faster approach for most people. Second, the smart aspect would improve the chance of the help system finding the information you want even if you do not state the query using the correct or most appropriate terms.

The state of the art in this style of interface is a human user consultant. Does it matter that the user consultant has a face and that the face can have expressions and convey a personality? Would a face affect you in terms of your comfort and satisfaction with the interaction? Would the presence of a face make the help or advice you receive more persuasive? The answers to such questions have implications for the design of training systems, customer service, kiosks, and many other areas.

Agent-based interfaces, particularly those with an anthropomorphic appearance, are still relatively uncommon.

Human-like assistants who answer questions and perform tasks through conversational, natural language-style dialogs with users contrast the traditional view of computers as enabling tools for functional purposes.

Many people believe that anthropomorphic interfaces have great potential to be beneficial in HCI for a number of reasons. Agents could act as smart assistants, much like travel agents or investment advisors, aiding people in managing the ever-growing amount of information encountered today [16]. Further, a conversational interface appears to be a more natural dialog style in which the user does not have to learn complex command structure and functionality [15].

An anthropomorphic interface could use intonation, gaze patterns, facial expressions, hand gestures, and posture, in addition to words, for conveying information and affect. The human face seems to occupy a privileged position for conveying a great deal of information, including relatively subtle information, efficiently [10]. Finally, anthropomorphic interfaces could make a computer more human-like, engaging, entertaining, approachable, and understandable to the user, thus harboring potential to build trust and establish relationships with users, and make them feel more comfortable with computers.

These potential advantages are balanced by strong negatives. Anthropomorphic agent interfaces are viewed by some researchers as being impractical and inappropriate. Current speech recognition, natural language understanding, and learning capabilities of computers still fall far short of any human assistant.

More specifically, it has been proposed that agent systems disempower users by clouding issues such as who is responsible for a system' s actions [14]. Others feel that user interfaces are more beneficial when they clearly reflect the commands available to a user and present the objects that a user can act upon [26]. Furthermore, critics argue that agent interfaces may mislead both users and designers, increase user anxiety, reduce user control, undermine user responsibility, and destroy a user's sense of accomplishment [25]. Many current anthropomorphic or personified interfaces are viewed as being annoying, silly characters who hinder rather than enhance productivity (e.g., the Microsoft Office Paper Clip).

Although strong opinions have been voiced on both sides of this issue, relatively little careful empirical evaluation on anthropomorphic interfaces has been performed, and the results from this research have been contradictory or equivocal [1]. Erickson states: “First it must be acknowledged that in spite of the popularity of the agent metaphor, there is remarkably little research on how people react to agents.”[9] Shneiderman has echoed the need for more study: “Please, please, please do your studies—whether they are controlled scientific experiments, usability studies, or simply observations, and get past the wishful thinking and be a scientist and report on real users doing real tasks with these systems.”[25]

Even more specifically, other researchers have laid out a research agenda for the area. Laurel, in a 1990 article [15], discusses one component of such an agenda:

> In the theoretical arena, work must proceed on the analysis of user needs and preferences vis-a-vis applications and environments. What are the qualities of a task that make it a good candidate for an agent-like interface? What kinds of users will want them, and what are the differences among potential user populations? How might interface agents affect the working styles, expectations, productivity, knowledge, and personal power of those who use them?
>
> In terms of design, the meatiest problem is developing criteria that will allow us to elect the appropriate set of traits for a given agent—traits that can form coherent characters, provide useful cues to users, and give rise to all of the necessary and appropriate actions in a given context.

In a recent article reviewing many empirical studies on the impact of animated interface agents, Dehn and van Mulken state, “Systematic and well-conducted investigations on animated interface agents and their effect on the user's attitude and performance are still scarce. There appears to be a need for future studies that (1) overcome the methodological shortcomings of many existing studies and (2) take a more fine-grained perspective on the effect of employing animated agents on the user's motivation and cognition” [7].

Our goal is to develop a framework to systematically evaluate and understand the autonomous agent as a user interface paradigm. The present paper describes an experiment that explores two issues within this framework. The first is whether the degree to which an interface agent is anthropomorphic has a measurable effect on users. Note that anthropomorphism is not a dichotomy but rather a continuum. One can think of interfaces with full fidelity video or 3D images of people to more caricature-style characters to 2D cartoons of people or personified characters such as dogs or toasters.

The second issue is to what extent the nature of the task will influence a user's perception of an agent. Some tasks might be more likely to induce a user to imbue the agent with human-like qualities (such as if the user had to engage the agent in a debate) while other tasks might lead the user to view the agent simply as a reference tool (e.g., for providing reminders of keystrokes for a software application) with no “individuality.”

We used an existing agent tool (from Haptek Corp.), rather than one of our own creation. This detachment from the system-building technology of agents can be viewed as an advantage because we have neither an agenda nor a system that we seek to prove useful, enjoyable, and productive. This helps us examine these styles of interfaces in an objective manner.

2. RELATED WORK

An important event in the history of anthropomorphic agent research was Apple Computer's late 1980's production of the video titled *Knowledge Navigator*. The video showed a university faculty member in his office interacting with his computer. The computer's chief interface metaphor was an anthropomorphic, 3-D “talking head,” a computerized assistant named Phil with whom the professor interacted via natural language. Phil answered questions directed to him and took the initiative in carrying out important actions that would benefit the professor. *Knowledge Navigator* was a thought-provoking film, and it has been the source of much discussion in the HCI community since its production, gaining its share of both praise and criticism.

In the past 10-15 years, quite a bit of effort has been made toward building autonomous agents like “Phil” in the *Knowledge Navigator* video. Of course, these efforts have been initial steps toward that advanced vision, and much work remains.

Some of the most noteworthy projects on building agent interfaces have been the Guides project from Apple [22], the Persona project (with Peedy the Parrot) from Microsoft [1], Rea from MIT [3], work from FX Palo Alto Labs [23,5], the PPP persona and AiA projects from DFKI in Germany [1], Baldi from CSLU [18] and Gandalf from MIT [29]. A good source for articles describing further agent projects as well as references to other efforts is [4].

It is probably safe to say that the most widely used system in this space, at least in spirit, is the Paper Clip assistant from Microsoft. The Paper Clip's pervasive presence in Microsoft Office tools such as Word or PowerPoint has influenced many people's opinions of user interface agents like this, often negatively.

Another recent boom in anthropomorphic/personified characters has occurred on the World Wide Web. Sites seek to provide human-like hosts or guides that will assist a person browsing web pages or that will read news much like an evening TV anchor. Noteworthy sites utilizing or providing such capabilities are Virtual Personalities, Inc. (www.vperson.com), Artificial Life (www.artificial-life.com), FaceWorks (interface.digital.com/overview/default.htm), Ananova (www.ananova.com), and Haptek (www.haptek.com).

While quite a bit of research into building anthropomorphic agents has occurred, relatively little evaluative empirical study has accompanied the system building. Cassell recently commented, “To date, few researchers have empirically investigated embodied interfaces, and their results have been equivocal” [3].

A few studies have revealed that anthropomorphic agents are attention-grabbing and people make natural assumptions about the intelligence and abilities of those agents. King and Ohya found that a dynamic 3D human form whose eyes blinked was rated more intelligent than any other form, including non-blinking 3D forms, caricatures, and geometric shapes [11].

One common trend discovered in studies is that anthropomorphic interfaces appear to command people's attention, both in positive and negative senses. Takeuchi and Nagao created conversational style interaction systems that allowed corresponding facial displays to be included or omitted [27,28]. According to their metrics, the conversations with a face present were more “successful.” Across two experiments they found that the presence of a face provided important extra conversational cues, but that this also required more effort from the human interacting with the system and sometimes served as a distraction.

Other studies have shown that the attention garnered by an anthropomorphic interface had a more positive, desired effect. Walker, Sproull, and Subramani found that people who interacted with a talking face spent more time on an on-line questionnaire, made fewer mistakes, and wrote more comments than those who answered a text questionnaire [30]. Koda created a Web-based poker game in which a human user could compete with other personified computer characters including a realistic image, cartoon male and female characters, a smiley face, no face, and a dog [12]. She gathered data on people's subjective impressions of the characters and found that people's impressions of a character were different in a task context than in isolation and were strongly influenced by perceived agent competence.

An influential body of related work is that of Nass, Reeves and their students at Stanford. Their efforts focus on the study of “Computers as Social Actors.” They have conducted a number of experiments that examined how people react to computer systems and applications that have certain personified characteristics [20,21,24]. Their chief finding is that people interact with and characterize computer systems in a social manner, much as they do with other people. This occurs even when the participants know that it is only a computer with which they are interacting. More specifically, Nass and Reeves found that existing, accepted sociological principles (e.g., individuals with similar personalities tend to get along better than do those with different personalities) apply even when one of the two participants is a machine.

The studies cited above, and others, suggest that people are inclined to attribute human-like characteristics to agents and that a variety of factors might influence how positively the agents are viewed. For a more extensive review of the studies discussed above as well as others, the interested reader can see [7]. Part of the conclusion to that article nicely summarizes the findings in this area to date:

> ...it would be oversimplifying to conclude that there is certainly no advantage in using animated agents. In the case of user's attitudes towards the system, some positive effects have been established. [...] Unfortunately though, the present studies do not enable us to make clear predictions as to what types of animations employed in what type of domain will result in positive attitudes towards the system.
>
> For changes in the user's behaviour, the existing evidence is even less compelling. In the light of the small number of studies investigating performance data, though, it would be premature to draw

conclusions with regard to the effects of animated agents that are too negatively biased. Furthermore, the lack of effects that can be uniquely attributed to the use of animated agents is often due to experimental confoundings rather than to the fact that there is no effect.

3. A FRAMEWORK FOR RESEARCH ON ANTHROPOMORPHIC INTERFACE AGENTS

To effectively and systematically investigate the use of anthropomorphic interface agents, one needs to consider the key factors that will affect the usefulness of such interfaces. We propose an investigative framework composed of three key components: characteristics of the user, attributes of the agent, and the task being performed.

We believe that serious empirical study in this area must systematically address each of these factors and understand how it affects human users. Below, we provide examples of individual variables within each factor that could potentially influence user performance and impressions.

Factor 1: Features of the User

Potential users vary, of course, in many ways. However, there are certain features that may be quite likely to affect how useful a user finds an agent. These features include:

Personality: Researchers have identified what are referred to as the "Big Five" traits that seem to be quite useful in describing human personalities: extraversion, openness, agreeableness, neuroticism, and conscientiousness (e.g.[19]). While any such breakdown is debatable, it seems reasonable to examine whether users' positions on these, or other, trait dimensions is predictive of how they will respond to agents.

Background knowledge: A user who has a good deal of background knowledge in a domain might prefer an agent that is reactive and that the user can call upon when he or she needs some low-level bit of information or has a low-level task that needs to be done. Conversely, a user who is learning how to carry out tasks in a particular domain might welcome strategy advice from an agent, particularly if the agent can analyze the strategy and provide reasons for why the strategy might be altered.

Capability: The ability of a user to cognitively understand the causes of agents' actions as well as the planning capacity of the user for problem solving may vary greatly across individuals. Other non-cognitive abilities may also need to be considered. For example, in order to develop interactive learning tools for language training with profoundly deaf children, visible speech instructions are crucial [18].

Goal: Users who intend to get a solution of good quality may evaluate the usefulness of agents based on the accuracy and completeness of the agent's help, whereas users who intend to get a quick reference may evaluate the usefulness of agents based on the completion time and efficiency of the operation. Other indicators such as whether the user feels comfortable with the agent and how engaging the interaction is may be applicable in other situations such as learning and entertainment.

Psychological States: Users' moods and emotional states have both positive and negative impact on their attitude and behavior towards agents in a conscious and subconscious manner. Comforting words from an agent may be valued when the user is struggling with a math puzzle overnight, whereas surprises from an agent may not be appreciated if the user is rushing to meet a deadline.

Gender: Although background knowledge and the Big Five personality measures are likely to account for much of the user-determined usefulness of agents, it is also possible that gender will play a role. There has been some research suggesting gender differences in advice-taking, so it seems prudent to consider gender effects on evaluations of agents.

Other variables: Other user-related variables include age and computer experience.

Factor 2: Features of the Agent

Like users, agents can vary on a wide variety of features. These features include:

Visual Appearance: Empirical evidence provided by Dryer suggests that rounder shapes, bigger faces, and happier expressions are perceived by humans as extraverted and agreeable, while bold colors, big bodies, and erect posture characters are perceived by human as extraverted and disagreeable [8]. In other words, visual stimuli may influence users' perception of agents' personalities and should be carefully chosen.

Fidelity: Earlier studies suggest that more realistic-appearing, 3D human representations are perceived as being more intelligent, which could be viewed positively or negatively. Furthermore, realistic-appearing agents are more difficult to implement, so if user performance is improved by the presence of an agent, but does not vary according to appearance, simpler caricature style characters would be advantageous.

Expressiveness: Within realistic-appearing agents, we might vary the level of facial expressions, gestures, emotions, and movements of a particular character. Animated, expressive agents again may be viewed as more realistic and intelligent, but they might also unduly draw the viewer' s attention and thus be distracting and annoying.

Personality: A further important component of an agent's profile is its personality. Should it be a dominant expert or humble servant? Should we adapt the personality of the agent according to the preferences of different users? As we have mentioned before, design decisions on the agent's personality should be made consistently with other characteristics of the agent, such as appearance.

Presence: Is an agent's face always present on the screen or does the agent only appear when it is engaged in a dialog by the user? One might hypothesize that an ever-present agent would make users uneasy by producing an effect of being watched or evaluated all the time.

Role: Should an agent act as a partner in the task or should it contribute only in clearly specified ways? For instance, an agent might be able to offer strategy guidance for design tasks. Alternatively, it might provide only lower-level procedural "how to" information.

Initiative: Related to the "role" dimension is the degree to which an agent initiates interactions. Should it proactively make suggestions and offer guidance or should it only respond when directly addressed? A proactive agent might be viewed as being "pushy" and might bother users, or it could be viewed as being extremely helpful and intelligent if it acts in situations in which the user is unsure of how to proceed or is so confused that he or she is unable to form a coherent help request.

Speech quality: Does the quality of the agent' s speech affect user impressions of the agent? We speculate that poor quality spoken output might negatively influence user views of an agent. Research on user perceptions of speech quality already exists [13], and that work can provide guidance in designing our agent experiments.

Other variables: Other agent-related variables to consider are "gender", competence, and integration with the accompanying software application or environment.

Factor 3: Features of the Task

Tasks can also vary in many different ways. Some tasks can be opinion-like (e.g., choosing what to bring on a trip) while others are more objective (e.g., solving a puzzle) in terms of assessing the quality of a solution. Some involve a good deal of high-level planning (e.g., writing a talk) while others are more rote (e.g., changing boldface words into italics). Tasks must be classified along some or all of the dimensions listed below:

Intent: The user could have a learning goal or alternatively may be carrying out a set of steps in a familiar domain. In the latter, the user might need help with low-level details whereas in the former the user is looking for guidance as to the structure of the domain.

Objectiveness: The situation might be an opinion-based one in which the user is seeking advice and recommendations on some topic (e.g., which items to pack for a trip to Europe). Alternatively, the user might be carrying out an objective task such as acquiring facts (e.g., finding the keystroke combination for a particular command in a software application).

Domain: The domain in which the user is working (e.g., editing a paper vs. building a garage) might matter even if all other relevant features (e.g., objectiveness, intent) are held constant.

Focus/Context: An agent’s assistance may be directly involved with the primary task upon which a user is engaged. On the other hand, agents might be helpful with “side” tasks such as looking up a phone number quickly while a user attends to some other primary task. Would people perceive an agent as being more useful in one of the scenarios compared to the other?

Timing: While some tasks, such as monitoring events, require regular or constant attention, other tasks such as guiding a presentation require some degree of cooperation between the agent and the user. Some tasks may have a significant delay between their initiation and completion while in other tasks the delay of an agent’s actions may arouse user’s suspicion.

Other variables: Other task-related variables to consider are duration, consequences of the quality of task performance, and environment.

The number of variables within each factor is certainly larger than the number we identify here. No doubt these factors will also interact. For instance, a novice attempting to carry out a task in a particular domain might welcome proactive comments/advice from an agent while someone with more experience could get annoyed; these reactions might be reversed in another domain or task.

With respect to measuring the usefulness of an agent, we have to consider which dependent measures are most appropriate. Towards the more objective end, a user's performance on a task in terms of accuracy and time--when such measures are meaningful--can give one indication of usefulness. Thus, time and errors would be appropriate measures for a text-editing task. Towards the more subjective end, a user is likely to have a number of affective reactions to an agent. These reactions might manifest themselves in terms of how much users liked the agent, how intrusive they found the agent, how they perceived the agent's personality, and how willing they are to use the agent in the future. We can certainly assess a user's liking and satisfaction towards an agent, but if the user can carry out the tasks more effectively with the agent, then how important are liking and satisfaction? On the other hand, long-term use of an agent might be predicted by liking and satisfaction.

The likelihood of a user following an agent's advice might be another interesting measure of the usefulness of an agent. While advice-following would certainly be at least partly a function of the quality of the advice, it will also be impacted by how the user feels about the agent (how many children ignore the advice of their parents merely because it is the parents giving the advice?).

4. EXPERIMENT

4.1 Overview

One fundamental issue in the quality of agent interfaces is competence [17]. It appears obvious that perceptions of anthropomorphic agent interfaces will be strongly influenced by the competence of the supporting software system and the quality of the replies and suggestions made by the agent. While a set of experiments could examine how differing levels of competence affect user performance and impression, we chose to factor out competence as an influence in this particular study. If our experiments uncover that people's performance is not enhanced and they dislike anthropomorphic user interfaces even though the system is competent, then that is an important and strong result that other researchers and developers need to understand. To remove competence as a factor, we employed a “Wizard of Oz” [6] experimental methodology (described below).

The experiment manipulated the agent appearance and the task objectiveness variables because prior work and our framework suggest they seemed likely candidates to have an affect on the perception of agents. Usefulness was evaluated via both the performance and satisfaction dimensions. We hypothesized that user reactions to the agent would vary as a function of the objectiveness of task. A task that required the user to debate the merits of his or her opinion (about items to pack on a trip) might lead the user to feel the agent had more of a personality (for good or for bad) compared to a task in which the user made use of the agent more as a reference tool (i.e., reminding the user of keystroke commands for a text editor). We also hypothesized that users might find the agent to be more useful in its role as a reference source rather than as an entity that provides opinions. Finally, we expected that the more life-like the agent appeared, the more likely the user might be to ascribe qualities such as personality and intelligence to the agent, but objective performance would likely not be affected by appearance.

4.2 Participants

Thirty-nine undergraduates participated for course credit and were randomly assigned to conditions. Participants had a variety of majors and computer-experience backgrounds.

4.3 Procedure and Design

Participants were run individually using a computer equipped with a microphone and speaker. An experimenter read an introductory script that was identical for all participants. Participants then began the first of two tasks. The two tasks were a travel task and an editing task. The travel task was chosen to be a type of creative, opinion-based task in which interacting with an agent might be viewed as an opportunity to think more deeply about the task by discussing points of view about the importance of travel items. The editing task was chosen to represent an opportunity to use an agent primarily as a reference source rather than as a guide or teacher.

The travel task involved a hypothetical situation in which the participant had a friend who was flying overseas on his first international trip. The task was to recommend six items for the person to take with him from a pool of 12 items and to rank the six items in order of importance. This task was similar to the desert island survival problem used in studies by Nass [20].

After the participant did the initial ranking using a simple software interface, a computer agent who supposedly had knowledge about international trips appeared. The agent made a predefined set of suggestions in which it recommended changing the rankings of four of the six choices and it agreed with the ranking of two other items. For example, the agent first suggested promoting the person’s fourth item to the first position, demoting the first item but keeping it in the top six. The agent explained the reasoning for its suggestion at every stage and asked the participant what he or she thought about the suggestion. After the participant responded to the agent's comment on a particular item, the agent would say one of several conversational conventions (e.g., “OK, let's continue”) so that it could move on to the next suggestion. After the agent finished providing feedback on the rankings, the original rankings were displayed on the screen and the participant was given the opportunity to change the rankings. After doing the re-ranking, participants filled out a questionnaire about the agent and were asked a few questions about the agent and related issues by the experimenter (a subset of which are mentioned in the results section).

The editing task required participants to use an unfamiliar text editor to modify an existing document by making a set of prescribed changes to the document. Participants first viewed a short video that described the various functions (e.g., copy, paste) and the specific key combinations needed to issue the commands. Participants were then shown a marked-up document that required a set of changes such as deletions, insertions, and moves, and they were instructed that if at any time they could not remember the keystrokes for a particular function, they could ask the agent for help. Pilot testing was conducted to ensure that the tasks were of appropriate difficulty and that the number of commands was sufficiently large so that participants would be likely to need to ask the agent for help. After completing the editing tasks, participants again filled out a questionnaire about the agent and answered questions from the experimenter.

As mentioned earlier, the agent was controlled through a Wizard of Oz technique. One experimenter was in the room with the participant to introduce the experimental materials, and a second experimenter was in an adjacent room, monitoring the questions and responses made by the participant. The second experimenter insured that the agent responded in a consistent, predefined manner using a prepared set of replies.

Two between-subjects variables were manipulated: type of agent (animated, stiff, iconic) and task order (travel task then editing task or vice versa). Figure 1 shows the face of the agent in the animated and stiff conditions. The animated agent was a 3D, female appearance (though somewhat androgynous) that blinked, moved its head, and produced certain facial expressions in addition to moving its mouth in synchronization with the synthesized voice. The stiff agent had the same face as the animated agent but moved only its mouth. The iconic agent (see the right side of Figure 1) was a light-bulb icon that had arrows appear whenever it spoke.

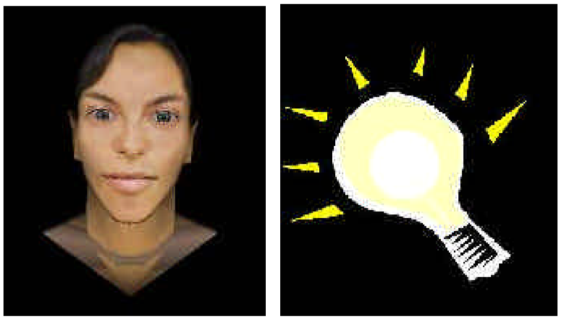

Figure 1: Appearance of Agent in Animated and Stiff Conditions (left) and Iconic Condition (right)

One design issue about the pilot study should be flagged here. Although our key task manipulation was the “objectiveness” of the task (i.e., the travel task being less objective and the editing task being more objective), the nature of the agent also was varied as a function of the task. The agent was completely reactive in the editing task; it provided information only when requested. However, in the travel task the agent provided feedback regardless of the participant's desire and even attempted to make a small joke about going on the trip. A cleaner version of the experiment would have been to hold the “nature” of the agent constant across the tasks. We allowed this confounding to occur here because were interested in getting participants' reactions to certain human-like attributes of the agent but did not have the resources to run the additional conditions that would have been required to completely cross this factor with the task and appearance manipulations. In future work we plan to systematically investigate this reactive/proactive dimension.

4.4 Measures

Both objective and subjective measures were used. One objective measure was, for the travel task, whether participants changed their rankings as a function of the agent's feedback. For the editing tasks we measured how long it took participants to complete the tasks. The primary subjective variables in the experiment were the responses to the individual items in the questionnaires and the answers to the questions posed by the experimenter. The questionnaire items used a five-point Likert scale (1 = strongly agree, 5 = strongly disagree) that addressed a number of qualities of the agent (see Table 2). The questions posed by the experimenter were open-ended and provided participants an opportunity to give their impressions about the agent's personality, helpfulness, and intelligence.

5. RESULTS

In the data analyses we found that the task order manipulation did not have an effect, so in the interest of simplicity we will collapse across that factor in the presentation and discussion of the results.

5.1 Performance Measures

With respect to more objective measures, Table 1 shows that participants were more likely to change the rankings of items that the agent disagreed with compared to items that the agent agreed with, $F(1, 36) = 38.48$, $MSE = .07$, $p < .0001$). There was no effect of type of agent, $F(2, 36) = 0.9$, $p = .42$. Finer-grained analyses of how much the rankings changed could be conducted, but given that changes are not independent, such additional analyses might not have much validity.

The time (in seconds) to do the editing task did not differ as a function of agent (animated: 714.8, stiff: 568.7, iconic: 671.1); $F(2, 31) = 1.78$, $MSE = 37637.22$, $p = .19$ (5 participants failed to complete or chose not to do the editing tasks).

Table 1: Proportion of Travel Items with Changed Rankings as a Function of Type of Agent and Agent Advice (n = number of participants)

	Animated (*n* =14)	**Stiff** (*n* =12)	**Iconic** (*n* =13)
Agent Suggested Change	.82	.90	.77
Agent Concurred with Rank	.57	.43	.38

5.2 Questionnaire Responses

Table 2 shows the mean responses to the questionnaire items for the different agent conditions after the travel and editing tasks (there were 5 participants who did not do both tasks and they are excluded from Table 2). There was no effect of agent type for any of the questions. For two of the items, worthwhile and intrusive, there was an effect of task (worthwhile: *F*(1, 31) = 15.68, *MSE* = .45, *p* = .0004; intrusive: *F*(1, 31) = 20.28, *MSE* = .23, *p* = .0001). The agent was rated more worthwhile and less intrusive after the editing task compared to the travel task. These results make sense. First, the editing task required most participants to rely heavily on the agent to remind them of commands, thus making the agent seem worthwhile. Second, the uninvited critique of participants' rankings of travel items could certainly have seemed particularly intrusive.

While group differences did not exist on most of the questionnaire items, it is interesting to note that for most of the items, the average response tended to be in the positive direction. Participants felt, on average, that the agent was useful.

5.3 Interview Responses

While participants made a number of interesting and insightful comments about the agent in response to questions from the experimenter, a simple tally of responses shows reactions to the agent that again vary as a function of task. Table 3 shows that virtually all participants found the agent helpful for both tasks. Participants were much less likely to consider the agent to have a personality after doing the editing tasks compared to the travel task. This makes sense because the agent was merely providing subjects with information on commands in the editing task. In the travel task the agent expressed its “opinions” and made comments about how it wished it was going on the trip and how it recognized that it and the participant might disagree on certain rankings.

Table 2: Responses to Questionnaire Items as a Function of Type of Agent and Task

	Animated (*n*=12)		**Stiff (*n* =12)**		**Iconic (*n* =10)**		**AVG**
Agent was…	Travel	Edit	Travel	Edit	Travel	Edit	Travel/Edit
Worthwhile	2.50	1.58	2.25	1.42	2.30	2.10	2.35/1.57
Intrusive	2.83	3.50	3.50	4.00	3.40	3.80	3.24/3.76
Friendly	2.67	2.67	2.42	2.50	2.40	2.80	2.50/2.65
Annoying	3.25	3.33	2.83	3.25	3.20	3.80	3.09/3.44
Intelligent	2.58	2.92	2.58	2.50	2.40	2.70	2.53/2.71
Cold	3.25	3.08	3.00	2.67	3.70	3.30	3.29/3.00
Agent has clear voice	2.33	2.58	2.58	2.33	2.50	2.40	2.47/2.44
Enjoyed interacting with Agent	3.08	3.17	2.75	2.83	2.70	2.90	2.85/2.97
Agent helped with task	2.25	1.50	1.67	1.50	2.00	2.30	1.97/1.74
Like to have agent	2.83	2.67	2.58	2.33	2.20	2.40	2.56/2.47

Note: Responses were on a scale from 1 (strongly agree) to 5 (strongly disagree). Not all participants answered questions for both tasks; the table contains data only from participants who answered the questionnaires after both tasks.

Table 3: Number of Participants Who Responded “Yes” to Interview Questions as a Function of Agent and Task

	Animated		**Stiff**		**Iconic**	
	Travel	**Edit**	**Travel**	**Edit**	**Travel**	**Edit**
Thought the agent was helpful	14/14	11/11	12/12	11/11	11/12	10/10
Thought the agent had a personality	7/14	1/12	5/12	0/12	7/12	0/10

Thought the agent was intelligent	11/14	6/11	8/11	6/11	10/12	6/8

Note: Due to problems with the videotape, responses were not recorded for all participants for all questions.

Finally, it is worth noting that in general the agent was perceived as more intelligent after the travel task than after the editing task. At one level this seems odd because the agent had all the answers for the editing task. However, as demonstrated by some participants' comments, the agent was perceived as very limited in the editing task; it knew about editing commands and probably little else (despite the fact that it also appeared to understand spoken language!). In the travel task though it presumably gave the impression of having sufficiently deep knowledge about travel such that it could give feedback on the importance of various items one might take on a trip. While some of the participants' responses to the agent indicated that they disagreed with its suggestions, they appeared to believe that the suggestions were at least thoughtful.

5.4 Observations

In addition to the results reported above, we learned a great deal by observing participants' behaviors and responses in the sessions. One key question we had was how would the participants interact with the agent in the two different tasks. In the editing task, participants seemed very comfortable asking the agent for assistance. Participants requested help an average of 6.5 times. However, in the travel task participants seemed reluctant to engage the agent in a dialog. Only a few replied with more than a few words when the agent engaged them. There was clearly awkwardness to the interaction.

We also noted interesting participant reactions with respect to four key attributes of the agent: its appearance, speech, intelligence, and social skills/personality. We will address these individually below.

Overall, not many of the participants made comments about the agent's appearance. Two of the 13 participants who saw the iconic agent suggested that a more realistic, video image of a person be used. Three of the 12 participants who saw the stiff agent commented about its awkwardness, one noting, "It's creepy that it didn't blink." Another added, "The face didn't move much, didn't sound like a real person, didn't look like a real person." Of the participants seeing the animated agent, a few commented about the life-like appearance, one stating, "[It] reminded me of HAL. It was eerie how lifelike it was."

One striking difference in behavior in the interviews was whether a person referred to the agent using words such as "agent" or "it," versus the gender pronouns "she," "her," "he," or "him." Eleven of the 39 participants used the gender pronouns. When asked after the travel task if the agent was intelligent, one participant responded, "Yes, because she said she was an experienced traveler, so I'm pretty sure I can trust what she was telling me." This behavior reinforced the notion of how people often treat computers as social actors [21].

Of the 11 participants using the gender pronouns, five saw the animated agent, four saw the stiff agent, and two saw the iconic agent. Thus, it appears that the 3D human-like appearance did promote this reaction to some degree.

The study participants included 15 women and 24 men. Curiously, eight of the 11 participants who used the gender pronouns were women and only three were men. Thus, over half the women in the study referred to the agent this way and only 13% of the men did so. This gender-based difference deserves further inquiry in future experiments.

The agent's (poor) speech quality was mentioned by 22 of the 39 participants, usually when they were asked to suggest improvements to the agent. (Note that we used the DecTalk speech generation system.) Two primary different negative comments were voiced frequently: the voice was choppy, not continuous, and the voice was cold, not having a personality.

As mentioned above, across both tasks nearly all the participants answered affirmatively when asked if the agent was helpful. In fact, only one participant (in the iconic travel condition) felt it was not helpful. One participant's comment after the travel task was typical, "The points it made were very valid. It brought up things I hadn't thought of." Another participant stated after the editing task: "[Asking about a command] was easier than looking it up in a help window."

Fewer, but still a majority, of the participants answered affirmatively when asked if the agent was intelligent. Most noteworthy here was not so much that fact, but the personal criteria used by the participants to characterize “intelligence.” Two schools of thought emerged.

First, some respondents felt that the agent was intelligent because it understood their comments and questions and replied appropriately. One participant responded, “Yes, it knew what I asked, and it knew what it was supposed to know.” To these participants, competence in comprehension and knowledge in one particular area was good enough to be considered intelligent behavior.

The other set of responses suggested a view of intelligence as a deeper, more autonomous and independent characteristic that the agent did not possess. One participant commented, “[It was] just spitting out things, spouting off programmed things to say.” Furthermore, a number of participants stated that the agent was not intelligent, but the person who programmed it was. Such people appear unlikely to consider agents intelligent because they will tend to attribute the intelligence to the human who created the agent. Conversely, one respondent used a quasi-“depth” argument to justify calling the travel-task agent intelligent: “He was giving opinions as opposed to facts.”

Finally, the agent’s social abilities and personality (or lack thereof) were noted by a number of the participants. In the travel task, we intentionally had the agent begin the session saying, “Hello, [person’s name].” After one of its suggestions, the agent also stated, “All this talk of the trip makes me wish I was going too.” The animated agent even smiled after saying this. Three participants explicitly mentioned these features, one stating, when asked if the agent had a personality, “Yes, respectful. It said, ‘[my name]’, and ‘I agree with this.’ It mentioned that all the talk makes me wish I was going. I thought that was very funny. That was really cool.” Other comments implying a personality included, “Seemed a lot like a travel agent that was in a hurry,” and “helpful, but kind of annoying,” and “he seemed almost irritated when I didn’t agree with him. Sensitive maybe.”

The majority of participants felt that the agent did not have a personality, many citing the poor speech quality as a reason. More than one participant stated that they felt the agent *tried* to have a personality. One participant who did the editing task first, stated after the task that the agent did not have a personality, “It was just directed at answering questions. It had no inflections.” But when asked again after the travel task, the participant responded, “It was still mechanical, but you could feel the attempt at being more personable. It acknowledged my responses, asking me to elaborate. The responses were at a more personal level. Mouth movement is nice. If there was eye movement it would be better.” Finally, one participant who used the iconic agent responded, “If you can’t talk to someone face to face, it’s kind of hard to say it has a personality.”

When asked to name potential improvements to the agent, three participants mentioned that it could be more proactive by suggesting more efficient commands to use on the editing task. On a related note, the last interview question of each session asked participants if they were familiar with the Microsoft Office Paper Clip helper agent and if so, what was their opinion of it. Of the 36 people who said they were familiar, 30 had what could be termed a “negative” impression and 6 had a “positive” impression. The most commonly cited reason for the negative view was being annoyed at the agent popping up and offering irrelevant advice. These observations suggest that the more proactive the agent's behavior, the more important it is that the behavior be competent. Intrusions better be worth it!

6. DISCUSSION AND LESSONS LEARNED

To summarize the things we learned from the study, we will use the analytic framework proposed earlier as an organizational aid.

User - While there was near a total view that the agents were helpful on the two tasks, other impressions of the agents varied widely across the different participants in the study. Participants’ views on the agent’s intelligence, in particular, were strongly influenced by the person’s preconceived notion of intelligence. We noted that females were more likely to refer to the agent in a personified manner using gender-based pronouns. This hints that there may be gender-based differences in reactions to and interactions with agents. This appears to be a good candidate for more careful study in the future.

Agent - User performance and satisfaction were not significantly affected by the three different appearances of the agent in the experiment. The lifelike appearance and behavior of the animated agent was noted by some of the participants. The unrealistic computer-generated speech was noted by a majority of the participants. It appears that if an agent's personality is an important feature to the developer of the agent then care in creating the agent's appearance, social skills, and voice should be taken.

Task - As we hypothesized, the user's task and the agent's role in the task strongly influence people's perceptions of the agent, its value, and its attributes. Future experiments should carefully consider and systematically manipulate tasks and agent roles because they seem likely to directly affect users' perceptions.

We can also consider lessons learned from the study with respect to user performance and impressions.

Performance - Objectively, the experiment did not uncover any major differences in user effectiveness on tasks as a function of agent. One modest finding concerned how users followed the guidance of the agent on the travel task. Participants were more likely to modify item ranks when the agent suggested a change than when the agent agreed with the ranking. Subjectively, it was clear that the participants felt that the agent was helpful in both tasks as this was articulated frequently by different participants.

We uncovered no changes in efficiency due to the agent in this experiment, but that was not the primary focus of the study. To address this factor in more detail, it would be useful to conduct an experiment to compare agent-assisted performance to performance using other user interface techniques. For instance, a future experiment could replicate the editing task with an added condition of participants being assisted by on-line documentation. A few study participants subjectively addressed this issue by stating that they preferred speaking to the agent over on-line help, but this was balanced by others who said that they would prefer traditional system help/documentation to the agent they encountered in this experiment.

Satisfaction - As mentioned earlier, participant impressions of the agents were generally favorable. We observed quite a variation in impressions of specific agent qualities and attributes. Some participants stated that they felt the agent should have more personality and attitude, and others felt that such qualities were strange in such a thing. It appears that the "one size fits all" notion of creating an agent that acts similarly for all users is unlikely to be successful.

7. CONCLUSION

Anthropomorphic interface agents might be one of the best interface approaches ever devised. Or they might not. Equivocal results from prior research make it virtually impossible to decide this matter. The difficulty with prior work has been its lack of systematicity in examining key factors and the use of dependent measures that often did not appropriately assess subjective experience and objective performance.

In this article, we introduced a framework for systematically examining the effects of anthropomorphic agents on user performance and subjective responses. We performed an experiment within this framework that suggested that type of task may play an outsized role in the perception of agents. We plan to use our framework to guide additional studies and hope other researchers find it useful and that it will allow future experiments to build on each other more effectively than in the past.

REFERENCES

1. André, E., Rist, T., van Mulken, S., Klesen, M., and Baldes, S., in *Embodied Conversational Agents*, Cassell, J. Prevost, S., Sullivan, J., and Churchill, E. (eds.), MIT Press, Cambridge, MA, 2000, 220-255.
2. Ball, G. Lifelike Computer Characters: The Persona Project at Microsoft Research, in *Software Agents*, Bradshaw, J. (ed.). MIT Press, Cambridge, MA, 1997, 220-255.
3. Cassell, J. Embodied conversational interface agents. *Communications of ACM 43*, 4 (April 2000), 70-78.

4. Casell, J., Sullivan, J., Prevost, S., and Churchill, E., (eds.). *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000.

5. Churchill, E., Cook, L., Hodgson, P., Prevost, S., and Sullivan, J. May I Help You?: Designing Embodied Conversational Agent Allies, in *Embodied Conversational Agents*, Cassell, J. Prevost, S., Sullivan, J., and Churchill, E. (eds.), MIT Press, Cambridge, MA, 2000, 3-46.

6. Dahlback, N., Jonsson, A. & Ahrenberg, L. Wizard of oz studies – why and how, in *Proceedings of the 1993 International Workshop on Intelligent User Interfaces*, (Orlando, FL, 1993), 193-200.

7. Dehn, D.M. & van Mulken, S. The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies 52*, 1 (January 1999), 1-22.

8. Dryer, D. C. Getting Personal with Computers: How to design personalities for agents. *Applied Artificial Intelligence 13*, 3 (1999), 273-295.

9. Erickson, T. Designing Agents as if People Mattered, in *Software Agents*, Bradshaw, J. (ed.). MIT Press, Cambridge, MA, 1997, 79-96.

10. Fridlund, A.J. & Gilbert, A.N. Emotions and facial expression. *Science* 230 (1985), 607–608.

11. King, W.J. & Ohya, J. The representation of agents: Anthropomorphism, agency and intelligence, in *CHI ' 96 Conference Companion*, (Vancouver, B.C., April 1996), 289-290.

12. Koda, T. Agents with faces: A study on the effect of personification of software agents. Master's thesis, MIT Media Lab, Cambridge, MA, 1996.

13. Lai, J., Wood D. & Considine, M. The effect of task conditions on the comprehensibility of synthetic speech, in *Proceedings of the ACM CHI 2000*, (The Hague, Netherlands, April 2000), 321–328.

14. Lanier, J. Agents of alienation. *Interactions 2*, 3 (July 1995), 66–72.

15. Laurel, B. Interface agents: Metaphors with character, in *The Art of Human-Computer Interface Design*, Laurel, B. (ed.), Addison-Wesley, Reading, MA, 1990, 355-365.

16. Lyman, P. & Varian, H. How Much Information? Available at http://www.sims.berkeley.edu/how-much-info/.

17. Maes, P. Agents that reduce work and information overload. *Communications of the ACM*, *37*, 3 (July 1994), 31-40.

18. Massaro, D.W., Cohen, M. M., Beskow, J., & Cole, R.A. Developing and evaluating conversational agents, in *Embodied Conversational Agents*, Cassell, J. Prevost, S., Sullivan, J., and Churchill, E. (eds.), MIT Press, Cambridge, MA, 2000, 287-318.

19. McCrae, R. & Costa, P. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 1 (1987), 81-90.

20. Nass, C., Isbister, K. & Lee, E. Truth is beauty: Researching embodied conversational agents, in *Embodied Conversational Agents*, Cassell, J., Prevost, S., Sullivan, J., and Churchill, E. (eds.), MIT Press, Cambridge, MA, 2000, 374-402.

21. Nass, C., Steuer, J., & Tauber, E. Computers are social actors, in *Proceedings of CHI '94*, (Boston, MA, April 1994), 72-78.

22. Oren, T., Salomon, G., Kreitman, K. and Abbe, D. Guides: Characterizing the interface, in *The Art of Human-Computer Interface Design*, Laurel, B. (ed.), Addison-Wesley, Reading, MA, 1990, 367-381.

23. Prevost, S., Hodgson, P., Cook, L. and Churchill, E. Face-to-face interfaces, in *CHI '99 Conference Extended Abstracts*, (Pittsburgh, PA., May 1999), 244-245.

24. Rickenberg, R. & Reeves, B. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces, in *Proceedings of CHI 2000*, (The Hague, Netherlands, April 2000), 329-336.

25. Shneiderman, B. & Maes, P. Direct manipulation vs. interface agents. *Interactions*, *4*, 6 (Nov. + Dec. 1997), 42-61.

26.Shneiderman, B., Direct Manipulation Versus Agents: Paths to Predictable, Controllable, and Comprehensible Interfaces, in *Software Agents*, Bradshaw, J.M. (ed.), MIT Press, Cambridge, MA, 1997, 97-106.

27.Takeuchi, A. & Nagao, K. Communicative facial displays as a new conversation modality, in *Proceedings of INTERCHI ' 93,* (Amsterdam, April 1993), 187-193.

28.Takeuchi, A. & Nagao, K. Situated facial displays: Towards social interaction, in *Proceedings of CHI ' 95 Conference*, (Denver, CO, May 1995), 450-455.

29.Thórisson, K. Real-time decision making in multimodal face-to-face communication, in Proceedings of the Second International Conference on Autonomous Agents, (Minneapolis, MN, May 1998), 16-23.

30.Walker, J.H., Sproull, L., & Subramani, R. Using a human face in an interface, in *Proceedings of CHI ' 94*, (Boston, MA, April 1994), 85-91.