Evaluation of an anthropomorphic user interface in a travel reservation context and affordances

Murano, P, Gee, A and Holt, PO

Evaluation of an Anthropomorphic User Interface in a Travel Reservation Context and Affordances

Dr Pietro Murano, Anthony Gee, and Prof. Patrik O'Brian Holt

Abstract— This paper describes an experiment and its results concerning research that has been going on for a number of years in the area of anthropomorphic user interface feedback. The main aims of the research have been to examine the effectiveness and user satisfaction of anthropomorphic feedback in various domains. The results are of use to all interactive systems designers, particularly when dealing with issues of user interface feedback design. There is currently some disagreement amongst computer scientists concerning the suitability of such types of feedback. This research is working to resolve this disagreement. The experiment detailed, concerns the specific software domain of Online Factual Delivery in the specific context of online hotel bookings. Anthropomorphic feedback was compared against an equivalent non-anthropomorphic feedback. Statistically significant results were obtained suggesting that the non-anthropomorphic feedback was more effective. The results for user satisfaction were however less clear. The results obtained are compared with previous research. This suggests that the observed results could be due to the issue of differing domains yielding different results. However the results may also be due to the affordances at the interface being more facilitated in the non-anthropomorphic feedback.

Index Terms— anthropomorphism, user interface feedback, evaluation, affordances.

1 Introduction

User interfaces and the feedback given to users are one of the most important aspects of any software system. This is because if the user interface and the feedback given is not usable, the users will either give up using the system, will be less efficient in using the system or will simply not enjoy using the system. This in turn can seriously affect the success of a software house and its sales. Also the growth and complexity of modern day software systems, in particular the tasks they are able to perform, results in the continual requirement for more usable interfaces to be developed.

The main objective of this research is to aid in the improvement of user interfaces by better understanding the effects of using anthropomorphic user interface feedback. Specific concentration is placed on comparing anthropomorphic and non-anthropomorphic user interfaces to address the issues of effectiveness and user satisfaction in relation to context and domain and to provide some explanation in terms of an appropriate theory such as the theory of affordances.

There are various opinions amongst the computer science community regarding the effectiveness and user approval of anthropomorphic feedback at the user interface. Some researchers are in favour of anthropomorphism, e.g. see [1-6]. However, some researchers are not generally in favour of anthropomorphism in most circumstances e.g. see [7]. Each of these researchers tends to base their opinions on various studies conducted in the area. Due to the inconclusive nature of the results of these studies, there is the need for more work in this area to gain a better understanding of such differences in opinion and experimental results.

The rest of this paper is composed of four main sections. Section 2 briefly reviews some key previous research. Section 3 describes in detail the experiment carried out including the observed results. Section 4 discusses the observed results in light of Affordances. The paper then concludes with section 5 proposing subsequent steps to further the research.

- *Dr Pietro Murano is with the University of Salford, School of Computing, Science and Engineering, Newton Building, Salford, M5 4WT, UK.*
- *A. Gee was with the University of Salford, School of Computing, Science and Engineering, Newton Building, Salford, M5 4WT, UK.*
- *Prof Patrik O'Brian Holt is with The Robert Gordon University, School of Computing, St Andrew Street, Aberdeen, AB25 1HG, Scotland.*

2 Some Key Previous Research

This section will aim to discuss research which has already been carried out by others and the authors of this paper on anthropomorphism and highlight some of the differences in results obtained by other researchers. Research about anthropomorphism spans various contexts including agent-based software (an interface software agent usually assists the user in some way in their tasks and in some cases may be an animated character).

The first paper to consider had an experimental study by Moreno et al [8] in the context of tutoring and learning about plants, they found that experimental participants using an anthropomorphic agent were better able to use their newly learned knowledge to solve similar problems in the same domain. They also found that participants in

this group had more motivation to continue learning about plants and had overall more interest in the subject area. No difference was found for actual memory capacity. The control group used the same information as in the anthropomorphic agent group, but the agent was substituted with text.

Another study in the area of tutoring by Moundridou and Virvou [9] tested 2 conditions in an algebra tutoring environment. The first condition had a talking synthetic face and the second was the same as the first condition with text replacing the synthetic face. The main results showed that there was no significant difference between the 2 conditions for task time completion. However the participants in the anthropomorphic condition enjoyed the experience more, found the system more useful and less difficult to use.

Also in a study by Maldonado et al [10] an environment for teaching Japanese students about American English idioms was used. The environment had an anthropomorphic tutor. They tested 3 conditions. The first used only the anthropomorphic tutor, the second had the tutor and an anthropomorphic peer learner with no emotions and the third had the tutor and an anthropomorphic peer learner with emotions. The idea was to emulate more closely the actual interactions involved between tutors, a learner and other learners. The results of this study, which had statistical significance, showed the third condition fellow learner to have overall more 'social skills'. Learning was also greater under the third condition. Furthermore the participants reported more enjoyment under the third condition.

In another study by Catrambone et al [11] an experiment was conducted using an editing environment and a travel items recommendation environment. Three conditions were tested with these 2 environments. The first condition was an animated agent, the second was a still photograph of the same agent used in the first condition and the third was a cartoon image of a lit light bulb. The tasks involved doing some editing in an unknown word processor and making some choices regarding what items to take on an international trip. During the editing task the agent was reactive in nature, while in the travel items task the agent was proactive. The main results were that for the travel task, the participants were generally influenced by the agent's suggestions. However no effect was recorded for type of agent. In the editing task, there was no difference in task time across the 3 conditions. Further, the participants felt that the agent was less intrusive and more worthwhile in the editing task than in the travel task. Also the participants were observed to be at ease in querying the agent for help in the editing task while the converse was true for the travel items task.

Furthermore in a related study by Xiao et al [12] an experiment was conducted in an editing environment testing 3 experimental conditions. The first was a reactive anthropomorphic agent, the second was also an agent that was reactive and proactive in nature and the third was a control condition consisting of an approximately equivalent paper based manual. The main results for the experiment show that there were no significant differences in task time and number of commands used across the 3 conditions. After the experiment participants were also asked to recall as many editor commands as possible. This aspect did not produce any significance across the 3 conditions. There were also no significant results in the participant opinions about the agents and paper manual.

In a study by David et al [13], the authors conducted a three condition experiment in the context of a quiz about ancient history. They were investigating different anthropomorphic cues in terms of character gender and attitude and user perceptions about the character in relation to quiz success (or not). The overall results of their experiment suggested that anthropomorphic cues led to users believing the character to be less friendly, intelligent and fair. This finding was linked with the male character and not with the female character.

An interesting investigation has also been carried out by Forluzzi et al [14]. Their main consideration concerned anthropomorphic form. They tested different anthropomorphic forms in terms of abstraction, two or three dimensional form and gender – cartoon based/realistic appearance. Their results suggest that the form of the anthropomorphic character is linked to its task and that users tend to prefer a character that fits with gender stereotypes and particular tasks.

The discussion so far clearly suggests that using anthropomorphic feedback in a given context does not guarantee better usability in an application. Clearly the above studies have shown that the results overall in various experiments spanning several years at times show anthropomorphism to be better or worse and in some cases not being any different to conventional type feedback. This pattern of inconsistent results in relation to using anthropomorphic feedback has also been observed in the authors' previous work (see [15-20]) on anthropomorphic feedback.

In Murano [18] it was shown that in the domain of software for in-depth learning, anthropomorphic feedback was significantly more effective. The results for user satisfaction were not so clear, but participant preferences tended towards the anthropomorphic feedback. This was specifically in the context of English as a Foreign Language pronunciation. Also in Murano [17] it was shown that in the domain of software for online systems usage, anthropomorphic feedback was significantly more effective and preferred by users. This context specifically involved the area of using UNIX commands at the UNIX shell.

Specifically related to this paper, are the results by Murano [16]. The paper investigated anthropomorphic feedback in the context of online factual delivery, using the area of direction finding as the specific context. This paper showed with statistically significant results, that non-anthropomorphic feedback was more effective. The results for user satisfaction were not so clear, but participant preferences tended towards the non-anthropomorphic feedback.

As mentioned in the introduction, this research is aiming to find more information regarding the usage of anthropomorphic feedback, particularly aiming to discover

if such feedback is appropriate in terms of effectiveness and user satisfaction. The research is being done in various software domains.

This paper therefore investigates the domain of online factual delivery further, describing an experiment set in this domain, using the context of online hotel bookings to test the user interface feedback. This context was chosen because it is a fairly common activity for users of all kinds to carry out over the Internet and was therefore considered to be useful and realistic, whilst maintaining the theme of the previous experiment conducted by Murano [16]. As with the previous experiments, effectiveness and user satisfaction were the aspects being investigated. Effectiveness was defined by the success rate in completing the tasks, a low error rate whilst carrying out the tasks and a low rate of hesitations/frustrations expressed by the participants during the experiment. The user approval aspects concerned the participants' subjective opinions regarding the user interface aspects. For this experiment, the anthropomorphic feedback consisted of an animated character supplied with MS Agent 2.0 (see Apparatus and Material section) called 'Merlin'. The non-anthropomorphic feedback consisted of guiding text. This was text of the kind one would expect to see on a 'real' online hotel booking site.

3 The Experiment - Hotel Bookings

3.1 Hypotheses

As stated in the previous section this research concerns determining the effectiveness and user satisfaction of anthropomorphic user interface feedback in various contexts. Hence the following hypotheses were derived:

H0a - There will be no difference in terms of user satisfaction between the anthropomorphic feedback (Merlin) and non-anthropomorphic feedback (guiding text).

H0b - There will be no difference in terms of effectiveness between the anthropomorphic feedback and non-anthropomorphic feedback.

H1a - The non-anthropomorphic (guiding text) feedback will be more effective than the anthropomorphic (Merlin) feedback.

H1b - Users will prefer the anthropomorphic (Merlin) feedback rather than the non-anthropomorphic (guiding text) feedback.

3.2 Users

The initial recruitment of the participants took place by means of a recruitment questionnaire. The participants were carefully selected so as to have similar profiles, with the aim of having an approximately level starting point of knowledge and experience in relation to the context and tasks of the experiment (This practice is used when participant background can have an effect on overall results (see [21]). Initially 40 individuals were selected, but only 20, with similar profiles, were actually used in the experiment. The main aspects of the profiles of the participants used were similar in the following ways: All participants had similar computing knowledge. They were not complete beginners or 'power' users. Complete novice users were not selected as they would have required basic training in the concepts of devices and Windows systems. Experienced participants were not used in the experiment as it was decided that such users would in reality not require feedback of the sort being tested in their every day usage patterns. Lastly all the participants were less than 36 years of age with English as their primary language.

3.3 Experimental and Task Design

For the purpose of the given experiment a between users design method was deployed. 10 of the participants were assigned to Group A, and the remaining 10 participants were assigned to Group B. Participants were assigned to groups randomly.

Group A participants tested the anthropomorphic feedback (MS Merlin) as part of their experiment session.

Group B participants tested the non-anthropomorphic feedback (guiding text) as part of their experiment session.

The experiment involved each participant attempting the following tasks: Task 1 required participants to make a specific booking for a hotel and theatre performance. Participants would use the prototype online hotel reservation user interface to make the bookings according to specific details supplied. Task 2 required participants to cancel the booking they had just made using the hotel reservation user interface.

The tasks outlined are representative of realistic tasks commonly carried out by users booking a hotel or holiday, using the Internet. For tasks 1 and 2 all participants were initially shown a brief tutorial explaining how to book and cancel a hotel using the interface. The content of the tutorials shown was identical regardless of the feedback being given to ensure there was no bias.

3.4 Variables

For the purpose of the experiment the associated independent variables were determined as being the two different methods of feedback that were available, i.e. Animated Microsoft® Merlin with speech and text (anthropomorphic) and guiding text (non-anthropomorphic).

The dependent variables were the participants' performance in dealing with the hotel bookings and their subjective opinions.

The dependent measures were that the performance was measured by counting the number of errors incurred, observing whether participants completed the tasks and counting the number of times hesitation and frustration were manifested. These factors were then used in a scoring formula (see Scoring section below for a description of the formula). Specifically performance was measured in the following manner:

1. Tasks carried out with some deviation from the instructions given. The participants were given a task sheet with specific instructions regarding the booking they should make (e.g. given dates and number of rooms required etc.). Deviation from this was considered to be a complete task but with some incorrect details.

2. Tasks completed. This refers to the overall successful completion of the two prescribed tasks.
3. Number of times participants showed very clear signs of hesitation, e.g. their facial expression appearing puzzled or requesting help from the experimenter.
4. Number of times participants showed very clear signs of frustration, e.g. verbally commenting on an aspect of the user interface which caused them some 'anger'.
5. The number of times participants used the feedback help, and then went on to do an error.

These factors were recorded by means of an observation protocol.

The subjective opinions were measured by means of a post-experiment questionnaire. Participants were asked to rate various aspects of the user interface using a Likert type scale, where 9 was the most positive score regarding some opinion, and 1 was the most negative score available. The aspects covered by the questionnaire, included several questions on the general user interface features of the prototype, the error messages used by the prototype, the tutorial material viewed and 'emotional' feelings of the participant.

3.5 Apparatus and Materials

The equipment used in the experiment involved a laptop with, 128MB RAM, 20Gb disk and Windows™ XP. Also Microsoft® Agent 2.0, the "Merlin" character and Lernout & Hauspie TruVoice Text-to-Speech (TTS) engine (American English) were used. Supplementary hardware used consisted of an external mouse and external speakers. Further, a paper notepad was available for each participant, for use in the experiment (see Procedure section). Each prototype was developed using Visual Basic 6. The Anthropomorphic interface required the use of the Microsoft® Agent 2.0 Active X™ component.

Two questionnaires were designed for the experiment. The first was a pre-experiment questionnaire for recruitment purposes and the second was a post-experiment questionnaire for eliciting subjective opinions from the participants. An observation protocol was also designed for recording observed errors and participant behaviour. This consisted of a categorised grid where the observer could quickly record the number of errors etc. using a tally system. Obvious participant behaviour was recorded, e.g. a participant exhibiting clear annoyance whilst doing a task. Having one trained observer and defining in advance what were categorised as 'errors' and the kind of participant 'behaviour' that would be recorded, ensured a more consistent set of data.

3.6 Procedure

The experiment itself took approximately 30 minutes to complete. The procedure involved ensuring that each participant was treated in the same way, with the following outlined procedure being identical for each of the participants. Also all the questionnaires and observation techniques used were the same for each participant, with the aim of minimizing confounding variables.

The experiment took place in a carefully controlled environment, ensuring that there were no distractions and that the participants felt at ease.

Upon entering the room each participant was greeted by the experimenter and was made to feel comfortable and relaxed. To make them feel more at ease, light refreshments were also offered at this time. The participants received a short verbal introduction to the experiment, explaining the purpose of the study, with reassurance that the software was the focus of the study and not themselves. At this time participants were informed that they would be observed by the experimenter who would be present in the room throughout the experiment. When the participant felt relaxed, a task sheet was given to them, which contained a brief introduction to the experiment along with Tasks 1 and 2 (see Experimental and Task Design section). Having read through the task sheet participants were again assured that they were not being examined and they were subsequently asked if they had any immediate concerns regarding the tasks. Participants were then instructed as to which method of feedback they would be testing.

Once the participant was ready the program began with a brief tutorial using the relevant method of feedback (Group A - anthropomorphic and Group B - non-anthropomorphic in terms of feedback). Both tutorials, regardless of the feedback, were the same in content. The only differences involved the anthropomorphic character referring to itself as 'I', while the non-anthropomorphic feedback was neutral in nature. The tutorial informed the participant how to book and cancel a hotel using the prototype. When the tutorial was started, the relevant mode of feedback 'explained' how to use each screen and its features. All the screens involved in the tutorial dealt with bookings and the cancellation of bookings. For the anthropomorphic condition the character uttered the information and this was also concurrently viewable by means of corresponding speech bubbles. Further, the character moved on the screen and 'pointed' with a hand to the features of each screen as it was being 'described'. For the non-anthropomorphic condition, the same information appeared in text boxes with arrows pointing to the various features of the screens.

Upon completion of the tutorial participants were then asked whether the tasks were clear, and when the participants felt ready the first task began.

Upon completion of task 1 participants were asked if they had any immediate comments as to the task they had completed, such comments being recorded in the observation notes. The participants were then asked whether they were ready to begin task 2, once comfortable, task 2 proceeded. It was determined that the task was complete when the participants had successfully cancelled the booking they had made during task 1. Following the task, completion comments and opinions were sought from the participants.

Errors were categorised by recording whether a participant completed the task according to the specifications given on the task sheet. If the participants deviated from the instructions given, e.g. the hotel was booked for the

party arriving on the wrong day, or not enough rooms booked etc, this was recorded as a participant completing a task but with some incorrect details (see Variables section above).

At times when the participants hesitated as to what they were required to do at a particular point, these hesitations were recorded (see Variables section above). At any point during the experiment if a participant asked the experimenter present in the room for guidance, no additional help was given. Instead participants were instructed that they should consult the feedback integrated into the interface, which was of the same kind as found in the tutorial and had the same condition being tested. If at any time a participant did consult the feedback, and still subsequently made an error regarding the problem they were trying to overcome, this was recorded. However, if the participant did consult the feedback and this solved the problem, this was also recorded. The number of times participants expressed clear frustration was also recorded. Such frustration included occurrences where participants would make remarks regarding certain aspects of the interface or feedback that caused them anger.

A particular aspect of the second task was to enter the booking reference supplied when participants made a booking during Task 1, so that the correct booking information could be retrieved to enable the booking to be cancelled. If a participant was unable to remember the booking reference (the software instructed the participant to note the reference during Task 1), having not written it down on the notepad provided, this would be seen as an error and subsequently resulted in the participant not fully completing Task 2.

Once all tasks had been completed the experimenter debriefed each participant. This included the completion of the post experiment questionnaire, elicitation of participants' immediate comments as well as the experimenter informing the participant how the results of the study will be made available if required.

3.7 Scoring

The effectiveness variables described (see Variables section) were carefully recorded for each participant. For each task completed/not completed, a score was assigned for use in the statistical analyses. The score for each task was based on a similar points system as published in [17]. For each task, each participant (unknown to them) was started on 10 points.

Events which caused the score to reduce were observations of the following types: Signs of frustration (negative physical attitude) or hesitation resulted in 0.5 points being deducted from the score. If the participant carried out an incorrect action, causing the system to display an error message, 0.5 points were deducted. If the participant consulted the feedback in a particular situation and despite the help, continued to make a mistake, 0.5 points were deducted from the running score.

Occurrences when the participant had completed the task but made a mistake in the booking, resulted in 1.5 points being deducted from the score. If the participant was unable to complete the task, 1.5 points were deducted. Finally if the participant completed the task with none of the noted penalties the score would remain at 10. Consequently, at the end of each task the participant obtained a final score.

The formula was devised because it was felt that all the factors being measured potentially had a direct effect on overall success. However the authors are also aware that such an approach can lead to the hiding of certain effects. Therefore the authors did the analysis again with the data decoupled from the formula. This extra analysis did not reveal anything useful or contradictory compared to the findings when using the data within the formula. Hence, for brevity, the next section will present the most interesting findings in conjunction with the analysis of the data using the formula described above.

3.8 Results

The data obtained for this experiment concerned effectiveness and subjective user opinions issues. The effectiveness issues and subjective opinions were statistically analysed by means of MANOVA testing and when significance was observed, the data was then subjected to post-hoc testing using t-tests for confirmation purposes. The data simultaneously used in the MANOVA calculation were the two experimental conditions (described above), gender and number of tasks done. These were analysed with 43 factors elicited from the post-experiment questionnaire (briefly described in Section 3.4 above) and from the data collected by observation (e.g. errors and other user actions such as manifesting anger etc.).

Firstly the tables of means and standard deviations (SD) for the MANOVA results presented later in this section are shown below in Tables 1 to 4.

TABLE 1
MEANS AND SD - FINAL SCORE

Anthropomorphic	
Mean	15.25
Std Dev	1.46
Std Err Mean	0.46
upper 95% Mean	16.29
lower 95% Mean	14.21
N	10
Non-Anthropomorphic	
Mean	17.90
Std Dev	0.88
Std Err Mean	0.28
upper 95% Mean	18.53
lower 95% Mean	17.27
N	10

TABLE 2
MEANS AND SD – TASK 1 SCORE

Anthropomorphic	
Mean	7.25
Std Dev	1.55
Std Err Mean	0.49
upper 95% Mean	8.36
lower 95% Mean	6.14
N	10
Non-Anthropomorphic	
Mean	8.85
Std Dev	0.53
Std Err Mean	0.17
upper 95% Mean	9.23
lower 95% Mean	8.47
N	1

TABLE 3
MEANS AND SD – PLEASANT ERROR MESSAGES

Anthropomorphic	
Mean	7.8
Std Dev	0.79
Std Err Mean	0.25
upper 95% Mean	8.36
lower 95% Mean	7.24
N	10
Non-Anthropomorphic	
Mean	8.6
Std Dev	0.52
Std Err Mean	0.16
upper 95% Mean	8.97
lower 95% Mean	8.23
N	10

TABLE 4
MEANS AND SD – DETAILED TUTORIAL

Anthropomorphic	
Mean	7.7
Std Dev	0.48
Std Err Mean	0.15
upper 95% Mean	8.05
lower 95% Mean	7.35
N	10
Non-Anthropomorphic	
Mean	8.4
Std Dev	0.52
Std Err Mean	0.16
upper 95% Mean	8.77
lower 95% Mean	8.03
N	10

For the 20 participants, 10 using the anthropomorphic feedback (MS Merlin) and 10 using the non-anthropomorphic feedback (guiding text), data gathered concerning effectiveness issues showed some significance as described below.

For the variables 'task 1 score' and 'group' there is a significant difference. The non-anthropomorphic group scored significantly ($p<0.05$) higher than the anthropomorphic group in task 1, with an F-ratio of 3.19*. This can be seen in Table 5 below:

TABLE 5
MANOVA – TASK 1 – SCORE AND GROUP, GENDER, NO OF TASKS DONE

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	16.97	4.24	3.19
Error	15	19.98	1.33	Prob > F
C. Total	19	36.95		0.04

A post-hoc t-test was conducted, which confirmed the above result with t = 2.13*, Alpha = 0.05. There were no significant effects in relation to the gender and the number of tasks done (Note: gender data was collected as part of the recruitment process, but was not a main aspect of the research. However it was included in the analysis for thoroughness and interest).

For the variables 'final score' and 'group', there is a significant difference. The non-anthropomorphic group scored significantly ($p<0.01$) higher than the anthropomorphic group, with an F-ratio of 5.62**. This can be seen in Table 6 below:

TABLE 6
MANOVA – FINAL SCORE AND GROUP, GENDER, NO OF TASKS DONE

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	36.67	9.17	5.62
Error	15	24.46	1.63	Prob > F
C. Total	19	61.14		0.006

A post-hoc t-test was conducted, which confirmed the above result with t = 2.13*, Alpha = 0.05. There were no significant effects in relation to gender and the number of tasks done.

For the data collected concerning the subjective opinions by means of the post-experiment questionnaire, mostly the results showed no significance. The main exceptions where some significance is observed are shown below.

For the variable 'pleasant error messages' and 'group' there is a significant difference ($p<0.05$). The non-anthropomorphic group subjectively scored the pleasantness of error messages significantly higher than the anthropomorphic group, with an F-ratio of 3.30*. The main results are shown in table 7 below:

TABLE 7
MANOVA – PLEASANT ERROR MESSAGES AND GROUP, GENDER, NO OF TASKS DONE

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	5.24	1.31	3.30
Error	15	5.96	0.40	Prob > F
C. Total	19	11.20		0.04

A post-hoc t-test was conducted, which confirmed the above result with t = 2.13*, Alpha = 0.05. There were no significant effects in relation to gender and the number of

tasks done.

For the variables 'detailed tutorial', 'group' and 'gender' there is a significant difference. The non-anthropomorphic group subjectively scored the tutorial as being significantly more ($p<0.05$) detailed, than the anthropomorphic group. Furthermore, although this is not a gender study, the males significantly scored the tutorial as being more detailed than the females. The F-ratio is 4.45* shown below in Table 8.

TABLE 8
MANOVA – DETAILED TUTORIAL AND GROUP, GENDER, NO OF TASKS DONE

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	3.77	0.94	4.45
Error	15	3.18	0.21	Prob > F
C. Total	19	6.95		0.01

A post-hoc t-test was conducted, which confirmed the above result with t = 2.13*, Alpha = 0.05. There were no significant effects in relation to the number of tasks done.

Participants were also asked their opinions regarding potential future use of the interface feedback they used during the experiment. For the non-anthropomorphic group (guiding text), 6 out of 10 participants said they would use the feedback again if made available. For the anthropomorphic group (MS Merlin), 8 out of 10 participants said they would use the feedback again if made available.

3.9 Experimental Conclusions

The results from the individual tasks show statistical significance in favour of the non-anthropomorphic (guiding text) feedback. Participants completed the tasks more successfully and with less errors/hesitations in the hotel booking context.

Consequently, with reference to the hypotheses stated earlier in this paper, it is now possible to reject the (H0b) null hypothesis, with the results showing that there is a clear difference between the feedbacks in terms of effectiveness. Statistical significance in the results enables the (H1a) positive hypothesis to be accepted. This postulated that the non-anthropomorphic feedback would be more effective.

Assessment of the user satisfaction of the two types of feedback in terms of the tutorial and help sub-system implies that overall there were not many significant differences between the 2 experimental groups. Some of the exceptions to this are seen in tables 7 and 8 above. The differences were in favour of the non-anthropomorphic feedback. Although the aspect of the tutorial being detailed could be positive if viewed from the perspective that more detail is better for accomplishing tasks. However another view could postulate that more detail can create confusion if not essential to the context and tasks. Unfortunately the experimental design and procedure did not take this aspect further in the post-experiment questions. Also the reason for the male participants rating the tutorial as significantly more detailed compared to the females' opinions is unclear. However, while it is an interesting question, since this is not primarily gender research, we leave the matter for further research in the future.

Therefore although the subjective responses are tending to slightly favour the non-anthropomorphic feedback, the results overall are not strong enough to categorically accept the positive hypothesis. This is because a limited number of factors are implying user satisfaction towards the non-anthropomorphic feedback with statistical significance. The positive hypothesis (H1b) is therefore rejected.

Some interesting comments and observations were made by the participants during and after the experiment. The anthropomorphic feedback seemed to 'fascinate' some of the participants in this group. Some commented that they 'sat back' and 'watched'. Some also commented that they felt they were not actually learning anything. Certain individuals seemed to concentrate more on the 'appearance' of the Merlin character rather than concentrating on the words being uttered. These participant comments were reasonable as their observed behaviour matched their self-evaluation. Another interesting aspect concerns the fact that some participants in the anthropomorphic group stated that their experience with this feedback was 'engaging', 'involving' and 'fun'. The converse was true of some of the comments made by the non-anthropomorphic group, where some stated their experience was 'uninspiring' and 'normal'. These aspects were also evident as the participants were being observed. These participant comments and observations could explain why the anthropomorphic feedback was rated very closely to the non-anthropomorphic feedback. It could simply be that the anthropomorphic feedback had more of a novelty factor. However the authors suggest that this novelty factor would disappear with regular use of such a system.

4 OVERALL DISCUSSION

These results generally follow the results obtained by Murano [16] in the different context of direction finding – within the same domain of online factual delivery. This could suggest that similar domains or contexts would yield similar results. In Dehn and van Mulken [22] it was suggested that the context or domain of concern could influence the effectiveness and user approval of anthropomorphic interfaces. Something similar was suggested by Catrambone et al [11] regarding context and task type. While context and task type may be factors affecting the usage of anthropomorphic feedback, this may not be the whole explanation – despite this direction being the initial approach of the authors. One reason for this cautious approach is that as can be seen from some of the studies summarised at the beginning of this paper, the domains were similar, but the results were disparate. One such example concerns the plant design context [8] which seemed to show better results when the anthropomorphic feedback was used, but no differences were found when an anthropomorphic feedback was used in the algebra context [9]. Although these are two different subject areas

(biology and mathematics), they are both in the same domain of tutoring software. If the issue was simply a matter of different domains, one would have expected more similar results in the tutoring domain. A further example concerns the study summarised above regarding various editing tasks [12]. This showed no differences between the conditions. However the study regarding UNIX commands [17] showed the anthropomorphic feedback to be more effective. These two studies although having slightly different task types, are within the same kind of domain for systems usage. If the issue was simply about domain, one would expect more closely aligned patterns of results.

If context does not provide a complete enough explanation, this clearly raises the question regarding what does provide a more complete explanation or complimentary explanation. The principal author of this paper has been investigating the theory of affordances as a likely explanation for the observed effects. The concept of affordances was initially introduced by Gibson [23] in terms of how organisms (e.g. humans) interact and react to the environment. The theory was then reinterpreted in terms of computer systems and particularly user interfaces. Norman [24, 25] and Hartson [26] are the main sources of the reinterpretations, with more lightweight contributions from Gaver [27] and McGrenere and Ho [28], where they started to apply affordances to computer systems and to decompose affordances into different components.

Norman [24] describes what he means by affordances. Generally, but not exclusively, he talks about perceived affordances - in contrast with Gibson's physical affordances. In a typical user interface, an example of a real or physical affordance would be if the designers of a system only allowed the mouse cursor to be visible when it was over a clickable area. In contrast a perceived affordance is of the kind where one would have a screen with various icons present etc. Another way of thinking about this, is that a screen with various icons visible is actually 'visual feedback' which makes known the affordances (perceived affordances) to the user [24].

However, the most substantial reinterpretation has been conducted by Hartson [26]. He identifies cognitive, physical, functional and sensory affordances. His rationale is that when doing some computer related task, the users are using cognitive, physical and sensory actions. Cognitive affordances involve 'a design feature that helps, supports, facilitates, or enables thinking and/or knowing about something' [26]. One example of this aspect concerns giving feedback to a user that is clear and precise. If one labels a button, the label should convey to the user what will happen if the button is clicked. Physical affordances are 'a design feature that helps, aids, supports, facilitates, or enables physically doing something' [26]. According to Hartson a button that can be clicked by a user is a physical object acted on by a human and its size should be large enough to elicit easy clicking. This would therefore be a physical affordance characteristic. Functional affordances concern having some purpose in relation to a physical affordance. One example is that clicking on a button should have some purpose with a goal in mind. The converse is that indiscriminately clicking somewhere on the screen is not purposeful and has no goal in mind. This idea is also mentioned by McGrenere and Ho [28]. Lastly, sensory affordances concern 'a design feature that helps, aids, supports, facilitates or enables the user in sensing (e.g. seeing, feeling, hearing) something' [26]. Sensory affordances are linked to the earlier cognitive and physical affordances as they complement one another. This means that the users need to be able to 'sense' the cognitive and physical affordances so that these affordances can help the user.

In terms of the experiment presented in this paper concerning hotel bookings the results showed significance for effectiveness in favour of the textual feedback (non-anthropomorphic). The user satisfaction however was inconclusive. Therefore looking at the interaction that took place, it must be noted that all the information required to complete the tasks was presented at the beginning of the session as a tutorial (see procedure description above). Then the participants were asked to carry out the prescribed tasks based on the information given at the outset. This may have affected the results in favour of the textual non-anthropomorphic feedback. This is possible because the cognitive affordances in the textual condition could have facilitated the 'thinking' and 'knowing' processes for carrying out the task. This may have happened because in the textual non-anthropomorphic condition the participants were able to go through the tutorial material at their own pace. However in the Merlin anthropomorphic condition the character went through the various stages at a 'conversational' pace with accompanying speech bubbles. Therefore it is possible that the cognitive affordances could have been positively affected in the non-anthropomorphic condition. In turn the sensory affordances could have been facilitated by the participants being able to 'see' or read better what had to be done to accomplish the tasks. One of the statistically significant results of the experiment concerning the fact that participants in the non-anthropomorphic group thought the tutorial was more detailed than the participants in the anthropomorphic group could give credence to this argument. In contrast the Merlin anthropomorphic feedback could have negatively affected the cognitive affordances by not supporting as well the act of 'thinking' or 'knowing' what to do to accomplish the tasks, perhaps due to being obliged to proceed at Merlin's pace (this was not too fast but could have had an effect) and not at the participants' pace. This in turn could have affected the sensory affordances of not being able to 'see' or perceive as well what would need to be done to accomplish the tasks of making a booking and cancelling a booking. This could suggest that the sensory affordances were positively affected in the non-anthropomorphic condition. In this experiment the physical affordances were also the same under both conditions and therefore should not have affected the results under some specific condition. The various interaction screens were the same under both conditions. The functional affordances would therefore have been the same under both conditions, but obviously how the various screens were explained under each con-

dition differed. The labels of fields and buttons were the same under both conditions, thus equalising the functional affordances. Their explanations differed in how the information was presented, but not in their content.

5 Concluding Remarks

Having considered some of the research in the area of anthropomorphic interface feedback and presented the results of an experiment, the authors accept, based on the evidence of several years' research by others and by the authors that the results observed could be to do with factors of differences of context or domain. However according to the authors this is probably not the only explanation. An explanation that complements the supposition of contextual and domain differences is that differences in a user interface, such as the ones mentioned above, can lead to the affordances being either facilitated or hindered. If they are hindered then the results will probably mean a reduction in effectiveness and user satisfaction. It is therefore suggested that the issue of affordances continues to be analysed in future research to either confirm or disprove this apparent emerging link. If the link can be more firmly established [29], it will help user interface developers to produce better interfaces regardless of using anthropomorphism or not as a medium. Furthermore any future experiment conducted would ideally have a larger sample size than the one used for this experiment.

References

[1] T. Koda, and P. Maes, "Agents with faces: the effect of personification," *Proc. of the 5th IEEE International Workshop on Robot and Human Communication,* p 189-194, RO-MAN 1996.

[2] P. Maes, "Agents that reduce work and information overload," *Communications of the ACM*, 37(7), pp. 31-40, 1994.

[3] B. Laurel, "Interface agents: metaphors with character," *In*: J.M. Bradshaw, ed. *Software Agents*, MIT Press, London, p67-77, 1997.

[4] A. Agarwal, "Raw computation." *Scientific American*, 281, 44-47, 1999.

[5] V. Zue, "Talking with your computer," *Scientific American*, 281, pp. 40-41, 1999.

[6] A. Takeuchi, and T. Naito, "Situated facial displays: towards social interaction," *Proc. CHI'95 Human Factors in Computing Systems*, pp. 450-454, 1995.

[7] B. Shneiderman, and C. Plaisant, "*Designing the user interface: strategies for effective human computer interaction,"* Pearson Education, 2005.

[8] R. Moreno, R.E. Mayer, and J.C. Lester, "Life-like pedagogical agents in constructivist multimedia environments: cognitive consequences of their interaction," *ED-MEDIA Proceedings*, p.741-746, AACE Press, *2000*.

[9] M. Moundridou, and M. Virvou, "Evaluating the persona effect of an interface agent in a tutoring system," *Journal of Computer Assisted Learning*, 18, p253-261, Blackwell Science, 2002.

[10] H. Maldonado, J.R. Lee, S. Brave, C. Nass, H. Nakajima, K. Yamada, I. Iwamura, and Y. Morishima, "We learn better together: enhancing elearning with emotional characters," *In:* T. Koschmann, D. Suthers and T. W. Chan. Eds. *Computer Supportive Collaborative Learning: The Next Ten Years*, Proceedings of the Sixth International Computer Supported Collaborative Learning Conference, p408-417, Erlbaum, 2005.

[11] R. Catrambone, J. Stasko, and J. Xiao, "Anthropomorphic agents as a user interface paradigm: experimental findings and a framework for research," *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, p166-171, 2002.

[12] J. Xiao, R. Catrambone, and J. Stasko, "Be quiet? evaluating proactive and reactive user interface assistants," *Proceedings of INTERACT*, IOS Press, p383-390, *2003*.

[13] P. David, T. Lu, S. Kline, and L. Cai, "Social Effects of an Anthropomorphic Help Agent: Humans Versus Computers," Cyber Psychology and Behaviour, 10, 3. Mary Ann Liebert Inc, 2007.

[14] J. Forluzzi, J. Zimmermann, V. Mancuso, and S. Kwak, "How Interface Agents Affect Interaction Between Humans and Computers," Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces. Helsinki, Finland, Aug. 22-25, 2007.

[15] P. Murano, "Why anthropomorphic user interface feedback can be effective and preferred by users," *7th International Conference on Enterprise Information Systems*, (c) – INSTICC, Miami, 2005.

[16] P. Murano, "Anthropomorphic vs non-anthropomorphic software interface feedback for online factual delivery," *7th. International Conference on Information Visualisation,* IEEE, p. 138, London, 2003.

[17] P. Murano, "Anthropomorphic vs non-anthropomorphic software interface feedback for online systems usage," 7th ERCIM International workshop on user interfaces for all, p339-349, Paris, 2002a.

[18] P. Murano, "Effectiveness of mapping human-oriented information to feedback from a software interface," *Proc. 24th International Conference on Information Technology Interfaces*, Cavtat, Croatia, 2002b.

[19] P. Murano, "A new software agent 'learning' algorithm," *People in Control: An International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, IEE, Manchester, 2001a.

[20] P. Murano, "Mapping human-oriented information to software agents for online systems usage," *People in Control: An International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, IEE, Manchester, 2001b.

[21] N. Hayes, "Doing Psychological Research," Open University Press, 2000.

[22] D.M. Dehn, and S. van Mulken, "The impact of animated interface agents: a review of empirical research," *International Journal of Human-Computer Studies,* 52: p1-22, *2000*.

[23] J.J. Gibson, "*The ecological approach to visual perception,*"Houghton Mifflin Co, 1979.

[24] D.A. Norman, "Affordance, conventions, and design," *Interactions*, 1999, May-June, p.39-42, 1999.

[25] D.A. Norman, "*The design of everyday things*,"Basic Books, 2002.

[26] H.R. Hartson, "Cognitive, physical, sensory and functional affordances in interaction design," *Behaviour and Information Technology*, Sept-Oct, 22 (5), p.315-338, 2003.

[27] W.W. Gaver, "Technology affordances," *Proceedings of the ACM, CHI 91, Human Factors in Computing Systems Conference*, April 27 – May 2, New Orleans, Louisiana, USA, p79-84, 1991.

[28] J. McGrenere, and W. Ho, "Affordances: clarifying and evolving a concept," *Proceedings of Graphics Interface*, May, Montreal, Canada, 2000.

[29] P. Murano, and T. Sethi, "Anthropomorphic User Interface Feedback in a Sewing Context and Affordances", International Journal of Advanced Computer Science and Applications, Vol 2, Issue 4, April 2011.

Dr Pietro Murano is a Computer Scientist at the University of Salford, UK. Amongst other academic and professional qualifications he holds a PhD in Computer Science. His specific research areas are in Human Computer Interaction and Usability of software systems. He publishes regularly at refereed international levels and also supervises PhD level work at the University of Salford. Dr Murano is also a full member of the British Computer Society, a Chartered Engineer and is FEANI registered. Further Dr Murano is a reviewer for various international conferences/journals. (www.pietromurano.org).

Anthony Gee was a Computer Science and Information Systems student at the University of Salford. He has done Computer Science work in various areas including Human Computer Interaction.

Prof Patrik O'Brian Holt holds a PhD as well as other academic qualifications. He is a Research Professor of Computing (Cognitive Engineering) at Robert Gordon University, Aberdeen. His research interests include interactive visualisation of complex data and cognitive modelling of human performance. Prof. Holt is the Director of the Joint Research Institute for Computational Systems which is part of the Northern Research Partnership in Engineering, a Scottish research pooling initiative.