Ergonomics
Vol. 55, No. 9, September 2012, 1059–1072

Taylor & Francis
Taylor & Francis Group

Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults

Richard Pak*, Nicole Fink, Margaux Price, Brock Bass and Lindsay Sturre

Department of Psychology, Clemson University, Clemson, SC, USA

(*Received 30 January 2012; final version received 3 May 2012*)

This study examined the use of deliberately anthropomorphic automation on younger and older adults' trust, dependence and performance on a diabetes decision-making task. Research with anthropomorphic interface agents has shown mixed effects in judgments of preferences but has rarely examined effects on performance. Meanwhile, research in automation has shown some forms of anthropomorphism (e.g. etiquette) have effects on trust and dependence on automation. Participants answered diabetes questions with no-aid, a non-anthropomorphic aid or an anthropomorphised aid. Trust and dependence in the aid was measured. A minimally anthropomorphic aide primarily affected younger adults' trust in the aid. Dependence, however, for both age groups was influenced by the anthropomorphic aid. Automation that deliberately embodies person-like characteristics can influence trust and dependence on reasonably reliable automation. However, further research is necessary to better understand the specific aspects of the aid that affect different age groups. Automation that embodies human-like characteristics may be useful in situations where there is under-utilisation of reasonably reliable aids by enhancing trust and dependence in that aid.

Practitioner Summary: The design of decision-support aids on consumer devices (e.g. smartphones) may influence the level of trust that users place in that system and their amount of use. This study is the first step in articulating how the design of aids may influence user's trust and use of such systems

Keywords: automation; trust; ageing; personification; health; diabetes

Introduction

Research in social computing has shown that users, often unknowingly, will treat their computers as social beings, attributing human-like characteristics to them (Nass *et al.* 1993). This effect of anthropomorphism is large enough in some cases that social rules that govern human–human interaction have been found to also apply in human–computer interactions (e.g. Nass *et al.* 1996). Anthropomorphism of technology has been extensively studied and even capitalised upon to make interfaces more user-friendly and approachable with limited success (e.g. Microsoft's Clippy).

Lee (2007) and Lee and See (2004) discussed the importance of understanding the boundary conditions under which social emotions influence users' response to technology, particularly automation. By its nature, automated systems can possess human-like traits even if that was not the design intent (e.g. agency). The aim of this study was to examine how a deliberately anthropomorphised health decision aid would affect trust and dependence in the aid, as well as performance, in younger and older adults.

Based on our review of the literature on anthropomorphism and automation, we hypothesised that: (a) anthropomorphic aids would elicit more trust than a non-anthropomorphic aid, (b) this higher trust engendered by the anthropomorphic aid would manifest itself as dependence and (c) anthropomorphism would have greater effects for older adults (i.e. dependence and performance will be enhanced).

Anthropomorphic interface agents

Interface anthropomorphism occurs when computer automation is given human-like traits and characteristics (Gong 2008). The implementation of anthropomorphism can take many forms, such as: how similar the interface appears visually to be human (e.g. Catrambone *et al.* 2002), the language it uses (e.g. Schulman and Bickmore 2009), how it sounds (Nass and Lee 2001) and the presence of life-like animation (Berry *et al.* 2005). Studies have examined the effect of anthropomorphic agents (usually compared to a text-based interface) on user's perceptions of an interface (Sproull *et al.* 1996) and subjective ratings of satisfaction with the system (Wexelblat 1998, Murano 2003) with mixed findings.

*Corresponding author. Email: richpak@clemson.edu

ISSN 0014-0139 print/ISSN 1366-5847 online

http://dx.doi.org/10.1080/00140139.2012.691554
http://www.tandfonline.com

Sproull *et al.* (1996) directly compared participants' responses to the image of a face versus plain text in an interface. In their experiment participants consulted with a computer-based career counsellor. Participants judged the anthropomorphic counsellor to have more personality attributes compared to the text-based counsellor despite the equivalence in information. Because of the nature of the task (career counselling with no correct answer) objective performance in the task was not examined.

Anthropomorphic interfaces also appear to influence older adults' perceptions. Bickmore *et al.* (2005) compared whether older adults would increase their physical activity (walking) when they interacted with (1) an anthropomorphic agent that engaged in dialogue, (2) an agent that did not engage in dialogue or (3) no agent, instead received pamphlets about the benefits of walking. Older adults in the anthropomorphic agent condition had a significant increase in the number of steps over the two-month period, while the control group (pamphlets only) did not. These findings suggest that agents can be used and accepted by older adults, as well as lead to significant behaviour change. However, it was unclear if the beneficial effect of anthropomorphism was due to the purported social aspects of the relational agent or simply more frequent reminders to walk (compared to reading a pamphlet). In addition, the primary function of the relational agent was not as an automated aid (that might help the user work through a decision-making problem) but more as a coach to enhance motivation and engagement in the task.

Despite the evidence supporting the positive effects of anthropomorphism, some studies suggest a negative effect. Bengtsson *et al.* (1999) compared having a computer partner (with varying levels of anthropomorphism) or a human partner on a decision-making task. Interactions with a human partner resulted in the highest level of social judgments (likability, truthfulness and perceived competence) compared to a computer interface. More puzzling was that within the computer partner condition, more anthropomorphism was associated with comparatively negative social judgments.

More recently, Yee *et al.* (2007) conducted a meta-analysis of 25 studies and found that human-like anthropomorphism (specifically the presence of a face or not) had consistent yet small effects on subjective preferences and in fewer cases objective measures of performance. They found that only about 2.6% of the variance in subjective measures was attributable to the presence of an anthropomorphic agent in the interface ($r = 0.16$). The variance accounted for in performance measures by anthropomorphism was even smaller at 0.008% ($r = 0.09$).

However, there are reasons why the results of the meta-analysis should be approached with some caution. First, as Yee *et al.* (2007) have acknowledged, meta-analytic techniques attempt to generalise across a disparate set of studies that contain different tasks, stimuli and dependent variables. In a more specific task context, subject population or dependent variable, it is logical to assume that effect sizes may be larger than what a meta-analysis might suggest.

In an attempt to resolve some of the disparate findings in anthropomorphic effects, Gong (2008) manipulated level of anthropomorphism and type of task. Gong (2008) surmised that methodological inconsistencies led to the previously reported paradoxical findings. First, the modality (computer versus person) and anthropomorphism (text versus face) were confounded; that is, the face-to-face interactions were always high in anthropomorphism while the computer-based text interactions were always low. In addition, the tasks used in previous studies did not seem amenable to influence by social aspects (e.g. web-based searching). This may explain much of the inconsistency and weak effects (c.f. Yee *et al.* 2007) found in the literature on anthropomorphic agents. Gong (2008) concluded that social responses did increase with increasing level of anthropomorphism and effects were strongest for situations that were amenable to social influence.

Gong's (2008) study clarified the conditions under which anthropomorphic interfaces may be maximally beneficial in terms of social influence, however it, along with most prior research, did not examine effects on performance. The nature of Gong's (2008) study precluded any assessment of performance, per se, because it was primarily a social judgment task that did not necessarily have a correct answer. Given the potential usefulness of anthropomorphic agents or interfaces in manipulating perceptions of computerised systems such as trust, would some of the social effects of anthropomorphic interfaces be visible as better decision-making?

Decision support systems (DSS)

Automation is the execution by a machine agent (usually a computer) of a function that was previously carried out by a human (Parasuraman and Riley 1997). Automation (of the DSS type) can be beneficial because it can alleviate the complex cognitive burden of interpreting cues and analysing possible outcomes. Examples include global positioning systems (GPS) that advise a driver the best route to take, or medical decision aids that analyse a myriad of health indicators and instruct the doctor on the best course of action. Whether a user opts to follow the automation's instructions can be influenced by a number of factors, particularly the level of trust the user has placed in the system (Lee and Moray 1994, Meyer 2001, Wiegmann *et al.* 2001).

Just as trust is an important factor in how humans deal with other humans, it also can determine how users interact with computerised systems. Trust can be influenced by both system-factors (e.g. reliability, level of automation) and user-related factors (experience, age). For example, a system that consistently gives unreliable advice or automates a task that is easy may be perceived as less trustworthy by the user (Madhavan *et al.* 2006). A lack of trust in the decision aid may explain why users tend to not rely on automated decision aids. Studies with both younger (e.g. Dzindolet *et al.* 2002) and older adults (Ezer *et al.* 2008) show that users tend to under-use automated aids although they are as accurate as they are in the task. This under-use of a reasonably reliable automated aid occurred even when there was a high cost associated with obtaining a wrong answer.

One aspect of automation or DSSs that has been less well researched is the role of automation appearance on human-automation trust and dependence. That is, how does the appearance of the automated system, independent of other system-factors such as reliability, affect trust and behaviour? Although automation that uses the speech modality have been studied (e.g. Cassell and Bickmore 2000), little research has examined the role of a human-like appearance on trust in an automated aid and performance, particularly for older adults. Guidance about the possible role of anthropomorphic automation on trust and performance comes from Parasurman and Miller (2004) who examined the role of automation etiquette, or how politely the DSS advised the user. They found that trust and user performance was higher with polite automation (one that did not interrupt the user and exhibited patience) than impolite automation. These results show promise in the idea that some level of anthropomorphism can engender trust in an automated aid, but it should be noted that the authors deliberately chose to limit their manipulation to only communication style; no other aspect of the aid appeared anthropomorphic. Thus, it is still unclear if a moderate to high level of visual anthropomorphism has similar effects on trust and performance as observed by Parasuraman and Miller (2004).

It is implied that the previously mentioned studies are primarily history-based measures of trust; that is trust that is developed after some period of interaction (Merritt and Ilgen 2008). This is contrasted with dispositional trust which is the level of trust that a user might bring to a system that they have not used before. Dispositional trust may be more related to pre-existing notions of technology's reliability or other personality difference. The focus of the current study is on history-based trust and how human-automation behaviour is altered after some exposure and interaction with a decision aid.

Current study

The aim of this study was to explore how medical decision-making would be affected by an anthropomorphic aid. We were primarily interested in the level of trust engendered by intentionally anthropomorphic automation, dependence on the automation and decision-making performance. This study is set in the context of at-home diabetes management. Managing diabetes requires constant monitoring of one's blood glucose levels (A1C's), blood pressure and other diabetic-related symptoms (e.g. swollen feet due to poor circulation). Furthermore, diabetics are often faced with time-dependent decisions that need to be made before a doctor can be consulted. They may also be under the stress of symptoms that can affect reasoning ability (e.g. low A1C's leading to a hypoglycaemic episode). The complexity of managing diabetes was reflected in a study by Glasgow and Strycker (2000) who found considerable variability in behavioural self-management of diabetic patients. In some cases, less than 40% of diabetic patients received care that met recommended guidelines. Thus, the at-home management of diabetes represents a complex task that can benefit from the adoption and use of decision-making automation to enhance patient education and adherence to guidelines. Using the levels of automation concept proposed by Parasuraman *et al.* (2000) the aid used in the current study would be classified as a stage 3 aid – one that proposes a course of action but allows the patient to veto.

Despite the domain of the task (diabetes management), we deliberately selected non-diabetics for this study to control the amount of expertise that participants would bring to the task. For example, older diabetics would naturally have much more information and experience with the condition compared to younger diabetics. This decision severely limits the generalisability of our study in regards to medication aids but we felt it was important to study the unique influence of automated aids on performance.

As suggested by Gong (2008) and Parasuraman and Miller (2004), it is expected that the human-like anthropomorphic agent will evoke a positive social response such that the diabetes decision maker will come to depend on the automation more frequently, perceive it to be more trustful, be more confident in their answers and perform better (faster and more accurately) in the decision-making task than participants who are presented with no decision aid or a non-anthropomorphic aid. It should be noted that the term 'anthropomorphic' could encompass quite a wide continuum of 'human-like' characteristics (from very shallow appearance and simple behaviour on the low end to highly complex, photo-realistic and interactive). For the purposes of this study, anthropomorphic was operationalised in a very shallow and easily implementable graphical manner that focused on the presence of a pseudo-agent, not the behaviour, language or other complex interaction.

Method

Participants

Forty-five undergraduate students (20 females, 25 males) ranging in age from 18 to 26 ($M = 19.82$, SD = 1.54) and 45 older adults (28 females, 17 males) ranging in age from 64 to 83 ($M = 72.98$, SD = 5.18) participated in the study. The older adults were recruited from the community through newspaper advertisements. None of the participants reported that they had diabetes and all reported more than five years of experience with computers. The undergraduate participants received extra credit in a psychology course for their participation while the older adults received $7/hour for their participation.

Design

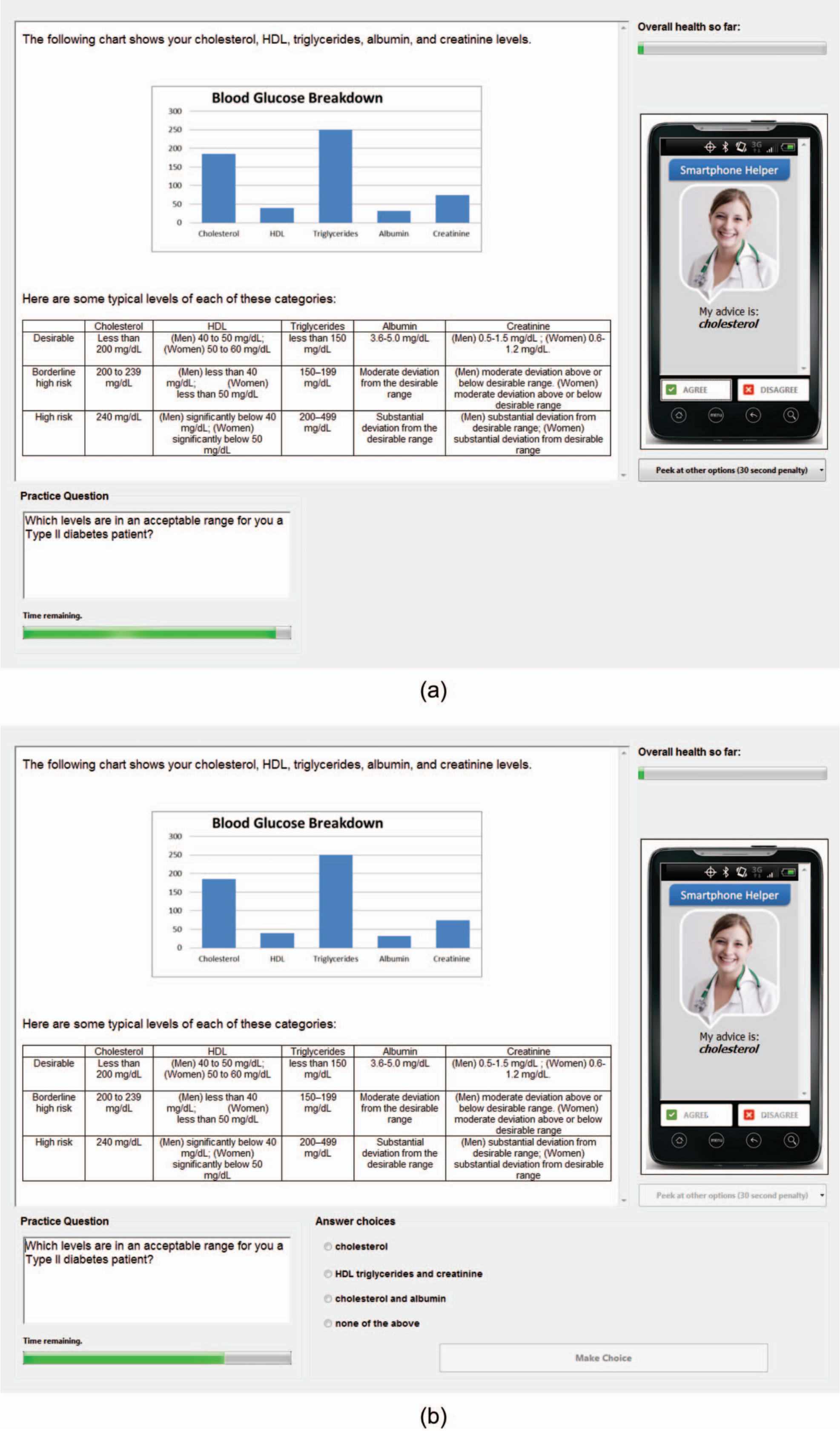
The study was a 2 (age group: young, old) × 3 (aid: no-aid, text-only non-anthropomorphic aid and anthropomorphic aid (heretofore referred to as 'anthro')) factorial with age group as a quasi-independent grouping variable and aid type as a between-subjects variable. The dependent variables were task time, proportion correct and confidence in the chosen answer (after every trial; Likert scale). In addition, for the conditions with an aid (non-anthro, anthro), trust in the aid (after every trial; Likert scale) was assessed. An 'objective' measure of trust (what we call 'behavioural trust') was also computed by examining the pattern of dependence on the aid: proportions of times that participants agreed or disagreed with the smartphone aid or peeked at alternatives choices. Peeking behaviour could be considered a type of automation verification. Automation verification or monitoring as a more objective measure of trust has been used in prior studies (Moray and Inagaki 2000, Bahner *et al.* 2008). The behavioural trust measure is described in more detail in the results section. Finally, subjective workload (NASA TLX; Hart and Staveland 1988) measures were collected in all aid conditions.

Task and materials

The task was to answer 30 questions related to a diabetes scenario. All necessary information for answering the question was provided in the scenario. The questions were adapted from scenarios in a diabetes education workbook (Drucquer and McNally 1998) and an online diabetes support forum. Participants viewed the task on 19-inch LCD monitors and made all responses using the mouse. They were seated in office chairs about 18–24 inches from the screen in an office environment. In the no-aid condition participants answered the questions using only the information contained in the scenario or they guessed. In the two aid-present conditions (Figure 1), participants were told that they had access to an advanced smartphone app that could help them make their decisions. In the non-anthro aid condition the text was accompanied by an image of gears (Figure 2) while in the anthro condition the gears were replaced with a static smiling female doctor. The decision to use such extremes was to maximise the possibility of detecting any effect, if present. Participants in the no-aid condition did not see the image of the smartphone; only the scenario, question and possible courses of action.

Reliability was 67% so that the aid's suggestion was incorrect in 10 of the 30 questions (see Table 1 for a distribution of aid failures). This was chosen to be high enough for the automation to be perceived as trustable, but low enough to not be perceived as perfect. The reliability is between what Parasuraman and Miller (2004) called low reliability (60%) and high reliability (80%). The first aid failure did not occur until the 8th question in order to let users build trust in the system and the remaining aid failures were placed randomly throughout the remaining 22 questions.

Figure 1(a) illustrates the programme that presented the diabetes question. The upper left contained a large text box with the scenario that set up the question in the lower left. Below the question was a bar that slowly moved left to indicate how much time was available to answer the question (two minutes for younger adults, four minutes for older adults). Since the task required quite a bit of reading and mouse usage, more time was allotted for older adults to compensate for normative age-related declines in perceptual and psychomotor speed (e.g. Salthouse 1992). On the upper right was another bar graph that indicated 'overall health'. As participants answered questions



The following chart shows your cholesterol, HDL, triglycerides, albumin, and creatinine levels.

Here are some typical levels of each of these categories:

	Cholesterol	HDL	Triglycerides	Albumin	Creatinine
Desirable	Less than 200 mg/dL	(Men) 40 to 50 mg/dL; (Women) 50 to 60 mg/dL	less than 150 mg/dL	3.6-5.0 mg/dL	(Men) 0.5-1.5 mg/dL ; (Women) 0.6-1.2 mg/dL.
Borderline high risk	200 to 239 mg/dL	(Men) less than 40 mg/dL; (Women) less than 50 mg/dL	150–199 mg/dL	Moderate deviation from the desirable range	(Men) moderate deviation above or below desirable range. (Women) moderate deviation above or below desirable range
High risk	240 mg/dL	(Men) significantly below 40 mg/dL; (Women) significantly below 50 mg/dL	200–499 mg/dL	Substantial deviation from the desirable range	(Men) substantial deviation from desirable range; (Women) substantial deviation from desirable range

Practice Question

Which levels are in an acceptable range for you a Type II diabetes patient?

The following chart shows your cholesterol, HDL, triglycerides, albumin, and creatinine levels.

Here are some typical levels of each of these categories:

	Cholesterol	HDL	Triglycerides	Albumin	Creatinine
Desirable	Less than 200 mg/dL	(Men) 40 to 50 mg/dL; (Women) 50 to 60 mg/dL	less than 150 mg/dL	3.6-5.0 mg/dL	(Men) 0.5-1.5 mg/dL ; (Women) 0.6-1.2 mg/dL.
Borderline high risk	200 to 239 mg/dL	(Men) less than 40 mg/dL; (Women) less than 50 mg/dL	150–199 mg/dL	Moderate deviation from the desirable range	(Men) moderate deviation above or below desirable range. (Women) moderate deviation above or below desirable range
High risk	240 mg/dL	(Men) significantly below 40 mg/dL; (Women) significantly below 50 mg/dL	200–499 mg/dL	Substantial deviation from the desirable range	(Men) substantial deviation from desirable range; (Women) substantial deviation from desirable range

Practice Question

Which levels are in an acceptable range for you a Type II diabetes patient?

Answer choices

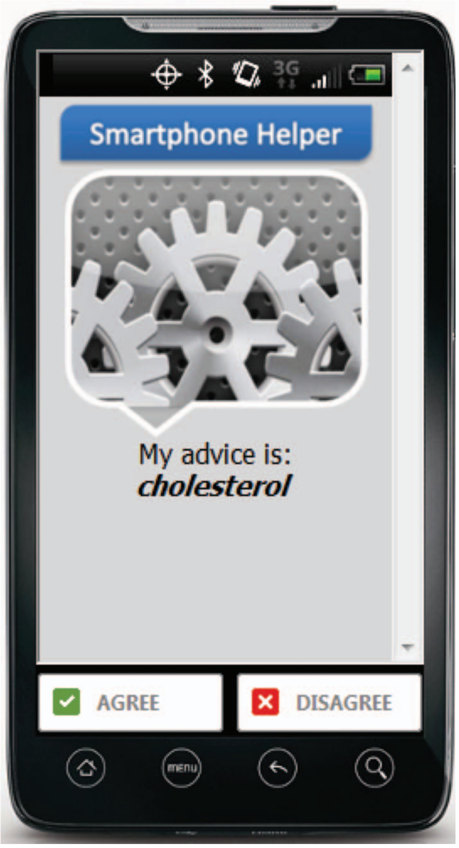
○ cholesterol

○ HDL triglycerides and creatinine

○ cholesterol and albumin

○ none of the above

Figure 1. (a) Experiment screen, anthropomorphic condition. Scenario in upper left, question in lower right, decision aid on right. (b) The participant has clicked 'disagree' and is presented with three other options (in addition to the suggested course).


Smartphone Helper

My advice is:
cholesterol

Figure 2. Image of the smartphone aid in the non-anthro automation condition.

correctly, this bar grew in length. The health score was a combination of whether the answer was correct, and how quickly the participants answered (e.g. a higher score was obtained on each trial if the correct answer was chosen quickly). Finally, on the lower right was the decision aid that presented its suggestion. Within the interface were the options to agree or disagree with the suggestion. Agreeing finished the trial, but disagreeing revealed further options on the lower portion of the screen (Figure 1b). Below the interface was the option to 'peek' at the other question answers that were available. This option was meant to simulate asking a friend or doctor about other options and could be considered a type of distrust of automation (i.e. reliance).

Figure 2 shows how the smartphone aid differed in the text condition. The response possibilities (agree, disagree and peek) were identical in both conditions. Finally, in the no-automation condition, the smartphone was hidden from view. When the question and scenario appeared, the four possible answers were shown immediately with the scenario and question.

Procedure

The procedure is illustrated in Figure 3. Participants in the no-aid condition were told that they were recently diagnosed with Type II diabetes and the problems presented in the experiment should be treated like everyday problems they might encounter. If they did not know the correct answer, they should guess. Participants in both aid-present conditions were told that they were recently diagnosed with Type II diabetes and in an attempt to better manage their health, they have downloaded an app on their smartphone designed to assist them in making everyday diabetes management decisions. They were told that just like technological aids in everyday life, the system is not always reliable.

In the no-aid condition, participants were shown the scenario, question and four possible answers. As soon as the question appeared the timer bar on the lower left started counting down two or four minutes. If the participant did not respond in time, their answer, whether it was correct or incorrect, did not contribute to their health score

Table 1. Distribution of aid failures for text and anthro conditions by question.

Question sequence	Aid's suggestion
Practice 1	Correct
Practice 2	Correct
Practice 3	Correct
1	Correct
2	Correct
3	Correct
4	Correct
5	Correct
6	Correct
7	Correct
8	Incorrect
9	Correct
10	Incorrect
11	Correct
12	Correct
13	Incorrect
14	Incorrect
15	Correct
16	Correct
17	Incorrect
18	Correct
19	Incorrect
20	Correct
21	Incorrect
22	Correct
23	Incorrect
24	Correct
25	Correct
26	Incorrect
27	Correct
28	Correct
29	Incorrect
30	Correct

although the response was recorded for analysis. After making a response, the question, scenario and responses disappeared and a Likert scale appeared asking about their confidence in the answer they selected. After participants responded to the confidence question, they were told whether their answer was correct or incorrect. If their answer was incorrect they were told what the correct answer was. After confirming this feedback, the next question was presented.

Participants in the aid conditions followed a similar procedure, but their response options were different. After being presented with the scenario, question and aid advice, participants could do one of the three things: (1) agree with the aid's advice, 2) disagree or 3) peek at the other options. If participants *agreed* with the aid, they were presented with a Likert scale asking about their confidence in their choice and their trust in the aid. After receiving feedback of correctness of their response they were presented with the next trial. If participants instead *disagreed* with the aid, the lower right portion of the screen revealed three alternate answers (in addition to what was presented by the aid) for a total of four choices. After participants made a choice, they responded to the confidence and trust Likert scale and then received feedback. Finally, participants could (before they agree or disagree) *peek* at the possible answers to the question. Peeking allowed participants to view other answers without committing to choosing disagree, which was the only other way to see the other answers. When participants clicked the peek button, they were presented with the four possible answers in a drop-down box and then only allowed to either agree or disagree with the aid's suggestion. However, clicking the peek button resulted in a 30 s penalty. We instituted this penalty to simulate the time penalty that one might suffer if they double-checked or asked someone for help (cost of verification; Ezer *et al*. 2008). After answering all 30 questions, participants completed the computerised version of the NASA TLX subjective workload scale.

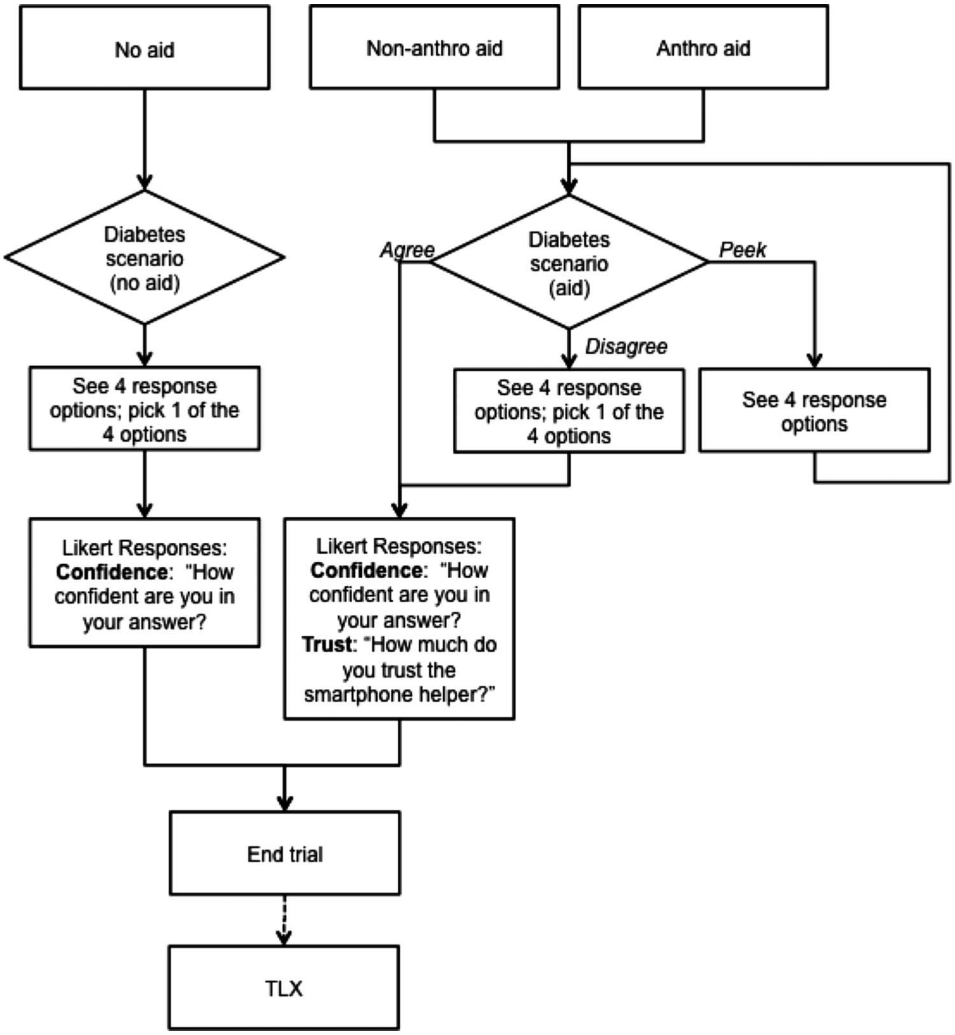

No aid
Non-anthro aid
Anthro aid
Diabetes scenario (no aid)
Agree
Diabetes scenario (aid)
Peek
Disagree
See 4 response options; pick 1 of the 4 options
See 4 response options; pick 1 of the 4 options
See 4 response options
Likert Responses: **Confidence**: "How confident are you in your answer?
Likert Responses: **Confidence**: "How confident are you in your answer? **Trust**: "How much do you trust the smartphone helper?"
End trial
TLX

Figure 3. Procedure for each condition (no-aid, non-anthro aid and anthro aid).

Results

Table 2 shows the categorised dependent variables by condition (no-aid, non-anthro aid and anthro aid). The variables were organised into two main categories: performance variables common to all three conditions (task time, proportion correct, confidence in answer and perceived workload), and subjective and 'objective' trust variables that only applied to the two aid conditions: perceived trust and a measure of dependence we call behavioural trust (primarily composed of how participants responded to the agree/disagree/peek).

Although we did not intend to investigate the effect of participant gender in our study, we discovered after the study was run, the possibility of gender effects because of our choice to use a female human agent in the anthro condition. Prior research has shown that gender of anthropomorphic agent can affect some aspects of perceptions (e.g. Rosenbergkima *et al.* 2008), thus our analysis includes gender as a grouping variable. First, we examined the dependent variables related to task performance and workload that were common to all three conditions: task time, proportion correct, perceived workload and confidence in the chosen answer. An alpha level of $p < 0.05$ was the criterion for statistical significance.

Performance and workload

To examine differences in the performance and workload as a function of age group, condition and gender, a 2 (age group: young, old) × 3 (condition: no-aid, non-anthro aid and anthro aid) × 2 (gender: male, female) ANOVA was conducted. For decision accuracy (proportion correct) there was a significant main effect of condition,

Table 2. Performance across the categorised dependent variables by condition and (no-aid, non-anthro aid and anthro aid).

	Younger (N = 45)						Older (N = 45)					
	No-aid		Non-anthro aid		Anthro aid		No-aid		Non-anthro aid		Anthro aid	
	M	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Performance and workload												
Proportion correct	0.49	0.08	0.62	0.10	0.65	0.06	0.54	0.15	0.63	0.08	0.66	0.08
Task time (s)	40.47	12.96	42.88	16.16	26.76	9.64	89.18	26.41	78.29	24.58	75.96	25.92
Confidence in answer (after every trial)[1]	4.76	0.57	4.99	0.93	5.48	0.88	4.41	1.27	5.31	1.47	5.23	0.75
Overall workload (TLX)[2]	63.87	12.32	62.09	17.36	53.42	8.29	68.69	12.52	63.80	13.04	63.93	15.73
Perceived trust												
Trust in aid (after every trial)[1]	–	–	4.46	1.18	5.40	0.88	–	–	5.20	0.88	5.12	0.84
Behavioural trust (dependence)[3]	–	–	2.98	0.45	3.29	0.26	–	–	3.09	0.39	3.34	0.29

Notes: [1]Likert scale: 1 = not at all, 7 = completely. [2]Composite NASA TLX score. [3]Computed based on interaction with aid; higher is more trust (1 = disagree, 2 = disagree but peek, 3 = agree but peek, 4 = agree).

$F(2,78) = 18.75$, $\eta_p{}^2 = 0.33$, but no effects of age group or gender. Figure 4 illustrates proportion correct as a function of condition. Pairwise comparisons show that the accuracy was significantly lower in the no-aid condition compared to the two aid conditions. However, there was no difference in accuracy between the two aid conditions. In addition, the absence of a main effect or interaction with age group shows that in terms of accuracy both age groups were able to benefit from the presence of an aid.

There was a significant main effect of aid condition on task time, $F(2,78) = 4.50$, $\eta_p{}^2 = 0.10$. Pairwise comparisons showed that participants answered questions more quickly with the anthro aid compared to no-aid, but the difference between anthro and non-anthro aids were not different (Figure 5). Answer time also did not differ between the no-aid and non-anthro aid conditions. In light of the accuracy results, this shows a unique benefit of the anthro aid condition over the non-anthro aid condition: equal accuracy, but faster performance.

The analysis of confidence showed a significant main effect of condition on confidence, $F(2,78) = 4.54$, $\eta_p{}^2 = 0.76$. Pairwise comparisons showed that confidence was significantly higher in the anthro condition compared to the no-aid condition. No other pairwise comparisons were significantly different.

There was a significant main effect of aid condition on overall workload, $F(2,77) = 3.47$, $\eta_p{}^2 = 0.08$, but no main effects of age group or gender. However, the three-way interaction between age group, condition and gender was significant, $F(2,77) = 4.08$, $\eta_p{}^2 = 0.71$. This interaction is illustrated in Figure 6. The source of the interaction was a significant, age group × gender interaction within the non-anthro aid condition, $F(1,26) = 4.28$, $\eta_p{}^2 = 0.51$. Pairwise comparisons showed that within the non-anthro aid condition, younger females perceived higher workload than younger males, $F(1,13) = 4.83$, $\eta_p{}^2 = 0.43$. This interaction was not observed in the anthropomorphic or no-aid condition for younger adults (nor for older adults).

Perceived trust and dependence

Mean trust (measured after every trial)

The trust data was subjected to a 2 (age group) × 2 (aid condition) × 2 (gender) ANOVA. There was no main effect of any variable, but there was a significant age × condition interaction, $F(1, 52) = 5.04$, $\eta_p{}^2 = 0.60$, illustrated in Figure 7. For younger adults, trust was significantly lower with a non-anthro aid compared to an anthro aid. For older adults, there was no significant difference in trust by aids. In terms of trust, younger adults were subject to the anthropomorphic effect previously observed. However, older adults seem impervious to the effects of anthropomorphism.

Behavioural trust

In addition, participant's behaviour (whether they agree, disagree or peek) was analysed and used as an 'objective' measure of trust. The rationale was that if participants immediately agreed with the aid without peeking; that could

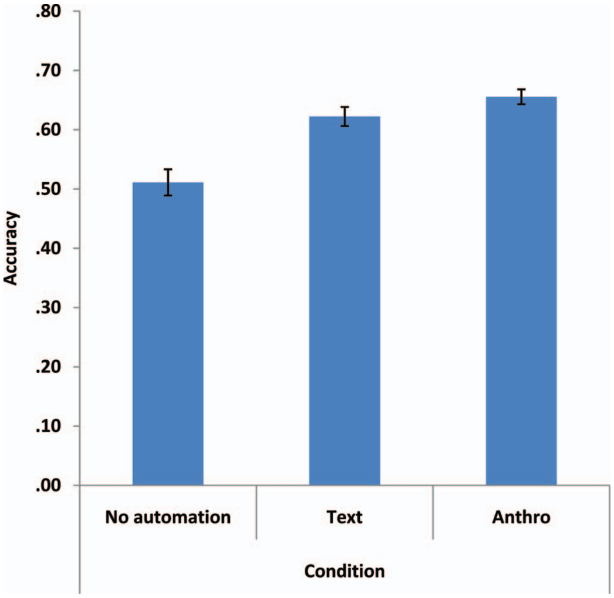


Figure 4. Proportion correct as a function of aid condition.

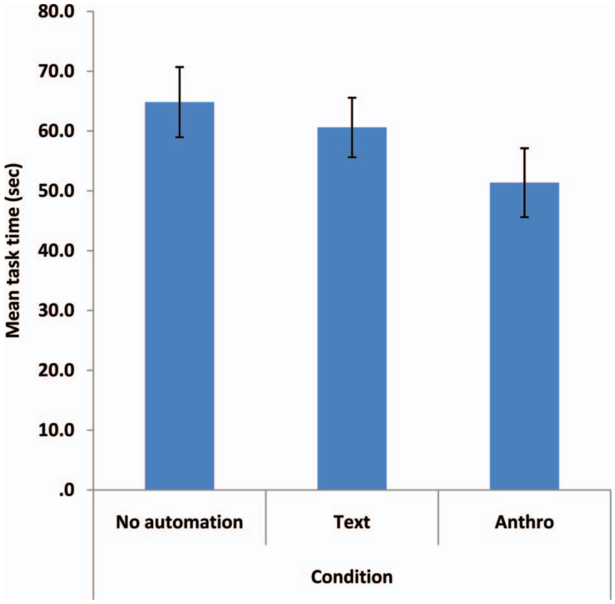


Figure 5. Mean task time as a function of aid condition.

be considered a high level of trust in the aid. However, if participants disagreed without peeking it would indicate a complete lack of trust. If participants peeked (whether they agreed or disagreed) it could represent moderate levels of trust. Behavioural trust was a scale from 1 to 4. If participants immediately clicked disagree that was given a value of 1 (no trust). If they peeked and eventually clicked disagree, that trial was assigned a 2 (moderate distrust). Peeking and agreeing was assigned a 3 (trust but verify) and clicking agree a 4 (trust). This measure of objective trust was significantly correlated with subjective trust ($r = 0.35$). A 2 (age group) × 2 (aid condition) × 2 (gender) ANOVA showed only a main effect of condition on behavioural trust, $F(1,52) = 11.06$, $\eta_p^2 = 0.90$, illustrated in Figure 8. Regardless of age or gender, participants in the non-anthro aid condition had a lower level of behavioural trust compared to participants in the anthro condition.

Finally, Lee and Moray (1992, 1994) analysed the relationships between trust, confidence and automation usage in a process-control task and found that when trust exceeded confidence, participants tended to use automation but

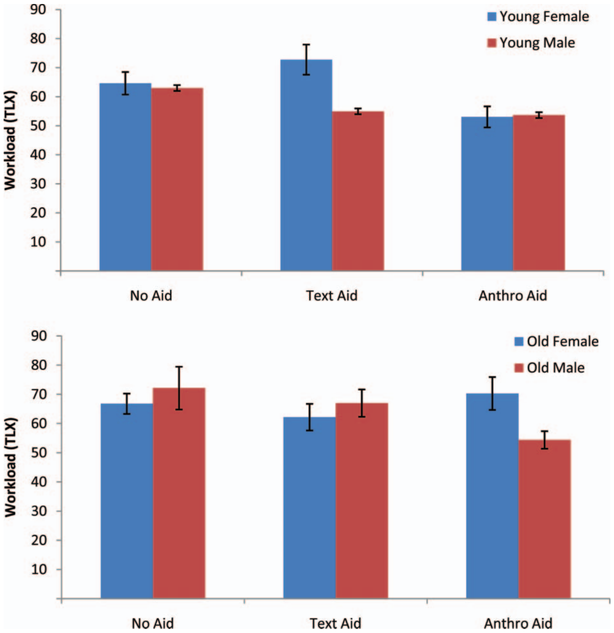


Figure 6. Perceived workload as a function of aid condition, gender and age group. (a) Younger adults, (b) older adults.

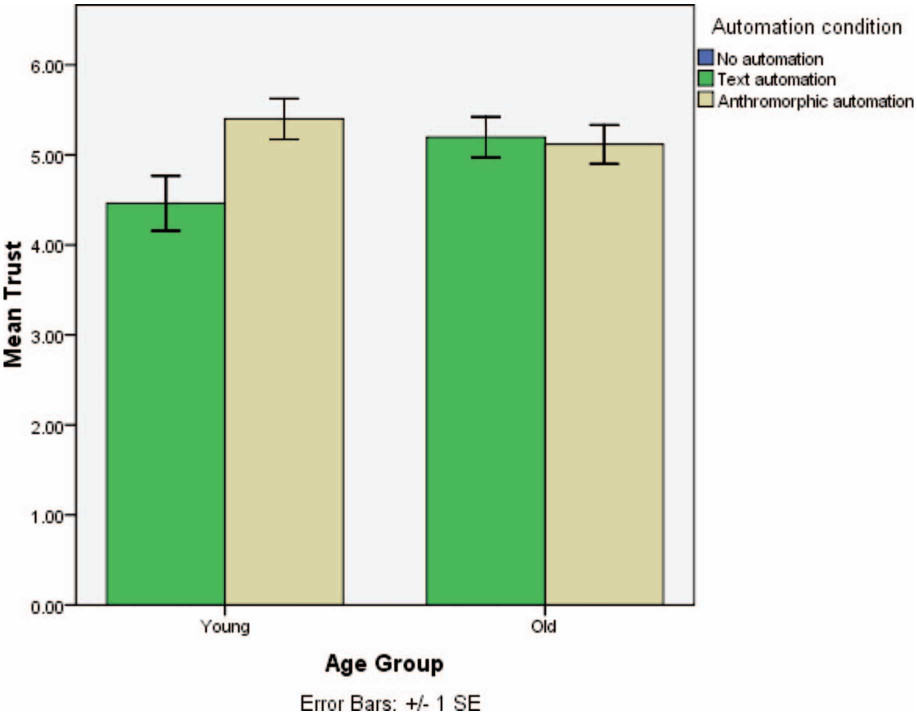


Figure 7. Mean trust as a function of aid condition.

when the reverse was true, they tended to ignore automation. We analysed trial-level data (by condition and age group) by first categorising each trial based on whether trust (scaled from 1 to 7) was less than or greater than or equal to confidence (also scaled from 1 to 7). Consistent with prior literature, when trust exceeded confidence, participants agreement with the aid was significantly higher than when trust fell below confidence ($M = 0.80$ and $M = 0.34$, respectively; $F(1{,}109) = 73.9$, $\eta_p^2 = 0.40$.

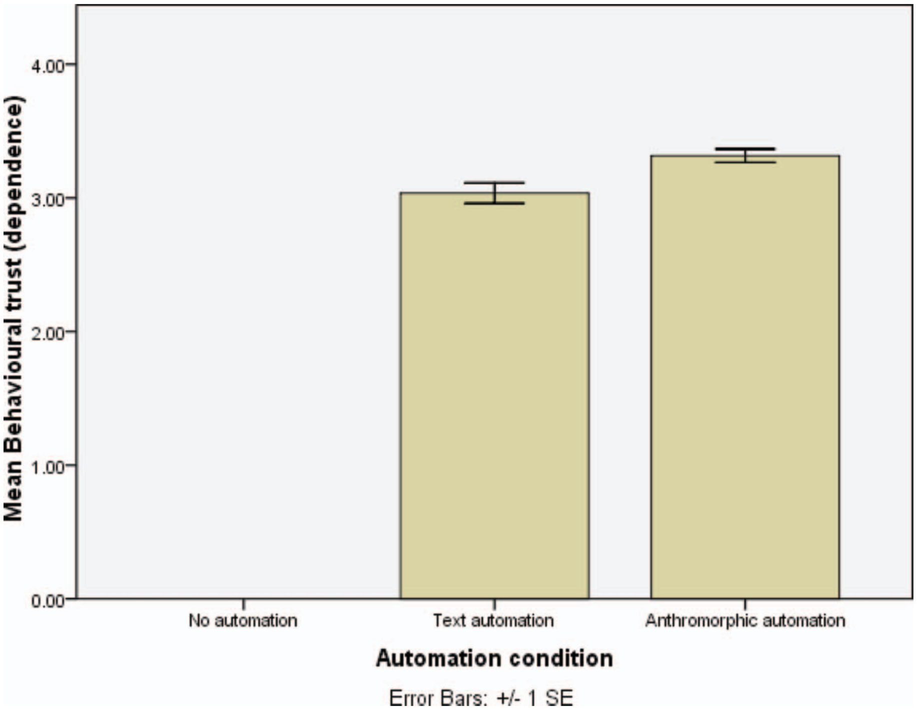


Figure 8. Mean behavioural trust as a function of aid condition.

General discussion

Consistent with the theory of computers as social actors (Nass *et al*. 1993), the results of this study showed that the simple inclusion of an image of a person can significantly alter perceptions of an automated aid even when there is no difference in aid reliability or information presentation. Younger participants' trust was enhanced when using an anthropomorphic aid compared to a non-anthro aid. Younger and older participants' question answer time was reduced with an anthropomorphic aid over a non-anthro aid. A plausible explanation is that increases in trust led to an increased dependence on the aid, which led to better and faster performance. The finding of an increase in trust mirrors Parasurman and Miller's (2004) study that showed another minimal type of anthropomorphism (politeness) increased trust and performance.

Our mixed results with older adults also stress the need to more precisely examine the design of age-sensitive anthropomorphic aids. In a study examining age-related differences in trust with non-anthropomorphised medical decision aids, Ho *et al*. (2005) found that older adults tended to trust (non-anthropomorphised) decision aids more than younger adults. A speculative possibility, based on our data, is that older adults' trust is less malleable in the presence of an agent that does not share such an obvious trait (age) with the user. Nass and Lee (2001) earlier found some confirmation of the so-called similarity-attraction hypothesis in the context of computerised synthetic voices. They found that when the personality of a computerised voice matched the user, judgments of liking and credibility were higher than when the computer/user voice personality was non-matched.

Another possible reason why most of the positive effects were only evident for younger adults is that the age of the image used in our study was a young (mid 20s) female and it may have induced stereotype threat in our older participants. Stereotype threat is the concept that an individual's awareness of the negative stereotype of his group negatively impacts his performance on a task that might confirm that stereotype (Steele 1997). Thus, when an older adult becomes primed or aware of a negative age-related stereotype (e.g. older adults are bad at complex decision-making) and they are put in a task situation that might confirm that (a complex decision-making task), increased anxiety leads to reduced performance. Hess *et al*. (2003) have observed this age-related equivalent of stereotype threat with regard to memory performance.

Another appearance-related factor is that of ethnicity. Ethnicity was neither controlled nor examined in our study (100% of the older participants were Caucasian and our decision aid was represented as a Caucasian female), but recent research by Qiu and Benbasat (2010) suggests that the ethnicity match between the decision agent and user may be more important than other factors such as gender. Their study looked at recommendation agents one might find on an e-commerce website. They found that perceived enjoyment and perceived usefulness of the recommendation agent was highest in ethnicity-matched conditions while gender matching had no effect. A future study can also examine other specific aspects of the anthropomorphic representation that have the most effect on automation interaction. If the presence of an anthropomorphic automation agent can engender social responses to automation (i.e. more trust), the effect may also be enhanced when the anthropomorphic agent appears more authoritative (more doctor-like than not).

One potential explanation for the unique anthropomorphic effects might be due to demand characteristics; that is, the participants in the anthropomorphised condition, having little knowledge of the domain and task, simply agreed with the friendly-faced anthro aid more thinking; that is what the experimenter wanted. However, since this was a between-groups design and participants were not able to see the other conditions, identical demand characteristics should also be evident in the non-anthro automation condition. Clearly, that is not what our data show. Instead, there seems to be a slight increase in the use of anthro automation compared to non-anthro automation. In addition, perceptions of trust were different. To suggest that demand characteristics were the sole driver in automation aid differences would require an explanation of why perceived trust also differed between automation-present conditions.

Finally, the deliberate selection of non-diabetic participants may be a controversial choice but we felt it was critical to control the amount of domain knowledge that participants had as it would naturally affect their trust in an aid. This design decision reduces the generalisability of our results in regards to the design of medical health decision aids but clarifies the investigation of the relationship between aid design and performance. In spite of this, the current study may have implications in the design of aids that improve patient self-care. Even beyond diabetes, patient adherence to recommended guidelines for self-care (which may include recommendations for diet, exercise and medications) may be less than 50% (Rapoff 1999). A highly reliable decision aid that provides decision support (integrating of information) and recommendations may alleviate non-adherence in self-care activities.

In addition, we did not assess participant's experience with smartphones (just technology in general). Age-related differences in smartphone ownership and experience may have affected the perceived utility of the aid used in this study. In the United States, as of 2011, 52% of those ages 18–29 owned a smartphone while ownership was only 11% for those over age 65 (Smith 2011). Future studies should investigate the role of user knowledge and experience on reactions to an anthropomorphised aid.

Lee (2007) described the importance of better understanding how human interaction with more autonomous technology could be affected by such technology that elicits affective or social responses. This study was an attempt to deliberately capitalise on social/affective responses in the design of an aid by a visual anthropomorphisation. Given prior research with anthropomorphic interfaces, such agents should engender a highly social response in the very social domain of decision-making. A highly social response means trusting the automation to make the right recommendation in decision tasks that are otherwise very difficult and error-prone. At a minimum, the current results suggest that anthropomorphically designed automation has effects on trust, dependence and performance. It is important to note that while our automated aid was designed to be deliberately anthropomorphic, all automation is in a sense anthropomorphic. Even aids that are ostensibly designed without anthropomorphism in mind must communicate to the user in a specific tone that can be polite or authoritative (e.g. Parasuraman and Miller 2004).

Acknowledgements

This research was supported by a Google Research Award to the first author. We would like to thank two anonymous reviewers for their helpful comments.

References

Bahner, J.E., Hüper, A., and Manzey, D., 2008. Misuse of automated decision aids: complacency, automation bias and the impact of training experience. *International Journal of Human–Computer Interaction*, 66, 688–699.

Bengtsson, B., Burgoon, J.K., Cederberg, C., Bonito, J., and Lundeberg, M., 1999. The impact of anthropomorphic interfaces on influence, understanding, and credibility. *In*: *Proceedings of the 32nd annual Hawaii international conference on systems sciences*, 5–8 January, Maui, HI. New York: IEEE. 1–15.

Berry, D., Butler, L., and deRosis, F., 2005. Evaluating a realistic agent in an advice-giving task. *International Journal of Human–Computer Studies*, 63 (3), 304–327.

Bickmore, T.W., Caruso, L., Clough-Gorr, K., and Heeren, T., 2005. 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers*, 17 (6), 711–735.

Cassell, J. and Bickmore, T., 2000. External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43 (12), 50–56.

Catrambone, R., Stasko, J., and Xiao, J., 2002. Anthropomorphic agents as a user interface paradigm: experimental findings and a framework for research. *In*: *Proceedings of the 24th annual conference of the cognitive science society*, 7–10 August, Fairfax, VA. Hillsdale, NJ: Lawrence Erlbaum, 166–171.

Drucquer, M.H. and McNally, P.G., 1998. *Diabetes management: step by step*. Osney Mead, Oxford: Blackwell Science Ltd.

Dzindolet, M., Pierce, L.G., Beck, H.P., and Dawe, L.A., 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44 (1), 79–94.

Ezer, N., Fisk, A.D., and Rogers, W.A., 2008. Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50 (6), 853–863.

Glasgow, R.E. and Strycker, L.A., 2000. Preventative care practices for diabetes in two primary care samples. *American Journal of Preventative Medicine*, 19 (1), 9–14.

Gong, L., 2008. How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, 24 (4), 1494–1509.

Hart, S. and Staveland, L., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *In*: P.A. Hancock and N. Meshkati, eds. *Human mental workload*. Amsterdam: North Holland Press, 239–250.

Ho, G., Wheatley, D., and Scialfa, C.T., 2005. Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17 (6), 690–710.

Lee, J. and Moray, N., 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human–Computer Studies*, 40 (1), 153–184.

Lee, J.D. and Moray, N., 1992. Trust, control strategies and allocation of function in human–machine systems. *Ergonomics*, 35 (10), 1243–1270.

Lee, J.D. and See, K.A., 2004. Trust in automation: designing for appropriate reliance. *Human Factors*, 46 (1), 50–80.

Lee, J.D., 2007. *Affect, attention, and automation. Attention: from theory to practice*, Vol. 4. USA: Oxford University Press, 73.

Madhavan, P., Wiegmann, D.A., and Lacson, F.C., 2006. Automation failures on tasks easily performed by operators undermines trust in automated aids. *Human Factors*, 48 (2), 241–256.

Merritt, S. and Ilgen, D.R., 2008. Not all trust is created equal: dispositional and history-based trust in human–automation interactions. *Human Factors*, 50, 194–210.

Meyer, J., 2001. Effects of warning validity and proximity on responses to warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43 (4), 563–572.

Moray, N. and Inagaki, T., 2000. Attention and complacency. *Theoretical Issues in Ergonomics Science*, 1 (4), 354–365.

Murano, P., 2003. Anthropomorphic vs non-anthropomorphic software interface feedback for online factual delivery. *In*: *Information visualization 2003 proceedings of the seventh international conference*, 16–18 July, London. London, UK: IEEE, 138–143.

Nass, C. and Lee, K.M., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7 (3), 171–181.

Nass, C., Fogg, B., and Moon, Y., 1996. Can computers be teammates? *International Journal of Human–Computer Studies*, 40, 543–559.

Nass, C., *et al*., 1993. INTERACT '93 and CHI '93 conference companion on human factors in computing systems – CHI '93. *In:* Computers as social actors. New York: ACM Press, 111–112.

Parasuraman, R. and Miller, C.A., 2004. Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47 (4), 51–55.

Parasuraman, R. and Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39 (2), 230–253.

Parasuraman, R., Sheridan, T.B., and Wickens, C.D., 2000. A model of types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30 (3), 286–297.

Qiu, L. and Benbasat, I., 2010. A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human–Computer Studies*, 68 (10), 669–688.

Rapoff, M.A., 2009. *Adherence to pediatric medical regimens*. New York: Springer Verlag.

Rosenbergkima, R., Baylor, A., Plant, E., and Doerr, C., 2008. Interface agents as social models for female students: the effects of agent visual presence and appearance on female students' attitudes and beliefs. *Computers in Human Behavior*, 24 (6), 2741–2756.

Salthouse, T.A., 1992. What do adult age differences in the Digit Symbol Substitution Test reflect? *Journal of Gerontology*, 47 (3), P121.

Schulman, D. and Bickmore, T., 2009. Persuading users through counseling dialogue with a conversational agent. *In*: *Proceedings of the 4th international conference on persuasive technology*, 26–29 April, Claremont, CA. New York, NY: ACM Press, 1–8.

Smith, A., 2011. *35% of American adults own a smartphone* [online]. Pew Internet & American Life Project. Available from: http://pewinternet.org/∼/media//Files/Reports/2011/PIP_Smartphones.pdf [Accessed 23 March 2011].

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., and Water, K., 1996. When the interface is a face. *Human–Computer Interaction*, 11 (2), 97–124.

Steele, C.M., 1997. A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist*, 52 (6), 613–629.

Wexelblat, A., 1998. *Don't make that face: a report on anthropomorphizing an interface*. From AAAI Technical Report SS-98-02. American Association for Artificial Intelligence.

Wiegmann, D.A., Rich, A., and Zhang, H., 2001. Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomic Science*, 2 (4), 352–367.

Yee, N., Bailenson, J.N., and Rickertsen, K., 2007. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. *In*: *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '07)*. New York: ACM Press, 1–10.