This article was downloaded by: [Clemson University]
On: 24 July 2014, At: 04:43
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Ergonomics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/terg20

A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults

Richard Pak[a], Anne Collins McLaughlin[b] & Brock Bass[a]
[a] Department of Psychology, Clemson University, Clemson, USA
[b] Department of Psychology, North Carolina State University, Raleigh, USA
Published online: 17 Jun 2014.

To cite this article: Richard Pak, Anne Collins McLaughlin & Brock Bass (2014): A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults, Ergonomics, DOI: 10.1080/00140139.2014.928750

To link to this article: http://dx.doi.org/10.1080/00140139.2014.928750

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Ergonomics, 2014
http://dx.doi.org/10.1080/00140139.2014.928750

Taylor & Francis
Taylor & Francis Group

A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults

Richard Pak[a]*, Anne Collins McLaughlin[b] and Brock Bass[a]

[a]*Department of Psychology, Clemson University, Clemson, USA;* [b]*Department of Psychology, North Carolina State University, Raleigh, USA*

(*Received 17 December 2013; accepted 19 May 2014*)

Previous research has shown that gender stereotypes, elicited by the appearance of the anthropomorphic technology, can alter perceptions of system reliability. The current study examined whether stereotypes about the perceived age and gender of anthropomorphic technology interacted with reliability to affect trust in such technology. Participants included a cross-section of younger and older adults. Through a factorial survey, participants responded to health-related vignettes containing anthropomorphic technology with a specific age, gender, and level of past reliability by rating their trust in the system. Trust in the technology was affected by the age and gender of the user as well as its appearance and reliability. Perceptions of anthropomorphic technology can be affected by pre-existing stereotypes about the capability of a specific age or gender.

Practitioner Summary: The perceived age and gender of automation can alter perceptions of the anthropomorphic technology such as trust. Thus, designers of automation should design anthropomorphic interfaces with an awareness that the perceived age and gender will interact with the user's age and gender.

Keywords: automation; trust; aging; stereotypes; mobile; health

Downloaded by [Clemson University] at 04:43 24 July 2014

1. Anthropomorphic technology can elicit stereotypes

Interactive computer systems that exhibit human-like, or anthropomorphic, traits can lead users to perceive and treat them differently than non-human-like systems (Nass, Steuer, and Tauber 1994). Thus, it is imperative to understand how users' perceptions of the system might be affected by their social reactions to anthropomorphic technology. One way in which a system may elicit social reactions is by eliciting stereotypes (Yee, Bailenson, and Rickerson 2007).

Stereotypes are preconceptions about the traits, behaviour, or abilities of a group and can set expectations of a stereotyped individual. Stereotypes can have both negative and positive connotations that may be inconsistent with real group attributes but provide adaptive value because they filter and organise incoming information, thereby easing processing and interpretation (Hilton and von Hippel 1996). Stereotypes can be activated and applied with or without conscious awareness (Banaji, Hardin, and Rothman 1993; Greenwald and Banaji 1995). Unfortunately, when the stereotype is highly simplified or inaccurate, it can lead to errors in perceptions and behaviour.

Nass, Steuer, and Tauber (1994) tested whether users would apply gender-related stereotypes when interacting with a computer that exhibited a gender. Their participants were first tutored by a computer on a specific topic. Tutored topics were either stereotypically female (love and relationships) or stereotypically male (computers and technology). They then moved to a non-gendered computer for testing and to a gendered computer for evaluation of their test responses. When gender of the tutor matched the stereotypic topic, participants rated it as a better teacher. This finding was echoed by Lee (2003) in a study where participants answered difficult trivia questions that were either stereotypically feminine or masculine. After answering the trivia question, participants viewed a female or male computerised agent that presented its own answer and then were allowed to change their answer. More participants changed their answers to agree with the agent when the gender of the agent matched the stereotypical topic.

Stereotype activation for computerised agents can also interact with individual differences, such as physical characteristics. Qiu and Benbasat (2009) found that an anthropomorphic decision aid significantly increased perceptions of social presence and led to increased trust of the agent. The strength of these effects was influenced by the degree to which the decision aid agent was similar to the user on a visible factor, such as ethnicity. The link between trust and apparent physical characteristics was explained via similarity-attraction theory that predicted that people would be more attracted to those similar to them (Byrne 1971). The user may have attributed their attraction to a similar ethnicity as trustworthiness of the agent.

*Corresponding author. Email: richpak@clemson.edu

© 2014 Taylor & Francis

In another example of the moderating role of individual differences in susceptibility to anthropomorphic effects, susceptibility to flattery (insincere praise) depended on the level of computer experience of the user (Johnson, Gardner, and Wiles 2004). Johnson, Gardner, and Wiles found that susceptibility to flattery from a computer depended on the user's experience level with computers – the judgments of highly experienced users were more affected by flattery than less experienced users. Furthermore, Lee (2010) found that people who exhibited less analytical and more intuitive cognitive style were more susceptible to flattery from a computer.

In summary, stereotypes can affect user perceptions of a computer or automated aid and can be moderated by individual differences. Some of the aids described in the previous studies were forms of automation that functioned in a decision-support capacity; thus, some automation bias may be based on stereotypes (Skitka, Mosier, and Burdick 1999). However, no research has explicitly examined how these factors might interact with machine-related factors of automation, such as reliability of the automation or how various activated stereotypes might interact (e.g. age and gender).

2. Age stereotypes in technology?

Age is one of the first and most salient attributes noticed of a person (Fiske 1998) which suggests it may also be true with anthropomorphic agents. Furthermore, stereotypes about age are stronger (Kite, Deaux, and Miele 1991) and more complex than gender stereotypes (Kite et al. 2005). In Kite, Deaux, and Miele's study assessing age and gender stereotypes using free response, participants viewed a younger (35-year-old) male or female and older (65-year-old) male or female and provided characteristics of the target person. Analysis showed that when negative stereotypes were generated, they were much more likely to be due to the age of the target than the gender. Finally, according to the similarity-attraction hypothesis (Qui and Benbasat 2009), older and younger adults should exhibit positive anthropomorphic effects with automation that matches their age group. However, it may also be that an older-looking automated agent may prime negative stereotypes about age, particularly when the reliability of the automation is perceived to be low. This may explain why a previous study found that a young female agent enhanced younger adults' trust in automation but not older adults' when participants interacted with a health decision aid (Pak et al. 2012). The authors hypothesised that the dissimilarity between a younger female decision agent and an older participant may have muted any potential anthropomorphic effect on trust due to violation of the similarity-attraction. An alternative explanation is that older adults hold negative stereotypes of the capabilities of younger, female doctors but younger adults do not.

3. Age and gender stereotypes of physicians

People hold stereotypes that older workers have lower ability, are less motivated, and are less productive than younger workers (Posthuman and Campion 2009). Older workers are also seen as less adaptable to changing work situations and uncertainty than younger workers (DeArmond et al. 2006). Although aging studies show that these views may be exaggerated (e.g. see Czaja and Sharit 1998), they are widely held by people of all ages and affect workplace hiring decisions and evaluations (DeArmond et al. 2006; Posthuma and Campion 2009). Negative age stereotypes about older workers are even held by older adults themselves (Rosen and Jerdee 1976; Finkelstein and Burke 1998; Wrenn and Maurer 2004). Finally, these stereotypes may be activated without awareness (Devine 1989; Perdue and Gurtman 1990; Banaji and Hardin 1996).

Activation of age stereotypes may be moderated by individuating past behaviour or context (Kunda and Sherman-Williams 1993). Individuating information such as context (e.g. interacting with a doctor) may determine which aspect of a stereotype gets activated (Casper, Rothermund, and Wentura 2011). Knowing the occupation of an individual is a type of individuating information that seems to alter some negative age stereotypes. For example, although some occupations seem more negatively age stereotyped (e.g. Cleveland and Hollman 1990), the occupation of physician is moderately seen as a stereotypically older male occupation (Singer 1986) even though it is an occupation that may require adaptability and is faced with uncertainty. In contrast, when stereotypes of doctors were more recently assessed (Shah and Ogden, 2006), younger female doctors were perceived as having better personal manner and technical skill than older doctors of either gender. The scant literature on physician age stereotypes seems to suggest that the stereotype of older doctors is less negative than the stereotype for older adults in general, but still present (McKinstry and Yang 1994), demonstrating the power of individuating information on the otherwise powerful age stereotype.

In summary, person-judgment based on stereotypes can depend on individuating information, including profession, past performance (i.e. reliability), gender, and age. Similarly, assessment of computer-based automation with human-like characteristics may also be subject to pre-existing stereotypes consistent with the human-like qualities (e.g. age, gender). Anthropomorphic automation with ambiguous reliability may be more likely to activate pre-existing stereotypes. That is, when automation is unambiguously reliable or unreliable, stereotypes should not affect perceptions. But when automation is ambiguous, stereotypes will affect perceptions of the automation such as trust. The idea that imperfect automation may

Downloaded by [Clemson University] at 04:43 24 July 2014

engender the expression of implicit attitudes has been suggested by other automation researchers (Lee and See 2004; Merritt, Heimbaugh, and LaChapell 2012).

4. Anthropomorphism and automation characteristics

Studies of human–automation interaction have demonstrated that many factors related to the person, automated system, and task interact to determine trust in and performance with automation. For example, individual differences in attitudes towards automation (e.g. Mosier et al. 1998; Dzindolet et al. 2003; Merritt and Ilgen 2008) interacted with machine characteristics such as reliability and error types (e.g. Madhavan, Wiegmann, and Lacson 2006; Rovira, McGarry, and Parasuraman 2007) and task or situational factors such as workload (e.g. Röttger, Bali, and Manzey 2009) to affect behaviour with and perceptions of automation.

Research investigating the influence of anthropomorphic aspects specifically on human–automation interaction (Parasuraman and Miller 2004, Pak et al. 2012) found that various implementations of anthropomorphism such as etiquette (Bickmore 2011; Zhang, Zhu, and Kaber 2011) affected perceptions of trust and automation behaviour. For example, in aircraft engine diagnosis, the automation either presented advice in a rude or polite manner (Parasuraman and Miller 2004). As expected, perceived trust and performance in the diagnosis task was better when the automation was 80% reliable compared to 60% reliable. However, engine diagnosis performance and trust with polite but less reliable automation was the same as rude but highly reliable automation. It was not speculated why etiquette would interact with reliability but it may be that politeness affected an internal belief that artificially adjusted expectations of the automation that influenced attributions of responsibility (e.g. Marakas, Johnson, and Palmer 2000).

Thus, behaviour with anthropomorphic automation is affected by how it is perceived in addition to its reliability. The literature in computer-mediated communication has demonstrated the computers as social actors effect (e.g. stereotype elicitation, susceptibility to flattery) as well as the moderating influence of individual differences (e.g. cognitive style, ethnicity). Complementing these findings, the automation literature has shown that overt anthropomorphic elements (etiquette, human-like appearance) in automation can interact with machine-related factors such as automation reliability to influence trust and performance. The conceptual link between these two literatures is the finding that implicit attitudes about automation itself, or beliefs about the capabilities of automation held without conscious awareness, significantly affect trust in automation but only when reliability of the automation was uncertain (Lee and See 2004; Merritt, Heimbaugh, and LaChapell 2012).

Merritt, Heimbaugh, and LaChapell (2012) theorised that implicit general attitudes about automation affected the propensity to trust machines and an individual's trust in a specific automated system. Perceptions of the behaviour of any automation will be filtered through these explicit and implicit pre-existing beliefs about automation (Dzindolet et al. 2002). Merritt et al. found that when automation reliability was ambiguous, implicit beliefs about automation and stereotypes were more influential in determining trust than explicit beliefs. Presumably, in the face of ambiguity, individuals made attributions that were consistent with their implicit, schematic pre-existing beliefs about automation. This paralleled findings from the social cognition literature that stereotypic reasoning was common when an individual was faced with conflicting or ambiguous information (Kunda and Thaggard 1996).

Reframing the results of Parasuraman and Miller (2004) in light of the findings of Merritt, Heimbaugh, and LaChapell (2012), it may be that when automation performance was ambiguous/of low reliability participants fell back to their *newly* formed positive *implicit* beliefs about the automation (that the automation was polite), and the participants made more situational rather than dispositional attributions (i.e. attributed fault to the situation, not the automation). For the present study, Merritt et al's and Parasuraman and Miller's studies are crucial for several reasons. First, they showed that implicitly held beliefs influence explicit perceptions of trust in automation. Second, the implicit attitudes interacted with automation reliability to determine trust and behaviour. Factors at the person-level (stereotypes) and task-level (automation reliability) interacted to affect judgments and perceptions of technology. There is a wealth of research examining the role of etiquette on automation perceptions (Hayes and Miller 2011) but the current work extends the concept that another type of implicitly held perception (stereotypes) may affect how users perceive automation. The present study extended previous work on gender stereotypes on automation behaviour by examining another potential stereotype: age.

5. Overview of the study

Using participants in younger and older adult age groups, we collected judgments of trust of a simulated agent embedded within a decision aid that varied in gender, age, and reliability using a factorial survey with concrete health-related vignettes. Following the social cognition literature, we expected that age and gender stereotypes would most affect trust in the decision aid when system performance was ambiguous, but that there would be different effects for different age groups and genders of users. Specific research aims were as follows: (1) Determine the amount of variance in trust due to within-person variation

Downloaded by [Clemson University] at 04:43 24 July 2014

compared to between-person variation, (2) Determine how age of the agent, gender of the agent, and reliability of the decision aid agent affected judgments of trust in the aid, and (3) Determine how individual differences such as age and gender of the participant affected trust ratings of various decision aids. The results informed basic knowledge of how differing age and gender groups responded to stereotypes as well as informing the design of decision aids targeting particular groups of users.

We presented scenarios involving a decision aid (a smartphone 'app') for diabetes management via a factorial survey. The decision aid contained a simulated anthropomorphised agent. Factorial surveys have been widely used to examine how beliefs, judgments, and decision-making are influenced by situational factors (Rossi and Anderson 1982). Specific factors of the scenario were manipulated (in a factorial manner) and the participant rated all combinations of factors. The agent was a health-care provider offering advice on a specific diabetes-related dilemma. Because our dependent variable (trust) was a social judgment about a situation, a factorial survey was an ideal way to measure the influence of manipulated variables (age, gender, reliability of automation) as well as individual differences of the participants (Rossi and Anderson 1982; Hox, Kreft, and Hermkens 1991).

5.1 Method

5.1.1 Participants

Sixty younger adults and 47 older adults completed the study. The mean age of the younger group was 18.6 (SD = 0.9) while the older group was 72.7 (SD = 5.3). Younger adults were undergraduate college students whereas older participants were independently living, community-dwelling older adults. The younger participants chose to receive either course credit or $7 per hour and the older participants received $7 per hour. Descriptive statistics of participant characteristics are shown in Table 1.

5.1.2 Materials

Equipment. PC-compatible (Windows 7) computers running at 3.2 GHz with 4 GB of RAM were used with a 19-inch (48.3-cm) LCD monitor set at a resolution of 1024 × 1280 pixels. Participants were seated approximately 18 inches from the monitor and interacted primarily with a mouse (on the preferred side) and a keyboard.

Individual difference measures. In addition to participant age group and gender, we were interested in two individual difference measures: automation complacency and prior diabetes knowledge. The Complacency Potential Rating Scale (CPRS; Singh, Molloy, and Parasuraman 1993) is a 16-item scale designed to measure complacency towards common types of automation (e.g. automated teller machines). Participants responded to the extent they agreed with statements about automation on a scale of 1–5. The CPRS score was a sum of these responses and ranged from 16 (low complacency potential) to 80 (high complacency potential). We were primarily interested in CPRS to compare our sample to other studies that show higher complacency potential in older adults (e.g. Ho, Wheatley, and Scialfa 2005). Diabetes knowledge was assessed with the Diabetes Knowledge Test (DKT; Fitzgerald et al. 1998). The 23 questions of the DKT assessed basic knowledge about diabetes and diabetes management. Computerised versions of both the CPRS and DKT were used in this study.

Task. In a factorial survey, independent variables are called dimensions. The dimensions are orthogonal and can have multiple levels. Orthogonal dimensions allowed us to disentangle the unique effects of each dimension on judgments of trust. Our dimensions were agent gender (male, female), agent age (younger, older), and aid reliability (low, medium, high).

Table 1. Participant characteristics by age group and gender.

	Younger adults ($n = 60$)				Older adults ($n = 47$)			
	Female ($n = 37$)		Male ($n = 23$)		Female ($n = 25$)		Male ($n = 22$)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	18.49	0.72	18.74	1.15	72.00	5.29	73.45	5.27
CPRS[a]*	43.73	3.83	43.00	5.38	48.52	5.31	46.09	4.04
Diabetes knowledge[b]*	11.68	2.02	11.48	2.52	14.24	2.81	13.41	2.84

*Significant age group difference, $p < 0.05$ (no significant gender differences).
[a] Scores could range from 16 indicating low complacency potential to 80 indicating high complacency potential (Singh, Molloy, and Parasuraman 1993).
[b] The DKT scores could range from 0 indicating no knowledge to 23 indicating high knowledge (Fitzgerald et al. 1998).

Downloaded by [Clemson University] at 04:43 24 July 2014

Table 2. Dimensions (independent variables) of interest and resulting scenarios.

Scenario	Agent age (2)	Agent gender (2)	Stated reliability (3)
1	Young	Female	45%
2	Young	Female	70%
3	Young	Female	95%
4	Young	Male	45%
5	Young	Male	70%
6	Young	Male	95%
7	Older	Female	45%
8	Older	Female	70%
9	Older	Female	95%
10	Older	Male	45%
11	Older	Male	70%
12	Older	Male	95%

Note: Each scenario was presented twice resulting in 24 unique vignettes.

The dimensions of interest, their levels, and the factorial combinations resulting in 12 possible scenarios are shown in Table 2. Each scenario was replicated twice to create 24 unique vignettes. This resulted in 12 measurements of each dimension per participant. In their review of the literature, Wickens and Dixon (2007) proposed that an automation reliability of about 70% represented a critical inflection point; less than about 70% reliable was not relied upon while reliabilities higher than 70% led to complacency. For this reason, we chose high and low values that were well above and below 70% (45%, 70%, and 95%) to represent low, medium, and high reliabilities, respectively. Participants never actually experienced the levels of automation reliability; they were only told the past reliability of the particular app that was shown. No matter the stated past reliability of an app, the advice given by the app in every scenario was correct.

The possible combinations of agent age and gender are shown in Figure 1. An example vignette (containing older female, high reliability) is illustrated in Figure 2. The diabetes dilemma was presented in the upper left of screen. On the right, a diagnostic smartphone app gave a possible solution via an agent. The size of the smartphone was larger than actual size (approximately 30% larger) to be easily viewable from seated distance. Also, on the screen was a statement about the past reliability of the particular app (low, medium, or high). On the lower third, participants rated on a Likert scale their perception of trust and likelihood of following the advice of the aid.

The diabetes scenarios were used in a prior study (Pak et al. 2012) and were developed by adapting questions from a diabetes education workbook (Drucquer and McNally 1998), and reading diabetes support forums. They were designed to represent realistic scenarios that someone with Type II diabetes might experience. The presentation of the factorial survey was programmed in the Real Studio environment (Real Software 2013).

5.1.3. *Design and procedure*

The study was a 2 (age group of respondent: younger, older) × 2 (gender of respondent: male, female) × 2 (agent age: young, old) × 2 (agent gender: male, female) × 3 (aid reliability: low, medium, high) mixed-model design, with within-

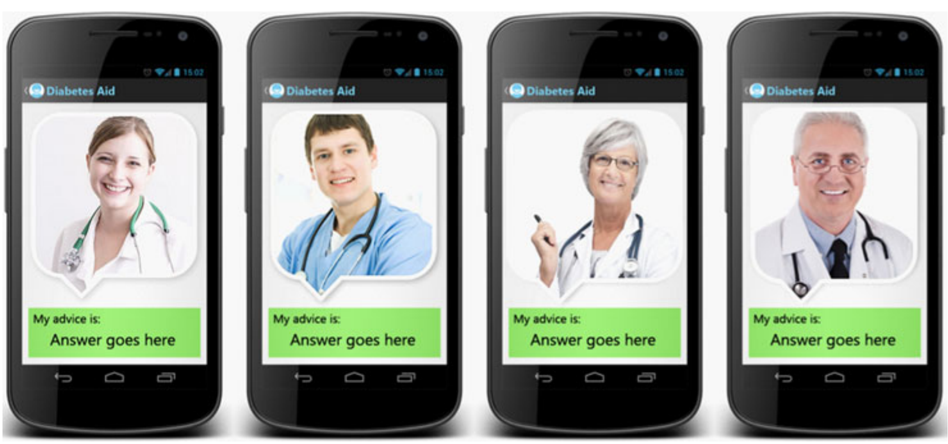

Figure 1. Illustration of the four possible smartphone agent conditions (young female, young male, older female, older male).

Downloaded by [Clemson University] at 04:43 24 July 2014

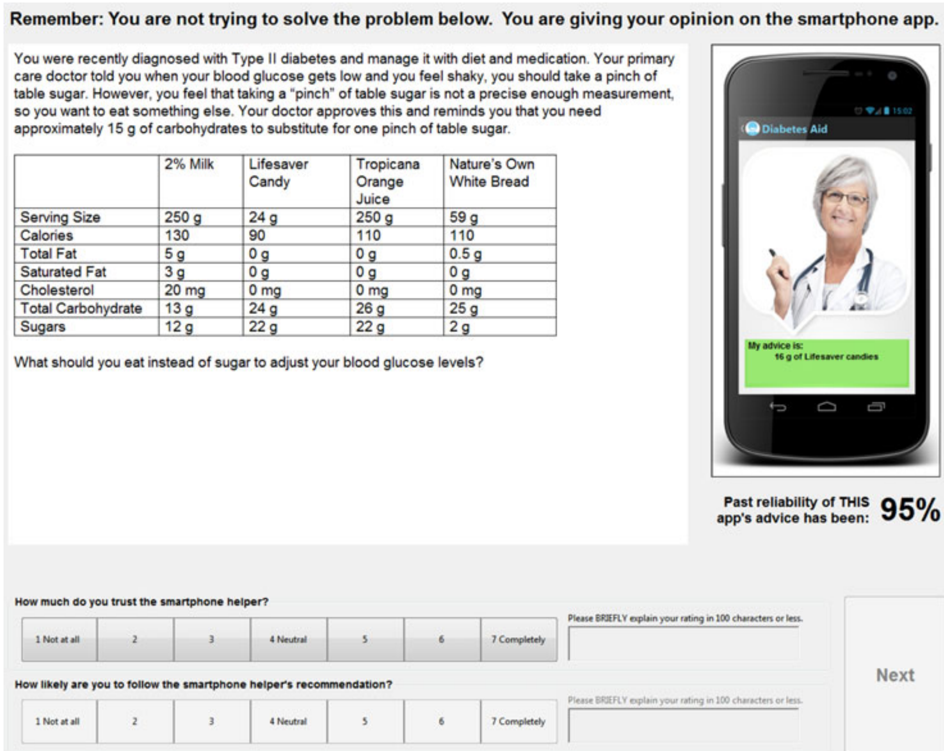

Figure 2. Image of the factorial survey response screen.

participant factors manipulated in the factorial survey. The first two variables (age group and gender of respondent) were quasi-independent grouping variables while the last three were within-groups manipulations of the decision aid and agent. The dependent variables were trust, likelihood of following advice, and diabetes knowledge.

Participants first completed a diabetes knowledge questionnaire administered on a computer. Next, participants started the factorial survey and were told:

> You are playing the part of a newly diagnosed diabetic. Your doctor has given you a variety of different smartphone apps that may help you with your diabetes care. Your task involves giving us your opinion of the different smartphone apps. Just like many technological aids, the different apps will only sometimes seem reliable. Your performance is not being tested so you do not have to try to solve every problem. Instead, you are making judgments of the smartphone apps as quickly as possible.

After acknowledging the instructions and answering any remaining questions they began the survey.

In the survey, participants viewed a randomly presented vignette and were asked the following questions: (1) how much they trusted the smartphone app on a scale from 1 (not at all) to 7 (very much), and (2) whether they would follow or actually use the advice of the app (1–7). After the trust and decision aid usage questions, participants were also asked to briefly explain their ratings. To reinforce the notion that the smartphone app was a real decision aid and not just a pre-computed image, the smartphone app did not reveal its answer for 1.5 seconds (in the interim the message, 'Analysing the scenario. Just a moment …' appeared on the smartphone screen). After responding to 24 vignettes, participants completed the CPRS. Finally, participants answered the question, 'What do you think the study was about?' to assess whether they were aware of the purpose of the study. None of our participants were able to accurately state the purpose of the study other than what was told to them in the instructions (evaluating different apps). Because the trust and likelihood to follow ratings were highly correlated ($r = 0.83$, $p < 0.05$) only trust ratings were analysed.

5.2 Results

5.2.1 Hypothesised model

To answer our original research questions a two-level hierarchical model assessed the effects of agent gender and age, decision aid reliability, and diabetes knowledge on perceptions of trust in the decision aid. To review, our questions were (1) How is trust in an anthropomorphic decision aid affected by a user's age and gender?, (2) How is trust in the smartphone app affected by its appearance and reliability?, and (3) How is trust affected by domain knowledge?

Downloaded by [Clemson University] at 04:43 24 July 2014

Multiple responses were nested within the 107 participants: Each participant judged 24 vignettes resulting in a total of 2568 judgments for analysis. These judgments were nested within the manipulations performed on the survey (agent age, agent gender, reliability), which were in turn nested within the attributes of the participant (participant age, participant gender, diabetes knowledge score, CPRS score). Multi-level modelling was implemented through SAS, version 9.2.

Multi-level models are appropriate for data that exhibit hierarchical structure as they account for variability between and within participants and allow for examination of cross-level interactions (Raudenbush and Bryk 2002). Because respondents repeatedly made judgments on varying vignettes, those judgments of trust were not independent of each other; in fact, they were highly likely to be correlated which violates the independence of error variances assumption of analysis of variance (ANOVA) and regression (Hox and Bechger 1998; Tabachnick and Fidell 2007). There were also likely to be correlations between different levels (response level, group level). For example, trust responses on a vignette would likely be correlated to the responders group (gender, age group). That is, males may have a different stereotype than females (or older respondents versus younger ones) that they applied to the situation. Ignoring this hierarchical structure, or nesting, (i.e. by using ordinary least squares regression) can lead to an inflated Type I error rate, or detecting effects when there are none (Tabachnick and Fidell, 2007). Multi-level modelling solves this problem by allowing intercepts and slopes between levels to vary. Variability at one level is treated as a dependent variable at the next level. Hoffman and Rovine (2007) provided an accessible description of the usefulness of multi-level linear models in experimental psychology and human factors and Hox, Kreft, and Hermkens (1991) detailed why multi-level modelling is preferred for the analysis of factorial surveys.

A fully unconditional (non-multivariate) model (Model 1) was used to discover the amount of variance in trust found within participants at the survey level (Level 1; variance due to app appearance) and the amount of variance at the person level (Level 2; variance due to individual differences). This model represented a baseline to assess the fit of subsequent multivariate models (Models 2 and 3; equations in Appendix). Results (Table 3) revealed significant variance at both levels, with 94% of the variance at the survey level ($\sigma^2 = 3.04$, $z = 35.08$, $p < 0.0001$) and 6% of the variance at the person level ($\tau_{00} = 0.19$, $z = 4.39$, $p < 0.0001$).

Model 2 examined the effects of the survey manipulations on judgments of trust: agent gender, agent age, reliability, and all Level 1 interactions. Results revealed significant effects for all survey manipulations. Participants trusted male agents

Table 3. Unstandardised coefficients of multi-level models of the within and between-person effects of predictors on trust.

	Model 1		Model 2		Model 3	
	Unconditional Model		Random Coefficients Regression		Slopes and Intercepts	
	Estimate	SE	Estimate	SE	Estimate	SE
Fixed effects						
Intercept	4.75***	0.05	3.52***	0.11	3.43***	0.13
Between-person						
Age Group					0.35*	0.14
Gender					−0.13	0.12
Diabetes knowledge score					−0.06**	0.02
CPRS					−0.01	0.01
Within-person						
Agent gender			0.67***	0.14	0.67***	0.14
Agent age			0.38**	0.14	0.38**	0.14
Reliability of agent			1.09***	0.08	1.09***	0.08
Agent gender × agent age			−0.42*	0.20	−0.42*	0.20
Agent gender × reliability			−0.43***	0.11	−0.44***	0.12
Agent age × reliability			−0.36***	0.11	−0.19	0.12
Agent gender × agent age × reliability			0.47**	0.15	0.47**	0.15
Cross-level						
Age group × agent age group × reliability					−0.35***	0.09
Gender × agent gender × reliability					0.09	0.09
Age group × agent gender × reliability					−0.06	0.09
Gender × age group × reliability					−0.04	0.09
R^2 within-person			16.02		16.55	
R^2 between-person			<0.01		<0.01	
Random effects						
σ^2	3.04***	0.09	2.56***	0.07	2.54***	0.07
τ_{oo}	0.19***	0.04	0.21***	0.04	0.20***	0.04

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. All between-person predictors were grand-mean centred.

Downloaded by [Clemson University] at 04:43 24 July 2014

more than female ones, older agents more than younger ones, and more reliable apps than less reliable ones. However, multiple significant interactions further refined this story. The three-way interaction of agent gender, agent age, and app reliability was significant – illustrated in Figure 3 – such that when the app was of low reliability, the younger female agent was trusted significantly less than the younger male aid, $F(1{,}1272) = 24.64$, $p < 0.05$, $\eta_p^2 = 0.2$, although there were no significant differences of agent gender for the younger agent at other reliability levels. For the older aid, the female agent was rated as less trusted, but this difference occurred only at the medium reliability level, $F(1{,}1272) = 13.91$, $p < 0.05$, $\eta_p^2 = 0.01$. These findings are consistent with our hypothesis that stereotypes would affect trust judgments when the reliability of a system was ambiguous (i.e. low or medium reliability).

A third model was conducted to include the individual difference predictors of participant age group, participant gender, CPRS, and diabetes knowledge and to examine hypothesised cross-level interactions. Our hypothesis was that participant

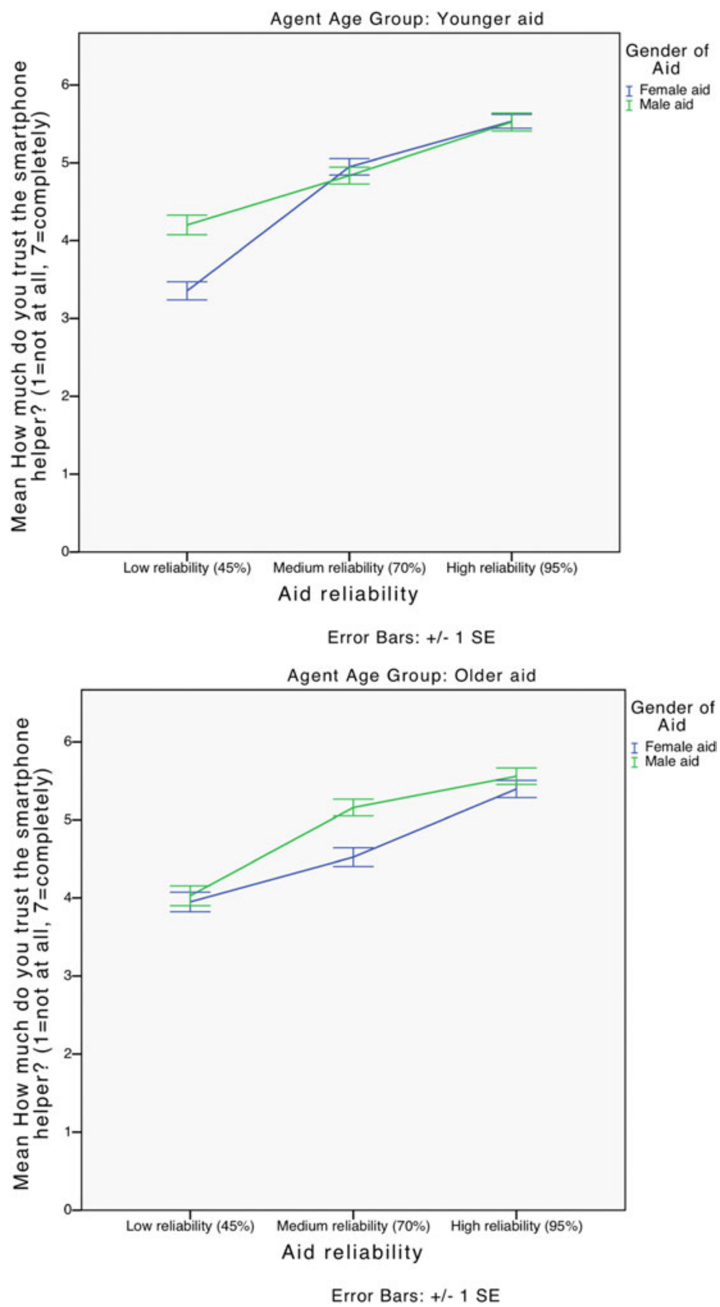

Figure 3. Three-way interaction of agent age group, agent gender, and reliability (from Model 2).

Downloaded by [Clemson University] at 04:43 24 July 2014

age group would interact with the age of the agent to differentially affect trust. The similarity-attraction hypothesis (Byrne 1971) would predict that the user's trust would be highest with agents that appear similar to them, particularly in age-appearance. We examined all cross-level interactions in Model 3.

In Model 3, those with higher diabetes knowledge rated the agents as less trusted overall. Older participants generally rated the agents as more trusted than did younger participants. This may be a manifestation of the generally higher complacency that older adults have with automation than younger adults (Ho, Wheatley, and Scialfa 2005). Gender of the participant and CPRS score had no effect on trust ratings. By entering these variables in the model they were controlled for when examining the cross-level interactions. Using the Akaike's information criterion, Model 3 was determined to better fit the data than Model 2 (it accounted for variance beyond Model 2). The three-way interaction among participant age group, agent age, and app reliability was significant (Figure 4). The source of the interaction was that younger adults in the low reliability condition tended to trust older agents significantly more than younger agents, $F(1,1434) = 16.88$, $p < 0.05$,

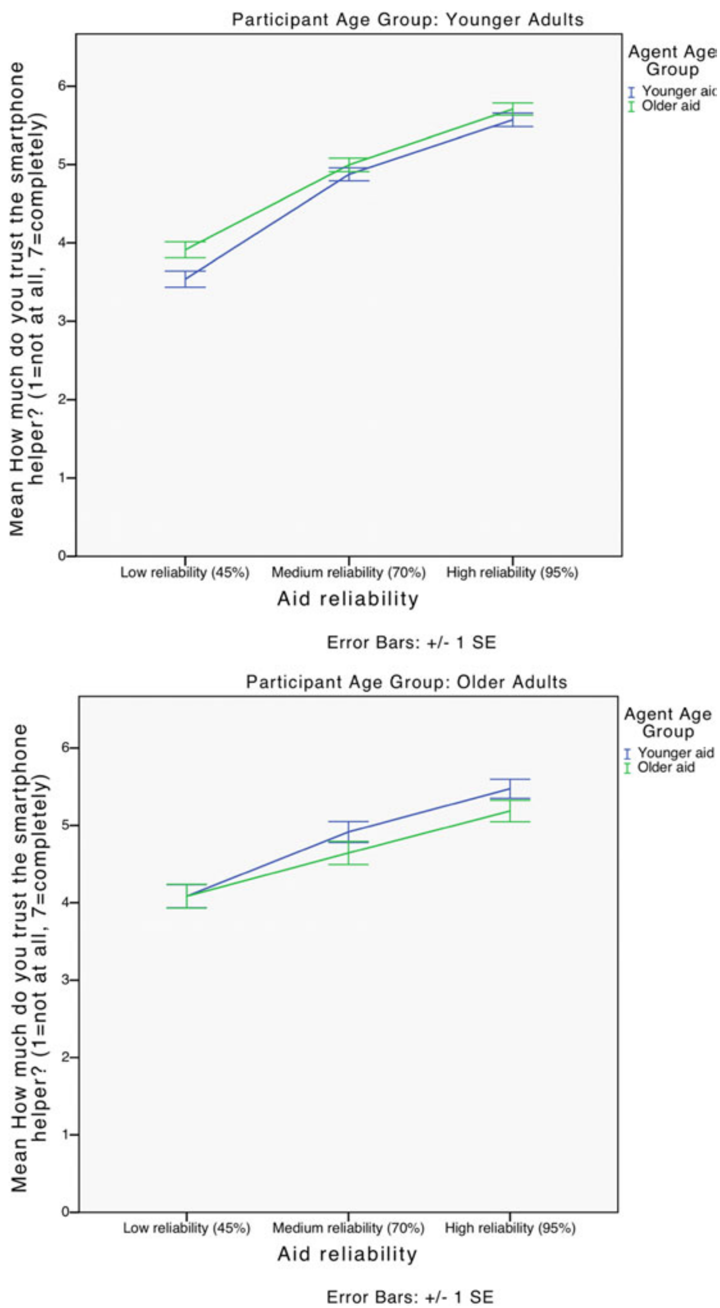

Figure 4. Cross-level interaction of participant age group, agent age group, and reliability (from Model 3).

Downloaded by [Clemson University] at 04:43 24 July 2014

$\eta_p^2 = 0.006$. There was no significant difference in trust by younger adults as a function of agent age in the medium or high reliability conditions. For older adults, there was no significant difference in trust as a function of agent age in any of the reliability conditions. Finally, to more directly test the possibility, presented in the introduction, that older adults may specifically hold negative stereotypes of young female agents, we examined the four-way interaction of agent age, agent gender, age group, and gender and found it to be not significant.

6. General discussion

As automation in consumer products and systems embodies human-like traits (e.g. anthropomorphic agents), stereotypes that users hold of age and gender may play an important role in trust and use of that automation. Prior research established that people apply gender stereotypes to computers but the purpose of this study was to examine if powerful and pervasive age stereotypes, as well as gender stereotypes, would be applied to anthropomorphic agents.

The finding that trust varies with reliability is not surprising; with higher levels of perceived reliability, users, particularly older adults, may become complacent (Mouloua et al. 2002; Ho, Wheatley, and Scialfa 2005). What is surprising is that this relationship between trust and complacency interacts with attributes of technology and individual differences in a way that is roughly consistent with the stereotype literature, specifically, age and gender stereotypes of doctors. However, perceived age group and gender of the agent and its reliability moderated the application of stereotypes (Model 2). When the agent appeared young, male agents were more trusted than female agents only when reliability was low. This gender difference disappeared at other levels of reliability. This pattern might suggest that unless the reliability of the system is catastrophically low (45%), most participants do not exhibit gender stereotypic thinking; perceptions of trust are primarily driven by reliability. However, when the reliability is very low, participants clearly shift to more stereotypic thinking and seem to attribute low performance to gender.

When the agent appeared older, male agents were more trusted than female agents only at medium levels of reliability. That is, stereotypic judgments appear at more moderate levels of reliability (70% versus 45%) if the aid is older rather than younger. The finding of gender stereotypic effects at 45% reliability when the agent is young, but at 70% when the agent is old seems to suggest that older female agents are judged more harshly than younger female agents. Given this finding one design recommendation is that when it is crucial for users to maintain high levels of trust in imperfect automation, a younger male agent is optimal because it seems less susceptible to large fluctuations in perceptions of trust as a function of gender (i.e. gender stereotypic thinking). More specifically, if it is undesirable to have users exhibit gender differences (or bias) in trust then using younger agents was preferable to older agents. A male agent was recommended over female because trust in female agents appeared more erratic as a function of reliability compared to male agents (e.g. the steep plunge in trust at 45% reliability for young females). However, this design recommendation does not take into account the gender or age group of the user. As the significant cross-level interaction of Model 3 shows, individual differences also seem to interact with the agent characteristics.

Model 3 showed that some anthropomorphic aspects of the aid did interact with participant individual differences to affect trust. Younger adults in low reliability conditions tended to trust older agents over younger agents while older adults did not show any significant differences in trust as a function of agent age. Based on Model 3, if the goal is to maintain high levels of trust in imperfect automation in young adult users, older agents (regardless of agent gender) are preferred. For older adult users, there was no significant difference in trust as a function of agent age group. However, there did appear to be a trend towards higher trust of younger agents with increasing reliability so for older users, a young agent may be optimal.

One caveat is that we did not assess *a priori* the pre-existing stereotypes held by our participants (as such an assessment might have influenced their behaviour in the experiment.) However, the stereotype literature is replete with research that shows the pervasiveness of the 'warm but not competent' stereotype of older adults not only in the USA but worldwide (Cuddy, Norton, and Fiske 2005). Another limitation is the use of a diabetes scenario. Although none of the participants in our study reported having diabetes, older adults may be more aware of diabetes simply because it is more common in their cohort than among younger adults (26.9% versus 11.3%, respectively; American Diabetes Association, 2011). Thus, simply being in a cohort that is more affected by diabetes may influence how one perceives diabetes advice. Another limitation was that because we assessed subjective perceptions of the automation (trust) because it is uncertain if trust translates to behaviour. However, past research has shown that perceptions of trust in automation are strongly correlated with behaviour (e.g. Lee and Moray 1994).

Acknowledgements

We thank Meghan Goodwin and Kayla Brennan for their help in collecting data and Peg Tyler for her help in manuscript preparation.

Downloaded by [Clemson University] at 04:43 24 July 2014

Funding

This research was supported by a grant from the Air Force Office of Scientific Research [award number FA9550-12-1-0385].

Notes on contributors

Richard Pak is currently an Associate Professor in the Department of Psychology at Clemson University. He received his PhD in psychology in 2005 from the Georgia Institute of Technology.

Anne McLaughlin is an Associate Professor in the Department of Psychology at North Carolina State University. She received her PhD in psychology in 2007 from the Georgia Institute of Technology.

Brock Bass received his MS in Human Factors from Clemson University.

References

American Diabetes Association. 2011. "Statistics About Diabetes." Retrieved from http://www.diabetes.org/diabetes-basics/statistics/?loc=db-slabnav

Banaji, M., and C. D. Hardin. 1996. "Automatic Stereotyping." *Psychological Science* 7 (3): 136–141.

Banaji, M. R., C. Hardin, and A. J. Rothman. 1993. "Implicit Stereotyping in Person Judgment." *Journal of Personality and Social Psychology* 65 (2): 272–281.

Bickmore, T. 2011. "Etiquette in Motivational Agents." In *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*, edited by C. C. Hayes and C. A. Miller, 206–226. Boca Raton: Auerbach.

Byrne, D. E. 1971. *The Attraction Paradigm (Vol. 11)*. New York: Academic Press.

Casper, C., K. Rothermund, and D. Wentura. 2011. "The Activation of Specific Facets of Age Stereotypes Depends on Individuating Information." *Social Cognition* 29 (4): 393–414.

Cleveland, J. N., and G. Hollmann. 1990. "The Effects of the Age-Type of Tasks and Incumbent Age Compositions on Job Perceptions." *Journal of Vocational Behaviour* 36 (2): 181–194.

Cuddy, A. J., M. I. Norton, and S. T. Fiske. 2005. "This Old Stereotype: The Pervasiveness and Persistence of the Elderly Stereotype." *Journal of Social Issues* 61 (2): 267–285.

Czaja, S. J., and J. Sharit. 1998. "Age Differences in Attitudes toward Computers." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 53 (5): 329–340.

DeArmond, S., M. Tye, P. Y. Chen, A. Krauss, D. Apryl Rogers, and E. Sintek. 2006. "Age and Gender Stereotypes: New Challenges in a Changing Workplace and Workforce." *Journal of Applied Social Psychology* 36 (9): 2184–2214.

Devine, P. G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56 (1): 5–18.

Drucquer, M. H., and P. G. McNally. 1998. *Diabetes Management: Step by Step*. Osney Mead, Oxford: Blackwell Science.

Dzindolet, M. T., S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. 2003. "The Role of Trust in Automation Reliance." *International Journal of Human Computer Studies* 58 (6): 697–718.

Dzindolet, M. T., L. G. Pierce, H. P. Beck, and L. A. Dawe. 2002. "The Perceived Utility of Human and Automated Aids in a Visual Detection Task." *Human Factors* 44 (1): 79–94.

Finkelstein, L. M., and M. J. Burke. 1998. "Age Stereotyping at Work: The Role of Rater and Contextual Factors on Evaluations of Job Applicants." *The Journal of General Psychology* 125 (4): 317–345.

Fiske, S. T. 1998. "Stereotyping, Prejudice, and Discrimination." In *The Handbook of Social Psychology*, edited by D. T. Gilbert, S. T. Fiske, and G. Linclzey. 4th ed. New York: McGraw-Hill.

Fitzgerald, J. T., R. M. Anderson, M. M. Funnell, R. G. Hiss, G. E. Hess, W. K. Davis, and P. A. Barr. 1998. "The Reliability and Validity of a Brief Diabetes Knowledge Test." *Diabetes Care* 21 (5): 706–710.

Greenwald, A. G., and M. R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-esteem, and Stereotypes." *Psychological Review* 102 (1): 4–27.

Hayes, C. C., and C. A. Miller. 2011. *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*. Boca Raton: Auerbach.

Hilton, J. L., and W. von Hippel. 1996. "Stereotypes." *Annual Review of Psychology* 47 (1): 237–271.

Ho, G., D. Wheatley, and C. T. Scialfa. 2005. "Age Differences in Trust and Reliance of a Medication Management System." *Interacting with Computers* 17 (6): 690–710.

Hoffman, L., and M. J. Rovine. 2007. "Multilevel Models for the Experimental Psychologist: Foundations and Illustrative Examples." *Behaviour Research Methods* 39 (1): 101–117.

Hox, J. J., and T. M. Bechger. 1998. "An Introduction to Structural Equation Modeling." *Family Science Review* 11: 354–373.

Hox, J. J., I. G. G. Kreft, and P. L. J. Hermkens. 1991. "The Analysis of Factorial Surveys." *Sociological Methods & Research* 19: 493–510.

Johnson, D., J. Gardner, and J. Wiles. 2004. "Experience as a Moderator of the Media Equation: The Impact of Flattery and Praise." *International Journal of Human-Computer Studies* 61 (3): 237–258.

Kite, M. E., K. Deaux, and M. Miele. 1991. "Stereotypes of Young and Old: Does Age Outweigh Gender?" *Psychology and Aging* 6 (1): 19–27.

Kite, M. E., G. D. Stockdale, B. E. Whitley, and B. T. Johnson. 2005. "Attitudes toward younger and Older Adults: An Updated Meta-analytic Review." *Journal of Social Issues* 61 (2): 241–266.

Downloaded by [Clemson University] at 04:43 24 July 2014

Kunda, Z., and B. Sherman-Williams. 1993. "Stereotypes and the Construal of Individuating Information." *Personality and Social Psychology Bulletin* 19 (1): 90–99.
Kunda, Z., and P. Thagard. 1996. "Forming Impressions from Stereotypes, Traits, and Behaviours: A Parallel Constraint Satisfaction Theory." *Psychological Review* 103 (2): 284–308.
Lee, E. J. 2003. "Effects of 'Gender' of the Computer on Informational Social Influence: The Moderating Role of Task Type." *International Journal of Human-Computer Studies* 58 (4): 347–362.
Lee, E.-J. 2010. "The More Humanlike, the Better? How Speech Type and Users' Cognitive Style Affect Social Responses to Computers." *Computers in Human Behaviour* 26 (4): 665–672.
Lee, J. D., and N. Moray. 1994. "Trust, Self-confidence, and Operators' Adaptation to Automation." *International Journal of Human Computer Studies* 40 (1): 153–184.
Lee, J. D., and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80.
Madhavan, P., D. A. Wiegmann, and F. C. Lacson. 2006. "Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids." *Human Factors* 48 (2): 241–256.
Marakas, G. M., R. D. Johnson, and J. W. Palmer. 2000. "A Theoretical Model of Differential Social Attributions toward Computing Technology: When the Metaphor Becomes the Model." *International Journal of Human-Computer Studies* 52 (4): 719–750.
McKinstry, B., and S. Y. Yang. 1994. "Do Patients Care about the Age of Their General Practitioner? A Questionnaire Survey in Five Practices." *The British Journal of General Practice* 44 (385): 349–351.
Merritt, S. M., H. Heimbaugh, and J. LaChapell. 2012. "I Trust It, but I Don't Know Why: Effects of Implicit Attitudes toward Automation on Trust in an Automated System." *Human Factors* 55 (3): 520–534.
Merritt, S. M., and D. R. Ilgen. 2008. "Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human Automation Interactions." *Human Factors* 50 (2): 194–210.
Mosier, K. L., L. J. Skitka, M. D. Burdick, and S. T. Heers. 1996. "Automation Bias, Accountability, and Verification Behaviours." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40 (4): 204–208.
Mouloua, M., J. A. -A. Smither, D. A. Vincenzi, and L. Smith. 2002. "Automation and Aging: Issues and Considerations." *Advances in Human Performance and Cognitive Engineering Research* 2: 213–237.
Nass, C., J. Steuer, and E. R. Tauber. 1994. "Computers Are Social Actors." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence*, 72–78.
Pak, R., N. Fink, M. Price, B. Bass, and L. Sturre. 2012. "Decision Support Aids with Anthropomorphic Characteristics Influence Trust and Performance in Younger and Older Adults." *Ergonomics* 55 (9): 1059–1072.
Parasuraman, R., and C. A. Miller. 2004. "Trust and Etiquette in High-Criticality Automated Systems." *Communications of the ACM* 47 (4): 51–55.
Perdue, C. W., and M. B. Gurtman. 1990. "Evidence for the Automaticity of Ageism." *Journal of Experimental Social Psychology* 26 (3): 199–216.
Posthuma, R. A., and M. A. Campion. 2009. "Age Stereotypes in the Workplace: Common Stereotypes, Moderators, and Future Research Directions?" *Journal of Management* 35 (1): 158–188.
Qiu, L., and I. Benbasat. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems." *Journal of Management Information Systems* 25 (4): 145–182.
Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models*. 2nd ed. Thousand Oaks, CA: Sage.
Real Softtware. 2013. [Computer software]. Austin, TX.
Rosen, B., and T. H. Jerdee. 1976. "The Nature of Job-Related Age Stereotypes." *Journal of Applied Psychology* 61 (2): 180–183.
Rossi, P. H., and A. B. Anderson. 1982. "The Factorial Survey Approach: An Introduction." In *Measuring Social Judgments*, edited by P. H. Rossi and S. L. Nock. Beverly: Hills Sage.
Rovira, E., K. McGarry, and R. Parasuraman. 2007. "Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task." *Human Factors* 49 (1): 76–87.
Röttger, S., K. Bali, and D. Manzey. 2009. "Impact of Automated Decision Aids on Performance, Operator Behaviour and Workload in a Simulated Supervisory Control Task." *Ergonomics* 52 (5): 512–523.
Shah, R., and J. Ogden. 2006. "What's in a Face? The Role of Doctor Ethnicity, Age and Gender in the Formation of Patients' Judgements: An Experimental Study." *Patient Education and Counseling* 60 (2): 136–141.
Singer, M. S. 1986. "Age Stereotypes as a Function of Profession." *The Journal of Social Psychology* 126 (5): 691–692.
Singh, I. L., R. Molloy, and R. Parasuraman. 1993. "Individual Differences in Monitoring Failures of Automation." *The Journal of General Psychology* 120 (3): 357–373.
Skitka, L., K. Mosier, and M. Burdick. 1999. "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51 (5): 991–1006.
Tabachnick, B. G., and L. S. Fidell. 2007. "Multi-Level Linear Modeling." In *Using Multivariate Statistics*. 5th ed., 781–857. Boston, MA: Pearson.
Wickens, C. D., and S. R. Dixon. 2007. "The Benefits of Imperfect Diagnostic Automation: A Synthesis of the Literature." *Theoretical Issues in Ergonomics Science* 8 (3): 201–212.
Wrenn, K. A., and T. J. Maurer. 2004. "Beliefs about Older Workers' Learning and Development Behavior in Relation to Beliefs about Malleability of Skills, Age-Related Decline, and Control." *Journal of Applied Social Psychology* 34 (2): 223–242.
Yee, N., J. N. Bailenson, and K. Rickertsen. 2007. "A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces." In *Presented at the CHI 07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY: ACM.
Zhang, T., B. Zhu, and D. B. Kaber. 2011. "Anthropomorphism and Social Robots." In *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*, edited by C. C. Hayes and C. A. Miller, 231–255. Boca Raton: Auerbach.

Downloaded by [Clemson University] at 04:43 24 July 2014

Appendix. Multi-level model

Model 2:	
Level 1:	$TRUST_{it} = \beta_{0it} + \beta_{1it}(AgntGndr) + \beta_{2it}(AgntAge) + \beta_{3it}(Reliab) + \beta_{4it}(AgntGndr*AgntAge)$ $+ \beta_{5it}(AgntGndr*Reliab) + \beta_{6it}(AgntAge*Reliab) + \beta_{7it}(AgntGndr*AgntAge*Reliab) + r_{it}$
Level 2:	$\beta_{0i} = \gamma_{00} + u_{0i}$
	$\beta_{1i} = \gamma_{10}$
	$\beta_{2i} = \gamma_{20}$
	$\beta_{3i} = \gamma_{30}$
	$\beta_{4i} = \gamma_{40}$
	$\beta_{5i} = \gamma_{50}$
	$\beta_{6i} = \gamma_{60}$
	$\beta_{7i} = \gamma_{70}$
Model 3:	
Level 1:	$TRUST_{it} = \beta_{0it} + \beta_{1it}(AgntGndr) + \beta_{2it}(AgntAge) + \beta_{3it}(Reliab) + \beta_{4it}(AgntGndr*AgntAge)$ $+ \beta_{5it}(AgntGndr*Reliab) + \beta_{6it}(AgntAge*Reliab) + \beta_{7it}(AgntGndr*AgntAge*Reliab) + r_{it}$
Level 2:	$\beta_{0i} = \gamma_{00} + \gamma_{01}(AGE) + \gamma_{02}(GENDER) + \gamma_{03}(DKS) + \gamma_{03}(CPRS) + u_{0i}$
	$\beta_{1i} = \gamma_{10}$
	$\beta_{2i} = \gamma_{20}$
	$\beta_{3i} = \gamma_{30} + \gamma_{31}(AGE*GENDER)$
	$\beta_{4i} = \gamma_{40}$
	$\beta_{5i} = \gamma_{50} + \gamma_{51}(GENDER) + \gamma_{52}(AGE)$
	$\beta_{6i} = \gamma_{60} + \gamma_{61}(AGE)$
	$\beta_{7i} = \gamma_{70}$

Downloaded by [Clemson University] at 04:43 24 July 2014