# Brains or Beauty: How to Engender Trust in User-Agent Interactions



BESTE F. YUKSEL, Tufts University, University of San Francisco
PENNY COLLISSON, Microsoft
MARY CZERWINSKI, Microsoft Research

Software-based agents are becoming increasingly ubiquitous and automated. However, current technology and algorithms are still fallible, which considerably affects users' trust and interaction with such agents. In this article, we investigate two factors that can engender user trust in agents: reliability and attractiveness of agents. We show that agent reliability is not more important than agent attractiveness. Subjective user ratings of agent trust and perceived accuracy suggest that attractiveness may be even more important than reliability.

CCS Concepts: • **Human-centered computing** → **Web-based interaction**; **User centered design**; **Interaction design theory, concepts and paradigms**; • **Applied computing** → **Psychology**; **Sociology**

Additional Key Words and Phrases: Intelligent personal assistants, software agents, trust, reliability, attractiveness, intelligent agents

**ACM Reference Format:**
Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. ACM Trans. Internet Technol. 17, 1, Article 2 (January 2017), 20 pages.
DOI: http://dx.doi.org/10.1145/2998572

## 1. INTRODUCTION

Software agents or intelligent personal assistants are playing an increasingly prominent and ubiquitous role in our lives. Examples of such smart personal agents include Microsoft's Cortana [2016], Apple's Siri [2016], and Google's Google Now [2016]. Agents can collaborate with users or perform tasks on users' behalf autonomously, gradually becoming more effective as they learn the users' interests, habits, and preferences. Maes [1994] states that two main problems need to be solved when building software agents: (1) *competence*: decisions of when to help users and what to help them with, and (2) *trust*: how we can guarantee that users will trust the agent in making those decisions.

Humans have been shown to have "algorithm aversion," where they show greater intolerance to an algorithmic error than a human error [Dietvorst et al. 2014]. While current artificial intelligence remains fallible, it is crucial that as researchers in user-agent interactions, we understand and develop the second problem area defined by Maes [1994]: how to build and maintain trust in the human-computer relationship even when the system inevitably makes errors. Trust has been defined as "an evolving,

Authors' addresses: B. F. Yuksel, Department of Computer Science, Tufts University, 161 College Avenue, Medford MA 02155; Department of Computer Science, University of San Francisco, Harney Science Center, 2130 Fulton Street, San Francisco CA 94117; email: byuksel@usfca.edu; P. Collison and M. Czerwinski, One Microsoft Way, Redmond WA 98052; emails: {penny.marsh, marycz}@microsoft.com.

DOI: http://dx.doi.org/10.1145/2998572

affective state including both cognitive and affective elements and emerges from the perceptions of competence and a positive, caring motivation in the relationship partner to be trusted" [Young and Albaum 2002]. What is interesting about this definition is its intrinsic ties with user affect. This idea of trust containing both affective and cognitive elements was first introduced in the field of sociology [Lewis and Weigert 1985]. In the context of user-agent interactions, *affective trust* is related to the perception that an agent's actions are intrinsically motivated and is based on affect generated by care, concern, and security [Rempel et al. 1985]. *Cognitive trust* is based on rational thought processes and empirical evidence [Lewis and Weigert 1985]. Cognitive trust is very closely related to an intelligent agent's competence as defined by Maes [1994], whereas affective trust could be developed by ensuring and enhancing other factors in the agent.

In order to engender trust in intelligent agents, researchers have been turning to paradigms developed in social psychology that investigate human-to-human relationships and applying them to human-computer interaction [Reeves and Nass 1996; Bickmore and Picard 2005]. A very well-established paradigm is that of the physical attractiveness stereotype where people attribute other positive personality traits onto attractive people [Dion et al. 1972]. A correlation between physical attractiveness and trustworthiness has been shown across several fields. Perhaps most interestingly, it has been demonstrated in acts of transgression or breaches of trust. Adults view more attractive children's transgressions less negatively and view less attractive children's offenses as more of an enduring dispositional fault [Dion 1972]. Mock juror studies show that physically attractive defendants are found to be guilty less often and are given shorter sentences [Efran 1974; Mazzella and Feingold 1994]. Studies of judicial decisions in real courtrooms show that attractiveness is highly predictive of both minimum and maximum sentences [Stewart 1980, 1985] and that judges exhibited a strong attractiveness bias in the bails and fines that they set [Downs and Lyons 1991].

As software agents will become increasingly embodied in the future, it seems unlikely that unattractive intelligent personal assistants will become the norm. However, in this article, we investigate whether the power of the physical attractiveness stereotype can be leveraged to cope with deficiencies in the machine. We investigate the effects of reliability and attractiveness of embodied agents on users' trust. Thus, the contributions of this article are as follows:

—*Build a trust evaluation system consisting of a Q&A search task with virtual agents to investigate the factors of reliability and physical attractiveness.*
—*Demonstrate through questionnaire data that attractiveness is more important than reliability for user perceptions of trust and accuracy.*
—*Demonstrate through user-agent interactions that reliable agents also need to be attractive for users to engage in behavior that denotes trust.*

We present our contributions by describing how the materials were required and assessed before building our experimental design. We discuss the mapping of the participants' responses to agents to the measure of trust in our evaluation of agent reliability and attractiveness. Finally, we discuss the implications of the findings in terms of future virtual agent design and human-computer interaction.

## 2. MOTIVATION AND BACKGROUND

### 2.1. Trust Between Humans and Computers

Research on trust between humans and intelligent systems can be divided into two broad categories: (1) *the algorithms of the decision making*—that is, the *what* and *when* of the system's adaptations, and (2) *the interaction style and design*—that is, *how* the system interacts with the user.

These two categories can be linked to the two categories defined in cognitive and affective trust. Trust in algorithmic decision making, that is, the artificial intelligence of the system-based agent, is akin to cognitive trust, whereby trust is based on rational thinking. However, current artificial intelligence technology is fallible. While errors still exist in intelligent systems, as user-agent interaction researchers, we can draw from the power of the second category, the interaction style and design, to cope with deficiencies in the machine. In other words, we could build affective trust in system-based agents by designing the system in a way that will engender user trust, even if some mistakes are made. We discuss the two categories in further detail later.

*2.1.1. Trust and the Decision-Making Algorithms.* The importance of the adaptations made by the system can determine whether the intelligent system should be entrusted to intervene and what actions it should take [Dabbish and Baker 2003]. Transparency of such algorithms behind the system's decisions, knowledge of the system's resources,[Glass et al. 2008], and inclusion of comprehensive product information [Lee et al. 2000] have been found to be important to engendering trust with the user. Not only do users want to understand the decision-making processes, but also they would like mechanisms to override erroneous behavior when necessary [Glass et al. 2008]. They also are much more reluctant to trust intelligent systems to make important changes such as making a purchase with a credit card [Glass et al. 2008].

System reliability is crucial to building trust between the human and computer and needs to be built early on Maltz and Meyer [2000] and LeeTiernan et al. [2001]. Maltz and Meyer [2000] provided users carrying out a visual task with potentially helpful cues that varied in their reliability. They found that only the users who received highly reliable cues in the first set of trials continued to accept suggestions in the second set of trials. LeeTiernan et al. [2001] found that when users' first impression of system reliability was high, they continued to use system suggestions, even when reliability decreased. Conversely, when first impressions denoted low system reliability, users would not accept system suggestions even when reliability increased [LeeTiernan et al. 2001]. Similarly, when users viewed an algorithmic error in a forecasting system, they lost confidence in the system and chose an inferior human forecaster over it [Dietvorst et al. 2014].

*2.1.2. Trust and the Interaction Style and Design.* HCI researchers have been taking paradigms in social psychology used in human-to-human relationships and applying them to develop human-to-computer relationships [Reeves and Nass 1996; Bickmore and Picard 2005]. Transfer of results has been found using these paradigms as people demonstrate anthropomorphic behavior toward computers [Nass and Moon 2000]. Behaviors include applying human social categories such as gender and ethnicity to computers or engaging in social behaviors such as politeness and reciprocity [Reeves and Nass 1996; Nass and Moon 2000].

Social psychology has shown that humans are persuaded through association with good feelings [Schwarz et al. 1991; Mackie and Worth 1991] and that they prefer people who like them [Kenny and Nasby 1980]. So it is not surprising that flattery can be a powerful tool to increase likability [Berscheid and Hatfield 1969]. For example, Fogg and Nass [1997] found that humans are susceptible to flattery from computers and that the effects are the same as flattery from humans. Participants subjected to flattery reported greater performance and more positive evaluations of the human-computer interaction [Fogg and Nass 1997]. Lee [2008] also found that flattery increased positive impressions of the computer; however, flattery was also found to increase user suspicion about the validity of computer feedback and actually lowered acceptance of computer suggestions. A suggested explanation for this has been that flattery temporarily fosters mindfulness, thus making users scrutinize the available information [Lee 2009].

The similarity or familiarity paradigm in social psychology has demonstrated that people will like others who are similar to them [Griffitt and Veitch 1974; Kandel 1978]. Users prefer computers that match them in personality over those that do not [Reeves and Nass 1996]. Interestingly, users actually prefer computers that become more like them over time more than computers that maintain a consistent level of similarity [Reeves and Nass 1996].

Other relational behaviors used in human relationships have been adopted by HCI researchers. Virtual agents that use humor are rated as being more likable, competent, and cooperative [Morkes et al. 1998]. Computers that engage in reciprocal and deepening self-disclosure during conversation engender more personal information from the user and increase the chances of a product sale [Moon 1998]. Highly expressive pedagogical agents increase students' perception of trust [Lester et al. 1997]. Bickmore and Picard [2005] found that a relational embodied agent that used social dialogue, meta-relational dialogue, politeness, nonverbal empathy exchanges, and continuity behaviors was trusted, respected, and liked more than a nonrelational agent.

Another human trait that has been investigated is the physical attractiveness stereotype, which is a long-held paradigm in social psychology stemming from "what is beautiful is good," whereby people attribute other positive traits onto attractive people [Dion et al. 1972]. We will now discuss it in further detail.

### 2.2. Physical Attractiveness Stereotype

The unfortunate and uncomfortable truth is that, as human beings, we are positively biased toward physically attractive people. Physically attractive people are thought to possess more favorable characteristics such as being more sociable, happier, sexually warmer [Miller 1970; Dion et al. 1972; Alley and Hildebrandt 1988], better adjusted and more self-assertive [Eagly et al. 1991], more intelligent [Jackson et al. 1995], more successful in life [Dion et al. 1972; Nida and Williams 1977], and better marital partners [Nida and Williams 1977], as well as having greater social power [Mills and Aronson 1965; Sigall et al. 1969], having more positive effects on other people and receiving more positive responses from others including requests for help and requests at work [McGuire 1969], being more persuasive [Baker and Churchill Jr 1977; Dion and Stein 1978; Chaiken 1979], and being more qualified for jobs [Hosoda et al. 2003] than less physically attractive people.

The physical attractiveness stereotype has held up consistently across different populations and exists in older, more sophisticated groups such as health professionals [Nordholm 1980], teachers [Elovitz and Salvia 1983], and other professional groups [Alley and Hildebrandt 1988]. The stereotype also holds in young children who are biased toward more attractive children [Dion 1973; Dion and Berscheid 1974; Langlois and Stephan 1981]. Even 3-month-old infants show selective preference for attractive over unattractive adult faces [Langlois et al. 1987; Samuels and Ewy 1985], suggesting that attractiveness standards are more inherently acquired.

While two reviews have found the physical attractiveness stereotype to be stronger for certain personality traits, with less correlation across integrity of character and attractiveness [Eagly et al. 1991; Feingold 1992], there is a wealth of evidence across many fields to suggest that there is certainly some favorable bias in the area of trustworthiness and physical attractiveness. Patzer [1983] found that communicators with higher levels of physical attractiveness were perceived as more trustworthy and of higher expertise than communicators of lower physical attractiveness. Higher communicator physical attractiveness also had a significant effect for increased liking of the communicator [Patzer 1983]. This ties in with findings that celebrities who are liked will also be trusted [Friedman et al. 1978]. Friedman and Friedman [1976] found a

significant correlation between physical attractiveness and trustworthiness of political figures.

This correlation between physical attractiveness and trustworthiness carries into breaches of trust as well. Adults view more attractive children's transgressions less negatively. Moreover, when the child is unattractive, the offense is seen to be more of an enduring dispositional fault [Dion 1972].

This bias translates into later life where mock juror studies show that physically attractive defendants are treated more leniently than unattractive defendants for many types of crimes such as robbery, rape, cheating, and negligent homicide [Efran 1974; Sigall and Ostrove 1975; Mazzella and Feingold 1994]. Physically attractive defendants are found to be guilty less often and are given shorter sentences [Efran 1974; Mazzella and Feingold 1994]. This has not been found to be the case for the crime of swindling, where the crime is attractiveness related [Sigall and Ostrove 1975; Mazzella and Feingold 1994]. While Mazzella and Feingold [1994] suggest that such results are linked to defendant likability in that attractiveness may be promoting liking, Abwender and Hough [2001] controlled for likability and found unchanged results, suggesting that jurors are responding to something more than just likability.

The correlation between defendant physical attractiveness is also externally valid as corroborated by length of sentencing in real courtrooms by judges and juries. Stewart [1980] found that attractiveness was highly predictive of both minimum and maximum sentences—the more attractive the defendant, the less serious the sentence imposed. These findings were confirmed by Stewart [1985] and were found to remain significantly correlated even when the seriousness of the crime was controlled. Downs and Lyons [1991] investigated judges' sentencing of 915 female and 1,320 male defendants who were unable to alter their appearance levels prior to court appearances. They found, once again, that judges exhibited a strong attractiveness bias in the bails and fines that they set (for misdemeanor charges) [Downs and Lyons 1991].

### 2.3. Hypothesis

In light of the importance of physical attractiveness in the literature, our goal was to investigate and compare it to the importance of reliability in virtual agents. We took the conservative view that, even though attractiveness has been shown to be an important factor in engendering trust, it would not be *more* important than agent reliability. Thus, we set out to investigate the following hypothesis:

*Hypothesis: Reliability is more important than attractiveness in building trust in user-agent interactions.*

## 3. METHOD

### 3.1. Materials

We chose to test highly attractive versus neutrally attractive agents and avoid the unattractive condition as it would be highly unlikely that system designers would purposefully create unattractive agents. We prepared the material for agency appearance and voice accordingly.

*3.1.1. Attractiveness of Physical Appearance.* Research has shown that certain facial features are optimized for attractiveness, such as when the distance between the eyes and mouth is 36% of the face's length, or when the distance between the eyes is 46% of the face's width [Pallett et al. 2010]. A large increase in the distance between the eyes and mouth is perceived as grotesque [Searcy and Bartlett 1996]. The eyes and mouth have consistently been the specific facial regions found to be most influential for facial attractiveness ratings [Alley and Hildebrandt 1988].

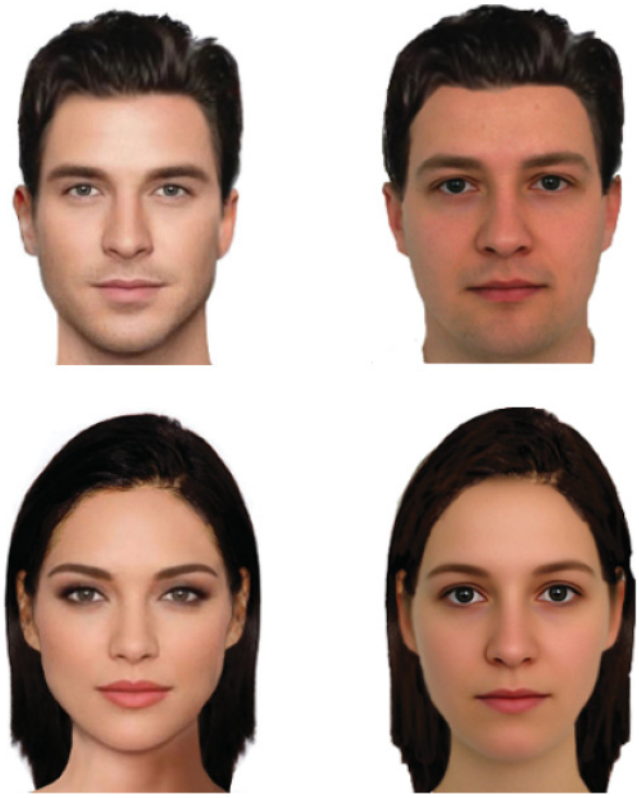

Fig. 1. Left: Highly attractive agents courtesy of © Dr. Stuart Gibson and Dr. Chris Solomon, University of Kent. Right: Neutrally attractive agents courtesy of © Dr. Martin Gruendl, University of Regensburg [Beautycheck 2016].

We turned to two external sources that have worked in the area of facial features and attractiveness, who kindly gave us their permission to use faces that they had already synthesized.

The two attractive agents' faces were provided courtesy of Dr. Stuart Gibson and Dr. Chris Solomon from the University of Kent. They created the faces as archetypal models of beauty from externally rated features of beauty by 100 people in the United Kingdom (http://www.telegraph.co.uk/news/11502572/Are-these-the-most-beautiful-faces-in-the-world.html). Important features of beauty included distance between the eyes, nose length and width, thickness of lips, and width of mouth. Figure 1 shows the highly attractive agents on the left. The two neutrally attractive agents' faces were provided courtesy of Dr. Martin Gruendl from the University of Regensburg [Beautycheck 2016]. We transplanted the same hair onto both the attractive and neutral faces as pilot studies showed that participants' ratings of agent attractiveness was being affected by differences in hair (Figure 1). We then ran our own preliminary study to verify the attractiveness of the agents. We asked 20 participants (11 female) to rate the physical attractiveness of the agents on a scale of 1 to 7 with 1 being "very unattractive" and 7 being "very attractive." We provided participants with agents of their main gender of preference to facilitate agent attractiveness (all participants were heterosexual, so they rated the opposite gender's physical attractiveness). There is research to show

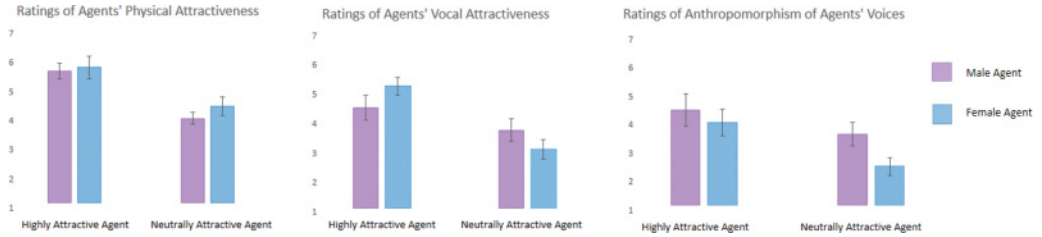


Fig. 2. Mean and standard error of participants' ratings of the highly and neutrally attractive agents: *left*: physical attractiveness, *center:* vocal attractiveness, and *right:* vocal human-likeness.

that indiscretions carried out by attractive people of the same sex can be more harshly judged [Abwender and Hough 2001].

Figure 2 shows that participants rated the agents that were designed to be more physically attractive higher than those designed to be neutrally attractive. We note here that these faces may be viewed differently in different cultures; however, there is evidence that the physical beauty of female faces translates across cultures [Cunningham et al. 1995].

*3.1.2. Attractiveness of Voice.* Vocal attractiveness has been shown to increase the attractiveness of face plus voice agency [Zuckerman and Driver 1989]; it was therefore important that we maintained a differentiation of attractiveness in the vocalization of the agents. A stereotype for vocal attractiveness has been shown [Zuckerman and Driver 1989; Zuckerman et al. 1990] where attractive voices are rated more favorably in voice-only and face-plus-voice conditions.

Studies from biological psychology have shown that men seem to prefer women with higher-pitched voices, which is associated with smaller body size [Liu and Xu 2011], and women actually speak in a higher pitch to men they find attractive [Fraccaro et al. 2011]. Conversely, women seem to find men with lower-pitched voices more attractive [Collins 2000; Apicella et al. 2007] as this seems to indicate larger body size and other masculine traits [Evans et al. 2006].

We used these findings when building the voices of our agents using an internal Microsoft text-to-speech toolkit software. The female attractive voice was higher in pitch than the female neutral voice. The male attractive voice was lower in pitch than the male neutral voice. Furthermore, both the male and female attractive voices had prosodic contours with slight rises and falls in pitch, emphasis in places, and slight changes in the rhythm of speech. The neutral voices simply had a neutral tone throughout. We chose a neutral tone to match the neutral level of physical attractiveness. As discussed earlier, we are comparing attractiveness to neutrality rather than unattractiveness as it seems unlikely that unattractive virtual agents would be purposefully designed and created.

We asked 21 participants (11 female) to rate the vocal attractiveness of the voices on a scale of 1 to 7 with 1 being "very unattractive" and 7 being "very Attractive." Participants rated their main gender of sexual preference (all participants were heterosexual so they rated the opposite gender's voices). Figure 2 shows that participants rated the voices that were designed to be more attractive higher than those designed to be neutrally attractive.

Due to the differences in prosodic contours creating more human-like speech patterns, we also asked participants to rate how human-like versus machine-like the voices sounded, with 1 denoting "very machine-like" and 7 being "very human-like." Figure 2 shows that participants thought that the neutral voices sounded more machine-like

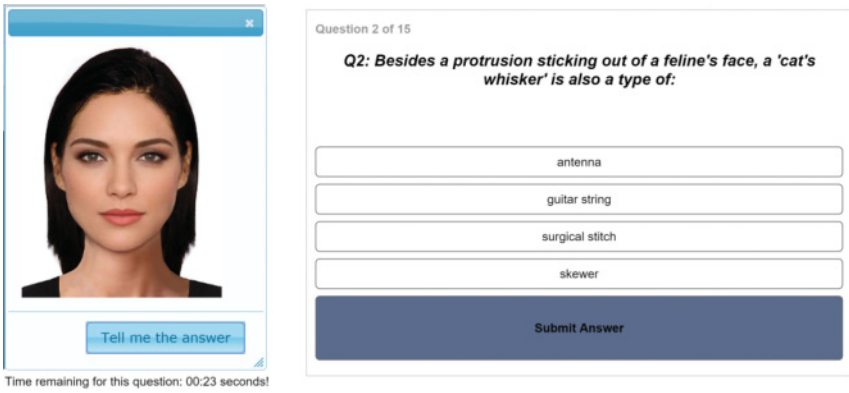


Fig. 3. Example of a question with an agent offering a suggestion. Participants would hear the agent offering a suggestion as well. If they wished, the participant could hear the suggestion by pressing on the "Tell me the answer" button.

than the attractive voices. It is possible, therefore, that this factor affects participants' preferences, which we consider when evaluating the results.

### 3.2. Experimental Design

Participants took part in a within-subject 2 (high vs. low reliability)× 2 (high vs. neutral attractiveness) agency design. Participants answered four sets of 15 questions. Each set had one of the four conditions:

—A highly reliable and highly attractive agent
(Reliable$_{HIGH}$ Attractive$_{HIGH}$)
—A highly reliable and neutrally attractive agent
(Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$)
—A lesser reliable and highly attractive agent
(Reliable$_{LOW}$ Attractive$_{HIGH}$)
—A lesser reliable and neutrally attractive agent
(Reliable$_{LOW}$ Attractive$_{NEUTRAL}$)

The ordering of the conditions was alternated so that there was a fair distribution. We stressed to participants that the agent presented at each trial was independent of agents from any previous trials in order to avoid attribution of perceived trustworthiness from previous agents. We controlled for agents to be participants' main gender of sexual preference. All participants reported being heterosexual and therefore only interacted with agents of the opposite gender.

In the high-reliability conditions, **five out of the six** suggestions offered by the agent were correct. In the low-reliability conditions, **three out of the six** suggestions offered by the agent were correct. The percentages for high and low reliability (83% and 50%, respectively) were taken from user trust literature [LeeTiernan et al. 2001] (to clarify, LeeTiernan et al. [2001] used 80% for high reliability). The order of the correct and incorrect answers were randomized in each trial.

The questions were multiple choice and were general knowledge. Questions were on a broad range of topics and were designed to a difficulty level that required searching on most questions for most participants [Hyperion 2000; Wood and Kovalchik 2012]. An example is included in Figure 3. Participants could search for the answers in a search engine on a separate monitor. They had 40 seconds to answer each question and they could see how much time they had left for each question with a countdown showing

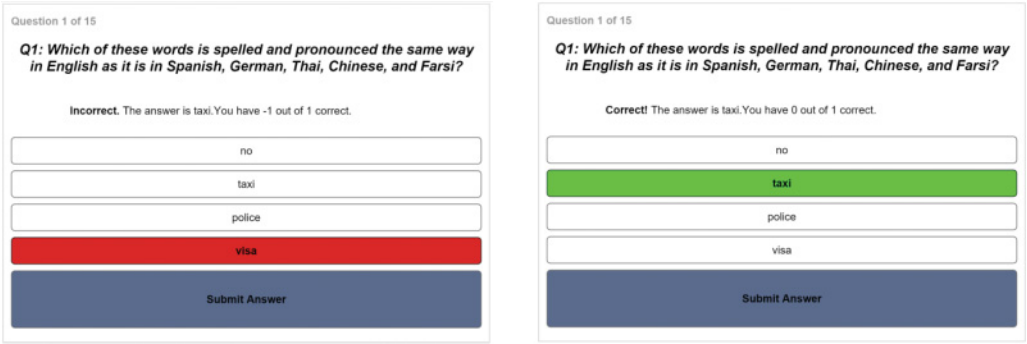


Fig. 4. Examples of feedback for an incorrect (left) and correct (right) answer. Participants gained a point for every correct answer and lost a point for every incorrect answer and every time they timed out in order to discourage guesswork.

on the screen (Figure 3). For every question they answered correctly, they would gain a point, and they could move on to the next question. If they answered a question incorrectly, they would lose a point (to discourage guesswork) and would have to keep attempting the question until they got the right answer or timed out. If they timed out they would also lose a point. They received a 2-minute break in between each set of questions. Participants were incentivized to reach high scores in the questions with extra gratuity. Participants always received immediate feedback when they answered a question (Figure 4).

Participants were told that every now and then an agent would appear on the screen to help them with the answer to a question. They were told that the agents would offer a suggestion for the answer and that they could click on the "Tell me the answer" button to hear the suggestion (Figure 3). Before starting the experiment, participants were introduced to the two agents (highly and neutrally attractive) of their gender of preference from Figure 1.

For each set of 15 questions, six questions were picked pseudo-randomly on each trial to have suggestions. The suggestions were spread out so that the first five questions had two suggestions, the second five questions had two suggestions, and the last five questions had two suggestions.

*3.2.1. Participants.* Twenty participants took part in the study. One participant had to be excluded from the dataset as the participant did not follow the instructions. Out of the remaining 19 (11 female), there was a wide range in ages: 17 to 51 years ($\mu = 30, \sigma = 11.24$). Ten participants were recruited internally from Microsoft Research, and nine were recruited externally from the general public. Participants received a gift certificate for their participation. They were also offered an additional incentive of a $25 Amazon gift card if they could achieve 75% accuracy or higher in one of the sets of questions.

## 4. MAPPING OF MEASURES TO TRUST

Trust is a complex issue, and measuring it is no less simple. In this section, we present a mapping of how each measure relates to the topic of trust in user-agent interactions.

*Number of Questions with Button Presses:* Users could choose to hear an agent's suggestion by pressing the "Tell me the answer" button. We analyzed the number of questions with button presses to verify how often users opted to hear agents' suggestions. This is different from the number of button presses in total because in some trials

users pressed the button more than once to relisten to the suggestion for clarification or reminder purposes. There was no cost to listening to agent suggestions.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

*Self-Reported Measures of Trust in Agent:* This measured users' trust in the agents.

*Self-Reported Measures of Perceived Accuracy of Agent:* This measured users' perceived accuracy of agents. This reveals whether perceived agent reliability can be affected by other factors such as attractiveness, hence indicating where trust is being directed.

*Number of Correct Suggestions Accepted:* Due to the difference in the correct answers supplied by the agents with high versus low reliability, this measure should show this difference. A difference in the number of accepted correct answers between the high- and low-reliable agents will indicate a difference in trust between agents. Each correct answer gained the user a point.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

*Number of Incorrect Suggestions Accepted:* Similarly, this measure should show a difference between agents with high and low reliability due to the difference in incorrect answers supplied by those two groups of agents. A difference in the number of accepted correct answers between the high- and low-reliable agents will indicate a difference in trust between agents. This measure is different from the number of correct suggestions accepted as users can select more than one incorrect answer for each question. Each incorrect answer cost the user a point.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

*Length of Time Spent on Questions with Suggestions:* We took this measure to suggest trust by whether a user will accept the agent's suggestion or verify the answer with an Internet search, the assumption being that for the agents with high reliability, generally the longer the time taken, the less trust in the agent. However, due to agents with less reliability offering a greater number of incorrect suggestions, greater time spent may indicate that users were accepting incorrect suggestions, which is still indicative of trust. Therefore, this measure is best compared between the agents with high and low reliability.

*Number of Timeouts with Suggestions:* Each question timed out after 40 seconds, costing the user a point. Therefore, if a user timed out on a question where an agent was offering a suggestion, this suggests a lack of trust. This is true across all conditions.

*Number of Correct Answers:* Agents with higher reliability offered a higher percentage of correct answers. Therefore, this measure should show a difference between the agents with high and low reliability. A deviation from this would indicate a difference in trust between agents with high and low reliability.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

*Number of Incorrect Answers:* This measure should show an even greater difference between the agents with high and low reliability due to the fact that users can submit multiple incorrect answers per question. Therefore, questions with suggestions from an agent with low reliability can result in several incorrect answers.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

## 5. RESULTS

### 5.1. Verification of Attractiveness Independent Variable

We first examined participants' ratings for the attractiveness of the agent's appearance and voice in order to verify the independent variables of attractiveness. Participants rated agents' physical and vocal attractiveness on two separate scales of 1 to 7 with 1 being "very unattractive" and 7 denoting "very Attractive."

*5.1.1. Attractiveness of Physical Appearance.* A paired t-test showed extremely significant different responses in participants' ratings of agents' physical attractiveness ($t(37) = 2.03$, $p = .00005$). Participants rated the agents designed to be highly attractive ($\mu = 4.8, \sigma = 0.8$) as significantly more physically attractive than the agents designed to be neutrally attractive ($\mu = 3.9, \sigma = 1.1$).

*5.1.2. Attractiveness of Voice.* A paired t-test also showed significant results for participants' ratings of agents' vocal attractiveness ($t(37) = 2.03$, $p = .02$). Participants rated the voices designed to be highly attractive ($\mu = 4.3, \sigma = 1.3$) as significantly more attractive than the voices designed to be neutrally attractive ($\mu = 3.9, \sigma = 1.2$).

These findings may be due, at least partially, to participants responding to the more human-like quality of the more attractive voices, as rated during preliminary studies of agents' voices (Figure 2).

### 5.2. Questionnaire Data on Trust and Accuracy

We carried out various 2 (high vs. low reliability) $\times$ 2 (high vs. neutral attractiveness) repeated measure (RM) ANOVAs with a Greenhouse-Geisser correction across all dependent variables. For all significant results, we calculated post hoc analyses using Bonferroni corrections.

Results on *self-reported ratings of agent trust* showed a significant difference for the attractive agents ($F(1, 18) = 11.473$, $p = .003$, $\eta^2 = .389$) (Table I). Post hoc analysis revealed that users trusted the $\text{Reliable}_{HIGH}\text{Attractive}_{HIGH}$ agent significantly more than both of the neutrally attractive agents (including the $\text{Reliable}_{HIGH}$ $\text{Attractive}_{NEUTRAL}$ ($p = .011$) and $\text{Reliable}_{LOW}$ $\text{Attractive}_{NEUTRAL}$ ($p = .005$) agents). Interestingly, there was no significant difference between the $\text{Reliable}_{LOW}$ $\text{Attractive}_{HIGH}$ agent and the highly reliable agents (Figure 5) (while Likert scales of 1 to 7 were given to participants with 1 denoting a low value and 7 denoting a high value, graphs show a scale of 0 to 6 to show better differences between conditions).

Results also showed that *users' self-reporting of agent accuracy* was significantly different for more attractive agents ($F(1, 18) = 26.645$, $p < 0.001$, $\eta^2 = .510$) (Table I). Post hoc analysis showed that users thought both of the highly attractive agents were more accurate than the $\text{Reliable}_{HIGH}$ $\text{Attractive}_{NEUTRAL}$ agent. Both the $\text{Reliable}_{HIGH}$ $\text{Attractive}_{HIGH}$ agent ($p = .005$) and the $\text{Reliable}_{LOW}$ $\text{Attractive}_{HIGH}$ agent ($p = .006$) were significantly more accurate than the $\text{Reliable}_{HIGH}$ $\text{Attractive}_{NEUTRAL}$ agent (Figure 5).

These results indicate that attractiveness is more important than reliability for user perceptions of trust and accuracy. We turn to data gathered from user-agent interaction to investigate these findings further.

### 5.3. User-Agent Interaction Data

There was a significant difference in the *number of correct suggestions accepted* for the reliability variable ($F(1, 18) = 25.352$, $p < 0.001$, $\eta^2 = .585$) and the interaction between reliability and attractiveness ($F(1, 18) = 4.569$, $p = .047$) (Table I). As expected, post hoc analysis showed that users accepted a significantly higher number of correct suggestions from the $\text{Reliable}_{HIGH}$ $\text{Attractive}_{HIGH}$ agent than for the lower-reliability

Table I. Results of 2x2 RM-ANOVAs on Agent Reliability and Attractiveness

| Dependent Variable | Factor | F-Value | p-Value | $\eta^2$ |
|---|---|---|---|---|
| Number of Questions with Button Presses | Reliability | 3.485 | .078 | .162 |
| Number of Questions with Button Presses | Attractiveness | 3.547 | .076 | .165 |
| Number of Questions with Button Presses | Interaction | 4.455 | .049 | .198 |
| Questionnaire on Trust | Reliability | 0.548 | .469 | .030 |
| **Questionnaire on Trust** | **Attractiveness** | 9.642 | **.006** | .389 |
| Questionnaire on Trust | Interaction | 0.986 | .334 | .052 |
| Questionnaire on Accuracy | Reliability | 1.536 | .231 | .079 |
| **Questionnaire on Accuracy** | **Attractiveness** | 18.731 | **<.001** | .510 |
| Questionnaire on Accuracy | Interaction | 0.706 | .412 | .038 |
| **Number of Correct Suggestions Accepted** | **Reliability** | 25.352 | **<.001** | .585 |
| Number of Correct Suggestions Accepted | Attractiveness | 3.888 | .064 | .178 |
| Number of Correct Suggestions Accepted | Interaction | 4.569 | .047 | .202 |
| **Number of Incorrect Suggestions Accepted** | **Reliability** | 65.681 | **<.001** | .785 |
| Number of Incorrect Suggestions Accepted | Attractiveness | 0.029 | .867 | .002 |
| Number of Incorrect Suggestions Accepted | Interaction | 0.073 | .790 | .004 |
| **Time Spent on Questions with Suggestions** | **Reliability** | 13.027 | **.002** | .420 |
| **Time Spent on Questions with Suggestions** | **Attractiveness** | 6.321 | **.022** | .260 |
| Time Spent on Questions with Suggestions | Interaction | 0.073 | .790 | .004 |
| Number of Timeouts with Suggestions | Reliability | 6.892 | .017 | .277 |
| Number of Timeouts with Suggestions | Attractiveness | 1.000 | .331 | .053 |
| Number of Timeouts with Suggestions | Interaction | .053 | .821 | .003 |
| Number of Correct Answers | Reliability | 3.512 | .077 | .163 |
| Number of Correct Answers | Attractiveness | 3.337 | .084 | .156 |
| Number of Correct Answers | Interaction | .476 | .499 | .026 |
| **Number of Incorrect Answers** | **Reliability** | 5.773 | **.027** | .243 |
| Number of Incorrect Answers | Attractiveness | 0.585 | .454 | .031 |
| Number of Incorrect Answers | Interaction | .249 | .624 | .014 |

*Note*: All measures that remained significant *after* the Bonferroni correction are highlighted in bold. All results have a df of (1,18).

agents, regardless of attractiveness ($p < 0.001$). This was expected as the highly reliable agents provided a higher number of correct suggestions in the first place than the agents with lesser reliability.

However, we did not expect for there to be *no significance* between the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent and the lower-reliability agents (Figure 5). This occurred even though the agents with lower reliability had many fewer correct suggestions to offer.

There was also a significant difference in the *number of incorrect suggestions accepted* for the reliability variable ($F(1, 18) = 65.681$, $p < 0.001$, $\eta^2 = .785$) (Table I) as expected.

Results for the *time spent on questions when the agent offered suggestions* showed that there was a significant difference in both reliability ($F(1, 18) = 13.027$, $p = .002$, $\eta^2 = .420$) and attractiveness ($F(1, 18) = 6.321$, $p = .022$, $\eta^2 = .260$) (Table I). Post hoc analysis revealed that users spent significantly less time on questions with the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent than the Reliable$_{LOW}$ Attractive$_{NEUTRAL}$ ($p = .003$) agent (Figure 5). Interestingly, once again there were no significant differences between the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent and the low-reliability agents. There were also no significant differences between the Reliable$_{LOW}$ Attractive$_{HIGH}$ agent and the high-reliability agents (Figure 5).

We used the median when measuring time as is standard practice when looking at reaction time distributions due to its robustness to skewness in the data. Figure 5 shows the mean of the medians.


Fig. 5. Mean and standard error of users' self-reported trust in agents (top left), user perception of agent accuracy (top right), number of correct suggestions accepted (bottom left), and time spent on questions with suggestions (bottom right). Significant differences are indicated by an asterisk. Results indicate significant differences based on attractiveness in users' trust in agents ($F(1, 18) = 11.473$, $p = .003$) and user perception of agent accuracy ($p = .005$ and $p = .006$). Contradictory to expectations, there were no significant differences between the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent and the low-reliability agents in either the number of correct suggestions accepted or in the time spent on questions with suggestions.

There was a significant difference in the *number of incorrect answers* in reliability ($F(1, 18) = 5.773, p = .027, \eta^2 = .243$) (Table I) as expected due to the difference in incorrect suggestions offered between agents with high and low reliability.

However, there were no significant differences in the *number of correct answers* across conditions (Table I).

There were also no significant differences in the *number of questions with button presses* (indicating the number of times users asked to hear the suggestion from the agent) (Table I), which is not surprising considering there was no cost or penalty associated with hearing the suggestion. There were also no differences in the *number of timeouts in questions with suggestions* (Table I).

We also examined all measures for gender differences with $2\times2\times2$ RM-ANOVAs and found *no differences in gender* throughout the results.

## 6. DISCUSSION

### 6.1. Reliability Manipulation Check

Users seem to trust agents with high attractiveness more and think that they are more accurate. We first verify that these results were not due to skewed exposure from different agents by examining the dependent variables.

*6.1.1. Number of Button Presses.* There was no significant difference in the number of button presses to hear agents' suggestions between conditions (Table I). Therefore, users saw and heard the same number of suggestions across all conditions.

*6.1.2. Number of Correct Suggestions Accepted.* There was no significant difference in the number of correct suggestions accepted between the low-reliability agents and the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent (Figure 5). Therefore, differences in data between those three agents are not due to a skew in exposure to correct answers being accepted.

*6.1.3. Number of Incorrect Suggestions Accepted.* There was a significant difference in the number of incorrect suggestions accepted between the high- and low-reliability agents (Table I), which was expected due to the higher number of incorrect suggestions provided by low-reliability agents. Therefore, any differences between highly and neutrally attractive agents were not due to skewed feedback in incorrect suggestions.

*6.1.4. Number of Correct Answers.* There was no significant difference in the number of correct answers across conditions, which once again verifies that the findings were not due to skewed exposure.

*6.1.5. Number of Incorrect Answers.* There was a significant difference in the number of incorrect answers between the high- and low-reliability agents (Table I), which was to be expected due to the higher number of incorrect suggestions in the low-reliability agents. Therefore, any differences between highly and neutrally attractive agents were not due to skewed feedback in incorrect answers.

It seems, therefore, that the results are due to users' own reactions to and perceptions of the agents. We now examine the data in terms of the hypotheses stated at the start.

### 6.2. Is Reliability More Important Than Attractiveness in Building Trust in User-Agent Interactions?

We now address the hypothesis:

*Reliability is more important than attractiveness in building trust in user-agent interactions.*

Questionnaire data of users' perceived trust and accuracy of virtual agents indicate a rejection of this hypothesis.

Questionnaire data indicated users trusted the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent significantly more than the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent. Users also found the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent to be significantly more accurate than the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent. This subjective evaluation of agents from users sheds light on how important agent attractiveness is in user-human interactions and relationships.

The user-agent interaction data showed us that the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent was not treated differently from the agents with low reliability (Figure 5), which suggests that reliability alone is not enough.

While there were significant differences between the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent and the agents with low reliability in the number of correct suggestions accepted, there were no such significant differences between the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent and the agents with low reliability. This was unusual because we had been expecting a significant difference in the number of correct suggestions accepted between the high- and low-reliability agents because the agents with lower reliability had fewer correct suggestions to offer.

This interaction data strongly suggests that users trusted the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent less than the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent by accepting fewer of their correct suggestions. It is also very interesting that users accepted the same number of correct suggestions from the agents with lower reliability as the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent. Additionally, results showed that there was no significant difference in the length of time spent on questions between the Reliable$_{LOW}$

Attractive$_{HIGH}$ agent and the highly reliable agents. Longer lengths of time spent considering answers when an agent offers a suggestion is indicative that the user is not as trusting, whether by searching for the answers online to verify the agent's answer or by some other decision-making process that involves questioning the agent's suggestion. Together, these findings indicate a distinct lack of trust in user interactions with this highly reliable yet neutrally attractive agent. This result is really quite surprising and demonstrates how even a highly reliable agent's suggestions can be ignored by users if it is not accompanied by high levels of physical attractiveness.

It is possible to draw the following conclusions from the two types of data gathered:

—*Questionnaire data of users' reports of perceived trust and accuracy indicate that attractiveness is more important than reliability. Hence, this is a rejection of the hypothesis that reliability is more important than attractiveness in building trust.*
—*User-agent interaction data showed that reliability alone was not enough to build user trust; that is, the agent had to be attractive as well as reliable. Therefore, this also rejects the hypothesis that reliability is more important than attractiveness in building trust.*

Thus, there is an overall rejection of the hypothesis that reliability is more important than attractiveness in building trust in user-agent interactions.

These findings are representative of the complexity of human nature in interactions with software agents. They highlight the importance and value of acknowledging and exploring traits such as physical attractiveness. We recommend the leveraging of any trait that will help users trust software agents enough to engage with them, including physical attractiveness. This is not to deceive users, but instead, to encourage interaction. Otherwise, as demonstrated in these results, even highly reliable agents can be ignored by users.

### 6.3. Cognitive and Affective Trust

A possible explanation for greater user trust in the more attractive agents, even over the highly reliable but neutrally attractive agent, is the concept of affective trust over cognitive trust. Dowell et al. [2015] found that in the early phase of a relationship cycle in business (in the first 12 months), affective trust was especially important. In later, more mature phases of the relationship cycle (after the first 12 months), cognitive trust was a more important factor than before [Dowell et al. 2015].

When there is less substantive evidence to build trust upon, it is thought that trust is based on the hope or expectation that the task will be fulfilled; therefore, cognitive trust would be less influential early on Dowell et al. [2015]. It is therefore possible that users were basing their trust on the agents in terms of how they felt about the agents (affective trust), which could have been influenced by the level of agent attractiveness. Lewis and Weigert [1985] suggest that the stronger the affective trust in relation to the cognitive trust, the less likely behavioral evidence to the contrary will weaken the relationship. This could have occurred with the agent with lesser reliability but higher attractiveness, and with the agent with high reliability but lesser attractiveness. If users had established an affective attitude toward them based on appearance, this could have created a blindness to evidence-based agent reliability. Examining users' report of trust at two separate moments during each trial, once at the beginning before any interaction and once at the end of the interaction, could also highlight any discrepancies, or lack thereof, in trust over the course of the user-agent interaction. It would also be important to investigate the longer-term relationships between agents and users in terms of cognitive and affective trust.

Affective trust has been modeled into two subcategories by Dowell et al. [2015]: (1) *relational trust,* which is related to the "leap of faith" aspect of affective trust where

faith is placed in the other individual that they will act in a trustworthy way, and (2) *intuitive trust,* which is based on moods and feelings about the other individual. Dowell et al. [2015] found that relational trust was particularly important in the early phase of a relationship. The physical attractiveness stereotype has shown that people often attribute other positive traits onto physically attractive people, which could lend itself to this "leap of faith" category.

It is important to note, however, that trust is a *mixture* of feeling and rational thinking [Weigert 1981]. In any user-agent interaction, it is not possible to exclude one or the other from the analysis of trust, nor is it easy to analyze the effects of the two components separately [Calefato et al. 2015]. However, affective trust seems to play a more important role than cognitive trust in primary group relations (long-lasting, close, personal relationships), whereas cognitive trust seems to play a greater role in secondary group relations (more temporary and less personal relationships) [Lewis and Weigert 1985]. Considering the deeply personal relationship most users have with their computers, it is foreseeable that an intelligent personal assistant or software agent will be included in the user's primary social group, thereby making the affective trust aspect of the relationship even more important. This would need to be investigated in a more long-term study of trust in user-agent interactions.

### 6.4. Physical Versus Vocal Attractiveness in User-Agent Interactions

Is this study, we examined attractiveness as a combination of both physical and vocal attractiveness. It is possible that the two factors had an enhancing effect on the other. For example, vocal attractiveness has been shown to increase the attractiveness of face-plus-voice agency [Zuckerman and Driver 1989]. A separate study needs to be carried out to differentiate between the effectiveness of the two agencies.

After the completion of the main experiment, we carried out a pilot study to this effect to investigate the impact of vocal attractiveness *only* on engendering trust in user-agent interactions. Four new participants (one female) carried out the same experiment with only the agent's voice as stimuli. No visuals of the agents were presented.

Results showed no significant differences for any of the dependent variables across any conditions. The users also reported being so engrossed in the task that they could not tell a marked difference in the agents' voices. This was in contrast to previous findings where 10 participants who heard the agents' voices (also with no physical appearance) in an environment where they were not carrying out any tasks could clearly hear the differences in the voices (Figure 2).

These are preliminary results, however, which require further investigation to differentiate which agency factors elicit the attractiveness stereotype in humans.

### 6.5. Limitations

It may have been easier for users in our study to accept the agents because they were merely 2D graphical representations of a real person and therefore avoided problems related to the uncanny valley. Systems with embodied virtual agents are currently being created that are able to move, gesture, and lip sync quite accurately when interacting with users (e.g., Wang et al. [2011]). Examining the effects of reliability and attractiveness on the development of trust when agents are animated with natural gestures, facial expressions, and voice intonations would allow for a huge increase in agency parameters. This would provide a much more complex framework in which to design and investigate agent attractiveness and its effects on user-agent interactions such as emotional expressivity, which has been shown to relate to trust [Boone and Buck 2003] and decision making [De Melo et al. 2012].

Our interactions with users were merely in the form of question/answer types of interactions. While we used search scenarios and motivated participants with money

to take the tasks seriously, more real-world, complex interaction between the user and the agent, when stakes are higher, would also be a ripe area for future exploration. For example, it has been shown that varying scenarios affects users' preference for agency, such as embodied conversational agents versus social robots [De Carolis et al. 2010].

However, in this work, we feel we have contributed to knowledge about agent design and its effects on humans, highlighting that agent attractiveness is an important consideration that should be kept in mind in future user-agent interactions.

## 7. CONCLUSION

We investigated the effects of reliability and attractiveness on the development of trust in user-agent interactions. We found that we had to reject our hypothesis that reliability was more important than attractiveness. Users reported trusting the highly attractive agents more and found them to be subjectively more accurate, even over the more reliable but less attractive agent. User-agent interaction data from the number of correct suggestions accepted and length of time spent on questions showed that the highly reliable yet neutrally attractive agent was treated in the same way as the agents with low reliability. In addition, this agent was treated significantly differently from the agent who was highly reliable *and* highly attractive.

These results demonstrate that agent reliability is not more important than agent attractiveness in developing trust in user-agent interactions. In fact, attractiveness is at least as important as reliability, and perhaps more so. These findings may be due to users having greater affective trust than cognitive trust in the highly attractive agents, especially as these were early-phase interactions.

We hope this article demonstrates that reliability by itself should not be the only point of focus in developing intelligent software agents. There has been a wealth of research conducted in social psychology for several decades highlighting the power and importance of the physical attractiveness stereotype. There has been further research demonstrating that humans respond to computers as if they were humans. Our study suggests that the attractiveness stereotype will follow humans into the future with our relationships and interactions with intelligent virtual agents.

## ACKNOWLEDGMENTS

The authors would like to thank Steve Macbeth, Ivan Tashev, and Gwen Hardiman from Microsoft. We also thank Drs. Stuart Gibson and Chris Solomon from the University of Kent and Dr. Martin Gruendl from the University of Regensburg for the physical representation of the agents.

## REFERENCES

David A. Abwender and Kenyatta Hough. 2001. Interactive effects of characteristics of defendant and mock juror on U.S. participants' judgment and sentencing recommendations. *Journal of Social Psychology* 141, 5 (2001), 603–615. DOI:http://dx.doi.org/10.1080/00224540109600574

Thomas R. Alley and Katherine A. Hildebrandt. 1988. *Social and Applied Aspects of Perceiving Faces*. Lawrence Erlbaum Associates, Hillsdale, NJ, 101–140.

Coren L. Apicella, David R. Feinberg, and Frank W. Marlowe. 2007. Voice pitch predicts reproductive success in male hunter-gatherers. *Biology Letters* 3, 6 (2007), 682–684.

Michael J. Baker and Gilbert A. Churchill Jr. 1977. The impact of physically attractive models on advertising evaluations. *Journal of Marketing Research* 14, 4 (1977), 538–555.

Beautycheck. 2016. https://www.beautycheck.de.

Ellen Berscheid and Elaine Hatfield. 1969. Interpersonal attraction. Addison Wesley.

Timothy W. Bickmore and Rosalind Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12, 2 (2005), 293–327.

R. Thomas Boone and Ross Buck. 2003. Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior* 27, 3 (2003), 163–182.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2015. The role of social media in affective trust building in customer–supplier relationships. *Electronic Commerce Research* 15, 4 (2015), 453–482.

Shelly Chaiken. 1979. Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology* 37, 8 (1979), 1387.

Sarah A. Collins. 2000. Men's voices and women's choices. *Animal Behaviour* 60, 6 (2000), 773–780.

Michael R. Cunningham, Alan R. Roberts, Anita P. Barbee, Perri B. Druen, and Cheng-Huan Wu. 1995. "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology* 68, 2 (1995), 261.

Laura A. Dabbish and Ryan S. Baker. 2003. Administrative assistants as interruption mediators. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1020–1021.

Berardina De Carolis, Irene Mazzotta, Nicole Novielli, and Sebastiano Pizzutilo. 2010. Social robots and ECAs for accessing smart environments services. In *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, 275–278.

Celso M. De Melo, Peter Carnevale, Stephen Read, Dimitrios Antos, and Jonathan Gratch. 2012. Bayesian model of the social effects of emotion in decision-making in multiagent systems. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 55–62.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2014. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 143, 6 (2014), 1–13.

Karen Dion. 1972. Physical attractiveness and evaluation of children's transgressions*. *Journal of Personality and Social Psychology* 24, 2 (1972), 207–213. DOI: http://dx.doi.org/10.1037/h0033372

Karen Dion, Ellen Berscheid, and Elaine Walster. 1972. What is beautiful is good. *Journal of Personality and Social Psychology* 24, 3 (1972), 285–290. DOI: http://dx.doi.org/10.1037/h0033731

Karen K. Dion. 1973. Young children's stereotyping of facial attractiveness. *Developmental Psychology* 9, 2 (1973), 183.

Karen K. Dion and Ellen Berscheid. 1974. Physical attractiveness and peer perception among children. *Sociometry* (1974), 1–12.

Karen K. Dion and Steven Stein. 1978. Physical attractiveness and interpersonal influence. *Journal of Experimental Social Psychology* 14, 1 (1978), 97–108.

David Dowell, Mark Morrison, and Troy Heffernan. 2015. The changing importance of affective trust and cognitive trust across the relationship lifecycle: A study of business-to-business relationships. *Industrial Marketing Management* 44 (2015), 119–130.

A. Chris Downs and Phillip M. Lyons. 1991. Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin* 17, 5 (1991), 541–547.

Alice H. Eagly, Richard D. Ashmore, Mona G. Makhijani, and Laura C. Longo. 1991. What is beautiful is good, but: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin* 110, 1 (1991), 109.

Michael G. Efran. 1974. The effect of guilt, of physical appearance on the judgment severity interpersonal attraction, and of recommended punishment in a simulated. *Journal of Research in Personality* 8 (1974), 45–54.

Gerald P. Elovitz and John Salvia. 1983. Attractiveness as a biasing factor in the judgments of school psychologists. *Journal of School Psychology* 20, 4 (1983), 339–345.

Sarah Evans, Nick Neave, and Delia Wakelin. 2006. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology* 72, 2 (2006), 160–163.

Alan Feingold. 1992. Good-looking people are not what we think. *Psychological Bulletin* 111, 2 (1992), 304.

Brian J. Fogg and Clifford Nass. 1997. Silicon sycophants: The effects of computers that flatter. *International Journal of Human-Computer Studies* 46, 5 (1997), 551–561.

Paul Fraccaro, Benedict Jones, Jovana Vukovic, Finlay Smith, Christopher Watkins, David Feinberg, Anthony Little, and Lisa Debruine. 2011. Experimental evidence that women speak in a higher voice pitch to men they find attractive. *Journal of Evolutionary Psychology* 9, 1 (2011), 57–67.

Hershey H. Friedman, Michael J. Santeramo, and Anthony Traina. 1978. Correlates of trustworthiness for celebrities. *Journal of the Academy of Marketing Science* 6, 4 (1978), 291–299.

Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. 227–236.

William Griffitt and Russell Veitch. 1974. Preacquaintance attitude similarity and attraction revisited: Ten days in a fall-out shelter. *Sociometry* 37, 2 (1974), 163–173.

Megumi Hosoda, Eugene F. Stone-Romero, and Gwen Coats. 2003. The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology* 56, 2 (2003), 431.

Hyperion. 2000. *Who Wants to be a Millionaire*. Hyperion, New York, NY.

Linda A. Jackson, John E. Hunter, and Carole N. Hodge. 1995. Physical attractiveness and intellectual competence: A meta-analytic review. *Social Psychology Quarterly* 58, 2 (1995), 108–122.

Denise B. Kandel. 1978. Similarity in real-life adolescent friendship pairs. *Journal of Personality and Social Psychology* 36, 3 (1978), 306.

David A. Kenny and William Nasby. 1980. Splitting the reciprocity correlation. *Journal of Personality and Social Psychology* 38, 2 (1980), 249.

Judith H. Langlois, Lori A. Roggman, Rita J. Casey, Jean M. Ritter, Loretta A. Rieser-Danner, and Vivian Y. Jenkins. 1987. Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental Psychology* 23, 3 (1987), 363.

Judith H. Langlois and Cookie White Stephan. 1981. Beauty and the beast: The role of physical attractiveness in the development of peer relations and social behavior. *Developmental Social Psychology: Theory and Research* (1981), Oxford University Press New York, 152–168.

Eun J. Lee. 2009. I like you, but I won't listen to you: Effects of rationality on affective and behavioral responses to computers that flatter. *International Journal of Human Computer Studies* 67, 8 (2009), 628–638. DOI: http://dx.doi.org/10.1016/j.ijhcs.2009.03.003

Jungwon Lee, Jinwoo Kim, and Jae Yun Moon. 2000. What makes internet users visit cyber stores again? key design factors for customer loyalty. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 305–312.

Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter Hoffman. 2001. Effective notification systems depend on user trust. In *Proceedings of Human-Computer Interaction–Interact*. 54–69.

James C. Lester, Sharolyn A. Converse, Susan E. Kahler, S. Todd Barlow, Brian A. Stone, and Ravinder S. Bhogal. 1997. The persona effect: Affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 359–366.

J. David Lewis and Andrew Weigert. 1985. Trust as a social reality. *Social Forces* 63, 4 (1985), 967–985.

Xuan Liu and Yi Xu. 2011. What makes a female voice attractive? *Proceedings of the XVIIth International Congress of Phonetic Sciences.* 2174–1277.

Diane M. Mackie and Leila T. Worth. 1991. Feeling good, but not thinking straight: The impact of positive mood on persuasion. *Emotion and Social Judgments* 23 (1991), 210–219.

Pattie Maes. 1994. Agents that reduce work and information overload. *Communications of the ACM* 37, 7 (1994), 30–40.

Masha Maltz and Joachim Meyer. 2000. Cue utilization in a visually demanding task. In *Proceedings of the Human Factors and Ergonomics Society... Annual Meeting*, Vol. 1. Human Factors and Ergonomics Society, 283.

Ronald Mazzella and Alan Feingold. 1994. The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of mock jurors: A meta-analysis. *Journal of Applied Social Psychology* 24, 15 (1994), 1315–1338.

William J. McGuire. 1969. The nature of attitudes and attitude change. *Handbook of Social Psychology* 3, 2 (1969), 136–314.

Arthur G. Miller. 1970. Role of physical attractiveness in impression formation. *Psychonomic Science* 19, 4 (1970), 241–243.

Judson Mills and Elliot Aronson. 1965. Opinion change as a function of the communicator's attractiveness and desire to influence. *Journal of Personality and Social Psychology* 1, 2 (1965), 173.

Youngme Moon. 1998. *Intimate Self-Disclosure Exchanges: Using Computers to Build Reciprocal Relationships with Consumers*. Division of Research, Harvard Business School.

John Morkes, Hadyn K. Kernal, and Clifford Nass. 1998. Humor in task-oriented computer-mediated communication and human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 215–216.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1 (January 2000), 81–103. DOI: http://dx.doi.org/10.1111/0022-4537.00153

Steve A. Nida and John E. Williams. 1977. Sex-stereotyped traits, physical attractiveness, and interpersonal attraction. *Psychological Reports* 41, 3f (1977), 1311–1322.

Lena A. Nordholm. 1980. Beautiful patients are good patients: Evidence for the physical attractiveness stereotype in first impressions of patients. *Social Science & Medicine Part A: Medical Psychology & Medical Sociology* 14, 1 (1980), 81–83.

Pamela M. Pallett, Stephen Link, and Kang Lee. 2010. New "golden" ratios for facial beauty. *Vision Research* 50, 2 (2010), 149–154. DOI: http://dx.doi.org/10.1016/j.visres.2009.11.003

Gordon L. Patzer. 1983. Source credibility as a function of communicator physical attractiveness. *Journal of Business Research* 11, 2 (1983), 229–241.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* CLSI and Cambridge University Press, Cambridge.

John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95.

Curtis A. Samuels and Richard Ewy. 1985. Aesthetic perception of faces during infancy. *British Journal of Developmental Psychology* 3, 3 (1985), 221–228.

Norbert Schwarz, Herbert Bless, and Gerd Bohner. 1991. Mood and persuasion: Affective states influence the processing of persuasive communications. *Advances in Experimental Social Psychology* 24 (1991), 161–199.

Jean H. Searcy and James C. Bartlett. 1996. Inversion and processing of component and spatial–relational information in faces. *Journal of Experimental Psychology: Human Perception and Performance* 22, 4 (1996), 904.

Harold Sigall and Nancy Ostrove. 1975. Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. *Journal of Personality and Social Psychology* 31, 3 (1975), 410.

Harold Sigall, Richard Page, and Ann C Brown. 1969. The effects of physical attractiveness and evaluation on effort expenditure and work output. In *Proceedings of the Annual Convention of the American Psychological Association*.

John E. Stewart. 1980. Defendant's attractiveness as a factor in the outcome of criminal trials: An observational study. *Journal of Applied Social Psychology* 10, 4 (1980), 348–361.

John E. Stewart. 1985. Appearance and punishment: The attraction-leniency effect in the courtroom. *Journal of Social Psychology* 125, 3 (1985), 373–378.

Lijuan Wang, Wei Han, Frank K. Soong, and Qiang Huo. 2011. Text driven 3D photo-realistic talking head. In *INTERSPEECH*. 3307–3308.

Andrew J. Weigert. 1981. *Sociology of Everyday Life*. Longman.

Sandy Wood and Kara Kovalchik. 2012. *Mental Floss Trivia: Brisk Refreshing Facts Without the Ice Cream*. Puzzlewright, New York, NY.

Louise C. Young and Gerald S. Albaum. 2002. *Developing a Measure of Trust in Retail Relationships: A Direct Selling Application*. School of Marketing, University of Technology, Sydney.

Miron Zuckerman and Robert E. Driver. 1989. What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior* 13, 2 (1989), 67–82.

Miron Zuckerman, Holley Hodgins, and Kunitate Miyake. 1990. The vocal attractiveness stereotype: Replication and elaboration. *Journal of Nonverbal Behavior* 14, 2 (1990), 97–112.

Received December 2015; revised September 2016; accepted September 2016