Truth Is Beauty: Researching Embodied Conversational Agents

Clifford Nass, Katherine Isbister, and Eun-Ju Lee

G. H. Hardy (1941) argues that the sole criterion for excellent research is that the researcher produces "beauty." While seemingly ineffable and frustratingly imprecise, Hardy instead suggests that creating beauty is straightforward. First, the work must be accurate: erroneous results are useless. Second, one's peers must recognize the work to be interesting, exciting, elegant, and "cool." While this second criterion might seem arbitrary, there is generally good agreement between scholars in a given community about "interesting" work (see Cole and Cole 1973 for a discussion), so one need not survey numerous researchers to ensure research is beautiful; asking a couple is equivalent to asking them all.

With certain caveats, the work in embodied conversational agents (ECA) can make claims to beauty. ECAs are phenomenologically "accurate" to the extent that the agent's outward appearance objectively matches the appearance, language, attitudes and behavior of humans. Thus, questions that address manifestation accuracy include "Does the agent walk like a person walk?" and "Does the agent use language and make grammatical errors the same way a person does?"

An alternative approach to accuracy, generally associated more with the artificial intelligence literature than with the ECA literature, assesses the extent to which the *processes* that produce aspects of the ECA are the same as the processes in humans. For example, "Does the muscle model of the character match how human muscles work?" or

"Does the character generate language using the same production models that humans have?" With current technology, success in one approach can lead to less success in the other. For example, the most phenomenologically human ECAs often are animated, scripted, and informed by visual "tricks" (Thomas and Johnston 1981), while models that incorporate the best understanding of physical and psychological processes often create representations that are ironically not "lifelike." Under either definition, ECAs are becoming increasingly accurate (and hence beautiful), even though they have not met the standard of absolute accuracy on any dimension of humanness.

Despite the incredible diversity of disciplines, problems, and techniques that are brought to bear in creating ECAs, the research community generally agrees on which work is "interesting" or "cool." Somehow, researchers fundamentally agree on what "should" be done and when it is done well, even though one might argue that ECA is a preparadigmatic discipline (Cole and Cole 1973; Levitt and Nass 1989). Hence, if a few colleagues endorse a particular ECA, the researcher can be confident that he or she has passed the second beauty hurdle.

X.1 ECA Research Requires a Special Criterion for Beauty

Unlike the simple two-pronged beauty test appropriate to most other areas of research in engineering, physical science, and social science, researchers in ECA have a third beauty test: Does the ECA satisfy *users* of the technology? While this question seems deceptively similar to the question about researchers' reactions, users differ from researchers in two fundamental ways: (1) variance in the community, and (2) to what the ECA should be compared. In contrast to the homogeneity of perspectives in the research

community, users responding to ECAs exhibit enormous variety in their assessment dimensions (Nass and Mason 1990) and the particular values they assign to these dimensions. Similarly, while all researchers have exposure to the same relevant range of interfaces, both with and without ECAs, users' experiences vary enormously, one of the reasons for the aforementioned heterogeneity in perspective. Thus, a conversation with two or three researchers covers that community; a chat with ten or one hundred times that many users might not provide reliable judgments.

Does this mean that one must abandon practically obtainable definitions of beauty? Happily not. Experimental research (which has its own claims to beauty) provides objective and reliable measures of whether a particular ECA has satisfied users. A detailed discussion of experimental research is obviously beyond the scope of this chapter, but a few general guidelines seem useful.

X.2 Experimental Research and ECAs

The definition and guiding principle of experimental research is as follows: *random assignment to varied conditions*. "Varied conditions" means that one cannot simply show users an ECA and ask questions; instead, responses to a particular ECA must be *compared* to one or more other instantiations of an ECA, an interface without an ECA, or an actual person. This is often emotionally difficult for a designer. After working so hard to build an ECA, the demands of experimentation means that the work is only (at most) half done. "Random assignment" means that neither the user nor the experimenter consciously choose the interfaces that they are exposed to, nor do they choose the order

in which the interfaces are presented (when a given user experiences multiple instantiations).

Why are varied conditions and random assignment important? Unlike many theories in the physical sciences, for which theories provide absolute values and external metrics, virtually all theories concerning ECAs are *relative* and *comparative*. ECA theories include such statements as "Users will like this ECA more than none at all," "Users will better remember statements by ECAs with synthesized speech than ECAs with word balloons," "My ECA will lead to greater efficiency than your ECA," and so forth. The theories do not make claims such as "users' hearts will beat an average of 72 beats per minute when shown this ECA" or "This ECA will lead to an average of 3.6 errors on the task." Unlike absolute statements that only require the assessment of one interface, relative statements require (empirical) comparison.

The comparison ECAs (or an ECA and a non-ECA or human) can be virtually identical, differing on one or two characteristics (as in the experiments described below), or radically different. When an experiment on ECAs presents limited and well-defined differences (e.g., ethnicity, personality of language, synthesized speech vs. recorded speech), it becomes easy to draw highly specific conclusions, at the expense of failing to capture all of the exciting aspects of a particular ECA. On the other hand, gross differences in interfaces—for example, a particular instantiation of an ECA versus no ECA at all, can generate the very broad conclusions that may be appropriate at the early stages of a technology's development but provide less specific help for designers or theorists.

Random assignment, the second aspect of experiments, is necessary for two reasons. First, if the participants in an experiment are assigned to see a particular ECA based on some systematic criteria (e.g., gender, computer experience), one will not know whether the observed differences between conditions are the result of the manipulation or the prior characteristic. Second, the act of choice has a number of psychological consequences, including the conflicting tendencies of postdecision justification (people like the alternative they choose) and buyer's remorse (people like the alternative they *didn't* choose), so that without random experimental assignment, these effects could obfuscate the results of interest.

Having correctly designed an experiment, an ECA researcher must choose the criteria that define "user satisfaction." In the homogenous research community, satisfaction criteria do not have to be stated because they are shared. In the diverse user community, however, the questions "Are you satisfied?" or "Is it cool?" are much too ambiguous to be answered in a valid and reliable way. In each experiment, the experimenter must choose from an enormous list of user attitudes and behaviors: emotional judgments of liking and arousal, general judgments of similarity (which can be used to impute relevant dimensions), assessments of attractiveness, personality, competence, and similarity to humans, behavioral measures of performance, attention, memory, and so forth. Having chosen the criteria, a researcher has created a "beautiful" ECA when the participants in the experiment provide more positive responses to the ECA than to the comparative agent/interface/human on the dimensions that are of interest to the researcher.

X.3 Examples of Experimental Determination of Beauty in ECAs

To illustrate how to perform the third beauty test, the assessment of user satisfaction, we present two experiments. The two studies address, somewhat tongue-in-cheek, an ECA's *appearance*, the conventional approach to beauty.

The two studies focus on opposite extremes of ECA technology. The first study employs a presently unobtainable ECA: a full-motion video representation of a character with perfect language production and understanding, unequivocal human appearance, and so forth; indeed, they are video recordings of a human face. The second study employs what many would argue is a bare minimum ECA: a stick figure character without a face, communicating through text input and word balloon output. The studies explore very different phenomena with very different metrics, but both incorporate the critical criteria for experiments: clearly specified variation in the representation of the ECAs, random assignment as to which people see which character, and clearly defined metrics that indicate whether one ECA is "better" or "worse" than another.

Both studies are based on the Computers Are Social Actors (CASA) paradigm (Nass and Moon n.d.; Reeves and Nass 1996). This paradigm argues that one can take both theories and methods from social psychology and directly apply them to human-technology interaction. These are, in fact, the first studies to explore the CASA paradigm with respect to human-ECA interaction; previous studies addressed social responses to simple text-based interfaces only (see Reeves and Nass 1996 for a review).

In the CASA paradigm, one does not directly ask users whether they are applying social rules to computers or ECAs; they consistently deny that they do. Instead, the paradigm directs one to place users in a situation in which social rules dictate particular

responses, while common sense would suggest different responses. To the extent that individuals apply social rules, even though it is foolish to do so (and they deny doing so), one has evidence for social responses to computers and ECAs (see Nass and Moon n.d.).

X.4 Does the Ethnicity of ECAs Matter?

The first study asked questions about the ethnicity of ECAs (see Lee and Nass 1998 for a more complete discussion of this study). When we meet someone, one of the first things we do is to classify that person as "in-group" or "out-group." This categorization is not always based on a thorough examination of others' beliefs, thoughts, and value systems. Rather, readily observable physical cues such as ethnicity often work as the most salient and strong basis for social categorization (Biernet and Vescio 1993). If one can extend the literature on human-human interaction to human-ECA interaction, as specified by the CASA paradigm, one might expect that users will quickly assess the ethnic identity of an ECA. Having determined the ethnicity of the ECA, the critical question is whether that determination will affects users' attitudes and behaviors. Logically, there should be no effect. Computers do not have ethnicities, and ECAs are not socialized or acculturated into any particular ethnicity. For ECAs, unlike people, ethnicity is essentially arbitrary. Hence, common sense would dictate that ethnicity would be irrelevant to users' responses to ECAs. Conversely, the CASA model would predict the same responses to ethnically identified ECAs as people direct toward ethnically identified humans.

The literature on ethnicity suggests that it does not work monotonically; instead, it operates in conjunction with the ethnicity of the interaction partner. Specifically, individuals assess another's ethnicity primarily to determine whether they are part of the

same group or a different group (Tajfel 1978; Turner 1985). Members of the in-group are more "beautiful" on a number of dimensions (Gerard and Hoyt 1974; Whitehead, Smith, and Eichhorn, 1982). Individuals agree with in-group members more than out-group members (Clark and Maass 1988) and in-group members are perceived as having the same values as the individual (Allen and Wilder 1979). Furthermore, when someone is identified as part of the in-group as opposed to the out-group, he or she is perceived as more socially attractive and better liked (Y. T. Lee 1993; Stephan and Beane 1978), more trustworthy (Clark and Maass 1988), and more competent (Stephan and Beane 1978). If CASA is correct, these effects should obtained for both participants who believe that they are interacting with an ECA as well as those who believe that they are interacting with a person.

Despite the empirical evidence that demonstrates the critical role ethnicity plays in social interaction, the effects of ethnically diverse computer agents have never been explored. Thus, the first question we address in this study is as follows: *Are agents that ethnically match users more "beautiful" than ethnically different ECAs?* Or, put another way, *Does the ethnicity of a computer agent have an effect on users' attitudes and behaviors?* The experiment also addresses a foundational question for the CASA paradigm: *Does the belief that one is interacting with a person (via video conferencing) as opposed to an ECA affect users' reactions?*

X.4.1 Design of the Ethnicity Experiment

To examine these questions, we created an experiment in which participants interacted with a full-motion video of a person; the only difference in the ECA was whether it was

of a similar or different ethnicity than the user. The other varied dimension was whether participants were told they were interacting with a computer agent (HCI condition) or via video conference software with a person in another room (CMC condition). The interactions (described later) were identical for all participants.

To maximize the salience of ethnicity as an identity-defining factor, and because members of the minority tend to identify more strongly with their in-group than do those of the majority (Wilder and Shapiro 1984), individuals from an ethnic minority (40 Korean students born in Korea and with strong ethnic identity) participated in this experiment. To control for the possible effects of gender, only male participants participated in this study (we were unable to obtain enough female students to permit a balanced design). Participants were randomly assigned to one of four conditions in a 2 x 2 design: HCI-in-group, HCI-out-group, CMC-in-group, or CMC-out-group.

Upon arrival, the participant was told either that he would interact with a computer agent that had speech recognition capacity (HCI condition) or with another participant in another room via a video conferencing system (CMC condition). To emphasize that these were not pre-recorded responses (although they were), the participant was asked to choose one of ten different packets, each composed of eight choice-dilemma situations. In fact, all packets were identical, so that every participant went through the same scenarios. Choice-dilemma situations are hypothetical situations in which an individual has to decide what to do between two courses of actions (Kogan and Wallach 1967), one of which has the potential for both greater benefit and greater harm. For example, one of the situations depicted the dilemma of a college football

player who could go for either a risky play that would win or a cautious play that would tie.

After choosing the packet, the participant was instructed to read the situation on the questionnaire silently and then, using the microphone, ask the agent/partner, "Do you think Mr. A (the person in the scenario) should do B (one of the possible choices)?" At this point, one of two Korean (in-group condition) or Caucasian (out-group condition) male confederates popped up on the screen and presented his decision and the arguments in favor of that decision. (We used two different faces to control at least minimally for the fact that every face has unique characteristics that might be more relevant than ethnicity.) After listening to the agent's/partner's decision and arguments (which was prevideotaped, unbeknownst to the participants), the participant answered a paper-and-pencil questionnaire concerning his perception of the interactant's decision, the quality of the arguments, and his own decision. The questionnaire items were based on a ten-point Likert scale. When he was done answering the questions, he went on to the next scenario. This procedure was repeated for the eight different situations.

In order to make it more like a real-time interaction, a couple of tricks were used. The agent/partner asked the participant on one occasion to repeat his question during the interaction; at another point, the agent/partner asked for more time to prepare his arguments. The choice of packets and the request for repetition of the question and for more time for the interaction were very effective in making people believe that the interaction was not preprogrammed. When the interaction was over, participants filled out a final paper-and-pencil questionnaire regarding value congruence (how much the participant perceived agreement between themselves and the other interactant) and source

perception (the participant's assessment of the other interactant).

X.4.2 Measuring the Possible Consequences of ECA Ethnicity We attempted to measure many of the characteristics of in-group/out-group differences noted above. When indices were created, they were very reliable.

Value congruence was computed by summing two self-reported similarity scores asked after all of the choice dilemmas were completed: "How similar were the computer agent's/your partner's decisions to yours?" (decision similarity), and "How similar were the computer agent's/your partner's reasons for its/his decisions to yours?" (reasoning similarity). Both items were responded to on ten-point scales ranging from "not at all similar"(1) to "very similar" (10) (r = .77)**.**

The indices for *social attractiveness* and *trustworthiness* were based on the question "How well does each of these adjectives describe the computer agent/partner you worked with?," which appeared on the final paper-and-pencil questionnaire. Responses were provided on a ten-point Likert scale ranging from "describes very poorly" (1) to "describes very well"(10). The index of social attractiveness was comprised of four items: "likable," "sociable," "pleasant," and "friendly" (Cronbach's A = .88). Trustworthiness was an index comprised of two items: "trustworthy" and "reliable" (r = .65).

The *quality of arguments* (competence) was measured by creating an index based on four adjectives from the final paper-and-pencil questionnaire that described the arguments participants had heard during the interaction: "persuasive," "clever," "analytical," and "creative" (Cronbach's A = .68). To assess *conformity*, we examined

the correlation between the agent's/partner's decision and their own decision across the eight situations for each person.

Our analytical strategy was to compare the two HCI conditions directly, followed by the two CMC conditions. We then determined whether individuals reacted to ethnicity differently in the HCI case than in the CMC case; a significant interaction terms in the 2 x 2 ANOVA would suggest that they do.

X.4.3 Responses to the Ethnicity of ECAs

Consistent with the equivalence of human-ECA interaction and the social psychological literature, participants who worked with the in-group agent believed that it matched their opinions more than did those who interacted with the out-group agent, $t(18) = 2.35$, $p < .05$ (see fig. X.1). Greater perceived in-group value congruence was also evident in the CMC case, $t(18) = 2.47$, $p < .05$. There was no significant interaction between perceived ontology of the interaction partner and group identity of the source, although CMC participants attributed more attitudinal similarity to their partners than HCI participants did to the computer agents, $F(1,37) = 6.54$, $p < .05$.

Fig. X.1 here

Consistent with the idea that in-group agents/participants are more beautiful, the in-group agents, $t(18) = 6.03$, $p < .001$, and the in-group partners, $t(18) = 2.65$, $p < .05$, were perceived to be more socially attractive than their out-group counterparts. There

was no interaction and no main effect for HCI versus CMC with respect to social attractiveness.

In-group agents, $t(18) = 5.77$, $p < .001$, and partners, $t(18) = 2.94$, $p < .01$, were perceived as more trustworthy than their out-group counterparts. There was no interaction, but CMC participants considered their partner to be more trustworthy than did HCI participants, $F(1, 37) = 5.01$, $\underline{p} < .05$. Again consistent with CASA, in-group computer agents were perceived as providing higher quality arguments than out-group agents, $t(18) = 2.48$, $\underline{p} < .05$. There was no effect of group identity in the CMC conditions, $t(18) = 1.31$, $p > .10$, although the results were in the expected direction. There was no interaction and no main effect for HCI versus CMC with respect to argument quality.

Consistent with the expectation that in-group members would obtain greater conformity than out-group members, a higher average correlation existed between the computer agent's decision and participant's own decision for in-group participants, $t(14) = 1.85$, $p < .05$ (see fig. X.2). Similarly, in-group partners in CMC condition elicited more conformity from the participants than did their out-group counterparts, $t(14) = 2.20$, $p < .05$. There was no interaction, but there was a main effect for perceived ontology. People agreed more with the computer agents than with the CMC interaction partners, $F(1, 29) = 2.35$, $\underline{p} < .05$.

Fig. X.2 here

The foregoing results provide convincing evidence that ethnicity of computer agents has significant and consistent effects on user's attitudes and behaviors. In-group participants perceived the computer agents to be more similar to themselves and more socially attractive and trustworthy. Participants also conformed more to the decision of their in-group partner and perceived the agent's arguments to be better. Given that in-group favoritism is more likely to occur when the group identity becomes salient due to intergroup conflict/competition (Taylor and Moriarty 1987; Wagner and Ward 1993), our findings obtained in the absence of intergroup contrast lend strong support to the existence of in-group favoritism in HCI.

Before we address the broader theoretical and design implications of this study, we present a second study that uses a very different kind of ECA but also focuses on whether beauty can only be "skin-deep."

X.5 Personality in ECAs

Whether it is a screenwriting guide or a book about how to make successful animated features, artists seem to agree that developing an appealing "personality" is an important part of creating successful characters. What exactly is personality? Media practitioners have working definitions of personality that they use to try to explain what they do. For instance, Thomas and Johnston (1981) discuss how an animated character's personality consists of characteristic attitudes and actions that people learn to associate with that character, as revealed during the story, through the character's motions and conversations and interaction with other characters. Hoffner and Cantor (1981) say that people use a character's physical appearance, speech characteristics, and behaviors to determine what

the characters' traits are. Field (1994) contends that one develops a character's personality by establishing attitudes and behaviors people come to expect from the character. Laurel (1993) explains that the traditional Aristotelian understanding of dramatic characters is as "bundles of traits, predispositions, and choices that, when taken together, form coherent entities" (60). Judging from these descriptions, character personality seems to be related to predictability in the character's actions and attitudes that people use to understand how the character works within the media they are watching or reading.

This working definition of personality from the arts is corroborated by the understanding of personality within the field of psychology. The opening definition from a standard psychology textbook reads "Personality represents those characteristics of the person that account for consistent patterns of feeling, thinking, and behaving" (Pervin and John 1997, 4). Consistent with artists' working knowledge of why a character's personality matters, psychologists have found that personality is a predictor of many important things about a person. For example, one's personality is related to the kinds of social situations one is comfortable in, how others will choose to interact with one, and a host of other important life activities (Campbell and Hawley 1982; Eysenck and Long 1986). Personality is something that everyday people recognize and discuss about others and that they feel is a valuable piece of information about a person (Pervin and John 1997).

Because of its importance in both traditional media and in psychology, and informed by the CASA paradigm, we wanted to see whether personality was important in interactive character design as well. To understand how computer users would respond to

personality-rich characters in interfaces, we performed an experiment in which the character's appearance and language both presented a particular personality type, although in some cases, the appearance and language suggested conflicting personalities (for complete details of this study, see Isbister and Nass n.d.). Drawing on the CASA paradigm outlined above, we turned to the psychology literature to derive a set of predictions of how people would interact with ECAs that manifested personality in their language and in their appearance.

Which personality trait did we choose to manipulate? Of the many dimensions of personality that trait psychologists have identified, two are particularly important during interaction: extroversion and agreeableness (McCrae and Costa 1989). People quickly assess how extroverted and how friendly a person is, and this affects how they feel about the interaction. Because it is quickly and easily assessed, important to interpersonal interaction, and readily discerned from nonverbal behavior (Gallaher 1992), we selected extroversion as the personality trait that we would examine in our experiment.

X.5.1 Manifesting Personality in ECAs

To manipulate the characters' expression of personality, we were guided by the ways people normally read personality in others. What cues do people use to make assessments about another's extroversion? Confirming artists' intuitions, psychologists have discovered that people use a variety of cues depending upon the situation (Ekman et al. 1980). However, people consistently rely on verbal style and nonverbal cues to guide the determination of personality.

Verbal style includes choice of words and types of sentences and fluidity of speech, as well as how the person refers to another while speaking. For example, an extroverted person might use strong, confident words and phrasing and speak very fluidly, whereas an introverted person might be more hesitant in speech and use less direct and confident phrasing (Jung 1971; Nass et al. 1995).

Nonverbal cues include posture as well as the way that the person moves his or her body when interacting with others. For example, an extroverted person is more likely to use gestures that are expansive and may approach more readily, whereas an introvert may keep limbs close to the body and avoid approaching (Gallaher 1992).

In this study, we independently manipulated both verbal and nonverbal cues to convey the interactive characters' extroversion or introversion. No one has yet demonstrated experimentally that people will read personality cues in an interactive character in the same way that they will read them in people, although there is evidence from television research that people apply the same personality traits to TV characters as they do to other people (Hoffner and Cantor 1981; Reeves and Greenberg 1977). We predicted that people would successfully label introverted and extroverted verbal and nonverbal cues from interactive characters, just as they identified the verbal cues of dominance-submissiveness in previous research in human-computer interaction (Moon and Nass 1996; Nass et al. 1995).

X.5.2 Inconsistent Personalities in ECAs

Because people judge a person's personality from a host of different cues, the possibility of conflicting cues arises: What happens if a person is suggesting one personality with the

way that he or she speaks, and an entirely different personality with the way that he or she moves? It is clear that people prefer to engage with others whom they can label consistently. Consistency in others allows people to predict what will happen when they engage with them (Fiske and Taylor 1991), makes it easier to remember a person accurately (Cantor and Mischel 1979), and generally lightens cognitive load (Fiske and Taylor 1991). In addition, studies that looked at how people detect deception have found that people turn to nonverbal cues to see if they are inconsistent with the verbal ones. This suggests that discrepancies among cues is a big problem in others (Ekman and Friesen 1974). Cassell, McNeill, and McCullough (1998) note that even though people may not be aware of mismatches between verbal and gestural cues, they will still make combined use of these cues to form an integrated understanding of what was said. Literature also suggests that adults use mismatched verbal and gestural cues in children to help determine the child's knowledge state (Goldin-Meadow, Alibali, and Church 1993).

Character consistency is of great concern to traditional character crafters. Guidelines for creating characters often include a caveat that everything a character does should convey the same general impression about the character to the viewer (Field 1994; Thomas and Johnston 1981). These caveats are needed because it is easy for inconsistencies to creep in during the development process. This is especially the case for complex character creation involving a large team of people, as is often found in the development of interactive characters.

What happens when a person is confronted with inconsistent cues from an on-screen character? From the psychological literature and CASA, one can predict that the person will dislike inconsistent cues and thus will like the character less. This would

indeed be a problem that character designers should avoid. However, inconsistency might not work in the same way for characters as for actual people. Perhaps people average the two sets of conflicting cues to arrive at an overall impression of the character. If so, it might be better to design a character with mixed cues, to ensure that all users, regardless of personality, would be at least partially satisfied with the character, on the assumption that all individuals have a preference for one personality type over another. We sought to determine which of these hypotheses would hold true for interactive characters. In sum, the study had two goals: (1) to determine whether users can recognize personality in both verbal and nonverbal cues of interactive characters; and (2) to determine whether inconsistent characters are universally disliked (consistency theory) or perceived as neutral (averaging theory)—that is, whether consistent characters are more beautiful than inconsistent characters.

One complication in addressing these questions is that studies in interpersonal psychology have shown that people tend to prefer others based on the match or mismatch to their own personality. Two conflicting hypotheses exist in this literature: the similarity-attraction hypothesis and the complementarity principle. Similarity-attraction holds that people prefer those with personalities similar to their own (Blankenship et al. 1984; Byrne 1969). Complementarity, conversely, holds that people will tend to behave in complementary ways in their interpersonal interactions and will seek out others who elicit complementary behavior from them (Leary 1957; Sullivan 1953). Both similarity-attraction and complementarity have significant experimental confirmation in the psychological literature (see Isbister 1998). Rather than attempt to resolve these ambiguous results, we simply ensured that an equal number of introverts and extroverts

assessed both of the consistent (introverted and extroverted) ECAs as well as the two mixed ECA (introverted verbal with extroverted nonverbal, or vice versa). This balancing ensured that the effects of user/ECA match or mismatch would be washed out.

X.5.3 Design of the Personality Experiment

To address our two core questions adequately, we created a balanced, between-participants design in which introverted or extroverted individuals were randomly assigned to one of four conditions: (1) wholly matching character (verbal and nonverbal cues were consistent and matches the user); (2) wholly mismatched character (verbal and nonverbal cues were consistent but were opposite the user); (3) matching verbal and mismatching nonverbal; and (4) mismatching verbal, matching nonverbal. Examining the two main effects (verbal personality and nonverbal personality) allowed us to address our first goal of recognizing personality. A comparison of conditions (1) and (2) versus conditions (3) and (4) answered whether consistency was "beautiful" or not, which was our second goal.

Our participants were students from two West Coast universities who had been asked to be in various studies as part of their coursework. There were forty students in all, with students from both schools balanced evenly across the conditions.

We assigned students to conditions in our study based on their own introversion/extroversion (this is not a violation of the principle of random assignment, as equal numbers of introverts and extroverts were randomly assigned to experience each type of character). A few weeks before the experiment ran, we had every student in both classes complete a section of the Myers-Briggs personality inventory (see Murray 1990

for a review) as well as a portion of the Wiggins personality adjective set (Wiggins 1979), as part of a packet of questionnaires administered to the entire class. Students who fell above the class median on the Myers-Briggs (higher than 4 out of a possible score of 9) and on the Wiggins introversion scale (higher than 27 out of a possible score of 54) were classified as introverted; students who fell below the class median on the Myers-Briggs (lower than 4) and above it on the Wiggins extroversion scale (higher than 38 out of a possible score of 54) were classified as extroverted.

Twenty students from the extroverted group and twenty students from the introverted group were asked to participate. They were simply told that they would be participating in a study examining how people work with computer characters to accomplish a task. Everyone signed informed consent forms, was debriefed at the end of the experimental session, and was awarded class credit for participating in the study.

When they arrived, each participant was first asked to complete the Desert Survival Problem (DSP) (Lafferty and Eady 1974) using pencil and paper. The DSP is a problem-solving task that has been used in a variety of studies involving interpersonal interaction and human-computer interaction (see Reeves and Nass 1996). It asks participants to rank a series of twelve items (compress kit, book, raincoat, flashlight, vodka, parachute, water, mirror, jackknife, magnetic compass, salt tablets, and air map), according to their assessment of the items' importance in a desert survival situation.

After finishing this initial ranking, the participant was introduced to an on-screen computer character. The experimenter explained that the participant would get to exchange information about each of the twelve desert survival items with the computer

character. In addition, after completing the interaction with the character, the participant would have the opportunity to change his or her initial ranking of all the items.

The on-screen character was in a format similar to comic books: the figure was a still image in each turn, with a word balloon with text in it that represented its own "voice." The character stayed in one place on each screen, creating the impression that one was working through an interaction with a comic-book-like character. The character had no face and was a simple stick figure, in contrast to the rich video images in the previous experiment. Participants typed their own words into their own text word balloon, which also stayed on screen in the same place throughout the interaction (see fig. X.3).

Fig. X.3 here

After a single practice round (discussing the pistol, an item not on the actual list), the experimenter left the room. The participant was left alone to exchange information with the computer character about each of the twelve desert survival items. After the interaction was complete, the student made a final ranking of the desert survival items, on paper. Then, the student was given a questionnaire to fill out. This questionnaire asked for his or her assessment of both the computer character and the interaction itself. After completing the questionnaire, the student was debriefed, thanked, and asked not to discuss the experiment with other classmates until the study was completed.

X.5.3.1 Creating Introversion and Extroversion *Verbal* extroversion or introversion was operationalized (implemented) by manipulating the phrasing of the text displayed in the character's word balloons during the interaction. The extroverted computer character used strong and friendly language expressed in the form of confident assertions. This manipulation is consistent with the theoretical definition of extroversion as being the tendency to be assertive, outgoing, and friendly. The introverted computer character used weaker language expressed in the form of questions and suggestions. This manipulation is consistent with the theoretical definition of introversion as behavior that indicates less ease in socializing and less assertiveness.

For example, the introverted computer character would display the following text: "What about maybe rating the pistol a bit higher? It seems like by the end of the second day, speech may be seriously impaired. Perhaps the pistol could be used as a signaling device?" In contrast, the extroverted character would display the following text: "Friend, I'd say the pistol should definitely be rated higher. By the end of the second day, speech will be impaired and the pistol will be an important signaling device." All text for this manipulation was pretested by individuals who did not participate in this experiment, using a web form. Pretest participants were randomly assigned to read one of two sets of statements and to rate the person who made the statements on the same extroversion/introversion scales that we used in the study itself. We tried to control for undesirable personality trait manipulations. To do this, we also asked those who filled out the form to rank the speech giver on adjectives representing undesirable traits ("sly," "conceited," "big-headed").

Nonverbal extroversion or introversion was operationalized by manipulating the postures of the computer characters. The extroverted character body had poses with its limbs spread wide from its body, and some postures made the character seem to have moved closer to the participant (see fig. X.4). This is consistent with the literature on nonverbal cues of extroversion that indicate that extroverts tend to make wider movements and to approach others more freely in space. The introverted character body had poses with its limbs closer in to its body and did not ever appear to approach the participant. This is consistent with the literature on nonverbal cues of introversion that indicate that introverts tend to keep their limbs closer to their bodies, gesture less freely, and avoid approaching others in space. The character itself was a simple stick figure, which allowed us to avoid possible effects of other cues of personality and personal qualities that arise from things like age, clothing, or gender. As with the verbal cues, the nonverbal cues were pretested to confirm that they were being read properly.

Fig. X.4 here

The fundamental information conveyed by the computer character was *not* manipulated; that is, in all four conditions, the computer character conveyed the same type and amount of information about the items being discussed in the task. Only the *style* of communication was manipulated. Moreover, all responses were preprogrammed. No natural language processing or artificial intelligence was employed. To create a smooth interaction, the character always went first in discussing an item, then the participant responded with his or her own information about an item.

X.5.3.2 Measuring the Possible Consequences of ECA Personality As in the previous study, the dependent variables were measured using a paper-and-pencil questionnaire.

The first set of questions asked, “For each word below, please indicate how well it describes your interaction with the character on the computer. Note that you are evaluating the actual interaction, not the character itself.” This was followed by a list of adjectives (e.g., “fun,” “interesting,” “useful”), each of which had a nine-point Likert scale that ranged from “describes very poorly” to “describes very well.”

The second set of questions, which also used a nine-point Likert scale, were aimed at allowing participants to rank their satisfaction with the character and its perceived value.

The third set of questions asked, “For each word below, please indicate how well it describes the character that you just worked with. Note that you are evaluating the character now, NOT the interaction.” This was followed by a list of adjectives (e.g., “assertive,” “friendly,” “bashful”), each of which had a nine-point Likert scale that ranged from “describes very poorly” to “describes very well.” This list of adjectives included all those used in the Wiggins introversion and extroversion scales. Participants then rated the character’s *body language* on the Wiggins scales, then its *verbal style* on these same measures.

Based on factor analysis, four indices were created from the questionnaire items. All indices were reliable. *Fun* was an index of four adjectives used to characterize the interaction: enjoyable, exciting, fun, and satisfying (A = 0.90). *Liking* was an index of two items**:** “Would you enjoy working with this character in another experiment?” and

"How much did you like this character?" (A = 0.82). *Usefulness of the Interaction* was an index of two items used to characterize the interaction: helpful and useful (A = 0.91). *Usefulness of the Character* was an index comprised of three questions: "How much did the character improve your ranking of the items?," "How much did you learn from interacting with this character?" and "How helpful did you find this character?" (A = .89).

To assess the perception of the character's personality, we created an index of the Wiggins introversion and extroversion scales for the verbal and nonverbal cues, respectively, to reflect a general extroverted versus introverted assessment (hereafter referred to as "extroverted").

X.5.2 Responses to the Personality of ECAs

Our first goal was to find out if people would be able to identify both verbal and nonverbal personality cues. Consistent with previous research, participants accurately identified the extroverted language as significantly more extroverted than the introverted language, $F(1,38) = 5.26$, $p < .05$ (see fig. X.5). (All figures have standardized the indices to reflect nine-point Likert scales.) Consistent with the power of nonverbal cues, the extroverted postures were perceived as significantly more extroverted than the introverted postures, $F(1,38) = 8.90$, $p < .01$, even though the characters were faceless stick figures.

Fig. X.5 here

Our second goal was to better understand the impact of inconsistent verbal and nonverbal cues. If individuals adopt a holistic approach toward the character and are disturbed by inconsistency, we should see significant interactions between the verbal and nonverbal cues. If, however, individuals assess verbal and nonverbal cues independently, no interactions should occur, and main effects should occur only to the extent that similarity-attraction and complementarity have differential effects. Supporting the idea that consistency is important (the CASA prediction), text and body cues showed consistent interactions for all four indices, giving support to the whole-impression model of how we read mixed cues.

People liked the ECA more when it was consistent than when it was inconsistent, reflected in a significant interaction, $F(1,39) = 4.21$, $p < .05$. There were no significant main effects (see fig. X.6). The interaction was also perceived to be more useful when the ECA was consistent, $F(1,39) = 6.87$, $p < .02$. There were no main effects.

Fig. X.6 here

The ECAs with the consistent personality were also more fun to interact with, $F(1,39) = 3.50$, $p < .07$. Consistent with similarity-attraction, individuals had more fun with the character whose nonverbal cues matched their own personality, $F(1,39) = 5.47$, $p < .03$. Finally, the consistent character was perceived as significantly more useful than the inconsistent character, even though the content was identical, $F(1,39) = 3.29$, $p < .08$. There were no main effects for text or body.

Another type of confirmation for the superiority of consistent characters comes from an examination of change in the participants' rankings of the items in the desert survival item list. In a comparison of initial rankings to final rankings, a significantly larger change in rankings was found in the consistent character conditions, $F(1, 39) = 7.9$, $p < .01$, suggesting that the information from the matched character had a greater effect on participants (see fig. X.7). We then performed an analysis looking at what the direction of the change in rankings was in relation to the character partner and found that participants with a consistent character partner changed their answers much more toward their character partner's answers than those with an inconsistent partner, $M = 14.42$ average change closer to partner after interaction versus $M = 9.75$, $F(1,39) = 11.44$, $p < .05$.

Fig. X.7 here

This study gives additional support to the growing body of evidence that people apply the same interpretive strategies to interaction with ECAs as they do to interaction with other people (Reeves and Nass 1996). In this study, people labeled postures and verbal styles in interactive characters the same way they would label postures and verbal styles in other people. And, just as is the case in interaction with other people, participants preferred consistency in the characters they interacted with, and they used all the cues from the characters to form an overall impression by which the character was judged.

X.6 Summary and Discussion

This chapter has urged that appropriately designed assessments of user satisfaction, based on experimental methodology, are a critical component of creating beautiful research. Fortunately, a good experiment does more than give one a pat on the back for doing a research job well; it can inform general principles of both theory and design. As an example of the power of the methodology, we briefly highlight a few of the conclusions that can be derived from our studies.

X.6.1 Research Contributions

The two studies presented here show that people apply social rules and expectations to ECAs, even when doing so is not logical. In the first study, participants responded to ECAs as if their ethnicity had meaning beyond an arbitrary representation; the same words meant different things when coming from ECA that was similar as opposed to different. In the second study, the posture of a stick figure had a significant influence on individuals' assessments of the content and the interaction, even though the character was in no way the "source" of the information. Individuals mindlessly (Nass and Moon n.d.) applied social rules to ECAs, even though they were experienced computer users (though they did not have significant experience with ECAs) who knew that these rules were not logical. Thus, ECAs are clearly social actors.

Each study provides its own unique extension to the CASA paradigm. The ethnicity study is one of the first to directly compare human-computer interaction with computer-mediated communication. Having determined that people treat computers in a social manner, the next step must be to examine *how socially* people respond to

computers. For example, people feel good when they are flattered by a computer (Fogg and Nass 1997) but possibly not as much as when they are flattered by a person. Similarly, the gender stereotypes people unconsciously apply to computers (Nass, Moon, and Green 1997) might not be as strong as those they have about people. In other words, CASA had previously replicated the *pattern* of social rules that govern human-human interaction in the context of HCI but had not tested the *degree* of socialness in people's response to computers (Morkes, Kernal, and Nass n.d. and E.-J. Lee 1999 are exceptions). By juxtaposing HCI and (perceived) CMC, we can now address the "degree" question.

The results provide strong evidence for surprising similarities in the way people respond to ethnicity in ECAs and in humans. Critically, there were no significant interactions between the (seeming) ontology of the source and the source's ethnic identity, suggesting that the processing of ethnicity is similar in the two cases. Thus, this study demonstrates that when one taps into a basic social category (e.g., ethnicity), that category overrides any skepticism the user may have of the relevance of social categories to ECAs as compared to actual people.

The personality study demonstrates that even the most ersatz representation of a person is enough to encourage the user to bring to bear a subtle and complex apparatus that can assign personality to nonverbal cues. In previous CASA studies, the interface was plain text. For users who were experienced with e-mail (virtually all of the participants), this interaction felt no different than an interaction with an actual person. However, the stick figures employed in the present experiment were obviously not human and only moved between screens; indeed, they seemed more like a piece of wood

than a person. Yet individuals used their expectation about personality consistency to interpret the meaning of the ECA's words. This provides additional evidence that there is no "on-off" switch in the brain that allows one to process media differently than real people (Reeves and Nass 1996); if there were, the clearly nonhuman representations would surely have tripped it.

X.6.2 Design Contributions

Almost everyone involved in research on ECAs is motivated by the desire to improve interface design. The studies presented here both have numerous implications for design; we will touch on only a few of these implications (for a complete discussion, see Lee and Nass 1998 and Isbister and Nass n.d.).

Perhaps the most general and important take-away from these studies is that "beauty" matters; appearance is a critical component of how people access ECAs. It is perhaps not surprising that individuals prefer to look at or even interact with ECAs that are more "attractive." What is more compelling is that appearance influenced more cognitive assessments, even when the content was identical. ECAs that shared the ethnicity of the user were perceived as more competent; ECAs that presented consistent personality cues were perceived as more useful. Thus, perceptions of seemingly objective criteria, such as intelligence, can be influenced by attractiveness.

These direct assessments of intelligence had indirect effects as well. In both studies, the beautiful characters were more convincing and obtained greater compliance from their users. Thus, designers of interfaces that involve persuasion (Fogg 1998), from

an e-commerce web site to medical advice software, should be concerned with whether their ECAs are attractive or not.

Current research also suggests that ECA design must be a highly coordinated activity. Because of the complexity of creating a multimodal ECA, the task is often divided into small groups; integration occurs at a fairly late stage. These results suggest that coordination across functional units is absolutely critical. Each aspect of the ECA should be as similar as possible to the user groups that will be interacting with the agent, requiring interaction with the marketing efforts, and each aspect of the ECA should be consistent with each other aspect, requiring coordinated development efforts in crafting consistent interactive characters.

What should one do when the goals of similarity clash with the goals of consistency, as when there is a varied user base and only one ECA? It is important to note that the desires for consistency were obtained regardless of the characters' similarity to the user. Thus, it may be more important that the character is consistent than that it complement or match the individual using the software. While early industry research and subsequent software (e.g., Microsoft BOB), focused on matching the software user's personality with the character's, given the conflicting picture about similarity versus complementarity, it may in fact be more important that the character sends a clear message about its personality, than that it matches the user.

Overall, these studies are an important step toward a more comprehensive picture of how humans use their person perception skills in interpreting and evaluating interactive onscreen characters. As the use of these characters in software increases, it is essential that designers take into account both how person perception works in human-

human interaction, and how these skills play out in responses to ECAs.

X.6.3 Final Thoughts on Beautiful ECAs

We began this chapter by establishing three criteria for beautiful research: (1) create ECAs that accurately mirror humans, (2) obtain the admiration of one's peers, and (3) demonstrate that the ECA satisfies users. As the first two criteria have long been goals of the ECA community, this chapter focused on experimental research as the means for objectively and reliably assessing user satisfaction. The principles of varied conditions, random assignment, and clear specification of the satisfaction criteria were shown to be the first steps in ensuring that the creation of ECAs meets the third standard for beauty.

The reader is likely thinking that if that is all there was to experimental research, the ECA research community would have adopted the technique long ago. There is some truth to this concern. There are, of course, numerous other issues in designing experimental research for judging the beauty of ECA research: the number and choice of dimensions that are presented to the participants, and the number and range of values along each dimension; the number of participants in each condition; whether participants will be exposed to all of the conditions (within-participants design) or only a subset of the conditions (between-participants design); the appropriate number and range of effects (dependent variables) to be obtained; the ability to draw conclusions form null results; the ability to generalize from one or more experiment; and so forth. With that said, if researchers simply created more than one stimulus, randomly assigned the presentation of the stimuli, and provided clear metrics for assessments, we would have both beautiful research and beautiful agents, and that, after all, is the justification for our field.

Note

All authors participated equally in writing this chapter. The ethnicity study was executed by Lee and Nass (1998). The personality study was executed by Isbister and Nass (n.d.).

References

Allen, V. L., and D. A. Wilder. 1979. Group categorization and attribution of belief similarity. *Small Group Behavior* 10:73–80.

Biernet, M., and T. K. Vescio. 1993. Categorization and stereotyping: Effects of group context on memory and social judgment. *Journal of Experimental Social Psychology* 29:166–202.

Blankenship, V., S. M. Hnat, T. G. Hess, and D. R. Brown. 1984. Reciprocal interaction and similarity of personality attributes. *Journal of Social and Personal Relationships* 1:415–432.

Byrne, D. 1969. Attitudes and attraction. In L. Berkowitz, ed., *Advances in experimental social psychology 4.* Orlando, Fla.: Academic Press.

Campbell, J. B., and C. W. Hawley. 1982. Study habits and Eysenck's theory of extroversion-introversion. *Journal of Research in Personality* 16:139–146.

Cantor, N., and W. Mischel. 1979. Prototypes in person perception. *Advances in Experimental Social Psychology*, 12:3–52.

Cassell, J., D. McNeill, and K. E. McCullough. 1998. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition* 6(2):1-34.

Clark, R. D., and A. Maass. 1988. The role of social categorization and perceived source credibility in minority influence. *European Journal of Social Psychology* 18:381–394.

Cole, J. R., and S. Cole. 1973. *Social stratification in science*. Chicago: University of Chicago Press.

Ekman, P., and W. V. Friesen. 1974. Detecting deception from the body or face. *Journal of Personality and Social Psychology* 29:288–298.

Ekman, P., W. V. Friesen, M. O'Sullivan, and K. Scherer. 1980. Relative importance of face, body, and speech in judgments of personality and affect. *Journal of Personality and Social Psychology* 38(2):270–277.

Eysenck, S. B. G., and F. Y. Long. 1986. A cross-cultural comparison of personality in adults and children: Singapore and England. *Journal of Personality and Social*

Psychology 50:124–130.

Field, S. 1994. *Screenplay: The foundations of screenwriting*. New York: Bantam Doubleday Dell.

Fiske, S. T., and S. E. Taylor. 1991. *Social cognition*. New York: McGraw-Hill.

Fogg, B. J. 1998. Persuasive computers: Perspectives and research directions. In *Proceedings of the CHI98 Conference of the ACM/SIGCHI*, 225–232. New York: ACM Press

Fogg, B. J., and C. Nass. 1997. Silicon sycophants: Effects of computers that flatter. *International Journal of Human-Computer Studies* 46:551–561.

Gallaher, P. E. 1992. Individual differences in nonverbal behavior: Dimensions of style. *Journal of Personality and Social Psychology* 63(1):133–145.

Gerard, H. B., and M. F. Hoyt. 1974. Distinctiveness of social categorization and attitude toward in-group members. *Journal of Personality and Social Psychology* 29:836–842.

Goldin-Meadow, S., M. Alibali, and R. B. Church. 1993. Transitions in concept acquisition: Using the hands to read the mind. *Psychological Review* 100(2):279–297.

Hardy, G. H. 1941. *A mathematician's apology.* Cambridge: Cambridge University Press.

Hoffner, C., and J. Cantor. 1981. Perceiving and responding to mass media characters. In J. Bryant and D. Zillmann, eds., *Responding to the screen: Reception and reaction processes.* Hillsdale, N.J.: Erlbaum.

Isbister, K. 1998. Reading personality in onscreen interactive characters: An examination of social psychological principles of consistency, personality match, and situational attribution applied to interaction with characters. Ph.D. diss., Communication Department, Stanford University, Stanford, California.

Isbister, K., and C. Nass. N.d. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies.* Forthcoming.

Jung, C. G. 1971. *Psychological types.* Princeton: Princeton University Press.

Kogan, N., and M. A. Wallach. 1967. Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology* 3:75–84.

Lafferty, J. C., and P. M. Eady. 1974. *The desert survival problem.* Plymouth, Mich.: Experimental Learning Methods.

Laurel, B. 1993. *Computers as theater*. Reading, Mass.: Addison-Wesley.

Leary, T. F. 1957. *Interpersonal diagnosis of personality*. New York: Ronald Press.

Lee, E-J. 1999. *Effects of number, ontology, and representation of influencing agents on public compliance and private conformity*. Ph.D. diss., Stanford University, Stanford, California.

Lee, E.-J., and C. Nass. 1998. Does the ethnicity of a computer agent matter? An experimental comparison of human-computer interaction and computer-mediated communication. In *Proceedings of the Workshop on Embedded Conversational Characters Conference* (Lake Tahoe, Calif.).

Lee, Y. T. 1993. In-group preference and homogeneity among African American and Chinese American students. *Journal of Social Psychology* 133:225–235.

Levitt, B., and C. Nass. 1989. The lid on the garbage can: Institutional constraints on decision making in the textbook publishing industry. *Administrative Science Quarterly* 34(2):190–207.

McCrae, R. R., and P. T. Costa Jr. 1989. The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology* 56:586–595.