Psychological Science OnlineFirst, published on November 5, 2012 as doi:10.1177/0956797612448793

Research Article

Psychological Science
XX(X) 1–8

DOI: 10.1177/0956797612448793
http://pss.sagepub.com

Detecting the Trustworthiness of Novel Partners in Economic Exchange

David DeSteno[1], Cynthia Breazeal[2], Robert H. Frank[3], David Pizarro[4], Jolie Baumann[1], Leah Dickens[1], and Jin Joo Lee[2]

[1]Department of Psychology, Northeastern University; [2]Media Lab, Massachusetts Institute of Technology; [3]Johnson Graduate School of Management, Cornell University; and [4]Department of Psychology, Cornell University

Abstract

Because trusting strangers can entail high risk, an ability to infer a potential partner's trustworthiness would be highly advantageous. To date, however, little evidence indicates that humans are able to accurately assess the cooperative intentions of novel partners by using nonverbal signals. In two studies involving human-human and human-robot interactions, we found that accuracy in judging the trustworthiness of novel partners is heightened through exposure to nonverbal cues and identified a specific set of cues that are predictive of economic behavior. Employing the precision offered by robotics technology to model and control humanlike movements, we demonstrated not only that experimental manipulation of the identified cues directly affects perceptions of trustworthiness and subsequent exchange behavior, but also that the human mind will utilize such cues to ascribe social intentions to technological entities.

Keywords

decision making, cooperation, social cognition

Received 1/2/12; Revision accepted 4/11/12

People must often decide whether to trust new potential partners in the absence of reliable information about their past behavior (Axelrod, 1984; Delton, Krasnow, Cosmides, & Tooby, 2011; Frank, 1988). Although forming an entirely new exchange relationship can be extremely advantageous, a decision to trust also entails the risk of substantial loss if one's partner acts in an untrustworthy manner. Consequently, any capacity that enhances accuracy in detecting the trustworthiness of other people would offer a significant competitive advantage.

In the absence of reliable information about an individual's reputation, nonverbal cues may serve as one possible source of information about his or her likely actions. Indeed, ample evidence indicates that humans regularly use specific cues, often without conscious awareness, to infer the motivations of others with some level of accuracy (Ambady & Weisbuch, 2010; Knapp & Hall, 2010). To date, however, the nature of the cues that might predict trustworthy or untrustworthy behavior, or even whether such cues exist, remains unclear. Yet, for cooperation to occur to the degree it does in humans, it appears theoretically necessary that people have access to information related to reliable, albeit imperfect, signals of trustworthy intent of potential partners or to the likelihood of subsequent encounters with them (Delton et al., 2011; Frank, 1988).[1] If people lack access to trust-relevant signals in situations in which reputational information is absent (e.g., interactions with strangers) and in which the likelihood of repeated interactions is low, it is likely that the advantages of acting opportunistically would reduce cooperation substantially.

Given the theoretical import and adaptive advantages of an ability to assess trustworthiness, the search for trust-relevant signals has long occupied the attention of scholars from many fields (e.g., psychology, behavioral economics, evolutionary biology). Researchers have looked in vain for a single dynamic "golden cue" that predicts whether a person can actually be trusted (Ambady & Weisbuch, 2010; Knapp & Hall, 2010). In a similar vein, researchers have also looked for and identified certain markers that affect judgments of trustworthiness of static faces (Todorov, Baron, & Oosterhof, 2008); however, there is little reliable evidence linking such markers to people's actual behaviors (Todorov, 2008).[2]

Corresponding Author:
David DeSteno, Department of Psychology, Northeastern University, Boston, MA 02115
E-mail: d.desteno@gmail.com

We believe that past difficulties in identifying trust-relevant signals may stem from attempts to look at cues individually or in isolation from social interaction. We suspect that if reliable trust-related signals are to be found, they will likely emerge dynamically and be processed intuitively within the context of interpersonal situations between individuals who are unfamiliar with one another. Given that each member of such a dyad is attempting to assess the intentions of an unfamiliar other, signals will likely be subtle and unfold over time as each person evaluates his or her potential partner. As Frank (1988) noted, interpreting signals of cooperation must either be costly or entail some level of uncertainty, because a costless and perfectly reliable signal would have long since relegated opportunistic individuals to extinction.

Given the increasing volume of work suggesting that the interpretation of nonverbal cues is highly context dependent (Ambady & Weisbuch, 2010; Barrett, Mesquita, & Gendron, 2011), we expect that no single cue will possess substantial trust-related predictive power on its own. As Keltner and his colleagues have shown, nonverbal signals of complex social states, such as embarrassment, are often composed of a set of cues (e.g., Keltner & Buswell, 1997). Indeed, sets, by their nature, can be more informative than any of their individual components because of their ability to resolve ambiguities inherent in the interpretation of single cues (cf. Hall, Coats, & Smith Lebeau, 2005). Accordingly, we expect any trust-relevant signal to be composed of a set of cues that are emitted in close temporal proximity and that, when taken together, convey reliable information about an individual's intentions.

To examine if and how individuals can assess the likelihood that a novel partner will cooperate, we designed a two-phase strategy. The goal of the first phase was to demonstrate that exposure to nonverbal cues increases accuracy in assessing trustworthiness and to identify a set of cues that are reliably predictive of trust-relevant behavior. The goal of the second phase was to manipulate the expression of the candidate cues with exacting precision in order to assess their causal impact on subsequent decisions to trust a partner.

Experiment 1

Experiment 1 constituted the exploratory phase of this project. The primary goal was to identify a set of nonverbal cues that constitute a signal related to the trustworthiness of a novel partner. To accomplish this goal, we constructed a paradigm in which individuals would interact with a previously unknown other in a "get to know you" conversation (either face-to-face or over a Web-based chat) and then play an economic game with this individual that pitted self-interest against joint interest.

Our strategy in this phase was quite straightforward. First, if information about the trustworthiness of another person is conveyed through nonverbal cues, then accuracy in judging the cooperative intent of a partner should be greater when an interaction occurs face-to-face, as opposed to over a Web-based chat in which only semantic information is available. Second, if such nonverbal information exists, one should be able to identify a candidate cue set by linking expressions of specific cues to actual economic behavior. Identification of such a signal would, of course, be an initial step, with final confirmation of a cue set requiring validation through experimental manipulation (which we undertook in Experiment 2).

Method

Participants. Eighty-six individuals (34 male, 52 female) from the undergraduate participant pool at Northeastern University agreed to take part in the experiment. They were assigned randomly to 43 dyads, with the only requirement for assignment being unfamiliarity with the assigned partner.

Procedure. Dyads were randomly assigned to one of two conditions: face-to-face interaction or Web-based chat. In the face-to-face condition, the members of each pair were brought into a single lab and seated at a table. In the Web-based condition, the 2 participants were instead brought individually into two separate rooms. Participants in both conditions were told that the purpose of the study was to explore how people form impressions of one another, but were not told any details about the economic-exchange game that they would play following their initial interaction.

In this first part of the experiment, participants were asked to have a conversation for 5 min. Participants in the face-to-face condition spoke face-to-face, whereas participants in the Web-based condition spoke over the Internet using AOL instant messenger (AIM). Participants using AIM were asked to refrain from using emoticons. In both conditions, participants were encouraged to speak about whatever they liked, with the exception that they should not discuss what tasks might be coming next. They were given several conversation probes (e.g., "What are your plans for spring/summer break?" "What do you like about living in Boston?") but were told that they should not feel limited to these topics.

The reason for prohibiting any discussion of the upcoming economic game was to remove any possibility for strategic deception. Given that partners could not discuss the game, and did not even know its rules or form, active deception was unlikely. Rather, it was our goal to determine if a partner's general level of cooperative intent could be discerned prior to engaging in any type of negotiation with him or her.

After providing the instructions, the experimenter left the participants alone to have their conversation for 5 min. Participants in the face-to-face condition were recorded digitally for the duration of their conversation by three different cameras: two that captured a head-on view of each participant and one that captured a side view of the dyad as a whole. In the Web-based condition, participants' AIM dialogue was recorded as text.

Following the initial interaction, participants played a single round of the Give-Some Game (cf. DeSteno, Bartlett, Baumann, Williams, & Dickens, 2010; van Lange & Kuhlman, 1994). Participants in the face-to-face condition were removed to separate rooms to complete the game; those in the Web-based condition remained in their separate rooms to complete it. The Give-Some Game is an analogue of a typical prisoner's dilemma but allows for a greater range of behaviors. Each participant is given four tokens, each worth $1 to the participant if he or she keeps it, but $2 if given to the partner. Maximal cooperation and communal gain occur if each individual gives all four tokens to his or her partner (i.e., an $8 payoff for each person). Maximal individual (i.e., selfish) gain accrues to someone who gives no tokens and whose partner gives all four (i.e., a payoff of $12 for the receiver and $0 for the giver). The game thus provides an incremental measure of cooperative intent, with intermediate levels of giving corresponding to degrees of cooperative (i.e., trustworthy) or selfish (i.e., untrustworthy) behavior.

In addition to reporting how many tokens they would offer, participants estimated the number of tokens they believed their partner would offer. Once offers were complete, participants received their respective payoffs. There was no expectation that partners would see each other again; therefore, individuals had no reason to feel pressure to reciprocate.

Coding of nonverbal cues. The digitally recorded interactions from the face-to-face condition were coded by a set of independent coders; each interaction was coded by two separate individuals (interrater agreement: $\rho = .87$). Using all three camera angles and Noldus Observer XT software (Noldus Corp., Leesburg, VA), coders marked the start and stop times for each of 12 types of cues throughout each interaction. The cues to be coded were selected on the basis of their frequencies of appearance; a cue had to be expressed by at least 5 participants in the data set (although in practice, most cues were expressed by many more individuals) to be coded. The final coding provided a time-synchronized stream of nonverbal cues that were emitted by each participant in each dyad. The 12 nonverbal cues in the coding scheme were the following: smile, laugh, lean forward, lean back, arms crossed, arms open, face touch, hand touch, body touch, head shake, head nod, and look away.[3]

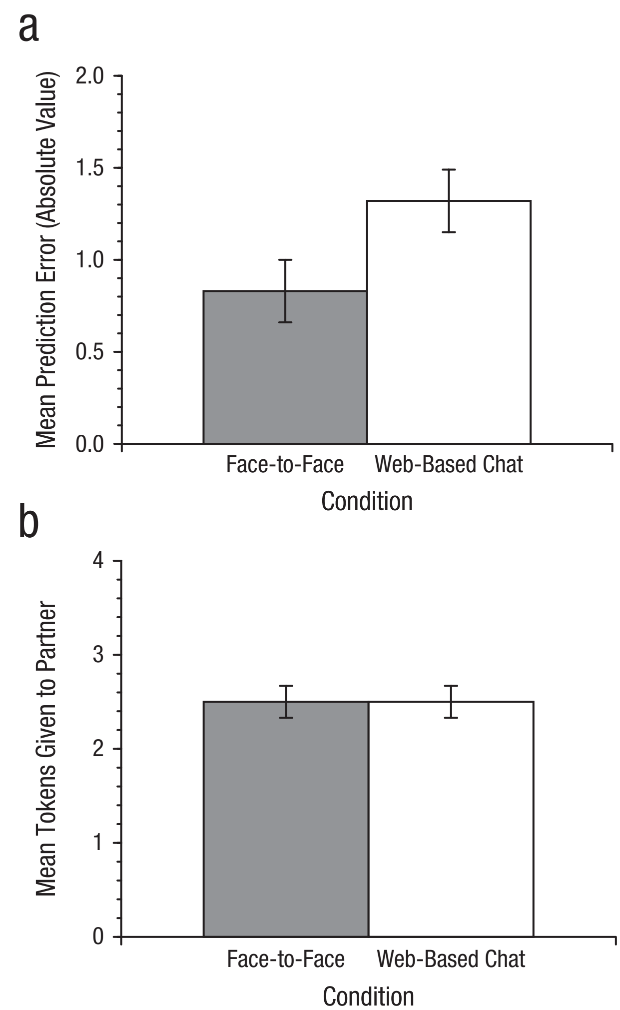

Fig. 1. Results from Experiment 1. The graph in (a) shows mean prediction error (i.e., absolute value of the difference between the number of tokens participants expected they would receive and the number they actually received) as a function of condition. The graph in (b) shows mean number of tokens given to the partner as a function of condition. Error bars signify ±1 *SE*.

Results

A mixed-model analysis of variance treating dyad as a random factor to control for nested dependencies among participants confirmed that, in accord with our primary hypothesis, accuracy in predicting trustworthy behavior was greater when individuals had access to the nonverbal cues of their partners, $F(1, 41) = 3.99$, $p = .05$. As depicted in Figure 1, the average prediction error (i.e., average absolute value of the discrepancy between the predicted and received number of tokens) was smaller in the face-to-face condition than in the Web-based condition. However, general levels of giving were equivalent in the two groups (see Fig. 1). Thus, access to nonverbal cues enhanced accuracy in assessing subsequent trustworthy behavior but did not influence the actual occurrence of such behavior.

In order to identify a candidate signal that was predictive of trustworthy or untrustworthy behavior, we constructed candidate sets of cues based on both existing knowledge relating cues to affiliative or avoidant behavior (Ambady & Weisbuch, 2010; Knapp & Hall, 2010) and examination of zero-order correlations between frequencies of the 12 cues and the actual economic behaviors of participants. We examined the ability of combinations of cues to predict cooperative behavior using multilevel models that allowed for the control of dyadic dependencies within the data. The general model used to assess the signal value of cue sets with respect to self- and partner-expressed cues, respectively, took the following form:

$$\hat{Y}_{ij} = \beta_{0j} + \beta_{1j}X + r_{ij},$$

where

$$\beta_{0j} = \varphi_{00} + \mu_{0j}$$

and

$$\beta_{1j} = \varphi_{10}.$$

Y_{ij} refers to the number of tokens offered for exchange by (depending on the analysis) participants or partners (i) nested in dyads (j); X refers to the mean frequency of the cues (expressed either by participants or by partners, depending on the analysis) in the set; r_{ij} refers to participant-level error; φ_{00} and φ_{10} refer, respectively, to population-value estimates for the intercepts and slopes linking frequencies of the cues to the number of tokens given; and μ_{0j} refers to dyad-level variability in intercept values of the dependent variable (i.e., tokens given by participants or partners, depending on the analysis). The final candidate set for a trust-relevant signal, determined on the basis of power to predict cooperative behavior and established knowledge relating cues to affiliation- and avoidant-relevant intentions, consisted of four cues: hand touch, face touch, arms crossed, and lean away.

As expected, none of these cues offered significant predictive ability when examined in isolation. However, when the cues were taken together in a unit-weighted manner (i.e., the mean value of occurrence across the set of four cues), the resulting signal was predictive of trust-relevant behavior. The more frequently an individual expressed these cues, the less trustworthy was his or her behavior (i.e., the fewer tokens the individual offered to his or her partner), $\varphi_{10} = -0.15$, $p = .03$. Similarly, the more frequently an individual's partner expressed these cues, the fewer tokens the individual decided to offer the partner, $\varphi_{10} = -0.13$, $p = .03$.

Discussion

In Experiment 1, exposure to nonverbal information increased accuracy in assessing the trustworthiness of another person by approximately 37%. Thus, the results of this experiment support the existence of a trust-relevant signal. In addition, we were able to identify a specific candidate cue set that was directly associated with trust-related economic behavior. It is important to note, however, that although the identified cue set received empirical validation in this sample, the study was clearly exploratory. The findings could have stemmed from random sample variations or from spurious correlations between specific cues. For example, given that most people unconsciously generate a multitude of potential cues, certain cues might covary with others (e.g., nodding the head while leaning away), which would make spurious correlations between some subsets of cues nearly inevitable. In order to validate the cue set, we therefore needed to achieve exacting control over all potentially relevant cues so that we could manipulate and isolate them experimentally. This strategy would be truly confirmatory and would allow for a clean test of the cue set's causality.

Experiment 2

Experiment 2 constituted the confirmatory phase of this project. Specifically, its goals were to allow for experimental validation of the causal efficacy of the target cue set identified in Experiment 1. Accomplishing this goal required the ability to manipulate target cues orthogonally to nontarget ones with high precision. However, a fundamental challenge inherent in this research design is that individuals regularly emit cues outside of their own awareness, which makes it difficult even for trained confederates to express individual cues in a reliable and orthogonal fashion. Our strategy for meeting this challenge was to employ a social robotics platform that allowed the specific cues emitted by one member of each dyad (the robot) to be controlled to a degree not possible with humans.

The procedure for Experiment 2 thus closely mirrored that of Experiment 1, with the principal exception being that one of the members of each dyad was replaced with the robot Nexi (see Fig. 2). The primary manipulation centered on whether

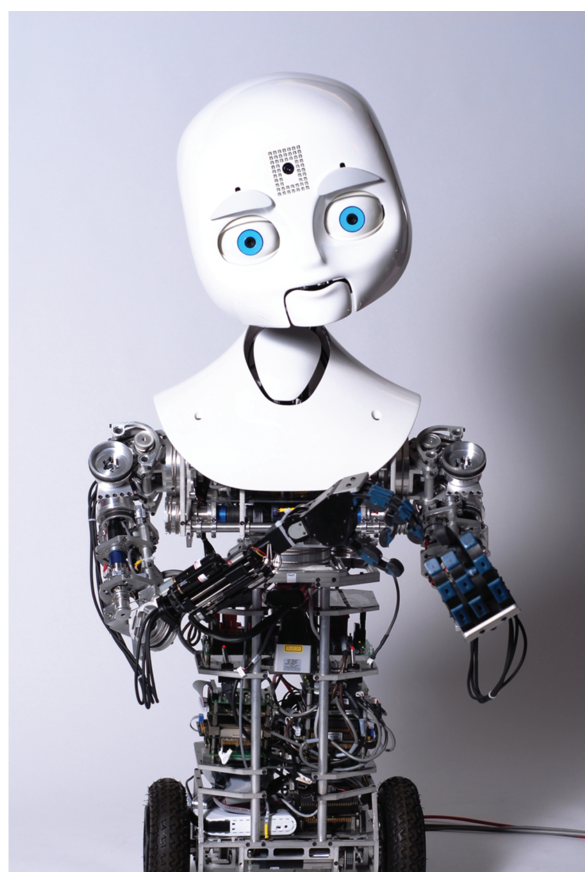

Fig. 2. Nexi, the humanoid robot tele-operated in Experiment 2.

Nexi expressed the target cue set identified in Experiment 1. The causal power of the cue set would be confirmed to the extent that Nexi's expression of the target cues resulted in reduced perceptions of her trustworthiness and subsequent reductions in the number of tokens expected from and offered to her.

Method

Participants. Sixty-four individuals from the greater Boston community agreed to take part in the experiment (22 male, 42 female; mean age = 21 years, *SD* = 2.06 years). These individuals were randomly assigned to one of two conditions: target cues versus control.

Procedure. The procedure for the second experiment was based on that of the first, but with several noteworthy exceptions. The primary difference was that we employed a Wizard-of-Oz paradigm, so called because Nexi was controlled by two operators in a separate room. Participants completed the experiment individually rather than in dyads (i.e., Nexi was the partner for each participant). They were brought into a lab where the humanoid robot was already positioned. Participants were seated across from Nexi, with a small, low table separating them. Nexi welcomed each participant with a wave, saying, "Hello. It's nice to meet you." Participants were then told that for the first part of the experiment, they would have a conversation with Nexi for 10 min. As before, they were not given any specific details about the second half of the experiment, which again involved playing the Give-Some Game. The duration of the conversation was extended from the previous study's 5 min so that participants would have time to adjust to interacting with a robot. At the start of the conversational segment, Nexi provided a few details about herself (e.g., where she was built) to allow participants to become comfortable with conversing with a robot.

As in Experiment 1, participants were given a set of conversation probes (e.g., "What do you like about living in Boston?"), but in this experiment, they were asked to stick to these topics during their interaction. Participants were not told that Nexi was being tele-operated by experimenters in an adjacent room. The experimenter left participants in the room with Nexi for the duration of their interaction. Three different cameras were again used to record all interactions.

Thirty-one of the 64 participants were assigned to the target-cues condition; the remaining 33 were assigned to the control condition. In the control condition, Nexi made several conversational gestures throughout the interaction, but did not engage in any of the target cues that were found to be predictive of untrustworthy behavior in Experiment 1. In the target-cues condition, some of the conversational gestures from the control condition were replaced with the target cues (i.e., hand touch, face touch, arms crossed, and lean away); each target cue occurred one to three times (i.e., frequencies similar to those observed in Experiment 1). The robot's expression of all cues was based on prototypical human motions. That is, the gestures were created by an animator who distilled them from examples of human motions.[4] Care was taken to keep the robot's overall amount of movement consistent across conditions.

Prior to beginning the experiment, we prepared scripted responses to each of the conversation probes. A single female experimenter served as the voice of Nexi and interacted with all participants. Her speech and head movements were tracked and mapped to the robot's head and mouth movements in real time. Cameras and microphones embedded in the robot allowed the experimenters who tele-operated Nexi to see and hear participants conversing with the robot, and face-tracking software was utilized to keep Nexi's gaze centered on participants' faces. A different experimenter operated Nexi's torso and arms throughout the conversation, using a graphical user interface (GUI) to do so. In this way, the experimenter who spoke with the participants could remain blind to each participant's experimental condition (i.e., target cues vs. control). For detailed specifications of the technology underlying Nexi's animation and control, see Robot System Design in the Supplemental Material; Video S1 in the Supplemental Material provides a video example of Nexi's interaction with participants.

Following the "get to know you" period, participants were moved to a separate room, where they played the Give-Some Game assuming Nexi was their partner. They also completed several questionnaires that probed their views of and familiarity with Nexi; the questionnaires included items assessing how much they trusted Nexi and how much they liked the robot (7-point scales).

Results

The causal efficacy of the signal in question would be confirmed to the extent that Nexi's expression of the cues in the set caused participants to judge her to be less trustworthy and correspondingly led them to expect and offer fewer tokens in the economic game. We used the structural equation model depicted in Figure 3 to examine whether this was the case. Results confirmed the cues' signal value; Nexi's expression of the target cues was associated with a significant decrement in her perceived trustworthiness. This decrement directly predicted lower expectations for the number of tokens Nexi would give and also directly predicted a reduction in the number of tokens that participants offered to her. The cues, however, had no effect on economic decisions other than via their impact on perceived trustworthiness. Trimming the direct paths linking condition to the number of tokens expected to be received and the number of tokens given did not diminish the model's goodness of fit, $\chi^2(2) = 3.76$, $p = .15$.[5]

Finally, the effects of the cues appear to have been quite narrowly focused on trust, as their presence or absence did not influence the degree to which participants liked Nexi ($t < 1$). This finding suggests that the presence of the cues did not produce a general "antihalo" effect such that evaluations of the

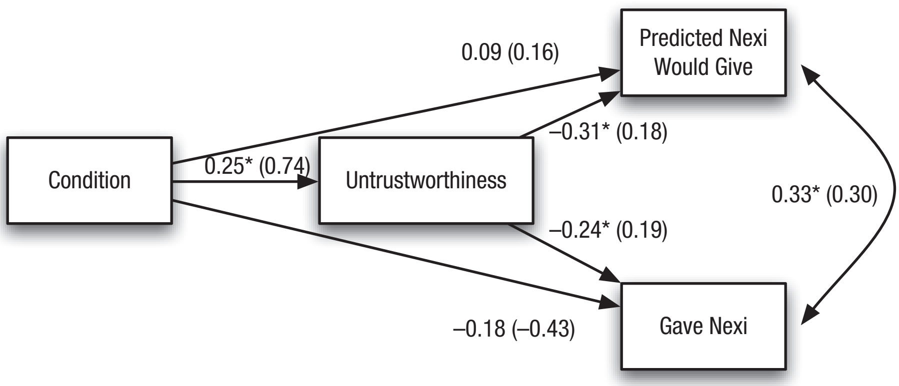

Fig. 3. Path model depicting participants' decisions and expectations in Experiment 2 as a function of condition (dummy coded: 0 = control condition; 1 = target-cues condition) and Nexi's perceived untrustworthiness. Standardized parameters are presented (raw coefficients are inside parentheses). Asterisks indicate significant parameters, $p \leq .05$.

robot on any social dimension became more negative. In many ways, this finding mirrors a familiar experience for many people in that most can point to individuals whom they like but with whom they would not trust their money.

Discussion

These findings are noteworthy for two primary reasons. First, they provide a stringent confirmatory test for the validity of the nonverbal signal identified in Experiment 1. The robotic system allowed us to gain precise control over the cues in order to manipulate them in an experimental manner (i.e., as orthogonal from any other cues) and, thereby, test their causal impact. As predicted on the basis of the correlational data from Experiment 1, presence of the cue signal caused individuals to perceive Nexi to be less trustworthy, which directly affected their economic behavior toward her.

Second, these findings also offer the first evidence that the human mind will respond to trust-relevant signals emitted by humanoid robots in the same manner as they respond to similar signals emitted by humans. It remains to be explored whether the impact of these cues stemmed from the ascription of moral intentions to Nexi's "mind" (cf. Gray, Young, & Waytz, in press; Waytz, Epley, & Cacioppo, 2010; Waytz, Gray, Epley, & Wegner, 2010), or simply from nonconscious utilization of the cues as predictors of Nexi's subsequent moral behavior. Irrespective of mechanism, however, these findings clearly indicate the readiness of the human mind to respond in the expected manner to humanlike biological motion emitted by robotic entities.

General Discussion

Taken together, these findings are among the first to identify a human capacity to assess whether an unfamiliar individual is likely to behave cooperatively in a given situation.[6] They offer empirical support for a phenomenon that has long been theorized to allow for the existence of cooperation, especially in one-shot dilemmas. Yet it is also important to note that the cues we have identified are almost surely not the only ones that affect judgments of trustworthiness. In the circumstances we examined, cooperation was fairly common, but the default expectation in many contexts may be that cheating will occur. In such situations, a different set of cues, such as those typically associated with motivations for affiliation (e.g., leaning forward, affirmative head nods), might hold greater predictive power to detect fair and cooperative, as opposed to selfish and opportunistic, tendencies. Indeed, the informational value of any set of cues is likely to depend on context-based expectations about partners' intentions.

As noted, these findings also offer initial evidence that the human mind will use nonverbal cues to predict the trustworthiness of humanoid robots, thereby opening many avenues for increasing the capacity of robots to function as interaction partners capable of building trust and social bonds with humans through either the presence or the absence of specific gestures. In so doing, our findings support the view that robotics technology has now reached a level where its mirroring of human social cues, though imperfect, is nonetheless sufficient to embody the basic components necessary to engage the social mind's interpretive machinery.

We readily acknowledge the view that robotics might not constitute a valid method to study human behavior, as robots clearly do not look or move exactly like their human counterparts. Such imperfections might bias the human mind's responses to nonverbal cues. Although this is certainly a valid concern, it is one that, to our minds, can be addressed empirically as opposed to being based on subjective impressions of the "humanness" of any robot. Indeed, within the paradigm of "computers as social actors," a wide range of technological embodiments capable of expressing social cues (e.g., from computers that interact via text or speech, to animated avatars, to physical robots) has been shown to evoke natural human social responses and social judgments, depending on how the cues and embodiments are technologically implemented (Blascovich & Bailenson, 2011; Kidd & Breazeal, 2008; Sidner, Lee, Kidd, Lesh, & Rich, 2005; Siegel, Breazeal, & Norton, 2009).

In the present case, the data clearly suggest that participants' minds responded to Nexi's nonverbal cues as they would to those of a human. Had Experiment 2 failed to confirm the findings of Experiment 1, any number of reasons could have been posited. For example, the findings from the human-human interactions could have been incorrect, or the technology of the robot might not have been capable of adequately mirroring human movement. However, when the findings of the two experiments are viewed as a whole (i.e., the expression of the cues by the robot confirmed the cues' predicted impact based on human-human interactions), we believe that the most parsimonious explanation is the proffered one. Any alternative explanation would require one to accept the view that a biasing agent (e.g., timing of movements) inserted itself such that it not only produced the predicted behavioral effects (i.e., effects matching those from the human-human interactions), but also did so via the predicted mediator.

We maintain that interdisciplinary techniques, such as those used here, hold great promise for the study of human social dynamics. Although it is true that the use of interdisciplinary techniques can often be accompanied by increased ambiguities in the interpretation of any single experiment, as the simple act of combining different technologies and paradigms can similarly combine the methodological shortcomings of each, the potential risks that result can be offset through the use of multiple approaches. When findings converge across methodologies, confidence in their robustness is greatly increased, as the number of ways in which replication can erroneously emerge becomes vanishingly small.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

We acknowledge the support of National Science Foundation Research Grants BCS 0827084, 0827088, and 0827094.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. Delton et al. have presented a model simulation showing that cooperation in one-off interactions can occur regularly if expectations for future interaction are nonzero. However, their model also suggests that an adaptive advantage in deciding to cooperate may stem from an ability to gain insight into the trustworthy intent of a partner.

2. One exception has been evidence suggesting that individual differences in facial width are associated with differences in selfish behavior (Stirrat & Perrett, 2010). Such a marker, however, is static, and thus unable to account for differences in interactions across situations.

3. Supplementary Methods and Analyses, in the Supplemental Material available online, provides additional information on cue selection and cue frequencies.

4. Prototypical examples of target cues were derived from the video recordings of Experiment 1.

5. Note that the initial model was saturated, and thus fit perfectly. Therefore, the nonsignificant chi-square test for the reduced model indicates not only a lack of substantive change in fit for the reduced model, but also an acceptable fit for the reduced model overall.

6. Frank, Gilovich, and Regan (1993) presented initial findings demonstrating that trustworthiness (operationalized within the context of a prisoner's dilemma) could be assessed with greater-than-chance accuracy; however, this seminal work did not address the mechanisms by which such judgments might occur.

References

Ambady, N., & Weisbuch, M. (2010). Nonverbal behavior. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 464–497). Hoboken, NJ: John Wiley & Sons.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*, 286–290.

Blascovich, J., & Bailenson, J. N. (2011). *Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution*. New York, NY: William Morrow.

Delton, K. W., Krasnow, M. X., Cosmides, L., & Tooby, J. (2011). The evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, USA*, *108*, 13335–13340.

DeSteno, D., Bartlett, M., Baumann, J., Williams, L., & Dickens, L. (2010). Gratitude as moral sentiment: Emotion-guided cooperation in economic exchange. *Emotion*, *10*, 289–293.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York, NY: W. W. Norton.

Frank, R. H., Gilovich, T., & Regan, D. (1993). The evolution of one-shot cooperation. *Ethnology and Sociobiology*, *14*, 247–256.

Gray, K., Young, L., & Waytz, A. (in press). Mind perception is the essence of morality. *Psychological Inquiry*.

Hall, J. A., Coats, E. J., & Smith Lebeau, L. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, *131*, 898–924.

Keltner, D., & Buswell, B. N. (1997). Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin*, *122*, 250–270.

Kidd, C. D., & Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3230–3235). Piscataway, NJ: IEEE Press.

Knapp, M. L., & Hall, J. A. (2010). *Nonverbal communication in human interaction* (7th ed.). Belmont, CA: Wadsworth.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*, 140–164.

Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive robotics: The influence of robot gender on human behavior. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2563–2568). Piscataway, NJ: IEEE Press.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*, 349–354.

Todorov, A. (2008). Evaluating faces on trustworthiness. *Annals of the New York Academy of Sciences*, *1124*, 208–224.

Todorov, A., Baron, S., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience*, *3*, 119–127.

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Psychological insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*, 58–62.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*, 383–388.

Supporting Online Material

SUPPORTING TEXT AND ANALYSES

Identification and Analysis of Cue Set in Experiment 1

Selection of the set of target cues occurred through several stages. The first stage involved examination of the mean frequencies of cues and correlation matrices for associative links between specific cues and trust-relevant economic variables (i.e. tokens given by individuals expressing cues and corresponding tokens given by partners). Tables S1 and S2 present the relevant data. Note that significance values are not provided, as due to the dyadic nesting of the data, standard errors are not valid; coefficient estimates, however, are valid.

As expected and readily noted from the data, the frequencies of expression of many individual cues covary. Consequently, use of the zero-order correlation matrix on its own to suggest predictive power represented only an initial starting point, but not a firm decision basis (note, correlation values were also screened to ensure that they were not overly influenced by a few outliers). Moreover, we firmly believed that a strict reliance on an item-by-item significance-testing basis to form a composite would be misguided even if the standard errors for zero-order correlations were not biased due to the inherent dyadic dependencies. It was our *a priori* view that predictive power would come from a set of cues, as opposed to single cues, as the meaning of any one cue in isolation is often difficult to interpret. Therefore, we did not rely on a mechanistic quantitative methodology (e.g., stepwise regression) to build the regression model (due to the correlations among the variables and the necessity such a method would have required with respect to lower power levels due to the need to control for family-wide error). Indeed, traditional stepwise methodologies are not readily available for the multilevel model analyses that are needed to calculate the appropriate standard errors for significance testing. We therefore began to choose sets of cues based both on the above correlation matrix and *a prior* expectations about the general meanings of certain cues with regard to motivation. Indeed, examining sets is quite important, as they can disambiguate the meaning of specific single cues in a given instance. For example, leaning away could be a marker for non-engagement or simply a posture change for comfort; however, when it co-occurs with crossing arms, it is more likely a marker of nonafiliation (and a simple zero-order correlation between this cue and trust-relevant behavior would be unable to pick up the contextual difference and thereby be attenuated). Once a candidate set of cues was selected as a signal (based on the multilevel modeling), we would then rely on the cross-validation of the findings in a separate sample to rule out concerns about capitalization on chance or spurious relations. Thus, although there is some necessary subjectivity inherent in selecting the final set of cues based on data from Experiment 1, tests of validity of those cues in Experiment 2 constitute a stringent test of their causal impact. Through manipulating their presence or absence with great precision, the ability of the corresponding signal to cause alterations in perceived trustworthiness and subsequent economic behavior could be rigorously examined.

Candidate sets of cues were subjected to analysis using multilevel models of the form identified in the main text. The final set of cues (based on the criteria noted above and a requirement of explaining significant variance in economic behavior [alpha criterion < .05 when combined]) consisted of four items: face touch, hand touch, crossing arms, leaning back. It is instructive to note that although this set of cues has not been previously associated with untrustworthy perceptions, it does possess partial overlap with sets of cues previously identified to signal a sense of unease within interpersonal interactions (Knapp & Hall, 2010). The independent variable consisted of the mean frequency across these items. As such, the model may be viewed as a causal indicator one wherein increasing frequencies of specific individual indicators increases scores on the overarching construct (here *untrustworthiness*). That is, increasing frequencies of these cues, especially when coinciding the expression of the others in the set, functions to signify greater untrustworthiness.

Recognition of Cues by Nexi in Experiment 2

Finally, in order to confirm that the robot-emitted cues were recognized as corresponding to their human counterparts, we had participants subsequently (i.e., after their participation in the primary parts of the experiment) view videos of Nexi expressing the target cues. After viewing each cue, participants were presented with 5 multiple-choice options to identify it; one option was always "none of the above."

If participants selected "none of the above," they were asked to write in a description of what they believed Nexi was doing. For the "Lean Back" item, we included descriptions of the nature "twisted away" as acceptable, as Nexi's leaning back (modeled after the prototypical one expressed by humans in Experiment 1), consisted of drawing the neck and upper torso back while rotating slightly to one side.

To facilitate analysis of the findings, responses were dummy-coded: 0=incorrect answer, 1=correct answer. Given the 5 options for each question, chance recognition would be 20%. For each of the four cues, a one-sample t-test comparing mean recognition values against a chance level of .20 confirmed that recognition for all cues was significantly greater than chance (all t's > 20; all p's < .001).

<u>Cue (mean proportion correct; standard error)</u>
Arms Crossed (.91; 0.04)
Face Touch (.96; 0.03)
Hand Touch (.97; 0.02)
Lean Back (.64; .06)[§]

[§]The lower recognition level for leaning back is to be expected given the fact that Nexi leans backward in a direction aligned with participants' lines of sight in the video. Given the reduction in accurate depth perception that accompanies viewing Nexi on a two-dimensional monitor as opposed to in person, the lower accuracy level makes sense. Of great import, the level of recognition was still greatly above chance.

Table S1

Descriptive Statistics for Cue Frequencies

Cue	N	Minimum	Maximum	Mean	Standard Deviation
Smile	42	4	31	18.71	6.954
Laugh	42	0	23	8.52	6.383
Lean Forward	42	0	5	1.52	1.435
Lean Back	42	0	3	.55	.916
Arms Crossed	42	0	14	1.76	3.230
Arms Open	42	0	3	.19	.552
Face Touch	42	0	20	4.60	4.013
Hand Touch	42	0	16	3.76	4.089
Body Touch	42	1	13	4.93	3.316
Head Nod	42	1	34	18.05	7.865
Head Shake	42	0	18	5.33	3.613
Look Away	42	4	55	26.07	14.477

*Note minimum and maximum refer to the number of occurrences within a single individual.

Table 2

Correlations of Nonverbal Behaviors and Tokens Given in Give Some Game

	Tokens Given	Partner Gave	Smile	Laugh	Lean Forward	Lean Back	Arms Crossed	Arms Open	Face Touch	Hand Touch	Body Touch	Head Shake	Head Nod
Tokens Given	-												
Partner Gave	.12	-											
Smile	-.07	.18	-										
Laugh	.03	.17	.62	-									
Lean Forward	-.18	.02	-.21	-.27	-								
Lean Back	-.07	.08	-.19	-.10	.59	-							
Arms Crossed	-.21	-.13	.20	.16	.08	-.02	-						
Arms Open	-.10	-.10	.09	-.02	.09	.32	-.07	-					
Face Touch	-.24	-.32	-.01	-.19	.26	-.08	.13	-.11	-				
Hand Touch	-.22	-.16	.13	-.11	-.02	-.04	-.03	.37	.39	-			
Body Touch	-.08	.06	.08	.16	.11	-.01	.22	-.17	.38	.07	-		
Head Shake	.03	.11	-.12	.01	.44	.32	.21	.27	.24	.05	.26	-	
Head Nod	-.31	-.09	.09	-.22	.26	.11	.29	-.04	.23	-.04	.14	.26	-
Look Away	-.20	.07	-.22	-.20	.10	.11	-.02	.00	.12	.33	-.06	-.01	-.26

ROBOTIC SYSTEM DESIGN

Overview

We are investigating the use of nonverbal cues as a powerful signal that can reveal information about the dynamics of trustworthiness not only between humans but also between robots and humans. As such, by using a robotic system, we are taking advantage of its consistent and programmed behavior to tightly control the nonverbal cues that are displayed. By having this control, we can eliminate other subtle gestures that a confederate, or a human actor pretending to be a subject, could unintentionally exhibit. Furthermore, by using robots, we can also investigate whether this dynamic can translate and replicate onto humanoid robots.

We have developed a system that allows remote robot operators to have a conversation with a person through the robot while simultaneously controlling the robot's movement as effortlessly as possible. In order to achieve as natural a social interaction as possible, our system combines various fully-autonomous and semi-autonomous interfaces to teleoperate our humanoid robot, Nexi.

Research Platform

Our hardware platform is a mobile, upper torso humanoid robot named Nexi, who is part of our MDS (Mobile-Dexterous-Social) robot line as seen in Figure S1. The MDS robot is 4-feet tall and has 15 facial degrees of freedom (DoFs), 4 neck DoFs, a pair of 3 DoF shoulders, a pair of 5 DoF lower arms and hands, DoFs, a mobile wheel base with 2 DoFs, and a torso DoF. The face has several facial features (eyes, eyelids, eyebrows, and jaw) to support a diverse range of facial expressions and an articulate mandible for expressive speech. The neck mechanism has 4 DoFs to support a lower bending at the base of the neck as well as pan-tilt-yaw of the head. The head can move at human-like speeds to support human head gestures such as nodding, shaking, orienting, craning the neck forward, and recoiling the neck back.

In the following sections below, we explain the various fully-autonomous and semi-autonomous interfaces implemented in order to control Nexi's head, mouth, eyebrows, eyes, and body.

Head Control for Communicative Cues

The mode of control for the robot's head is similar to that used in previous work with the MeBot robot (Adalgeirsson & Breazeal, 2010). Adalgeirsson et al.'s main motivation in their design was to create an interface that would capture the natural head movements of the operator while also not costing the operator too much cognitive load. Thus by using faceAPI, a realtime face-tracking software library, to estimate the operator's 3D head orientation, we automatically captured the natural head movements as the operator simply sat in front of a camera. Head movements like nods and shakes, which convey important communicative information, were easily generated through this interface.

To teleoperate the robot's head movements, we mapped the operator's x,y,z head orientations estimated from faceAPI to the robot's pitch, yaw, and roll head joints, which is demonstrated in Figure S2 (see below). By controlling the robot's head through this interface, we were able to achieve fairly natural movements that corresponded in realtime to that of the operator's head movements. In our experiment, the operator responsible for talking with the participants was the one whose head movements were tracked and mapped to the robot's head movements.

Mouth Control for Lipsyncing

By using Annosoft's realtime lipsync sdk, we extracted the visemes, or the visible mouth positions that occur in speech, from an incoming audio source. We then perform the appropriate mappings between the human-visemes to robot-visemes to have the robot speak with human-like mouth movements. The robot's mouth animation visemes were designed by a professional animator. And it is important to note that this mapping is not one-to-one as our robot's mouth has only 3 DOFs that allow the jaw to move up and down, jut in and out, and roll left and right. However, the resulting articulated jaw produced readable mouth positions that corresponded in realtime to the operator's speech (as seen in Figure S3).

Eye Control for Gaze

In interpersonal communication, establishing eye-contact serves as a signal that the channel for communication is open between the participants, and the individuals involved are attending to each other. As such, for the robot to establish eye-contact, we used OpenCV, an open source computer vision library, to detect and track faces viewed from the camera in the robot's right eye. Once a face is found, the robot maintains eye-contact by centering the face in its field of vision.

OpenCV possesses some limitations in tracking face targets, especially when they are in motion (i.e., faces of individuals who are moving around a room). Fortunately, the setup of the experiment had the participant sit in a chair, at a suitable distance from the robot, making the face much easier to track. Nonetheless we instituted two override modes that a monitoring operator could activate in cases of unpredicted failures of OpenCV. For the first override, we used a graphical interface in which an operator could literally see through the eyes of the robot using embedded cameras in Nexi. Whenever the participant's face was not centered in the image, the operator could manually click in the interface where the robot's eyes should be correctly centered (allowing autonomous tracking to resume from that point). Second, if the autonomous tracking continued to fail even after manual corrections, we would turnoff autonomous tracking completely and fix the gaze point at a predetermined location where an average height participant would most likely be located on the fixed-positioned chair. This secondary override was used only once (participant in the control [no target cue] condition) and caused no appreciable effect on the participant's data with respect to deviations from others in this condition.

Eyebrow Control for Expression

To make the robot's face appear engaged in the conversations, we moved the robot's eyebrows according to the energy of the operator's voice. As the energy of speech increased (measured using the root-mean-squared (rms) value of the signal's amplitude), the robot's eyebrows rose. At first, we anticipated that the eyebrow movement might seem too exaggerated, but the resulting movement only made the robot's facial expressions appear more conversational and lively.

Body Control for Gestures

To control the robot's arms and neck, we developed a button interface that can easily activate and deactivate various body gestures (shown in Figure S4). The robot's arm and body gesture animations were designed by a professional animator to resemble human-like movements as much as possible. These included animations for initiating each gesture and also for releasing them. The Gesture Controller window contains start and stop buttons for the following predictive gestures: hand-touch (low, middle), face-touch (left, right), cross-arms, lean-back (left, middle, right), and a few combination gestures like cross-arms plus lean-back (pictures of Nexi demonstrating these gestures can be found in Figure S5). In addition to these predictive gestures, we also have other communicative gestures for only conversational purposes like hand waving and emphasis gestures. To provide the operator with direct visual feedback, we graphically display the robot performing the current gesture in motion. This feedback mechanism allows the operator to see the direct effects of his/her control and know when the robot has completed a gesture.

Reference

S.O. Adalgeirsson, C. Breazeal, *Proceedings of the Fifth ACM/IEEE Intl. Conf. on Human-Robot Inter. 10,* 15 (2010).

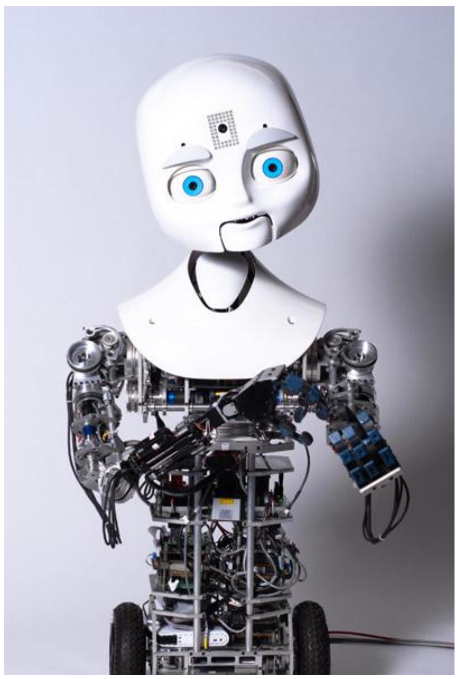

Figure S1: Research platform Nexi, an upper torso humanoid robot designed to be mobile, dexterous, and social.

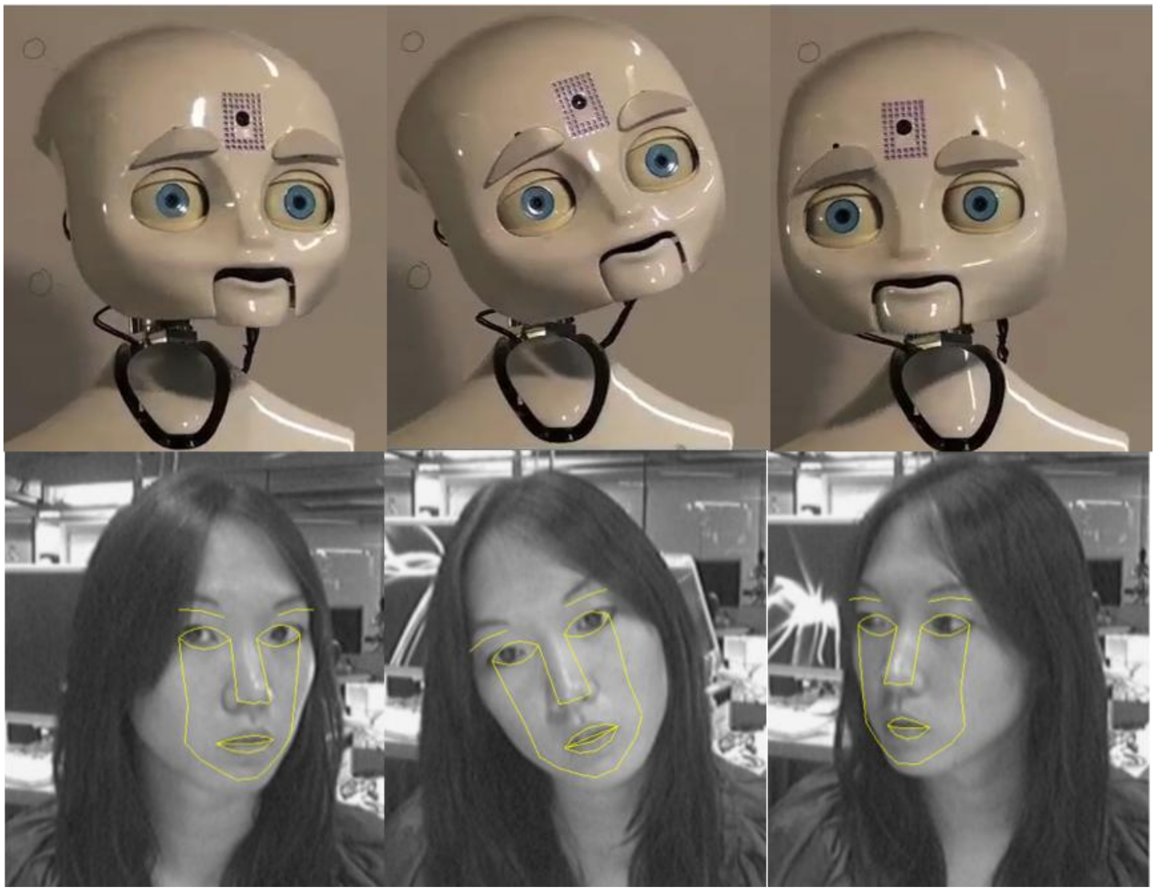

Figure S2: Using faceAPI to detect the operator's 3D head orientation, we can capture the natural head movements of the operator and project them onto our humanoid robot.

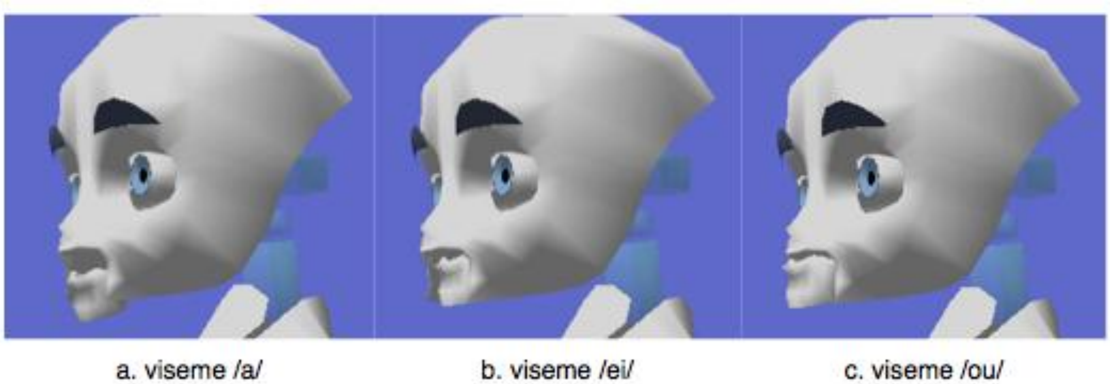

Figure S3: Using Annosoft's realtime lipsync sdk, we can detect visemes like /a/, /ei/, and /ou/ from speech. And with appropriate mappings, the robot can display human-like mouth positions that correlate with the speech.

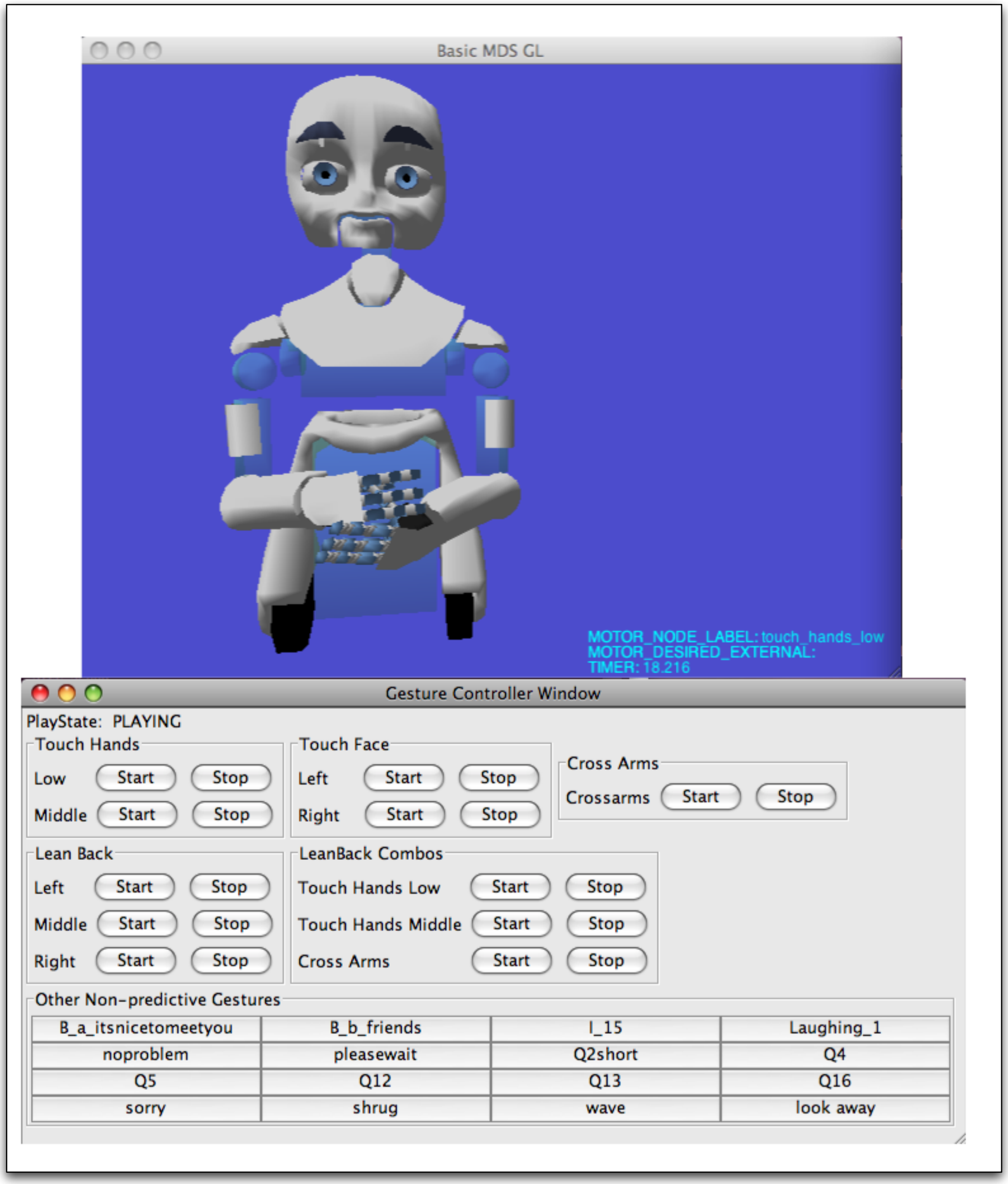

Figure S4: The body control interface consists of the Basic MDS GL window and the Gesture Controller window. The Basic MDS GL window serves as visual feedback for the operator. And the Gesture Controller is a button interface to engage and disengage both the predictive gestures and the non-predictive communicative gestures.

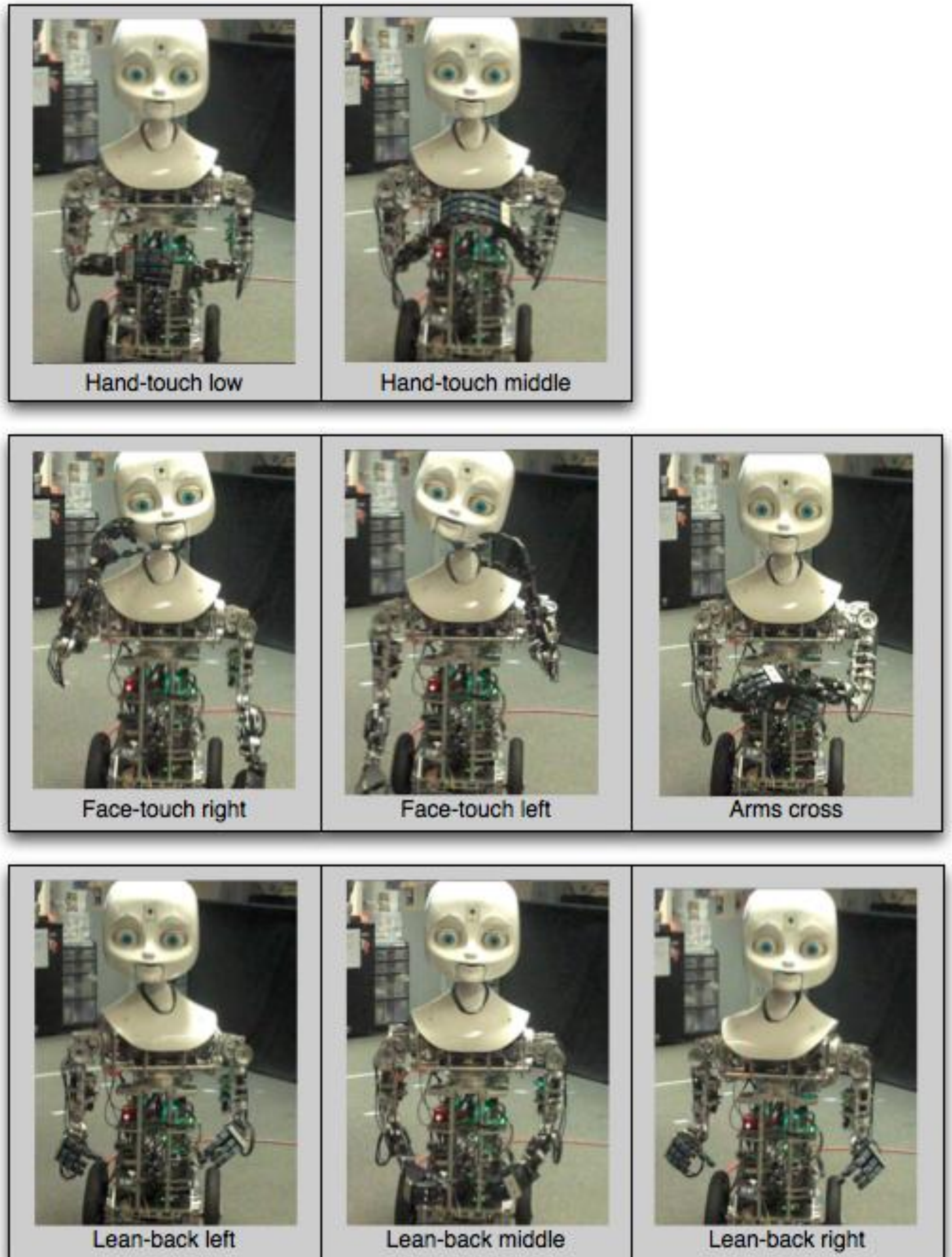

Figure S5: Nexi demonstrating the predictive gestural cues: hand touch, face touch, arms cross, and lean back. Note: the neck also recoils back in the lean-back gestures.