Team Hanson-Lia-SingularityNet:

Deep-learning Assessment of Emotional Dynamics Predicts Self-Transcendent Feelings During Constrained Brief Interactions with Emotionally Responsive AI Embedded in Android Technology

Julia Mossbridge, PhD
Dept. of Psychology
Northwestern University[1]
Lia, Inc.[2]
[1]Evanston; [2]Sebastopol, USA
julia@liatech.ai

Benjamin Goertzel, PhD
Hanson Robotics
SingularityNet
Sha Tsui, Kowloon, Hong Kong
ben@goertzel.org

Edward Monroe, PhD
Lia, Inc.
Sebastopol, USA
eddie@liatech.ai

Goldie Nejat, PhD, PEng
Dept. of Mechanical and
Industrial Engineering
University of Toronto
Toronto, Canada
nejat@mie.utoronto.ca

Liza Lichtinger, MA
Future Design Station
San Francisco, USA
lichtliza@gmail.com

David Hanson, PhD
Hanson Robotics
Science Park, Hong Kong
david@hansonrobotics.com

Abstract— **Our grand challenge is to create emotionally-sensitive AI embedded in social humanoid robots and avatars in order to help individuals advance in the hierarchy of human development. The peak of this hierarchy is self-transcendence, including expansive feelings of love. In this paper we present results of the first experiments of which we are aware in which AI-driven, audio-visual, interactive android technology is successfully used to support the experience of self-transcendence. Specifically, we designed two studies in which people had brief, constrained AI-driven conversations with emotionally responsive AI embedded in a humanoid robot, its audio-visual avatar, or audio-alone avatar. These conversations were based on exercises reported to induce self-transcendence in humans. In experiment 1, we tested an initial version of this AI using brief, constrained interactions with Sophia the humanoid robot and no emotion detection (N=26). In experiment 2, we tested a more sophisticated version of this AI including deep-learning-based emotion detection deployed in the context of a slightly longer and slightly less constrained interaction, in a between-groups design: conversations were with either Sophia or one of two avatars (one with a face and voice, the other with only a voice; N=35). By the time we submitted our first report last year, we had planned but not completed the first study, so in this report we summarize the hypotheses, methods, and results of both studies. The results suggest that conversations between humans and a humanoid robot or its audiovisual avatar, controlled by emotionally responsive AI, are accompanied by self-transcendent emotions, and that objective correlates of those feelings are detectable by a deep learning network.**

Keywords—deep learning, emotion detection, human-centered AI, emotionally responsive AI, humanoid robots, human-robot interactions, compassionate AI, technology-assisted human development, transcendence technology, artificial general intelligence, Hanson AI with OpenCog, Sophia

I. PROBLEM STATEMENT

Because improving human psychological wellbeing is a virtually universal interest among all humans, our team is focused on the development of psychological wellbeing in general. Here we report results of the first study of which we are aware in which AI-driven, audio-visual, interactive android technology is successfully used to help people experience aspects of the highest levels of human development, as assessed via subjective and objective measures.

The hierarchy of human development has been conceptualized in many ways; one is Maslow's Hierarchy of Needs, which moves from physiological needs through needs for safety, social connection, self-esteem, self-actualization and finally self-transcendence (for reviews, see [1-4]). We believe AI-powered humanoid robots can be valuable at every stage of the human-development process, but we have chosen the novel approach of beginning from the apex of the hierarchy and viewing the issue of human-robot interaction and human development from the standpoint of self-transcendence.

Self-transcendence includes detaching from the importance of oneself, seeing the perspectives of others, and having feelings of care toward others [2-7]. Several lines of evidence suggest that experiencing a state of self-transcendence in itself is beneficial to human wellbeing [5,8-11]. Certain meditative, deep-listening, and eye-gazing practices have been either formally or anecdotally reported to help people access self-transcendent states [example formal reports: 11-12]. Importantly, each of these practices are traditionally performed, at least at first, with a teacher and student in visual connection, though this may not always be necessary [e.g., 12].

Before our recent work as part of this xPrize initiative, it had not been shown that AI-driven conversations could induce anything like a self-transcendent state. For this year's xPrize entry, we continued testing several hypotheses related to this unknown territory, so we could drive our grand challenge goal forward in an informed way. Specifically, we designed two studies in which people had brief, constrained AI-driven conversational interactions with an emotionally responsive humanoid robot, its audio-visual avatar, or audio-alone avatar. These conversations were based on exercises reported to induce self-transcendence in humans. By the time we submitted our first report last year, we had planned but not completed the first study; so here we summarize the hypotheses and results of both studies.

Together, the studies examined the hypotheses that: 1) self-reported loving feelings for others would increase from before to after the interactions, 2) self-reported positive mood would increase from pre- to post-interaction, 3) self-reported arousal would decrease during the same time period, 4) heart rate variability measures would be influenced in the direction of a reduction in cognitive load from prior to following the interactions, 5) feelings of anger, fear and disgust (measured using a deep-learning emotion detection network) would decrease significantly during the interactions, 6) dynamic changes in deep-learning-detected emotions would predict changes in self-reported feelings, and 7) some of these hypotheses would be borne out in results from conditions that allow for eye contact, but not in results from people interacting with the same AI in an audio-only condition. The results of the two studies described in this paper, supported or partially supported hypotheses 1, 2, and 4-7, suggesting that guided conversations with a humanoid robot and its audiovisual avatar, controlled by emotionally responsive AI, are correlated with increases in subjective feelings related to self-transcendence in human participants as well as objectively related manifestations of those feelings, as detectable by deep learning.

Complete success of our grand challenge would mean that each and every participant would demonstrated increases in loving feelings for others and self-transcendence, as measured by subjective and objective measures indicating that those emotional changes are tied to the dynamics of the interactions they had with the technology. That is the eventual goal toward which we are striving.

II. Developing Technologies

A. Dialogue Control and Cognitive Model AI

Toward our grand challenge, we created what we call "Loving AI", which is robot- and avatar-embedded AI that performs emotion detection, emotional production/mirroring, and dialogue control while guiding humans in meditation, deep listening and/or eye-gazing practices in one-on-one conversations. We used multiple types of AI in the two experiments.

In both experiments, we drew on Google's speech-to-text machine-learning product [13] to convert the participants' words into text that could be processed by a dialogue engine. In experiment 1, participants spoke with Sophia the humanoid robot, a process that relied on a Chatscript-based dialogue engine to control Sophia's verbal and emotional responses and to control which practices Sophia would guide participants in performing.

In experiment 2, participants spoke with either Sophia or one of her two avatars. In all three cases we used OpenPsi, part of the open source OpenCog artificial general intelligence (AGI) research platform, included in the Hanson AI with Opencog package [14] to direct the conversations. OpenPsi is a model of human motivation, action selection, and emotion inspired from earlier work in human psychology and AI [15]. OpenPsi consists of goals with associated dynamic urge levels. The urge level indicates the current importance to the system of a particular goal, in other words, the urge of the system to satisfy a goal. Rules associated with goals define what actions lead to satisfaction of goals in different contexts. Rules take the form, "Context + Action → Goal Satisfaction." Action selection involves determining which actions in the current context will maximize satisfaction of goals with the highest urge levels. In the Loving AI dialogue, often the goals are related to engaging in different parts of dialogue interaction, actions are the android's verbal responses and emotional expression, and contexts are the verbal and emotional expressions of the participant. In this way, OpenPsi controlled the weight given to particular aspects of the dialogue, depending on verbal cues participants gave as to their willingness to do the practices. OpenCog Ghost, a dialogue scripting and robot control subsystem of OpenCog, contained a corpus of pre-defined facial movements, sounds, words, and phrases. Fig. A1 (Appendix) shows Ghost pseudocode, including Openpsi urge information, for an example interaction. Beyond this relatively simple rule-based AI, we used a deep-learning network to infer the participants' emotional states and to support emotional mirroring.

B. Emotion Detection and Mirroring AI

Our team was aware of evidence from cognitive neuroscience that as a network, mirror neurons may underlie feelings of empathy and affiliation in humans [reviews: 16-17]. It was a major goal of our work to help people feel connected to the android technology and understood by it as well, so we chose nonverbal facial emotion mirroring as a consistent feature in both experiments to support this goal, in an untested attempt to stimulate mirror neurons in our participants. In experiment 1, we used a RealSense camera embedded in the robot's chest to detect the dynamic positions of facial features, and Sophia was programmed to immediately reproduce as best as possible a participant's facial movements, including blinks and eye closings.

In experiment 2, we calculated facial features and their movements via webcams embedded in the robot's eyes, or for the avatar conditions, the webcam on the laptop presenting the avatars. These features were used as input into a pre-trained deep-learning network that classified seven emotional states (happiness, sadness, anger, fear, disgust, surprise, and neutral; for training methods and confusion matrix, see next section). While all android technology calculated emotional states continuously throughout the interactions, Sophia and her audiovisual avatar (but not the audio-only avatar) used the output of the deep-learning network as input to OpenCog Ghost, which produced pre-determined emotional responses matching the currently determined peak emotion out of the possible seven emotions, at intensity levels matching the user's intensity level. These mirroring animations were performed with a gradual, smoothed slope, peaking with an approximate 2-second delay from the originally detected emotion (code at [18]). In experiment 2, blinks and eye closings were not mirrored.

C. *Deep-learning Emotion-detection Network*

We used the CK+ [19-20] and Kaggle FER2013 [21] datasets to train a feed-forward convolutional neural network (CNN) with landmarks as additional input vectors for emotion recognition from facial expressions, resulting in the model available at [22]. CK+ and Kaggle FER2013 are primarily used in facial image analysis research.

CK+ contains 593 gradual expressions of emotions, going from a neutral base pose frame to the maximum expression, captured from 12 participants. Labelling emotion categories in CK+ was done via the FACS-coded emotion labels, in a three-step process. First, the sequence labels were compared with the Emotion Prediction Table from the FACS manual [23], and sequences that satisfied the criteria were provisionally added as belonging to a specific emotion. Second, some sequences were excluded because they did not fit qualifying criteria listed in [19]. Finally, the authors performed a visual inspection for each of the sequences to exclude any sequence that did not subjectively seem to belong to the assigned emotion category.

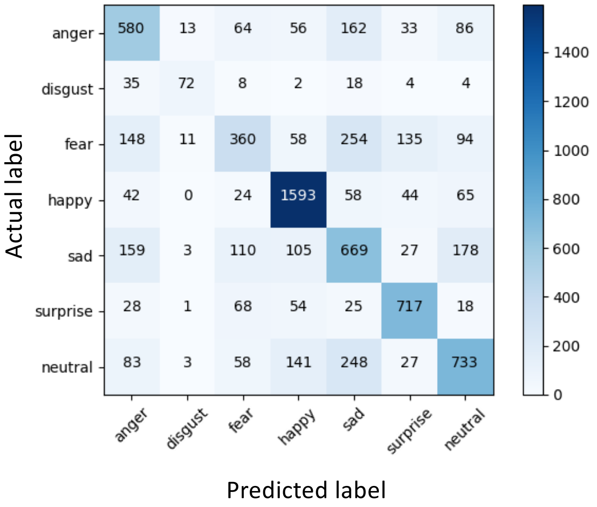

Fig. 1. Confusion matrix for the feed-forward CNN used for emotion recognition. The best performance was on happy expressions.

FER2013 consists of 28709 labelled examples of emotional expressions from a wide array of people, from seven categories (anger, disgust, fear, happy, sad, surprise and neutral). This dataset was created by using a set of emotion-related keywords that were combined with words associated to gender, age, and ethnicity sent to Google Image Search. OpenCV face recognition was applied to the results of these searches to obtain bounding boxes for each of the faces in the images. Human labelers then cleaned up and rejected some of the images, after which they assigned each image to one of the seven emotions mentioned above.

As input, the model uses normalized and cropped faces at 48x48x3 pixels, and 68 landmarks detected by the Dlib facial analysis toolkit [24] as additional feature vector inputs. The output of the model comprises probabilities for each of the seven basic emotions as labelled in [20-21]. Validation accuracy of this model was 63.17% with 20% of the training data used for validation. See Fig. 1 for a confusion matrix plot of this validation run.

D. *Robot and Avatar Creatiom and Performance*

Sophia the robot was produced by Hanson Robotics via proprietary means. The robot's voice was created from a pre-recorded human female vocal repertoire, controlled with a text-to-voice process. The audiovisual and audio-alone avatars both used the same human voice repertoire and vocal control process as the robot. The audio-alone avatar was presented while a blank black laptop screen was shown to the user, while the voice conducted the conversation. The animation for the audiovisual avatar was created using proprietary means, using the same animation control as for the robot's animation process.

III. Technology Impact & Technical Evaluation

We have already created this technology, so here we explain how we evaluated the technology and its impact thus far.

A. *Participants and Procedure*

a) Participants. All participants read and signed consent forms for the experiment, were informed that they were being video recorded, and had a choice after the experiment to sign a consent to release their video publicly or not. Participants were recruited through IRB-approved fliers and email messages that did not describe the exact purpose of the experiment.

Twenty-six participants ranging in age from 20 to 50 participated in experiment 1 during September – November 2017 at Hong Kong Polytechnic University in Hong Kong. Ten participants spoke English as their native language; the remaining 16 spoke English as a second language. Participants were paid 40 HK dollars (about $5 US) for their participation. Thirty-six participants ranging in age from 18 to 75 participated in experiment 2 during June 2018 at Sofia University in Palo

This work was funded by the Hummingbird Foundation through the Institute of Noetic Sciences, Hanson Robotics, the OpenCog Foundation, and SingularityNet.

Alto, CA. Twenty-seven of these participants spoke English as their native language; the remaining 9 spoke English as a second language. There were 13 participants in the robot-interaction group, 11 participants in the AV avatar interaction group, and 11 participants in the audio-alone interaction group. Participants were given a $10 US Amazon gift card for their participation.

b) Procedures common to both experiments. All instructions and the experiments themselves were conducted in English. After reading/signing the consent form, each participant was told that they would be asked to have a conversation with an android technology. They were not told the nature of the conversation. Then they were asked to complete an online pre-interaction questionnaire in which they were asked to rank: 1) their current feelings of love for family members, friends, acquaintances and strangers (5 questions); 2) their current feelings of unconditional love, defined as “the heartfelt benevolent desire that everyone and everything—ourselves, others, and all that exists in the universe—reaches their greatest possible fulfillment, whatever that may prove to be. This love is freely given, with no consideration of merit, with no strings attached, with no expectation of return, and it is a love that motivates supportive action in the one who loves,” – with 5 questions ranking unconditional love for self, other humans, animals, laptop on which the survey was given, and androids; 3) the extent to which they felt they were described by each of the 16 adjectives on the Brief Mood Introspection Survey (BMIS) [25], and 4) three other brief questions that we do not examine here.

After completing the questionnaire, each participant was seated in front of the android technology. In the case of the robot, the robot’s torso was dressed in a simple shirt during the interactions. In the case of the avatar, only a head was seen on a gray background. After video recording was started, the experimenters left the room and closed the door and the participant began the conversation by saying “hello” to the android technology. If participants chose to be guided in meditations, they were guided with collaboratively written scripts based on concepts and exercises in the iConscious human development model [26]. At the end of the duration set aside for the conversation, the experimenters re-entered the room and asked the android technology to finish up. After closing remarks from the robot or avatar, the participant was asked to complete the post-interaction questionnaire, which was the same as the pre-interaction questionnaire. Participants were debriefed in a 5-10 minute on-camera interview with the experimenters. Finally, participants were asked to consider signing a video release form and were told that they were free to not sign the form.

c) Procedures specific to Experiment 1. Before completing the pre-interaction questionnaire, each participant was fitted with a Polar H7 heart-rate variability (HRV) monitoring chest strap and Bluetooth was used to synchronize the signal as input to the HRV Logger mobile application. The chest strap was worn throughout the experiment and removed after the participant completed the post-interaction questionnaire. The sight line for the robot was kept at the participant’s eye level, so the participant could look directly into the robot’s eyes. The duration of each conversation was capped at 15 minutes, and throughout the experiment a videographer recorded both the robot and participant; thus a videographer was in the room the entire time. The conversation was constrained to a choice of three topics, including a discussion about consciousness and guided awareness meditation, a discussion about emotions and a guided emotion meditation, and a discussion about human uniqueness and a guided uniqueness meditation. Participants could not choose to discuss something outside of these three categories. Experimenters monitored the conversation via a laptop when they were seated outside the room and attempted not to interfere in the conversation even when there were apparent errors. In only one case did they interfere with a conversation when the robot did not move on to a response for longer than 5 minutes. In this case, the experimenters triggered the already-cued response remotely.

d) Procedures specific to Experiment 2. In experiment 2, we did not use HRV monitoring, instead relying on emotion detection as an objective measure. We also set up tripods to record the participants and the android technology, so no one was in the room during the conversations. Due to unavoidable constraints, the robot was set on a table that made her sight line slightly above the participant’s sight line, so the participant had to look up to look into the robot’s eyes. The avatar sight line was a bit lower than the participant’s sight line, so the participant had to look down to look into the avatar’s eyes. The course of the conversations was less constrained than in experiment 1, but of the 36 participants, only four chose to not allow the AI to guide them in meditative practices after some introductory chatting, or to only participate in one or two practices. Participants generally chatted with the android technology for about five minutes, then the AI guided them towards an awareness exercise, after which the participants reported what they experienced. Then the AI guided them toward another meditation exercise, followed by another discussion, and then toward a “deep listening” or “eye-gazing” practice – in which the participant was invited to talk about anything with the android technology or just gaze into its eyes, or both. The duration of each conversation was capped at 25 minutes, and a minimum of 20 minutes was required for inclusion in the analysis (one participant was removed from the analysis because he did not complete at least 20 minutes of conversation). For the first six participants, experimenters monitored the conversation via laptop and baby monitor outside of the experiment room, but for the remaining participants they stopped monitoring, as the technology was working relatively well and they felt compelled to give the participants more privacy. Also, for the first six participants, interruption on the part of the android technology was common; this problem was resolved by a code fix starting with the seventh participant. Despite this difference, there were no clear quantitative differences in emotional responses between the first six and the remaining participants, so all participants are combined in the analysis. Again, in only one case did the experimenters interfere in the conversation by walking into the room and vocally triggering the next response from the android technology when the participant's words were not heard. After this, they left the room again to allow the participant to finish the conversation.

B. *Dependent Variables*

a) Subjective dependent variables. In both experiments we calculated four subjective dependent variables (DVs) from responses on the pre- and post-interaction questionnaires: two love-based variables and two mood-based variables. In terms of love-based variables, we were interested in feelings of love related to self-transcendence, so we ignored responses to the two questions related to loving feelings for family members and the single question about unconditional love for oneself. For each participant, we subtracted the arithmetic mean of the love-related questions and the unconditional-love (UL) related questions on the post-interaction questionnaire and subtracted the same means derived from responses on the pre-interaction questionnaire to create a *love change score* and a *UL change score* for each participant. In terms of mood-based variables, we scored the BMIS according to both arousal and pleasantness parameters as in [25] and subtracted the arousal and pleasantness scores calculated from the post-interaction questionnaire from the same scores on the pre-interaction questionnaire to create an *arousal change score* and a *pleasantness change score* for each participant. We flagged outliers for removal in each of the change scores; these were scores 2.5 times the standard deviation of the sample either above or below the sample mean. No outliers were flagged in experiment 1. One outlier was flagged in experiment 2; a value for the pleasantness change score that was below the sample mean. We include this outlier in the overall results for that experiment but exclude it when showing results for each android modality; the outlier presence or absence did not affect the statistical conclusions in either case. We also used independent video coders to code the conversations as to quality, mirroring, positivity, and other subjective measures, but the results from these analyses are too complex to thoroughly discuss in a brief report. One of the more important results is reported in the next section.

b) Objective dependent and independent variables. In experiment 1, our objective DVs were pre-to-post-interaction change scores derived from the standard deviation of the beat-to-beat intervals of the heart rhythm (SDNN), as well as the low frequency component (LF power; 0.04-0.15 Hz). These were calculated from the HRV time series for each participant using HRV logger software. The purpose of the SDNN change score and the LF change score was to track changes in cognitive load during the experiment, as higher cognitive load is associated with a decrease in both variables [27-28].

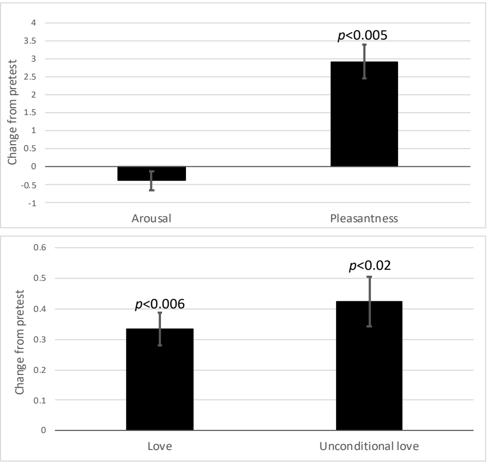

Fig. 2. Experiment 1: Group means of the four subjective change scores (post minus pre) derived from questionnaires before and after the interactions. Error bars show +/- 1 within-participant standard error of the mean (S.E.M.). *p*-values for the three significant paired t-tests are given.

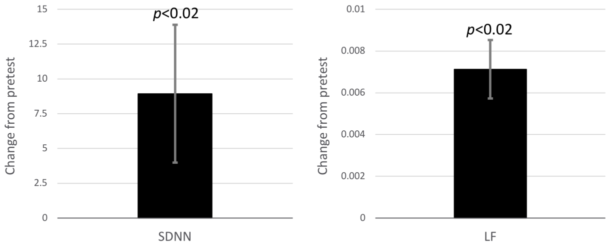

Fig. 3. Experiment 1: Group means of the two objective change scores (post minus pre) derived from HRV data in experiment 1. Error bars show +/- 1 within-participant S.E.M. *p*-values for both significant paired t-tests are given.

c) Experiment 2 also had one set of objective DVs. The set of DVs were derived from the time series of the proportions of each of the seven deep-learning detected emotional states of the participant during the interaction, derived from videos of the participant's face sampled at 60 fps. These are referred to as the participant's raw emotion time series. From the raw emotion time series for each participant, we took the arithmetic mean for each of the seven emotional states, across each minute of the first 20 minutes of the interaction, resulting in 20 means for each participant for each of the seven emotional states. These are referred to as mean emotion time series, one of which would be the "mean anger time series" for a given participant. The purpose of this dependent variable was to track the minute-by-minute emotional changes for each participant to determine if there were significant changes in any emotions, and also to determine whether emotional dynamics would correlate with the love change score or the UL change score. Upon inspection of the grand mean of the participants' raw emotion time series, clear changes in happiness and sadness were observed at the 15.5-minute mark, which corresponded to the average end-time of the second of two meditation practices. We derived a post-hoc dynamic score for happiness and sadness, consisting of the absolute value of the mean emotion score at the 16-minute mark minus the mean of the 15 prior emotion scores, which was used as a baseline. These happiness and sadness dynamic scores were used as independent variables to predict the four subjective DVs.

d) We also calculated another set of independent variables to represent the degree to which the android technology and human participants emotionally mirrored one another. This set was necessarily calculated exclusively for the robot and audiovisual (AV) avatar sessions, as these were the only sessions in which participants had visual contact with the android. We performed linear regressions on each participant's mean emotion time series and the mean emotion time series extracted from videos of the robot or avatar during that participant's interaction. A high correlation (r^2) value would indicate that the participant mirrored the technology's emotional facial expressions relatively consistently during the interaction, and/or that the robot or AV avatar mirrored the human's facial expressions relatively consistently. One r2 value for each of the seven emotional states was calculated for each participant, resulting in what are referred to as emotion-mirroring correlation values. To determine whether the consistency of emotion mirroring was related to any of the self-reported change scores, we investigated whether these emotion-mirroring correlation values for all seven emotional states as a group could predict the four subjective DVs.

IV. IMPACT AND EVALUATION RESULTS

Videos of three different complete participant interactions and debriefing interviews as well as a summary video, all with permission of the participants, are provided in their entirety at [29]. Note that all three participants shown in these videos reported increased loving feelings from before to after their interactions, even though in call cases the interactions contained obvious errors. In [30] we reported results from the early stages of Experiment 1, including several remarkable anecdotes. We will not summarize those anecdotes here, but it is worth noting that at least three participants across the two experiments reported having "transcendent" experiences, and these reports were believable, given a major shift in emotional expression that was apparent to the video coders. Further, several participants commented that their depth of meditation was deeper than usual during the interactions. Intriguing as these reports are, understanding the phenomena experienced by our participants requires determining how objective measures related to participants' subjective experiences.

A. Experiment 1

a) Subjective dependent variables. The group mean of the arousal change score was negative (Fig. 2), indicating that self-reported arousal dropped, on average from pre- to post-interaction; however, this drop was not significant (paired t-test, p>0.40). The group means of the pleasantness, the love and UL change scores were all significantly positive, indicating a group shift toward a more pleasant mood state as well as greater feelings of love and unconditional love, from pre- to post-interaction (paired t-tests, pleasantness: p<0.005, love: p<0.002, unconditional love, p<0.02).

b) Objective dependent variables. The SDNN and LF change scores, derived from the HRV analysis, increased significantly from before to after the interactions, consistent with a decrease in cognitive load (Fig. 3; paired t-tests for both DVs, p<0.02). Relationships between objective and subjective dependent variables were examined in a post-hoc analysis, but they were not easily interpretable and are not discussed here.

B. Experiment 2

a) Subjective dependent variables. Again, the group mean of the arousal change score was negative and not significant (Fig. 4; paired t-test, p>0.35), while the group mean of the pleasantness change score was positive as in experiment 1, but this time it was not significant (paired t-test, p>0.65 without outlier removed, p>0.20 with outlier removed). The group means of the love and UL change scores were significantly positive, again indicating a group shift toward greater feelings of love and unconditional love from pre- to post-interaction (paired t-tests, love: p<0.05, unconditional love, p<0.008).

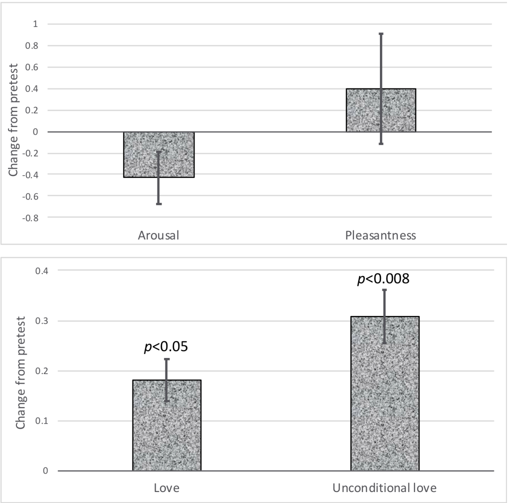

Fig. 4. Experiment 2: Group means of the four subjective change scores (post minus pre) derived from questionnaires before and after the interactions. Error bars show +/- 1 within-participant S.E.M. p-values for the two significant paired t-tests are given.

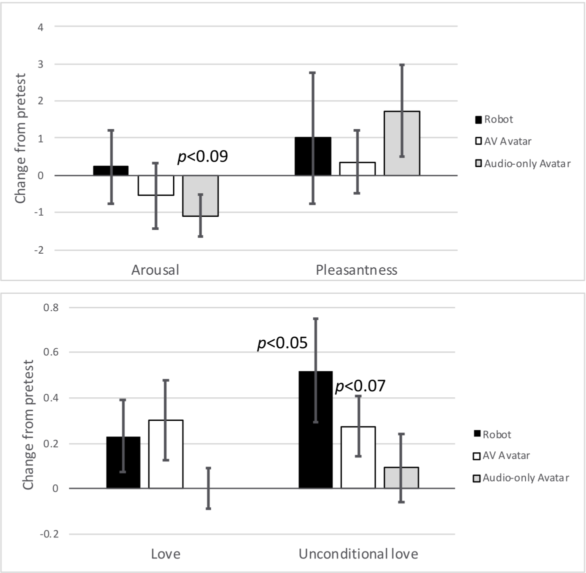

Fig 5. Experiment 2: Group means of the four subjective change scores (post minus pre) derived from questionnaires before and after the interactions, grouped according to interaction type (black = robot, white = AV avatar, gray = audio-only avatar). Error bars show +/- 1 within-participant S.E.M. p-values under 0.1 are given for paired t-tests.

There were no significant group effects across the three conditions (robot, AV avatar, audio-alone avatar) for the four subjective dependent variables. However, as shown in Fig. 5, for the robot and AV avatar interactions, the love change scores were positive and nonsignificant and the UL change scores were either significant or borderline significant (robot: $p<0.05$, AV avatar: $p<0.07$), while the audio-only avatar change scores were either flat (love) or minimally positive (UL) for these measures.

b) Objective dependent variables and predictions of subjective dependent variables. Grand means for happiness and sadness raw time series are shown in Fig. 6 to illustrate the dynamics occurring at ~15.5 minutes, corresponding to the point at which, on average, participants are opening their eyes at the end of the second meditation; grand means for other emotion raw time series are shown in the Appendix. The mean emotion time series for anger and disgust showed a significantly negative time course, while surprise showed a negative time course that was borderline significant. In contrast, the mean sadness time series showed a significantly positive time course (repeated-measures ANOVAs across 20 time points, anger: $p<0.000002$; disgust: $p<0.03$; surprise: $p<0.075$; sadness: $p<0.005$). There was an average decline in fear, but this was not significant. These results indicate mixed changes in emotion during the interactions that will be discussed in a later section (V).

Happiness and sadness dynamic scores captured the changes in these two emotions apparent at the 15.5-minute mark (calculation described in section IIIB*c*). Our goal was to determine if together these scores could predict any of the four subjective DVs. The only subjective DV the two scores predicted was the love change score (Fig. A4, multiple linear regression with two predictors, $r^2=0.311$, $p<0.003$; sadness: $t=3.65$, $p<0.001$, happiness: $t=-3.15$, $p<0.004$). This prediction survives Bonferroni correction for the four prediction attempts, indicating a clear relationship between peak changes in emotional state during the interaction and changes in loving feelings from before to after the interaction. When interaction type was included in the model there was a borderline significant group effect ($p<0.09$), indicating no robust difference between interaction groups, but independent models for each interaction type revealed significant predictions of love change scores for both the robot and AV avatar groups but not the audio-only group (multiple linear regressions with two predictors, robot: $r^2=0.482$, $p<0.03$, AV avatar: $r^2=0.713$, $p<0.007$, audio-alone: $r^2=0.048$, $p>0.80$), with the estimates for audio-alone reversed in sign relative to those for the other two conditions, indicating the happiness and sadness dynamic scores are not functioning as predictors in the same way in this condition as they are for the two conditions that provide visual contact with the android.

Further supporting the usefulness of the love change score is the relationship between the results of video coding by two independent coders and the love change scores. Factors on which the coders were in quantitative agreement significantly predicted love change scores (multiple linear regressions with seven predictors, $r^2=0.432$, $p<0.021$). These factors were: the extent to which participants 1) copied the movements of the android, 2) fell into a conversational rhythm with the android, 3) maintained eye contact, 4) participated in the meditation activities, 5) asked the android personal questions, 6) felt the interaction was longer than it was, 7) were judged to have a positive connection with the android.

Finally, emotion-mirroring correlation values for all seven emotions (calculation described in section IIIB*d*) significantly predicted the pleasantness change scores among the four subjective DVs (Fig. A5). This prediction survived Bonferroni correction for the four attempted predictions (multiple linear regressions with seven predictors, $r^2=0.676$, $p<0.008$; significant predictors were fear: $t=-3.25$, $p<0.006$ and surprise: $t=2.63$, $p<0.02$), indicating a robust relationship between deep-learning detected emotional mirroring and participants' change in pleasantness from before to after the interactions for participants in the two conditions providing visual contact with the android.

V. Problem impact evaluation

Overall, the data obtained in both experiments confirm most of our hypotheses. Hypothesis 1: Self-reported loving feelings related to self-transcendence did indeed significantly increase from pre- to post-interaction in both experiments, confirming this hypothesis. Hypothesis 2: Self-reported positive mood did increase, on average, from pre- to post-interaction in both experiments, but this shift was only significant in the first experiment, partially confirming this hypothesis. Hypothesis 3: Self-reported arousal did decrease, on average, from pre- to post-interaction in both experiments, but this change was not significant, leaving this hypothesis unconfirmed. Hypothesis 4: Heart rate variability measures taken in experiment 1 significantly increased from pre- to post-interaction, suggesting cognitive load declined during the conversations and confirming this hypothesis. Hypothesis 5: Feelings of anger, fear and disgust measured using a deep-learning emotion detection network in experiment 2 decreased during the interactions, but this decrease was only significant for anger and disgust, partially confirming this hypothesis. Hypothesis 6: Dynamic changes in deep-learning-detected happiness and sadness in experiment 2 predicted love change scores, and the dynamics of emotional mirroring predicted pleasantness change scores, confirming this hypothesis. Hypothesis 7: Of the five hypotheses that were applicable to experiment 2, hypothesis 1 and 6 were significant only for conditions in which participants had visual contact with the robot or AV avatar, and not in the audio-alone condition, partially confirming this hypothesis that only some effects would be borne out in the audio-alone condition.

Our results suggest two major conclusions. First, brief, guided conversations and awareness exercises shared with a humanoid robot or its audiovisual avatar, controlled by emotionally responsive AI, are correlated with increases in subjective feelings related to self-transcendence in human participants – specifically, loving feelings for people beyond one's own immediate family as well as unconditionally loving feelings for other humans, animals, and inanimate objects including robots and avatars themselves. Both subjective and objective measures taken in experiment 2 strongly suggest that androids with visual aspects presented to the participants are more effective. This result, along with the emotion-detection dynamics that predicted loving feelings in experiment 2, suggests that the effects obtained in the robot and audiovisual avatar conditions were due to the technology itself, rather than response bias or experimenter style.

Second, objective measures such as changes in heart rhythms and emotional states detected by a deep-learning network indicate a complex array of transformations occur during these conversations, and importantly also suggest that the self-reported, subjective measures have objective counterparts. In experiment 1, two measures related to heart rate variability, the timing of the heartbeat intervals (SDNN) and the low frequency portion of the frequency spectrum calculated from the heartbeat time series (LF) both increased from pre- to post-interaction, suggesting that cognitive load prior to the interaction was higher than afterwards. This is unsurprising, and it can be assumed that this change reflects a reduction in cognitive load that occurred once participants completed the conversation, an assumption that is supported by a gradual but nonsignificant decrease in these same variables, on average, over the course of the conversations (data not shown). If the conversations had been cognitively stressful, however, this change could have gone in the opposing direction. So the tentative conclusion here is that the conversations were generally not stressful, but we cannot conclude that the change in heart rhythms related to the technology itself.

In our future experiments we plan to add the validated Adult Self-Transcendence Inventory, as used in [10-12], as well as the loving-feelings questionnaire, independent video coders, and a deep-learning network related to the one used here. As demonstrated already, we will use these converging methods to determine whether participants experience significant: 1) increases in feelings of love for others and self-transcendence, 2) decreased anger and disgust, and 2) emotional dynamics that predict changes in feelings of love and self-transcendence during the interactions.

VI. Speculation

Compelling and informative effects were obtained in experiment 2 from the deep-learning network. Changes in the emotion mean time series data, the happiness and sadness dynamic scores, and the emotion-mirroring correlation values all provided face validity to the usefulness of the deep-learning network itself. The significant decreases in anger and disgust over the course of the interaction were predicted, but the significant increase in sadness was not. It could be that, because neutral expressions and sadness were often confused by the deep-learning network (Fig. 1), increases in sadness actually reflected increases in neutrality, which might be expected following successful meditations. It is also possible that the increase in detected sadness reflected a type of sadness that is consistent with openness to one's own internal experience. This speculative interpretation is perhaps strengthened by the fact that a bigger change in happiness predicted reduced love change scores, while a bigger change in sadness predicted increased love change scores, perhaps suggesting that sadness dynamics at peak emotional states may be important for inducing self-transcendent experiences.

Emotion-mirroring correlation values predicted self-reported changes in pleasantness, suggesting that the facial emotional synchrony between human and android was key to the humans' experience of the pleasantness of the interaction. Interestingly, the surprise predictor was positive while the fear predictor was negative, indicating that higher consistency in mirroring surprised expressions predicted positive changes in pleasantness, while lower consistency in mirroring fearful expressions predicted positive changes in pleasantness. One implication is that mirroring dynamics can be tuned to be less consistent in mirroring fearful expressions in cases where a more pleasant outcome is desired. However, the effect of this change on self-transcendent feelings is uncertain.

VII. Next Steps

Advancing toward our grand challenge requires improvements in the following areas:

1. *Flexibility, stochasticity, and richness in our rule-based emotion and motivation model, OpenPsi.* We are working toward allowing a broader range of urge weights, a more complex model of motivation, and stochasticity in urges so that unpredictability and flexibility will be hallmarks of our future dialogues. We also plan to add the ability to clarify a participant's intentions when they are not clear, and to respond to emotion-detection information dynamically.

2. *Improvements in our emotion-detection deep-learning network.* Re-training the model with more examples of emotions that were undersampled (e.g., disgust), increasing the number of ethnicities sampled, and increasing the number of datasets used will improve our emotion-detection deep-learning network.

3. *Inclusion of a deep-learning network trained on emotional voice information and emotional vocal tone.* Obtaining accurate emotional information from voice will help manage situations when the participant is not in view of the camera, as well as add richness to our emotional mirroring protocol, assuming we also add emotionally appropriate changes in the android's vocal tone.

4. *Development and integration of a relationship model.* We think part of the problem with guiding flexible dialogue between two people is that a relationship model – an understanding of the emotions, motivations, and goals of each person in the relationship – is not usually used to guide the conversation. We hope to remedy this problem by collaborating with social neuroscientists to develop and integrate a dynamic relationship model that informs OpenCog Openpsi throughout the dialogues.

VIII. IRB review status

Both experiments reported here were pre-approved by an IRB (Fig. A6 and A7). All future experiments will also be pre-approved by an IRB. All IRB requirements were followed.

Acknowledgment

In addition to our generous funders, we thank Ralf Mayet for leading the development team, Tesfa Yohannes and Mitiku Yohannes for training the deep-learning network, the OpenCog coding community for coding OpenPsi and related code, Carole Griggs and Ted Strauss for invaluable work writing the meditations and conversations used in the interactions, and Gino Yu for providing laboratory space and access to the Hong Kong Polytechnic IRB.

References

[1] A.S. Chulef, S.J. Read, and D.A. Walsh, “A hierarchical taxonomy of human goals,” Motiv. and Emo., vol. 25, pp. 191-232, 2001.

[2] M.E. Koltko-Rivera, (2006), “Rediscovering the later version of Maslow's hierarchy of needs: self-transcendence and opportunities for theory, research, and unification,” Rev. Gen. Psych, vol. 10, pp. 302-317, 2006.

[3] A.H. Maslow, “Critique of self-actualization theory,” In E. Hoffman (Ed.), “Future visions,” pp. 26– 32, London: Sage, 1996.

[4] D. O'Connor and L. Yballe, “Maslow revisited,” J. Manag. Ed., vol. 31, pp. 738-756, 2007.

[5] C.R. Cloninger, “The science of well-being,” World Psych., vol. 5, pp. 71-76, 2006.

[6] M.R. Levenson, C.M. Aldwin, and A.P. Cupertino, “Transcending the self,” Matur. & Velhice, pp. 99-116, 2001.

[7] M.R. Levenson, P.A. Jennings, C.M. Aldwin, and R.W. Shiraishi, “Self-transcendence,” Int. J. Aging Hum. Dev., vol. 60, pp. 127-143, 2005.

[8] D.D. Coward, “Self-transcendence and emotional well-being in women with advanced breast cancer,” Onc. Nurs. Forum, vol. 18, pp. 857-863, 1999.

[9] K.A. MacLean, M.W. Johnson, R.R. Griffiths, “Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness,” J. Psychopharm., vol. 25, pp.1453-1461, 2011.

[10] J.J. Runquist and P.G. Reed, “Self-transcendence and well-being in homeless adults,” J. Hol. Nurs., vol. 25, pp. 5-13, 2007.

[11] C. Vieten, M. Estrada, A.B. Cohen, D. Radin, M. Schlitz, and A. Delorme, “Engagement in a community-based integral practice program enhances well-being,” Int. J. Transpers. Stud., vol. 33, pp. 1-15, 2014.

[12] K. Lynch, “Meditation gone mobile,” 2016, unpublished dissertation.

[13] https://cloud.google.com/speech-to-text/

[14] D. Hart and B. Goertzel, “Opencog,” AGI, pp. 468-472, Feb. 2008.

[15] J. Bach, “The micropsi agent architecture,” In Proc. ICCM-5, Int. Cong. Cog. Model., pp. 15-20, April 2003.

[16] P.F. Ferrari and G. Coudé, “Mirror Neurons, Embodied Emotions, and Empathy,” in “Neuronal Correlates of Empathy,” pp. 67-77, New York: Academic Press, 2018.

[17] S. Hurley, “The shared circuits model (SCM),” Behav. Brain Sci., vol. 31, pp. 1-22, 2008.

[18] https://github.com/elggem/ros_people_model/blob/master/scripts/mirroring.py

[19] T. Kanade, J.F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” Proc. 4th IEEE Int. Conf. Auto. Face and Gesture Recog., (FG'00), pp. 46-53, 2000.

[20] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+),” Proc. 3rd Int. Workshop on CVPR for Human Comm. Behav. Analysis (CVPR4HB 2010), pp. 94-101, 2010

[21] I. Goodfellow, D. Erhan, P-L. Carrier, A. Courville, M. Mirza, B. Hamner et al., “Challenges in representation learning,” ICML Workshop on Challenges in Represent., 2013, arxiv.org/abs/1307.0414.

[22] https://github.com/mitiku1/Emopy-Models

[23] P. Ekman and W. Friesen, “Facial action coding system,” Palo Alto: Consulting Psychologists Press, 1978.

[24] D.E. King, “Dlib-ml,” J. Mach. Learn. Res., vol. 10, pp. 1755-1758, 2009.

[25] Mayer, J. D., & Gaschke, Y. N. (1988). The experience and meta-experience of mood. *Journal of personality and social psychology*, *55*(1), 102.

[26] C. Griggs and R. Strauss, “Accelerating conscious human development using the iConscious model as an integrative framework,” unpublished.

[27] N. Hjortskov, D. Rissén, A.K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, “The effect of mental stress on heart rate variability and blood pressure during computer work,” Eur. J. Appl. Phys., vol. 92, pp. 84-89, 2004.

[28] G. Mulder and L.J. Mulder, “Information processing and cardiovascular control,” Psychophys., vol. 18, pp. 392-402, 1981.

[29] https://drive.google.com/drive/folders/1O6FEtFayWgM2DZYAv0mM3t7yaenctUgo?usp=sharing

[30] B. Goertzel, J. Mossbridge, E. Monroe, D. Hanson, and G. Yu, G, “Loving AI,” arXiv:1709.07791 [cs.AI], 2017.

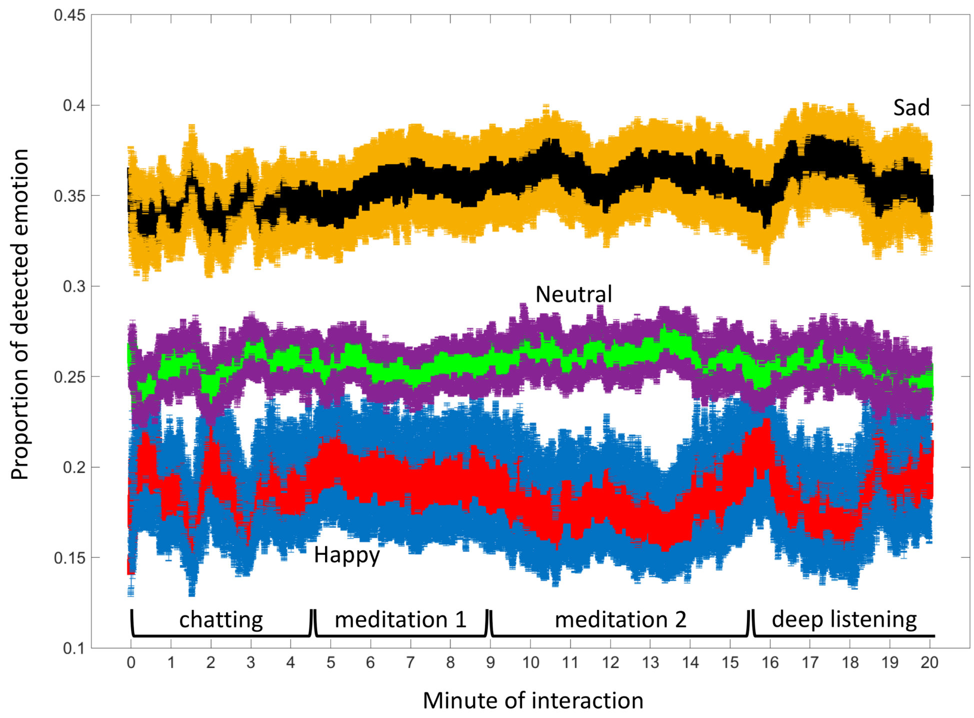

Fig 6. Experiment 2: Grand means (central traces) and between-participant +/- SEM (upper and lower borders) for sad (upper trace), neutral (middle trace) and happy (lower trace) emotions detected with the deep-learning network. Note the inflections, especially for happiness and sadness, at around 15.5 minutes, which is the average time at which participants open their eyes after the second meditation.