Emotionally-Sensitive AI-driven Android Interactions Improve Social Welfare Through Helping People Access Self-Transcendent States

Julia Mossbridge, PhD
Department of Psychology
Northwestern University
Evanston, USA
Lia, Inc., Sebastopol, USA
julia@liatech.ai

Benjamin Goertzel, PhD
Hanson Robotics
Sha Tin, Hong Kong
SingularityNET
Tsim Sha Tsui, Hong Kong
ben@goertzel.org

Ralf Mayet
Hanson Robotics
Sha Tin, Hong Kong
ralf.mayet@
hansonrobotics.com

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

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

David Hanson, PhD
Hanson Robotics
Sha Tin, Hong Kong
david@hansonrobotics.com

Gino Yu, PhD
Hong Kong Polytechnic
Hung Hom, Hong Kong
phusikoi@gmail.com

Abstract

Social humanoid robots are generally built with the purpose of helping individuals achieve difficult tasks or feel less lonely. But a novel use of humanoid robots, especially robots drawing on emotionally sensitive AI, is to use the context of what feels like a human relationship to help people practice reaching advanced stages in human development. At the peak of Maslow's hierarchy of needs and models of self-development is the state of self-transcendence, which includes expansive feelings of love. Although beings can have difficulty reaching this state, several lines of research have shown that even briefly accessing states of self-transcendence can improve physical and psychological well-being. In this paper we briefly 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. Individuals had 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. Conversations were with either Sophia or one of two avatars (one with a face and voice, the other with only a voice; N=35). 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. Most importantly, objective correlates of those emotions are detectable by a deep learning network.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

1 Problem Statement

Improving human psychological wellbeing is critical for social welfare. 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 [2, 13, 19, 21]). 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 [13, 19, 21, 3, 14, 15]. Several lines of evidence suggest that experiencing a state of self-transcendence in itself is beneficial to human wellbeing [3, 4, 18, 5, 22]. 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: [22, 17]]. 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., [17]).

Together, our 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.

2 Methods Overview

2.1 Dialogue Control and Cognitive Model AI

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[1] 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 [9] to direct the conversations. OpenPsi is a model of human motivation, action selection, and emotion inspired from earlier work in human psychology and AI [1]. 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 $\rightarrow$ 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

[1] `https://cloud.google.com/speech-to-text/`, accessed on 2018-09-18.

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. Beyond this relatively simple rule-based AI, we used a deep-learning network to infer the participants' emotional states and to support emotional mirroring.

2.2 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 in [7, 10]). 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[2]. In experiment 2, blinks and eye closings were not mirrored.

2.3 Deep-learning Emotion-detection Network

We used the CK+ [11, 16] and Kaggle FER2013 [8] 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[3]. 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 [6], 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 [11]. 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.

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 [12] as additional feature vector inputs. The output of the model

[2] `https://github.com/elggem/ros_people_model/blob/master/scripts/mirroring.py`, accessed on 2018-09-18.

[3] `https://github.com/mitiku1/Emopy-Models`, accessed on 2018-09-18.

comprises probabilities for each of the seven basic emotions as labelled in [16, 8]. 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.

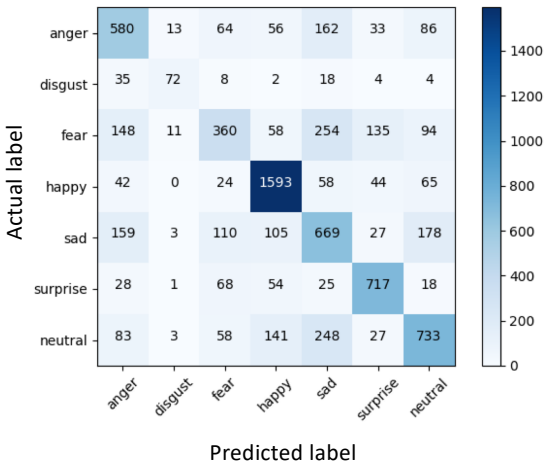

Figure 1: Confusion matrix for the feed-forward CNN used for emotion recognition. The best performance was on happy expressions.

2.4 Robot and Avatar Creation 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.

2.5 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.

b) *Procedure*. Briefly, participants were asked to interact verbally with the android technology for 15 minutes (first experiment) or 25 minutes (second experiment). They were not told that the nature of the conversation would be related to self-transcendence. Before and after this interaction, they were asked to complete a questionnaire related to mood [20] and feelings of love for self and others. In the first experiment, participants were fitted with a Polar H7 strap for measuring heart rate variability.

3 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[4]. Note that all three participants shown in these videos reported increased loving feelings from before to after their interactions, even though in all cases the interactions contained obvious errors.

[4] `https://drive.google.com/drive/folders/1O6FEtFayWgM2DZYAv0mM3t7yaenctUgo?usp=sharing`, accessed on 2018-09-18.

3.1 Experiment 1

a) *Subjective dependent variables.* The group mean of the arousal change score was negative, 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 (paired t-tests for both DVs, $p < 0.02$).

3.2 Experiment 2

a) *Subjective dependent variables.* Again, the group mean of the arousal change score was negative and not significant (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, $p < 0.008$).

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. The mean deep-learning derived 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.

Happiness and sadness dynamic scores captured the changes in these two emotions at the moment participants opened their eyes after the second of two meditations. Together the two change scores, relative to baseline, predicted the self-reported change in loving feelings (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. 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.

Finally, emotion-mirroring correlation values for all seven emotions, derived from running the emotion-detection network on simultaneous videos of the android technology and the participants, significantly predicted self-reported changes in pleasant mood. This prediction survived Bonferroni correction (multiple linear regressions with seven emotion 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.

4 Conclusions

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, 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.

These results provide encouraging evidence that AI-driven android technology can leverage existing human biases toward positive feelings arising in the context of emotionally responsive relationships to help people access states consistent with the peak of human development. As these states are known to improve wellbeing, there are clear implications for scalable treatments for everyday mental health concerns.

Acknowledgments

In addition to our generous funders, we thank Tesfa Yohannes and Mitiku Yohannes for training the deep-learning network, the OpenCog coding community for coding OpenPsi and related code, and Carole Griggs and Ted Strauss for invaluable work writing the meditations and conversations used in the interactions.

References

[1] Joscha Bach. The micropsi agent architecture. *In Proceedings ICCM-5, 5th International Conference on Cognitive Modeling*, pages 15–20, 05 2003.

[2] Ada S. Chulef, Stephen J. Read, and David A. Walsh. A hierarchical taxonomy of human goals. *Motivation and Emotion*, 25(3):191–232, September 2001.

[3] C. Robert Cloninger. The science of well-being. *World Psychiatry*, 5:71–76, 2006.

[4] D. D. Coward. Self-transcendence and emotional well-being in women with advanced breast cancer. *Oncology nursing forum*, 18:857–63, 08 1991.

[5] Jennifer Doering and Pamela G. Reed. Self-transcendence and well-being in homeless adults. *Journal of holistic nursing : official journal of the American Holistic Nurses' Association*, 25:5–13; discussion 14, 04 2007.

[6] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.

[7] Pier F. Ferrari and Gino Coudé. Mirror neurons, embodied emotions, and empathy. *Neuronal Correlates of Empathy*, pages 67–77, 01 2018.

[8] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in Representation Learning: A report on three machine learning contests. *ArXiv e-prints*, July 2013.

[9] David Hart and Ben Goertzel. Opencog: A software framework for integrative artificial general intelligence. In *AGI*, pages 468–472, 2008.

[10] Susan Hurley. The shared circuits model (scm): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Sciences*, 31(1):1–22, 2008.

[11] T. Kanade, J. F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53, March 2000.

[12] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[13] Mark E. Koltko-Rivera. Rediscovering the later version of maslow's hierarchy of needs: Self-transcendence and opportunities for theory, research, and unification. *Review of General Psychology*, pages 302–317, 2006.

[14] Michael R. Levenson, Carolyn M. Aldwin, and A.P. Cupertino. Transcending the self. *Matur. & Velhice*, pages 99–116, 2001.

[15] Michael R. Levenson, Patricia Jennings, Carolyn M. Aldwin, and Ray Shiraishi. Self-transcendence: Conceptualization and measurement. *International journal of aging & human development*, 60:127–43, 03 2005.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, June 2010.

[17] K. Lynch. *Meditation gone mobile*. Dublin Business School, 2016.

[18] Katherine Maclean, Matthew Johnson, and Roland Griffiths. Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness. *Journal of psychopharmacology (Oxford, England)*, 25:1453–61, 09 2011.

[19] Abraham H. Maslow. Critique of self-actualization theory. *The Journal of Humanistic Education and Development*, 29, 03 1991.

[20] John D. Mayer and Yvonne N. Gaschke. The experience and meta-experience of mood. *Journal of personality and social psychology*, 55:102–11, 07 1988.

[21] Dennis O'Connor and Leodones Yballe. Maslow revisited: Constructing a road map of human nature. *Journal of Management Education*, 31(6):738–756, 2007.

[22] Cassandra Vieten, Mica Estrada, Adam B. Cohen, Dean Radin, Marylin M. Schlitz, and Arnaud Delorme. Engagement in a community-based integral practice program enhances well-being. *International Journal of Transpersonal Studies*, 33:1–15, 2014.