AFRL-AFOSR-VA-TR-2019-0102

Social Function of Emotions Modeling and Exploiting the Social Function of Emotions in Mixed Human-Machine Teams

Jonathan Gratch
UNIVERSITY OF SOUTHERN CALIFORNIA
3720 S FLOWER STREET, THIRD FLOOR
LOS ANGELES, CA 90007

04/16/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/RTA2

Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
16-04-2019	Final Performance	30 Sep 2014 to 29 Sep 2018

4. TITLE AND SUBTITLE
Social Function of Emotions Modeling and Exploiting the Social Function of Emotions in Mixed Human-Machine Teams

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-14-1-0364

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Jonathan Gratch

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF SOUTHERN CALIFORNIA
3720 S FLOWER STREET, THIRD FLOOR
LOS ANGELES, CA 90007 US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2019-0102

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Advances in autonomy raise the potential for rich partnerships between humans and machines. This proposal examines the potential costs and benefits of incorporating human-like social and emotional capabilities into machine teammates in potentially high-stakes situations. The research has made progress in three areas: Mind perception theory (illustrating that the extent to which one responds socially to autonomy depends on the 'mind' that is attributed to the machine, with people categorizing other minds in terms of agency/intelligence and experience/emotion); Representation effects (illustrating that people act differently towards others when they interact with them through an autonomous agent that represents their own interests); and Predictive Neural models (showing that there are neural signatures that predict these attributions and social responses).

15. SUBJECT TERMS
Human-machine teaming, Autonomy, Robotics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			RIECKEN, RICHARD
Unclassified	Unclassified	Unclassified	UU		19b. TELEPHONE NUMBER *(Include area code)* 703-941-1100

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Grant Number: AFOSR FA9550-14-1-0364

Project Title: Modeling and exploiting the social function of emotions in mixed human-machine teams

Start Date of Project: September 30, 2014

End Date of Project: September 29, 2018

ICT Technical Point of Contact: Jonathan Gratch

Final Technical Report

Proposal Title: Modeling and exploiting the social function of emotions in mixed human-machine teams

Technical Point of Contact: Jonathan Gratch, 310-440-0306, gratch@ict.usc.edu

Date: November 26, 2018

Abstract: Advances in autonomy raise the potential for rich partnerships between humans and machines. This proposal examines the potential costs and benefits of incorporating human-like social and emotional capabilities into machine teammates in potentially high-stakes situations. The research has made progress in three areas: Mind perception theory (illustrating that the extent to which one responds socially to autonomy depends on the "mind" that is attributed to the machine, with people categorizing other minds in terms of agency/intelligence and experience/emotion); Representation effects (illustrating that people act differently towards other when they interact with them through an autonomous agent that represents their own interests); and Predictive Neural models (showing that there are neural signatures that predict these attributions and social responses).

Project Goals: Advances in autonomy raise the potential for rich partnerships between humans and machines. Human-robot teams are bound to emerge across a range of potentially high-stakes situations including military operations, first-responders and caring for vulnerable populations. Such situations can be challenging even for pure-human teams as they can involve multiple competing concerns (such as tradeoffs between material and moral priorities) and combine a mixture of mundane low-stakes decision-making punctuated by moments of high-consequence decisions. Recent scholarship has emphasized the important and under-researched role that social and emotional factors can play in determining human-team effectiveness in such situations. This proposal examines the potential costs and benefits of incorporating human-like social and emotional capabilities into machine teammates in potentially high-stakes situations.

More specifically, the object of this proposal is to 1) unpack the broad concept of "human-like" into a taxonomy of specific mechanisms that shape human team outcomes (e.g., envy that a teammate disproportionately benefits from an outcome) and the machine-traits that might enhance or mitigate these mechanisms (e.g., is the machine able to accrue material benefits?); 2) build computer teammates that exhibit these traits across a range of standard social decision-making tasks; and 3) conduct a series of empirical studies, (initially with scripted machine teammates then transitioning to model-driven teammates) that examine how different machine-traits impact team outcomes across a range of social situations; and 4) build towards a general theoretical framework that could provide a firm basis for

designing human-machine teams for specific classes of situations. In accomplishing these aims, we build on our prior AFOSR-funded results related to the role emotion processes play in human-machine decision-making.

Primary Accomplishments and experimental findings: The research performed as part of this grant served to advance understanding of how the incorporation of human-like characteristics into machines shapes behaviors, perceptions and outcomes in mixed human-machine teams. Specifically, the grant sought to answer the following questions:

- Do people treat machines "as if" they interacting with a human?
- Do they treat them differently?
- Does this change if the machine includes "human-like" traits
- And ultimately, does this benefit or harm human-machine collaboration?

The grant approached these questions through three interrelated tasks. The first task examined Mind Perception Theory (Gray, Gray and Wegner, 2007) as a possible mechanism, on the one hand to explain differences in how people treat different machine teammates and, on the other hand, to point to manipulations that can systematically shape people's behavior towards machines. The second task, referred to as Representation Effects, examined how behavior towards machine teammates is shaped, not only by inferences about the mind of the machine, but also by beliefs of who the machine represents. Except in science fiction, machines are not truly autonomous but rather serve the needs of another human user or human-controlled organization. In other words, a machine acts as a representative for another person, and acts to fulfill the needs and goals of its "client". This task examines how the salience of the person-behind-the-machine shapes outcomes in the human-machine team. Finally, the third task examined neural correlates of the effects observed in these other two tasks.

Mind Perception Theory: The research examined the impact of both shallow and "deep" traits. An example of a shallow trait is adding facial expressions or providing a backstory that the agent has advanced mental capabilities. Deep traits involve programming the machine to act as though it possesses emotions or advanced mental capabilities. The research findings emphasize the mediating role of mind perception theory of Gray, Gray and Wegner (2007). Mind perception theory argues that, regardless of whether deep or shallow traits are manipulated, the impact of these traits depends on the extent that it shapes inferences concerning the mental capacities of the machine. These capacities are broadly organized into two dimensions: agency (the ability to think and execute free will) and experience (the ability to experience pain and emotion). Based on where a machine falls in this two-dimensional space, it will elicit different perceptions and behavior. For example, simple machines are assumed to possess neither agency nor experience. Users will tend to have low regard for the concerns/needs/goals of such a machine, but neither will they tend to blame the machine when it causes them harm. In contrast, a machine with high agency and low experience will be assigned blame, but again, users will tend to disregard the machines needs and goals (e.g., see de Melo and Gratch, 2015). Shallow traits will tend to have a stronger and more predictable impact on mind-perceptions, but these impressions will be extinguished over time if the agents "deep" behavior fails to be consistent with these surface impressions.

Representation effects: Human-machine interaction is typically conceived as a person interacting with an autonomous machine, but (unless and until machines become truly independent of human oversight), there is always a person or organization behind the machine, for which the machine is a representative. Research in this task examined how decision making is altered when these representatives are salient. Research examined these effects from two perspectives. On the one hand, we considered how people treat a machine that represents the interests of another person. Broadly speaking, people do not consider the needs/goals of a machine: then tend to exploit the machine to maximize their own goals and feel little willingness to employ human social norms like fairness. However, when people come to believe that the machine is serving the interests of another person, these self-interested effects are mitigated (although not completely eliminated). On the other hand, we consider how people "program" their machines to represent their own interests with other human actors. Our initial assumption was that people would task their machines to be less fair and hold lest regard for others (compared with when they interact with others directly) as the other person would be less salient. In contrast, in a series of studies, we found that people program their machines to exhibit higher regard for others than these people would who if they were to interact with others directly (e.g., see de Melo, Marsella and Gratch, 2018). This has been validated across several tasks including economic games and self-driving cars. The mechanism behind these findings is unclear and uncovering them serves as a focus of a subsequent AFOSR-funded MINERA grant (Organizational Implications of Autonomy-Mediated Interaction) when Current results show, contrary to initial expectations, that people are more fair when interacting through representatives.

Predictive neural models: Whereas the previous studies focus on behavioral effects, here we examine the neural underpinnings of different behaviors shown in human-agent interactions. In a human-machine negotiation task, were able to find distinct brain patterns in emotion-related regions of the brain when people interacted with what they believed was a human versus agent opponent. Further, we show that it is possible to predict whether the negotiator concedes, does not change, or asks for more during the negotiation. Results may yield insights into the cognitive underpinnings of why people behave differently towards machine teammates. This has been recently replicated in a second study accepted to cognitive science.

Staff supported under grant: The following personal were supported under the period of performance of the grant

- Faculty:
  - Jonathan Gratch (USC CS, Psychology)
  - Stacy Marsella (Northeastern CS, Psychology)
  - Peter Carnevale (USC Business, Psychology, Communications)
  - Morteza Dehghani (USC CS, Psychology)
- Graduate Research Assistants
  - Zahra Nazari (USC)
  - Johnathan Mell (USC)
  - Eunkyung Kim (USC)

    - Yuyu Xu (Northeastern)
- Visiting student researchers
    - Celso de Melo
    - Rens Hogen (University of Twente)
    - Bexy Alfonso (Universitat Politècnica de València)
    - Lilly Marie Leisse (University of Duisburg-Essen)
- Research Staff
    - Jill Boberg
    - Alesia Gainer
    - Sharon Mozgai
    - Rachel Wood
- External (unfunded) collaborators
    - Brian Parkinson (Oxford University)
    - Anthony Manstead (Cardiff University)
    - Ion Juvina (Wright State University)
    - Milind Tambe (USC)

Grant related publications :

- Johnathan Mell, Gale Lucas and Jonathan Gratch. Welcome to the real world: How agent strategy increases human willingness to deceive. *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*. Stockholm, Sweden 2018
- Celso de Melo, Stacy Marsella and Jonathan Gratch. Social decisions and fairness change when people's interests are represented by autonomous agents. *Journal of Autonomous Agents and Multiagent Systems* 32(1), 2018. pp. 163-187
- Saba Khashe, Gale Lucas, Burcin Becerik and Jonathan Gratch. Buildings with persona: Towards effective building-occupant communication. *Computers in Human Behavior*, v. 75. 2017. pp. 607-618
- YuYu Xu, Pedro Sequeira, and Stacy Marsella Marsella. Towards modeling agent negotiators by analyzing human negotiation behavior. *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017.
- Eunkyung Kim, Jared Gilbert, Charlotte Horowitz, Jonathan Gratch, Jonas Kaplan, and Morteza Dehghani. Decoding Partner Type in Human-Agent Negotiation using functional MRI. *39th Annual Meeting of the Cognitive Science Society*. London 2017
- Eunkyung Kim, Sarah Gimbel, Aleksandra Litvinova, Jonas Kaplan, and Morteza Dehghani. Decoding Virtual Agent's Emotion and Strategy from Brain Patterns. *39th Annual Meeting of the Cognitive Science Society*, London. 2017
- Dan Feng, David Jeong, Nicole Krämer, Lynn Miller and Stacy. Marsella. Is It Just Me?: Evaluating Attribution of Negative Feedback as a Function of Virtual Instructor's Gender and Proxemics. *16th Conference on Autonomous Agents and Multiagent Systems*, Sao Paulo, Brazil 2017
- Celso de Melo, Stacy Marsella and Jonathan Gratch. Increasing Fairness by Delegating Decisions to Autonomous Agents. *16th International Conference on Autonomous Agents and Multiagent Systems*, Sao Paulo, Brazil 2017.

- Jonathan Gratch, Peter Friedland and Benjamin Knott. Recommendations for Research on Trust in Autonomy. *5th International Workshop on Human-Agent Interaction Design and Models*. NYC. 2016
- Celso de Melo, Stacy Marsella and Jonathan Gratch. "Do As I Say, Not As I Do:" Challenges in Delegating Decisions to Automated Agents. *15th International Conference on Autonomous Agents and Multiagent Systems*, Singapore, 2016.
- Jonathan Gratch, Peter Friedland and Benjamin Knott. Recommendations for Research on Trust in Autonomy. *5th International Workshop on Human-Agent Interaction Design and Models*. NYC. 2016.
- Eunkyung Kim, Sarah Gimbel, Aleksandra Litvinova, Jonas Kaplan, and Morteza Dehghani Predicting decision in human-agent negotiation using functional MRI. *38th Annual Meeting of the Cognitive Science Society*. 2016.
- Ion Juvina, M. Collins, O. Larue and Celso de Melo. Toward a Unified Theory of Learned Trust. In D. Reitter & F. E. Ritter (Eds.), *14th International Conference on Cognitive Modeling (ICCM 2016)*. University Park, PA. 2016
- Jonathan Gratch, David DeVault and Gale Lucas. The benefits of virtual humans for teaching negotiation. *16th International Conference on Intelligent Virtual Agents*. Los Angeles, CA, September 2016
- Celso de Melo and Jonathan Gratch. Beyond Believability: Quantifying the Differences between Real and Virtual Humans, *International Conference on Intelligent Virtual Agents,* Delft, The Netherlands. 2015
- Rens Hoegen. *Human Behavior towards Virtual Humans.* (Masters Thesis), Human Media Interaction, University of Twente. 2015
- Rens Hogens, Giota Stratou, Gale Lucas and Jonathan Gratch. Comparing behavior towards humans and virtual humans in a social dilemma, *International Conference on Intelligent Virtual Agents,* Delft, The Netherlands. 2015
- Nazari, Lucas, Gratch. Opponent Modeling for Virtual Human Negotiators, *International Conference on Intelligent Virtual Agents,* Delft, The Netherlands 2015
- Jonathan Gratch, Stacy Marsella and Lin Cheng. The Appraisal Equivalence Hypothesis:
- Verifying the domain-independence of a computational model of emotion dynamics. *International Conference on Affective Computing and Intelligent Interaction.* Xi'an, China. 2015
- Bexy Alfonso, David V. Pynadath, Margot Lhommet, Stacy Marsella. Emotional Perception for Updating Agents' Beliefs. *International Conference on Affective Computing and Intelligent Interaction.* Xi'an, China. 2015
- Celso de Melo and Jonathan Gratch. People Show Envy, Not Guilt, when Making Decisions with Machines. *International Conference on Affective Computing and Intelligent Interaction.* Xi'an, China. 2015

Grant related invited talks:

- Keynote Speaker, International Society for Research on Emotion, St. Lewis, July 2017
- Panelist, Microsoft Research Faculty Summit, June 2017

- Invited Seminar, Samsung Research, May 2017
- Keynote Speaker, International Conference on Computer Animation and Social Agents, Korea, May 2017
- Invited Seminar, Stanford University, January 2017
- Invited Colloquium, Naval Research Laboratory, Washington D.C., December 2016
- Distinguished Lecture, Northwestern University, Chicago, October 2016
- Invited Colloquium, University of California, Irvine, October 2016
- Gratch, Invited Speaker, National Academy of Sciences Science and Entertainment Exchange, Los Angeles, October, 2015
- Marsella, Invited Speaker, University of Glasgow Institute of Neuroscience and Psychology, 2015
- Gratch, Invited Panelists, Human Factors and Ergonomics Society Annual Meeting, Los Angeles, October, 2015
- Gratch, Keynote, 4th International Workshop on Human-Agent Interaction Design and Models, Istanbul Turkey, May 2015

Grant Related Awards:

Elected Fellow of the Cognitive Science Society, 2017