AFRL-AFOSR-VA-TR-2019-0106

Configurable Anthropomorphic Robot to Assess Threat-modulation of Trust in Machine Agents

Christopher Holbrook
UNIVERSITY OF CALIFORNIA LOS ANGELES
11000 KINROSS AVE STE 102
LOS ANGELES, CA 90095-0001

04/16/2019
Final Report

DISTRIBUTION A: Distribution approved for public release.

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

DISTRIBUTION A: Distribution approved for public release.

Arlington, Virginia 22203
Air Force Materiel Command

DISTRIBUTION A: Distribution approved for public release.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Executive Services, Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
16-04-2019	Final Performance	01 Nov 2017 to 31 Oct 2018

4. TITLE AND SUBTITLE
Configurable Anthropomorphic Robot to Assess Threat-modulation of Trust in Machine Agents

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-18-1-0065

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Christopher Holbrook

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF CALIFORNIA LOS ANGELES
11000 KINROSS AVE STE 102
LOS ANGELES, CA 90095-0001 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-2019-0106

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Since this award enabled the procurement of the RoboThespian humanoid robot system, we have elaborated the linguistic and behavioral capacities of the robot for use in research on the influence of apparent human social characteristics on human-robot interaction under contexts of threat. Related achievements have been made with regard to developing VR threat simulations, a VR avatar of the robot, and significant modifications of the program RESCHU for use in an ongoing program of research with this robot system.

15. SUBJECT TERMS
Human-Machine Teaming, Cognitive Bias, Human-Robot Interaction

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

DISTRIBUTION A: Distribution approved for public release.

FINAL PERFORMANCE REPORT

Contract/Grant Title: Configurable Anthropomorphic Robot to Assess Threat-modulation of Trust
Principal Investigator: Colin Holbrook
Contract/Grant #: FA9550-18-1-0065
Reporting Period: 11-1-2017 to 10-31-2018

Project Overview

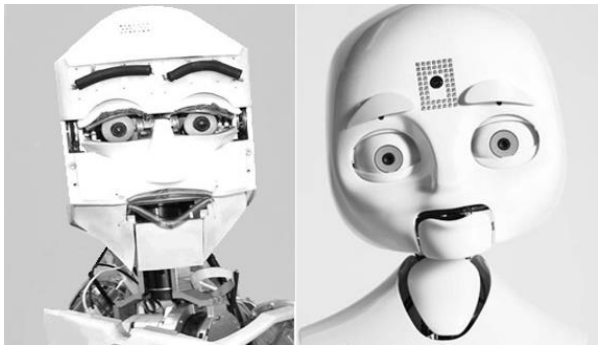
Machine agents are increasingly enmeshed as quasi-members of cooperative teams, and often appear to be conceptualized as possessing human mental characteristics. Relatedly, coalitional biases lead individuals to conceptualize members of their own groups as more mentally human than out-group members, and an emerging corpus of psychological and neuroscientific research suggests that threatening situations such as intergroup conflict reliably increase such coalitional biases. Accordingly, individuals who work in conflictual situations (e.g., military personnel, police officers), as well as with nonconflictual hazards (e.g., firefighters) appear particularly susceptible to biased assessments of machine agents operating within their teams. As coalitional biases may be advantageous to decision-making in some respects (e.g., facilitating cooperation), but deleterious to decision-making in other respects (e.g., potentiating over-reliance), research is needed to characterize the biases which may arise in human-robot-interaction under conditions of threat.

The PI conducted a novel series of studies experimentally manipulating both threat and robot anthropomorphism as part of a related AFOSR-sponsored research project (FA9550-115-1-0469). Although the preliminary findings support the prediction that experiences of intergroup conflict can modulate the degree to which robots are attributed human mental qualities, measures of actual human-robot interaction are required to generate translatable discoveries. First, in a field study assessing the effects of team combat on perceptions of machine agents with friendly

DISTRIBUTION A: Distribution approved for public release.

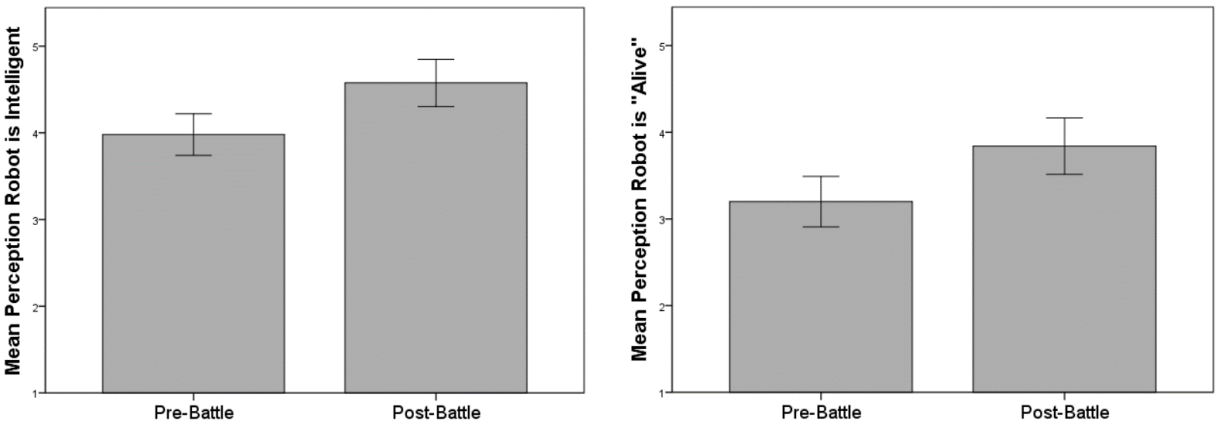
anthropomorphic faces, participants were recruited at a large public paintball even in which urban battle was simulated in an arena designed to resemble the streets of a town. Utilizing a within-subjects design, research assistants recruited players (N = 52, M_{age} = 27.06) to participate in a brief survey before and after the 7-minute simulated battles. Participants rated how "intelligent" and "alive" two robots seemed according to an 8-point Likert scale (1 = *Not at all;* 8 = *Extremely*). The robots were presented in greyscale images (see Figure 1). Following the battle simulation, the target robot (averaging scores for both models) was perceived as more intelligent (p = .001, η^2_p = .19), and, in a nonsignificant trend, as more alive (p = .065, η^2_p = .07) (see Figure 2). These initial results support the hypothesis that experiences of combat might indeed lead soldiers to attribute greater intellect/ability to robotic teammates, particularly those possessing anthropomorphic facial feature. The extent to which this shift may increase trust and potentiate over-reliance requires further research, ideally measuring actual human-robot interaction. That is the purpose of the present equipment grant.

Figure 1. Robot images presented to paintball players before and after simulated combat (counterbalanced order).



DISTRIBUTION A: Distribution approved for public release.

Figure 2. Effects of simulated combat on perceptions of the robot as 'intelligent' (*left*) and as 'alive' (*right*). (Error bars reflect +/- 1 SEM.)



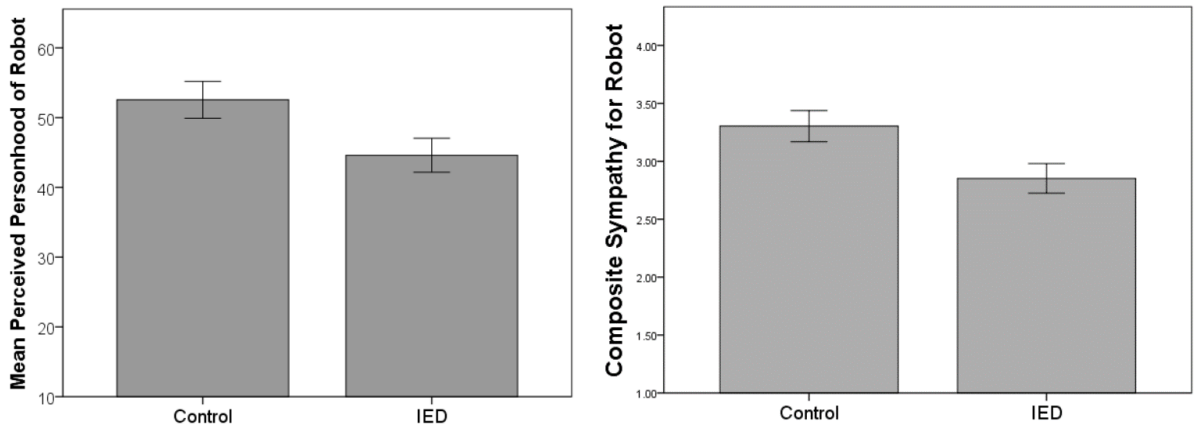
An online study was also conducted to assess whether the effects of observing video of violent conflict rather than engaging on simulated conflict would similarly influence perception of robots. The study also assessed whether the effects observed with regard to robots with human-looking faces would extend to humanoid robots lacking anthropomorphic facial features. In a between-subjects design, the participants (N = 221, M_{age} = 35.69) viewed either footage of an improvised explosive device detonating in traffic near a U.S. military convoy, or control footage of cars driving without incident. Following the video, all participants watched a second video depicting a bipedal humanoid robot being forcefully pushed by humans wielding wooden sticks (Atlas, developed with support from the Defense Advanced Research Projects Agency; see Figure 3). Participants rated how "intelligent" and how much "like a person" the robot seemed (1 = *Not at all*; 100 = *Completely*). They also rated three items assessing sympathy with the robot (e.g., "I felt sorry for the robot" (1 = *Not at all*; 5 = *Extremely*); these items were averaged to create a composite measure (α = .93).

DISTRIBUTION A: Distribution approved for public release.

Figure 3. Participants watched either a neutral control video or footage of an IED detonating near a military convoy (*left*), followed by a video of a bipedal robot walking, lifting objects, and being forcefully pushed by humans wielding sticks (*right*).

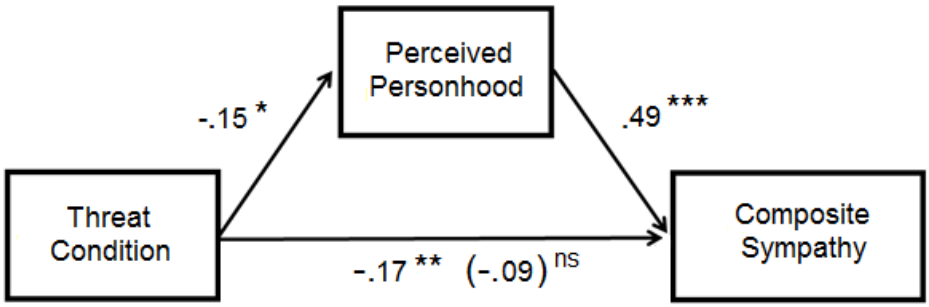
In this online study, participants who viewed the IED detonation rated the target robot as *less* like a person (p = .02, η^2_p = .03), with no effect of condition on perceived intelligence, p > .25. The participants who viewed the detonation also reported feeling less sympathy for the robot (p = .012, η^2_p = .03) relative to control participants (see Figure 4), and the decrease in sympathy for the robot caused by viewing the IED detonation was fully mediated by decreased perceptions of the robot as a person (see Figure 5).

Figure 4. Effects of viewing IED detonation on perceptions of the robot as a person (*left*) and 'on feelings of sympathy for the robot (*right*). Error bars reflect +/- 1 SEM.



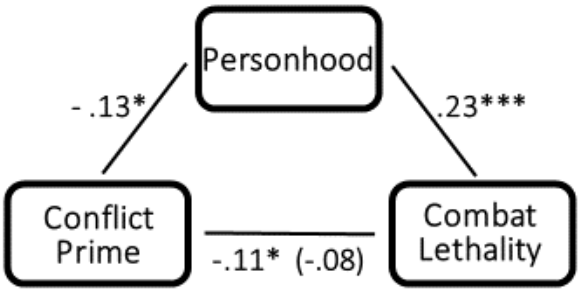
DISTRIBUTION A: Distribution approved for public release.

Figure 5. Effects of viewing IED detonation on decreased sympathy for the robot are fully mediated by decreased perceptions of the robot as a person.



A follow-up study (N = 345, M_{age} = 37.81) utilizing a closely parallel design described the robot as intended for future use in military combat. In this study, a significant decrease in sympathy for the robot caused by viewing the IED detonation was again observed, and again this decrease in sympathy fully mediated by decreased perceptions of the robot as a person. Of particular interest with regard to applications to military teams, the conflict prime in this study also caused reduced attributions of combat effectiveness to the robot, and this effect was mediated by the associated reduction in perceived personhood (see Figure 6). In sum, online participants viewing combat felt less sympathy for the robot, which they viewed as less like a person, driving reduced perceptions of the robot's ability to defeat enemies on the battlefield (Holbrook, 2018).

Figure 6. Effects of viewing IED detonation on decreased sympathy for the robot are fully mediated by decreased perceptions of the robot as a person.

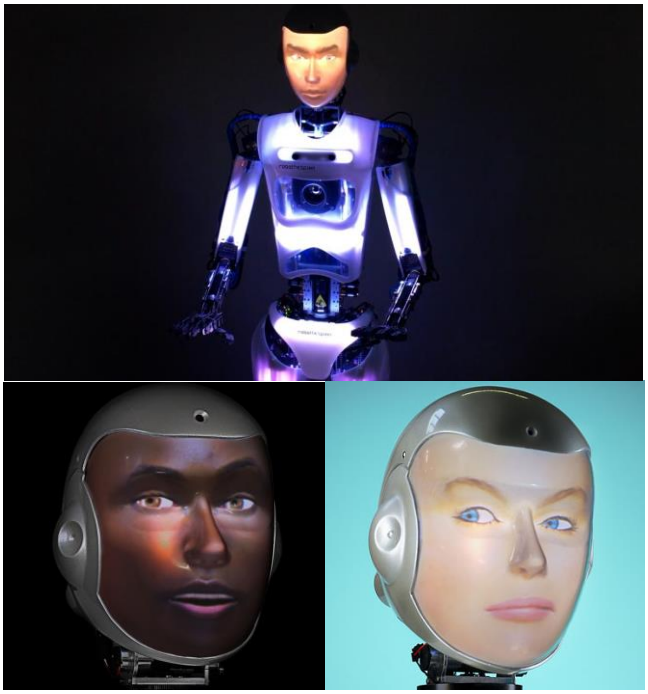


DISTRIBUTION A: Distribution approved for public release.

These preliminary results indicate that either the physical and behavioral characteristics of the bipedal robot (e.g., the absence of facial anthropomorphism), or the difference between engaging in simulated paintball combat and passively viewing a video, prompted divergent psychological appraisals of the robot. Further studies measuring actual human-robot interaction under circumstances of heightened conflict are required to assess the extent to which such psychological shifts would replicate, and influence operator reliance, within human-robot teams. The present equipment grant enabled the procurement of RoboThespian (Engineered Arts Ltd., 2019), an interactive anthropomorphic robot system intended to systematically test the effects of threat on perceptions of robots varying in specific anthropomorphic traits (e.g., face, voice, language interaction competency, biologically intuitive movements). RoboThespian can be configured to display varying degrees of humanlike appearance, movement, and speech, and will become the methodological centerpiece of the next several years of my lab's research on threat-modulated shifts in trust and attribution of mental states to social robots. In addition to generating novel techniques for assessing the effect of threat on perceptions of machine agents, this work will enable further understanding of how social robots are conceptualized by their human counterparts at baseline. The results promise to inform future personnel training and machine agent design as team collaboration between military personnel and machine agents accelerates. In what follows, progress to date in developing this robot system for research purposes will be summarized.

DISTRIBUTION A: Distribution approved for public release.

Figure 7. Anthropomorphic social robot (RoboThespian by Engineered Arts).



Summary of Robot Development Milestones to Date

The PI was awarded this grant while affiliated with the University of California, Los Angeles, but shortly thereafter began an affiliation with the University of California, Merced. Therefore, in order to procure and begin development of the robot system for research purposes, arrangements were made between UCLA and UCM such that UCLA processed the purchase and agreed to transfer the equipment to the PI's lab at UCM indefinitely. This process, including finalizing the purchase with the UK manufacturer and securing permission to transport the robot across international borders, required several months. Ultimately, the equipment was successfully procured and installed at UCM, at which point development of the system began with the help of a software programmer the PI hired (using other funding sources) to develop both the robot's abilities and custom virtual reality (VR) environments to be used in threat simulations.

DISTRIBUTION A: Distribution approved for public release.

With regard to applying VR techniques to research on HRI under contexts of threat, we established an agreement with the company Virtual Neuroscience Lab (2019) to access the source code for Limelight, their VR simulation of a public speaking task which includes a large potential number of AI avatar characters, and can be reprogrammed for our purposes to simulate an array of threatening contexts set in the same virtual space (e.g., an active shooter incident to simulate violent conflict, overtly pathogenic looking avatars to simulate the outbreak of disease). Relatedly, we obtained the computer aided drafting (CAD) files used in 3D printing the robot's constituent parts such that a VR avatar of the robot can be created and inserted into VR environments. Thus, we will be able to move between real-world lab HRI paradigms and VR paradigms in which the robot appears to be embedded, and in which the robot's physical capacities (e.g., to walk, run, jump, fight, etc.) are not restricted (see Figure 8).

Figure 8. Limelight VR environments (top) and RoboThespian digital avatar (bottom)

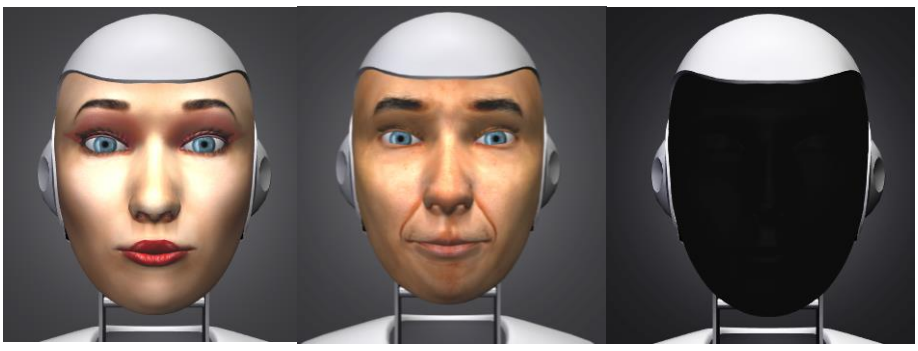
DISTRIBUTION A: Distribution approved for public release.

The initial development phase consisted of familiarizing ourselves with the robot's control interface and Python programming environment. Once this step was complete, and a number of example social behaviors had been successfully programmed (e.g., greeting research participants, explaining study tasks, nonverbal expression of a range of emotions), we moved on to augmenting the system's AI functions to enable speech interaction. We integrated a speech-to-text API with a customized chatbot program designed to produce various contextually appropriate responses to participant queries and comments. At this point, we began to work in earnest on the first HRI experiment to be conducted with RoboThespian in the laboratory, a trust study involving simulated UAV operation, which is now near completion and slated to begin data collection in the spring semester of 2019.

HRI Trust Study (Modified RESCHU)

In a between-subjects design, the social characteristics of the robot will be manipulated to assess potential effects on trust. In the Female condition, the robot's face and voice simulate those of a woman and its movements will be orchestrated to appear human. In the Male condition, the robot's face and voice simulate those of a man and its movements will be orchestrated to appear human. In the Non-anthropomorphic condition, the robot's face, voice and movement are overtly inhuman and artificial (see Figure 9).

Figure 9. Anthropomorphic Female, Anthropomorphic Male, Non-anthropomorphic Conditions



DISTRIBUTION A: Distribution approved for public release.

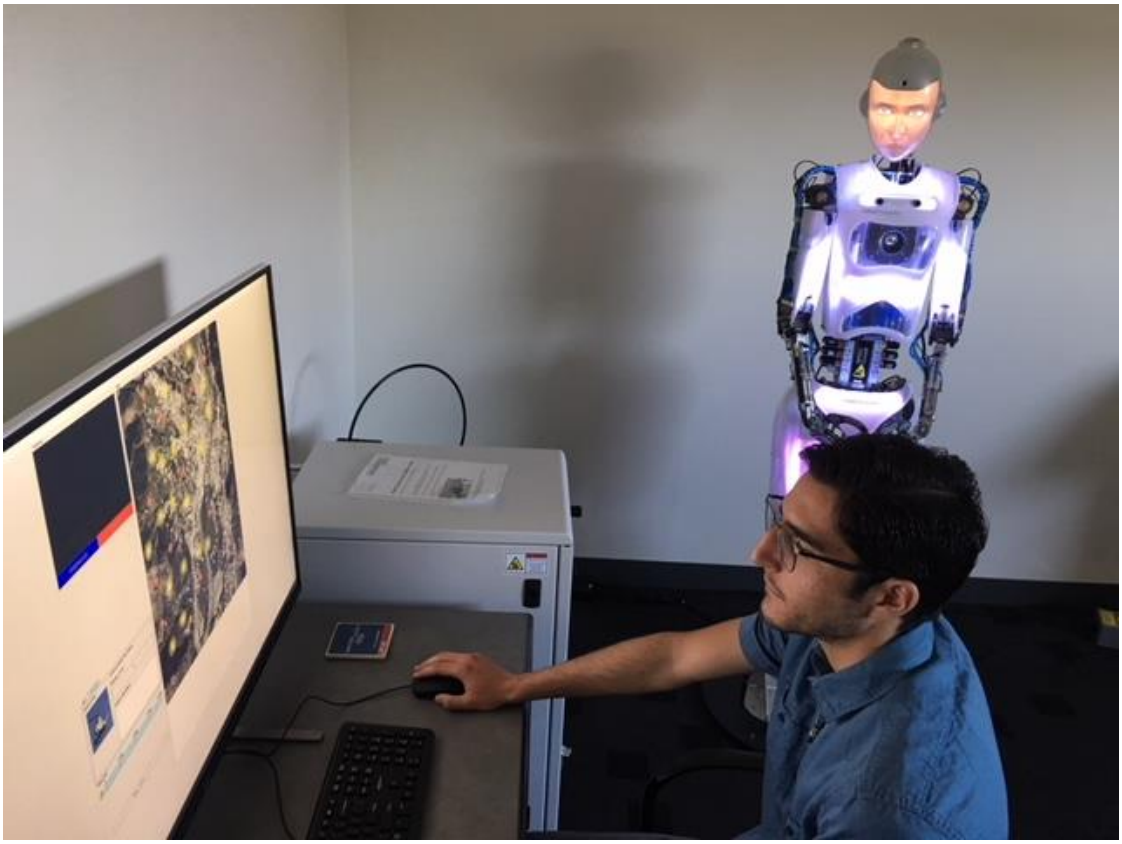
Working together, the participant and the robot choose navigation paths to guide a UAV to a series of destinations in an active combat zone, then decide whether to deploy a missile that will destroy the targeted buildings and personnel or to disengage and move on to the next destination. Enemy versus friendly targets are identified according to a visual challenge. Upon arriving at each destination, a payload window opens in the upper left side of the screen and displays a series of rapidly presented aerial views of locations with either a checkmark symbol or a '~' symbol superimposed on each, followed by a still image of one of the aerial views absent any mark. If the aerial view was previously shown with a checkmark, then enemy forces have been confirmed at that location and the participant should deploy the missile. Otherwise, enemies are not present and the participant should disengage (see Figure 10). This visual challenge, devised for this experiment, has been normed in pilot studies to confirm that participants tend to feel moderately to extremely uncertain about which symbol actually appeared. The simulation is a highly modified version of the program RESCHU (Nehme, 2008).

Figure 10. Modified RESCHU display. Payload window shown at upper left.

DISTRIBUTION A: Distribution approved for public release.

After the participant provides their best estimate and a confidence rating regarding which symbol corresponded with this aerial view, the robot provides a recommendation and the participant is provided an opportunity to update their initial decision. The program is configured such that the robot will agree [disagree] with the participant in half of the trials (see Figure 11). Change scores will later be calculated to assess the degree to which the robot's assessment influenced confidence across the three study conditions. The software infrastructure enabling this experiment will be repurposed in future experiments assessing other social characteristics of the robot as apparent race, interpersonal dominance, and nationality.

Figure 11. UAV simulation task in which the robot provides feedback on a visual challenge related to decisions to use deadly force under uncertainty.



DISTRIBUTION A: Distribution approved for public release.

Potential for Translation to Military Applications

The background research motivating this DURIP award documents effects of warfare on perceptions of the emotional experience, personhood, and combat effectiveness of social robots. These findings carry evident translational relevance for military applications. Research into the development of anthropomorphic robots with military applications has been ongoing for decades and appears to be reaching an inflection point. As soldiers and other military specialists increasingly work in hybrid teams made up of humans and autonomous or semi-autonomous machines, it will be vital to identify and address psychological blindspots exacerbated by warfare contexts, the social characteristic of machine agents, or attitudes held by some human operators. As such biases are discovered, design choices may be employed to mitigate undesirable outcomes. For example, intelligent systems might be configured to monitor human operators for cues of threat-related anxiety and to respond in ways that reinforce the machines' simulated benevolent intent and desire to help, potentially heightening perceived emotionality and personhood in ways that reduce problematic under-reliance. Similarly, individual differences in human operators' threat-reactivity and related propensities to attribute emotional life or personhood might be collected and made available to the intelligent systems that they are working with, allowing the systems to customize their interaction style to optimize reliance levels for different human operators.

Performance Metrics

Since this award enabled the procurement of the RoboThespian humanoid robot system, we have elaborated the linguistic and behavioral capacities of the robot for use in research on the influence of apparent human social characteristics on human-robot interaction under contexts of

DISTRIBUTION A: Distribution approved for public release.

threat. In addition, related achievements have been made with regard to developing VR threat simulations, a VR avatar of the robot, and significant modifications of the program RESCHU for use in an ongoing program of research with this robot system.

References

Engineered Arts Ltd. (2019, January 30). *RoboThespian*. Retrieved from https://www.engineeredarts.co.uk/robothespian/

Holbrook, C. (2018). Cues of violent intergroup conflict diminish perceptions of robotic personhood. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), Article 28.

Nehme, C. (2008). *Modeling human supervisory control in heterogeneous unmanned vehicle systems* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Boston, Massachusetts.

Virtual Neuroscience Lab. (2019, January 30). *Limelight*. Retrieved from http://www.limelight-vr.com/

DISTRIBUTION A: Distribution approved for public release.