AFRL-AFOSR-VA-TR-2019-0009

Group Bias and the Attribution of Mental Properties to Allies, Antagonists, and Automata

Christopher Holbrook
UNIVERSITY OF CALIFORNIA LOS ANGELES

01/04/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

DISTRIBUTION A: Distribution approved for public release.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
07-01-2019	Final Performance	30 Sep 2015 to 29 Sep 2018

4. TITLE AND SUBTITLE
Group Bias and the Attribution of Mental Properties to Allies, Antagonists, and Automata

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-15-1-0469

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-2019-0009

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

To date, this project has generated novel evidence that i) threat cues heighten perceptions of in-group allies as strategically capable, and that ii) individual differences linked with threat-reactivity similarly predict confidence in coalitional competence and the prospect of victory. The third year of the project has yielded further support for the core hypotheses motivating the grant. In particular, the findings of Year 3 have produced further evidence that the threat of violent intergroup conflict leads individuals to attribute greater intellect to allies relative to adversaries (Holbrook et al. 2018, Social Psychological and Personality Science), and to believe claims regarding potential hazards (Samore et al. 2018, PLoS ONE). Likewise, individual differences in perceived supernatural support (Holbrook et al. 2018, Religion, Brain and Behavior; Pollack et al. 2018, Human Nature) were observed to heightened confidence in the coalition's prospects for victory in a series of unusually ecologically valid field studies involving realistically simulated gun and knife combat. The studies of the effects of threat on perceptions of social robots conducted in previous years were successfully published (Holbrook 2018; ACM Transactions on Interactive Intelligent Systems), and the DURIP equipment grant seeking funds for a highly configurable anthropomorphic robot to follow up on those findings was purchased and installed in my laboratory. Considerable effort has gone into custom-programming the robot and preparing virtual reality threat simulations for the planned series of laboratory studies of the effects of threat on mind attribution and related aspects of human-robot interaction, which are now slated to begin in the spring of 2019.

15. SUBJECT TERMS
Human-Machine Teaming, decision-making, Transcranial Magnetic Stimulation

16. SECURITY CLASSIFICATION OF:				

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

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

FINAL PERFORMANCE REPORT

Contract/Grant Title: Group Bias and the Attribution of Mental Properties to Allies, Antagonists, and Automata
Principal Investigator: Colin Holbrook
Contract/Grant #: FA9550-115-1-0469
Reporting Period: 09-30-2015 to 09-29-2018

Project Overview

The functional logic of coalitional assortment provides a framework within which to understand judgment biases occurring under contexts of active conflict. Members of coalitions gain access to both material and informational resources (i.e., culturally transmitted knowledge), and are therefore motivated to regard in-group members as more valuable than out-group members, thereby enhancing mutual resource-sharing, trust and cooperation. Conversely, individuals aligned with out-groups should be expected to be perceived as less deserving of in-group resources, less trustworthy, or even as threats. If in-group favoritism advanced reproductive fitness over deep time, then natural selection may have shaped the human brain not only to support ethnocentrism, but to intensify baseline coalitional biases when threatened in order to increase the individual's ability to draw on group alliances.

People are generally overconfident in their expectations of coalitional victory, a pattern theorized to reflect an evolved bias that, although disastrous in particular cases, has generated aggregate adaptive benefits (e.g., by boosting resolve to fight) over our species' long history of warfare. Relatedly, and consistent with chauvinistic group attitudes, individuals typically conceptualize members of their own coalition as relatively more mentally sophisticated than out-group members. Accordingly, intergroup conflict may be expected to heighten baseline biases in the perceived intellect of in-group fighters relative to their out-group adversaries. Importantly, intergroup conflict should heighten perceptions of the intellectual ability of allies moreso than

diminish perceptions of the intellectual ability of adversaries, as underestimating the strategic cunning of adversaries would be maladaptive. Cues of violent conflict may similarly bias perceptions of the mental attributes and performance abilities of anthropomorphic machine agents, given that people intuitively ascribe human qualities to machines and that threat mobilizes functional shifts in social perceptions. Specifically, machine agents may be intuitively categorized as out-group members as a by-product of the evolved coalitional psychology engaged when evaluating human beings.

Much as exposure to cues of violent conflict may be expected to modulate perceptions of the in-group's capacity to possess superior mental capacities and to win, stable individual differences in threat-assessment should similarly predict biased perceptions of allies and adversaries. For example, political orientation indexes individual differences in both prioritizing the welfare of the in-group and sensitivity to potential hazards. Relative to liberals, conservatives evince greater physiological reactivity to threatening imagery or noise bursts, invest more time attending to threats, and are more implicitly distracted by threatening imagery. Such threat vigilance should not be mistaken for timidity, as conservatives generally favor relatively aggressive responses to intergroup conflict. Accordingly, conservatives should display confidence in victory and view in-group fighters as relatively more intellectually capable than out-group enemies to a greater extent than liberals. Similarly, perceiving supernatural agents as sources of aid can inspire confident aggressive responses to conflict by bolstering confidence in victory, much as one might expect to follow from perceptions of access to powerful earthly allies, weapons, or abilities.

Synthesizing these considerations, the present project investigated how threat-detection interacts with coalitional affiliation to bias assessments of both human and machine agents with

regard to appraisals of intellectual ability and/or confidence in victory. As the summary of specific findings that follows will detail, convergent evidence was obtained indicating that both temporary exposure to cues of violent conflict and individual differences related to threat-reactivity predict perceptions of in-group allies as more strategically capable and likely to defeat out-group enemies. With regard to machine agents, group warfare stimuli were found to diminish perceptions of the personhood and military effectiveness of social robots. Finally, investigations into corollary hypotheses revealed a consonant pattern of findings wherein individual differences in political orientation predicted tendencies to believe claims about threats, propensities for risk-taking were correlated in disparate domains, and cues of poor group cooperation inhibited inclinations to cooperate. Highlights of the projects are briefly presented in what follows, organized thematically rather than chronological order in the interest of clarity. The publication numbers provided refer to the publications listed at the conclusion of this report.

Summary of Specific Findings

Threat and Mind Attribution

The core hypothesis motivating this project is that threat facultatively alters social perceptions such that allies are viewed as more competent than adversaries. Consistent with this prediction, experimental assignment to view a brief video depicting intergroup warfare caused participants to envision an ally (an in-group soldier) as more intelligent than an adversary (an out-group militant) to a greater extent than in a control manipulation (Publication 4). This finding has been robustly reproduced in multiple studies (some of which are currently in preparation for publication), and conceptually replicates observations that individuals become

more ideologically committed to their in-groups in the aftermath of reminders of death, in an effect linked with neural threat-detection mechanisms discussed in Publication 6.

Extending this work to perceptions of machine agents, a similar warfare video paradigm was used to assess the effect of threat on attributions of mental qualities and performance ability to robots (Publication 9). These studies produced a number of provocative results. In the first two studies, participants evaluated a large, bipedal all-terrain robot (Atlas by Boston Dynamics), described as either intended for disaster rescue functions or battlefield combat. In both studies, participants who had been exposed to the warfare stimulus attributed less emotional awareness to the robot, and this reduction fully mediated ratings of the robot as seeming less like a person. Further, participants' estimates of the combat lethality of the military robot were positively correlated with the personhood ascribed, such that the reduction in perceived personhood in the warfare prime condition fully mediated the diminished ratings of combat lethality that were observed in the warfare prime condition. A third study utilizing a less imposing social robot with a face and humanlike speech (Nexi) yielded a comparable pattern of results. These studies are the first to observe effects of violent conflict on perceptions of the mental life, personhood, or abilities of machine agents, and demand further study. To this end, a Defense University Research Instrumentation Program grant was sought (FA9550-18-1-0065) to procure an advanced social robot for a programmatic series of human-robot interaction studies in which conflict conditions will be vividly simulated in virtual reality to systematically test the effects of threat on robots varying in particular anthropomorphic traits (e.g., face, voice, language interaction competency, biologically intuitive movements). The robot system and virtual reality apparatus are now custom-programmed and installed in my laboratory, with data collection set to begin in spring of 2019 (see Figure 1).

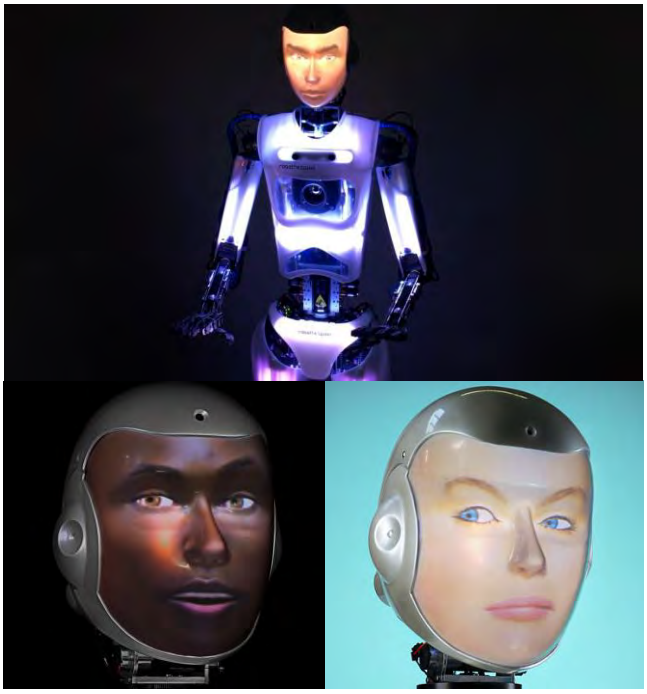

Figure 1. Anthropomorphic social robot (RoboThespian by Engineered Arts)

The aforementioned findings with respect to effects of threat on perceptions of intellect invite consideration of other mental qualities advantageous in combat. For example, although still in preparation for publication, this project has generated evidence over multiple studies that cues of threat heighten perceptions not only of the intellectual ability of allies, but also of their subjective resolve to fight. Another promising avenue of study concerns the hypothesized link between assessments of one's group as superior and processes of cooperation between group members. In a related recent finding, cues of poor group cooperation inhibited participants' inclinations to cooperatively render prosocial aid (Publication 10).

Individual differences in mind attribution and combat confidence

Just as exposure to cues of threat may be expected to functionally shift perceptions of mental capacities and abilities, individual differences in threat-reactivity may generate parallel appraisal tendencies. Political orientation has been identified in convergent research as a robust indicator of threat-reactivity, such that conservatism [liberalism] predicts a greater [lesser] tendency to identify ambiguous targets as threatening, as well as a greater [lesser] degree of confidence in the capacity to defeat threats. For example, Publication 2 reports the first evidence that conservatives are more prone to believe claims about a wide array of potential threats than are liberals; Publication 8 replicates and extends these results to show that the capacity among conservatives to believe in putative threats coincides with relatively greater confidence in the ability for their preferred political party to triumph in upcoming elections. Although these findings are broadly consistent with a depiction of political orientation as at once enhancing perception of the presence of threats and confidence in vanquishing them, little work has examined the influence of political orientation on real-world decision processes related to violent group conflict.

To begin to fill this gap, in a pair of studies reported in Publication 1, political orientation was correlated with assessments of the intent and physical size/strength of Syrian refugees. Physical size/strength were measured as an implicit assessment of the capacity of the individual to win in a violent conflict. As predicted, conservatism tracked categorizing the target refugee character as possessing more malevolent intent to commit acts of terror, but also as relatively physically diminutive, and the significant association between political conservatism and lower body size ratings was fully mediated in multiple studies by confidence in the ability of the in-group to militarily defeat terrorist groups. A comparable set of findings was obtained in

Publication 4, in which political orientation was correlated with ratings of the intellect of an in-group fighter and an unambiguously malevolent target—a fighter aligned with the terrorist group known as ISIS. Again consistent with predictions, and echoing the effects of exposure to warfare cues, conservatives rated the in-group fighter as relatively more strategically capable, but did not estimate the ISIS fighter to be less strategically capable. The specificity of the relationship between political orientation and appraisals of the in-group ally argues against an interpretation rooted in group prejudice alone, as such an account would predict derogation of the antagonistic enemy as less intelligent. Rather, political conservatism appears intrinsically related to inclinations to invest confidence in the leadership of in-group military leaders to fight enemies, but not to underestimate enemies' strategic capacities.

Religiosity is another key individual difference that has been broadly associated with confidence in the face of danger. To test whether belief in supportive supernatural agents bolstered coalitional confidence in contexts of violent conflict in a manner comparable to political orientation, and to methodologically extend the aforementioned findings to a more ecologically valid, immersive combat simulation, a series of field studies were conducted. Publication 7 describes field research conducted at martial arts training classes in which a community sample of participants were taught fundamentals of knife fighting. Participants were assigned to teams competing in group knife combat. Just before the battles commenced, participants listened to a brief (~90 second) guided visualization prime involving either support from unseen supernatural powers or a control visualization, then rated their confidence in the performance of themselves and their teams. Individual differences in religious faith and political orientation were also collected. Consistent with the hypothesis that perceived supernatural support enhances battle confidence, both experimental assignment to the supernatural

visualization and trait religiosity significantly predicted battle confidence. A similar pattern was observed with regard to political orientation, conceptually replicating the results of Publications 1 and 4 in a more realistic battle simulation paradigm. A nearly identical supernatural support versus control visualization paradigm was then utilized in a field study conducted with competitive team paintball players (Publication 5; see Figure 2). Although logistical constraints in this unusual study context prevented collection of detailed demographics, including individual differences in trait religiosity or political orientation, the same basic results emerged with regard to the impact of briefly visualizing the presence and aid of benevolent supernatural forces. Participants in the supernatural support condition were more confident of their team's victory and assessed their team as more skilled than the rival team, in effects paralleling the effects of religiosity observed in Publication 7 and of political orientation reported in Publications 1 and 4.

Figure 2. Photograph conveying a participant's point of view during the simulated coalitional combat (team paintball) field study (Publication 5).

With regard to whether and how individual differences in threat-reactivity influence judgment and behavior, one basic question concerns the extent to which effects generalize across domains of threat (e.g., pathogen exposure, violence, resource deprivation, social stigma, nonviolent risk-taking, etc.). Although there is no question that distinct threat domains exist within the mind, reflective of the unique functional responses appropriate to distinct threats, Publication 3 showed that individual differences in threat-reactivity correlate across distinct domains, such that individuals who are less averse to physical risks (e.g., combat) tend to be less averse to other sorts of risks (e.g., pathogen exposure).

Potential for Translation to Military Applications

This research investigated the impact of threat and threat-reactivity on coalitional bias with regard to humans and, by extension, social robots, revealing a number of associations with translational relevance to military operations. For example, the findings suggest that demographic differences in the political orientation or religiosity of military personnel, enemy groups, or other relevant populations will predict tendencies to seek confrontation versus negotiation. The differential between the estimated intellectual ability of members of antagonistic groups may also be taken as an indirect, intuitive measure of individuals' perceptions of the relative formidability of each party. For example, personnel might utilize these measures to assess the combat confidence of members of allied forces in a manner less susceptible to demand characteristics than simply asking them (i.e., to the extent that individuals might feel compelled to answer affirmatively). Likewise, assessments of relative intellect may be used to gauge combatants' assessments of the relative formidability of opposing parties.

Turning to human-machine interaction, this project's novel findings with regard to the effect of warfare cues on perceptions of the emotional experience, personhood, and combat effectiveness of social robots carry arguably the greatest translational potential 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. For example, the Battlefield Extraction-Assist Robot, developed for the U.S. Army, is a bipedal humanoid robot of approximately human height designed to carry supplies or wounded soldiers. 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 or intrinsic to the trait 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

This project was comprised of 17 successfully published studies, conducted over 36 months, and resulting in 10 published or in-press empirical journal articles, 4 under-review or in-

preparation papers, as well as additional data that are currently being analyzed. Reflecting the culmination of prior years' investments in data collection, eight of the ten completed items were published during the final year of the project. The datasets and full methods and materials accompanying the majority of the published works have been placed in the publically accessible Open Science Framework research archive (see individual papers for access information). Publications from this project have received coverage by major media outlets, including The Atlantic, The Los Angeles Times, New York Magazine, Pacific Standard and The Guardian (United Kingdom), among others. Twenty-two undergraduate research assistants also received training over the course of this research. Finally, the emerging evidence obtained regarding the influence of threat on attributions of competency and human mental qualities to robots (Publication 9) bolstered a successful DURIP grant to procure a cutting-edge humanoid social robot for upcoming laboratory work on teaming with machine agents (FA9550-18-1-0065; 'Configurable Anthropomorphic Robot to Assess Threat-modulation of Trust in Machine Agents'; $141,481).

Publications

1. Holbrook, C., López-Rodríguez, L., Fessler, D.M.T., Vázquez, A., Gómez. A. (2017). Gulliver's politics: Conservatives envision potential enemies as readily vanquished and physically small. *Social Psychological and Personality Science, 8*(6), 670-678.

2. Fessler, D.M.T., Pisor, A.C., & Holbrook, C. (2017). Political orientation predicts credulity regarding putative hazards. *Psychological Science, 28*(5), 651-660.

3. Sparks, A., Fessler, D.M.T., Chan, K.Q., Ashokkumar, A., & Holbrook, C. (2018). Disgust as a mechanism for decision making under risk: Illuminating sex differences and individual risk-taking correlates of disgust propensity. *Emotion, 18*(7), 942-958.

4. Holbrook, C., López-Rodríguez, L., Gómez, Á. (2018). Battle of wits: Militaristic conservatism and warfare cues enhance the perceived intellect of allies versus adversaries. *Social Psychological and Personality Science, 9*(3), 319-327.

5. Pollack, J., Holbrook, C., Fessler, D.M.T., Sparks, A. M., & Zerbe, J.G. (2018). May God guide our guns: Visualized supernatural aid heightens team confidence in a paintball battle simulation. *Human Nature, 29*(3), 311-327.

6. Holbrook, C., Gordon, C.G., & Iacoboni, M. (2018). Continuous theta burst stimulation of posterior medial frontal cortex to experimentally reduce ideological threat responses. *Journal of Visualized Experiments, 139,* e58204, doi:10.3791/58204

7. Holbrook, C., Pollack, J., Zerbe, J.G., & Hahn-Holbrook, J. (2018). Perceived supernatural support enhances battle confidence: A knife combat field study. *Religion, Brain & Behavior.* doi.org/10.1080/2153599X.2018.1464502

8. Samore, T., Fessler, D.M.T., Holbrook, C., & Sparks, A.M. (2018). Electoral fortunes reverse, mindsets do not: Political orientation, credulity, and conspiracism following the 2016 U.S. elections. *PLoS ONE, 13*(12), e0208653. doi.org/10.1371/journal.pone.0208653

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

10. Fessler, D.M.T., Sparks, A.M., Samore, T., & Holbrook, C. (in press). Does observing reciprocity or exploitation affect elevation, a mechanism driving prosociality? *Evolutionary Human Sciences.*