Cues of Violent Intergroup Conflict Diminish Perceptions of Robotic Personhood

COLIN HOLBROOK, University of California, Merced

Convergent lines of evidence indicate that anthropomorphic robots are represented using neurocognitive mechanisms typically employed in social reasoning about other people. Relatedly, a growing literature documents that contexts of threat can exacerbate coalitional biases in social perceptions. Integrating these research programs, the present studies test whether cues of violent intergroup conflict modulate perceptions of the intelligence, emotional experience, or overall personhood of robots. In Studies 1 and 2, participants evaluated a large, bipedal all-terrain robot; in Study 3, participants evaluated a small, social robot with humanlike facial and vocal characteristics. Across all studies, cues of violent conflict caused significant decreases in perceived robotic personhood, and these shifts were mediated by parallel reductions in emotional connection with the robot (with no significant effects of threat on attributions of intelligence/skill). In addition, in Study 2, participants in the conflict condition estimated the large bipedal robot to be less effective in military combat, and this difference was mediated by the reduction in perceived robotic personhood. These results are discussed as they motivate future investigation into the links between threat, coalitional bias and human-robot interaction.

• Human-centered computing~Human computer interaction (HCI) • Human-centered computing~HCI theory, concepts and models

Additional Key Words and Phrases: threat detection, human-robot interaction, empathy, group prejudice, Theory of Mind

1. INTRODUCTION

Humans spontaneously attribute mental states and emotional motivations to objects as rudimentary as moving geometric shapes [Heider and Simmel 1944]. As sophisticated robots and other machine agents become increasingly interwoven into modern life [Ishiguro and Nishio 2007; Coradeschi et al. 2006], it is imperative that we understand how humans conceptualize the machine partners with which we collaborate. To the extent that people tend to intuitively ascribe human qualities to machines [Knijnenburg and Willemsen 2016], and that threat mobilizes functional shifts in social perceptions upon detecting cues of hazard [Holbrook 2016], threat may be expected to bias perceptions of the mental attributes and personhood of robots.

Although no prior investigations have specifically targeted the impact of threat on the attribution of mental states or personhood to machines, there are several lines of indirect evidence for such a relationship. For example, people who report feeling a need for greater control over their environment—arguably an index of perceiving the world as threatening—attribute a greater degree of mind to inanimate objects [Waytz et al. 2010; Epley et al. 2007]. Further, in a stark demonstration of the potential for over-reliance when people feel threatened, an overwhelming majority of participants in a

This work is supported by the United States Air Force Office of Scientific Research under Award no. FA9550-115-1-0469. Author's address: C. Holbrook, Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Road Merced, CA 95343.

DOI:http://dx.doi.org/10.1145/0000000.0000000

recent study followed a robot guide in an emergency simulation, despite the fact that the robot's guidance was obviously wrong (e.g., leading them away from clearly marked exits) [Robinette et al. 2016]. Given the life or death stakes of judgment biases occurring during military or emergency response operations, understanding the psychology undergirding human-robot interaction under contexts of active threat is particularly vital [Ososky et al. 2012].

Research into the development of anthropomorphic robots with military applications has been ongoing for decades. For example, the Battlefield Extraction-Assist Robot, developed for the U.S. Army, is a bipedal humanoid robot of approximately human height that can lift up to 500 pounds, sufficient to carry supplies or wounded soldiers [Gilbert and Beebe 2010]. This and similar humanoid robot initiatives are still in development, but the day is fast approaching when soldiers and other military specialists will work in hybrid teams made up of humans and autonomous or semi-autonomous machines [Everett et al. 2004; Barnes et al. 2013; Carpenter 2013; Barnes et al. 2011]. Indeed, explosive ordnance disposal units currently incorporate semi-autonomous robots into their teams to help detect, inspect, and disarm explosive devices, and thereby reduce the likelihood of human injury or death. In a study of human-robot-interaction among explosive ordinance detonation technicians and the robots with which they work in close proximity, Fincannon and colleagues [2004] observed that the human operators maintained eye-to-robot contact and observed human-to-human interpersonal distance norms. Anecdotally, explosive ordnance disposal technicians have been known to name, paint faces onto, and express grief over these robots when they are damaged or destroyed [Garreau 2007; Barylick 2006], and have even awarded robots "posthumous" medals and organized elaborate funerals for them, including 21-gun salutes [Carpenter 2013]. In sum, humans working under hazardous conditions appear to reflexively import the conceptual schemas of human psychology into their interactions with robotic teammates.

Robots may be reflexively categorized as out-group members as a by-product of the evolved coalitional psychology engaged when evaluating people. Natural selection appears to have shaped the human mind to support ethnocentrism because group-based categorization and favoritism advanced reproductive fitness over deep time [De Dreu et al. 2011; Neuberg, Kenrick, and Schaller 2010; Hammond and Axelrod 2006]. Categorizing another person as sharing a positive investment in a common in-group—and hence, ethnocentrically, as more valuable and reliable than members of out-groups—is thought to privilege coordination between fellow in-group members by enhancing mutual resource-sharing and trust [Darwin 1873; Efferson, Lalive and Fehr 2008]. Conversely, ethnocentrism may also manifest as negative valuation of individuals who are perceived to be aligned with out-groups and hence as more likely to attempt to exploit or harm [Fiske 2002; Dovidio and Gaertner 2010; Holbrook, Fessler and Navarrete 2016]. With regard to perceptions of robots, these deep-seated coalitional tendencies may lead individuals to perceive robots as less sympathetic—and even as less human—in line with previously documented tendencies to regard out-group members as less sympathetic and less human [Brewer 1999; Haslam 2006]. The extent to which robots are categorized as coalition versus out-group members may therefore be a key determinant of the influence of threat on perceptions of their personhood. Furthermore, individuals may perceive robots as still less sympathetic or human under conditions of hazard, as numerous studies have found that coalitional biases in perceptions of out-group members magnify following experimental primes of threat [Holbrook, Izuma et al. 2016; Jonas et al. 2014]. Notably, the limited, largely anecdotal evidence to date bearing on human-robot interactions under threat primarily focuses on individuals likely to regard the robots as helpful members of their coalitions (e.g., explosive ordnance technicians). People who are not routinely embedded in human-robot teams, and who regard human-robot interaction as relatively exotic, may intuitively evaluate robots as out-group others that are less possessed of human mental states or personhood when under conditions of threat.

Another outstanding question with regard to perceptions of robotic personhood concerns the relative contributions of attributed intellectual versus emotional capacities. Minds appear to be represented along two broad dimensions corresponding to i) the cognitive capacity to formulate and execute plans in a strategic, efficacious manner, and ii) the subjective capacity for emotional experience [Gray et al. 2007]. Machine agents are generally regarded as possessing rich computational and memory capacities, but not emotional experience, and it is the absence of the latter facet of mental life that presumably leads individuals to regard machine agents as disposable non-persons. Gray and colleagues [2007] found that humans are conceptualized as distinct from robots primarily with respect to our capacity for emotions such as fear or pain, as well as the intuitively related capacity for subjective consciousness. In convergent evidence that it is the perceived deficit in emotional experience that primarily leads humans to regard machine agents as non-persons, representing other human beings as emotionally cold has been termed "mechanistic dehumanization," and found to motivate indifference to harm befalling these persons [Haslam, 2006]. Thus, to the extent that threat modulates perceptions of robotic personhood, such shifts should be primarily mediated by shifts in perceptions of the robot as an emotional being.

Here, the effects of cues of violent conflict on robot perception are explored, using a participant sample of individuals who lack pre-existing affiliative relationships with the robot evaluation targets. In addition to testing the hypothesis that shifts in perceived robotic personhood will be mediated by differences in attributed emotionality rather than intellect, these three studies are intended to shed light on three open questions regarding the influence of threat on perceptions of robots' mental states:

1. Will threat cues heighten or diminish perceptions of robotic personhood?
2. Will the effects of threat on robot perception stem from the negative affective reactions inherent to witnessing violent conflict?
3. Will perceptions of personhood influence assessments of robots' potential reliability in combat or other emergencies?

2. STUDY 1

2.1 Participants and overview of procedure

300 U.S. participants were recruited via Amazon's MechanicalTurk.com survey platform in exchange for $0.55 for a study titled "Visual Memory and Social Perceptions".[1] Data were pre-screened for completeness, repeat participation, U.S.

[1] The participants in Studies 1-3 were drawn from the online labor market Amazon Mechanical Turk (AMT), in which workers complete brief tasks in exchange for small sums of money [Paolacci et al. 2010]. AMT allows researchers to move beyond the convenience samples of university undergraduates that are typically used in psychological research [Horton et al. 2011; Buhrmester et al. 2011]. To address potential concerns about the comparability of online samples to conventional laboratory samples, a number of recent studies have tested and affirmed the validity of results generated with AMT samples [Rand 2012]. Furthermore, AMT participants have been found to evince a level of test-retest reliability comparable to that observed in laboratory samples on diverse measures such as political attitudes, self-esteem, and personality traits, as well as age, sex, education level, income, and religiosity [Mason and Suri 2011].

citizenship, reported video playback problems, spending at least 5 minutes taking the study, and correctly answering attention checks (detailed below).[2] The final sample consisted of 224 adults (51.8% female; 75.3% White) ranging in age from 18 to 70 (M = 35.71, SD = 12.35).

In a between-subjects design, participants were randomly assigned to view either a video of highway traffic (Control condition; N = 108) or of an improvised explosive device (IED; N = 116) detonating in Iraq (Conflict condition; see Figure 1; to access the video stimuli, see the Supplemental Electronic Material). The videos were presented with accompanying text:

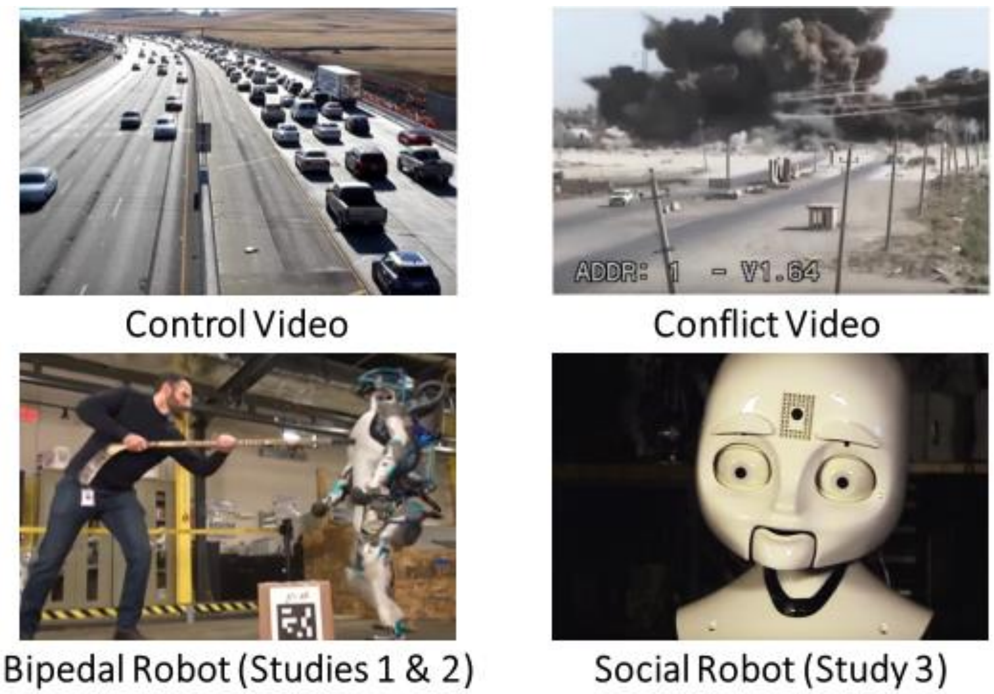

Fig. 1. Top panels: Still images taken from the Control and Conflict video stimuli used in all studies. Bottom panels: Still images taken from the highly mobile bipedal humanoid robot video stimuli used in Studies 1 and 2 [Boston Dynamics 2016] and from the highly expressive social robot video used in Study 3 [Chandler 2008].

Control Condition Text:

> "This is a recording taken from a traffic camera. This brief video provides a sense of what it is like on the highway. Watch the flow of traffic, and try to notice details such as the scenery, or any differences between the two sides of the highway. Can you guess in what country and year this video was taken?"

[2] Follow-up tests confirmed that the overall pattern of significant relationships observed in Studies 1-3 (including mediation results) persists if including all participants from the unfiltered samples, as l arge samples are relatively robust to noise. The full datasets are archived online and may be accessed at https://osf.io/kjw2e/.

Conflict Condition Text:

> "This is a recording taken from a traffic camera in Iraq. This brief video may provide some sense of what soldiers and local people experience in war zones. Watch the slow-moving truck in the top left corner moving down the road toward the camera. You will see that it blocks the road at first, then moves to one side. The truck is carrying an IED (improvised explosive device) and is about to be passed by a military convoy."

We next confirmed that participants had attended to the video by asking them to identify which of the following they had viewed: "trees," "cars," a "cloud of smoke," "airplanes," and a "shaking camera". Participants who failed to report viewing cars were dropped from the study prior to analysis, as were any individuals assigned to the Conflict condition who failed to report viewing a cloud of smoke or a shaking camera.

Composite negative affective response to the video was assessed by averaging reported feelings of sadness, anger, fear, and tension, presented in random order and rated on a 5-point scale (1 = *Not at all*; 2 = *A little*; 3 = *Moderately*; 4 = *Quite a bit*; 5 = *Extremely*; a = .93).

Next, participants viewed the video of the humanoid robot, accompanied by the following text:

> "This is a video of a humanoid robot that is currently under development. This sort of robot may soon be part of daily life, helping to perform routine work and even rescue services to save people in disasters."

The robot stimulus was modified from a recent video showcasing Atlas, a humanoid robot currently under development by Boston Dynamics (2016). The video begins by depicting the robot successfully traversing rough outdoor terrain. The robot is next depicted picking up and moving a box indoors. A human man then appears and, armed with a hockey stick, knocks the box out of the robot's grip and roughly pushes the robot in the chest, knocking it back several feet. Next, the man is shown battering the robot in the back still more forcefully, causing the robot to topple forward onto the ground. Finally, the robot is shown rising to a standing position (see the Supplemental Electronic Material to access the video).

Next, in random order, ratings of the robot's intelligence, sympathy for the robot, and perceptions of the robot as a person were solicited. Composite assessments of the robot as intelligent were created by averaging ratings of the robot as 'intelligent' and 'skillful', rated on the same 5-point scale used previously, $r(222) = .64$, $p < .001$. Composite assessments of feelings of sympathy for the robot were created by averaging self-reported responses to the portion of the video in which the man was shown forcefully hitting the robot. Items assessing the extent to which participants felt "sympathetic," "sorry," and "wished the man would stop" were presented in random order and rated on the same 5-point scale (a = .93). Representations of the robot as possessing emotional capacities—here, to suffer–were gauged based on participants' own spontaneous feelings of sympathy toward it because this may be a more effective way of tapping intuitions of the robot as a feeling agent, as opposed to asking participants to state their declarative beliefs regarding the robot's putative emotional states (i.e., the 'correct' answer would be that the robot lacks emotion). Participants were asked to rate the perceived personhood of the robot based on a 100-point scale using a horizontal slider interface: "Overall, how much like a person did the robot seem to you?" (1 = *Not at all*; 100 = *Completely*).

We next confirmed that participants had attended to the robot video by asking them to identify which of the following they had viewed: "Baseball," "Hockey stick," and "Boxes". Participants who failed to report viewing a hockey stick were dropped prior to analysis.

Finally, participants answered demographic items, two catch questions (e.g., "How many letters are in the English alphabet?"), and questions confirming that the video playback had functioned correctly and that they had watched both videos attentively.

Table 1. Mean Effects of Conflict Manipulation on Robot Perception (Studies 1 and 2)

	Control *Mean (SD)*	Conflict *Mean (SD)*	*F*	*p*	η^2_p	95% CI
Study 1:						
Sympathetic	3.24 (1.34)	2.83 (1.33)	5.06	.026	.02	.05, .75
Intelligent	3.46 (1.01)	3.37 (1.09)	.39	.535	.00	-.19, .37
Personhood	53.67 (26.92)	45.28 (26.32)	5.56	.019	.02	1.38, 15.40
Study 2:						
Sympathetic	2.68 (1.39)	2.36 (1.22)	5.27	.022	.02	.05, .60
Intelligent	3.29 (1.02)	3.08 (1.07)	3.54	.061	.01	-.01, .43
Personhood	50.36 (28.31)	42.97 (28.17)	5.87	.016	.02	1.39, 13.39
Combat lethality	62.89 (27.65)	56.74 (28.20)	4.31	.039	.01	.32, 11.97

Note. Study 1: $N = 224$. Study 2: $N = 345$. Study 2 contrasts control for Framing condition.

2.2 Results

2.2.1 Manipulation check: effects of Conflict prime on state negative affect.

In Study 1 (and in all subsequent studies), preliminary tests confirmed that the data distributions conformed with the assumptions required to conduct parametric tests. The full datasets for all studies are archived and available for download at https://osf.io/kjw2e/.

As intended, an analysis of variance revealed that state negative affect was significantly greater in the Conflict condition relative to control, confirming that participants experienced the experimental manipulation as subjectively threatening (for detailed descriptives and contrasts, see Appendix Table A1). However, as there were no significant associations between composite negative affective reactions and feelings of sympathy (p = .65), attributions of intelligence (p = .15), or perceptions of personhood (p = .32), negative affect is not included in subsequent analyses.

2.2.2 Effects of the Conflict prime on feelings of sympathy, attributions of intelligence, and perceptions of personhood.

A series of analyses of variance confirmed that Feelings of sympathy and perceptions of personhood were both significantly reduced in the Conflict condition relative to control, with no such effect obtaining with regard to attributions of intelligence (see Table 1). Feelings of sympathy, attributions of intelligence, and perceptions of personhood were all positively correlated (see Table 2).

2.2.2.1 Reduced feelings of sympathy mediate the effect of the Conflict prime on perceived personhood.

We conducted a mediation test to assess whether diminished sympathy mediated the decreased perception of personhood observed in the Conflict condition (see Figure 2, top panel). For these (and all subsequent) mediation analyses, we utilized the bias-corrected bootstrapping procedure (5,000 samples) found in the INDIRECT macro for SPSS (Preacher & Hayes, 2008).[3] We entered condition as the independent variable, sympathy as the mediating variable, and perceived personhood as the dependent variable. As predicted, feelings of relatively diminished sympathy fully mediated the effects of the video condition on the perceived personhood of the robot. The direct effect of condition on perceived personhood (b = -8.39, SE = 3.56, β = -.16, p = .019) was no longer significant in the model (b = -4.52, SE = 3.16, β = -.08, p = .154), whereas the indirect effect of sympathy on perceived personhood remained significant (b = 9.64, SE = 1.17, β = .48, p < .001), and the confidence intervals did not overlap with zero (95% CI = [-7.78, -.64]).

Table 2. Bivariate Correlations Among Feelings of Sympathy, Attributions of Intelligence, Perceptions of Personhood (Studies 1 and 2), and Estimated Combat Lethality (Study 2)

Study 1:	*1*	*2*	*3*	
1. Sympathy	--	.37	.50	
2. Intelligence		--	.50	
3. Personhood			--	
Study 2:	*1*	*2*	*3*	*4*
1. Sympathy	--	.32	.56	.11*
2. Intelligence		--	.56	.37
3. Personhood			--	.25
4. Combat lethality				--

Note. Study 1: N = 224. Study 2: N = 345. * p = .047; all other ps < .001. Study 2 correlations control for Framing condition.

[3] The INDIRECT process estimates the path coefficients in a mediator model and generates bootstrap confidence intervals for total and specific indirect effects of a predictor variable on an outcome variable through one or more mediator variable(s), adjusting all paths for the potential influence of covariates not categorized in the model as potential mediators (Preacher & Hayes, 2008).

2.3 Discussion

Study 1 was primarily designed to test whether exposure to cues of violent intergroup conflict would modulate perceptions of the robotic target as a person. Indeed, participants exposed to the video of the IED explosion reported that the bipedal robot seemed less like a person. The findings of Study 1 also provide support for the hypothesis that perceived robotic personhood is closely intertwined with attributions of subjective emotional qualia, as feelings of sympathy for the robot mediated the effect of the conflict condition on perceived personhood. However, consistent with Gray and colleagues' [2007] two-facet model of mind representation, attributions of intelligence, while unaffected by the conflict prime, were also equivalently positively correlated with perceived robotic personhood. Interestingly, although the conflict video elicited a striking increase in self-reported feelings of negative affect (see Appendix Table A1), there was no apparent relationship between the emotional component of the threat response and perceptions of the robot, indicating that the psychological pathway by which threat modulates robot perceptions is either non-affective in nature, or involves an implicit affective response which is not readily captured via self-report.

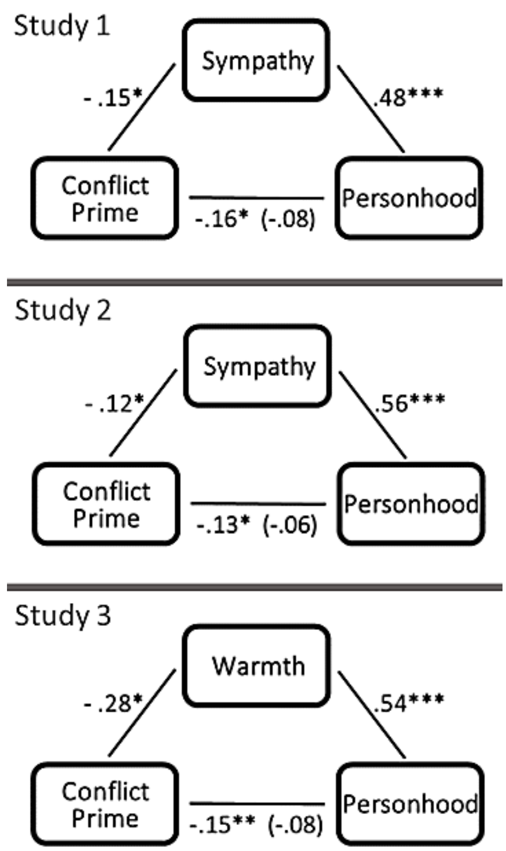

Fig. 2. Standardized regression coefficients for the relationship between the Conflict manipulation and the perceived personhood of the robot as mediated by feelings of sympathy (Studies 1 and 2) or warmth (Study 3) for the robot. The standardized regression coefficient between the Conflict prime and perceived personhood with the mediator included in the model is given in parentheses. * $p < .05$; ** $p < .01$; *** $p < .001$.

As sketched in the Introduction, a wide parallel literature on the role of threat in magnifying group bias suggests that identifying with robots as coalition members versus out-group members may importantly determine whether contexts of threat increase or decrease perceptions of robotic personhood. In Study 2, to attempt to test this possibility, the robot evaluation target was explicitly framed as either a coalitional ally or a non-ally from the point of view of a U.S. participant sample. In addition, to test whether cues of violent conflict influence perceptions of the efficacy of robots within conflictual contexts, a new measure of the robot's estimated battlefield combat lethality was also included.

3. STUDY 2

3.1 Participants and overview of procedure

400 U.S. participants were recruited, compensated and pre-screened as in Study 1, leaving a final sample of 345 adults (51.0% female; 81.2% White) ranging in age from 18 to 74 ($M = 37.81$, $SD = 12.12$).

In a 2 X 2 between-subjects design, participants were randomly assigned to view either the Control ($N = 182$) or Conflict ($N = 163$) videos utilized in Study 1, then evaluate the humanoid robot, now framed as in development for use in military combat alongside either United States (Ally condition; $N = 184$) or Russian (Non-Ally condition; $N = 161$) forces.

We next confirmed that participants had attended to the video using the same check questions utilized previously. Negative affective reactions to the video were then collected and composited as in Study 1 ($\alpha = .92$).

Next, participants viewed the humanoid robot using the same video as employed in Study 1, now framed as designed to serve an explicitly military function, and as either an Ally or Non-Ally:

Ally Condition Text:

> "This is a video of a humanoid robot that is currently under development by the United States Department of Defense. These robots will soon play a key role in American military combat, fighting alongside American soldiers on the battlefield."

Non-Ally Condition Text:

> "This is a video of a humanoid robot that is currently under development. Russia has shown a particularly strong interest in using humanoid robots in combat. These robots will soon play a key role in Russian military combat, fighting alongside Russian soldiers on the battlefield."

Next, in random order, ratings of the robot's intelligence, $r(343) = .65$ ($p < .001$), sympathy for the robot ($\alpha = .95$), and perceptions of the robot as a person were collected as in Study 1. In addition, participants were asked to rate the potential lethality posed by the robot in combat: "After the design is finished and the robot is deployed in combat, how potentially dangerous to its enemies will it be?" (1 = *Not at all*; 100 = *Extremely*). We next confirmed that participants had attended to the robot video using the same method as in Study 1.

Finally, participants answered demographic items, catch questions, and questions confirming that the video playback had functioned correctly and that they had watched both videos attentively.

3.2 Results

3.2.1 Manipulation check: effects of conflict prime on state negative affect.

As in Study 1, an analysis of variance revealed that state negative affect was significantly greater in the Conflict condition relative to control (for detailed descriptives and contrasts, see Appendix Table A1). Also as observed in Study 1, there were no significant associations between composite negative affective reactions and feelings of sympathy (p = .76), attributions of intelligence (p = .12), or perceptions of personhood (p = .26), nor was there an apparent association with the measure of estimated combat lethality added in Study 2 (p = .40).

3.2.2 Effects of the Framing manipulation on perceptions of the robot.

Analyses of variance revealed no main effects of the Framing manipulation (Ally vs. Non-Ally) on feelings of sympathy (p = .15), attributions of intelligence (p = .24), or perceptions of personhood (p = .94). However, participants did rate the Non-Ally robot to be significantly more lethal in combat (M = 65.54, SD = 26.85) than did participants who rated the robot framed as an Ally (M = 55.12, SD = 28.22), F(1, 343) = 12.25, p = .001, η^2_p = .03, 95% CI [-16.28, -4.57]. Against predictions, there were no significant two-way interactions between the Framing manipulation and the Conflict video manipulation on any of the ratings of the robot, ps .53 - .93. Consequently, the Framing condition is not considered further, but is included as a covariate in all subsequent analyses of the effects of the Conflict manipulation.

3.2.3 Effects of the Conflict prime on feelings of sympathy, attributions of intelligence, and perceptions of personhood.

As in Study 1, feelings of sympathy and perceptions of personhood were both significantly reduced in the Conflict condition relative to control, with no such effect obtaining with regard to attributions of intelligence (although, unlike in Study 1, there was a nonsignificant trend in the same direction; see Table 1). Also as observed in Study 1, feelings of sympathy, attributions of intelligence, and perceptions of personhood were all positively correlated (see Table 2).

3.2.3.1 Reduced feelings of sympathy mediate the effect of the Conflict prime on perceived personhood.

We conducted a mediation test to assess whether diminished sympathy mediated the decreased perceptions of personhood observed in the Conflict condition (see Figure 2, middle panel). We entered condition as the independent variable, sympathy as the mediating variable, and perceived personhood as the dependent variable. Replicating the pattern observed in Study 1, feelings of relatively diminished sympathy fully mediated the effects of the video condition on the perceived personhood of the robot. The direct effect of condition on perceived personhood (b = -7.39, SE = 3.05, β = -.13, p = .016) was no longer significant in the model (b = -3.51, SE = 2.56, β = -.06, p = .171), whereas the indirect effect of sympathy on perceived personhood remained significant (b = 11.98, SE = .98, β = .56, p < .001), and the confidence intervals did not overlap with zero (95% CI = [-7.24, -.73]). (Follow-up tests confirmed that dropping the

Framing condition as a covariate in the model does not alter the pattern or significance of the results.)

3.2.4 Effects of Conflict prime on estimated combat lethality.

Participants in the Conflict condition estimated the robot to be significantly less lethal in combat relative to control participants (see Table 1). The estimated combat lethality of the robot was positively correlated with attributions of intelligence, perceptions of personhood, and feelings of sympathy (see Table 2).

3.2.4.1 Reduced perceptions of personhood mediate the effect of the Conflict prime on estimated combat lethality.

We next conducted a mediation test to assess whether the significantly diminished perception of the robot as a person observed in the Conflict condition mediated the decrease in estimations of the robot's combat lethality (see Figure 3). We entered condition as the independent variable, perceived personhood as the mediating variable, and estimated lethality as the dependent variable. Indeed, diminished perceptions of the robot as a person fully mediated the effect of the video condition on the estimated lethality of the robot. The direct effect of condition on estimated lethality (b = -6.15, SE = 2.96, β = -.11, p = .039) was no longer significant in the model (b = -4.45, SE = 2.91, β = -.08, p = .127), whereas the indirect effect of perceived personhood remained significant (b = .23, SE = .05, β = .23, p < .001), and the confidence intervals did not overlap with zero (95% CI = [-3.86, -.35]). (Follow-up tests confirmed that neither dropping the Framing condition as a covariate in the model, nor including feelings of sympathy for the robot as an added covariate, alter the pattern or significance of the results.)

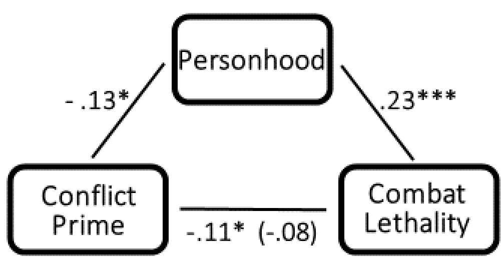

Fig. 3. Standardized regression coefficients for the relationship between the Conflict manipulation and the estimated combat lethality of the robot as mediated by perceptions of the robot as a person (Study 2). The standardized regression coefficient between the Conflict prime and estimated combat lethality with the mediator included in the model is given in parentheses. * p < .05; *** p < .001.

3.3 Discussion

Study 2 replicated and extended the findings of Study 1 by manipulating the coalitional affiliation of the robot evaluation target, now framed as designed for combat. Against expectations, framing the robot as designed for use by United States military forces (the Ally condition) did not alter perceptions of personhood, attributions of intelligence, or feelings of sympathy relative to framing the robot as designed for

Russian military use (the Non-Ally condition). Speculatively, electing to frame the robot as overtly military in function may have obscured otherwise detectable differences related to coalitional affiliation (e.g., by foregrounding the use of the robot in violent conflict). Consistent with this possibility, feelings of sympathy for the militaristically framed robot in Study 2 were lower than those elicited by the same robot framed as designed to assist in rescue services (see Table 1). Future work is required to test this interpretation, but regardless of the mechanism, the present data suggest that, at least among a sample that does not regularly interact with robotic teammates and in the absence of other information, military robots may be intuitively perceived to possess roughly equivalent personhood, intellectual ability, and emotional experience regardless of their coalitional status. Study 2 also provided evidence that non-allied military robots may tend to be regarded as more lethal than in-group robots, albeit for reasons that appear orthogonal to the attributions of mental states, as the robot was estimated to be more lethal when framed as Russian. However, it should also be kept in mind that the unexpected absence of effects of the coalitional manipulation on perceived personhood or sympathy may reflect variation among participants in the extent to which Russia was regarded as a potential ally or not with regard to joint military operations. Future attempts to manipulate coalitional affiliation might select a less ambiguously non-allied society (e.g., North Korea at the time of writing).

In Study 2, participants primed with cues of violent conflict rated the robot target as less potentially lethal in combat, an effect which was fully mediated by a parallel reduction in perceived personhood (see Figure 3). This finding is particularly striking given that, as in Study 1, shifts in feelings of sympathy for the robot fully mediated the effect of the conflict condition on the perceived personhood of the robot. Thus, it would seem that experiencing the robot target as less emotionally sympathetic—and hence less possessed of personhood—contributed to the reduction in the estimated fighting effectiveness of the robot. Tentatively, such a dynamic might be taken as suggestive that attributions of emotionality, possibly including affective motivations to fight, are incorporated into intuitions about fighting ability. However, the reader should bear in mind that in Study 2, albeit in a nonsignificant trend, attributions of intelligence to the robot target were also lower, which may have contributed to the link between reduced perceived personhood and reduced estimated lethality (i.e., inasmuch as intelligence connotes the ability to form and execute effective combat strategies).

Although the entire pattern of significant relationships observed in Study 1 replicated in Study 2, despite the shift to a militaristic framing and the introduction of explicit coalitional affiliations, the possibility remains that the results are artifacts of the robot video stimuli employed. In particular, the mobile bipedal robot depicted in the video lacks facial characteristics or voice interactive capabilities. Although the capacity to navigate variable, uneven terrain and dynamically track objects displayed by the bipedal robot is impressive, lay participant observers may not have appreciated the intelligence required to support these abilities. By contrast, a system capable of voice interaction and natural language may be regarded as inherently more intelligent, conceivably altering the impact of threat cues on perceptions of the robot. Moreover, robots with evident social intelligence may be immune to reductions in perceived personhood following threat cues. Therefore, to test the generalizability of the prior results to more anthropomorphic robots, a highly social robot equipped with expressive humanlike facial and vocal characteristics was employed in Study 3.

4. STUDY 3

4.1 Participants and overview of procedure

350 U.S. participants were recruited, compensated and pre-screened as in Studies 1 and 2, leaving a final sample of 286 adults (48.3% female; 82.2% White) ranging in age from 18 to 71 (M = 36.58, SD = 11.27).

In a between-subjects design, participants were randomly assigned to view either the Control (N = 141) or Conflict (N = 145) videos utilized previously. We confirmed that participants had attended to the video using the same check questions utilized previously, and negative affective reactions to the video were then collected and composited as before (α = .92).

Next, participants viewed a brief video of a humanlike robot featuring a neotenous, emotionally expressive face and a childlike speaking voice [Chandler 2008]. In the video, the robot demonstrates its capacity to move, grasp with its hands, and produce several facial expressions (sadness, anger, confusion, excitement, boredom) with accompanying shifts in voice prosody (see Figure 1; see the Supplemental Electronic Material to access the video). The video was framed with the following text:

> "This is a video of a humanoid robot that is currently under development. This sort of robot will soon play a key role in many aspects of life, including assisting in medical or emergency situations."

Next, ratings of the robot's intelligence, $r(284) = .66$ ($p < .001$), and perceptions of the robot as a person, were collected as in Studies 1 and 2, in addition to measures of participants' feelings of affiliative warmth toward the robot and estimates of its potential usefulness in emergencies. Participants' perceptions of the robot as interpersonally warm and affiliative were assessed by averaging agreement with feeling "sympathetic," "friendly," and "want to interact with it," presented in random order, and rated on the same 5-point scale used previously (1 = *Not at all*; 2 = *A little*; 3 = *Moderately*; 4 = *Quite a bit*; 5 = *Extremely*; α = .82). Given the childlike nature of the robot, rather than measuring its potential combat lethality, we asked participants to rate its utility in emergency situations, "After the design is finished and the robot is deployed in medical or emergency situations, how potentially helpful will it be?" (1 = *Not at all*; 100 = *Extremely*).

We next confirmed that participants had attended to the robot video. Participants who did not report viewing a white robot were dropped prior to analysis, as were participants who erroneously reported viewing an orange robot. Finally, participants answered demographic items, catch questions, and questions confirming that the video playback had functioned correctly and that they had watched both videos attentively.

Table 3. Mean Effects of Conflict Manipulation on Robot Perception (Study 3)

	Control *Mean (SD)*	Conflict *Mean (SD)*	*F*	*p*	η^2_p	95% CI
Warmth	2.83 (1.07)	2.55 (.98)	5.20	.023	.02	.04, .51
Intelligence	3.46 (.95)	3.23 (.94)	3.37	.068	.01	-.02, .43
Personhood	51.40 (25.94)	43.57 (24.82)	6.81	.010	.02	1.92, 13.74
Emergency utility	71.38 (23.65)	67.60 (22.69)	1.91	.169	.01	-1.32, 9.18

Note. N = 286.

4.2 Results

4.2.1 Manipulation check: effects of the Conflict prime on state negative affect.

As in Studies 1 and 2, an analysis of variance revealed that state negative affect was again significantly greater in the Conflict condition relative to control (for detailed descriptives and contrasts, see Appendix Table S1). As previously, there were no significant associations between negative affect and perceptions of personhood (p = .32), no were there significant correlations with the measures of warmth (p = .65) or estimated emergency utility (p = .32) added in Study 3. However, in a departure from the prior results, negative affect did positively correlate with attributions of intelligence, $r(284) = .14$, $p = .019$.

4.2.2 Effects of the Conflict prime on feelings of warmth, attributions of intelligence, and perceptions of personhood.

Comparably to the pattern observed in Studies 1 and 2, feelings of emotional warmth and perceptions of personhood were both significantly reduced in the Conflict condition relative to control, with no such effect obtaining with regard to attributions of intelligence or estimates of emergency utility (see Table 3). Also comparably to the results of Studies 1 and 2, feelings of warmth, attributions of intelligence, perceptions of personhood, and estimated emergency utility were all positively correlated (see Table 4).

4.2.2.1 Reduced feelings of warmth mediate the effect of the Conflict prime on perceived personhood.

We conducted a mediation test to assess whether diminished feelings of warmth mediated the decreased perception of personhood observed in the Conflict condition (see Figure 2, bottom panel). We entered condition as the independent variable, feelings of warmth as the mediating variable, and perceived personhood as the dependent variable. As predicted, feelings of relatively diminished warmth fully mediated the effects of the video condition on the perceived personhood of the robot. The direct effect of condition on perceived personhood (b = -7.83, SE = 3.00, β = -.15, p = .010) was no longer significant in the model (b = -4.15, SE = 2.56, β = -.08, p = .106), whereas the indirect effect of feelings of warmth on perceived personhood remained significant (b = 13.37, SE = 1.24, β = .54, p < .001), and the confidence intervals did not overlap with zero (95% CI = [-6.94, -.46]).

Table 4. Bivariate Correlations Among Feelings of Warmth, Attributions of Intelligence, Perceptions of Personhood, and Estimated Emergency Utility (Study 3)

	1	*2*	*3*	*4*
1. Warmth	--	.52	.55	.42
2. Intelligence		--	.56	.54
3. Personhood			--	.50
4. Emergency utility				--

Note. N = 286; all ps < .001.

4.3 Discussion

In Study 3, participants evaluated a relatively small and distinctly social robot, in contrast to the large, faceless bipedal robot utilized in Studies 1 and 2. Despite this change, and despite updating the measure of perceived emotionality of the robot from feelings of sympathy to feelings of emotional warmth, the same pattern replicated once again. Participants exposed to conflict cues responded less warmly to the social robot, and this reduction in warmth mediated an observed reduction in perceived personhood. Also as in the prior studies, there was no significant reduction in attributions of intelligence, nor were the reductions in perceived personhood or feelings of warmth related to negative affect evoked by the threat induction.

Departing from the prior results of Study 2, in which the conflict prime caused a decrease in the estimated combat lethality of the robot, no parallel effect on the estimated utility of the social robot in medical or emergency situations was found in Study 3. Of potential relevance, there was a marked difference in the strength of the correlations between perceived personhood, emotional responsiveness, and task effectiveness between Studies 2 and 3. Whereas estimated combat lethality was positively associated with feelings of sympathy and perceived personhood in Study 2, estimated emergency utility was more strongly correlated with both feelings of emotional warmth and perceived personhood in Study 3 (i.e., as though the robot's apparent warm emotional qualities would bolster its helpfulness in a crisis; compare Tables 2 and 4). Follow-up work is needed to confirm whether conflict cues diminished estimated efficacy in Study 2, but not Study 3, due to differences in the strength of the links between perceived emotionality and personhood in reasoning about military versus non-military domains of performance.

5. CONCLUSIONS

These experiments comprise the first direct tests of whether cues of threat can influence perceptions of the personhood, intelligence, or emotionality of machine agents. Across multiple studies, utilizing physically and behaviorally distinct robot evaluation targets, participants primed by witnessing violent conflict perceived the robot as less like a person, a reduction that was consistently mediated by reduced feelings of emotional connection. By contrast, there were no significant effects of the threat manipulation on attributions of intellectual ability in any study. This dissociation between attributions of intelligence and emotionality in driving perceptions of robotic personhood agrees with prior work linking robot anthropomorphism with neural mechanisms thought to represent emotional experience in machine agents [Takahashi et al. 2014; Krach et al. 2008], as well as with the previous findings of Gray and colleagues [2007].

The present observations may appear somewhat at odds with prior findings that, for example, military personnel working under threatening circumstances appear to form personal relationships with robotic teammates [Carpenter 2013; Fincannon et al. 2004]. However, there are likely to be critical differences between the effects of brief exposure to video stimuli depicting robots and the gradual development of a social rapport with the postulated mind of a robot teammate over hours, days, weeks, or months. As most individuals at the time of writing have very limited exposure to robots, the participants in the present studies were presumably unfamiliar with robots as prospective partners in collaborative endeavors, and may have implicitly categorized the robots they were asked to evaluate as akin to out-group members. Indeed, the present results with regard to the impact of threat cues on perceptions of robots as less human parallel prior findings that out-group members are viewed as less human than in-group members [e.g., Brewer 1999; Haslam 2006], and that cues of threat exacerbate coalitional biases in perceptions of out-group members [Jonas et al. 2014]. As robots become increasingly ubiquitous and anthropomorphic, baseline attitudes towards robots may evolve to the point that robots are not reflexively regarded as out-group members, and hence are not viewed as less human under conditions of threat. Future research examining the impact of familiarity, anthropomorphism and implicit group categorization will be needed to clarify these potentially changing dynamics.

Further investigation of the effects of threat on the perceived personhood of robots should also incorporate actual human-robot interaction. Whereas the present studies rely on 2D video recordings of robots, recent work has found that participants in an emergency situation exhibited extreme levels of trust in a robot in their presence [Robinette et al. 2016]. To the extent that such over-reliance indexes heightened perceptions of personhood, intellect, or emotional warmth, it may be the case that physical robots are intuitively viewed as more possessed of personhood merely because they are physically instantiated. If so, then the present studies may indicate that threat cues exert differing, or even opposite effects on perceptions of machine agents when experienced via screen-mediated software versus physical human-robot interaction. Alternately, it may be the case that actively interacting with a physical robot (as the participants in Robinette et al.'s study had done prior to the simulated emergency) induces affiliation with the robot as a kind of ally, and it is this experience of affiliative interaction which increases trust and concomitant mental attributions once threat arises. Research comparing actual and virtual human-robot interaction is required to adjudicate between these possibilities.

The studies reported here provide the first evidence that threatening stimuli can diminish feelings of emotional connection with robots, perceptions of their personhood, and, at least in conflictual contexts (Study 2), confidence in their operational capacity. These findings militate for studies targeting whether these effects translate to meaningful differences in real-world human-robot team performance. On the one hand, the effects of threat on actual reliance on machine agents may be negligible and safely disregarded. On the other, even small psychological biases may lead to dramatic consequences in life-or-death circumstances. Future work should employ behavioral measures of reliance on the recommendations of robotic partners, ideally focusing on decisions of relevance under conditions of conflict (e.g., distinguishing enemies from allies or noncombatants, determining whether to use force, navigation).

The present results were also obtained using a notably weak threat induction stimulus—a brief, silent video viewed far from actual hazard. Thus, future work in this area should employ immersive threat simulations (e.g., using virtual reality). Although the threat induction utilized across these large-sample studies yielded highly significant and replicable shifts in judgment, the effect sizes were relatively small, possibly due to the weak nature of the stimulus. Immersive, realistic experiences of threat may evoke larger biases. Should substantive effects be observed, it will be equally important to identify design choices that 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 sympathy and perceived 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. In short, the present findings are valuable primarily insofar as they inspire future research and design directions to gain traction over a potentially important bias coloring humans' perceptions of machine agents when under threat.

Finally, although the experimental threat induction did reliably decrease perceptions of robotic personhood in the present findings, participants in both the control and threat conditions of all three studies also exhibited remarkable tendencies to attribute personhood and experience warm emotional responses to machines. The propensity for individuals to perceive the automata that they operate or work beside as human will almost certainly increase as machine agents become more overtly human in appearance and linguistic ability, given that cognitive [Knijnenburg and Willemsen 2016] and neural [Takahashi et al. 2014] mechanisms related to perceptions of mind have been found to activate in proportion to human resemblance. The present studies demonstrate that threat can reduce perceptions of robotic personhood and mental states, but further research is required to ascertain the extent to which threat-modulated shifts in mind perception behaviorally influence trust and reliance in automated agents, and to map the range of situational, personality, and design factors shaping our intuitive understanding of tools as teammates.

APPENDIX

Table A1. Mean Effects of Conflict Manipulation on State Negative Affect (Studies 1-3)

	Control Mean (SD)	*Conflict Mean (SD)*	*F*	*p*	η^2_p	*95% CI*
Study 1:						
Anger	1.07 (.35)	3.10 (1.39)	218.37	<.001	.50	-2.30, -1.76
Sadness	1.10 (.30)	3.53 (1.23)	394.77	<.001	.64	-2.66, -2.18
Fear	1.14 (.40)	2.88 (1.33)	171.55	<.001	.44	-2.00, -1.48
Tension	1.53 (.84)	3.57 (1.11)	237.99	<.001	.52	-2.30, -1.78
Composite Neg, Affect	1.21 (.34)	3.27 (1.04)	388.62	<.001	.64	-2.27, -1.85
Study 2:						
Anger	1.17 (.56)	2.98 (1.31)	288.38	<.001	.46	-2.02, -1.60
Sadness	1.17 (.56)	3.47 (1.34)	450.32	<.001	.57	-2.52, -2.09
Fear	1.15 (.51)	2.91 (1.30)	280.53	<.001	.45	-1.96, -1.55
Tension	1.46 (.77)	3.58 (1.13)	420.29	<.001	.55	-2.33, -1.92
Composite Neg, Affect	1.24 (.48)	3.24 (1.01)	564.11	<.001	.62	-2.16, -1.83
Study 3:						
Anger	1.14 (.41)	2.97 (1.46)	204.61	<.001	.42	-2.08, -1.57
Sadness	1.23 (.58)	3.28 (1.29)	298.31	<.001	.51	-2.28, -1.82
Fear	1.16 (.56)	2.48 (1.33)	117.64	<.001	.29	-1.55, -1.08
Tension	1.52 (.88)	3.41 (1.19)	232.09	<.001	.45	-2.13, -1.65
Composite Neg, Affect	1.26 (.47)	3.03 (1.10)	308.99	<.001	.52	-1.97, -1.57

Note. Study 1: *N* = 224. Study 2: *N* = 345. Study 3: *N* = 286. Study 2 contrasts control for Framing condition.

ACKNOWLEDGMENTS

The author would like to thank Jennifer Hahn-Holbrook, the members of UCLA's Experimental Biological Anthropology Lab, and members of the AFOSR Trust and Influence Program for helpful feedback.

REFERENCES

Michael J. Barnes, Jessie Y.C. Chen, Florian Jentsch, and Elizabeth S. Redden. 2011. Designing effective soldier-robot teams in complex environments: Training, interfaces, and individual differences. *Engineering Psychology and Cognitive Ergonomics Lecture Notes in Computer Science.* 484–493. DOI:http://dx.doi.org/10.1007/978-3-642-21741-8_51

Michael J. Barnes, Jessie Y.C. Chen, Florian Jentsch, Elizabeth Redden, and Kenneth Light. 2013. An overview of humans and autonomy for military environments: Safety, types of autonomy, agents, and user interfaces. *Lecture Notes in Computer Science Engineering Psychology and Cognitive Ergonomics. Applications and Services.* 243–252. DOI:http://dx.doi.org/10.1007/978-3-642-39354-9_27

Barylick, C. (February 2006). iRobot's PackBot on the front lines. *United Press International.* Retrieved from: http://www.upi.com/Hi-Tech/view.php?StoryID=20060223-113933-8397

Boston Dynamics (November 2016). Atlas – The agile anthropomorphic robot. Retrieved from: http://www.bostondynamics.com/robot_Atlas.html

Marilynn B. Brewer. 1999. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, 55, 429-444. DOI:http://dx.doi.org/10.1111/0022-4537.00126

Michael Buhrmester, Tracy Kwang, and Sam Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science,* 6, 3-5. DOI:http://dx.doi.org/10.1037/e527772014-223

Julie Carpenter. 2013. The quiet professional: An investigation of US military explosive ordnance disposal personnel interactions with everyday field robots (Unpublished doctoral dissertation). University of Washington: Seattle.

Chandler, L. (April 2008). Meet Nexi, the Media Lab's latest robot and Internet star. *MIT News.* Retrieved from: http://news.mit.edu/2008/nexi-0409

Silvia Coradeschi, et al. 2006. Human-inspired robots. *Intelligent Systems, IEEE,* 21, 74-85. DOI:10.1109/MIS.2006.72

Charles Darwin. 1873. *The descent of man.* New York: Appleton.

Carsten De Dreu, Lindred L. Greer, Gerben A. Van Kleef, Shaul Shalvi and Michael J. J. Handgraaf. 2011. Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences of the United States of America,* 108, 1262-1266.

John F. Dovidio and Sam L. Gaertner, 2010. Intergroup bias. In S. T. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (Vol. 2, 5th ed., pp. 1084-1121). New York: Wiley.

Charles Efferson, Rafael Lalive, and Ernst Fehr. 2008. The coevolution of cultural groups and ingroup favoritism. *Science*, 321, 1844-1849.

Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114, 864-886. DOI:http://dx.doi.org/10.1037/0033-295x.114.4.864

Hobart R. Everett, Estrellina B. Pacis, Greg Kogut, Nathan M. Farrington, and S. Khurana. 2004. Toward a warfighter's associate: Eliminating the operator control unit. In *Proceedings of SPIE: Mobile Robots 18. Philadelphia: International Society for Optics and Photonics.* DOI:http://dx.doi.org/10.1117/12.571458

Thomas Fincannon, Laura E. Barnes, Robin R. Murphy, and Daniel Riddle. 2004. Evidence of the need for social intelligence in rescue robots. In *Proceedings from International Conference on Intelligent Robots and Systems, 2004* (Vol. 2, pp. 1089-1095). Piscataway, NJ: IEEE. DOI:http://dx.doi.org/10.1109/iros.2004.1389542

Joel Garreau. (May 2007). Bots on the ground; In the field of battle (or even above it), robots are a soldier's best friend. *Washington Post.* Retrieved from: http://msl1.mit.edu/furdlog/docs/washpost/2007-05-06_washpost_human_bots.pdf

Gary R. Gilbert and Michael K. Beebe. 2010. United States Department of Defense research in robotic unmanned systems for combat casualty care (No. RTO-MP-HFM-182). U.S. Army Medical Research and Metriel Command Telemedicine and Advanced Technology Research Center. Fort Fetrick, MD: Defense Technical Information Center.

Heather M. Gray, Kurt Gray, and Daniel M. Wegner. 2007. Dimensions of mind perception. *Science*, 315, 619. DOI:http://dx.doi.org/10.1126/science.1134475

Ross A. Hammond and Robert Axelrod. 2006. The evolution of ethnocentrism. *Journal of Conflict Resolution,* 50, 926-936.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and Social Psychology Review,* 10, 252-264. DOI:http://dx.doi.org/10.1207/s15327957pspr1003_4

Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American Journal of Psychology,* 57, 243-259. DOI:http://dx.doi.org/10.2307/1416950

Colin Holbrook. 2016. Branches of a twisting tree: Domain-specific threat psychologies derive from shared mechanisms. *Current Opinion in Psychology,* 7, 81-86. DOI:http://dx.doi.org/10.1016/j.copsyc.2015.08.006

Colin Holbrook, Daniel M. T. Fessler, and Carlos D. Navarrete. 2016. Looming large in others' eyes: Racial stereotypes illuminate dual adaptations for representing threat versus prestige as physical size. *Evolution and Human Behavior*, 37, 67-78. DOI:https://doi.org/10.1016/j.evolhumbehav.2015.08.004

Colin Holbrook, Keise Izuma, Choi Deblieck, Daniel M. T. Fessler, and Marco Iacoboni. 2016. Neuromodulation of group prejudice and religious belief. *Social Cognitive and Affective Neuroscience,* 11, 387-394. DOI:http://dx.doi.org/10.1093/scan/nsv107

John Horton, David Rand, and Richard Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics,* 14, 399-425. DOI:http://dx.doi.org/10.3386/w15961

Hiroshi Ishiguro and Shuichi Nishio. 2007. Building artificial humans to understand humans. *Journal of Artificial Organs,* 10, 133-142. DOI:http://dx.doi.org/10.1007/s10047-007-0381-4

Eva Jonas et al. 2014. Threat and defense: From anxiety to approach. *Advances in Experimental Social Psychology*, 49, 219-286.

Bart P. Knijnenburg and Martijn C. Willemsen. 2016. Inferring capabilities of intelligent agents from their external traits. *ACM Transactions on Interactive Intelligent Systems,* 6, 1–25. DOI:http://dx.doi.org/10.1145/2963106

Karl F. Macdorman and Hiroshi Ishiguro. 2006. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies,* 7, 297-337. DOI:http://dx.doi.org/10.1075/is.7.3.03mac

Winter Mason and Siddharth Suri. 2011. Conducting behavioral research on Amazon's Mechanical Turk. *Behavioral Research Methods,* 44, 1-23. DOI:http://dx.doi.org/10.3758/s13428-011-0124-6

Stephen L. Neuberg, Douglas T. Kenrick, and Mark Schaller. 2010. Evolutionary social psychology. In S. T. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (Vol. 2, 5th ed., pp. 761-797). New York: Wiley.

Scott Ososky et al. 2012. The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. In *Proceedings of SPIE Defense, Security, and Sensing. Baltimore: International Society for Optics and Photonics.* DOI:http://dx.doi.org/10.1117/12.923283

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making,* 5, 411-419.

Kristopher J. Preacher and Andrew F. Hayes. 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods,* 40, 879–891. DOI:http://dx.doi.org/10.3758/brm.40.3.879

David G. Rand. 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology,* 299, 172-179. DOI:http://dx.doi.org/10.1016/j.jtbi.2011.03.004

Paul Robinette et al. 2016. Overtrust of robots in emergency evacuation scenarios. *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* DOI:http://dx.doi.org/10.1109/hri.2016.7451740

Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M. Wegner. 2010. Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14, 383-388. DOI:http://dx.doi.org/10.1016/j.tics.2010.05.006