A Testbed for Evaluating Human and Robot Power Dynamics in Small Group Decision-Making

Campbell Dorsey[1], Marlee Smith[1], Nathan Crouse[1], Brian Bradley[1], Levi Brown[1], Roger I. Baeza[1], Wei-Hung Liao[1], Marcus Berrette[1], Katherine Dingwall[1], Daniel Holman[1], Malte Jung[2], Elizabeth Phillips[3], Ewart de Visser[1*], Richard Niemeyer[1]
[*] *Corresponding Author: ewartdevisser@gmail.com*

[1]*Department of Behavioral Sciences and Leadership*
U.S. Air Force Academy
Colorado Springs CO, USA

[2]*Department of Information Science*
Cornell University
Ithaca NY, USA

[3]*Department of Psychology*
George Mason University
Fairfax VA, USA

***Abstract*— Power shapes team decision-making, yet little is known about how machine-based expert power compares to human expertise in dynamic environments. This study investigates how expert type (human versus robot) and interaction environment (virtual reality versus physical embodiment) influence team perceptions, social dynamics and decision-making during a Desert Survival Task. Using a between-subjects design, teams interact with a standardized expert agent while behavioral logs, self-reports, and performance metrics are collected. Data collection is on-going and findings are expected to clarify how machine expertise alters influence dynamics and decision authority, informing the design and governance of human–machine teams in high-stakes operational contexts.**

Keywords*— Human-machine teaming; human-robot interaction; expert power; virtual reality

I. Introduction

A. Power in Human-Robot Interaction

Power is a central organizing feature of teams, shaping influence, coordination, and decision outcomes through multiple bases, including expert, legitimate, reward, coercive, and referent power [1,2]. As Artificial Intelligence (AI) systems are integrated into human-autonomy teams (HATs, [3]) during mission-critical tasks, machines may be perceived as holders of these power bases, especially expert power, leading humans to defer to their recommendations with little knowledge of their actual capabilities [4,5]. These effects may be further shaped by interaction context, as power cues can vary across virtual reality (VR) and physically embodied environments and within dynamic team structures [6-8]. However, it remains unclear how the expression of expert power by a machine, relative to a human expert, and the environment in which that expertise is encountered jointly affect team perceptions, cohesion, interaction quality, decision outcomes, and responsibility attribution. Clarifying these relationships is essential for evaluating the operational use of machine teammates in military and other high-stakes operational environments.

Previous research has shown that participants demonstrate greater compliance, time investment, and effort when coached by a human compared to various types of robots; however, robots were still able to elicit substantial effort from participants, indicating meaningful machine effectiveness [9]. At the same time, studies on mixed human and machine teams suggest that social structure and power dynamics can outweigh ‘agent type’ as predictors of influence, meaning a powerful robot can sometimes be more influential than a human teammate [10]. In high-stress, high-stakes environments, robots can also introduce uncertainty, causing human teammates to distrust or reject them as potential sources of risk [11]. However, these dynamics shifts when robots demonstrate high levels of knowledge or authority, exhibiting strong expert power [4]. Under these conditions, robots have been shown to exert even greater influence than human teammates, particularly when providing confident or highly structured recommendations [10].

Recent findings further demonstrate that different explanation styles, such as using stronger belief markers or the third person, can shift the degree to which humans rely on an AI’s recommendations, even when trust ratings remain unchanged [12]. This suggests that human reliance on AI may be more sensitive to perceived expertise and authoritative cues than to subjective trust alone. Additionally, Zhang’s study showed that when an AI’s expertise complements human knowledge, reliance on the AI significantly improves team performance, and explanations that reveal the AI’s reasoning can enhance human learning and independent decision accuracy. Furthermore, communication in virtual, text-based environments demonstrate team cohesion and influence patterns [13].

Recent work on human–AI group dynamics further shows that virtual agents can exert measurable social influence through their behavior. That research found that an AI agent exhibiting high task-focused engagement increased group synergy, while agents displaying stronger affective behaviors unexpectedly reduced trust and perceived social presence [14]. Experiments on AI group influence show that when multiple AI agents act as a coordinated bloc, they generate normative pressure that shifts human judgments, increases conformity, and may even trigger resistance once the AI “group” becomes too large or too forceful [14].

This material is based upon work supported by the Air Force Office of Scientific Research under award number 25RT0821

Fig. 1. Virtual Reality Environment of the Desert Survival Task

These findings have solidified the understanding that both humans and robots are capable of effectively influencing human compliance and rapport in team settings. However, it is still not understood how different HAT power dynamics can influence team performance. Therefore, our study will investigate the application of expert power by a robot to better understand the extent of its impact on HAT performance in high stress and high-pressure environments.

B. The Current Study

The objective of this study is to examine how the level of expert power demonstrated by a team's designated expert, either a human or a machine agent (or 'robot'), influences team performance during a virtual Desert Survival Task [15], while also assessing the role of interaction environment. Participants will complete the task in one of two experimental team configurations (see Fig. 2). This will be tested in a virtual reality environment where the Desert Survival Task is played among one of two types of teams in a singular run (see Fig. 1).

The study employs a between-subjects experimental design with two independent variables: expert type (human expert versus robot expert) and interaction environment (VR-based interaction vs. physical embodied interaction). In all conditions, teams consist of two cadet participants and a designated expert advisor. In embodied conditions, cadets will interact with a human confederate or a humanoid robot. To ensure consistency across teams, the human confederate will follow scripted, standardized behaviors and will likewise operate the humanoid robot using a Wizard-of-Oz (WoZ) methodology [16]. The task includes individual item ranking, team discussion, expert input with rationale, and final group ranking. Data collection includes self-report survey measures of team perceptions and cohesion, behavioral and interaction logs, as well as objective group decision quality based on comparison with established expert survival rankings.

Participant performance will be represented with five measures: (a) Level of Positive Team Member Perceptions, (b) Level of Positive Team Member Interactions, (c) Level of Team Cohesion, (d) Level of Group Decision Quality, and (e) Probability of Attributing Responsibility to Humans.

Our experiment will have the following two hypotheses:

H1: Compared to a human agent, the level of expert power utilized by a machine agent is associated with significant changes in (a) Level of Positive Team Member Perceptions, (b) Level of Positive Team Member Interactions, (c) level of Team Cohesion, (d) Level of Group Decision Quality, and (e) Probability of Attributing Responsibility to Humans.

H0: Compared to a human agent, the level of expert power utilized by a machine agent is not associated with significant changes in (a) Level of Positive Team Member Perceptions, (b) Level of Positive Team Member Interactions, (c) level of Team Cohesion, (d) Level of Group Decision Quality, and (e) Probability of Attributing Responsibility to Humans.

II. Methods

A. Participants

During the Spring 2026 semester, participants will be recruited from the Behavioral Science 110 and Leadership 400 classes at the U.S. Air Force Academy (USAFA), with participation remaining strictly voluntary and may be offered as an extra credit opportunity.

Upon arrival, participants will be randomly assigned to an experiment condition using a predetermined split/schedule. This ensures that there will be ample participation in all conditions. We aim to recruit 30 participants for each condition for a total of 120 expected participants. Participation in this study will be entirely voluntary and offered solely as an optional extra-credit opportunity. The capstone advisor, Dr. Niemeyer, will coordinate directly with course instructors to ensure proper administration of extra credit.

This study has been approved by USAFA's IRB. Data will be collected with a survey approved by the IRB. Because these courses primarily enroll fourth-class cadets (C4Cs) at a military institution, particular attention will be given to mitigating concerns related to diminished autonomy. Recruitment materials will clearly articulate that participation is optional, that cadets may decline without any penalty or impact on their course standing, and that members of the research team hold no grading, supervisory, or command authority over participants. No recruitment will occur in person by researchers; all invitations will be distributed by course instructors to further reduce perceived pressure. Cadets will be invited to sign up only if they have availability on "T-days", one of a USAFA's alternate academic days, aligning with scheduled access to the VR Lab and availability of the VR headsets.

B. Experimental Design

Our experiment is a between-subjects design with two independent variables: expert type (human expert versus robot expert) and interaction environment (VR-based interaction versus physical embodied interaction). Expert power is the influence an individual or entity has based on their knowledge in a specific field [1,2]. This power will be measured by having participants interact with an expert during the Desert Survival Task. Expert power will be manipulated in this experiment by either a human expert or a robot expert (see Fig. 2). For the human expert, cadets will receive guidance from a human. As for the robot expert, cadets will receive guidance from the robot in a WoZ style [16], where the robot has pre-scripted instructions and responses. Both agents will use scripted responses to maintain ecological validity. All participants will ultimately be completing the same task, however, they will randomly be assigned an expert agent type unbeknownst to them. Randomly assigning the expert type allows for an even split of participants in each condition to get sufficient data. For the interaction environment, cadets will either participate in a VR-based environment or a physical embodied environment (see Fig. 1). The VR-based environment

Fig. 3. Physical environment of the Desert Survival Task with Furhat Robot

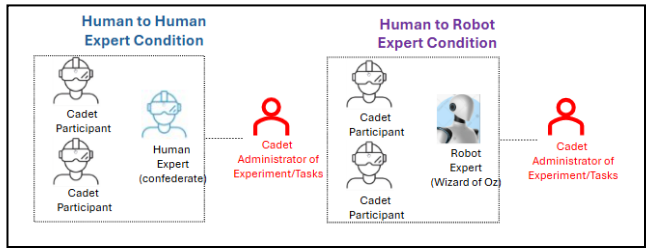

Fig. 2. Experimental Design with Human (left) and Robot (right) experts.

closely mimics the physical environment so an accurate comparison can be made about the actual effect of virtual reality versus being in a physical environment with a robot.

C. Desert Survival Testbed

The Desert Survival Task is a collaborative decision-making exercise in which participants are presented with a list of twenty items that could aid survival if stranded in a desert environment. The team must work collectively to rank the items in order of importance, after which their rankings are compared to an expert benchmark and evaluated based on the group's decision-making processes.

The team members operating the human expert power agent will follow a scripted and standardized list of behaviors to ensure that each group has the exact same experience. Based upon research on the five power types, expert power consists of brief, evidence-based suggestions which will be delivered during the team discussion phase. The study sessions will occur in the Behavioral Science VR Lab, deemed adequate for the VR operation of the study. Within the space are the VR headsets, ample walking room for participants to move around, a station for researchers to take notes and observe.

For the VR condition, a custom-made Desert Survival Task VR testbed will feature a shared desert environment with interactive item tables, embodied avatars, and virtual tablet interfaces that allow multiple participants to collaboratively rank survival resources while navigating physically or via Omni treadmills (see Fig. 1). Participants can gesture, communicate, and manipulate virtual objects, while a researcher-controlled virtual agent or robot provides expert input through a scripted control panel. The system is designed to record movement, interaction, and decision data in real time, supporting detailed analysis of group dynamics and performance. While we are still testing differences based on robotic and human expert agents, the robotic aspect will simply be a WoZ, faux robot. Within the environment the material list consists of: a flashlight, a jackknife, a map of the area, a plastic raincoat, a magnetic compass, a compress kit with gauze, a .45 loaded caliber pistol, a red and white parachute, a bottle with 1000 salt tablets, one quart of water per person, a book entitled *Edible Animals of the Desert,* a pair of sunglasses per person, two quarts of 180 Proof Vodka, one top coat per person, and a cosmetic mirror.

For the physical condition, participants will be seated around a table in a shared physical space and will receive standardized task materials, including the list of survival items and ranking sheets. A humanoid robot (i.e. Furhat), functioning as the expert teammate, will be positioned at the table alongside participants and will be introduced as a knowledgeable source during the scenario (see Fig. 3). The Furhat robot, produced by the company Furhat Robotics, has been used in previous research as an advisor [17,18]. The robot will use speech, gaze, and turn-taking behaviors to provide recommendations and interact with participants during the discussion phase of the task. The spatial arrangement will ensure that all human and nonhuman team members are mutually visible and able to interact in real time, creating a co-located decision-making environment while preserving experimental control.

D. Measures

Consistent with the conceptual model depicted in Fig. 4, the present study operationalizes five primary dependent variables capturing participants' perceptions, interactions, and performance outcomes in response to expert power (human versus machine). All self-report measures were assessed using a 5-point Likert scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree, with higher values indicating greater endorsement of the construct.

1) Level of Positive Team Member Perceptions. This is measured as participants' overall evaluative orientation toward the expert as a member of the team. This construct captures perceived openness, competence, and the general positivity of the team experience [19]. Items will assess whether team members were open in expressing thoughts and feelings, whether participation in the team was a positive experience, and whether team members were viewed as competent and capable.

2) Level of Positive Team Member Interactions. This assessed the quality and effectiveness of interpersonal dynamics between participants and the expert. This measure captures perceived equality of participation, coordination effectiveness, and willingness to engage in future collaboration [20]. Items evaluate whether all members participated equally, whether the team works effectively together, and whether participants will choose to work with the same team again.

3) Level of Team Cohesion. This is captured the extent to which participants experience psychological integration and unity with the expert. This construct reflects feelings of belonging and dyadic working relationships [20]. Items assess whether participants feel like part of the team and whether they work well with the expert teammate.

4) Level of Group Decision Quality. This is measured perceived and process-oriented aspects of collective decision-making. This construct reflects both deliberative quality and inclusivity in team decisions [21]. Items assess whether the group considers multiple ideas, whether all contributions are heard and considered, and whether discussions are dominated by a single individual (reverse-coded where appropriate).

5) Probability of Attributing Responsibility to Humans. This assesses participants' tendency to assign responsibility, trust, and decision authority to the expert, with a particular focus on differentiating attribution toward human versus robotic agents. Items captur perceptions of expert reliability and trustworthiness, perceived responsibility for task outcomes, alignment of decision-making processes with the participant's own, and willingness to delegate task completion to the expert [22].

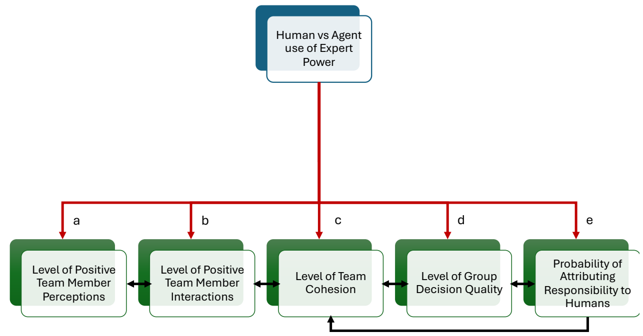

Fig. 4. Experiment model of human versus machine agent use of expert power with five dependent variables.

E. Procedures

Upon arrival, participants will be greeted by a member of the research team, who will verify group assignment and ensure that all participants are in the correct session. At this stage, participants are reminded verbally that participation is entirely voluntary. They are then escorted to the experimental area for onboarding. Next, participants complete an informed consent process using an IRB-approved consent form. Participants are given the opportunity to read the document, ask questions, and provide consent prior to any data collection. No experimental procedures begin until consent is obtained from all participants.

Following consent, participants receive an introduction to the experimental scenario, which consists of either the physical or a VR-adapted version of the Desert Survival Task. Instructions explain that participants will (a) review a list of survival items, (b) complete an individual ranking of those items, (c) discuss their rankings with a teammate, and (d) revise and submit a final group ranking. Participants are also informed of the presence of an "expert" agent who will participate in the task; however, the nature of this agent (human versus robot, including any WoZ implementation) is not disclosed to avoid demand characteristics. Participants are then seated in the physical setup or immersed in the VR environment and complete the individual ranking phase. During this phase, each participant independently evaluates and ranks survival items presented visually within the environment. Initial rankings are recorded for subsequent analysis of decision strategies and changes over time.

Following the individual phase, participants engage in a team discussion phase with the assigned expert agent. Over approximately 20 minutes, participants discuss their initial rankings, negotiate differences, and work toward a final group decision. Interactions between participants and the expert agent are automatically logged, including instances of suggestion acceptance, rejection, and justification. These behavioral indicators are used to assess influence dynamics and interaction patterns. The outcome of this phase is a finalized group ranking, which will serve as the basis for evaluating group decision quality.

TABLE I. FLOW AND TIMING OF THE EXPERIMENT

Experiment Activity	Allotted Time
1. Participants arrive and sign informed consent	3 minutes
2. Introduce desert survival scenario and assign roles	5 minutes
3. Participants complete individual rankings	8 minutes
4. Team discussion of survival items and ranking	20 minutes
5. Provide expert rank and rationale	10 minutes
6. Debrief experiment, dismiss participants	3 minutes
Total Time	**49 minutes**

After the team discussion, the expert agent (human confederate or robot) will reveal its own ranking of the survival items along with a rationale for its recommendations. The agent then prompts the team to confirm or revise their final submission, allowing for a final opportunity to incorporate expert input. Upon completion of the task, all participants will proceed to a debriefing session. During debriefing, a researcher will provide a full explanation of the study's purpose, clarify the nature of the expert agent (human or robot), and address any participant questions. Participants then complete any remaining survey measures before being formally dismissed.

The full session is expected to last approximately 49 minutes and will follow a standardized sequence to ensure consistency across experimental conditions (see Table 1).

F. Data Analysis

The purpose of the collected data is to determine whether the expert type, human expert or robot expert, significantly alters our measures of Positive Team Member Perceptions, Positive Team Member Interactions, Team Cohesion, Group Decision Quality, and Responsibility Attribution during the Desert Survival Task. Due to the experiment being a between-subjects experiment, the groups' data will be compared against each other, considering every participant will only experience one condition.

All data will be de-identified to prevent biases or faults and to get the truest results. As with all analyses, the data will first be cleaned and then organized to compare conditions better. For each dependent variable, each group's average scores will be compared against each other using a 2 x 2 factorial ANOVA. The differences between the two groups will be noted and also weighed against each other as overall totals.

III. ANTICIPATED RESULTS

Data collection is on-going and incomplete. Based on prior theory and preliminary work, we anticipate that robot expert advisors may be more persuasive to influence team decisions than human experts. Additionally, we expect that physical embodiment of the robot will increase this influence.

IV. DISCUSSION

The anticipated findings, if indeed confirmed by the collected data, would suggest that machine-based expert power can exert influence on team processes. If observed, this pattern would reinforce growing evidence that humans readily defer to algorithmic or AI-driven recommendations, even under conditions of limited transparency regarding underlying capabilities.

One potential key implication would be that expert power in human–autonomy teams may operate less as a fixed attribute of the agent and more as a socially constructed and context-sensitive phenomenon. If robot experts elicit equal or greater compliance, this would suggest that cues such as perceived objectivity, consistency, or technological authority can substitute for traditional markers of human expertise [23]. Moreover, if we observe a difference in responsibility attribution, such as a tendency to shift responsibility away from humans when interacting with machine experts, it would highlight a potential decoupling between influence and accountability [24].

If we observe that physically embodied robots amplify expert influence relative to VR-based agents, this would suggest that perceptual and social presence cues strengthen the perceived legitimacy and authority of machine teammates. Such findings would align with theories of social presence and human–robot interaction, indicating that the medium through which expertise is delivered shapes how that expertise is interpreted and acted upon [25].

Several limitations should be acknowledged. The use of a controlled task environment and a relatively small team size may constrain generalizability to more complex, real-world operational contexts. Additionally, the WoZ implementation, while necessary for experimental control, may not fully capture the variability of autonomous system behavior [26]. Future work should extend this paradigm to larger teams, more ecologically valid scenarios, and varying levels of system transparency and autonomy.

Overall, this study may advance understanding of how expert power functions when instantiated in machines.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number 25RT0821. The views expressed in this paper are those of the authors and do not reflect those of the U.S. Air Force, Department of Defense, or U.S. Government.

REFERENCES

[1] J. R. P. French, Jr., B. Raven, and D. Cartwright, "The bases of social power," in *Studies in Social Power*, D. Cartwright, Ed. Ann Arbor, MI, USA: Univ. Michigan Press, 1959, pp. 150–167.

[2] B. H. Raven, "The bases of power: Origins and recent developments," *J. Soc. Issues*, vol. 49, no. 4, pp. 227–251, 1993.

[3] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human-autonomy teaming: A review and analysis of the empirical literature," *Hum. Factors*, vol. 64, no. 5, pp. 904–938, 2022.

[4] Y. T.-Y. Hou and M. F. Jung, "Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, Art. no. 477, 2021, doi: 10.1145/3479864.

[5] Y. T.-Y. Hou, E. Cheon, and M. F. Jung, “Power in human-robot interaction,” in *Proc. 2024 ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, 2024, doi: 10.1145/3610977.3634949.

[6] W. G. Astley and P. S. Sachdeva, “Structural sources of intraorganizational power: A theoretical synthesis,” *Acad. Manage. Rev.*, vol. 9, no. 1, pp. 104–113, 1984.

[7] L. L. Greer, “Power in teams: Effects of team power structures on team conflict and team outcomes,” in *Handbook of Conflict Management Research*, O. B. Ayoko, N. M. Ashkanasy, and K. A. Jehn, Eds. Cheltenham, U.K.: Edward Elgar, 2014, pp. 93–108.

[8] M. Mortensen and M. R. Haas, “Rethinking teams: From bounded membership to dynamic participation,” *Organ. Sci.*, vol. 29, no. 3, pp. 341–355, 2018.

[9] K. S. Haring, K. M. Satterfield, C. C. Tossell, E. J. de Visser, J. R. Lyons, V. F. Mancuso, V. S. Finomore, and G. J. Funke, “Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance,” *Front. Psychol.*, vol. 12, Art. no. 625713, 2021, doi: 10.3389/fpsyg.2021.625713.

[10] Y. T.-Y. Hou, W.-Y. Lee, and M. F. Jung, “ ‘Should I follow the human, or follow the robot?’ Robots in power can have more influence than humans on decision-making,” in *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–13, doi: 10.1145/3544548.3581066.

[11] V. Groom and C. Nass, “Can robots be teammates? Benchmarks in human-robot teams,” *Interact. Stud.*, vol. 8, no. 3, pp. 483–500, 2007, doi: 10.1075/is.8.3.10gro.

[12] Q. Zhang, M. L. Lee, and S. Carter, “You complete me: Human-AI teams and complementary expertise,” in *CHI ’22: Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, doi: 10.1145/3491102.3517791.

[13] L. Alshammary, “What text can tell us about collaborative work: Team shared knowledge and cohesion in an online decision-making task,” Ph.D. dissertation, Univ. Bath, Bath, U.K., 2021.

[14] M. Jakesch, J. T. Hancock, and M. Naaman, “Human heuristics for AI-generated language are flawed,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 120, no. 11, Art. no. e2208839120, 2023.

[15] J. C. Lafferty, P. M. Eady, and J. Elmers, *The Desert Survival Problem*. Plymouth, MI, USA: Experiential Learning Methods, 1974.

[16] L. D. Riek, “Wizard of Oz studies in HRI: A systematic review and new reporting guidelines,” *J. Hum.-Robot Interact.*, vol. 1, no. 1, pp. 119–136, 2012.

[17] A. Momen, C. C. Tossell, R. E. Niemeyer, J. Walliser, M. Tolston, G. Funke, and E. J. de Visser, “Perceived trustworthiness and moral competence of a GenAI-enabled ethical robot advisor,” *Interact. Stud.*, vol. 26, no. 2, pp. 326–356, 2025.

[18] C. C. Tossell, C. Kuennen, A. Momen, G. Funke, M. Tolston, and E. J. de Visser, “Robots in the moral loop: A field study of AI advisors in ethical military decision-making,” *Front. Artif. Intell.*, vol. 8, Art. no. 1694772, 2025, doi: 10.3389/frai.2025.1694772.

[19] C. Nass, B. J. Fogg, and Y. Moon, “Can computers be teammates?” *Int. J. Hum.-Comput. Stud.*, vol. 45, no. 6, pp. 669–678, 1996.

[20] W. W. Huang, K. K. Wei, R. T. Watson, and B. C. Y. Tan, “Supporting virtual team-building with a GSS: An empirical investigation,” *Decis. Support Syst.*, vol. 34, no. 4, pp. 359–367, 2003.

[21] G. R. Bushe and G. Coetzer, “Appreciative inquiry as a team-development intervention: A controlled experiment,” *J. Appl. Behav. Sci.*, vol. 31, no. 1, pp. 13–30, 1995.

[22] J. Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an empirically determined scale of trust in automated systems,” *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, 2000.

[23] K. S. Haring, E. Phillips, E. H. Lazzara, D. Ullman, A. L. Baker, and J. R. Keebler, “Applying the swift trust model to human-robot teaming,” in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, Eds. Cambridge, MA, USA: Academic Press, 2021, pp. 407–427, doi: 10.1016/B978-0-12-819472-0.00017-4.

[24] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, “Sacrifice one for the good of many? People apply different moral norms to human and robot agents,” in *Proc. 10th Annu. ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, 2015, pp. 117–124, doi: 10.1145/2696454.2696458.

[25] E. Deng, B. Mutlu, and M. J. Matarić, “Embodiment in socially interactive robots,” *Found. Trends Robot.*, vol. 7, no. 4, pp. 251–356, 2019.

[26] A. Steinfeld, O. C. Jenkins, and B. Scassellati, “The Oz of Wizard: Simulating the human for interaction research,” in *Proc. 4th ACM/IEEE Int. Conf. Human-Robot Interact.*, 2009, pp. 101–108.