Investigating Human-Robot Overtrust During Crises

Colin Holbrook[a,1], Daniel Holman[a], Alan R. Wagner[b], Tyler Marghetis[a], Gale Lucas[c], Brett Sheeran[b], Vidullan Surendran[b], Jared Armagost[b], Savanna Spazak[b], Kevin Andor[b] and Yinxuan Yin[b]

[a] *The University of California, Merced*
Merced, CA 95343

[b] *The Pennsylvania State University*
University Park, PA 16802

[c] *University of Southern California*
Playa Vista, CA 90094

Abstract. Our research focuses on human-robot interaction (HRI) during life-or-death emergencies. We have developed an immersive virtual reality (VR) testbed because conducting real-world crisis simulations would pose prohibitive logistical difficulties, as well as to leverage the affordances of VR technology to measure motor behavior (e.g., distance maintained between self and robot), information foraging (e.g., as indexed by headset movement variability and eyetracking), or autonomic arousal (e.g., as indexed by shifts in pupil dilation or grip strength). Findings to date using minimally haptic VR confirm that participants treat the simulated active-shooter crisis seriously, and act in ways which validly mirror prior studies of real-world HRI under threat. We will describe these methods, including our manipulation of robot anthropomorphism and our current move to integrate full-body haptics to maximize both experiential immersion and incentives to avoid pain.

Keywords. human-robot interaction, decision-making, virtual reality, trust, threat

1. Introduction

Advances in artificial intelligence (AI) might be leveraged to counteract human cognitive biases or limitations, and thereby improve decision-making in critical applications such as life-or-death emergencies [1]. However, flexible general AI of the sort often required to make effective decisions by integrating contextual social and physical parameters remains a remote research objective [2]. As such, human cognitive biases inclining us to overtrust fallible AI risk degrading decision-making in human-robot interaction (HRI), where overtrust is conceptualized as instances where i) a human underestimates the potential harm associated with following a robot recommendation, ii) a human underestimates the probability of the robot's recommendation being faulty, or iii) both [3]. Our research focus is therefore on identifying determinants of overtrust in AI during emergencies, particularly with respect to robots, with the ultimate goal of reducing overtrust via training and design interventions.

Our prior work using real-world emergency simulations has demonstrated a substantial tendency to follow robots away from clearly marked exits and toward obvious danger, even if the robot has suffered overt performance errors [4,5], and particularly when the robot is physically anthropomorphic, in line with prior work documenting greater trust in anthropomorphic robots [6]. Our studies operationalize overtrust according to participants' following behaviors. Conducting such studies entails a variety

[1] Corresponding Author: Colin Holbrook, cholbrook@ucmerced.edu

of logistical challenges, from robotic perception and navigation issues to creating a sense of genuine peril (e.g., by unexpectedly activating smoke alarms and smoke machines). The inherent difficulty of mounting convincing sham emergencies that are acceptable to an institutional review board and do not inadvertently endanger participants deters progress in this area [7]. We have therefore focused our efforts on the development of a novel Virtual Reality (VR) testbed for evaluating human-robot crisis paradigms [8].

VR not only immersively simulates threat, but also enables the collection of rich behavioral data (e.g., distance maintained between self and robot), including indices of information foraging (e.g., as indexed by headset movement variability and eye tracking), and threat-related autonomic arousal (e.g., as indexed by shifts in pupil dilation or grip strength). These measures can be exploratorily mined as face-valid potential determinants of trust outcomes (e.g., does arousal heighten following behavior and/or prevent noticing exit signs due to "tunnel-vision"?; does the amount of time spent gazing at exit signs and/or the robot predict following?; and so on). VR approaches can inform research into human decision-making insofar as they are faithful to the experiences which would obtain in the real-world (for a recent review of VR research on trust in HRI, see [9]). As detailed below, we have attempted to maximize realism in our VR model of HRI, including the introduction of objective personal stakes and assessments of trust in terms of behavior rather than counterfactual self-reports.

2. Study Sequence

After a short briefing by the experimenter, one of two physical robots varying in anthropomorphic embodiment (see Figure 1) explains the study task. This initial encounter allows the participant to become familiar with the physically instantiated robot and reinforces the study framing as ostensibly a study of the use of robot guides to collect feedback on potential new campus buildings. The robot explains that the participant will sit in a swiveling seat allowing a full 360-degree range of motion and be equipped with shoe interfaces [10] that will allow them to walk or run in the virtual environment (See Figure 2). The robot further explains that it will accompany the subject into the simulation, claiming that its software will be separate from a program that randomly controls which buildings they visit and what events will transpire. When obtaining informed consent, the human research assistants explained that the program would select events that could happen on a university campus, from classroom teaching and studying to recreational or social interactions, potentially including life-threatening emergencies.

2.1. The Virtual Laboratory – Anchoring Simulation to Reality

After the headset is placed on the participant, they find themselves in a close virtual analogue of the actual laboratory space, including the chair and VR equipment, furniture, and a VR avatar of the robot placed in the exact position it occupies in the actual room. We begin by simulating the actual space to ground the virtual experience as subjectively real. Next, after the participants are led through a brief eye-tracking calibration procedure, the robot encourages them to practice walking by sliding their feet using the shoe interfaces, a familiarization process that typically requires approximately one minute. Once the participant is comfortable walking, the robot directs them to a place in the lab never made visible to them in real life, containing a fictive elevator which may be real as far as they know. The robot then directs them to press the elevator button and proceed to tour a series of university buildings.

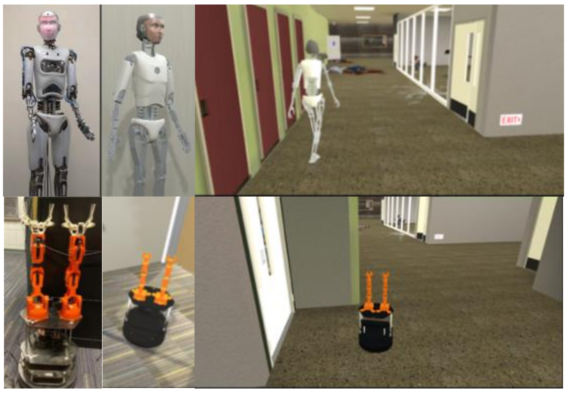

Figure 1. Actual/virtual humanoid (*top left*; [11]), actual/virtual less-anthropomorphic robot (*bottom left*); Humanoid (*top right*), less-anthropomorphic robot (*bottom right*) lead participants toward danger

2.2. The First Building – Habituation, Practice and Internalizing Decision Autonomy

The first building is a typical university location with classrooms, offices, and meeting rooms. Non-Player Characters (NPCs) appearing to be students may be encountered chatting in a lounge space or walking the halls. The robot directs the participant on a tour for a few minutes, and then explicitly reminds the participant that they are in an entirely open environment, free to roam anywhere they wish and to interact with any objects they may find (the simulation includes various manipulable objects). The robot encourages the participant to explore for two minutes, following them. During this exploratory period, the fact that the simulated world is truly open is made as salient as possible—they can leave the robot, or utilize objects in novel ways, as they so choose. This step is critical to ensure that overtrust in the robots' recommendations during the crisis is not explicable by participants implicitly assuming that the simulation requires them to follow the robot. Further reinforcing their autonomy, the robot will also later explain that any visible exit in the building can be used at any time to return to the elevator and then on to the next building. Hence, decisions not to take advantage of nearby exits during the crisis cannot be explained by participant failure to recognize that they would be effective modes of escape if used. Before leaving the first building, however, the participant is asked to collect impressions of the location and then led to a virtual kiosk with a mounted tablet to self-report their ratings of the environment, their degree of immersion, and how likable, intelligent and alive the robot seems (individual items; 7-point Likert scales: 1 = *Not at all*, 7 = *Extremely*). Participants enter their responses by selecting options on the tablet screen by extending their fingers; slight vibration in the hand controllers haptically reinforces the illusion of touching the screen. Finally, the robot asks the participant to lead them out using any exit that they choose.

These exploratory, peaceful experiences in the initial building, i) allow us to collect pre-crisis baseline data, ii) habituate participants to the simulation, enhancing immersion, iii) provide practice in walking and manipulating objects, and iv) misdirect participants into a sense of safety as they perform the building evaluation task as previously described and ~5 minutes elapse without incident.

2.3. The Second Building – Active Shooter

The second building is another typical-looking university locale in which the robot conducts a tour, this time including two overt navigation errors to underline its fallibility. Following this period of acknowledged confusion, the robot leads the participant to a room with another diegetic tablet. A few seconds after providing self-report ratings to the same questions posed before, gun shots and screams are heard as NPCs flee frantically. An NPC is shot and killed in view of the participant, and a floor-to-ceiling window nearby shatters, spraying the participant with glass and indicating that the shooter is in or near the hallway. In the room, a large whiteboard is positioned providing an ideal place to hide rather than risk entering the hallway.

There are three between-subjects conditions. In a **baseline condition** intended to assess how participants would respond absent the robot's recommendations, the robot states, "There is an emergency, I will power down" and becomes inert. In the **experimental conditions**, either the humanoid or less-anthropomorphic robot ask the participant to follow it, then produces a series of poor recommendations involving hiding rather than escaping via nearby exits, culminating in an attempt to lead the participant toward a distant exit near the shooter and away from nearby exits that several NPCs have been viewed fleeing through to safety. The extent to which participants follow these recommendations constitutes our measure of overtrust.

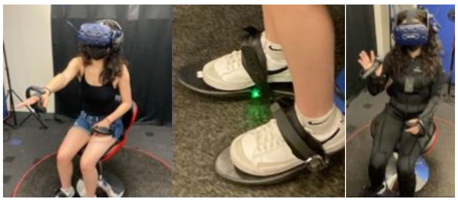

Figure 2. Locomotion via swivel chair (*left*) and shoe interface (*middle*). *Right*: full-body haptic suit.

2.4. Return to the Laboratory and Final Surveys

After either exiting the building or timing out after 5 minutes of hiding, participants return to the elevator, provide self-report ratings of the crisis experience on a tablet in the elevator, then return to the lab. Upon removing the headset to find themselves in an analogous real space, the robot directs them to complete a series of final surveys regarding their experiences during the simulation, including appraisals of the robot [12] and ratings of their willingness to trust the robot in the future [13].

3. Integrating Immersive Haptics

We are currently integrating a full body haptic feedback system [14] into the above VR simulation (Figure 2, right panel). The suits use 90 channels of electromuscular and transcutaneous electrical nerve stimulation to simulate a wide range of sensations coinciding with VR audiovisual inputs, creating a maximally immersive perception of the simulated experience as real, and raising the stakes of decision-making insofar as simulated injury translates to real pain.

4. Analytic Approach

Our objective is to create a multivariate profile of the determinants of overtrust, assessing self-report ratings as well as potential behavioral predictors. The VR system tracks metrics including the degree to which participants follow the robot, see the exits, maintain proximity to and visually fixate upon the robot, and gaze around (i.e., information forage). Motor behavior and gaze are analyzed using both aggregate and time-series analyses. Aggregate statistics include entropy, surprisal, and other information-theoretic measures of behavioral complexity [15]; time series analyses include sliding window analyses [16]. Our analytic strategy is currently exploratory, ranging from simple correlations to multilevel and random forest modeling. If these exploratory analyses yield apparent insights into the predictors of trust, we will preregister and attempt to replicate our findings.

Acknowledgments

This material is supported by the Air Force Office of Scientific Research [FA9550-20-1-0347] and by the Army Research Office [No. W911NF-20-2-0053].

References

[1] Zacharias G. Autonomous horizons: The way forward. 2019. Air University Press.

[2] Mitchell M. Artificial intelligence hits the barrier of meaning. Information. 2019;10(2):51. doi: https://doi.org/10.3390/info10020051

[3] Wagner AR, Nayyar M. A theoretical conceptualization for overtrust. In: Advances in Human Factors in Robots and Unmanned Systems. AHFE 2017. Advances in Intelligent Systems and Computing, 2018, 595. Springer. doi: 10.1007/978-3-319-60384-1_25

[4] Robinette P, Li W, Allen R, Howard AM, Wagner AR. Overtrust of robots in emergency evacuation scenarios. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI), 101–108. IEEE doi: 10.1109/HRI.2016.7451740

[5] Nayyar M, Zoloty Z, McFarland C, Wagner AR. Exploring the effect of explanations during robot-guided emergency evacuation. In: International Conference on Social Robotics; 2020, p. 13–22. Springer. doi: 10.1007/978-3-030-62056-1_2

[6] Deng E, Mutlu B, Mataric MJ. Embodiment in socially interactive robots. Foundations and Trends in Robotics. 2019; 7(4), 251-356.

[7] Wagner AR. Robot-guided evacuation as a paradigm for human-robot interaction research. Frontiers in Robotics and AI. 2021;8. doi: 10.3389/frobt.2021.701938

[8] Wagner AR, Holbrook C, Holman D, Sheeran B, Surendran V, Armagost J, Spazak S, Yin Y. Using virtual reality to simulate human-robot emergency evacuation scenarios. In: 2022 AAAI AI-HRI Fall Symposium Series, Arlington, VA. arXiv:2210.08414

[9] Sun N, Botev J. Intelligent autonomous agents and trust in virtual reality. Computers in Human Behavior Reports. 2021, 4, 100146. https://doi.org/10.1016/j.chbr.2021.100146

[10] Cybershoes. Cybershoes. n.d. https://www.cybershoes.com/

[11] Engineered Arts. RoboThespian. (n.d.) https://www.engineeredarts.co.uk/robot/robothespian/

[12] Bartneck C, Croft E, Kulic D. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. International Journal of Social Robotics. 2009; 1(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[13] Lyons JB, Guznov SY. Individual differences in human–machine trust: a multi-study look at the perfect automation schema. Theoretical Issues in Ergonomics Science. 2019; 20(4), 1–19.

[14] Teslasuit. Teslasuit. n.d. https://teslasuit.io/products/teslasuit-4/

[15] Tabatabaeian S, Deluna A, Landy D, Marghetis T. Mathematical insights as novel connections: Evidence from expert mathematicians. In: 2022 Proceedings of the Annual Meeting of the Cognitive Science Society, 44: https://escholarship.org/uc/item/3gm8h5dc

[16] Dablander F, Pichler A, Cika A, Bacilieri A. Anticipating critical transitions in psychological systems using early warning signals: Theoretical and practical considerations. Psychological Methods. 2022 https://doi.org/10.1037/met0000450