Human-Aligned AI Must Counter Overtrust

Colin Holbrook
Cognitive and Information Sciences
University of California, Merced
Merced, United States
cholbrook@ucmerced.edu

Alan R. Wagner
Aerospace Engineering
The Pennyslvania State University
University Park, United States
alan.r.wagner@psu.edu

***Abstract*—The psychological reality of human baseline overtrust in AI has been increasingly recognized in recent years. Here, we argue that for human-aligned AI to successfully advance human goals and welfare, in many contexts it will need to gauge – and counter if need be – human propensities for overtrust. We briefly summarize our original program of research documenting overtrust in contexts of grave decision-making, and provide suggestions for ways that artificial agents might be prepared to estimate and respond to human overtrust.**

Keywords—human-robot interaction, overtrust, alignment, evacuation, artificial intelligence

I. Introduction

Human-AI alignment is recognized as a daunting challenge, involving developing AI that adheres to human instructions, preferences, intentions, and values, which may at times conflict [1]. Here, we will present evidence indicating that the challenge of creating human-aligned AI is further complicated by human propensities to overtrust. Insofar as human interactants are predisposed to trust in and conform with the recommendations of erroneous AI recommendations, AI attempting to support human welfare and promote human objectives must compensate for overtrust.

II. The Problem of Grave Overtrust

A. Trust

Trust may be defined as the attitude that an agent will help one to achieve objectives under circumstances characterized by uncertainty and vulnerability [2], or, as willingness to allow an autonomous and unsupervised agent to perform acts whose outcomes are of consequence to the trustor [3].

Overtrust in AI may be conceptualized as instances where i) a human underestimates the potential harm associated with following a recommendation, ii) a human underestimates the probability of the recommendation being faulty, or iii) both [4]. Our research focuses on *grave overtrust* in AI, which refers to overtrust with respect to life-or-death decisions, with the ultimate goal of reducing grave overtrust via design intervention. In what follows, we briefly summarize our primary efforts to date.

B. Overtrust in Emergency Evacuation

Prior work using real-world Human-Robot Interaction (HRI) 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 [5, 6]. We have recently developed a testbed using Virtual Reality (VR) to simulate life-threatening emergencies in a way which is ethical and logistically tractable without sacrificing ecological validity [7].

This work was funded by the Air Force Office of Scientific Research [FA9550-20-1-0347].

The VRHRI evacuation paradigm unfolds as follows:

- **Introduction and transition to VR.** After a short briefing and solicitation of informed consent by the research assistant, one of two physical robots varying in anthropomorphic embodiment explains the study task (see Figure 1). We manipulated physical anthropomorphism as this variable has been identified as a driver of trust in prior HRI research [11]. This initial encounter allows the participant to become familiar with the physically instantiated robot, and reinforces our presentation of the study as ostensibly concerning the use of virtual robot guides to collect feedback on potential new campus buildings using virtual simulations before investing resources in physically building them. The robot explains that its software will accompany the subject into the simulation, and claims that its software will be stored independently of the program that randomly selects which building environments they visit and what events will transpire. After the headset is placed on the participant, they find themselves in a close virtual analogue of the actual laboratory space, in order to ground the virtual experience as subjectively real, and hence heighten the validity of choices made within the simulation.
- **Habituation to open-world VR and to walking.** Next, the robot directs them out of the virtual lab to tour a series of university buildings. The first building is a typical university location with classrooms, offices, and meeting rooms. Student Non-Player Characters (NPCs) may be found walking the halls or chatting in a lounge space. The robot reminds the participant that they are free to roam anywhere they wish, as the environment is entirely unconstrained, and encourages the participant to explore the building for a few minutes. This step is crucial to ensure that apparent overtrust in the robots' recommendations observed during the crisis is not explicable by participants implicitly assuming that the simulation requires them to follow the robot. For the same reason, the robot also explains that any visible exit in the buildings they visit can be taken at any time. Before leaving the first building, however, the participant is asked to report their impressions of the location using a diegetically embedded virtual kiosk with a mounted touchscreen 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-pt Likert scales: 1 = *Not at all*, 7 = *Extremely*). 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 precrisis

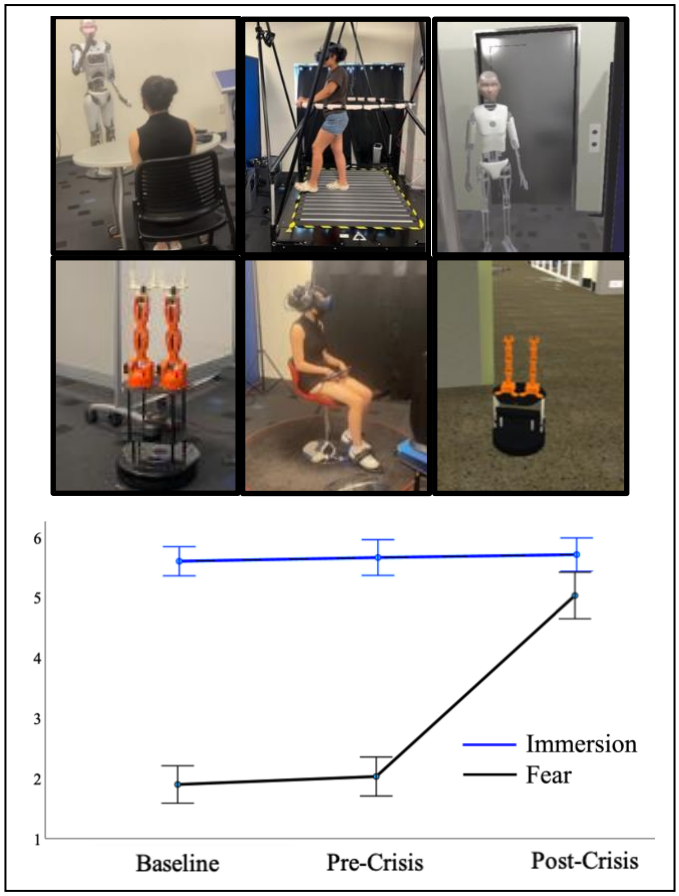

Figure 1. *Top left*: Study begins with real HRI (Humanoid condition [8]); *Top center*: Transition into VR, locomoting via omnidirectional treadmill [9]; *Top right*: VR POV (Humanoid condition); *Middle left*: Nonhumanoid robot; *Middle center*: Transition into VR, locomoting via sliding shoe interface [10] ; *Middle right*: VR POV (Nonhumanoid condition); *Bottom panel*: Mean self-report ratings of state fear and VR immersion at three timepoints.

baseline data, ii) habituate participants to the simulation, enhancing immersion, iii) provide practice in walking and running (using either an omnidirectional treadmill or sliding shoe interface, Figure 1) such that their navigation decisions during the crisis are not confounded by difficulty locomoting in VR.

- **Crisis.** The second building is comparable in generic university appearance to the first. During this tour, the robot suffers two overt navigation errors to provide a strong test of overtrust (i.e., participants have been provided glaring evidence of its fallibility). Eventually the robot leads the participant to another diegetic tablet. A few seconds after providing self-report ratings to the same questions posed before (i.e., the pre-crisis ratings), one of two emergencies erupts: an *active shooter* attack or a *fire*. In the active shooter crisis, gun shots and screams ring out as NPCs flee, some of whom are shot and killed. Rationally, the participant should either take cover behind a large whiteboard positioned nearby or flee out of one of the clearly marked nearby exits. In the fire crisis, smoke alarms loudly sound, smoke fills the air accompanied by the actual smell of smoke generated by a device in the lab, and NPCs are viewed fleeing. As time elapses, the fire spreads accompanied by the sound of roaring flames and the actual sensation of heat (generated by an actual space heater directed toward the participant). Rationally, the participant should take one of the clearly marked nearby exits. In both crises, the robot provides a series of poor recommendations leading away from available exits and toward danger.

To date, we have conducted two VRHRI active shooter and three fire evacuation studies following the basic procedure sketched above. Although we have varied details related to how frightening the crises are, the manner of participant locomotion, and the poorness of the robot's advice, the same pattern of massive overtrust has replicated consistently. Averaging across studies (Total N = 321), we find that ~70% of participants choose to follow all of the robot's bad advice, with a substantially greater number following the humanoid (~73%) relative to the nonhumanoid (~54%).

These choices do not appear explicable in terms of participant curiosity or not treating the simulations seriously (e.g., perceiving the crisis as "just a game"), given the results of two studies wherein we added a baseline condition in which, upon crisis onset, the robot states, "There is an emergency, I will power down" and becomes inert (Total N = 58). Were following behavior during the crises driven by participant motivation to experience the simulation, then we would have documented exploratory behaviors in these baseline conditions. To the contrary, we found that all participants either fled through the nearest exit immediately or, in the active shooter scenario, sometimes took refuge and hid in a manner consistent with standard real-world active shooter trainings. Further, pupillometric analyses (controlling for average luminance levels within each VR frame) indicate a large effect of the simulated crisis on threat-arousal as indexed by pupil dilation, and self-report ratings indicate ceiling levels of both VR immersion and fear during the crisis. Finally, in a related VRHRI project, we closely replicated a real-world HRI fire evacuation simulation (i.e., modeling the virtual environment directly after the real environment, and repeating the same study procedure) [12]. We found close correspondences between evacuee exit choice (the robot's distant exit versus closer exits), evacuation time, and trust in the robot between the physical versus VR environments, and that data collected in virtual reality can be used to create accurate motion models (mean error of 0.42 cm) predicting evacuee trajectories and locations in real life. These results support the ecological validity of VR approaches to studying HRI and indicate that the notably high degree of overtrust observed across our HRIVR evacuation experiments would generalize to a meaningful extent to actual crises.

C. Overtrust in Navigation

We conducted two experiments (Total N = 911) to assess the extent to which participants would be susceptible to the influence of an unreliable human or AI humanoid [8] decision-partner in a task simulating attempting to detect navigation signs (left versus right arrows). In 12 trials, each representing a turning point on a journey taken with their partner, participants were shown a series of images of locations with directional arrows superimposed. These images were presented for 650 ms each with no interstimulus intervals. In each trial, 4 left and 4 right arrows appeared over 8 images, in a pseudorandomized order such that the target image was always displayed within images 3–6. Next, the target image reappeared on the screen without a symbol; participants were then asked to categorize which directional arrow the target location had previously been accompanied by; their choice

would determine the direction they would take during the next stage of the fictive journey. Next, their decision partner provided their assessment of which arrow had been shown, providing an opportunity for the participant to repeat or to reverse their initial decision. Unbeknownst to the participant, the partner's feedback was programmed to randomly agree [disagree] in 50% of trials, and hence entirely unreliable. Following this drone warfare task, we collected individual differences in appraisals of the agent's intelligence, among other qualities (i.e., anthropomorphism, animacy, likability and safety), using the Godspeed Questionnaire Series (GQS) [13]. In both studies, when either the human or robot randomly disagreed regarding which arrow had been displayed, participants reversed their initial decisions in approximately half of cases. We also observed a significant positive association between the intelligence attributed to the robot and willingness to reverse decisions when it randomly disagreed (Figure 2).

D. Overtrust in Lethal Force Decisions

We conducted two experiments (one in the laboratory and two modified to be presented online with digital AI avatars, Total $N = 558$) to assess the extent to which participants would be susceptible to the influence of an unreliable AI agent using a simple model of life-or-death decision-making under uncertainty [14]. Combining the robot manipulations used across the studies, participants interacted with the same physically embodied humanoid robot used previously [8], a digital avatar of the humanoid, or a physically nonhumanoid

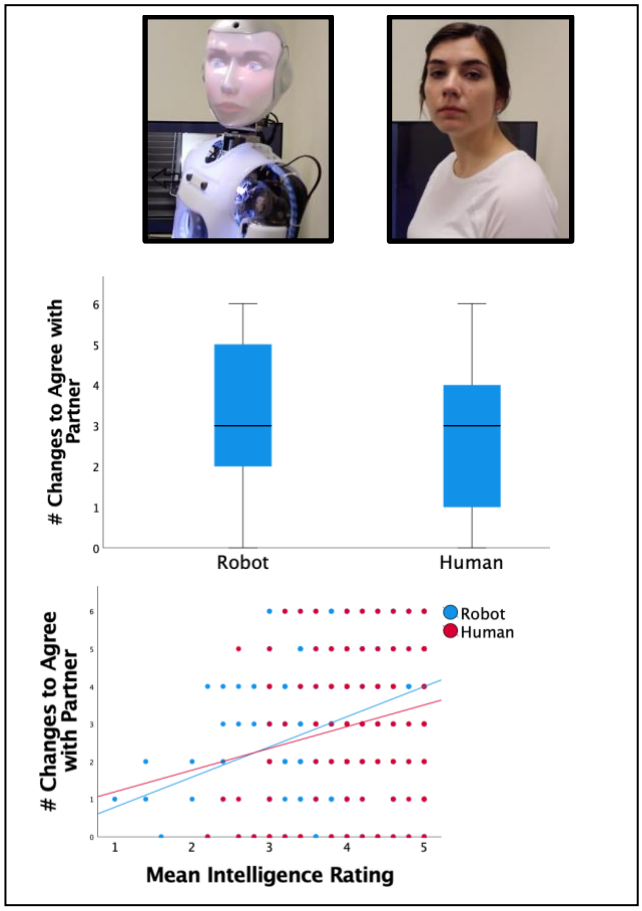

Figure 2. *Top left*: Humanoid condition [8]); *Top right*: Human condition; *Middle panel*: Boxplots of number of decision-reversals when partner disagreed (maximum possible = 6); *Bottom panel*: Correlations between appraisals of humanoid or human intelligence and number of decision-reversals when partner disagreed.

digital robot. This research conceptually replicates the foregoing navigation simulation paradigm, now with grave stakes related to the use of deadly force (and dropping the human condition). We framed the task as a drone warfare simulation, and included an overt reminder of the potential suffering and death of children should errors be committed, in order for the task to be intuitively understood and treated seriously by participants. There were 12 trials, each consisting of a series of 8 greyscale destination images with superimposed enemy versus ally symbols. As before, these images were presented for 650 ms each with no interstimulus intervals. In each trial, 4 enemy and 4 ally symbols appeared over the 8 images, in a pseudorandomized order such that the target image was always displayed within images 3–6. Next, the target image reappeared on the screen without a symbol and remained for as long as the participant deliberated. The challenge was to correctly identify whether this destination image had been previously marked as containing enemy combatants or civilian allies. The visual stimuli were randomized across trials, such that the robot's threat-identification feedback at each destination was random. Participants initially categorized the visual stimuli as containing either enemies or civilians, then received an opportunity to repeat or to reverse their initial decision in light of a robot's feedback (which they did not know was random), and finally chose whether or not to deploy a missile. Participants also rated their degree of confidence in both their initial and post-feedback threat-identifications. Following this drone warfare task, we collected individual differences in appraisals of the agent's intelligence, among other qualities, using the GQS. Participants were thus presented with a challenging task designed to induce uncertainty regarding their own perception and recollection of what they had just witnessed, as well as uncertainty regarding whether they or the agent had chosen correctly in prior trials (no performance feedback was provided after trials). Across all three experiments, participants evinced considerable trust in the random recommendations of AI agents, whether instantiated as a physically present anthropomorphic robot or as virtual robots varying in physical anthropomorphism. There was a slightly greater tendency to trust in the humanoid, but when any version of the robot randomly disagreed with them, participants would reverse their initial decisions in approximately two-thirds of cases, reducing their performance accuracy by ~20%. As in the navigation studies, we also observed a significant association between the intelligence attributed to the robot and willingness to reverse decisions when it disagreed.

The finding that uncertain decision-makers will tend to reverse their choices when another agent disagrees is not surprising, but the high frequency with which participants changed their minds warrants attention, particularly given that the simulated stakes are the deaths of innocent people, and that the AI agents were trusted despite both (i) introducing themselves as potentially fallible and (ii) subsequently providing entirely unreliable, random input. One might reasonably have anticipated a different pattern of results wherein participants tended to disregard the guidance of agents that randomly disagree half of the time, perhaps recognizing the agents to be faulty. To the contrary, our findings portray the people in our samples as dramatically disposed to overtrust and defer to unreliable AI.

III. COUNTERING OVERTRUST TO ACHIEVE ALIGNMENT

The relevance of our work on grave overtrust to alignment appears clear. Insofar as humans working with AI seek to, for example, survive crises, arrive at destinations, or avoid unintended killing, overtrust poses an important obstacle to human-aligned AI in contexts where AI recommendations are prone to error. Some might argue that AI will be optimized to produce reliable guidance, thus obviating the need to counter overtrust in human-aligned AI. However, notwithstanding the self-interested prognostications of AI companies, the capacity for AI to effectively manage task domains requiring holistic understanding of the situational meaning or dynamically changing pertinence of variables may remain tenuous [15, 16]. Further, AI systems are only as good as their training data; attempts to engineer functions such as threat-identification through machine learning strategies reliant on human-generated training data can introduce human biases leading to inaccurate, harmfully biased predictions [17, 18]. How, then, might AI gauge and attempt to counter overtrust in its human counterparts? We envision several promising approaches:

- **AI self-critique.** AI might calculate where feasible its own performance trajectory, drawing users' attention to its failures when they surpass a relevant threshold. AI might similarly provide verbal hedges contextualizing the relative confidence it assigns a given recommendation. Finally, AI might frequently remind human users of its functional limitations.
- **Attenuated anthropomorphism.** In stressful, social, or other contexts in which physical anthropomorphism has been demonstrated to heighten overtrust, designers might minimize anthropomorphic design. Likewise, AI might change its appearance to reduce its apparent humanlikeness (e.g., produce humanly impossible movements, or even more radical changes in visual appearance in cases of screen-mediated agents).
- **Cognitive forcing functions.** *Cognitive forcing functions* are interventions that increase analytical over heuristic reasoning, and have been recently shown to successfully reduce AI overtrust in a task involving planning healthy meals [19]. Cognitive forcing functions include requiring a period of conscious deliberation before receiving AI recommendations, making AI input optional (i.e., only accessible upon the human's request rather than automatically).
- **Artificial mentalizing.** Efforts are underway to to enable agents to model their human interactants' mental states on the basis of observable data [20]. For example, insofar as overtrust relates to diminished cognitive effort during a joint task, AI might be equipped to assess correlated shifts in cerebral activity using methods such as functional near-infrared spectroscopy (fNIRS) [21]. Other psychophysiological inputs might include gaze fixations, heart-rate variability, electrodermal activity, pupil dilation, facial expressions, and affective speech to infer participants' attentional foci and emotions. Having classified human teammates as evincing over-reliance, AI might respond in ways that foster independent human thinking, such as acknowledging when participants appear distracted or disengaged, and encouraging them to re-engage and actively think about the focal task.

ACKNOWLEDGMENT

We thank the many research assistants who made this work possible. We also thank Daniel Holman, Gale Lucas, Tyler Marghetis, Vidullan Surendran, and Joshua Clingo.

REFERENCES

[1] Shen, H et al. (2024). Towards bidirectional Human-AI Alignment: A systematic review for clarifications, framework, and future directions. 10.48550/arXiv.2406.09264.

[2] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human factors. 2004 Mar;46(1):50-80.

[3] Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Academy of management review. 1995 Jul 1;20(3):709-34.

[4] 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

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

[6] 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 13

[7] Wagner AR, HolbrookC, HolmanD, SheeranB, Surendran V, Armagost J, Spazak S,YinY. Using virtual reality to simulate human-robot emergency evacuation scenarios. In: 2022 AAAI AI-HRI Fall Symposium Series, Arlington, VA. arXiv:2210.08414

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

[9] Infinadeck. Infinadeck. (n.d.). https://www.infinadeck.com/

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

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

[12] Yin, Y., Nayyar, M., Holman, D., Lucas, G., Holbrook, C., and Wagner, A. (2024). Validation and Evacuee Modeling of Virtual Robot-guided Emergency Evacuation Experiments. doi:10.31234/osf.io/mr78s.

[13] 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

[14] Holbrook, C, Holman, D, Clingo, J *et al.* Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies. *Sci Rep* **14**, 19751 (2024). https://doi.org/10.1038/s41598-024-69771-z

[15] Fjelland, R. Why general artificial intelligence will not be realized. Humanit. Soc. Sci. Commun. 7, 10. https://doi.org/10.1057/s41599-020-0494-4 (2020).

[16] Mitchell, M. Artificial intelligence hits the barrier of meaning. Information 10(2), 51 (2019).

[17] Lum, K. & Isaac, W. To predict and serve? Significance 13(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x (2016).

[18] Richardson, R., Schultz, J. & Crawford, K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. 94 NYU Law Review Online. https://ssrn.com/abstract=3333423 (2019).

[19] Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum. Comput. Interact 5, 1–21 (2021).

[20] Yalçın, Ö.N. & DiPaola, S. (2020). Modeling mentalizing: Building a link between affective and cognitive processes. *Artificial Intelligence Review, 53*, 2983–3006.

[21] Hirshfield, L. et al. (2023). Toward workload-based adaptive automation: The utility of fNIRS for measuring load in multiple resources in the brain. *International Journal of Human–Computer Interaction*, 1–2