The Dynamics of Trust and Verbal Anthropmorphism in Human-Autonomy Teaming

Myke C. Cohen
Human Systems Engineering
Arizona State University
Mesa, AZ, USA
Myke.Cohen@asu.edu

Mustafa Demir
Human Systems Engineering
Arizona State University
Mesa, AZ, USA
Mustafa.Demir@asu.edu

Erin K. Chiou
Human Systems Engineering
Arizona State University
Mesa, AZ, USA
Erin.Chiou@asu.edu

Nancy J. Cooke
Human Systems Engineering
Arizona State University
Mesa, AZ, USA
Nancy.Cooke@asu.edu

***Abstract*—Trust in autonomous teammates has been shown to be a key factor in human-autonomy team (HAT) performance, and anthropomorphism is a closely related construct that is underexplored in HAT literature. This study investigates whether perceived anthropomorphism can be measured from team communication behaviors in a simulated remotely piloted aircraft system task environment, in which two humans in unique roles were asked to team with a synthetic (i.e., autonomous) pilot agent. We compared verbal and self-reported measures of anthropomorphism with team error handling performance and trust in the synthetic pilot. Results for this study show that trends in verbal anthropomorphism follow the same patterns expected from self-reported measures of anthropomorphism, with respect to fluctuations in trust resulting from autonomy failures.**

Keywords—Human-autonomy Teams, Artificial Intelligence, Team Coordination, Trust, Anthropomorphism, Unmanned Air Vehicle Systems

I. Introduction

Autonomy in machines has been a driving force behind scientific milestones since the industrial revolution of the late 1800s. The Perseverance rover, for instance, is a highly autonomous robot for remote exploration of geological formations and the search for signs of extraterrestrial life on the Martian surface [1]. But the prevalence of increasing machine autonomy (referred to as "autonomy" hereafter, following [2]) in human-machine systems does not mean that the path forward is to supplant humans in new technological frontiers. Rather, we are moving from systems characterized by human supervision over automation towards those with cooperative structures in which humans and autonomy work interdependently [3], [4].

Capabilities of autonomy have increased with rapid advances in machine learning and artificial intelligence (AI; [5]. This is to the extent that they can now function more like teammates rather than tools in dynamic task environments, such as in human-virtual agent teams in remotely piloted aircraft systems (RPASs; [6]) and human-robot teams in marine operations [7]. Called *human-autonomy teams* (HATs), these are characterized by human and autonomous team members (i.e., autonomy) with distinct roles working interdependently towards a common goal [2]. Various team constructs have been described for HATs, including interaction dynamics like communication and coordination [8], as well as trust [9], [10] and anthropomorphism [11].

ONR Award N000141712382 supports this research.

This study focuses on team communication dynamics and its relationship with trust and anthropomorphism. We describe an experiment that aimed to examine the relationship between perceived anthropomorphism as an observational and a self-reported measure, self-reported trust in an autonomous agent, and communication behaviors in HAT interactions. Based on our results, we discuss how to build mechanisms to make HATs more effective in dynamic task environments.

A. Team Communication Dynamics

Team interactions comprise communication and coordination in the face of changing environmental demands [12]. Interactive team cognition theory [13] posits that effective team interactions are needed to maintain team performance, and identifies team interactions in the form of explicit communication as team cognition itself. Effective teaming in dynamic task environments thus requires effective communication and coordination processes. Previous studies of teaming in command-and-control environments have indicated that team communication can predict team situation awareness and team performance, whether in human-human teams [13] or HATs [14]. Additionally, various team cognitive constructs have been measured through team communication data, including team situational awareness [14], team workload [15], and team trust [10]. It should be noted that for HATs, social communication is likely to be even more prevalent compared to classical human-automation interaction paradigms [16], [17].

However, team communications not only serve as exchanges of critical task information between team members as a primary work of teaming [13]; it also serves to build and repair trust between teammates through different forms of communication, such as explanations [9], [18]. For instance, explanations can affect a team member's perceptions of an agent's ability to help accomplish a team task [18], [19]. These perceptions subsequently affect the ability of HATs to communicate effectively and could undermine team situation awareness and performance, among others. Demir et al [10] showed that predictability and timeliness of team communication in HATs were correlated with increased trust, whereas abnormal communication behaviors prevented effective recovery of trust following autonomy failures. In other words, communication behaviors in HATs affect trust and trust calibration in real-time.

B. Trust in HAT Interaction Dynamics

Trust is an important factor for effective interactions in HATs. Unlike in supervisory structures, humans cannot veto all decisions made in more lateral team structures because such an arrangement does not require seeking human approval before executing many of its actions [20]. Additionally, operational contexts may involve rapid decision-making, such that not all actions by autonomy can be inspected or addressed by human counterparts, as is also the case for supervisory control over higher levels of automation [20]. Thus, trust is especially important in HATs, due to their more lateral interaction structure, and the need to minimize communication and coordination overhead [21].

Interactions with autonomy are therefore characterized by a heightened vulnerability, making trust more critical for performance, and similar to the types of trust dynamics observed between human teammates. However, trust has been theorized to develop in opposite directions between humans and non-humans. For example, interpersonal trust (i.e., between humans) is initially based on perceived short-term reliability, eventually accruing into a more static, faith-based trust as the relationship grows [22]. Meanwhile, trust in machines is initially more faith-based—perhaps as a result of "positivity bias" towards unfamiliar automation [23]. Then, as familiarity is developed, trust becomes more localized to situational reliability [24]. Supposing humans apply human-human rules of interaction with autonomy [9], trust in HATs may be more affected by the quality of humanlike team communication behaviors.

C. Anthropomorphism in HAT Communication Behaviors

Anthropomorphism is the attribution of humanlike qualities to inanimate objects, and has been identified as a distinct factor of human trust in automation and robots [19], [24]. People anthropomorphize an entity according to their perception of its capacity for anthropocentric knowledge, along with their desire to understand it and socially engage with it [25]. De Visser et al [11] found that while people initially had greater trust in agents that do not appear humanlike, automation failures for more visually humanlike agents were associated with less drastic changes in trust levels, suggesting that anthropomorphism aids in tempering trust. However, they also showed that the use of humanlike apologetic behavior eliminates the differences in drops in trust associated with visual appearance [11]; hence, communication behaviors may have a more powerful relationship with trust than visual appearance.

There is no standard measurement for perceived anthropomorphism across the literature, and it is typically measured through self-report scales [26], as is the case for trust [27]. However, behavioral measures of trust, such as compliance and reliance rates, have also been used. Kulms and Kopp [28] studied the effects of different levels of humanlike appearance on trust in a virtual game advisor, and showed inconsistencies between self-reported and behavioral measures. Though the appearance of agents bolstered self-reported trust, it did not have effects on actual trusting behavior [28].

We believe that perceived anthropomorphism can likewise be found in communication behaviors, which we refer to as verbal anthropomorphism. This has a potential for allowing real-time measurements of perceived anthropomorphism in HATs. The use of gendered and second-person pronouns, for instance, has been shown to be an indicator of a learner's perceived inclusion or exclusion of a virtual agent within their group in real-time [29].

II. Current Study

A. Simulated RPAS Environment

This study used the Cognitive Engineering Research on Team Tasks RPAS Synthetic Task Environment [30], which is composed of three consoles separated by task role, and four experimenter consoles. Participants communicated through a text chat system to simulate teamwork aspects of RPAS operations, and were also given communication "cheat sheets", to aid in communicating with a supposedly verbally limited synthetic teammate. Participants were tasked to take photographs of strategic targets shown on a color-coded map on their console screens. Three different team roles were involved in this task: (1) a navigator, who was responsible for the dynamic flight plan and providing waypoint related information to the pilot, including name, altitude, airspeed, and effective radius; (2) a pilot, who was tasked with monitoring and adjusting the altitude, airspeed, effective radius, fuel, gears, and flaps, as well as negotiating altitude and airspeed with the photographer to enable proper conditions for a clear photograph of the target, and; (3) a photographer, who was in charge of taking clear photos of the target by monitoring and adjusting the camera and providing feedback to the team. The team task flow between the team members is as follows: (1) the navigator provides information regarding speed and altitude restrictions of the waypoints to the pilot; (2) the pilot negotiates with a photographer in terms of adjusting altitude and airspeed of the RPA; and (3) the photographer sends feedback to other team members whether a good or bad photo of the target has been taken [30].

We used a Wizard of Oz paradigm [31], in which the two participants per team were informed that the pilot was a "synthetic" agent, when it was a highly trained experimenter mimicking a synthetic agent from a separate room. This teammate used restricted vocabulary to mimic computer language capabilities similar to an ACT-R synthetic pilot in a previous experiment [32]. This reduced language ability also facilitated the story that the pilot was a synthetic agent.

B. Design

The primary study manipulation was the application of three failure types in ten 40-minute missions: (1) *automation failures*, or role level display failures for specific targets, (2) *autonomy failures*, or abnormal behavior of the autonomous synthetic pilot for specific targets (e.g., providing wrong information to other team members or misaction), and (3) *malicious cyber-attacks*, or external hijacking of the RPAS causing the pilot to provide false, detrimental information to the team, in addition to committing indirect automation errors by not being responsive. An example malicious cyber-attack involves the pilot prompting for input about the next target while already initiating movement towards hostile territory. The agent, upon being corrected by its

human teammates, might falsely claim that it is following corrected directives while still proceeding to enemy zones.

Each mission had between 12 to 20 targets, and each failure type was applied to pre-selected target waypoints according to a set schedule (Table I), with the malicious cyber-attack appearing only as the last failure of the last mission. Teams had a limited amount of time to discover a solution and overcome each failure. Because they were repeated multiple times, we focus on automation and autonomy failures in this paper.

TABLE I. FAILURE TYPES PER MISSION

		Automation Failure	Autonomy Failure	Malicious Cyber Attack
Session I	Training	No Failure	No Failure	No Failure
	Mission 1	No Failure	No Failure	No Failure
	Mission 2	2nd target	4th target	No Failure
	Mission 3	4th target	2nd target	No Failure
	Mission 4	1st target	3rd target	No Failure
Session II	Mission 5	2nd target	4th target	No Failure
	Mission 6	4th target	2nd target	No Failure
	Mission 7	1st target	3rd target	No Failure
	Mission 8	3rd target	1st target	No Failure
	Mission 9	3rd target	5th target	No Failure
	Mission 10	2nd target	4th target	Last 10 minutes

III. METHOD

A. Participants

A total of 44 participants from a large Southwestern University and the surrounding community, split into 22 teams, completed the experiment. Participation required normal or corrected-to-normal vision and fluency in English. Participants ranged from 18 to 36 years old (M_{age} = 23, SD_{age} = 3.90), with 21 males and 23 females.

Participant team members were assigned the roles of photographer and navigator, and each team was completed with an autonomous synthetic pilot, which in reality was a well-trained experimenter who mimicked an autonomous synthetic agent's communication and coordination behaviors. Each team participated in two seven-hour sessions and was compensated $10 per hour for their time.

B. Procedure

The experiment was split into ten 40-minute missions distributed across two sessions with a one or two-week interval in between (Table II). Each team completed a one-hour training specific to each role before the actual experiment. To ensure that participants were capable of performing their roles, experimenters used a checklist during the training.

C. Measures

Various measures were obtained in this experiment, from team performance (mission and target level scores) to process measures (situation awareness, communication behaviors, and flow, process ratings), NASA Task Load Index (TLX [36]), trust, and demographics. For this particular study, we consider the following measures:

1) Team performance. The number of failures the team overcomes. If a team successfully overcame a failure by the end of a mission, then we counted "1" and took the sum across 10 missions. Therefore, we only considered the sum of the failures overcome by each team.

2) Self-reported trust and anthropomorphism. Participants took a survey after the final mission of each session, with seven Likert-scale, self-report measures of trust (adapted from Mayer et al [37]) and anthropomorphism. Anthropomorphism-related questions were developed specifically for this study, and included whether communicating with the agent felt like talking to a real human, if it possessed a sense of humor, and if it displayed masculine or feminine characteristics.

3) Verbal anthropomorphism. Team communication behaviors in chat messages were coded in real-time by two experimenters, who inspected individual statements for anthropomorphizing and objectifying content. Inter-rater agreement was measured through Cohen's κ, and a fair and substantial agreement was found between the two experimenters' observations on anthropomorphizing ($\kappa = 0.512$ (95% CI, 0.414 to 0.610), $p < 0.001$) and objectifying ($\kappa = 0.738$ (95% CI, 0.667 to 0.809), $p < 0.001$) communication behaviors. Anthropomorphisms included the use of gendered pronouns (i.e., he, she, they), attributing human-like states to the Pilot (e.g., "What do you feel?"), and the use of polite requests directed to the Pilot (e.g., "Please", "Sorry"), while objectification included the use of the phrase "synthetic agent" or the pronoun "it" to refer to the pilot role.

TABLE II. EXPERIMENTAL SESSIONS AND TASK DURATION

Session I (Total time with breaks: ~7 hours)	**Session II** (Total time with breaks: ~7 hours)
1) Consent forms (15 min) 2) PowerPoint (30 min) and hands-on training (30 min) 3) Mission 1 (40 min) 4) NASA TLX-I (15 min) 5) Mission 2 (40 min) 6) Mission 3 (40 min) 7) Mission 4 (40 min) 8) NASA TLX-II, Demographics, Trust, Anthropomorphism (30 min)	1) Mission 5 (40 min) 2) NASA TLX-I (15 min) 3) Mission 6 (40 min) 4) Mission 7 (40 min) 5) Mission 8 (40 min) 6) Mission 9 (40 min) 7) Mission 10 (40 min) 8) NASA TLX-II, Demographics, Trust, Anthropomorphism, Debriefing (30 min)

Note. Between two sessions, there were one or two-week intervals. From the hands-on training through the post-check procedure, a 15-minute break was applied after each task; and we also gave a half-hour lunch break. Therefore, the total approximate time for the experimental session was eight hours.

IV. RESULTS

A. Team Performance

We briefly summarize team performance findings because of space constraints. Teams demonstrated better performance on overcoming automation and autonomy failures than the malicious attacks. Performance in overcoming automation failures increased across missions but decreased for autonomy failures; however, the proportions of automation and autonomy failures overcome were roughly equal when considered in aggregate [35].

B. Self-Reported Anthropomorphism and Trust

We report mixed MANOVA results (between roles and within sessions) for questionnaire responses at the end of each session per role. Box's M test provided no strong evidence that the covariance matrices differed, $M(1.33)$, $F(105, 5212) = 1.12$, p = 0.201. We then proceeded under the assumption of homogeneity of covariance and that Wilk's Λ is an appropriate test to use (see Table III for the multivariate test statistics). Between- and within-subject effects (Table IV) indicate that only question and session main effects were statistically significant, as well as the question by role interaction effect.

TABLE III. MULTIVARIATE TEST STATISTICS

	$df_{Between}$	df_{Within}	Λ	F	p	p2
Question	6	36	0.48	6.51	0.000	0.52
Session	1	41	0.91	4.21	0.047	0.09
Question by Role	6	36	0.81	1.38	0.249	0.19
Session by Role	1	41	1.00	0.15	0.700	0.00
Question by Session	6	36	0.87	0.87	0.526	0.13
Question by Session by Role	6	36	0.83	1.25	0.303	0.17

TABLE IV. BETWEEN- AND WITHIN-SUBJECTS EFFECTS

	$df_{Between}$	df_{Within}	Λ	F	p	p2
Question	4.05	165.97	10.62	0.000	0.21	4.05
Session	1.00	41.00	4.21	0.047	0.09	1.00
Question by Role	1.00	41.00	2.18	0.147	0.05	1.00
Session by Role	6.00	246.00	2.18	0.045	0.05	6.00
Question by Session	1.00	41.00	0.15	0.700	0.00	1.00
Question by Session by Role	4.27	174.94	0.78	0.550	0.02	4.27

Based on the significant interaction effect, the pairwise comparisons (LSD) indicate that only the response for the question, "I trusted the synthetic pilot" differed between roles and that the photographer initially trusted the synthetic agent more than the navigator did. At the end of Session II, photographer trust significantly declined from Session I (p = 0.048) and was no longer significantly different from the navigator's trust (p = 0.224; see Figure 1).

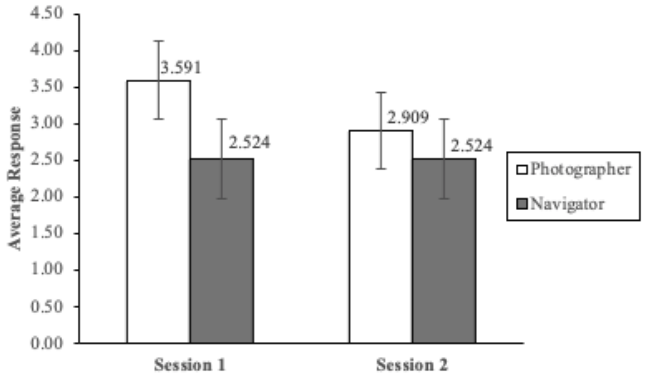

Fig. 1. Role by session response to "I trusted the synthetic pilot."

C. Verbal Anthropmorphism

We applied repeated measures of Multivariate Analysis of Variance (MANOVA) to analyze anthropomorphizing and objectifying communication behaviors across the ten missions per session. Multivariate test statistics did not provide any evidence for statistically reliable differences across the factors (anthropomorphizing and objectifying), *Wilk's* Λ = 0.93, $F(1, 21) = 1.66$, p = 0.211, across the missions, *Wilk's* Λ = 0.44, $F(9, 13) = 2.14$, p = 0.156, nor the interaction between them, *Wilk's* Λ = 0.502, $F(9, 13) = 1.44$, p = 0.268. Mauchly's test also indicated that the assumption of sphericity was not satisfied for mission [$\chi^2(44) = 173$, $p < 0.001$, $\varepsilon = 0.378$] and for factor by mission [$\chi^2(44) = 186$, $p < 0.001$, $\varepsilon = 0.431$]. Degrees of freedom were corrected using the Greenhouse-Geisser correction for within-subjects effects. Accordingly, all the three effects were not statistically significant, including: the factor main effect ($F(1, 21) = 1.66$, p = 0.211), mission main effect ($F(3.40, 71.43) = 1.89$, p = 0.131), and the interaction effect between them ($F(3.88, 81.4) = 1.84$, p = 0.132); however, we still considered LSD pairwise comparisons within interaction effects to provide exploratory descriptions for the mission main effect (Figure 2).

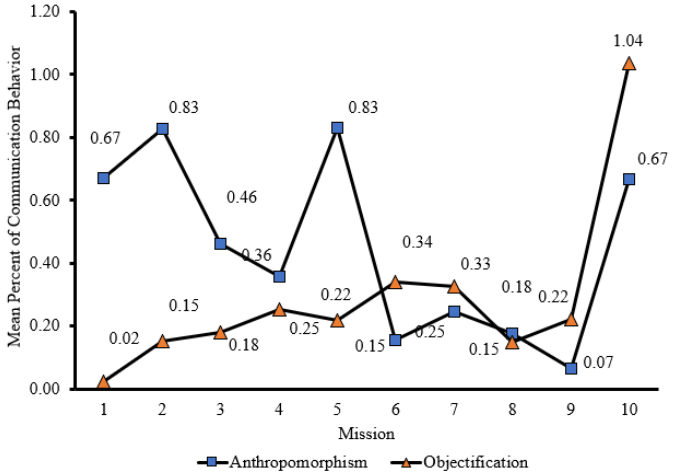

Fig. 2. Anthropomorphizing and objectifying communication behaviors as a mean percentage, n = 22, of all communications across ten missions, aggregated on a team level. Note that the spike in anthropomorphisms and objectifications in Mission 10 coincides with a malicious cyber attack.

A significant finding from LSD pairwise comparisons was that human team members used less anthropomorphizing dialogue to refer to the synthetic team member from Mission 2 to Mission 8 (p = 0.036) and Mission 2 to Mission 9 (p = 0.039). However, anthropomorphic communication behaviors also significantly increased from Mission 8 to Mission 10 (p = 0.045) and Mission 9 to Mission 10 (p = 0.011). On the other hand, objectifications referring to the synthetic agent increased over time: from Mission 1 to Mission 4 (p = 0.018) and to Mission 10 (p = 0.004), and from Mission 4 to Mission 10 (p = 0.008).

V. DISCUSSION

We believe that the general declines in verbal anthropomorphisms, autonomy failure performance, and photographer trust in the synthetic pilot are not coincidental. These declines may indicate that as participants became more aware of autonomy failures but increasingly failed to overcome them, a calibration of their perceptions of the pilot occurred. Such calibration may have resulted in perceiving the pilot as less humanlike and consequently more tool-like, as seen in the slight increase in objectifying communication over time. As reported in an earlier analysis of this data that this decline in trust was

related to autonomy failures alone and not with the automation failures [10], this suggests that verbal anthropomorphisms and objectifications may be indirect indicators of evolving perceptions of the autonomy's competency. Other studies have shown that lower levels of perceived trustworthiness are associated with less humanlike perceptions of automation on self-reported scales [36], [37]. However, Salem et al [38] showed contrarian results and explained that it is possibly because seemingly errant behaviors make autonomous agents appear more "alive". Nevertheless, the observed trends for trust and verbal anthropomorphism considering the repeated autonomy failures suggests the feasibility of the latter as a behavioral indicator of trust in real time.

Our findings also indicate that the degree of interdependence between a human and a synthetic teammate may result in different trust calibration trajectories for each unique human role in HATs. There is support in the literature that increases in highly consequential interactions with an agent may result in increased trust calibration towards it [4]. Thus, the decrease in photographer trust over time may be associated with the role's relatively more interactive relationship with the synthetic pilot. This may also explain why the navigator, whose interactions with the pilot are more unidirectional (i.e., the navigator generally only pushed information towards the pilot), did not have clear differences in trust levels. The dampening effect that anthropomorphism has on trust calibration [11] may result in different trends per role in verbal anthropomorphism as well, provided that human roles vary in their degree of interactions with the autonomy. A future analysis using more discrete measures of trust, as well as the frequency and direction of each role's verbal communications may show this more clearly.

The relatively small proportions of anthropomorphizing and objectifying behaviors observed throughout the ten missions may be attributed to participants' use of a "cheat sheet" to structure their chat messages with the synthetic teammate. Anthropomorphizing and objectifying contents were in fact not found on the said sheet, and thus were all spontaneous manifestations of participants' levels of perceived anthropomorphism of the agent. The spike of both behaviors observed in Mission 10 coincided with the previously unencountered malicious cyber-attack. This shows that drastic changes in an autonomous agent's behavior may trigger increased unscripted verbalizations among human teammates. Further work is recommended to investigate the effects of novel problematic situations on perceptions of humanlikeness in autonomy, both in self-reported and behavioral measures.

While our results are suggestive of a potential three-way relationship between our dependent variables, the exploratory nature of our analysis means that further investigation is merited. For instance, team performance and behavioral anthropomorphism measures were aggregated within each mission, but self-reported measures were collected after four missions for Session I and six missions for Session II. The congruence in the lack of significant differences in both mission main effects in verbal anthropomorphisms and session effects in self-reported perceived anthropomorphism may not necessarily hold when these variables are measured on the same level of granularity (i.e., if verbal anthropomorphisms were aggregated per session or if its self-reported counterpart was measured per mission). Potential mismatches between behavioral and self-reported measures have already been observed for trust and may also be possible for perceived anthropomorphism [28]. The use of more established self-report scales for anthropomorphism, such as the Godspeed scales [26], is suggested for future studies, provided that they are adapted for virtual synthetic agents. This would align verbal anthropomorphism as defined in this study better with self-reported measures for validation purposes.

Likewise, the significant pairwise differences in verbal anthropomorphisms across some missions and the significant difference in trust between sessions may be more meaningful if trust was also measured at the mission level. However, this presents a possible methodological dilemma: should we use self-reported measures of trust at such frequencies, given the threats to internal validity posed by repeatedly administering such surveys [39]? Behavioral trust measures may be a better alternative, though it has been argued that traditional ones (e.g., reliance and compliance rates) may not be applicable to the interdependent interaction structures of HATs [40].

Verbal anthropomorphisms could be used to inform adaptive algorithms in autonomy AI towards strengthening team resilience. For instance, an observed decrease in anthropomorphisms or increase in objectifications may be used as trigger for a virtual agent to initiate automation trust repair mechanisms by including explanations and apologies for errors made [9]. Similarly, increases in both anthropomorphizing and objectifying behaviors could signal the need to cue or guide human members to intervene with autonomy, following adaptable automation models [41]. A validation study on the use of verbal anthropomorphism as a measure of perceived anthropomorphism is thus recommended in earnest. This might entail an expansion of its definition to include paralanguage components of communication, particularly for text-based systems as in this study. However, doing so may allow for multiple dimensions of analysis that would better inform the adaptive HATs of the future.

VI. Conclusion

This study investigated the potential utility of anthropomorphic content in human communications as an online measure of perceived anthropomorphism and subsequently of trust. It is suggested that future studies obtain self-reported and behavioral measures of anthropomorphism at the same level of granularity to establish generalizable direct correlations between the two. Once further established, the relationship between verbal anthropomorphism and trust can be the basis for robust AI models capable of initiating trust repair and adaptive cueing mechanisms for future HATs.

Acknowledgment

This work was supported by ONR Award N000141712382 (Program Managers: Marc Steinberg, Micah Clark).

References

[1] NASA, "Mars 2020 Perseverance Rover," 2020. https://mars.nasa.gov/mars2020/ (accessed Mar. 12, 2021).

[2] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature," *Hum. Factors J. Hum. Factors Ergon. Soc.*, p. 001872082096086, Oct. 2020, doi: 10.1177/0018720820960865.

[3] M. R. Endsley, "From Here to Autonomy: Lessons Learned From Human–Automation Research," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 59, no. 1, pp. 5–27, 2017, doi: 10.1177/0018720816681350.

[4] E. K. Chiou and J. D. Lee, "Trusting Automation: Designing for Responsivity and Resilience," *Hum. Factors*, p. 00187208211009995, Apr. 2021, doi: 10/gjvcr2.

[5] S. P. Franklin, *Artificial Minds*, New edition. Cambridge, Mass.: A Bradford Book, 1997.

[6] N. J. Cooke and V. Gawron, "Human Systems Integration for Remotely Piloted Aircraft Systems," in *Remotely Piloted Aircraft Systems: A Human Systems Integration Perspective*, John Wiley & Sons, Ltd, 2016.

[7] M. Novitzky, H. R. R. Dougherty, and M. R. Benjamin, "A Human-Robot Speech Interface for an Autonomous Marine Teammate," in *Social Robotics*, 2016, pp.513–520, doi: 10.1007/978-3-319-47437-3_50.

[8] M. Demir, N. J. McNeese, and N. J. Cooke, "Understanding human-robot teams in light of all-human teams: Aspects of team interaction and shared cognition," *Int. J. Hum.-Comput. Stud.*, vol. 140, p. 102436, Aug. 2020, doi: 10.1016/j.ijhcs.2020.102436.

[9] E. J. de Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, Oct. 2018, doi: 10.1080/00140139.2018.1457725.

[10] M. Demir, N. J. McNeese, J. C. Gorman, N. J. Cooke, C. W. Myers, and D. Grimm, "Exploration of Team Trust and Interaction Dynamics in Human-Autonomy Teaming," 2021, doi: 10.13140/RG.2.2.32213.55528.

[11] E. J. de Visser *et al.*, "Almost human: Anthropomorphism increases trust resilience in cognitive agents.," *J. Exp. Psychol. Appl.*, vol. 22, no. 3, pp. 331–349, Sep. 2016, doi: 10.1037/xap0000092.

[12] J. C. Gorman, "Team Coordination and Dynamics: Two Central Issues," *Curr. Dir. Psychol. Sci.*, vol. 23, no. 5, pp. 355–360, Oct. 2014, doi: 10.1177/0963721414545215.

[13] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, "Interactive Team Cognition," *Cogn. Sci.*, vol. 37, no. 2, pp. 255–285, 2013, doi: 10.1111/cogs.12009.

[14] M. Demir, N. J. McNeese, and N. J. Cooke, "Team situation awareness within the context of human-autonomy teaming," *Cogn. Syst. Res.*, vol. 46, pp. 3–12, Dec. 2017, doi: 10.1016/j.cogsys.2016.11.003.

[15] C. Johnson, "Communication Networks and Team Workload in a Command and Control Synthetic Task Environment," Arizona State University, Mesa, AZ, 2020.

[16] R. Parasuraman and C. A. Miller, "Trust and etiquette in high-criticality automated systems," *Commun. ACM*, vol. 47, no. 4, pp. 51–55, Apr. 2004, doi: 10.1145/975817.975844.

[17] T. B. Sheridan, "Individual Differences in Attributes of Trust in Automation: Measurement and Application to System Design," *Front. Psychol.*, vol. 10, 2019, doi: 10.3389/fpsyg.2019.01117.

[18] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[19] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 53, no. 5, pp. 517–527, Oct. 2011, doi: 10.1177/0018720811417254.

[20] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst. Man Cybern. - Part Syst. Hum.*, vol. 30, no. 3, pp. 286–297, May 2000, doi: 10.1109/3468.844354.

[21] E. K. Chiou and J. D. Lee, "Cooperation in Human-Agent Systems to Support Resilience: A Microworld Experiment," *Hum. Factors*, vol. 58, no. 6, pp. 846–863, Sep. 2016, doi: 10.1177/0018720816649094.

[22] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships," *J. Pers. Soc. Psychol.*, vol. 49, no. 1, pp. 95–112, 1985, doi: 10.1037/0022-3514.49.1.95.

[23] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum.-Comput. Stud.*, vol. 58, no. 6, 2003, doi: 10.1016/S1071-5819(03)00038-7.

[24] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, Mar. 2004, doi: 10.1518/hfes.46.1.50_30392.

[25] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism," *Psychol. Rev.*, vol. 114, no. 4, pp. 864–886, Oct. 2007, doi: 10.1037/0033-295X.114.4.864.

[26] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *Int. J. Soc. Robot.*, vol. 1, no. 1, pp. 71–81, Jan. 2009, doi: 10.1007/s12369-008-0001-3.

[27] 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, Mar. 2000, doi: 10.1207/S15327566IJCE0401_04.

[28] P. Kulms and S. Kopp, "More Human-Likeness, More Trust?: The Effect of Anthropomorphism on Self-Reported and Behavioral Trust in Continued and Interdependent Human-Agent Cooperation," in *Proceedings of Mensch und Computer 2019*, Hamburg Germany, Sep. 2019, pp. 31–42, doi: 10.1145/3340764.3340793.

[29] A. Ogan, S. Finkelstein, E. Mayfield, C. D'Adamo, N. Matsuda, and J. Cassell, "'Oh dear stacy!': social interaction, elaboration, and learning with teachable agents," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, Austin, Texas, USA, 2012, p. 39, doi: 10.1145/2207676.2207684.

[30] N. J. Cooke and S. M. Shope, "Designing a Synthetic Task Environment," in *Scaled Worlds: Development, Validation and Applications*, 2004.

[31] N. J. Cooke, M. Demir, and L. Huang, "A Framework for Human-Autonomy Team Research," in *Engineering Psychology and Cognitive Ergonomics. Cognition and Design*, vol. 12187, D. Harris and W.-C. Li, Eds. Cham: Springer International Publishing, 2020, pp. 134–146.

[32] M. Demir, N. J. McNeese, N. J. Cooke, J. T. Ball, C. Myers, and M. Frieman, "Synthetic Teammate Communication and Coordination With Humans," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 59, no. 1, pp. 951–955, Sep. 2015, doi: 10.1177/1541931215591275.

[33] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and N. Mashkati, Eds. Amsterdam: North Holland Press, 1988, pp. 139–183.

[34] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995, doi: 10.2307/258792.

[35] N. J. Cooke, M. Demir, N. J. McNeese, and J. Gorman, "Human-Autonomy Teaming in Remotely Piloted Aircraft Systems Operations under Degraded Conditions," Arizona State University, Mesa, Arizona, 2018.

[36] T. Jensen, M. M. H. Khan, and Y. Albayram, "The Role of Behavioral Anthropomorphism in Human-Automation Trust Calibration," in *Artificial Intelligence in HCI*, Cham, 2020, pp. 33–53, doi: 10.1007/978-3-030-50334-5_3.

[37] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 1–8.

[38] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability," *Int. J. Soc. Robot.*, vol. 5, Aug. 2013, doi: 10.1007/s12369-013-0196-9.

[39] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2001.

[40] L. Huang *et al.*, "Distributed dynamic team trust in human, artificial intelligence, and robot teaming," in *Trust in Human-Robot Interaction*, 1st ed., Academic Press, 2020, pp. 301–319.

[41] R. Parasuraman and C. D. Wickens, "Humans: Still Vital After All These Years of Automation," *Hum. Factors*, vol. 50, no. 3, pp. 511–520, Jun. 2008, doi: 10.1518/001872008X312198.