Exploring the Effects of Swarm Degradations on Trustworthiness Perceptions, Reliance Intentions, and Reliance Behaviors

Izz aldin Hamdan[1], August Capiola[2], Gene M. Alarcon[2], Joseph B. Lyons[2], Keitaro Nishimura[3], Katia Sycara[3], & Michael Lewis[4]
[1]General Dynamics Information Technology, [2]Air Force Research Laboratory, [3]Carnegie Mellon University, [4]University of Pittsburgh

Swarms comprise robotic assets operating autonomously through local control laws. Research on human-swarm interaction (HSwI) investigates how human operators collaborate with swarms to accomplish shared goals. Researchers have begun to investigate the role of trust in HSwI, specifically which aspects of robotic swarms affect human trust. Through a human factors lens, the present research builds on earlier HSwI work and investigates the effect of swarm asset degradations on trustworthiness perceptions, reliance intentions, and reliance behaviors. Results showed that trustworthiness perceptions of and intentions to rely on swarms (but not reliance behaviors) were correlated, demonstrating the relation between theoretically relevant antecedents to trust in HSwI contexts. Contrary to past work, the results showed no statistical evidence that asset degradations differentially affect trustworthiness perceptions, reliance intentions, or reliance behaviors. Limitations of the current work (e.g., heterogeneity of post-intervention foraging behavior, sample size) are discussed and followed with future research suggestions.

INTRODUCTION

Automation and robotics are rapidly becoming more prevalent in modern life. The proper use of these systems by operators is, in part, dependent on trust. Trust is defined as one's willingness to be vulnerable to another based on the expectation that they will perform a particular action, regardless of one's ability to monitor/control them (Mayer et al., 1995). Over the last 25 years, significant research has been conducted on trust in the human factors literature, which comprises in part the specific literatures of trust in automation (Hoff & Bashir, 2015) and trust in human-robot interaction (HRI; Hancock et al., 2011). Each of these literatures investigates—to some extent—the willingness of a human to accept vulnerability to an automated system that may or may not be mobile, personified, and predictable.

With regards to HRIs, recent work has begun to investigate human trust toward robotic swarms (Kolling et al., 2015). Swarms comprise individual robotic assets which operate via local control laws (e.g., nearest neighbor algorithms) to self-organize and form emergent properties to complete tasks such as target foraging (Walker et al., 2012) and shape configuration optimization (Nagavalli et al., 2015). These local control algorithms allow the swarm as a whole to complete complex tasks that could not be completed by a single asset, while maintaining continuity in the midst of obstacles (Kolling et al., 2015), which arguably results in a more robust and resilient approach to adapting to one's environmental constraints versus a top-down (i.e., pre-planned) approach. Researchers have explored how to adapt the algorithms and other functionality underlying swarm operations (Ferrer, 2018; Haasdijk et al. 2014; & Nagavalli et al., 2017); however, the literature has largely neglected investigating user perceptions of robotic swarms.

Trust toward Automation and HRI

Researchers (Mayer et al., 1995; Lee & See, 2004, Schoorman et al., 2007; Schoorman et al., 2016) have explicated the trust process, separating the antecedents to and consequences of trust. In the human factors literature, perceptions of an automated aide's ability, benevolence, and integrity are thought of in terms of the aide's performance (what does the automation do and how well does it do that?), purpose (why does it do it?), and process (how does it do it?), respectively (see Lee & See, 2004, p. 59). These antecedents predict intentions to trust and the actual behavior that follows (i.e., reliance behavior). This relationship is theorized to be moderated by perceived risk in the situation. Previous research has demonstrated user's trustworthiness perceptions in robots to be an important factor (Alarcon et al., 2021). These context-specific antecedents to trust have demonstrated predictive validity for trust and reliance across a wide range of studies in the interpersonal trust literature (Colquitt et al., 2007) and more recently have been investigated in the trust in automation literature (Calhoun et al., 2019).

Although the literature is beginning to investigate the role of robot behavior on perceptions of trustworthiness, previous studies have focused on humans interacting with a single robot (Alarcon et al., 2020). Robotic swarms comprise tens, even hundreds, of unique assets which flock together to complete complex tasks (Kolling et al., 2015). As such, specific research on human-swarm interaction (HSwI) is needed.

Trust and HSwI

Past research shows that people may struggle to perceive and comprehend the behaviors of swarms (Nam et al., 2017). Thus, proper levels of trust toward swarms is important for HSwI, especially when the agents behave unexpectedly (de Visser et al., 2018). Loss of operator trust in the swarm may lead to premature abortion of the mission by the operator. In contrast, if an operator does not abandon a swarm when they should, resources ranging from robot assets to human life may

be at risk. These aforementioned scenarios demonstrate deviations (under- and over-trust, respectively) from calibrated trust, which describes user trust aligning with that of the automation's actual trustworthiness (Lee & See, 2004).

Thus, research is needed to isolate the factors that shape trust toward swarms to inform more optimal design of and training with swarm technologies. Recent research has demonstrated that differential proportions of asset degradations impact operators' trust toward swarms in target foraging tasks (Capiola et al., 2020). Although previous research has demonstrated manipulating the swarm's degradations can influence intentions to rely on those swarms (Capiola et al., 2020), participants were not given the opportunity to interact with the swarm. Further, the performance of the swarm did not impact the earnings participant received. We sought to replicate and expand on past work by addressing these limitations.

The Present Research

Capiola and colleagues (2020) explored if reliance intentions can be modulated based on swarm degradation (i.e., proportional loss of robotic assets). Participant's observed recorded swarms foraging at varying levels of degradation and were instructed to rate their intentions to rely on the swarm in a future target foraging task. The results demonstrated asset degradations influenced reliance intentions, but this study contained limitations. Notable limitations from their study were addressed and augmented in the present study.

In Capiola and colleagues (2020) study, participants watched recordings of the swarm and were not able to operate the swarm. Also, the performance of the swarm did not affect the participant's compensation; in other words, participants were not vulnerable to the swarm. Conversely, in the present study participants collaborated with the swarm and were given the opportunity to change the swarm's heading direction in each simulation. That is, they were given the opportunity to make a single input during each trial following a degradation. Participants could also exercise neglect benevolence, which is when a participant decides to let the swarm forage without interfering with its trajectory (Walker et. al, 2012). Further, participants received additional compensation for the amount of targets the swarm identified. These adjustments were made to involve the participants in the task as a collaborator instead of a spectator, increasing the trust relevance.

Capiola and colleagues (2020) found that reliance intentions increased as the percent of degraded assets decreased. We look to replicate and extend these findings by addressing the limitations of past work and including other trust-relevant criterion. Involving participants in the task and having swarm performance affect their compensation may make it easier for participants to establish differential trustworthiness perceptions as well as reliance on the swarm. Based on the extent trust in automation and HRI literatures, we explored the following hypotheses:

H1: Trustworthiness perceptions of, intentions to rely on, and response latencies toward swarms will be positively related.

H2: As swarm degradations increase, human perceptions of overall swarm trustworthiness will decrease.

H3: As swarm degradations increase, human intentions to rely on the swarm will decrease.

H4: As swarm degradations increase, participants' latencies of response times (change of heading direction, an instantiation of reliance behavior) will decrease.

METHOD

Design

Participants engaged in several rounds of the swarm foraging task (see Capiola et al., 2020) within a simulator developed in past literature (Walker et al., 2012). Participants viewed simulations of swarms comprising 256 assets forage for targets in an unknown space. One minute into each trial, participants were given the opportunity to make a single input to change the swarm's heading direction. Participants could make this input to the swarm heading at any time after the initial minute. A within-subjects design was used, such that all participants were presented with six degrees of degradation (5, 10, 15, 20, 25, and 50% of assets were degraded). Each degradation took place 1 minute into a 3-minute foraging task. Participants viewed six trials in a randomized order and could decide when to intervene (following a degradation). Following each trial, self-report measures were administered, including assessments of participants' trustworthiness perceptions of and intentions to rely on the swarm.

Simulator

The simulator (Walker et al., 2012) was produced in Microsoft Visual Studio 2017. The same parameters (i.e., percent of degradations, number of targets, no-fly zone locations) used by Capiola et al. (2020) were also used in this study. Participants interacted directly with the simulator in each trial (see Figure 1).

Participants

A total of 26 participants were recruited from the general population of a midwestern city via a combination of craigslist and word of mouth. Three participants were removed for having incomplete data. This resulted in a final sample of 23 participants (9 female), aged 25 – 62 years (M = 37.78) who participated for $20/60 minutes. In addition, participants were compensated $0.10 USD for each target collected in the foraging task.

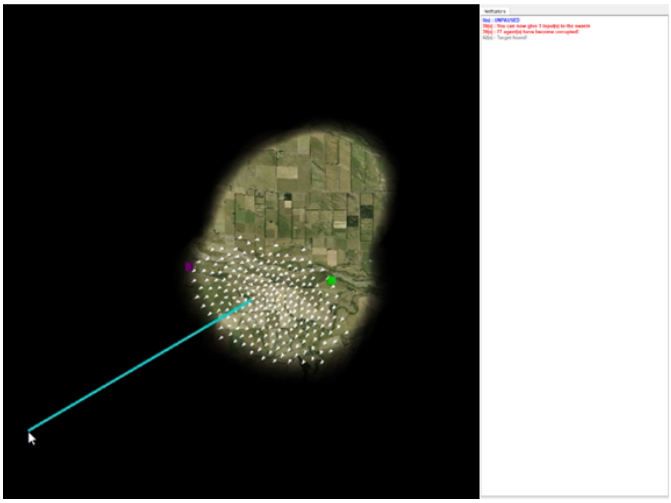
Manipulations

The focal manipulation in the current study was the percent of assets degraded per trial. Six trials were presented, with 5, 10, 15, 20, 25, and 50% of the assets degrading one minute into each 3-minute trial. Only one degradation occurred per trial. In an effort to maintain a true within-

subjects repeated-measures design, the order of trials was randomized per participant.

Figure 1

Simulator Snapshot



Note. Operator pointing, clicking, and dragging cursor to change the swarm heading direction following a degradation. The right-hand side of the image shows the notification panel, where participants are given updates regarding degradation occurrence, targets collected, and when an input may be offered.

Dependent Variables

Trustworthiness. To assess participants' trustworthiness perceptions of the swarm, we assessed self-reported perceptions of each swarm's performance (i.e., "The swarm was reliable in the target foraging task"), purpose (i.e., "I believe the swarm had my best interests in mind in the target foraging task"), and process (i.e., "The swarm adhered to stable principles in the target foraging task"). The items were adapted from Mayer et al. (1995) and Lee and See's (2004) explication of trust in automation. Each question was written to leverage the referent context (i.e., perceptions of swarms foraging). Participants responded to each item on a 5-point agreement scale (1 = *strongly disagree* to 5 = *strongly agree*).

Reliance Intentions. To assess participants' intention to rely on the swarm, we administered Lyons and Guznov's (2019) abbreviated four-item scale after each trial. A sample item was: "I think using the swarm will lead to positive outcomes." Participants responded to each item on a 5-point agreement scale (1 = *strongly disagree* to 5 = *strongly agree*).

Reliance Behavior. The swarm degraded 60 seconds into each trial. Once a degradation occurred, participants were given the opportunity to alter the heading direction of the swarm (see Figure 1). The response time of this change was recorded to assess participants' reliance behaviors. If participants altered the heading direction of the swarm immediately after degradation, this may indicate the participant has lost trust in the swarm. On the other hand, if participants withheld intervening and allowed the swarm to continue foraging (neglect benevolence; see Walker et al., 2012), this may indicate the participant trusts the swarm to complete the target foraging task.

Design Control

The occurrence of asset degradation was described as the likelihood that swarms will "encounter unexpected countermeasures, and it is unknown how many assets will be affected by such occurrences." Each condition comprised 50 targets (2 with controlled locations; 48 were randomized); 30 no-fly zones were equal in size and location to evade confounding environment difficulty. The swarm began in the center of the environment in every trial.

Procedure

Participants were welcomed and escorted into the laboratory. Upon receiving participant consent, a demographic questionnaire was administered. Participants viewed a training slideshow with information on robotic swarms, their operations in target foraging tasks, and instructions on how to operate the simulator. A 3-minute practice simulation was then presented. This simulation did not include any degraded assets, no-fly zones, or an obscured display. Following the practice simulation, participants completed six trials of the experimental task. Self-reports were administered following each trial. Upon completion of the final trial, participants were debriefed and compensated based on their performance as well as participation.

RESULTS

Table 1 displays the average Pearson correlation between self-reported assessments of swarm trustworthiness and reliance intentions, as well as the time (seconds) a participant altered the heading direction of the swarm (reliance behavior). Table 2 displays the *mean (SD)* of the 3-item trustworthiness scale, the 4-item reliance intentions scale, and reliance behaviors at each trial. Table 1 shows that trustworthiness perceptions of and intentions to rely on swarms were correlated, demonstrating the relation between theoretically relevant antecedents to trust in HSwI contexts. However, neither trustworthiness perceptions nor reliance intentions were related to reliance behaviors, evidencing partial support for H1.

Table 1

Average correlations for Trustworthiness, Reliance Intentions, and Reliance Behaviors

	TW	RI	RB
TW	*(.88)*		
RI	0.35*	*(.94)*	
RB	-0.01	-0.10	

Note. TW = Trustworthiness; RI = Reliance Intentions; RB = Reliance Behaviors (response time latency in seconds). *$p < .01$.

Table 2

Mean (SD) Trustworthiness, Reliance Intentions, and Reliance Behaviors at each percent of assets degraded

	5%	10%	15%	20%	25%	50%
TW Mean *(SD)*	3.67 *(.85)*	3.36 *(1.1)*	3.27 *(.96)*	3.41 *(1.08)*	3.12 *(.95)*	3.09 *(1.19)*
RI Mean *(SD)*	3.6 *(.94)*	3.39 *(1.05)*	3.13 *(.92)*	3.37 *(.98)*	3.05 *(1.07)*	3.22 *(1)*
RB Mean *(SD)*	33.26 *(43.32)*	19.09 *(32.82)*	24.96 *(38.83)*	25.96 *(39.16)*	28.52 *(35.18)*	14.22 *(16.05)*

Note. TW = Trustworthiness; RI = Reliance Intentions; RB = Reliance Behaviors (response time latency in seconds).

We conducted a repeated measures analysis of variance (RM ANOVA) for trustworthiness, reliance intentions, and reliance behavior. All RM ANOVAs were conducted using the "afex" package (Singmann et al., 2015) in the R programming language (R Core Team, 2018). No criterion met the assumptions of sphericity based on Mauchley's *W* test. Thus we used a Greenhouse-Geisser correction (note the *df*s reported).

First, an omnibus test found the effect of asset degradation on participants' perception of overall swarm trustworthiness was not significant, $F(3.17, 69.71) = 2.01$, $p > .05$, $\eta^2 = .04$. Thus, H2 was not supported. Next, an omnibus test found the effect of asset degradation on participants' reliance intentions was not significant, $F(3.21, 70.65) = 1.62$, $p > .05$, $\eta^2 = .04$. Thus, H3 was not supported. Finally, an omnibus test found the effect of asset degradation on participants' reliance behavior was not significant, $F(3.46, 76.01) = 1.03$, $p > .05$, $\eta^2 = .03$. Thus, H4 was not supported.

DISCUSSION

Swarms are a unique aspect of the HRI literature as they are amalgamations of several robots acting as one entity. Swarms comprise assets that operate via local control laws based on nearest neighbor algorithms (Kolling et al., 2015). Past research shows that people find it difficult to understand swarm performance in tasks such as shape configuration (Nam et al., 2017). Recently, however, research has shown that in a target foraging tasks, people do ascribe differential intentions to rely on swarms with varying levels of asset degradation in a future target foraging task (Capiola et al., 2020). The current paper explored the relation between trust-relevant criterion in HSwI and whether swarm asset degradations influence trustworthiness perceptions, reliance intentions, and reliance behaviors.

Results showed that trustworthiness perceptions of and intentions to rely on swarms were correlated. This finding demonstrates the relevance of theoretically relevant antecedents to trust (i.e., trustworthiness antecedents) explicated in trust in automation (Lee & See, 2004) in HSwI contexts. However, neither trustworthiness perceptions nor reliance intentions were related to reliance behaviors, thus only partially supporting our first hypothesis. In addition, the results demonstrated asset degradations did not have a significant influence on these criterion. It is worth stating that our study was underpowered as we only had a sample of 23 participants, we discuss this further in our limitations. Our results contradict the findings of (Capiola et al., 2020) which showed that people ascribe differential intentions to rely on swarms based on the percentage of asset degradations, though in general, the means in Table 2 appear to be in the anticipated direction. In their study, all participants were presented with the same six recordings of the swarm experiencing different levels of degradation. Although participants experienced six levels of degradation in our study, each trial was unique since participants interacted directly with the simulator and influenced subsequent foraging behaviors after their intervention. This could mean other extraneous variables affected participants' perceptions of the swarms which influence subsequent outcomes. For example, a swarm which modifies its heading direction after experiencing a degradation will be differentially influenced by nearest neighbors depending on when an input is offered by a participant. Thus, emergent features may provide visual cues (e.g., convex hull) that, although relevant in predicting trust in past work (Nam et al., 2018), may differentially affect perceptions of and reliance on swarms. Thus, heterogeneity of post-intervention foraging behavior may have had unique impacts on participants' trustworthiness perceptions of and intentions to rely on swarms.

As noted earlier, participants often have trouble understanding swarms (Nam et al., 2017). The overall cohesion of the swarm, the number of targets collected, or the presence of no-fly zones may impact human perception of the swarm. Capiola and colleagues (2020) proposed that the amount of targets a swarm identifies may affect the human operator's reliance on the swarm. Incentivizing swarm performance may have changed the context from that explored by Capiola and colleagues (2020). In summary, the intervention affordance along with incentives for swarm performance, while offering a context where trust is germane, may have also introduced other extraneous variables into this context. This, coupled with a small sample size, further necessitates that future research replicate and expand on the present work with an appropriate sized sample.

Limitations and Future Research

The present work has several limitations. Data collection took place in-person and was halted due to the coronavirus pandemic (COVID-19). This resulted in a small sample of only 23 participants. Post-hoc analyses showed our analyses were conducted without adequate power. A post-hoc sensitivity analysis found that assuming sphericity, 80% power ($\alpha = .05$), 6 repeated measures, and correlations between measures of $r = .50$, we would only have enough power to detect a f of .22 (Faul et al., 2009), or $\eta^2 = .05$. In our study, the sphericity assumption was not met; therefore, we recommend that future research replicate this study with a larger sample size to achieve adequate power.

Additionally, manipulating the reason *why* the swarm degrades ought to be investigated. Leveraging Lee and See (2004), people may have difficulty ascribing intentionality to automation, let alone swarms of robots. Thus, manipulating the reason why the swarm degraded (e.g., an enemy agent has corrupted the swarm's algorithms; the terrain has led to asset damage and thus loss of signal) could lead to drastically different (and perhaps differential) perceptions of swarm trustworthiness, intentions to rely on the swarm, and also the behavioral input of a human operator. As such, future research should apply and test these postulates in HSwI.

REFERENCES

Alarcon, G. M., Gibson, A. M., and Jessup, S. A. (2020). Trust Repair in Performance, Process, and Purpose Factors of Human-Robot Trust," *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, (pp. 1-6).

Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Capiola, A. (2021). Exploring the differential effects of trust violations in human-human and human-robot interactions. *Applied Ergonomics*, *93*, 103350.

Calhoun, C. S., Bobko, P., Gallimore, J. J., & Lyons, J. B. (2019). Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. *Journal of Trust Research*, *9*(1), 28-46.

Capiola, A., Lyons, J., Hamdan, I. A., Nishimura, K., Sycara, K., Lewis, M., ... & Borders, M. (2020, July). The Effects of Asset Degradation on Human Trust in Swarms. In International Conference on Human-Computer Interaction (pp. 537-549).

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology*, *92*(4), 909-927.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409-1427.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Ferrer, E. C. (2018, November). The blockchain: a new framework for robotic swarm systems. In *Proceedings of the future technologies conference* (pp. 1037-1058).

Haasdijk, E., Bredeche, N., & Eiben, A. E. (2014). Combining environment-driven adaptation and task-driven optimisation in evolutionary robotics. *PloS one*, *9*(6), 1-14.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434.

Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. (2013). Designing clinical research: An epidemiologic approach. Lippincott Williams & Wilkins; Appendix 6C, page 79.

Keller, D., & Rice, S. (2009). System-wide versus component-specific trust using multiple aids. *The Journal of General Psychology: Experimental, Psychological, and Comparative Psychology*, *137*(1), 114-128.

Kolling, A., Walker, P., Chakraborty, N., Sycara, K., & Lewis, M. (2015). Human interaction with robot swarms: A survey. *IEEE Transactions on Human-Machine Systems, 46*(1), 9-26.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.

Lyons, J., & Guznov, S. (2019). Individual differences in human-machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*(4), 440–458.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust*, Academy of Management Review, 20*(3), 709-734.

Nagavalli, S., Chien, S. Y., Lewis, M., Chakraborty, N., & Sycara, K. (2015, March). Bounds of neglect benevolence in input timing for human interaction with robotic swarms. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 197-204). IEEE.

Nagavalli, S., Chakraborty, N., & Sycara, K. (2017, May). Automated sequencing of swarm behaviors for supervisory control of robotic swarms. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2674-2681). IEEE.

Nam, C., Walker, P., Lewis, M., & Sycara, K. (2017, August). Predicting trust in human control of swarms via inverse reinforcement learning. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 528-533). IEEE.

R Core Team R (2018): A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future, *Academy of Management, 32*(2), 344−354.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2016). Empowerment in veterinary clinics: The role of trust in delegation. *Journal of Trust Research*, *6*(1), 76-90.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2015). afex: Analysis of factorial experiments. *R package version 0.13–145*.

Walker, P., Nunnally, S., Lewis, M., Kolling, A., Chakraborty, N., & Sycara, K. (2012, October). Neglect benevolence in human control of swarms in the presence of latency. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 3009-3014). IEEE.