Trust and Distrust Spreading in Interconnected Human–AI Teams: An Ergonomics Approach to Analyzing, Designing, and Evaluating Multi-Team Collaboration with AI Agents

Xiaoyun Yin
Role: Lead Researcher (Experimental Design, Data Collection, Analysis, Reporting)
Advisor: Jamie C. Gorman, PhD

Human Systems Engineering, Arizona State University
Study Conducted: August 2022 to Present

Study Objective

As AI systems become common teammates in high stakes settings like military operations, healthcare, and space exploration, a new challenge comes up: how does trust or distrust in one AI teammate spread across interconnected teams? This project investigated the mechanisms of trust and distrust spreading in a Team of Teams (ToT) structure where two human–AI teams (HATs) worked together on simulated reconnaissance missions. The goal was to understand how manipulating trust or distrust toward one AI agent in one team affects trust attitudes, coordination patterns, and performance not just within that team but also across teams that only interact through communication channels.

Applied Context: The Real World Problem That Motivated This Study

Consider a military intelligence, surveillance, and reconnaissance (ISR) operation where multiple distributed teams each rely on an AI system to process sensor data and recommend flight paths for unmanned aerial vehicles. Team Alpha's AI makes a targeting error. The navigator tells the cross-team channel: "Our AI can't be trusted with route planning." Team Bravo, 500 miles away with a perfectly functioning AI, hears this and starts second-guessing their own AI's recommendations. Within hours, operators across both teams are manually overriding AI suggestions, slowing mission tempo and increasing cognitive workload, even though only one AI actually malfunctioned.

This scenario is not hypothetical. As organizations deploy AI teammates at scale, trust failures can cascade through communication networks far beyond the original point of failure. Similar dynamics play out in remote healthcare (telemedicine teams sharing AI diagnostic tools), space operations (distributed mission control relying on AI for crew scheduling), and emergency response (agencies coordinating through AI-assisted dispatch). In all these domains, the user population consists of skilled professionals who must decide when to rely on and when to override AI recommendations, often based on secondhand information from other teams.

Our laboratory study was designed to capture the core dynamics of this applied problem in a controlled setting, so that the findings could be translated into design recommendations for real

multi-team AI deployments. The design principles, theories, and evaluation methods described below were chosen specifically to ensure this translation from lab to field.

Methodology

Analysis: User Population and Task Demands

The target community for this research is organizations that deploy AI agents as teammates in distributed, multi-team settings: military command and control, remote healthcare, space operations, and emergency response. In these domains, the user population consists of operators who work alongside AI systems for mission critical tasks while coordinating with other teams through limited communication channels. The core ergonomic challenge is that trust calibration (matching trust to actual AI reliability) must happen not just at the individual level but across entire multi-team systems.

The problem is significant. Trust is the primary factor determining whether human operators rely on, comply with, or reject AI recommendations. Miscalibrated trust leads to two failure modes: over-reliance (trusting an unreliable AI, leading to missed errors) and under-reliance (rejecting a reliable AI, leading to unnecessary manual work and cognitive overload). Most trust research has focused on a single human working with a single AI. But in real operational settings, trust attitudes spread between people through communication and observed behavior. If one team member communicates distrust about an AI, that distrust can propagate to other teams who have never directly experienced the failure. A single negative experience can cascade through an entire multi-team system.

We conducted a systematic analysis of the user population and task demands using the Cognitive Engineering Research on Team Tasks framework (CERTT; Cooke & Shope, 2004), which was specifically designed to study distributed team coordination in contexts analogous to military ISR operations. Our analysis identified the cognitive and communication requirements for each team role, the coordination demands of within-team and cross-team tasks, and the specific mechanisms (behavioral observation and verbal communication) through which trust attitudes could spread in a distributed environment. This analysis was grounded in interactive team cognition theory (Cooke et al., 2013), which treats team coordination as an emergent property of interaction rather than a sum of individual mental states.

Design: Experimental Intervention Grounded in Applied Principles

Based on our analysis, we designed an experiment using the CERTT Remotely Piloted Aircraft System Synthetic Task Environment (CERTT-RPAS-STE). The design choices were guided by two principles from cognitive engineering: (1) ecological validity, ensuring the laboratory task preserves the coordination demands of real ISR operations, and (2) experimental control, enabling us to isolate the specific mechanisms of trust spreading. The platform simulates a

reconnaissance mission where teams operate virtual unmanned aerial vehicles (UAVs) to photograph ground targets. The task requires tight coordination among three roles:

Pilot (AVO): Operates the UAV, controlling airspeed, altitude, and heading. The AVO was presented to participants as an AI teammate but was actually a human confederate using the Wizard of Oz method (Kelley, 1983), allowing precise behavioral control while maintaining the participant's belief they were working with a real AI.

Navigator (DEMPC): Responsible for mission planning, data exploitation, and communications. Also a confederate, who served as the trust/distrust "spreader" by communicating positive or negative statements about the AI to both teams.

Photographer (PLO): The payload operator controlling the UAV sensors. This was the participant's role. The PLO depended on accurate information from both the AVO and DEMPC to photograph targets, creating genuine task interdependence.

We recruited 40 participants (29 male, 11 female; ages 18 to 63; $M = 28.30$, $SD = 11.30$) from two geographically distributed sites: Arizona State University (Team A) and the University of Dayton Research Institute (Team B). Each session involved one participant from each site, forming a ToT of two interconnected HATs with 20 total sessions.

The experimental design was a 2 (Spreading Condition: Trust vs. Distrust) × 2 (Team: A vs. B) × 4 (Mission: 1 through 4) mixed nested factorial. Spreading condition was between subjects. Team was nested within ToT. Mission was within subjects. Team A received the direct manipulation; Team B received no direct manipulation. This is the key design feature: any changes in Team B's trust must reflect cross-team contagion rather than direct experience with modified AI behavior, isolating the spreading mechanism that is most relevant to applied settings where teams share information about AI systems they have not personally used.

The spreading manipulation had two components. Communication spreading involved the DEMPC on Team A conveying trust or distrust through text messages over Zoom (e.g., "I think the AVO-A is dependable" or "I don't think the AVO-A is trustworthy"). Behavior spreading occurred within Team A where the AVO-A either matched or mismatched UAV settings (e.g., reporting airspeed at 2100 when it was set to 1900 in the distrust condition). Mission 1 served as baseline. Spreading was introduced at Mission 2.

Measures: Evaluating from Multiple Angles

A core principle of ergonomic evaluation is that no single measure captures the full picture of system performance. We used three categories of measures, each addressing a different aspect of how trust operates in applied settings.

Self-report trust measures capture what operators *say* about their trust. We used Cognitive Trust (McAllister, 1995), which captures competence-based trust; the Multidimensional Measure

of Trust (MDMT; Ullman & Malle, 2019), which assesses trust across capacity, ethics, and sincerity dimensions; and Jarvenpaa et al. (1998) Team Trust Scale at three levels: trust in Team A, trust in Team B, and trust in the overall multi-team system. In applied settings, these correspond to post-shift surveys or debrief questionnaires.

Average Mutual Information (AMI) captures what operators actually *do*. AMI is an information-theoretic metric that measures how much one member's actions predict another's actions, grounded in interactive team cognition theory (Cooke et al., 2013). AMI was computed for all 15 dyadic pairs (within-team and cross-team) for each mission. The calculation is:

$$H(X) = -\Sigma\, p(x)\, \log_2 p(x) \qquad H(Y) = -\Sigma\, p(y)\, \log_2 p(y)$$

$$H(X, Y) = -\Sigma\, p(x, y)\, \log_2 p(x, y)$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

where $p(x)$ and $p(y)$ are the probabilities of each team member's discrete actions, and $p(x, y)$ is their joint probability. When actions are independent, $I(X; Y) = 0$. When one member's actions fully predict the other's, mutual information reaches the entropy of either variable. In applied settings, AMI could be computed from logged operator actions in real time, providing continuous behavioral monitoring without interrupting the task.

We also computed partial eta squared ($\eta^2 p$) as our effect size measure: $\eta^2 p = SS\_effect / (SS\_effect + SS\_error)$, where values of .01, .06, and .14 correspond to small, medium, and large effects (Cohen, 1988).

Team performance scores were computed independently by each site (ASU Team Performance, UDRI Team Performance) and combined into an MTS Performance score based on mission completion, target accuracy, and coordination efficiency.

Results

Manipulation Check: Trust/Distrust Spreading Worked

The trust/distrust manipulation successfully altered trust attitudes. Using MDMT scores, we found a significant main effect of Condition, $F(1, 36) = 4.43$, $p = .042$, $\eta^2 = .081$, and a significant main effect of Team, $F(1, 36) = 5.77$, $p = .022$, $\eta^2 = .103$. Table 1 shows mean trust ratings. The trust effect was 1.45 for Team A (direct manipulation) and 0.47 for Team B (communication only). The contagion ratio was 0.32, meaning about 32% of the trust/distrust effect transferred across teams through communication alone.

Table 1

Mean Trust Ratings (MDMT) for AVO-A by Team and Condition

Team	Condition	M	SD	Trust Effect
Team A	Trust	6.53	0.48	1.45
Team A	Distrust	5.08	1.37	

Team B	Trust	5.09	0.95	0.47
Team B	Distrust	4.62	1.50	

Note. Trust Effect = Mean difference between Trust and Distrust conditions. Contagion ratio = 0.47/1.45 = 0.32 (32%). n = 10 per cell.

Which Trust Measures Were Most Sensitive?

Not all trust measures detected the manipulation equally. Cognitive Trust (McAllister, 1995) was the most sensitive, showing a large Condition effect for Team A members rating AVO-A (η^2p = .531) and a large Condition × Mission interaction (η^2p = .419). The MDMT showed large Condition sensitivity for Team A (η^2p = .482). Jarvenpaa Team Trust at the MTS level showed a large effect for Team B members (η^2p = .194), meaning the manipulation affected system-level trust perceptions even for the non-manipulated team. This is important for practitioners: different trust measures capture different aspects, and the choice of measure affects what you can detect.

Table 2

Sensitivity of Trust Measures to Experimental Manipulation (Partial Eta Squared)

Measure	Condition	Mission	C × M	Size
Cognitive Trust AVO-A (Mem. A)	.531**	.310**	.419**	Large
MDMT AVO-A (Mem. A)	.482**	.072	.134*	Large
Team Trust MTS (Mem. B)	.194	.116	.062	Large

Note. ** $p < .01$, * $p < .05$. C × M = Condition × Mission interaction.

The Coordination vs. Trust Dissociation

The most practically important finding was the dissociation between behavioral coordination and self-reported trust. AMI was strongly predictive of team performance across all dyadic pairings (rs = .24 to .83, all ps < .05). But AMI showed only modest associations with subjective trust: only 23 of 210 correlations were significant.

The manipulation significantly affected trust ratings but did not significantly alter AMI patterns (all Fs < 1.65, all ps > .20). Even when people reported lower trust, their actual behavioral coordination did not change. For practitioners, this means that survey based trust monitoring alone will miss whether teams are actually changing their behavior in response to trust concerns.

Even more striking: in the Distrust condition, maintaining high trust in the unreliable AI was associated with *worse* performance (r = –.56). Over-reliance on a malfunctioning AI is an active performance liability, not just a theoretical concern.

ToT Level Dynamics

A 2 (Site) × 4 (Mission) × 2 (Trust Target) × 2 (Condition) mixed ANOVA using MDMT scores revealed a significant Site effect, $F(1, 18) = 10.37$, $p = .005$, $\eta^2p = .37$; a significant Trust Target effect, $F(1, 18) = 12.58$, $p = .002$, $\eta^2p = .41$; and a significant Site × Condition interaction, $F(1,$

18) = 6.97, p = .017, η^2p = .28. In the Trust condition, Team A rated both AVOs higher than Team B, but in the Distrust condition, trust levels converged. This convergence pattern is relevant to applied settings: distrust seems to create a "floor" that equalizes perceptions across teams, while trust advantages remain localized.

Discussion

This study demonstrates the full ergonomics practice cycle applied to a pressing challenge in human–AI teaming. Below, we discuss each finding and explicitly describe how it would translate to the applied ISR scenario and similar operational contexts.

Trust contagion is real and measurable. About 32% of the effect transferred across teams through communication alone. In the ISR scenario, this means if Team Alpha's navigator broadcasts distrust of their AI, we can expect roughly a third of that distrust effect to land on Team Bravo's operators, even though Team Bravo's AI is performing perfectly. The design implication: communication protocols between teams should include structured AI status reports (e.g., "AI route planning accuracy was 87% this shift") rather than allowing unstructured sentiment ("our AI is terrible") to dominate cross-team channels.

Reported trust and enacted trust are different things. Teams reported lower trust but did not change their coordination behavior. In an applied setting, a post-shift survey might show alarming drops in confidence while operators are still coordinating effectively. Conversely, operators might report high trust while developing workarounds that bypass the AI. The validation approach for operational settings: implement dual monitoring with periodic surveys combined with continuous behavioral metrics computed from system logs.

Over-reliance is an active hazard. The negative trust-performance correlation (r = –.56) in the distrust condition tells us that operators who maintained high trust in a malfunctioning AI performed worse. In the field, AI systems should actively signal reduced confidence when reliability drops, and dashboards should flag teams where trust remains high despite degraded AI performance.

Trust measures serve different purposes. Cognitive Trust was most sensitive to behavioral changes (η^2p = .531), while MTS-level Team Trust detected cross-team contagion. For practitioners implementing trust assessment programs, the recommendation is to use cognition-based scales (like McAllister) for detecting rapid changes in operator confidence, and team-level scales (like Jarvenpaa) for monitoring system-wide trust health.

Impact: Implementation and Validation for Applied Settings

While this study was conducted in a laboratory, the findings can be validated in operational settings through a staged process. First, dual monitoring (self-report + behavioral AMI) should be piloted in a field exercise with actual AI systems; AMI can be computed from system action logs that most AI-assisted platforms already generate. Second, structured communication

protocols for cross-team AI status reports should be tested in tabletop exercises, measuring whether they reduce unwarranted trust contagion. Third, over-reliance detection algorithms can be built using the trust-performance patterns identified here: when an operator's AI acceptance rate stays above 90% but AI accuracy drops below 80%, the system should flag a potential over-reliance risk.

Specific design recommendations:

1. Dual monitoring. Combine periodic trust surveys with continuous behavioral coordination metrics computed from system logs. Neither measure alone captures the full picture of how operators work with AI.

2. Trust-aware communication protocols. Require structured AI performance reports in cross-team channels. Include accuracy metrics, corrective actions, and current AI status rather than allowing unstructured sentiment to propagate.

3. Over-reliance detection. Build dashboards that flag when trust remains high but AI reliability has dropped. Alert supervisors, not just operators.

4. Context-appropriate trust assessment. Use cognition-based scales for detecting rapid attitude shifts; team-level scales for system-wide monitoring.

5. System-level trust management. Trust dynamics differ at the individual, team, and multi-team levels. Management strategies should address all three.

Limitations

The sample size (N = 40, 10 per cell) limits power for smaller effects. The Wizard of Oz method means participants interacted with a simulated rather than real AI; generalizability needs field validation. The CERTT platform, while validated for team research, is a laboratory setting. The text-based communication channel may not represent richer operational communication. The 32% contagion ratio may vary with organizational culture and AI system type.

Ergonomics Core Competencies Demonstrated

Analysis. User research and assessment (Analysis 1): we conducted systematic analysis of the three operator roles within the CERTT platform, identifying cognitive and communication requirements for each and how task interdependence creates opportunities for trust contagion. Organizational factors (Analysis 2): we identified how multi-team structure, geographic distribution, and communication channel constraints affect trust dynamics across the ToT. Cognitive and behavioral characteristics (Analysis 4): we measured cognitive trust, team trust, and multidimensional trust using four validated psychometric instruments, capturing both individual and group-level attitudes. Cognitive aspects of human-technology interfaces (Analysis 5): we used AMI to evaluate how AI behavior and communication affected operator trust calibration, revealing the critical dissociation between reported and enacted trust.

Design. Applied ergonomic principles to develop a controlled intervention (Design 1): the experimental protocol translated real-world trust contagion scenarios into testable laboratory conditions while preserving ecological validity. Designed the task within human capabilities and workplace context (Design 4): the CERTT task required realistic cognitive demands including monitoring, communication, and coordination under time pressure. Designed the multi-team organization structure (Design 6): the ToT configuration with separate within-team and cross-team communication channels mirrored the architecture of distributed operational teams. The Wizard of Oz method balanced experimental control with participant engagement.

Integration. Implemented and tested the full system across four missions per session and 20 sessions total (Integration 1, 2), collecting trust ratings, behavioral coordination data, and performance metrics at each measurement point. Produced both expected findings (32% trust contagion transfer) and unexpected findings (trust-performance inversion at $r = –.56$), demonstrating the value of evaluation that captures adverse as well as beneficial effects. Translated results into five design recommendations with specific validation steps for operational deployment, including field exercise piloting and over-reliance detection algorithm development (Integration 3: organizational testing).

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. Cognitive Science, 37(2), 255–285.

Cooke, N. J., & Shope, S. M. (2004). Designing a synthetic task environment. In S. G. Schiflett, L. R. Elliott, E. Salas, & M. D. Coovert (Eds.), Scaled worlds: Development, validation, and applications (pp. 263–278). Ashgate.

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. Journal of Management Information Systems, 14(4), 29–64.

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 193–196.

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of Management Journal, 38(1), 24–59.

Ullman, D., & Malle, B. F. (2019). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. 14th ACM/IEEE International Conference on Human-Robot Interaction, 618–619.