Effects of Human-Machine Interface Recommendation Accuracy on Trust when Controlling Collaborative Combat Aircrafts in Complex Missions

Sandro Scielzo
CAE USA
Arlington, TX
sandro.scielzo@caemilusa.com

Hely Lin
CAE USA
Orlando, FL
hely.lin@caemilusa.com

ABSTRACT

Future warfare demands are rapidly pushing the need to optimize collaborative decision-making in Human-Machine Teams (HMT). A primary means to support effective collaboration is via adaptive Human-Machine Interfaces (HMI) that mediate mission-centric HMT interactions. A clear and present need involves fighter pilots operating Collaborative Combat Aircrafts (CCA) in complex missions. However, a key research gap centers on HMT trust and its impact on the warfighter and mission outcomes. Consequently, the current study builds on previous HMT research that validated the Continuous Online Numerical Score (CONS), a novel trust measure for applied HMT settings (Hartzler et al., 2023). Our goal was to verify and extend HMT trust-related findings in a realistic environment: an F-16 pilot operating CCAs in a Suppression of Enemy Air Defenses (SEAD) mission. Specifically, we developed a SEAD mission using the Modern Air Combat Environment (MACE) and developed an HMI to operate CCAs within an F-16 Unit Training Device (UTD) simulator. The HMI was integrated into the cockpit in tablet format. The experiment design involved two manipulations: presence or absence of a secondary task to impact workload (fishing boats that were either neutral or hostile), and HMI recommendation accuracy on fused sensor data for enemy tracks. We used a repeated-measures design whereby each participant conducted multiple missions across our main manipulations. Twenty-eight participants were split into a pilot group and a naïve group. Subjective data involved trust questionnaires, the NASA Task Load Index (NASA-TLX), and CONS. Objective data involved Measures of Performance (MOPs) and Measures of Effectiveness (MOEs), interaction data with the HMI, and biometric data obtained from a wrist-worn device. Results confirmed CONS trust diagnostic power and further identified key measures that are both predictive of trust and performance. Results are further discussed in terms of HMI design implications to support complex HMT missions.

ABOUT THE AUTHORS

Sandro Scielzo is a Human Systems Technical Authority and Learning Science Fellow at CAE USA. Dr. Scielzo received his PhD in Applied Experimental Human Factors and M.S. in Modeling & Simulation from the University of Central Florida in 2008 and 2005 respectively. Sandro has over 20 years of experience researching next-generation training solutions for military and commercial applications. His current focus is on advancing state-of-the-art Human-Machine Team research and technologies.

Hely Lin is a Machine Learning Engineer at CAE USA. Hely received her bachelor's degree in biomedical engineering from the University of Florida in 2023 and is currently pursuing a master's degree in computer science, specializing in machine learning, from Georgia Institute of Technology. She is currently part of the Artificial Intelligence Research & Development team and supports advanced data analysis and AI projects.

Effects of Human-Machine Interface Recommendation Accuracy on Trust when Controlling Collaborative Combat Aircrafts in Complex Missions

Sandro Scielzo
CAE USA
Arlington, TX
sandro.scielzo@caemilusa.com

Hely Lin
CAE USA
Orlando, FL
hely.lin@caemilusa.com

INTRODUCTION

Collaborative Combat Aircrafts (CCA) represent autonomous air platforms designed to operate alongside crewed fighter jets in contested environments and engage in collaborative decision-making to maximize mission outcomes and operational flexibility (Airforce Technology, 2024). The employment of CCAs is aimed at disrupting air warfare—and the future of air combat in general—to maintain air superiority against our adversaries (e.g., Sweetman, 2024). The Department of the Air Force (DAF) Scientific Advisory Board defined key characteristics of CCAs as being able to autonomously execute high-level orders, which reduces pilot workload, and being more cost-efficient by having lower costs and being uncrewed (DAF, 2022).

By design, CCAs are collaborative in nature, and thus are subjected to the "tools to teammates" paradigm shift whereby automated systems are less and less seen as subordinates that fully depend on human input, and more and more as synthetic agents with which decisions are made and carried out in support of complex mission goals, just like you would with human teammates (Gibbs et al., 2024). The human factors implications are vast in order to achieve high performing CCA/pilot Human-Machine Teams (HMT), and an extensive body of science identified trust and Human-Machine Interfaces (HMI) as key mediating factors enabling highly effective HMTs, which we turn to, next.

Human Machine Teaming and Trust

Due to the exquisite sophistication of Artificial Intelligence (AI)-driven CCAs, trust in automation—a critical human factor—takes an even more prominent role, including the need to objectively measure this complex socio-affective construct in real time to promote calibrated trust via, for example, the manipulation of information on an HMI (e.g., Hartzler et al., 2023; Scielzo & Kocak, 2020). Because CCAs are designed to execute complex orders in full autonomy, just like a human teammate would, failure to meet performance-based expectations can lead to distrust and an increase in cognitive workload, resulting in CCA micromanagement and poor HMT performance (e.g., Gibbs et al., 2024; Hartzler et al., 2023; Hall & Scielzo, 2022).

The criticality of measuring trust objectively and in real-time is exemplified by the Defense Advanced Research Projects Agency (DARPA) Air Combat Evolution (ACE) program which assessed trust based on pilot interactions with the AI pilot interface and eye tracking data (e.g., DeMay, 2022). Gibbs et al. (2024) further verified the viability of using a pilot's gaze patterns as an objective behavioral indicator of trust that can be processed in real time. As a result, both biometric and behavioral metrics represent a promising way to assess a pilot's trust in the system objectively and in real time. In our previous study, Hartzler et al. (2023) validated the Continuous Online Numerical Score (CONS), which provides a simple and effective way to capture operator subjective perceptions of trust in automation across a mission. Results highlighted the diagnostic nature of CONS across trust-related behaviors when reacting to HMI recommendations (i.e., operator compliance, verification, or rejection of system recommendations). A recommendation from that study was to verify and extend these findings with fighter pilots operating CCAs in a high-fidelity simulated environment, which is precisely what we are exploring in this paper.

Human-Machine Interfaces Mediating HMT Performance

The goal of measuring an operator's trust in real time is to enable adaptive HMIs that respond to overtrust or distrust by altering the level of information transparency in order to achieve proper trust calibration; that is, a state of trust that matches the autonomous capabilities of a given system (e.g., Hartzler et al., 2023; Gibbs et al., 2024.; Wischnewski et al., 2023). As a result, designing effective HMIs is a key enabler of HMT performance. In a CCA paradigm, this translates into the type of HMIs that pilots interface with, whether integrated in a glass cockpit or as a discrete element

such as a tablet. Furthermore, current HMIs are often restricted to one input modality, typically touch, which lacks the flexibility in HMT collaboration and robustness of communications afforded by multimodal HMIs (e.g., Scielzo & Kocak, 2020).

Because of the complications in implementing multimodal HMIs for CCA command and control (C2), visual interfaces must follow proper human factors military standards principles (e.g., MIL-STD-1472), and autonomy design guidelines and taxonomies across a pilot's Observe, Orient, Decide, and Act (OODA) loop to control one or more uncrewed vehicles (e.g., Endsley & Kaber, 2018; Hightower, 2014). Failure to adopt sound warfighter-centric design principles can result in increased cognitive workload, distrust in the system, and loss of situation awareness, mostly when operating multiple uncrewed systems (Cummings et al., 2006). As a result, this paper will introduce an HMI interface designed to facilitate the operation of several CCAs in complex missions.

Collaborative Combat Aircrafts in Complex Missions

An additional recommendation from our previous study was to extend our findings on trust in a realistic and tactically focused mission environment. The purpose is threefold: (1) to investigate the extent to which CCAs can support mission outcomes in modern warfare, (2) to verify trust in automation patterns via biobehavioral and subjective metrics, and (3) to identify key HMI characteristics that are conducive to overall mission success. Extending our findings in an applied and tactically relevant scenario is primarily aimed at guiding trust-based adaptive HMI development in fast jet use cases.

The study presented in this paper centers on a 4th generation crewed platform operating four CCAs in a complex and military relevant scenario. We selected a 4th generation fighter to further explore the viability of extending lower-capability fighters (in comparison to 5th and 6th gen aircrafts) with CCAs. Overall, the main objective of this study was to provide mission relevant insights using adaptive HMIs when controlling CCAs, while shedding light on the complex relationship between trust, workload, and performance in an applied air-domain military environment.

Present Study

This study sought to verify and extend our findings on HMT trust in an applied military air domain environment. We opted for developing Suppression of Enemy Air Defenses (SEAD) missions due to their fast-paced nature and tactical relevance. The simulator employed was an F-16 Block 30 Unit Training Device (UTD), with a CCA HMI tablet mounted in the cockpit. Participants also wore the Emotibit wrist-worn biometric device. This study is unique in that it uses a relevant air-domain use case with fighter pilots as well as naïve participants to investigate the impact of trust on workload and performance.

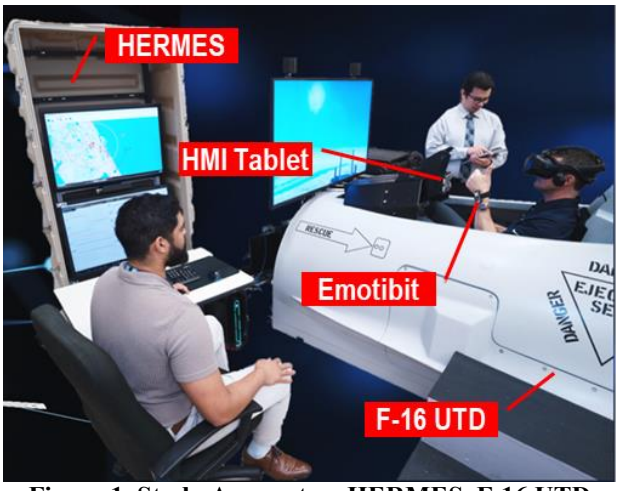


Figure 1. Study Apparatus: HERMES, F-16 UTD with CCA HMI, and Emotibit Biometric Device

The CONS measure, validated in our previous study (see Hartlzer et al. 2023), was integrated into the HMI tablet to allow participants to enter their subjective trust throughout the execution of SEAD missions (see Apparatus section). A dedicated experimenter station—the HMT Enhanced Research Manipulation & Experimentation Station (HERMES)—was developed to maximize experimental control and execution across our main manipulations: system accuracy and mission workload. HERMES facilitated the execution of complex human-subjects experiments using existing simulation-based training devices (such as the F-16 UTD used in this study), mission layouts using Modern Air Combat Environment (MACE), and biometric equipment. The controlled environment used in this study is shown in Figure 1. Although the figure shows a participant wearing a virtual reality headset, it was not used in the study for the practical purpose of using the CCA HMI, instead relying on an Out the Window (OTW) display. Finally, we developed behavior graphs for the CCAs using our Joint All Domain-AI (JAD-AI) technology to enable the realistic execution of orders from HMI inputs.

This experiment represents the culmination of an HMT Research & Development (R&D) effort, involving the adoption and development of multiple hardware and software assets integrated in a complex modeling and simulation environment to allow maximal experimental control. This overall setup will continue to serve as a "proving grounds" for HMT technologies in the future. Specific to this study, our main goals and corresponding hypotheses are as follows: (a) goal: validate CONS in an applied air domain environment; hypotheses: self-reported trust using CONS will significantly relate with recommendation accuracy and will exhibit convergence validity with established subjective trust questionnaires, (b) goal: identify biometric measures that are diagnostic of trust; hypotheses: biometrically-derived metrics from the wrist-worn device will significantly relate across trust-related events, and (c) goal: confirm relationship between trust, workload, and performance; hypotheses: low trust and high workload will have a negative impact on mission performance; conversely, high trust and low workload will positively impact mission outcomes.

METHOD

Participants

28 participants took part in our study. Participants were recruited without regard to age (M = 41.28, SD = 13.35) or gender (male = 24). The majority reported having a post graduate college degree (n = 16), while the remaining had an undergraduate degree (n = 10) or an associate degree (n = 2). Participants were evenly split between pilots and naïve groups. Pilots all had a military background flying various military combat air platforms (F-16, F18, F-15, F-35, AH-64) and trainers (T-6, T-38) with significant level of experience in terms of flight hours (M = 1596.61, SD = 347.27). Naïve participants had little to no experience flying, with one participant reporting 3 hours on a small DA-20 aircraft.

Study Design

A 2 System Accuracy (accurate or inaccurate sensor fusion) x 2 Workload (presence or absence of secondary task) within-subjects design was used with conditions counterbalanced across four mission events, each lasting a maximum of 15 minutes. That is, each participant completed all four mission events with different system accuracy profiles and mission workload. As shown in Table 1, the combination of both independent variables (recommendation accuracy and mission workload) manifested differentially across the four events.

Table 1. Summary of Mission Orders and Counterbalancing of Independent Variables

	Event 1							Event 2							Event 3							Event 4						
Order	**Msn**	**WL**	**Sys. Accuracy**					**Msn**	**WL**	**Sys. Accuracy**					**Msn**	**WL**	**Sys. Accuracy**					**Msn**	**WL**	**Sys. Accuracy**				
A	1	P	L	L	L	H	H	3	A	L	L	L	L	L	2	P	H	H	H	H	H	4	A	H	H	L	L	L
B	3	A	L	L	L	L	L	1	P	L	L	L	H	H	4	A	H	H	L	L	L	2	P	H	H	H	H	H
C	2	P	H	H	H	H	H	4	A	H	H	L	L	L	1	P	L	L	L	H	H	3	A	L	L	L	L	L
D	4	A	H	H	L	L	L	2	P	H	H	H	H	H	3	A	L	L	L	L	L	1	P	L	L	L	H	H

Note: participants completed the missions ("Msn") in one of four orders, A-D. Missions differed by workload ("WL") expressed by the presence ("P") or Absence ("A") of secondary task, and system accuracy (high "H" and Low "L") for each of the five mission tracks (i.e., tracks are either targets or decoys).

System accuracy was displayed in the CCA HMI as a sensor fusion confidence score (0-100%) for each potential enemy track, with a score of > 80% indicating high confidence that the track is an actual target versus a decoy. The Rule of Engagement (ROE) was to treat high confidence scores as *system recommendation* for target prosecution. This condition altered whether the CCAs combined sensors would classify the tracks correctly (each mission had five tracks). In the high system accuracy condition, the sensor fused confidence score was accurate for all tracks, while not accurate in low system accuracy, thereby potentially misleading participants as to which tracks were real targets or decoys. It was the participants' job to correctly prosecute real targets and dismiss decoys. Additionally, participants had the option to use the HMI and drill down to obtain individual CCAs' sensor data to "check" the math of the overall sensor fusion score. Mission workload was treated as a secondary task, that is, all missions had fishing boats present, but in the high workload condition a subset of these fishing boats would turn hostile and participants had to navigate around their threat rings to avoid detection, thus detracting from their primary SEAD objectives. As a result, workload was treated as presence (avoid enemy boats detection rings) or absence (all boats are neutral) of a secondary task.

Apparatus

Simulator, HMI Tablet, and Emotibit Wrist-Worn Device

We used a high-fidelity F-16 block 30 UTD and a tablet for the CCA HMI, which was mounted in the cockpit. Figure 2 provides a visual representation for both. Additionally, each participant was donned with an unobtrusive wireless Emotibit wrist-worn biometric device, primarily providing electrodermal activity (EDA), and photoplethysmogram (PPG) data, which uses an optical sensor to measure blood volume and used to derive Heart-Rate Variability (HRV).

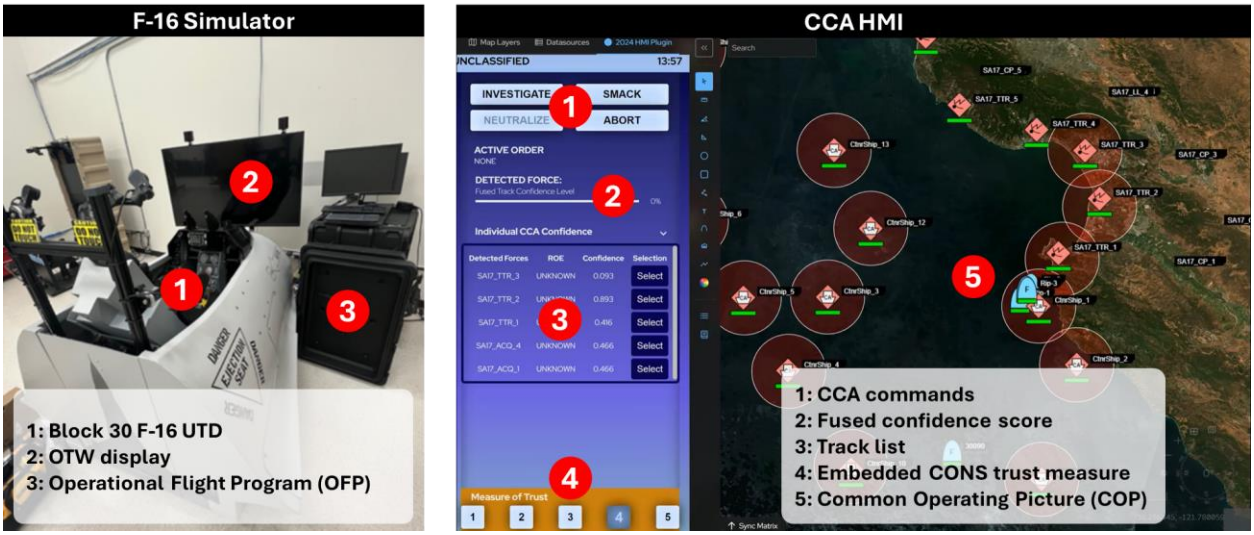


Figure 2. F-16 High-Fidelity Simulator and CCA HMI Participants Operated

SEAD Mission Control, Execution, and Objectives

As described earlier, HERMES (see Figure 1) was used by the experimenter to (a) control in real-time sensor fusion system accuracy for each track by following system accuracy profiles shown in Table 1, and (b) send Battle Damage Assessment (BDA) reports or reconnaissance reports to be displayed on the HMI. These reports notified participants if they correctly destroyed/classified a track, thus providing ground truth to participants after they destroyed a target or classified the track as a decoy. Additionally, MACE was used to develop a SEAD mission laydown populated with the F-16, CCAs, emitter tracks (targets and decoys), and surface boats. Surface boats were initially neutral and remained so in the low workload condition, whereas in the high workload condition the experimenter manually turned the boats hostile during the mission. We used a SEAD Subject Matter Experts (SME) to develop the mission and its objectives in a way that would maximize realism, while making it approachable for naïve participants.

Measures

Embedded CONS Measure

CONS was validated as a continuous self-reported measure of trust in our previous study (see Hartzler et al., 2023) and showed to be diagnostic of trust-related behaviors (i.e., system recommendation compliance, verification, or rejection) where operators controlled various autonomous drones in a civilian search and rescue operation. Additionally, Gibbs et al. (2024) used CONS data from that study as trust labels, which, when combined with operator gaze patterns, produced a preliminary model that could accurately classify trust. An important objective of this study was to verify the usefulness of CONS in a realistic and high-fidelity military simulation environment applied to the air domain with fighter pilots operating CCAs. As a result, CONS was embedded in the CCA HMI (see Figure 2, bottom left) as a 5-point scale, with 1 indicating the lowest level of trust and 5 indicating the highest level of trust. Participants were prompted to enter their trust levels as frequently as they desired. However, after 45 seconds, the CONS widget would blink to solicit input from participants. This resulted in an effective data capture of trust throughout all four missions for each participant.

Heart Rate Variability

A wireless Emotibit wrist-worn device gathered physiological data from each participant. It was placed on the left wrist, with sensors touching the skin under the forearm to maximize contact. Unfortunately, due to the metal shell of the F-16 UTD, we suffered significant data loss across many samples. Only PPG data had enough resolution and accuracy, thus the main biometric measure used in this study was HRV. The silver lining was that not many studies investigated the relationship between HRV and trust.

Surveys and Questionnaires

Participants initially completed a demographics questionnaire to gather individual differences information such as age, gender, and education. We also administered three surveys to assess automation bias: Perfect Automation Schema, PAS (Tschopp & Ruef, 2020); Propensity to Trust Technology, PTT (Jessup et al., 2019); and Trust in Automation Inventory, TAI (Chien et al., 2014). These surveys prompted participants to indicate their agreement to a series of statements using a 5-point Likert scale. Additionally, we used Shaefer's Trust Perception Scale – Human/Robot Interactions (TPS-HRI; Schaefer, 2016) as a validated questionnaire to measure participants' trust-related perception of the CCAs after each mission. The survey asks participants to rate 12 items on a 0-100 scale based on their last engagement with an automated system (i.e., to rate items based on automation perceptions during the mission). Items include assessing the system dependability, reliability, responsiveness, predictability, and consistency. Because this survey was administered after each mission, it was an ideal measure to test convergent validity with CONS. We also administered the National Aeronautics and Space Administration Task Load Index (NASA-TLX) after each mission to gather self-reports on perceived workload. The NASA-TLX is a gold standard measure of workload validated in the late 1980s and widely used in human factors studies (e.g., Hart, 2006). Finally, at the end of the study, participants completed a "Top 3, Bottom 3" survey asking them to describe the three best and worst parts of the study. All surveys and questionnaires were administered via SurveyMonkey.

Measures of Performance (MOPs) and Measures of Effectiveness (MOEs)

Mission performance and effectiveness are objective indicators that provide a detailed view of mission outcomes. The Commander's Handbook for Assessment Planning and Execution (Joint Staff, 2012) provides useful guidance for developing MOPs and MOEs, which was followed to develop our mission outcome measures, listed in Table 2.

Table 2. Mission Outcome Measures

MOPs *"doing things right"*	**MOEs** *"doing the right thing"*
• Total tracks classified	• Track classification accuracy
• Total mission time	• Percentage system recommendations compliance
• Ran out of time per mission (yes/no)	• Time to smack/neutralize tracks
• Arrived at rendezvous point (yes/no)	• Stayed clear of threat rings

Procedures

Study procedures are depicted in Figure 3. The top of the figure shows experimental events, with SEADs missions shown in dark blue. The bottom of the figure shows all measures collected in the study, with CONS and biometrics (bio.) in dark blue to denote continuous measures. From intro to debrief, the study lasted 120 minutes.

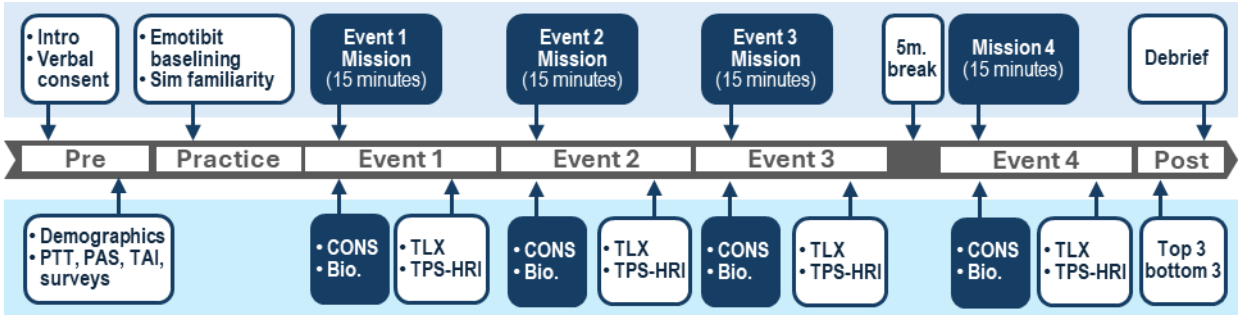


Figure 3. Timeline of Experimental Procedures

RESULTS

Our results center on five main areas of inquiry. First, we analyzed CONS data to verify its diagnosticity of trust-related behaviors. Second, we assessed convergence validity between CONS and TPS-HRI to further validate CONS in a military-relevant environment. Third, we analyzed heart rate as a predictor of trust. Fourth, we analyzed mission performance data against trust and workload. Finally, we assessed individual differences, trust biases, and study perceptions. Results for these areas of inquiry are provided next.

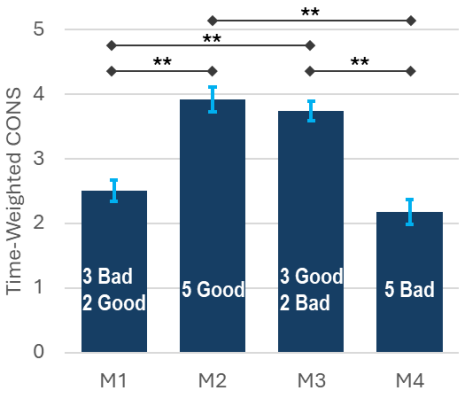
CONS Validation in Military Domain

A New Measure for CONS: Traditional and Time-Weighted

CONS was validated in a previous study (see Hartler et al., 2023) as a self-reported continuous measure of trust, gathering trust perceptions across a 1 (lowest trust) to 5 (highest trust) Likert-type scale at predetermined time intervals (e.g., every 40 seconds). This study implemented CONS as a widget embedded in the CCA HMI touchscreen tablet, and trust-related perceptions were gathered on the same 1-5 scale. This approach represents the traditional CONS methodology. However, for this study we also developed an alternative CONS measure, where participants inputs were weighted against time passed from their previous trust-related event. Trust-related events were defined as all instances in time (during the mission) when participants received ground truth messages regarding the outcome of targeting emitters and ignoring decoys. That is, every time a participant targeted (or ignored) what they thought was an emitter (or decoy) based on fused sensor confidence scores, they subsequently received ground truth reports regarding the accuracy of their decisions. We denote these as trust-related events because these reports could either be consistent with their decisions (e.g., prosecuting a real emitter) or discrepant (e.g., prosecuting a decoy). Thus, trust would either increase through confirmation or decrease through discrepant reports. As a result, with the time-weighted CONS measure, more weight was given to trust inputs with smaller time intervals with trust-related events, resulting in a measure with more granularity and bias towards inputs close in time from ground truth reports. This time-weighted measure also ranges from 1 to 5. In this result section, we are (a) using the original CONS measures to make it easier to interpret reported trust across different predictors, and (b) using the time-weighted CONS to facilitate detecting trends between predictors and trust reporting. Therefore, our analyses used both variations of CONS.

CONS Relationship with System Recommendation Accuracy

Exploratory analysis on CONS data was first performed to study the distribution of time-weighted reported trust across all missions. We were interested in overall trust response patterns when collapsed across all manipulations. What we found is a bimodal distribution of trust with peaks around trust values 2 and 3.8 (1-5 range). This indicates a cluster of trust inputs around a low trust value and a high trust value and suggests that missions had a differential impact on trust. Next, we took a closer look at where these peaks may be coming from by examining the trust distributions of each of the four missions. Remember that each mission was associated with a different schedule of system recommendation accuracy. Each mission had a maximum of 5 recommendations based on the 5 tracks participants had to identify as targets or decoys. Specifically, in mission 1 the first three recommendations were bad and the last two good. Mission 2 system recommendation accuracy was consistently good. In mission 3 the first three recommendations were good while the last two were bad. Finally, mission 4 consistently had poor recommendation accuracy. We hypothesized that recommendation accuracy would impact trust and that missions with a mix of good and bad recommendation accuracy would impact trust based on its initial level of accuracy. As a result, we conducted an Analysis of Variance (ANOVA) comparing time-weighted CONS trust scores across missions. A significant statistical difference was found for time-weighted CONS scores, with F (3, 84) = 23.99, $p < 0.01$, partial $\eta^2 = 0.47$, showing a large effect size. Pairwise t-test comparisons revealed significant differences between missions 1 and 2, 1 and 3, 2 and 4, and 3 and 4. Figure 4 provides missions' time-weighted CONS scores, with a pairwise comparisons table, and showing significant differences between missions via double asterisks (**). As hypothesized, recommendation accuracy impacted trust, with missions having good or initially good recommendations showing consistently higher trust when compared to missions with poor recommendations or initially bad recommendations.



Missions Compared		M_1	SE_1	M_2	SE_2	p
M1	M2	2.51	0.16	3.92	0.19	**<0.01**
M1	M3	2.51	0.16	3.74	0.16	**<0.01**
M1	M4	2.51	0.16	2.18	0.19	0.23
M2	M3	3.92	0.19	3.74	0.16	0.45
M2	M4	2.18	0.19	3.92	0.19	**<0.01**
M3	M4	3.74	0.16	2.18	0.19	**<0.01**

Figure 4. Impact of Recommendation Accuracy on Trust

Convergence Validity Between CONS and TPS-HRI

The TPS HRI survey was given after each mission and asked participants about their trust-related perceptions of the CCAs. First, we investigated item responses for the TPS-HRI and found strong multicollinearity between positively phrased items, as shown by cross correlation values ranging from 0.41 to 0.98 and variance inflation factor (VIF)

values ranging from 34.7 to 151.6. As a result, we proceeded to combine these items into one factor to produce a modified TPS-HRI. This was done by applying Principal Components Analysis (PCA) on the TPS-HRI positive description items. After fitting PCA on our set of TPS-HRI positive description factors and reducing the dimensionality to 1 component, we saw that the percentage of variance explained by the first component was 72.04%. This showed that the new combined feature captured sufficient information from the original dataset, allowing us to proceed with using this newly combined feature. Finally, we tested the correlation of the new feature against time-weighted mean trust and found a significant positive association, using Pearson correlation coefficient ($r(80) = 0.59$, $p < 0.01$), thus allowing us to claim convergence validity between CONS and an adapted version of TPS-HRI.

Heart Rate and Trust

The main biometric measure used in this study was HRV due to sensor data loss described above from procedural missteps (data collected only for pilots) and connectivity issues due to the interference of the metal shell of the F-16 UTD. After removing outliers, we applied a log transform on the heart rate dataset to filter out noise and reduce skewness. The transformed heart rate dataset was then ready to be analyzed. Our main goal was to investigate the relationship of heart rate across trust-related events versus all other mission activities (denoted as nominal events). However, we could not directly combine all participants' heart rates into one dataset because each participant may have a different resting heart rate. Factors such as age, fitness, and overall physiological state when starting a mission could all impact resting heart rate. Therefore, we instead used each participants' heart rate differences throughout a mission compared to their own resting heart rate as the measure for analysis. For each subject, we looked across a time window of 30 seconds before and after each trust-related event and labeled them as "trust event windows." Heart rate data outside trust event windows were labeled as "nominal heart rate." We assumed that each user's "nominal resting heart rate" is the average of their nominal heart rates. We then created two datasets: (1) "nominal heart rate": for each user, differences between each heart rate outside the trust event and the user's "nominal resting heart rate", and (2) "trust event heart rate": for each user, differences between each heart rate within a trust event and the user's "nominal resting heart rate". Figure 5 shows two trust event windows with log (heart rate) data plotted across time. We then analyzed differences in heart rate fluctuations between trust-related events and normal events. Specifically, a Mann-Whitney U test was conducted to see if there were significant differences in heart rate fluctuations around trust events versus outside. Results show a significant difference in log heart rate fluctuations during trust event windows ($M = -0.03$, $SD = 0.28$) versus outside those windows ($M = -0.01$, $SD = 0.27$) with $p < 0.01$, indicating that trust events likely had an impact on overall heart rate.

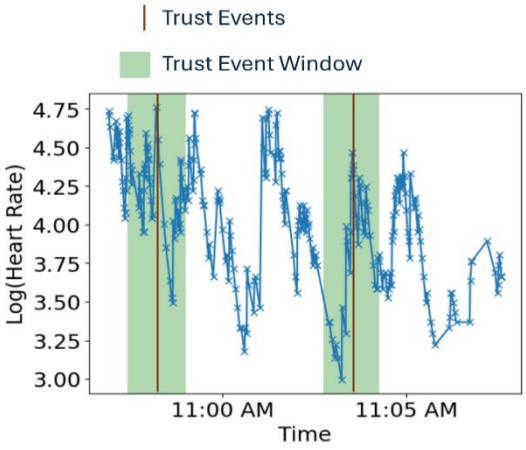


Figure 5. Mission Sample with Two Trust Event Windows

Mission Performance

From our list of MOPs and MOEs, one MOP, track identification accuracy, and one MOE, percentage recommendation compliance (to account for unequal recommendations), were used in our analyses. Other MOPs and MOEs either offered little variability or were deemed difficult to objectively quantify.

Performance and Recommendation Accuracy Across Missions

We conducted ANOVAs to look at percentage recommendation compliance (compliance) and track identification accuracy (accuracy) across missions. Compliance level across missions was statistically significant, with F (3, 72) = 32.38, $p < 0.01$, partial $\eta^2 = 0.58$, showing a large effect size. Accuracy across missions was also statistically significant, with F (3, 82) = 31.10, $p < 0.01$, partial $\eta^2 = 0.54$, again showing a large effect size. In support of our hypothesis that performance would be greater when the system provided good recommendations, pairwise t-test comparisons revealed significant differences in compliance and accuracy between missions that have good or initially good recommendations when compared to missions with bad or initially bad recommendations. Figure 6 summarizes these results, showing compliance and accuracy means across missions that provided overall good recommendations (good system) or overall bad recommendations (bad system).

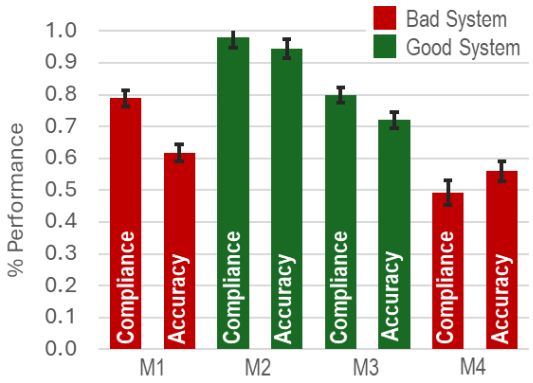
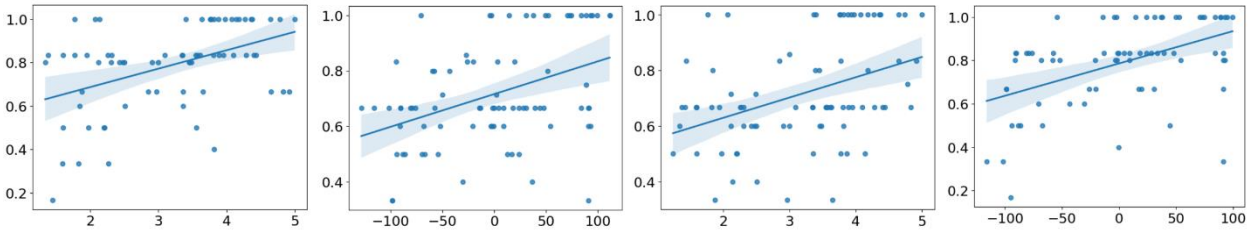


Figure 6. MOPs & MOEs Across Missions

Performance and Trust

We did not find significant relationships between CONS and performance, although correlations were positive. When correlating the modified TPS-HRI with performance, significant results emerged. Specifically, using Pearson's r we found a significant relationship between accuracy and TPS-HRI, with $r(79) = 0.41$, $p < 0.01$. We also found a significant correlation between compliance and TPS-HRI, with $r(69) = 0.48$, $p < 0.01$, (see Figure 7).



From left to right: (1) compliance across time-weighted mean trust, (2) accuracy across modified TPS-HRI, (3) accuracy across time-weighted reported trust, (4) compliance across modified TPS-HRI.

Figure 7. Correlating MOPs and MOEs with Trust

After proceeding to collapse missions 2 and 3 (good system) and missions 1 and 4 (bad system) to achieve a larger sample size, we conducted an ANOVA comparing time-weighted CONS across "good" and "bad" system missions. We found a significant statistical difference between "good mission" CONS scores ($M = 3.82$, $SE = 0.12$) and "bad missions" CONS scores ($M = 2.37$, $SD = 0.12$), with F (1, 84) = 69.53, $p < 0.01$, partial $\eta^2 = 0.46$, showing that, as hypothesized, participants trusted the system more when its recommendations were accurate.

Performance and Workload

No significant association was found between age and time-weighted reported trust during high workload missions. However, using Pearson's r we found that for low workload missions, there was a significant positive association between the two, with $r(39) = 0.46$, $p < 0.01$. We also found significant negative correlations on mission 1 using Pearson correlation coefficient between NASA-TLX and both system recommendation compliance ($r(26) = -0.51$, $p = 0.01$, one-tail) and track identification accuracy ($r(26) = -0.45$, $p = 0.03$, one-tail). Unfortunately, we did not find additional significant correlations across other missions, primarily due to poor data distributions. However, we were able to at least partially verify that as workload increased, performance decreased, as hypothesized.

Individual Differences, Trust Biases, and Study Perceptions

Participants' trust in automation biases were captured by the PTT, TAI, and PAS (see methods section). When correlating all three measures using Pearson correlation coefficient, only the PTT was found to be significantly positively correlated with the TAI, with $r(28) = 0.48$, $p < 0.01$. We then looked at the relationship between age and

trust biases. Unexpectedly, we found statistical trends showing a positive correlation between age and PTT ($r(26) = 0.31$, $p = 0.11$) and PAS ($r(26) = 0.31$, $p = 0.10$), going against the notion that younger folks are more predisposed to trust automation. We then conducted an ANOVA comparing PTT and TAI scores across pilots and naïve participants (also referred to as non-pilots). A significant statistical difference was found for PTT scores between pilots ($M = 3.96$, $SD = 0.42$) and non-pilots ($M = 3.58$, $SD = 0.41$), with F (1, 26) = 5.94, $p = 0.02$, partial $\eta^2 = 0.19$, showing pilots trusting automation more than non-pilots.

We also looked at the data from the "Top 3, Bottom 3" survey that gauged participants' perception of the overall study they just completed. Participants provided 95 total statements, nearly evenly split between 47 positive and 48 negative statements. Additionally, 43 statements were from pilots, whereas 52 were from naïve participants. We also found a significant age gap between pilots ($M = 50.00$, $SD = 9.51$) and non-pilots ($M = 32.73$, $SD = 10.42$) with $t(27) = 4.65$, $p < .001$, showing a generational gap between older pilots and younger non-pilots. All statements from the survey were then coded and sorted into six overall dimensions, each split into positive or negative statements. These dimensions were: mission, HMI, tablet, trust/distrust, F-16, and CCAs. Figure 8 provides a distribution of positive (+) or negative (-) statements across dimensions for pilots and naïve participants. Of particular interest were the positive statements about the mission from pilots, validating our attempt to create a realistic SEAD scenario with CCAs. Additionally, interesting patterns showed that younger, naïve participants provided more positive and negative statements for the HMI and F-16 dimensions when compared to pilots. When looking at the statements, younger participants were eager to provide interface feedback and expressed both excitement and difficulty of flying an F-16. Interestingly, distrust statements were similarly distributed across pilots and non-pilots. Also noteworthy was the higher volume of positive statements for the CCAs coming from pilots, indicating that CCA behaviors matched pilots' expectations, whereas non-pilots seemed to be less impressed with the technology.

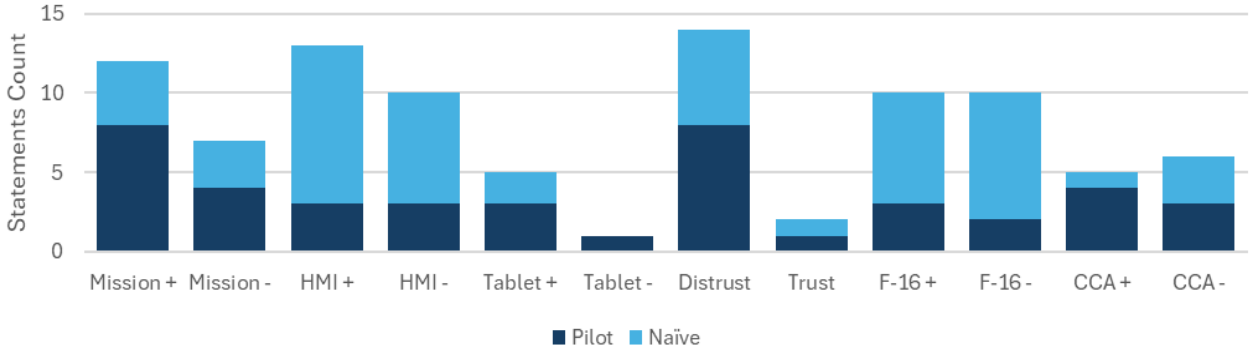


Figure 8. Distribution of Statements Across Dimensions for Pilots and Naïve Participants

DISCUSSION

Our main purpose was to verify and extend previous findings on HMT trust using a realistic air-domain military mission environment involving CCAs controlled via an HMI tablet mounted in a fast jet cockpit designed to facilitate HMT dynamics. Specifically, we sought to validate CONS as a diagnostic tool for trust-related behaviors, explore the viability for biometrically derived metrics to assess trust, and identify the relationships between trust, performance and workload in a tactically relevant mission environment. Additionally, we wanted to ascertain the overall impact trust biases and individual differences had on mission performance, including overall perceptions of HMT technologies from pilots and non-pilots alike. Overall, our findings from this study advance our understanding and ability to assess complex HMT dynamics, with result implications discussed next across our main areas of inquiry.

CONS and Trust-Related Behaviors

This study sought to extend CONS validity in an applied, air-domain military environment, using high-fidelity assets. Because the missions used in our experiment had specific and discrete trust-related events (i.e., receiving ground truth messages), we were able to enhance CONS by weighing scores based on elapsed time from trust-related events. Using time-weighted CONS, results from our experiment confirmed our hypothesis that CONS significantly related with system recommendation accuracy. An interesting finding was that system accuracy at the beginning of a mission anchored trust perception. When accuracy changed within a mission, it did not significantly change trust perception. This finding may fall under "first impressions" matter, but it also shows that trust biases from previous missions did not have a meaningful impact on a new mission (i.e., a new mission would "reset" trust perceptions). This could be

explained by the overall unpredictability of the system's accuracy. As a result, it is safe to infer that trust biases can be more prominent when a system is consistently "good" or "bad." Overall, CONS was successfully validated for pilots operating CCAs. Validation is further reinforced by convergence validity with the established TPS-HRI, and specifically via our enhanced TPS-HRI which only kept items that explained most of the variance in responses.

Biometrics and Real Time Trust

Gleaning biometric data in real time in conjunction with CONS can lead to real-time and objective measures of trust (see Gibbs et al., 2024). In this study we sought to collect biometric data via an unobtrusive wrist-worn device to identify potential biometric measures of trust. Unfortunately, we suffered from significant data loss due to experimenter error and the F-16 UTD metal shell, which interfered with the wireless connectivity, but were able to extract usable heart rate data. Heart rate data for pilots, although noisy, had enough resolution for processing after removing outliers and applying a log transformation to normalize overall distributions. Although clear inferences are difficult to make due to biometric data manipulations, we were able to establish a preliminary relationship between heart rate and trust. As a result, with more investigations and robust biometric datasets, heart rate, in conjunction with other metrics could become an objective indicator of trust.

Mission Performance, Trust, and Workload

An important aspect of this experiment was to confirm the relationship between trust, workload, and performance. A measure of performance, track identification accuracy, and a measure of effectiveness, system recommendation compliance, showed to be significantly different across system recommendation accuracy, which supported our hypothesis that performance would be greater when the system provided good recommendations. Specifically, both CONS and TPS-HRI showed to be related to performance, with higher trust leading to higher performance when system recommendations were good. Finally, we were not able to verify a negative relationship between workload and performance, with the exception of NASA-TLX being significantly higher in one of two high-workload missions. It may be that our workload manipulation (i.e., presence or absence of enemy boats) was not sufficiently tasking.

HMT Technology Perceptions and Biases

Two of our three trust bias questionnaires showed to be significantly related, thus indicating some validity in assessing trust biases. An interesting finding showed a significant correlation between age and trust biases, going against the notion that younger folks are more predisposed to trust automation. When coding our exit survey, we found that pilots had overall more positive statements about the SEAD mission, validating our need to create a realistic environment.

Study Control, Limitations, and Future Research

Exerting experimental control was achieved via HERMES, which allowed experimenters to manipulate, in real time, the level of system accuracy. Unfortunately, the loss of some biometric data restricted our analyses to heart rate data. Future experiments using biometrics need to thoroughly test data integrity. However, in spite of those limitations, we conducted a successful and complex experiment, which yielded results that can inform the future of training with CCAs and the development of adaptive HMIs to facilitate HMT collaboration. Specifically, future research on HMTs should include CONS as a validated measure to capture subjective trust perceptions. Additionally, future research should also verify the potential for heart rate variability to be a viable indicator of trust. The advantage of operationalizing trust via biometric indicators is to yield objective and real-time trust metrics. When validated against CONS these real-time trust measures can inform the adaptivity of an HMI to effectively calibrate a pilot's trust and improve HMT collaborative decision-making and mission outcomes. It is imperative for future research to identify design principles, rules and methods to optimally adapt HMIs, which is key to enable HMT performance.

CONCLUSION

This study extends our knowledge of mediating HMT trust via adaptive HMIs in support of training tactics with CCAs and achieving readiness in future warfare paradigms. Specifically, we recommend using CONS when developing adaptive HMIs that empower HMT collaboration, and to empirically verify that HMI design requirements can effectively support pilots across the OODA loop. Ultimately, CONS can be instrumental in the development and validation of biometrically derived real-time models of trust. In turn, these models can drive adaptive HMIs to trigger trust calibration methods, including transparency, level of automation, and decision-making support with the warfighter "*in*" or "*on*" the loop. Finally, we have shown the importance of holistically assessing HMT performance and effectiveness, which is essential to address the future of training with HMTs. In fact, the current HMT training landscape lacks defined HMT competencies, proficiency requirements, and effective mission rehearsal environments to ensure HMT readiness. We believe this paper offers valuable insights to address these gaps.

ACKNOWLEDGEMENTS

This study was made possible thanks to an internal project funded by CAE USA. We would like to acknowledge BGI pilots for participating in our study, and in particular Adam "Al" Kieda who piloted our experiment. Special thanks to Chris "SLAM" Duncan who served as our primary SEAD mission SME. Finally, a debt of gratitude goes to Alvin Abraham, who oversaw the human-subjects experiment, and Nicholas Crothers, who supported our data analyses.

REFERENCES

Airforce Technology (2024, June). Collaborative Combat Aircraft. USA - Airforce Technology

Chien, SY., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). Towards the development of an inter-cultural scale to measure trust in automation. In *International Conference on Cross-cultural Design*. Springer, Cham.

Cummings, M. L., Nehme, C. E., Crandall, J., & Mitchell, P. (2007). Predicting operator capacity for supervisory control of multiple UAVs. *Innovations in Intelligent Machines*, *1*, 11-37.

DAF Scientific Advisory Board (2022, December). Collaborative Combat Aircraft for Next Generation Air Dominance. DAF SAB FY22 Study ToRs_SecAF Final.pdf

DeMay, C. R., White, E. L., Dunham, W. D., & Pino, J. A. (2022). Alphadogfight trials: Bringing autonomy to air combat. *Johns Hopkins APL Technical Digest, 36(2)*, 154-163.

Endsley, M. R. (2018). Level of automation forms a key aspect of autonomy design. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 29-34.

Gibbs, S., Tanner, V., Larson, E., Scielzo, S., & Abraham, A. (2024). Towards a real-time model of trust in human-machine team paradigms. In *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL*.

Hall, B., & Scielzo, S. (2022). Red Rover, Red Rover, Send an F-35 Right Over: Assessing Synthetic Agent Trust in Humans to Optimize Mission Outcomes in Mosaic Warfare. *MODSIM World 2022*.

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage.

Hartzler, B. M., S. Scielzo, A. Abraham, R. Wong & S. Kohn (2023). Effects of Trust Calibration on Human-Machine Team Performance in Operational Environments. In *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL*.

Hightower, T. A. (2014). Boyd's OODA loop and how we use it. *Tactical Response*.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21* (pp. 476-489). Springer International Publishing.

Joint Staff (2012). *Commander's Handbook for Assessment Planning and Execution*. CreateSpace Independent Publishing Platform.

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "Trust Perception Scale-HRI". In *Robust Intelligence and Trust in Autonomous Systems* (pp. 191-218). Boston, MA: Springer US.

Scielzo, S., & Kocak, D. M. (2020). On the Need for Building Trust With Autonomous Underwater Vehicles. *Marine Technology Society Journal*, *54*(5), 15-20.

Sweetman, B. (2024). The Need for Collaborative Combat Aircraft for Disruptive Air Warfare. Final Article.pdf

Tschopp, M. & Ruef, M. (2020, May 07). PAS – The Perfect Automation Schema: Influencing Trust. PAS - The Perfect Automation Schema: Influencing Trust

Wischnewski, M., Krämer, N., & Müller, E. (2023, April). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).