1.1 HUMAN-AUTONOMY TEAMING IN A FLIGHT FOLLOWING TASK

1.1.1 Author and address/contact info

R. Jay Shively
UAS Integration into the NAS
Detect and Avoid, Sub-Project Manager
NASA-Ames Research Center

1.1.2 HAT approach description

This chapter summarizes two papers [1,2] that describe a study demonstrating and evaluating proposed tenets of Human-Autonomy Teaming (HAT) in an aviation context. In this study, HAT features were derived from three tenets and built into an automated recommender system on a ground station. These HAT tenets were:

- *Bi-directional Communication*: Communications between humans and autonomous agents are necessary for establishing and maintaining a shared knowledge of task goals, exchanging information, and reporting errors (human or autonomy-based) and status.

- *Automation Transparency*: The operator must understand the automation agent’s intent and reasoning, and be able to elicit additional factors on which the agent has based a recommendation for a course of action. Operators must be able to see what the agent is doing and, ideally, predict what the agent will do next. Transparency requires operator knowledge of the general logic of how the agent works and accurate mental models of its functioning [3]. At the same time, the autonomous agent must understand (to some extent) the current preferences, attitudes and states of human operators in the team. To be truly transparent, communication should use a shared language that matches the operator's mental model.

- *Operator Directed Interface*: Effective communication requires that the human operator can easily and accurately direct the automation, and override autonomous-system decisions, if necessary. Dynamic allocation of tasks based on operator direction and context also allows a more agile, flexible system and helps keep the operator in the loop.

This study focused primarily on interactions with one piece of automation, the Autonomous Constrained Flight Planner (ACFP). The ACFP is designed to support rapid diversion decisions for commercial pilots in off-nominal situations [4]. The ACFP compiles information from several sources such as ATIS broadcasts, METAR weather reports, GPS location and terrain, aircraft condition, and airport/runway characteristics. Evaluations are made for various factors (e.g., risk associated with the enroute, approach, and landing phases of flight, fuel usage, weather, terrain, distance, facilities). These evaluations are then aggregated to produce an overall score. Much effort has gone into enhancing this tool not only in capability but also in transparency [5,6,7]. In this study, participants used the ACFP at a ground station designed to aid dispatchers in a flight following role to reroute aircraft in situations such as inclement weather, system failures and medical emergencies. Participants performed this task both with HAT features enabled and without and provided feedback.

1.1.3 Method

1.1.3.1 Participants

Four dispatchers (median dispatch experience was 11 years) and two pilots (both active duty and each with over 10,000 hours flown as a line pilot) participated in this simulation.

1.1.3.2 Simulation environment

Our simulation ground station is described elsewhere [8,9,10]. Both the ground station and the concept of operations have evolved through a series of development and test cycles (i.e., “spiral development”). Our current concept of operations [10] envisions a role for ground support in monitoring and assisting aircraft in an advanced flight following mode, that requires increasingly sophisticated automation and enhanced collaboration between the operator and the automation. The following sections describe the components of the ground station focusing on the HAT features included for this simulation.

Aircraft Control List. The center of the station hosts an Aircraft Control List (ACL), the primary tool for managing multiple aircraft and switching the focus between aircraft (see Figure 1A). The ACL provides information crucial for situation awareness such as callsigns, departure and destination city pairs, estimated time of arrivals, flight plans, souls on board, and pilot details. This version of the ground station was designed to monitor, with the help of automation, a large number of aircraft (up to 30). Alerts were issued for failure to adhere to a clearance, failure to stay on path, environmental threats (weather on flight path or at the destination, and airport closures), system issues, and failure of the pilot to acknowledge a flight deck alert.

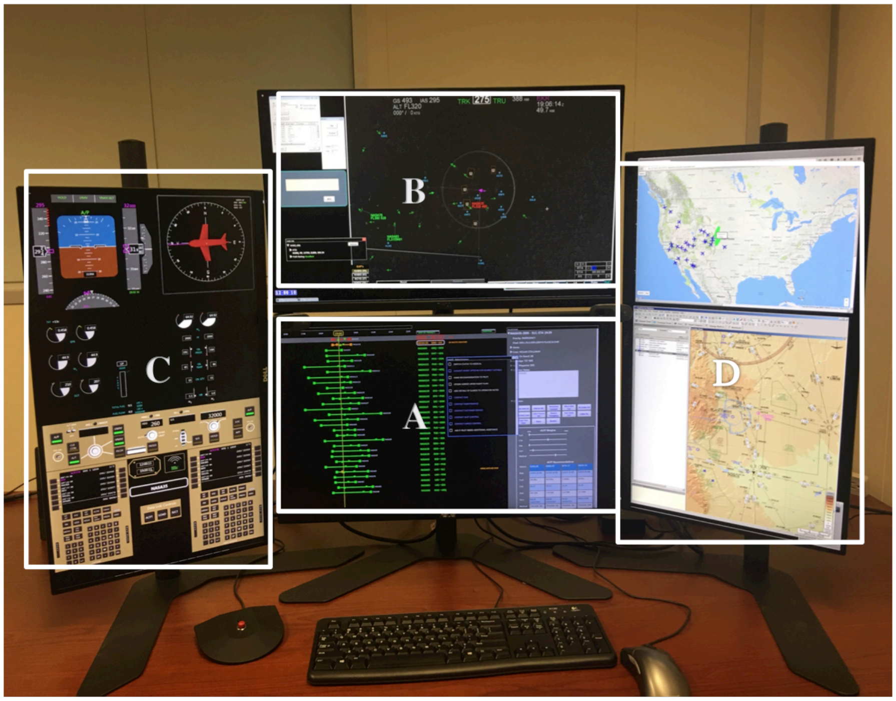
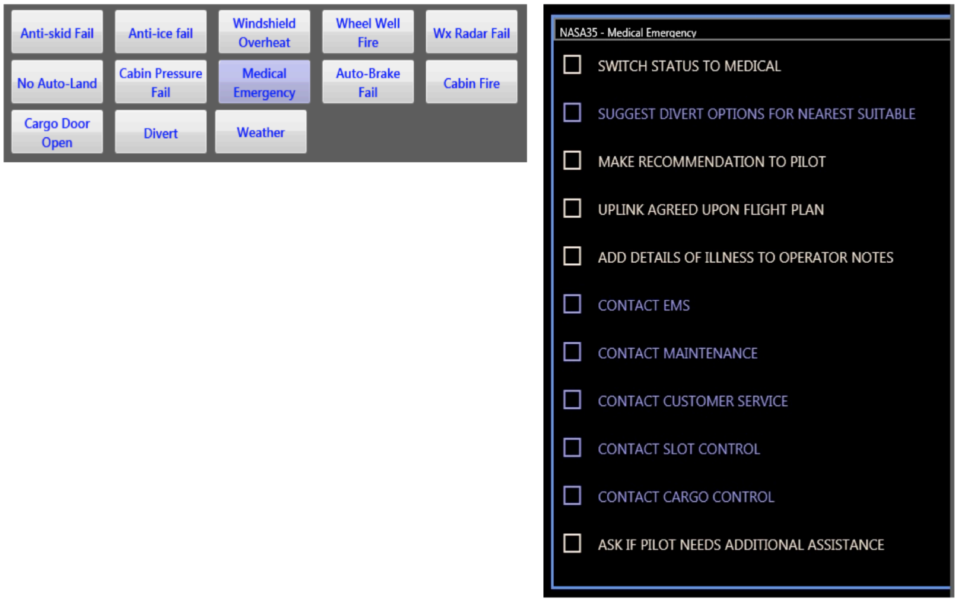


Figure 1. Simulation Ground Station. A: Aircraft Control List (ACL), augmented with timeline, alerting information and HAT features. B: Traffic Situation Display (TSD). C: Flight controls and displays for the selected aircraft in read-only mode. D: CONUS map and charts.

Operator Directed Interface: We have adopted the playbook approach to set system goals and manage roles and responsibilities between the operator and the automation [11]. When the operator selects a play, the ACFP is triggered with preset weights and the corresponding checklist appears on the display identifying operator tasks in white and automation tasks in blue (see Figure 2).



NASA35 - Medical Emergency

- □ SWITCH STATUS TO MEDICAL
- □ SUGGEST DIVERT OPTIONS FOR NEAREST SUITABLE
- □ MAKE RECOMMENDATION TO PILOT
- □ UPLINK AGREED UPON FLIGHT PLAN
- □ ADD DETAILS OF ILLNESS TO OPERATOR NOTES
- □ CONTACT EMS
- □ CONTACT MAINTENANCE
- □ CONTACT CUSTOMER SERVICE
- □ CONTACT SLOT CONTROL
- □ CONTACT CARGO CONTROL
- □ ASK IF PILOT NEEDS ADDITIONAL ASSISTANCE

Figure 2. Playbook and Checklist. Operators could call plays for various events and in the HAT condition were provided a checklist of roles and responsibilities. This is an example of a Medical Emergency play.

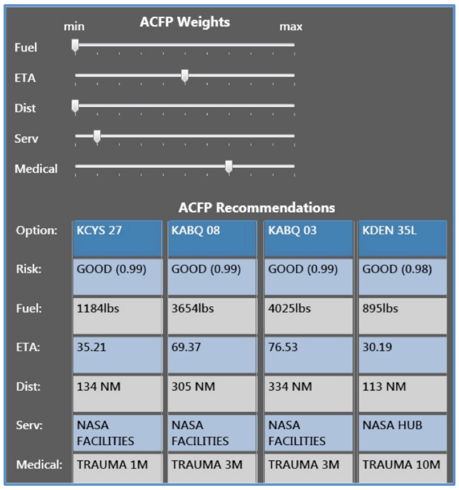
Building in Transparency. The ACFP has the capability of weighting the factors that go into its recommendations differently based on the situation. We increased transparency by explicitly showing these factors and weights when the ACFP is enabled (see Figure 3). Additionally, we translated the scores for the ACFP factors to more meaningful numbers and descriptors for the operator (e.g., presenting nautical miles (nm) instead of a distance score). In the example above, a Medical Emergency play was called which resulted in the distance to medical facilities (Medical row) and time to destination (ETA row) given more weight than other factors.

Building in Bi-directional Communication. We preset weights for each play and presented the weight settings (top of Figure 3). The operator can adjust the weights to better fit the situation and see how the recommendation is affected. Again, using the example in Figure 3, the operator could adjust the ACFP weight for estimated time of arrival if that was deemed a higher priority than quality of medical facilities.

Traffic Situation Display. The Traffic Situation Display (TSD) is a 3D map displaying company aircraft (see Figure 1B). Information such as flight plans, trajectories, weather, terrain, and airports can be displayed.

Building in Transparency. The ACFP was augmented to display ATIS at the recommended divert airport as well as indicate which of a number of risk factors are present in any potential divert location (see Figure 4) [6,7].

Building in Bi-directional Communication. Operators could also request such ratings for airports that are not recommended by the ACFP.

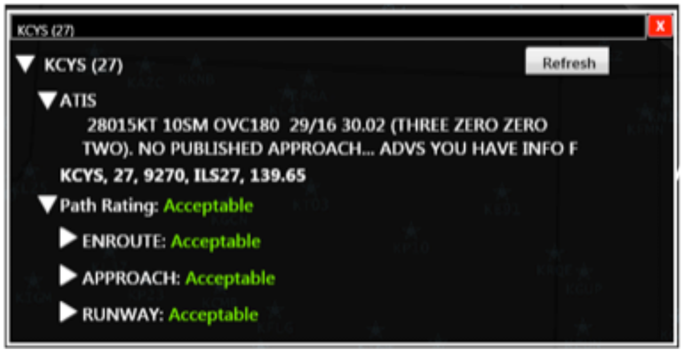


ACFP Recommendations

Option:	KCYS 27	KABQ 08	KABQ 03	KDEN 35L
Risk:	GOOD (0.99)	GOOD (0.99)	GOOD (0.99)	GOOD (0.98)
Fuel:	1184lbs	3654lbs	4025lbs	895lbs
ETA:	35.21	69.37	76.53	30.19
Dist:	134 NM	305 NM	334 NM	113 NM
Serv:	NASA FACILITIES	NASA FACILITIES	NASA FACILITIES	NASA HUB
Medical:	TRAUMA 1M	TRAUMA 3M	TRAUMA 3M	TRAUMA 10M

Figure 3. HAT Weights (top) and ACFP Factors (bottom). In the HAT condition, operators were provided ACFP factors (on bottom) and weights (on top) to increase transparency and bi-directional communication. This is an example from a Medical Emergency play.

Additional Displays. The left-side display containing aircraft flight controls and instrumentation for the selected aircraft (Figure 1C), and the right-side display containing a CONUS map with an overlay of company aircraft and weather and airport charts are provided below the map (Figure 1D), were unchanged from previous simulations.[10]



▼ KCYS (27)

▼ATIS

28015KT 10SM OVC180 29/16 30.02 (THREE ZERO ZERO TWO). NO PUBLISHED APPROACH... ADVS YOU HAVE INFO F

KCYS, 27, 9270, ILS27, 139.65

▼Path Rating: Acceptable

▶ ENROUTE: Acceptable

▶ APPROACH: Acceptable

▶ RUNWAY: Acceptable

Figure 4. Traffic Situation Display Transparency Window. When using the ACFP, regardless of condition, a transparency window appeared on the TSD for the recommended airport. An ATIS report, runway information, path rating and reasoning statements were also included.

1.1.3.3 Experimental design

The experimental design consisted of a single fixed factor, HAT, and a random factor, Participant. There were two levels of HAT: HAT (ground station with HAT features enabled) and No HAT (ground station without HAT features). Participants performed the flight-following task once in each condition with the order of trials counterbalanced across participants. We collected behavioral and subjective data.

One participant was tested per day. Each participant received approximately 3.5 hours of training before running two 50-minute experimental scenarios. Participants were provided with a concept of operation where automation and ground personnel provide “another set of eyes” monitoring aircraft. The role of ground in this flight following task was to support aircraft in high workload and off-nominal situations. Our primary interest was in participant feedback of the HAT features.

The operators had 30 aircraft to flight-follow from takeoff until landing. Shortly into the scenario, the operator began being notified of various events. An event started either with an alert on the ground station, elevating the aircraft priority and queuing the operator to contact the pilot, or with a radio call from the pilot. In either case, if it was determined that the aircraft needed to divert to a new destination, the operator would invoke the ACFP by selecting the appropriate play. In the HAT condition, once the play was selected, a checklist of procedures appeared indicating that the automation was responsible for a certain set of identified tasks. In the No HAT condition, operators had a paper checklist available for procedure items. In both conditions, the ACFP provided multiple recommendations in rank order and the transparency window was displayed on the TSD for the selected airport (see Fig. 4). The ACFP would select the highest rated airport based on the event and related factors, though the operator could explore additional airports and view both the suggested route and transparency window. In the HAT condition, the ACFP factors and weights were displayed on the ACL providing additional transparency and allowing for manipulation of factor weights. The operator would discuss options with the pilot and, when there was consensus on the new airport, the operator would datalink the route to the pilot who would, according to the concept of operations, contact air traffic control for approval. Operators could then determine to what extent they needed to follow that aircraft.

1.1.4 Results

1.1.4.1 Workload

Workload ratings were obtained at regular intervals throughout the scenarios. One subject did not respond to workload queries throughout the simulation; therefore, a repeated measures ANOVA was performed on workload ratings for the remaining five participants with the factors HAT condition and time in scenario. Overall, workload ratings were low. While the workload ratings in the HAT condition (M=1.65, SEM = 0.93) appeared lower overall than ratings in the No HAT condition (M=2.33, SEM = 1.24), the difference was not significant (p=.13; see Figure 5). The interaction between HAT condition and time interval was significant, however, F(5,11) = 3.42; p= 0.04. For the HAT condition, workload ratings were generally lower, and decreased slightly with time in the scenario. Workload ratings for the No HAT condition were more variable and roughly constant throughout. When asked post-simulation which displays and automation was preferred to reduce workload necessary for the task, all participants selected the HAT condition.

1.1.4.2 Eye gaze durations

Eye gaze data was obtained from two cameras mounted on the ground station monitors. Data were collected and analyzed in both conditions only for events resulting in a flight plan change based on the ACFP. The duration of the event was the time between activation of a play and uplink of a flight plan. This event time was subsequently broken down by the amount of time spent gazing at each of the five displays (Flight Controls, TSD, ACL, CONUS map, and Charts) during the event, by examining the gaze direction relative to each camera.

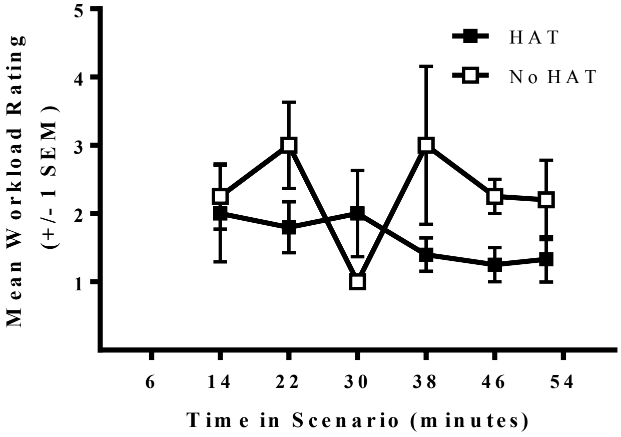


Figure 5. Workload Ratings for HAT and No HAT Conditions at Regular Intervals in the Scenarios.

Table 1 summarizes differences between conditions in terms of event and eye-gaze durations. From this table, it can be seen that participants did not always uplink a flight plan change for each event. On average only three flight-plan changes were made per scenario under each condition. However, in the HAT condition, the amount of time to select and uplink a flight change was significantly longer (40 s longer), compared with the No HAT condition. The increase in event times was due to participants gazing significantly longer at the ACFP (where the recommendations and weights were displayed) and the TSD (see Table 1) in the HAT condition. Some of the increased time must have been caused by slider movement; however, when restricting gaze times to events in which no slider movement occurred the difference between the mean gaze time for HAT was still higher (and marginally significant) compared with the No HAT condition. Very little time was spent gazing at the remaining displays in either condition.

Table 1. Summary of flight plan changes in response to off-nominal events and time spend on ground station displays for HAT and No HAT conditions

Summary Statistic	**HAT**	**No HAT**	**p***
Mean (SEM) number flight plan changes	3.3 (0.8)	3.0 (0.6)	.78
Mean (SEM) time per flight plan change (s)	101.2 (19.4)	63.7 (14.9)	.026
Mean (SEM) gaze time on ACFP (s)	53.6 (9.5)	27.5 (3.5)	.009
Mean (SEM) gaze time on ACFP - sliders not moved in HAT condition (s)	41.9 (8.7)	27.5 (3.5)	.08
Mean (SEM) gaze time on TSD (s)	38 (14.0)	20.7 (11.0)	.10
Mean (SEM) gaze time on other displays (s)	11.9 (13.0)	9.05 (12.3)	.28

*probability of obtained difference based on repeated measures t test (*df* = 5)

1.1.4.3 Weight adjustments

For the HAT condition, we counted the number of times the participant adjusted the weight factors prior to uploading a flight plan. As shown in Table 2, these weights were rarely adjusted, and these adjustments were restricted mostly to one off-nominal event.

Table 2. Number of participants who adjusted factor weights by event in the HAT Condition.

Event	N Participants	Factor(s) Adjusted
Fire in Lavatory	5	Distance
Airport Weather	2	Distance, ETA
Wheel Well Fire	1	ETA
Medical Emergency	1	Distance and ETA

1.1.4.4 Post scenario comparisons

A post-simulation questionnaire was administered after both trials were completed. Participants were asked to rate their preferred displays and automation on a 1 = No HAT to 9 = HAT scale. Participants unanimously preferred the HAT displays. Specifically, HAT displays were preferred with regard to:

- keeping up with operationally important issues (*M* = 8.67, *SD* = 0.52);
- ensuring the necessary situation awareness for the task (*M* = 8.67, *SD* = 0.52);
- integrating information from a variety of sources (*M* = 8.67, *SD* = 0.52); and
- efficiency (*M* = 8.33, *SD* = 0.82).

Participants were in agreement that overall they preferred interacting with the automation in the HAT condition (*M* = 8.50, *SD* = 0.55).

HAT Features. Items specific to the HAT tenets were rated both post-scenario and post-simulation.

Bi-directional Communication. Participants agreed that the ACFP weights improved the automation's ability to handle unusual situations (*M* = 7.83, *SD* = 1.60) and were useful in making divert decisions (*M* = 8.33, *SD* = 0.82). Participants liked having the weights (*M* = 8.33, *SD* = 1.21) and one participant commented that, *"[the display] gave me the ability to see why, gave me control to change weights in variable(s)."*

Transparency. Participants agreed that the ACFP table was helpful in making divert decisions (*M* = 7.67, *SD* = 1.51) and they liked having the table (*M* = 8.33, *SD* = 1.03). One participant commented that, "*This [table] is wonderful… You would not find a dispatcher who would just be comfortable with making a decision without knowing why*."

Operator Directed Interface with Plays. Participants liked having the electronic checklist for each play (*M* = 8.67, *SD* = 0.52) and prefered the electronic to the paper checklist (*M* = 6.33, *SD* = 1.97). One participant claimed that, *"The electronic list was easier because it was right there on the screen and it eliminated a couple of the steps."* Another participant was hesitant to rely solely on the electronic checklist and explained that s/he, *"found it necessary to have both on hand."*

1.1.5 Summary of any NATO communications/collaborations/interactions

	Planning/Design	***Execution***	***Analysis***
Communication	*X*	*X*	*X*
Coordination	*X*		
Collaboration			

1.1.6 Future research needs & plans in this area

We examined subjective and behavioral indicators of HAT collaborations using a proof-of-concept demonstration of HAT tenets: bi-directional communication between the human and automation, transparency for automation, and an operator directed interface. Because this was primarily a demonstration, the low number of participants resulted in limited statistical power. Nevertheless, the data collected that suggest potential advantages and disadvantages of HAT.

Overall, participant feedback was positive, supporting our implementation of HAT features. Participants liked having a recommender system with factors and weights and expressed interest in having similar automation for real flights. They valued the integration of the displays, commenting that this level of integration is not currently available. Participants found the electronic checklist useful as were a number of the alerts.

Operator workload appears to be lower in the HAT condition and decreased with time in the scenario. Seeing that participants completed only one scenario with the HAT agent, this may point to the need for additional training. In the No HAT condition, workload was higher and roughly constant throughout.

Interestingly, while workload was rated lower, participants took more time to uplink flight-plan changes in the HAT condition, even when no weighting adjustments were made. Typically longer task times indicate higher workload. However, given that workload overall was very low, and that very few flight plan changes were made in each scenario, it is possible that the increases in event times were insufficient to change subjective workload. If this is the case, we speculate that in the HAT condition, providing an opportunity to alter the recommendations of the HAT agent led participants to consider the initial recommendations more carefully before either accepting or adjusting them. In the No-HAT condition, if participants rejected recommendations of the ACFP, they would have to create a flight deviation with no automation assistance, and this may have led participants to accept recommendations with less prior evaluation. Of course, these speculations must be verified with further research.

These findings point out the difficulty in establishing behavioral and performance metrics for evaluating the effectiveness of HAT collaborations. On the one hand, longer event times might indicate higher workload and less efficient operations. Alternatively, longer event times might indicate more thorough processing, indicating that the HAT agent is improving operator and system performance. To understand the effect of HAT agents on operator performance and the quality of HAT collaboration, requires methods and measures be carefully developed and tested. We offer the following recommendations in this regard:

Evaluations of HAT must be performed at all levels of the system, and use a battery of assessment techniques. Evaluations must include system, autonomous-agent, and operator performance as well as behavioral precursors to operator performance [12]. The latter measures include workload, situation awareness and trust in automation. HAT evaluations should also utilize subjective, behavioral and performance measures of effectiveness. Subjective measures are simplest; they can be administered at different times during a scenario and capture the operator's perspective on the construct of interest. Note that, in our preliminary evaluation of the HAT agent for RCO, weight adjustments were made primarily by only one of our six participants. Validated subjective measures of HAT effectiveness are currently unavailable; however, validated subjective measures of human-human collaborations do exist [13], and they represent a reasonable starting point in this regard.

Finally, performance measures of HAT effectiveness will necessitate carefully designed tasks and scenarios that are sensitive to differences brought about by a HAT agent.

We continue to work on the implementation of our HAT tenets. In particular, in our current implementation, plays were flat, including only simple checklists. We are working towards making these more flexible with varying levels of automation, branch points, and opportunities for bi-directional communication. Our goal is to develop a framework for HAT, consisting of tenets and guidelines for implementing them. We eventually hope to create software libraries that make this implementation easier.

1.1.7 Acknowledgements

We would like to acknowledge NASA's Safe and Autonomous Systems Operations Project, which funded this research.

1.1.8 References

[1] Brandt, S., Lachter, J.B., Russel, R., Shively, R.J.: A Human-Autonomy Teaming Approach for a Flight-Following Task. In: Baldwin, C. (Eds). Advances in Neuroergonomics and Cognitive Engieneering. AHFE 2017. Advances in Intelligent Systems and Computing. 586, 12-22 (2017)

[2] Strybel, T., Keeler, J., Mattoon, N., Alvarez, A., Barakezyan, V., Barraza, E., Park, J., Vu, K.-P., Battiste, V.: Measuring the Effectiveness of Human Autonomy Teaming. In: Baldwin, C. (Eds). Advances in Neuroergonomics and Cognitive Engieneering. AHFE 2017. Advances in Intelligent Systems and Computing. 586, 23-33 (2017)

[3] Endsley, M.R.: From Here to Autonomy: Lessons Learned from Human–Automation Research. Human Factors. 59, 5-27 (2016)

[4] Meuleau, N., Plaunt, C., Smith, D.E., Smith T.B.: An Emergency Landing Planner for Damaged Aircraft. In: Proceedings of the Twenty-First Innovative Application of Artificial Intelligence Conference. 114-121 (2009)

[5] Lyons, J.B., Koltai, K.S., Ho, N.T., Johnson, W.W., Smith, D.E., Shively, R.J.: Engineering Trust in Complex Automated Systems. Ergonomics in Design. 24, 13-17 (2016)

[6] Lyons, J.B., Saddler, G.G., Koltai, K., Battiste, H., Ho, N.T., Hoffmann, L.C., Smith, D., Johnson, W., Shively, R.: Shaping Trust through Transparent Design: Theoretical and Experimental Guidelines. Advances in Human Factors in Robotics and Unmanned System. 499, 127-136 (2017)

[7] Sadler, G., Battiste, H., Ho, N., Hoffmann, L., Johnson, W., Shively, R., Lyons, J., Smith, D.: Effects of Transparency on Pilot Trust and Agreement in the Autonomous Constrained Flight Planner. In: Digital Avionics Systems Conference (DASC) IEEE/AIAA 35th. 1-9 (2016)

[8] Lachter, J., Brandt, S.L., Battiste, V., Ligda, S.V., Matessa, M., Johnson, W.W.: Toward Single Pilot Operations: Developing a Ground Station. In: Proceedings of the International Conference on Human-Computer Interaction in Aerospace, Santa Clara, CA (2014)

[9] Brandt, S.L., Lachter, J., Battiste, V., Johnson, W.W.: Pilot Situation Awareness and its Implications for Single Pilot Operations: Analysis of a Human-in-the-Loop Study. Procedia Manufacturing. 3, 3017-3024 (2015)

[10] Lachter, J., Brandt, S.L., Battiste, V., Matessa, M., Johnson, W.W.: Enhancing Ground Support: Lessons from Work on Reduced Crew Operations. Cognition, Technology & Work, 19(2-3), 279-288 (2017) doi: 10.1007/s10111-017-0422-6

[11] Miller, C.A., Parasuraman, R.: Designing for Flexible Interaction Between Humans and Automation: Delegation Interfaces for Supervisory Control. Human Factors. 49, 57-75 (2007)

[12] Cummings M.L., Stimpson, A., Clamann, M.: Functional Requirements for Onboard Intelligent Automation in Single Pilot Operations. AIAA. 1652, (2016)

[13] Dickinson T.L. & McIntyre R. M.: A Conceptual Framework for Teamwork Measurement. In: Team Performance Assessment and Measurement: Theory, Methods, and Applications. Lawrence Erlbaum, 19–43 (1997)