TRANSPARENCY AND CONFLICT RESOLUTION AUTOMATION RELIABILITY IN AIR TRFFIC CONTROL

Fitri Trapsilawati
Nanyang Technological University
Singapore
Christopher Wickens
Alion Science and Technology
Boulder, Colorado
Chun-Hsien Chen
Nanyang Technological University
Singapore
Xingda Qu
Shenzhen University
Shenzhen, China

This paper investigates the automation reliability and the transparency in automation conflict resolution advisories for air traffic control. Four general effects: those of traffic load, those of expertise and, those of imperfect automation and its mitigation by automation transparency in the context of the lumberjack analogy were examined. The results showed that the two automation functions, the conflict resolution advisor (CRA) and the vertical situation display (VSD) offer benefits for both novice and professional controllers's performance and increased situation awareness across traffic loads, even when the former is of imperfect reliability.

Automation Conflict Avoidance Aids

The next generation airspace procedures will be coupled with a wealth of new technology and automation tools, in order to accommodate the anticipated 2-fold growth in traffic density (IATA, 2016). One such automation tool of particular interest to our research is the air traffic control conflict resolution aid (CRA), and it is its evaluation that we report here. In the following, we briefly examine conflict avoidance automation tools for ground (ATC) operations. We then describe some of the general principles of human interaction with imperfect automation before presenting a synopsis of three experiments that have evaluated the imperfect CRA.

Conflict avoidance operations can benefit from support for two different predictive automation tools: conflict detection, and conflict resolution aids. On the ground, the air traffic controller's conflict detection tasks are well supported by the automation conflict alert (CA) system (Wickens, Rice, et al. 2009). However the operational controller is not currently supported by the tool corresponding to the airborn Traffic Alert and Collision Avoidance System (TCAS). That is, not supported by an automation-ground based conflict *resolution* advisor, although prototypes have been evaluated (Prevot, et al. 2012; SESAR, 2016).

One feature of all such conflict avoidance automation tools, an inevitable consequence of their functionality of predicting the future which is inherently uncertain (Herdener, et al. 2016) is that they are *imperfectly reliable*, prone especially to generate false alarms (or nuisance alarms), and more so at longer look-ahead times (Dixon & Wickens, 2007). However the TCAS false alarm rate appears generally to be low enough (reliability high enough) so as to still offer considerable benefits; and the same has been found for the CA for controllers (Wickens et al., 2009). In these cases the automation error rate is at a level below a threshold of around 25%, above which assistance is no longer proffered (Wickens & Dixon, 2007).

However it remains uncertain the extent to which an imperfect CRA will offer assistance relative to unaided conflict resolution, because such empirical research does not appear to have been conducted outside our laboratory. Instead the general R&D evaluations have implemented automated resolution advisories that will always increase separation, relative to the trajectory of the uncorrected aircraft (i.e.;, 100% relliablity). Yet because of the extreme complexity and density of the future airspace, it is likely that some such "automation errors" could occur. The pilot, receiving advice from the CRA-assisted controller may receive three categories of such errors: advice to maneuver in a manner that clearly decreases the anticipated minimum separation, advice to maneuver in a different direction or axis than one preferred by the pilot from the standpoint of energy management, fuel consumption or passenger comfort, and advice that avoids an immediate conflict, but now places the aircraft on a trajectory toward a new one. Our implementation of the imperfect CRA employs the third of these categories.

The Lumberjack Analogy

We can place the imperfect conflict avoidance aids (both detection and resolution) within the context of the automation stages & levels taxonomy initially applied to air traffic control automation by Parasuraman, Sheridan and Wickens (2000), and subsequently supported by strong empirical evidence from a meta-analysis carried out be Onnasch Wickens, et al., (2014), who coined the term "degree of automation" (DOA) defining automation that did "more cognitive work" relative to the human operator who is supported by that automation. While the full taxonomy is more complex than space allows here (See Sebok & Wickens, 2016), within the current conflict avoidance framework, we specify two degrees of automation. At a lower degree is conflict detection, a diagnostic or **situation assessment aid** that advises "what is". At a higher degree is conflict resolution, a **decision aid** that advises "what to do". This increase in DOA from SA support to decision support was predicted (Parasuraman, Sheridan, & Wickens, 2000), found (Onnasch et al., 2014) and modeled (Sebok & Wickens, 2016) to (a) *improve* nominal performance and reduce workload when automation functioned correctly, but (b) lead to *greater* problematic, and sometimes catastrophic consequences on the infrequent occasions when automation failed (or failed to operate as expected by the human supervisor). In terms of the lumberjack analogy: "the higher the tree, the harder it falls".

One of the key features revealed by the meta-analysis carried out by Onnasch et al (2014) is that the greater problematic response to automation failures, with the higher degree of imperfect automation, was paralleled by a loss of situation awareness in those circumstances. This loss is triggered by being more "out of the loop" in decision automation which enables automation to select or advise actions, compared to SA-support automation that still forces the operator to actively choose actions. Such active choices better implant the state of the system in the operators's memory, i.e., an increaese of situationn, via a phenomenon known in memory theory as the "generation effect" (Slamecka & Graf, 1989; Hopkin, 1995).

The final piece in our puzzle and basis of our current predictions, is that, if SA is lacking with decision support automation, it can be restored by effective automation **transparency**, or displays that provide more graphic information about the current state of the environment from which automation draws its action recommendations (Bizantz & Seeong, 2008; Mercado et al., 2016). Thus our argument in the current project is that, to the extent that controller-CRA interaction is hindered by the occurrence of occasional imperfections or automation errors (a prediction we expect to confirm), this problematic response can be mitigated by a display supporting controller situation awareness. What then should this display be? In the typical ATC console, the controller is well supported in lateral awareness by the "radar display" or plan view display (PVD). But less so in vertical (altitude) awareness because most information about altitude and relative altitude is depicted in symbolic digital data tags, a less than ideal way of conveying trend information about the relative altitude of multiple aircraft. Hence our transparency mitigation was designed to provide controllers with a *vertical situation display* (VSD), a concept receiving substantial research in the flight deck CDTI (e.g., Battiste & Johnsons, 2002; Thomas & Wickens, 2008), but less so in ATC (SESAR, 2013). In particular, to our knowledge, no research has been carried out joining the two automation concepts of the VSD and the CRA, let alone in circumstances in which the CRA is imperfect. Our program of research does this.

In the three experiments described below, we first show that the CRA can assist resolution performance, and can even do so when it is imperfect, relative to fully manual performance. We do this with modest traffic load (experiment 1; Trapsilawati et al., 2015) and then with much higher traffic load (experiment 2; Trapsilawati, et al., 2016) evaluating the greater dependence on the CRA in the latter conditions. Because both of these experiments are published, we only describe them briefly here. Then in experiment 3, we evaluate the possible mitigation effectiveness of transparency provided for some participants by the VSD to support the human response to automation failures, within the framework of the lumberjack analogy. Because we do not examine conflict detection aids here, our tree is always high (decision aiding); we document its fall, but also show that we can lessen the impact of the fall with the VSD.

Methods

All three experiments involved the same general simulation and methods, described in some detail in Trapsilawati et al., 2015, 2016, and only briefly here. Participants, either students within the Aeronautical and Aerospace programs, or professional controllers viewed the TRACON display in an NLR ATC simulator (NARSIM). They were responsible for moving traffic through the sector, and avoiding loss of separation (conflict avoidance). During a typical 1 hour session, 5 conflicts would be imposed, at unpredictable times, leading to an LOS if not control action was taken. This action was implemented by a voice input (e.g., "change heading to 030") and carried out by a pseudopilot, where the changed trajectory would be then visible on the display.

The four experimental sessions differed from each other in terms of the automation support offered by a CRA. In this regard, the CRA was either absent (manual performance only) present and fully reliable, or present and "imperfect" such that one of the advisories directed an aircraft to change trajectory and avoid an immediate conflict, but in the process, created a predicted conflict with a second aircraft. The latter predicted conflict did NOT trigger advice from the CRA. As such erroneous advice occurred in one trial out of 5 in the imperfect automation block, the overall CRA reliability could be said to be 80%; although prior to the first time a failure was observed in the imperfect session block, the controller would experience it as having 100% reliability, since no failures were imposed during the training blocks. This first failure will be particularly relevant to our evaluation of support for the lumberjack analogy. Participants were free to comply with or ignore the advice of the CRA if they felt that an alternative maneuver was preferable. During each session, participants were periodically probed with a SPAM situation awareness question regarding the current status of the airspace (Durso & Dattel, 2004). The latency to respond to the ready probe assessed overall workload (OWL), and the accuracy measure of the probe response assessed SA.

In all experiments, a generic TRACON space was employed. In Experiment 1, employing 12 controllers who were primarily students, traffic density was 30 aircraft per hour In Experiment 2, employing 24 participants, again primarily students, traffic density was increased to either 60 or 90 (between groups) to simulate the projected growth of airspace congestion that would benefit more from automation assistance. In Experiment 3, employing exclusively professional controllers, in which the VSD was imposed, traffic density was set at a constant level of 60 aircraft. In the following we refer to students as "novices" and to professional controllers as "experts".

Results

Figure 1 shows, on the X axis all three experiments juxtaposed, with the three automation conditions along the X axis defining the shape of each line. The relative scale of each of these three dependent variables (performance, top; situation awareness, middle; OWL, bottom), is arbitrary as each has been transformed so that they show minimum overlap within the figure. The important factors are the shape of the profiles of each 3-point line, and the relative position of the three profiles across experiments. These relative positions are connected by dashed lines. The following general observations can be made:

Differences, due to Traffic Load, between Experiments 1 and 2

Experiment 2, with its higher traffic load shows an overall reduction of performance compared to Experiment 1 (*p*= 0.02). However, the reduction of SA (*p*= 0.12) and the increased workload (*p*= 0.63) were not significant. Experiment 2, with greater traffic load shows OWL to be greater in the manual condition than with automation. Stated in other terms, in Experiment 2, with its higher traffic load, in fact, there is a greater benefit of CRA automation to reducing workload, whether the CRA was reliable or not, and the CRA automation in Experiment 2 actually **restores** workload to a level equivalent to that of the lower traffic load in experiment 1, as indicated by the significant interaction between the automation condition and experiment/traffic load (*p*= 0.04).

Differences in Profiles between Experiments 1&2 (Novices) and 3 (Experts)

To allow for direct comparison between novices and experts, we did the analysis between Experiment 2 with medium traffic condition (novice participants) vs Experiment 3 without the VSD condition (expert participants) where the air traffic loads were similar. We found that overall performance was not significantly different between novices and experts (*p*= 0.36). However, the interaction effect was significant (*p*= 0.03), showing much better performance of experts than novices in the manual condition. Novices' overall SA was marginally higher than that of the experts (*p*= 0.08). However no difference of SA was found across automation conditions for either novices or experts (*p*= 0.20). The experts' workload is considerably lower than novices although the trend was not significant (*p*= 0.14).

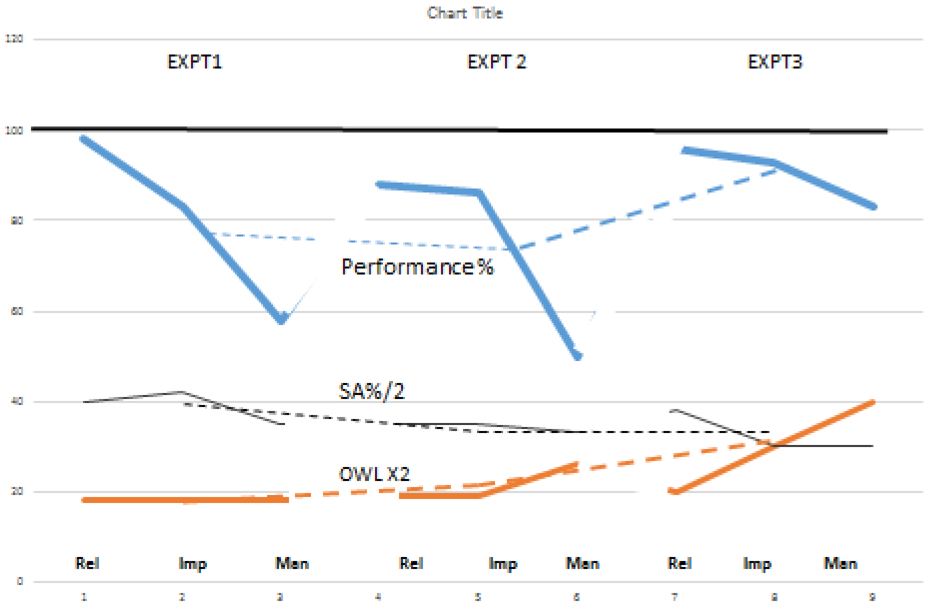

Figure 1 Results of the Three Experiments. The three color coded lines of different width within each experiment define each of the three critical independent variables; Performance (% resolved conflicts) at the top, Situation awareness (% correct) in the middle, and Objective workload (OWL: ready response latency) on the bottom**.**

The Examination of Lumberjack Analogy based on Data of Experiment 3

The presence of the VSD slightly improved conflict resolution performance (from 89% to 94%, F (1, 18) = 1.35, *p*= 0.26), substantially increased situation awareness (from 59% to 73%; F (1, 16) = 4.13, *p*= 0.059) and significantly lowered workload (from 7.78s to 5.38s; F (1, 14) = 8.57, *p*= 0.01). The VSD was found to have equivalent effects across all three automation conditions (i.e., no interaction between automation condition and display).

On the first failure trial, for the block in which the CRA was unreliable; performance accuracy was compared with all correct trials, in which automation was functioning correctly. Here the accuracy for the four combinations of automation accuracy and VSD support is shown in Table 1. Examining these data, we observed what could be interpreted as a significant interaction effect, in that a test of proportions revealed a substantial significant decrement of the 25% reduction without the VSD (Z= 2.36, *p*= 0.02), but no significant effect (Z= 1.08, *p*= 0.28) of the 7.5% decrement when the controller was supported by the VSD.

Table 1.
The First Failure Analysis.

Display Conditions	Automation Correct Trials	Automation Failure Trial
VSD	97.50%	90.00%
No VSD	95.00%	70.00%

Discussion

Overall the results allowed us to examine three general effects: those of traffic load, those of expertise and, most critically, those of imperfect automation in the context of the lumberjack analogy.

Workload/Traffic Load

In comparing experiment 1 (low density=30) with Experiment 2 (high density= 60 or 90), both using primarily novice controllers, we found that increasing density produced a decrease in performance, a trend of loss in situation awareness and only a very slight increase in objective workload. We might argue that, when these novice controllers confronted the high traffic density, their performance went over the "red line" of workload (Grier, 2008), which could not be rated higher (they were "maxed out"; and hence could give no more resources), even as the gap between resources demanded and those supplied increased, hence lowering performance. At maximum capacity in Experiment 2, the novices also diminished any resources available for maintaining SA. Hence there was a trend of SA degradation.

Controller Expertise

In comparing the overall results of Experiment 2 with those of Experiment 3, both at medium-high workload/traffic load levels, the most obvious difference is the increase in performance of the experts (Experiment 3), particularly when controlling manually (without CRA automation assistance). This is not surprising. Experts generally are better performers. This increase was attained with no change in workload, but with a marginal loss in SA, an effect that is somewhat surprising.

The Lumberjack Analogy

To examine the lumberjack analogy, we focus attention at greater depth only on the performance of the experts in Experiment 3, as this performance is most generalizable to the real world of air traffic control and only hee can we examine the mitigating effoects of the VSD. Here we find, as with the first two experiments, a benefit of automation, although this benefit was reduced, given the higher baseline level of manual performance of the experts in Experiment 3. Somewhat unexpectedly, we also found an increase in situation awareness with automation, contrary to the standard "folk lore" of automation (Sebok & Wickens, 2016) in which automation is assumed to produce an out-of-the-loop unfamiliarity syndrome, breeding complacency, dependency and "the automation bias" (Mosier & Skitka, 1996), and mediated by a **loss,** not a gain of situation awareness. In accounting for this departure from our expectations, we assume that, unlike some other cases, our expert professional controllers invested any resources saved by the CRA decision aid, into deeper processing the raw data from the display.

Insofar as the lumberjack analogy itself is concerned, we have partial support for its expression. On the one hand, experts did not perform significantly worse overall with imperfect (80% reliable), than with perfect automation blocks, even though there was a non-significant trend in that direction. On the other hand, on the single (and first) failure trial, they did perform worse, with a detection rate, when unsupported by the VSD that dropped from 95% (on the correct trials) to 70%. We also found that this problematic failure cost was mitigated by the automation transparency provided by the VSD relative to the control group. The former showed only a small (7.5%) non-significant loss of performance on the failure trial, while the latter showed a large loss of 25%. Finally, we ask if this failure recovery difference between the two groups was mediated by a difference in situation awareness. Here the interpretation is again ambiguous. On the one hand the VSD **did** substantially improve SA. But on the other hand, such an improvement was equally manifest on both manual trials and on those supported by perfect automaton. Hence we cannot infer that the differential performance improvement was associated with a differential increase in SA.

The ambivalence of theoretical interpretation notwithstanding, we can conclude with certainty that the two aspects of technology examined here, the automation of the CRA, and the SA support of the VSD are both of benefits to professional controllers, even when the former is of imperfect reliability.

Acknowledgements

This work was supported by grant No. [ATMRI:2014-R5-CHEN] from Civil Aviation Authority of Singapore (CAAS) and Air Traffic Management Research Institute (ATMRI), Nanyang Technological University (NTU), Singapore. We would like to thank Mr. Erik-Jan Hartlieb from Netherland Aerospace Lab for the technical assistance. The research presented in this paper was based on Fitri Trapsilawati's doctoral dissertation at the NTU, Singapore. Any opinions, findings, conclussions, recommendations expressed in this paper are those of authors and do not reflect the views of the employers or granting organization.

References

Battiste, V., & Johnson, N. H. (2002). An operation evaluation of ADS-B and CDTI during airport surface and final approach operations. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 36-40). Baltimore, MA: Human Factors and Ergonomics Society.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors, 49,* 564-572.

Durso, Frank T, Dattel, Andrew R, Banbury, S, & Tremblay, Sebastien. (2004). SPAM: The real-time assessment of SA. In S. Banburry & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (Vol. 1, pp. 137-154). Vermont: Ashgate Publishing.

Grier, R. (2008). The redline of workload: Theory, research and design. A panel. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* (pp. 54-58). Santa Monica, CA: Human Society.

Herdener, N., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2016). Overconfidence in projecting uncertain spatial trajectories. *Human Factors*, *58*, 899-914.

Hopkin, V.D., (1995) *The Human Factors of Air Traffic Control*. London: Taylor & Francis.

IATA. (2016). IATA Forecasts Passenger Demand to Double Over 20 Years: International Air Transport Association. Retrieved from http://www.iata.org/pressroom/pr/Pages/2016-10-18-02.aspx

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors, 58*, 401-415.

Mosier, K. L, & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other. *Automation and human performance: Theory and applications*, 201-220.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation An Integrated Meta-Analysis. *Human Factors, 56*, 476-488.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types & levels of human interaction with automation. *IEEE Transactions on systems, man, & cybernetics-Part A: Systems and Humans*, *30*, 286-297.

Prevot, T., Homola, J. R., Martin, L. H., Mercer, J. S., & Cabrall, C. D. (2012). Toward Automated Air Traffic Control—Investigating a Fundamental Paradigm Shift in Human/Systems Interaction. *International Journal of Human-Computer Interaction, 28,* 77-98.

Sebok, A., & Wickens, C. D. (2016). Implementing Lumberjacks and Black Swans Into Model-Based Tools to Support Human–Automation Interaction. *Human Factors*, 59, doi: 10.1177/0018720816665201.

SESAR. 2013. Final Project Report on the concept and benefits for improving TP using AOC data: Improved Airline Flight Plan Information into ATC Trajectory Prediction (TP) Tool. SESARJU Report. Retrieved from http://www.sesarju.eu/sites/default/files/ solutions/3_AOC_Data_for_TP_Final_Project_Report.pdf?issuusl=ignore.

SESAR. (2016). Automated support for conflict detection, resolution support information and conformance. ESSIP Plan Edition 2016. Retrieved from https://www.eurocontrol.in t/sites/default/files/content/documents/official-documents/reports/atc12-1.pdf

Slamecka, N. J., & Graf, P. (1978). The generation effect: delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, *4*, 592.

Trapsilawati, F., Qu, X, Wickens, C. D., & Chen, C.-H. (2015). Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control. *Ergonomics*, *58*, 897-908.

Trapsilawati, F., Wickens, C. D., Qu, X., & Chen, C. H. (2016). Benefits of imperfect conflict resolution advisory aids for future air traffic control. *Human factors*, *58*, 1007-1019.

Thomas L.C., & Wickens, C.D. (2006). Display dimensionality, conflict geometry, and time pressure effects on conflict detection and resolution performance using cockpit displays of traffic information. *International Journal of Aviation Psychology, 16*(3), 321-342.

Wickens, C. D, & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8,* 201-212.

Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a “cry wolf” effect? *Human Factors, 51,* 446-462.