Investigating Flight Crew Recovery Capabilities from System Failures in Highly Automated Fourth Generation Aircraft

Customer

National Aerospace Laboratory NLR

NLR-TP-2014-427 - October 2014

NLR – Dedicated to innovation in aerospace

National Aerospace Laboratory NLR
Anthony Fokkerweg 2
1059 CM Amsterdam
The Netherlands
Tel +31 (0)88 511 3113
www.nlr.nl

EXECUTIVE SUMMARY

Investigating Flight Crew Recovery Capabilities from System Failures in Highly Automated Fourth Generation Aircraft

Problem area

Modern transport category aircraft such as the Airbus A320, A330, A380, Boeing 777 and 787, employ an increasing amount of automation in the cockpit. These so-called fourth generation aircraft enable this automation for flight control tasks, flight guidance/planning but also for the management of aircraft subsystems. Because of the increasing complexity of aircraft and the environment they must operate in (e.g. traffic), these systems have evolved to work semi or even fully automated, in order to reduce pilot workload. However, studies into other contingent effects of automation show that the reduction in workload also induced a sense of complacency as it detaches the crew from regularly monitoring or operating the process being automated.

Report no.
NLR-TP-2014-427

Author(s)
J.F.W. Mohrmann
A.J.J. Lemmers
J.A.A.M. Stoop

Report classification
UNCLASSIFIED

Date
October 2014

Knowledge area(s)
Training, Missiesimulatie en Operator Performance
Vliegveiligheid (safety & security)
Vliegoperaties
Cockpit
Softwaretechnologie voor de luchtvaart

Descriptor(s)
Man4Gen
Automation
Decision Making
Situational Awareness
Control

This report is based on a paper submitted October 1, 2014 to the journal of Aviation Psychology and Applied Human Factors, by Hogrefe.

Description of work

A European research project, Manual Operations of Fourth Generation Airliners (Man4Gen), is poised at determining these emerging control aspects of a high level of automation in the cockpit, in particular the difficulties experienced in responding to unexpected situations which requiring a transition from monitoring very reliable systems to active and authoritative decision-making and exercising full control of the aircraft. The study detailed in this report runs parallel to the Man4Gen project, but is instead focused on the problems related to manual control of usually fully automated sub-systems of the aircraft, in particular the fuel system. The study has developed a conceptual framework which models the process of decision making which occurs upon the failure of automated systems. A subsequent experiment scenario makes use of the Airbus A330 fuel system, which normally operates nearly completely independent from the crew. The scenario creates a diffuse situation where the failures in automation attempt to obscure a more serious underlying failure which may severely impact flight safety. This forces various decisions to be made by the crew, and allows the observation of their problem-solving behaviour.

Results and conclusions

Subsequent comparison of crew behaviours with flight performance indicators reveals that there are clear and repeated examples of behaviours and interactions with automation which improve performance, but also those which deteriorate performance. Specific to failures of such aircraft sub-systems, in this case the fuel system, results indicate that improved technical knowledge of the automated systems and improved understanding of the (diagnostic) procedures related to them, in combination with a suitable level of trust in the automation can lead to significant performance gains. Therefore, this study concludes that, given certain circumstances, differences in flight crew interactions and competencies with automated systems can definitely have a profound influence on their operational performance.

Applicability

The effect of such crew-automation interaction is not to be underestimated in its potency to effect flight safety. This study attempts to provide a holistic approach in contributing to the understanding of human-computer interaction in the cockpit, in expectation of future advancement of such control structures.

National Aerospace Laboratory NLR
Anthony Fokkerweg 2, 1059 CM Amsterdam,
P.O. Box 90502, 1006 BM Amsterdam, The Netherlands
Telephone +31 (0)88 511 31 13, Fax +31 (0)88 511 32 10, www.nlr.nl

Investigating Flight Crew Recovery Capabilities from System Failures in Highly Automated Fourth Generation Aircraft

J.F.W. Mohrmann, A.J.J. Lemmers and J.A.A.M. Stoop[1]

[1] Kindunos Consultancy

Customer
National Aerospace Laboratory NLR
October 2014

This report is based on a paper submitted October 1, 2014 to the journal of Aviation Psychology and Applied Human Factors, by Hogrefe.

This publication has been refereed by the Advisory Committee AIR TRANSPORT.

Customer	National Aerospace Laboratory NLR
Contract number	-----
Owner	NLR + partner(s)
Division NLR	Air Transport
Distribution	Unlimited
Classification of title	Unclassified
Date	October 2014

Approved by:

Author	Reviewer	Managing department
J.F.W. Mohrmann	H. van Dijk	H.G.M. Bohnen
Date 10-10-2014	Date 17-10-2014	Date 13-10-2014

Contents

This page is intentionally left blank.

1 Introduction

Coping with growing operational demands and complexities, fourth generation commercial aircraft employ an increasing amount of automation for flight guidance, planning, system management, etc. However, according to Malinge (2011), the accident rate of fourth generation airlines has stagnated. Although automated systems have evolved to reduce pilot workload, recent concerns exist about their safety effectiveness. Automation introduces a paradox: providing crews with necessary operational assistance simultaneously dissociates the crew from those operations. Unfortunately, this shift in pilot tasking exhibits itself in many forms of adverse crew behavior such as automation induced complacency (Manzey et al., 2012), automation bias (Mosier et al., 1998), decision making errors (Orasanu et al., 1998), lack of (procedural and declarative) system knowledge and/or manual control skills (Potter et al., 2012), which are all in turn aggravated by overconfidence (Wood, 2004) and fatigue (Caldwell, 2012). These behaviors contribute to the loss of situational awareness (SA), which is defined by (Endsley, 1995). Often these effects are precursors to a deterioration of control of the aircraft, leading to Loss of Control in Flight (LOC-I). According to the European Aviation Safety Agency (EASA, 2013), LOC-I accidents are currently the leading type of fatal accidents. Notable accident cases which exhibit tell-tale signs of this problem are Colgan Air Flight 3407, Air France Flight 447, Air Transat Flight 236 and, more recently, Asiana Airlines Flight 214.

One initiative investigating the aforementioned automation concerns is the European project Manual Operations of Fourth Generation Airliners (Man4Gen). Man4Gen aims to understand the problems that crews face in the rapid transition from a monitoring role to an active decision making role. The project intends to develop and evaluate mitigating strategies in training, operations and system design. While Man4Gen will focus on crew response to automatic flight control systems failures and flight path related situations, this study complements it by investigating operations with the automation of aircraft sub-systems, in particular the fuel system. The fuel system is an interesting example of a highly automated system on a fourth generation aircraft, particularly given the overall acceptance of the high level of automation in this system. To this end, the research objective of this project is to evaluate the knowledge, awareness and problem solving capabilities of fourth generation aircraft flight crews, during unexpected failures in highly automated [fuel management] systems. The results should indicate whether the current use of automation in the cockpit environment could be contributing to the stagnation of the accident rate, and how. It should be noted that this paper aims to present an academic overview of this study. A more comprehensive explanation of the methodology and results can be found in (Mohrmann, 2013).

2 Method

2.1 Theoretical basis

In order to couple loss of control with a loss of awareness, a conceptual framework was developed based on the concepts of SA, as proposed by Endsley (1995), and sensemaking (SM), as proposed by Hollnagel and Woods (2005). For this study, SA was broadly defined as the quality of awareness that an individual has of his or her (aircraft) state, system state, environment, location, etc.. SA is divided into three successive levels:

1. Perception of reality
2. Understanding these perceptions
3. Projection this understanding into the near future

The quality of each level is for a large part dependent on the quality of a lower level. SM theory complements SA theory; in this study SM has been defined as the cyclic process by which reality is periodically reviewed in order to update ones model, and distil an interaction with reality, which completes the awareness feedback loop. This so called Contextual Control Model (COCOM) (Hollnagel and Woods, 2005) has been expanded to an Extended COCOM (ECOM) which proposes that multiple loops are stacked (Figure 1) and connected to create an array of control loops ranging from compensatory control to anticipatory control.

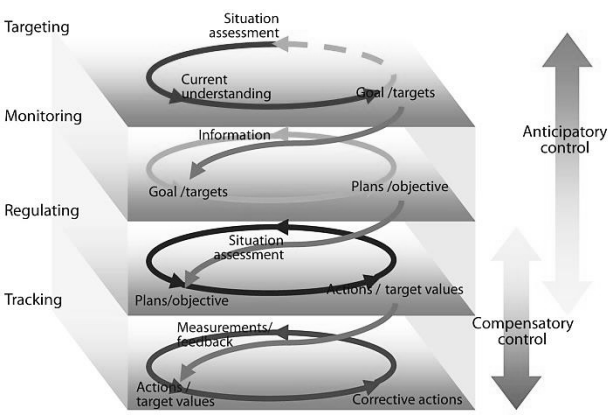

Figure 1: Extended Contextual Control Model. (Hollnagel, 2002)

The critical link between SA and SM is that compensatory action at the tracking level of the ECOM may be coupled to a non-complex SA (i.e. Perception only), whereas anticipatory control is only possible when the individual maintains excellent awareness (i.e. good Projection), as illustrated at the higher ECOM levels. It follows that a fault in perception will fragment the basis of subsequent awareness levels, and that any higher level of [anticipatory] control will not interact with reality as expected. Perceiving such a cognitive mismatch between expectations and reality is experienced as the universally recognized cognitive state of Surprise: when things do not go as planned.

The conceptual framework enabled the quality of a decision as a central concept, used as a medium to connect the level of control of the crew to their level of awareness. At this point two important assumptions were made concerning the quality of a decision:

1. *Decisions are rational choices completely based on the information that the decision maker perceives about his reality*
2. *Decisions within the context of a specific understanding, will have a positive effect on the situation, given that this understanding matches reality*

These two premises shed some light onto the relations between understanding reality, making decisions and expected results. The four statements below further built upon these two premises and formed a layout of the final logic structure applied in this paper:

1. *The level of SA is, per definition, inverse to the level of cognitive mismatch between understanding and reality*
2. *The quality of a decision is inverse to the level of cognitive mismatch*
3. *A particular level of control is dependent on the quality of decision*
4. *Therefore: The level of SA is causal to the level of control*

By this framework it was be possible to connect awareness with a level of control.

2.2 Conceptual framework

Automation awareness. In the context of subsystem automation, awareness was redubbed Automation Awareness (AA). The concept of AA related to areas of awareness specific to effectively operating the automation pertaining to a [fuel] sub-system. Good AA was defined as:

"Exhibiting sufficient knowledge, familiarity and awareness of the functioning of an automated process, in order to retain complete control of the process at any time."

Four behavioral categories were proposed as factors affecting AA: knowledge, rules, attitude and teamwork. The departure point for the definition of AA was the skills, rules and knowledge (SRK) based reasoning concept proposed by Rasmussen (1983). Rule and knowledge based reasoning were adapted from the SRK concept as two of the four behavioral categories. *Knowledge* referred to both declarative and procedural knowledge of the system(s) at hand, but also of other operational details which were relevant in diagnostic or recovery steps. *Rules* referred to a pilot's familiarity with and attitude towards procedures, checklists and other operational prescriptions and/or limitations. Skill based reasoning was omitted as such reflexive motor skills was not

expected to often come into play in resolving aircraft system automation problems such as those studied in this experiment. The third behavioral category was pilot *attitude towards automation*, based on several known effects such as automation complacency, (c)omission errors and trust. The fourth and last category was *teamwork*, in consideration of the effects of social/inter-personal aspects of awareness building, notably Team and/or Shared SA (Endsley and Jones, 1997). Especially when considering how complex diagnostics and task sharing often engage both crew members, it would be quite possible that shared awareness would be critical in effectively resolving such problems. These four discerning behavioral categories constituted AA, and served as the basis of awareness upon which pilots would make their decisions to act. They were abbreviated to the K-R-A-T decision making affectors.

Performance indicators. Upon making a decision, a crew's actions would interact with reality, and would result in some level of control, depending if the interaction with reality was productive or counter-productive to the situation. Working from Assumption #2, it was stated that a crew's decision would always be aimed at maximizing performance (the specific performance indicators will be discussed later in this section). Therefore, it was assumed that measurable performance is indicative of a level of control. Of course this was also dependent on the isolability and measurability of performance indicators, and that enough performance variability would be present. Additionally, considering the complexities of diagnostic and recovery strategies, two different crews may enable two different strategies which would result in the same level of overall performance. To mitigate such a nulling effect, performance was subdivided into smaller indicators centered about core decision making events in order to improve the fidelity of matching particular AA with particular performance.

Correction factors. The final aspect of this conceptual framework was the set of three correctors which were also considered as possible unintended affectors of the quality of decision making: workload, simulator realism and fatigue. A workload that would be too high or low could introduce effects such as excessive docility or too much stress, which in turn would affect decision making. Workload was measured by means of a RAW NASA TLX test (Hart, 2006). Simulator realism could affect crew immersion, where too little immersion would result in unrepresentative behavior for actual aircraft operations. Realism was scored on a continuous scale for five categories: simulator hardware, simulator software, scenario setting, scenario events and an overall score. Fatigue could also severely impair proper decision making due to reduced cognitive functioning. Fatigue was measured by means of a discrete seven-point Samn-Perelli scale. If any corrector correlated with performance, then the comparability of different

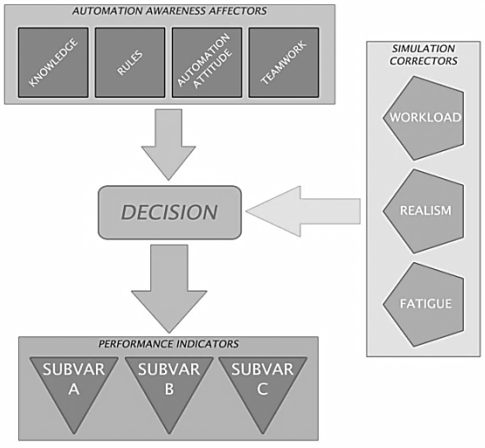

Figure 2: Conceptual Framework Model

crews may be compromised as their behavioral bases would not be the same. Such correlations would serve to correct performance ratings before their comparison to crew AA.

The three components of the conceptual framework are combined in Figure 2. This framework functioned by observing pilot behavior with respect to the four AA affectors, and correlating them to their respective performance ratings (albeit corrected).

2.3 Experiment design

This experiment diverged from a classic control and a test group setup, and focused on the observation of within-group variations of set of representative pilot participants. Where within-group variations were limited, the study could still provide a more global insight into crew behavior and performance. This study applied a simulated flight environment in order to create a situation in which crews were forced to demonstrate their decision making with respect to fuel system automation. The choice for a fuel sub-system was made considering the criticality and complexity of such a system. Criticality was important in order to build sufficient motivation in crews to thoroughly resolve the issue. Complexity was important in order to force more expansive diagnostic procedures, which in turn allowed for the more (precise) observation of decision making processes. The Airbus A330-200 fuel system ranked high on both factors, and featured the additional benefit of offering a larger group of type-rated participants to select from.

Scenario design. The A330 boasts very advanced automation in its fuel system. This made it well suited to create a very diffuse scenario with respect to the division of control between the crew and the system computers. The central objective of this scenario was to create multiple crew decision points where the outcome would be ambiguous to the crew. These decisions mostly pertained to their ability to control the fuel system. The scenario was set in a trans-Atlantic Extended Twin-engine Operational Performance Standards (ETOPS) flight from Europe to the Caribbean region, where the routing was very southerly, crossing the Azores and remaining equidistant between the African west coast and the U.S. east coast. The reason for this diversion was a tropical weather cell in the northern Caribbean region. The scenario would start about two-thirds into the cruise phase, approximately 25 minutes prior to the flight plan's critical fuel

point (beyond that point the fuel plan ensured 15 minutes of fuel upon arrival, given an emergency descent to FL100 and continuing with one engine inoperative - an ETOPS flight planning condition). Furthermore, the preflight briefing mentioned two components which would be unusable during the flight (legal and according to the Master Minimum Equipment List). One was the Fuel Control and Management Computer (FCMC) No. 2 (of 2), which was important in relation to the FCMC 1+2 FAULT (loss of automatic fuel management). The other unusable component was a hydraulic accumulator fault, which had no further consequences and only served to distract the crew from the fuel system. During the scenario several failures and events steered crew expectations by priming their mind-set, which set the stage for misleading, confusing or surprising the crew. In essence, the scenario consisted of a fuel leak which was difficult to accurately diagnose, given the A330's vast array of fuel tanks and piping. This failure was cloaked by faulted (fuel management) automation, which forced the crew to investigate manually/mentally. As the leak was central to the scenario, the leak rate was finely adjusted in order to provide the crew with sufficient time/fuel to reach several diversion fields, but to also be large enough in order to be noticed and evoke a sense of urgency. The leak rate of 1.5 kg/s was tuned to finely balance realism and challenge required to fully engage participants in their decision making process. The scenario was developed and validated together with operational experts on the A330 – both flight crew and system experts. All sessions were identical up to a particular *Event 7*, after which crews were able to diverge in their recovery attempts. Therefore, performance is only measured after Event 7. Behavior was monitored throughout, as behavior prior to Event 7 could also have had an impact on their course of action after Event 7. A detailed account of the scenario schema can be found in (Mohrmann, 2013).

Resources & Planning. The experiments were hosted at the National Aerospace Laboratory (NLR), in the Generic Research Aircraft Cockpit Environment (GRACE) simulator. The generic cockpit could be modified to accurately represent several common cockpit environments, including the A330. The simulator also featured functionally sufficient automation and FMS features, and radio contact with (ATCo) in the simulator control room. The observer was seated inside the cockpit aft of the two pilots, and also played the occasional role of purser/cabin crew. The participants were 20 active duty A330 pilots (five captains, 11 first officers and four second officers), forming 10 valid cruise crew combinations (no two second officers, or two captains). Confederate air traffic controller roles were fulfilled by other A330 pilots familiar with mid-Atlantic radio communication, although most of the ATC responses were anticipated and scripted ahead of time.

Participants were invited for an "observation study", and not informed about the true nature of the scenario. The experiments took place within one week to limit possible cross-contamination

of subject experience. Simulations were held during normal working hours to mitigate possible fatigue effects. Upon arrival, participants were familiarized with the simulator after which they were permitted to self-brief with a representative yet modified dispatch briefing document, and were informed they would start around two-thirds into the cruise phase. The experiment scenario duration varied between 50 and 75 minutes depending on the crew's decision making speed. Each scenario ended when the crews practically exhausted their diagnostics, and communicated a diversion plan.

Observations. Several different modes of information were recorded. The simulator recorded important flight and fuel system parameters (e.g. location, fuel tank quantities, failures, etc.). The simulator also facilitated audio and video recording of the pilots and displays. The observer recorded behavior, decisions and other notable occurrences during the flight. Participants filled in a post-hoc self-reflection form including measurements for corrector variables and a form asking about a pilot's attitude towards automation, with six questions pertaining to the respondent's feelings of trust, dependence, familiarity, etc. with respect to on-board automation in general. Lastly, the observer engaged both crew members together in a cognitive interview session, during which several important or unclear events were discussed, such that pilot rationale could better supplement observed behaviors.

3 Results

3.1 Performance indicators

Essentially, the most fundamental indication of good performance could be represented by the safety margin created by pilots. Given the nature of the failures, maximizing fuel on arrival (FOA) would be a heuristic for maximizing the safety margin during this emergency. Table 1 shows the range of FOA.

Table 1: Fuel on Arrival values ranked from best to worst

Best performance				**Fuel on Arrival (kg)**				**Worst performance**	
4470.5	4020.2	3886.5	3688.1	3542.0	2799.1	2520.5	512.0	409.6	-500.3*

However, a single fundamental indicator was not sufficient in comparing two different crew strategies leading to the same FOA. Therefore this study introduced three sub-performance variables (subvars): amount of fuel leaked, fuel burn efficiency (kg/NM) and the distance flown. The combination of these variables fully represented all the possible ways to reduce the fuel quantity on-board (FOB), which was directly causative to the safety margin achieved. At this point it is important to note that the scenario differed slightly between crews, due to certain events being triggered by (variable) crew actions. All events up to Event 7 were effectively priming events and did not warrant any differences in crew diversion actions. Therefore all performance (sub-)variables were only evaluated from Event 7 onwards. Additionally, because the crews did not fly the aircraft until landing, their intentions for the remaining 1-1.5 hours (clearly verbalized before the end of each session) were used to extrapolate performance according to Airbus A330 performance specifications.

Subvar A: Amount of fuel leaked. This variable indicated how much fuel was lost through the leak after Event 7 until landing. For many crews they (nearly) exhausted all fuel which was available to the leak (which is approximately half of all FOB). The fuel leaked was calculated by finding the difference between the total reduction of fuel (between Event 7 and landing), and the amount of fuel burned (i.e. fuel used: FU). This subvar was also corrected for the fact that not all fuel can be leaked, and that the “leakable” amount of fuel varies with Event 7 timing. Subvar A provided an objective score of how well the crew was able to minimize fuel loss through the leak. Because the leaking tank would always be empty well before arrival, difference in diversion distances and airspeeds did not affect this value. Table 2 shows subvar A scores.

Table 2: Percent of leaked fuel which leaked ranked from best to worst

Best performance			Percent of leakable fuel actually leaked						Worst performance
41%	57%	60%	61%	63%	68%	68%	69%	73%	94%

Subvar B: Fuel efficiency. This variable indicated how efficiently the crew configured their aircraft. This depended on a number of factors including the number of engines running, the altitude and airspeed. Some crews indicated actions that would affect the efficiency at some point well beyond the duration of the simulation. This information was used to correct the extrapolated efficiency. Using the distance flown since Event 7 plus the extrapolated distance to the diversion airport, this was combined with the total amount of fuel burned between Event 7 and landing, which together determined the amount of fuel used by the engine(s) per nautical mile (NM). Airbus performance data was used to extrapolate fuel burn after the end of the simulation depending on the chosen cruise altitude and engine configurations. Table 3 displays the subvar B performance rankings.

Table 3: Fuel efficiency ranked from best to worst

Best performance			Fuel efficiency (kg/NM)						Worst performance
7.56	7.93	9.03	9.41	10.11	11.18	12.21	12.48	12.68	13.74

Subvar C: Distance overflown. This last sub-variable indicated the quality of a crew's diversion decision. Ideally, a good decision was made without delay, and to the nearest airfield. Upon Event 7, a crew would have the minimum information available to make a diversion call, even without a diagnosis. This variable calculated the nearest valid diversion field at Event 7, and compared this reference distance to the actual distance flown after Event 7 until landing. This was expressed as a percentage in which 20% overflown indicated that the crew flew 1.2 times the minimum achievable distance from Event 7. Both the time to make a decision as well as choosing the nearest most suitable airport were represented by this statistic. Table 4 displays the subvar C performance rankings.

Table 4: Percent distance overflown ranked from best to worst

Best performance			Distance overflown (% of nominal Event 7 distance)						Worst performance
6%	13%	14%	18%	31%	35%	37%	39%	40%	49%

Regression analysis. The three subvars were constituents of the total FOA. However, the fundamental safety margin of FOA could not be used as it was not easily corrected for Event 7 timings. In order to correct for Event 7 timings, the FOA variable was replaced by a heuristic variable based on the amount of fuel used (FU: both leaked and burned) after Event 7. The FOA and FU/NM variables were regressed to each other and aligned very closely ($p<0.001$, $R^2=0.96$). Subsequently, the three sub variables were regressed to the FU/NM heuristic variable using an Ordinary Least Squares (OLS) stepwise regression by backward elimination, and also exhibited a strong (explanatory) correlation ($p<0.001$, $R^2=0.946$).

3.2 Correction factors

The three correction factors were regressed to the four performance variables (three subvars and FU/NM). All values were averaged per crew. All 12 regressions use crew-averaged values and followed stepwise regression by backward elimination of corrector sub-indices, with performance indices as the dependent variables. The only two regressions which exhibited marginally usable R^2 (0.67, 0.64) values were the workload and realism relations with subvar C. However, these two R^2 values were too inaccurate to warrant adapting the performance indicators with them. Even without clear regression results, average ratings could still shed light on the quality and suitability of the experimental setup for evaluation purposes. The results indicated that a mild workload was present, with a notable increase in mental demand. The realism scores showed moderate scores for hardware and software ratings, but much better ratings for the scenario. The fatigue ratings indicated that, according the Samn-Perelli scale, on average participants felt “OK, somewhat fresh” after the simulation session.

3.3 Automation awareness

Behavioral observations which constituted AA were sourced from live observer notes, audio/video recordings and interview notes, with the addition of an automation attitude questionnaire collected via the self-test form. Regression of the questionnaire results with the performance variables did not reveal any significant relations, however it was interesting to note

that on average pilots felt they are were control and had good oversight, but were less confident about knowing the role and necessity of automation in controlling the aircraft.

Besides the questionnaire, the majority of the observations underwent processing prior to being merged with performance data. Observations were clearly redefined and subsequently initially classified according to the four AA affectors, and secondly classified according to their positive or negative contribution to the three performance subvars, creating 24 categories in total. It should be noted that all observations were processed without knowledge of that session's performance. This was to prevent a performance indication bias during categorization.

3.4 Combining behavior with performance

Within each category, the ten sessions were ranked from best to worst pertaining to that specific performance subvar. This was the first time that performance data and behaviors were matched together. By aligning the results, it became clear that good performers tended to have more positive and less negative comments, and vice versa. At this point a process called Recursive Abstraction (RA) was started, which essentially performed several subsequent summarizing steps in order to combine observations into increasingly more general trends. For this, the "Ladder" technique was devised. The Ladder technique began with the poorest performer's positive remarks, and then observes which behavior the second-to-last poorest performer exhibited, inferring that this change in behavior was causal to the improvement in ranking. The same was done for the negative comments, but worked in the opposite direction. The Ladder technique also employed a second phase in which the actual quantitative changes in performance were used to indicate which behavioral changes were more important (i.e. resulted in greater changes in performance). As a validation step, these behavioral changes were compared to the pilot responses to open questions about their experiences of surprise, unclarity and/or uncertainty. Only a few minor adjustments were required; on the whole the pilot and observer comments aligned well. The resulting 24 small cohesive descriptions of suspected causal changes in behaviors underwent RA again and were integrated across all three performance variables, resulting in eight conclusive paragraphs describing how each AA category either positively or negatively affected performance. Table 5 provides a very brief overview of the main points in these conclusions.

Table 5: Summary of behavioral results

	Knowledge	Rules
Behavior to promote	- Extensive knowledge of relevant technical details - Work effectively with a mental model of the problem - Rules-of-thumb being recalled - Good mental model of the current aircraft state	- Assume worst case scenario - Complete understanding of structure/diagnostics - Understanding procedure limitations - Plan of action/read through
Behavior to avoid	- General lack of understanding of system operation - Poor mathematics and calculations - Unfamiliarity with various failure cases - Not suspecting unlikely occurrences	- Distrust procedures - Poor awareness of suitable checklists - Do not read through the checklist - Do not go beyond the checklist

	- **Automation attitude**	- **Teamwork**
Behavior to promote	- Familiar with non-normal ops increases trust - Level of wariness, noting other information sources - Increase in trust increases performance	- Generate options, have a plan - Active cross-checking - Setting priorities - Task sharing with multiple players
Behavior to avoid	- Lack of deeper interest - Attempt multiple resets - Assume problem will fix itself - Increased level of distrust; assume computer is always the culprit	- Level of distrust in others - Increased reliance on others - Poor idea sharing - Too explicit task splitting

4 Discussion

The conclusions drawn in this study based their validity and usefulness for a large part on the realism that the scenario obtained. The realism evaluation scores indicated that the scenario and its events rank as very realistic, but the simulator hardware and software less so. The low simulator rating was for the most part based on the warming-up scenario where the manual flight control model was not suitably calibrated for the A330. The experimental scenario only employed auto flight features which functioned normally, hence this study concluded that the low score did not weigh as much as it may seem. The high scenario scores were important in confirming crew immersion into the scenario. With the suitable exception of mental workload, results indicated a low overall workload during the scenario, which was sought after. In addition, the low correlation between correctors and performance scores indicated that the scenario was more or less equally challenging and realistic for all crews, ratifying the comparability between crews. To conclude, this scenario was arguably a sufficient basis to draw conclusions about crew behavior in real life.

The second challenge in this study was to obtain objective ratings of performance. The scenario design intended to limit the number of contingencies without making it too obvious which options were possible. Given the fact that the fuel system was critically impaired, it could be argued that reduced flight endurance represented the largest flight risk at that time, and the best possible actions were arguably those which increased the fuel margin upon arrival. Subsequent use of sub-variable performances provided the detail in explanatory performance indicators in order to differentiate and grade different crew strategies leading to the same fuel upon arrival. In this way the scenario succeeded in creating an objective basis (in this case safety-related) for condemning or promoting certain behavior.

The third challenge in the setup of this study was to facilitate fair grading of crew behavior. Inherent to the fact of humans making observations, bias and subjectivity are never completely removed. However, this methodology employed five techniques to mitigate subjectivity:

1. *The sheer number of initial observations dampened any observer opinions in specific observations*

2. *All observations were verified between three different sources (notes, recordings, interviews)*

3. *Early in the process, observations were stripped of metadata, and were always treated separately from performance data (they were analyzed before the performance data)*

4. *The Ladder technique performed an additional validation with pilot self-reflection remarks not used before*

5. *Recursive abstraction enabled many (24) rankings based on objective performance, such that biases not already neutralized were suppressed by the generalization process*

In addition to this, throughout the RA process, comments seemed to line up very well as every stage, and positive and negative comments would often show complementary behaviors. Furthermore, the fact that the number of positive and negative remarks correlated with performance as expected (i.e. good performers have more positive and less negative remarks, and vice versa), indicated consistency in the observation process.

The above provided evidence of a sound basis upon which conclusions may be made. Referring to the conclusive behaviors sampled in Table 5, many (underlying) behaviors showed clear overlaps with previously documented behaviors. Automation complacency (Manzey, 2012) was demonstrated by crews waiting for further warnings, expecting malfunctioning automation to still support their decision making and hesitating to take over from the automated system. Automation bias appeared more pronounced in the form of “checklist bias”, where crews blindly assumed that the procedures would deliver clarity and resolve the issue, when in many cases such overreliance only made the complex failure difficult to understand due to changing aircraft configurations. It should be noted that both automation and checklist biases were more prominent in pilots with less flight experience (in the A330). Crews which exhibited a better understanding of the way the systems work, in some cases extensive technical knowledge, as well as familiarity with the limitations of the procedures, noted these assets to be very beneficial in deciding which checklists to apply and setting an effective course-of-action. Both assessment and course-of-action decision making errors (Orasanu et al, 1998) were identified in, for example, misdiagnosing slowly changing system states (assessment error) and poor critical evaluation of the outcome of a checklist (a course-of-action error). In some sessions, certain cognitive challenges resulted in stress and tunneling which were expressed by cursing, holding on to the incorrect memory of a display and locking into an incorrect diagnosis. Crews performed better when they cross-checked actions with each other, and were able to check their expectations for system responses with observations from other information sources. Crew confidence worked well in some cases with experienced pilots, but overconfidence in those less experienced often

backfired into worsening surprise and confusion after noticing unexpected system responses upon checklist directives. This study also reinforced results from another study (De Brito and Boy, 1999), in which survey results indicated that crew were in many cases unsatisfied with the current complexity and user-unfriendliness of checklist procedures; although this could arguably be the outcome of a systemic lack of knowledge about the systems concerned in the checklist, which was demonstrated in this study. The few crews with a more extensive understanding of how a fuel system operates (in general terms, not specifically the A330) were able to translate this knowledge to the checklist procedures of the A330, and arguably apply procedures more effectively.

In summary, this study demonstrated that it is possible to effectively assess crew behavior in an ecological setting, using objective performance indicators. Based on the results of this study, four statements were made in lieu of the research objective:

1. *There were clear differences in crew behavior with automation, despite equal type-qualifications*

2. *Variations in AA have demonstrably resulted in a significant variance in safety margins – poor AA directly results in poorer decisions which, in this scenario, led to a reduction in safety margins*

3. *Negative AA behaviors detailed in literature have been demonstrated to exist in an ecological context and also result in degraded performance, as proposed*

4. *Variations in AA were not directly attributable to experience, rank or crew composition*

This study concluded that variations in crew-automation interaction should not be underestimated in their potency to affect flight safety. With aviation design and operational trends bolstering automation capabilities, minimizing training standards and simplifying the role of the pilot, the effects of AA on safety may become more pronounced without due consideration for the unintended side effects. In light of a stagnating accident rate, the methodology and results of this study may help identify which AA mechanisms and behaviors should be prevented but also, more importantly, identify positive examples of AA mechanism and behaviors which should be promoted. Further studies can use these analyses to direct new training, HMI and operations design initiatives to effectively improve the crew-automation interaction system. The Man4Gen project is already paving this path upon their experimental results.

Although this project has succeeded in identifying behaviors which impact performance and safety, it does not yet explain how these (variations in) behaviors come to be in the first place. An attempt at correlating behavior with specific attributes such as flight experience, leadership skills, technical knowledge, etc. may help in further defining which competencies could be spearheaded in training and selection programs. An example of such an attempt is Project Samurai (Gorter and Jaeger, 2014), which attempted to investigate whether biofeedback training could improve manual flight control by accelerating psychological and somatic recovery after stressful events. Finding the source of behaviors may be critical in defining the effective solution.

5 Acknowledgements

This research was co-funded by the National Aerospace Laboratory (NLR) and as part of the European Commission FP7 2012 Aeronautics and Air Transport program under EC contract ACP2-GA-2012-314765-Man4Gen. The views and opinions expressed in this paper are those of the authors and are not intended to represent the position and opinions of the Man4Gen consortium and/or any of the individual partner organizations.

6 References

1. de Brito, G. & Boy, G. (1999, September). Situation awareness and procedure following. In *CSAPC* (Vol. 99, pp. 9-14).

2. Caldwell, J. A. (2012). Crew schedules, sleep deprivation, and aviation performance. *Current Directions in Psychological Science*, *21*(2), 85-89.

3. European Aviation Safety Agency. (2013). Annual safety review 2013. Retrieved from http://www.easa.europa.eu.

4. Endsley, M. & Jones, W. M. (1997). *Situation Awareness Information Dominance & Information Warfare*. Logicon Technical Services Inc. Dayton, OH.

5. Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32-64.

6. Gorter, D. & Jaeger, C. (2014). Project Samurai Pilot. Cognitive Enhancement of Airline Pilots with Biofeedback Training. Delft University of Technology, Faculty of Aerospace Engineering, and University of Leiden, Faculty of Psychology.

7. Hart, S. G. (2006). 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 Publications.

8. Hollnagel, E. (2002). Understanding accidents-from root causes to performance variability. In *Human factors and power plants, 2002. proceedings of the 2002 ieee 7th conference on* (pp. 1-1). IEEE.

9. Hollnagel, E. & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC Press.

10. Malinge, Y. (2011). The views from an aircraft manufacturer; what have we learned?. Retrieved from http://www.enco.eu/Safetyworkshop/Malinge_Airbus_presentation%20handout.pdf.

11. Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making* (Vol. 6 No. 1, pp. 57-87).

12. Mohrmann, J.F.W., Lemmers, A.J.J., & Stoop, J.A.A.M. (2013) Investigating Flight Crew Recovery Capabilities from System Failures in Highly Automated Fourth Generation Aircraft. NLR-TR-2013-574. National Aerospace Laboratory.

13. Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International journal of aviation psychology*, *8*(1), 47-63.

14. Orasanu, J., Martin, L., & Davison, J. (1998). *Errors in aviation decision making: Bad decisions or bad luck?*. National Aeronautics and Space Administration, Ames Research Center.

15. Potter, M. et al. (2012). D1.1 literature review. Technical report, Man4Gen Consortium.

16. Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *Systems, Man and Cybernetics, IEEE Transactions on*, (3), 257-266.

17. Wood, S. (2004). Flight crew reliance on automation. CAA Paper 2004/10. http://www.caa.co.uk.

WHAT IS NLR?

The NLR is a Dutch organisation that identifies, develops and applies high-tech knowledge in the aerospace sector. The NLR's activities are socially relevant, market-orientated, and conducted not-for-profit. In this, the NLR serves to bolster the government's innovative capabilities, while also promoting the innovative and competitive capacities of its partner companies.

The NLR, renowned for its leading expertise, professional approach and independent consultancy, is staffed by client-orientated personnel who are not only highly skilled and educated, but also continuously strive to develop and improve their competencies. The NLR moreover possesses an impressive array of high quality research facilities.

***NLR** – Dedicated to innovation in aerospace*

www.nlr.nl