Effective Naturalistic Decision Support for Dynamic Reconfiguration Onboard Modern Aircraft

Giuseppe Montano
The University of York
Deramore Lane, York YO10 5GH
United Kingdom
giuseppe.montano@cs.york.ac.uk

John McDermid
The University of York
Deramore Lane, York YO10 5GH
United Kingdom
john.mcdermid@cs.york.ac.uk

Paul Cairns
The University of York
Deramore Lane, York YO10 5GH
United Kingdom
paul.cairns@cs.york.ac.uk

ABSTRACT
We propose a novel naturalistic decision support system for complex fault management procedures onboard modern aircraft. Two experiments involving 13 civil pilots are presented. The results show that the system proposed improves pilots' decision accuracy, decision performance and situation awareness, whilst reducing mental workload and complacency to system advisories.

Keywords
decision support systems; situation awareness; avionics; dynamic reconfiguration; complacency.

1. INTRODUCTION
The aviation industry is moving towards a new approach to the development of avionic systems: Integrated Modular Avionics (IMA) (Conmy & McDermid, 2001). IMA, in brief, is a term used to describe an airborne real-time computer network consisting of sensors, actuators and a number of computing modules capable of supporting numerous applications of differing criticality levels. The Boeing 787, the Airbus A380, the Lockheed Martin F-22 and F-35 aircraft all mount IMA technology.

The modularity and flexibility of the IMA architecture enables advantage to be taken of the possibility to reconfigure the avionics to adapt the functionality to the changing conditions. By pooling the computing resources and allowing them to be shared by different subsystems, at the occurrence of a fault or if the system gets damaged whilst airborne, the process of IMA Dynamic Reconfiguration (IMA-DR) allows relocation of affected functions to other healthy computing modules.

In a typical scenario, a fault or damage affects the computing resources available and an IMA-DR is automatically triggered. A timely reconfiguration decision has to be made, which typically entails choosing which functions should be deactivated because of the degraded operating conditions.

To get an idea of the complexity of the problem, consider that the IMA of the Airbus A380 amount to 80 computing modules, each of them runs up to 21 avionics functions that can be activated and deactivated during a reconfiguration (Itier, 2007). The functions in question are inter-dependent, they have different criticality levels which change with the operating conditions, the consequences of deactivating any of them are very uncertain and, at the same time, the risk is high given the safety-critical context (especially for military aircraft).

Whilst the IMA-DR process cannot be completely automated for safety reasons (Montano & McDermid, 2008), one way of supporting the pilot in this complex type of decisions is by providing her with the assistance of a decision support system (DSS).

Previous NDM studies revealed the effectiveness of DSS that parallel cognitive strategies used by decision makers during complex decisions characterised by time pressure, high risks and uncertainty, e.g. (Miller, Wolf, & Thordsen, 1992). A review of the literature shows that mental simulation and story generation play a critical role in the majority of NDM models, e.g. Recognition-Prime Decision (Klein, 1989), Image Theory (Beach, 1998), Noble's model (Noble, 1993), Explanation-based model (Pennington & Hastie, 1988).

This study investigates the effects of a novel DSS for IMA-DR designed on the basis of NDM principles which parallels human cognitive strategies by favouring mental simulation and story generation. Two experiments are presented which are part of a wider campaign of experiments performed in the context of a four-year long study that examined the issues with high autonomy and authority solutions to the design of dynamically reconfigurable avionics for next-generation aircraft. We first investigated how pilots make decisions during dynamic reconfiguration operations under different operating conditions, including time pressure, heightened stress, different types of decision support information content and framing, different cockpit displays

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Authors retain copyright of their work.

Proceedings of the 10th International Conference on Naturalistic Decision Making (NDM 2011). May 31st to June 3rd, 2011, Orlando, FL, USA. S. M. Fiore & M. Harper-Sciarini (Eds.). Orlando, FL: University of Central Florida.

configurations. Then we used the results obtained to develop an effective DSS for IMA-DR.

1.1 HYPOTHESIS

The following research hypothesis is investigated in this study:

> During the process of IMA-DR, decision support information that parallels cognitive strategies and includes *explanations*, *implications* and an *assessment of the uncertainty* associated with the reconfiguration advice provided by the system would have a positive effect on pilots *situation awareness*, *workload*, *decision accuracy* and *performance*, thus it would improve the overall decision making effectiveness of the pilot and the safety of the process.

Two experiments that address this hypothesis are described and discussed hereinafter.

2. METHOD

As part of this research work, we developed the Safe and Interactive Reconfiguration Architecture (SaIRA), a framework for the management of the IMA-DR process based on the Constraint Programming paradigm that a) generates applicable configurations at run-time by merging information coming from the aircraft sensors, and b) autonomously generates *effective* decision support information.

For this study, SaIRA was integrated in a flight simulation framework–which includes eye-tracking technology–that was used to perform the experiments (Figure 1).

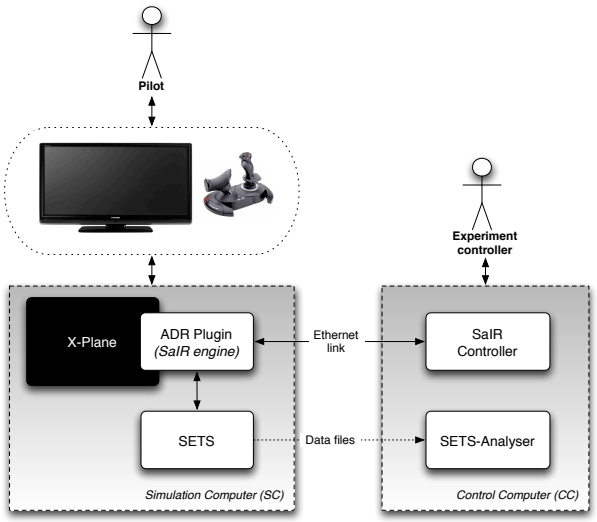

Figure 1. Simulation system architecture used in this study.

Technical details about the implementation of SaIRA, including the evaluation of the novel algorithms for automated decision support generation proposed, and about the SaIRA Eye-Tracking System (SETS) will be published separately and at a later time.

Thirteen civil pilots from two European airlines, certified to fly the Boeing 737 aircraft, participated in this study. At the time of writing, eleven pilots were resident in the United Kingdom and two in Italy. One of the pilots, of Italian nationality, served as a captain on the B737 and is now in retirement. All pilots were aged between 31 and 68; twelve of them are male, one is female.

Pilots were asked to perform a series of flight simulations in which the operating conditions were purposely manipulated in order to assess the research hypothesis. In a typical scenario, a fault was simulated during a critical manoeuvre of the flight (e.g. just before landing). A reconfiguration was required to mitigate the effects of the fault and the pilot was required to make a decision about whether accepting the advices of the system or not; in the positive case she also had to choose amongst two or more configuration options to apply amongst those suggested by the system.

We used two objective and two subjective metrics to characterise pilots' behaviour during IMA-DR decisions:

- *decision performance* (objective): this is a 'composite metric', made up of three sub-metrics: a) decision time, b) decision accuracy, and c) data exploration rate;
- *eye-movement* (objective): SETS is designed to record a large number of features of the eye movement. In the two experiments presented here, fixations duration (FD) is taken in consideration, and interpreted as an indication of task difficulty (Rayner, 1998);
- *mental workload* (subjective): the NASA-TLX (Hart & Staveland, 1988) technique has been used to assess pilots' mental workload (WL);
- *situation awareness* (subjective): the SA-SWORD (Vidulich & Hughes, 1991) technique was adopted for this study.

In addition, we conducted post-experiments interviews to verify the subjective results.

Figure 2 shows how SaIRA organises the decision support information on the Electronic Horizontal Situation Indicator (EHSI) of the Boeing 737-900ER cockpit display (used for the simulations). Additionally, schematics about the fault detected by the sensors temporarily replace the content of the Electronic Attitude Director Indicator (EADI) display, as shown in Figure 3.

SaIRA generates the following three cockpit conditions:

- *Description only* (baseline condition): only 'Fault information' and 'Diagnosis' data is displayed (upper portion of data in Figure 2). The original content of the EADI display is not modified.
- *Description & Schematics* (controlled condition): EHSI contains the same information of 'Description only' but the EADI shows schematics about the fault detected by SaIRA (Figure 2).

- *Full SaIRA Information* (controlled condition): full SaIRA decision support information is displayed, including explanation, implications, reliability figures and schematics, as shown in Figure 2 and Figure 3.

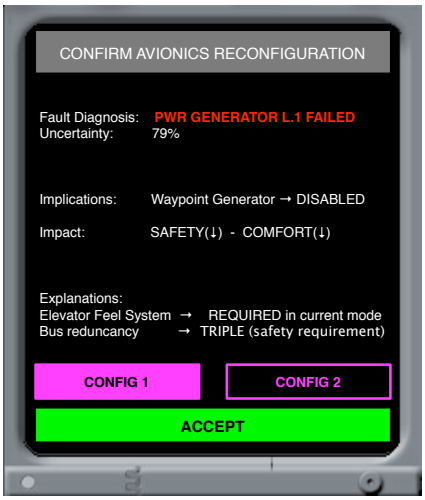

Figure 2. SaIRA decision support information ('Full SaIRA Information') on the EHSI display.

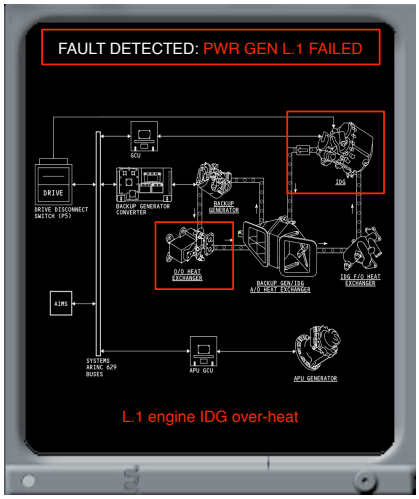

Figure 3. Schematics that describe the sub-systems mainly affected by the fault.

The eye tracking system superimposes a frame of 7 Areas Of Interest (AOI) on the B737 cockpit, as shown in Figure 4, which are used to characterise pilots' visual attention.

3. EXPERIMENT A

3.1 Description and Aim

Experiment A investigated the effect of *explanations*, *implications* and *schematics of the fault* on pilots' decision making behaviour. The effect of different conditions was examined in terms of decision accuracy, decision performance, frustration, workload and situation awareness.

3.2 Procedure

The pilot was asked to perform 6 simulations and complete any potential real-time fault management procedure correctly and in the shortest time possible.

Between 30 and 120 seconds after starting the scenario, a fault was simulated and a reconfiguration was automatically issued.

Two reconfiguration advisories were provided, one of which was *wrong*. The two advisories were always such that they required choosing between switching off one of two critical functions.

Figure 4. Definition of the AOIs on the cockpit of the Boeing 737-900ER.

The experiment was structured into 3 distinct tests. Being a within-subject test, each pilot ran all the simulations:

- *INFO_1 (Description only)*: pilots performed the first two simulations with 'Description only';
- *INFO_2 (Description & Schematics)*: pilots performed the following two simulations with 'Description & Schematics';
- *INFO_3 (Full SaIRA Information)*: pilots performed the last two simulations with 'Full SaIRA Information' (always showing 'FULL reliability').

Straight after the last test, both the NASA-TLX and the SA-SWORD questionnaires were submitted to the pilot.

3.3 Expectations

INFO_1 is the baseline condition. As a result of better decision support, we had the following expectations:

- E1: *decision accuracy* would have progressively improved with INFO_2 and INFO_3;
- E2: it was not possible to make any precise forecast concerning the *decision time* when the experiment was designed. On one hand better decision support should have reduced the time required by pilots to complete the procedure; on the other hand, more information to process could have increased the DT;
- E3: the *number of clicks on the reconfiguration buttons* would have progressively decreased with INFO_2 and INFO_3. We speculated that the number of times pilots switched from a configuration to

another to explore its characteristics would have been indicative of their confusion. Better decision support would have decreased pilots' confusion, hence this value should have decreased, too;

- E4: *fixation duration* would have progressively decreased with INFO_2 and INFO_3;
- E5: *workload* would have progressively decreased with INFO_2 and INFO_3;
- E6: *frustration* would have progressively decreased with INFO_2 and INFO_3;
- E7: *situation awareness* would have progressively improved with INFO_2 and INFO_3.

Altogether, expectations from E1 to E7 combine in the general expectation of obtaining improved decision performance with INFO_2 and, even more, with INFO_3.

3.4 Results

3.4.1 E1: decision accuracy (DA)

The Cochran's Q test reveals a statistically significant difference in terms of DA amongst INFO_1, INFO_2 and INFO_3 ($\chi^2(2)=7.091$, $p<0.029$). A pairwise comparison using the continuity-corrected McNemar's tests shows that the main improvement over INFO_1 (baseline) is provided by INFO_3. Table 1 contains the descriptive statistics.

Table 1. Decision accuracy under the effect of different types of decision support information. Columns 'Right' and 'Wrong' contain the number of pilots who made the right or wrong decision respectively.

	Right	Wrong	Decision accuracy
INFO_1	15	11	57.69%
INFO_2	20	6	76.92%
INFO_3	22	4	84.61%

3.4.2 E2: decision time (DT)

A significant effect of the type of decision support information on DT is revealed by the Friedman's test ($\chi^2(2)=13$, $p<0.02$). A post-hoc test using Wilcoxon Signed Rank tests with Bonferroni correction shows that the stronger decrease of DT is given by INFO_3 ($Z=-2.984$, $p<0.003$).

The descriptive statistics are provided in Table 2.

Table 2. Decision time (in seconds).

	Decision Time
INFO_1	36.78 (s.d. 6.36)
INFO_2	35.02 (s.d. 5.97)
INFO_3	28.63 (s.d. 8.61)

3.4.3 E3: number of clicks on the reconfiguration buttons (nrCL)

The statistical difference in terms of nrCL amongst the three conditions is confirmed by the Friedman's test ($\chi^2(2)=26.297$, $p<0.001$).

As expected, a progressive decrease of nrCL with INFO_2 and INFO_3 with respect to the baseline (INFO_1) is revealed by a post-hoc test performed through a series of Wilcoxon Signed Rank tests (INFO_2 vs INFO_1: $Z=-3.326$, $p<0.001$, $r=0.652$; INFO_3 vs INFO_1: $Z=-3.968$, $p<0.001$; INFO_3 vs INFO_2: $Z=-2.057$, $p<0.04$). These tests show that the biggest decrease of nrCL w.r.t the baseline is provided by INFO_3.

The descriptive statistics are provided in Table 3.

Table 3. Number of clicks on the reconfiguration buttons.

	nrCL
INFO_1	3.81 (s.d. 1.17)
INFO_2	2.88 (s.d. 1.07)
INFO_3	2.27 (s.d. 0.72)

3.4.4 E4: fixation duration (FD)

The Friedman's test reveals a significant influence of the independent variable on the FD ($\chi^2(2)=17.583$, $p<0.01$). The descriptive statistics are shown in Table 4.

Table 4. Fixation duration (in milliseconds).

	Fixation duration
INFO_1	410.01 (s.d. 10.55)
INFO_2	379.13 (s.d. 9.32)
INFO_3	354.89 (s.d. 7.04)

The biggest decrease of FD is provided by INFO_3 over INFO_1, as statistically confirmed by the Wilcoxon Signed Rank post-hoc test with Bonferroni correction ($Z=-4.229$, $p<0.001$).

3.4.5 E5 and E6: workload (WL) and frustration (FR)

Table 5 reports the results of the NASA-TLX test.

Table 5. NASA-TLX results. Parameters: Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Performance (PE), Effort (EF), Frustration (FR), Overall Workload (OWL).

	INFO_1	INFO_2	INFO_3
MD	71.92 (3.42)	63.46 (3.9)	50.00 (4.38)
PD	1.92 (1.21)	1.54 (0.87)	1.15 (0.61)
TD	32.31 (3.47)	27.69 (2.81)	31.15 (3.01)
PE	54.62 (3.94)	58.08 (4.1)	78.85 (2.34)

EF	54.23 (5.71)	46.54 (3.37)	34.23 (3.66)
FR	61.23 (5.72)	52.31 (4.03)	25.00 (2.59)
OWL	*52.23 (2.91)*	*47.15 (1.77)*	*39.1 (1.83)*

An one-way ANOVA test is run on each parameter of the NASA-TLX test. As a result, a strong, significant effect of the independent variable is found on all the parameters except PD and TD (see Table 6).

Table 6. Results of the one-way ANOVA test on the NASA-TLX results.

NASA-TLX Parameter	ANOVA result
MD	F(2,37)=7.95, p<0.001
PD	F(2,37)=0.171, n.s.
TD	F(2,37)=0.597, n.s.
PE	F(2,37)=13.605, p<0.001
EF	F(2,37)=5.32, p<0.009
FR	F(2,37)=19.219, p<0.001
OWL	*F(2,37)=8.802, p<0.001*

The Tukey HSD post-hoc test reveals that INFO_3 provides a stronger improvement then INFO_2 on the baseline INFO_1 (given the number of permutations, for the sake of brevity the figures are not reported here). Furthermore, a statistical improvement of INFO_3 is confirmed on INFO_2 for PE, FR and OWL.

As one of the parameters of the NASA-TLX method is frustration (FR), this technique allows collecting also concerning expectation E6. Table 5 and Table 6 reveal that FR decreases statistically with both INFO_2 and INFO_3, confirming the effectiveness of the decision support information produced by SaIRA.

3.4.6 E7: situation awareness (SA)

SA-SWORD does not provide a direct measure of SA but it is designed to give an assessment of which type of information gives the highest SA. As expected, the order for increasing level of SA is (1) INFO_1 (lowest SA), (2) INFO_2, and (3) INFO_3 (highest SA).

An one-way ANOVA test reveals a strong effect of the independent variable on the subjective assessment of SA ($F(2,37)=1860.943$, $p < 0.001$). The Tukey HSD post-hoc test shows that INFO_3 gives the strongest improvement.

3.5 Discussion

The main result is that, in general, the effectiveness of the complete set of decision support information generated by SaIRA (i.e. INFO_3) is strong in terms of all the dependent variables considered. To a certain extent, DA, nrCL, FD, WL, FR and SA all behaved as expected, providing evidence of a significant improvement in all the aspects of pilots' decision experience during ADR. An improvement is also found in terms of DT, which we were not in the position to predict.

The improvement brought by a graphical representation of the fault over the baseline, textual information set (i.e. INFO_2 *versus* INFO_1), is not as strong as in other studies like FAMSS (Hayashi, Huemer, & Lachter, 2006). It must be noted, however, that projects like FAMSS are specifically targeted to the design of effective graphical representations of the fault management information, whilst this study has a different objective: it is mainly tailored to the analysis of the effects of textual information made of explanations, implications and reliability information on the interactive fault management process in first instance. One possibility is that the graphical information generated by SaIRA is not as sophisticated and effective as the information produced by more advanced graphic engines like FAMSS. It would be interesting to analyse the combination of the two approaches.

An unexpected result comes from the NASA-TLX: pilots ranked their performance higher in the scale with INFO_3 than with the other information formats. In this regard, (Fox & Tversky, 1995) argue that feelings of competence occur when people have clear versus ambiguous knowledge. INFO_1 and INFO_2 provide less information than INFO_3, hence there is the possibility that the former two types of information leave room for ambiguities in pilots' minds. With reference to the *support theory* of reasoning (Tversky & Koehler, 1994), the content of INFO_3 is "unpacked" into more explicit disjunctions, a fact that, according to the theory, increases the "strength of belief" of the decision maker and decreases the ambiguity. We speculate that, as a result of this phenomenon, pilots would feel more competent and give themselves a higher performance score.

In the context of a general evaluation of SaIRA, it is particularly important to remark the positive effect of INFO_3 on frustration.

4. EXPERIMENT B

4.1 Description and Aim

The textual decision support information generated by SaIRA is made of explanations, implications and an assessment of the reliability of the reconfiguration advice generated by the system. Experiment A focused on explanations and implications (the information generated by the system was always assumed to be fully reliable); the assessment of the third component requires a different analysis, as shown hereinafter.

On-board modern aircraft, faults are detected and identified by sensor data fusion technology (e.g. 'Block 3.0 avionics' by Lockheed Martin (Caires & Stout, 2002)). SaIRA is designed to calculate the degree of uncertainty embedded in a fault assessment by using algorithms based on Constraint Programming and Evidential Reasoning techniques.

The aim of this experiment was collecting information about the potential effect of different degrees of reliability associated with decision support advices of dubious genuineness on pilots' decision behaviour.

We advance the following claim:

Providing reliability *figures would influence pilots' decision-making performance in the following ways:*

- *Because of the framing effect, pilots would feel more comfortable applying a configuration associated with high reliability than with low uncertainty;*
- *Evidently wrong IMA-DR advisories, when associated with low reliability, would be more easily spotted and avoided than without any reliability figure;*
- *Low and high reliability options would both be easier to process than medium reliability options, i.e. the decision time would increase with medium reliability options.*

4.2 Procedure

The pilot was asked to perform three flight simulations and complete any potential real-time fault management procedure correctly and in the shortest time possible. The pilot was also informed that the system could have potentially generated wrong decision support information as a result of technological limitations.

A safety-critical fault was simulated between 30 and 120 seconds after the start. The system was configured to generate only one configuration option; the pilot could either accept it or switch to safe mode.

Pilots were divided in two groups: Group A and Group B. All pilots performed Test 1; then Group A performed Test 2a and Group B performed Test 2b, as follows:

- *Test 1 - both Group A and Group B*: SaIRA generated *right* decision support information showing *FULL reliability*. This was the baseline test, aimed at building up pilots' confidence in the system before providing them with wrong information;
- *Test 2a - Group A only*: SaIRA generated *wrong* decision support information showing *LOW reliability*;
- *Test 2b - Group B only*: SaIRA generated *wrong* decision support information showing *MEDIUM reliability*;

4.3 Expectations

The following results were expected:

- E1: *workload* would have been higher with MEDIUM reliability than with LOW or FULL reliability;
- E2: *fixation duration* would have been higher with MEDIUM reliability than with LOW or FULL reliability;
- E3: *decision time* would have been higher with MEDIUM reliability than with LOW or FULL reliability.

4.4 Results

This experiment has a mixed factorial design. The two independent variables are the *correctness* of decision support information (which can be either 'correct' or 'incorrect', with the former being the baseline condition) and its *reliability* (either 'LOW', 'MEDIUM' or 'FULL', with FULL being the baseline condition). Correctness is the within-subjects independent variable (i.e. all pilots test both its conditions) and reliability is the between-subjects independent variable (i.e. Group A is tested with the 'LOW' reliability condition and Group B is tested with the 'MEDIUM' condition).

For FD and DT, the main effect of both correctness (C) and reliability (R) is assessed; when ANOVA is used (i.e. for WL), also the interaction between correctness and reliability factors is assessed (CR).

It must be noted that the main objective of this experiment, as previously stated, is investigating the effect of the 'reliability' factor. However, because of the nature of the decision support information, it was not possible to design this experiment without using both correct and incorrect information.

4.4.1 E1: workload (WL)

The results of the NASA-TLX test are show in Table 7. The factor 'Group' tests for the difference of reliability (degREL) whilst the factor 'Test' examines the effect of the correctness of the information provided. Physical demand is not reported because it was rated null by all pilots.

Table 7. NASA-TLX results.

	Test 1	Test 2	Group A	Group B
MD	66.67 (2.07)	74.14 (3.83)	64.58 (2.71)	76.25 (2.83)
TD	35.42 (4.15)	45.42 (4.01)	34.17 (3.53)	46.67 (4.28)
PE	83.33 (2.56)	59.17 (2.88)	71.25 (3.8)	71.25 (5.19)
EF	54.25 (2.85)	67.5 (5.06)	57.17 (4.39)	64.58 (4.46)
FR	25.83 (2.74)	61.67 (7.24)	34.58 (2.85)	52.92 (9.74)
OWL	*43.26 (2.23)*	*59.25 (3.56)*	*47.2 (1.51)*	*55.31 (4.9)*

A two-way split-plot ANOVA test was performed on each parameter of the NASA-TLX test except PD. The results for the main effect of correctness (C), reliability (R) and for their interaction (CR) are reported in Table 8, Table 9 and Table 10 respectively.

Table 8. Main effect of 'correctness' of the decision support information (two-way split-plot ANOVA).

NASA-TLX Parameter	Effect of 'correctness' (C)
MD	F(1,10)=5.031, p<0.049
TD	F(1,10)=5.294, p<0.044
PE	F(1,10)=40.239, p<0.001
EF	F(1,10)=7.28, p<0.022
FR	F(1,10)=80.742, p<0.001
OWL	*F(1,10)=66.87, p<0.001*

Table 9. Main effect of 'reliability' of the decision support information (two-way split-plot ANOVA).

NASA-TLX Parameter	Effect of 'reliability' (R)
MD	F(1,10)=13.517, p<0.004
TD	F(1,10)=4.556, n.s.
PE	F(1,10)=1.722, n.s.
EF	F(1,10)=1.686, n.s.
FR	F(1,10)=30.062, p<0.001
OWL	*F(1,10)=7.026, p<0.024*

Table 10. Interaction between 'correctness' and 'reliability' of the decision support information (two-way split-plot ANOVA).

NASA-TLX Parameter	Correctness/reliability interaction (CR)
MD	F(1,10)=6.211, p<0.032
TD	F(1,10)=2.353, n.s.
PE	F(1,10)=1.722, n.s.
EF	F(1,10)=4.941, p<0.05
FR	F(1,10)=44.716, p<0.001
OWL	*F(1,10)=51.563, p<0.001*

In line with E1, WL with MEDIUM reliability is higher than with the other two cases. It must be noted that WL is higher than the baseline also with LOW reliability.

Interestingly, a peak of temporal demand (TD) is recorded with MEDIUM reliability. This is an unexpected result because no time limits for decisions are set for this experiment. We speculate that the increased perception of TD is a by-product of the increased frustration and cognitive demand. NASA-TLX data was not processed in real-time (as eye movement data), hence it was not possible to make questions about this result to pilots in their post-experiments interviews.

Another interesting outcome is the statistical, negative effect of LOW reliability on pilots' perception of their performance. In practice, the results about their decision accuracy (DA) show that, contrary to the participants' perception, their performance–although lower in average–wasn't statistically worse than in the baseline case (Wilcoxon Signed-Rank test, Z=-1.171, n.s.).

4.4.2 E2: fixation duration (FD)

Table 11 reports the descriptive statistics concerning FD for Experiment E.

Table 11. Fixation duration (in milliseconds) under the effect of 'correctness of information' (Test 1 vs Test 2) and 'reliability of information' (Group A vs Group B).

	Fixation duration
Test 1	384.83 (s.d. 10.61)
Test 2	409.53 (s.d. 20.93)
Group A	371.52 (s.d. 8.56)
Group B	421.84 (s.d. 19.86)

The Wilcoxon Signed-Rank test reveals no statistical effect of 'correctness' of decision support information on pilots' FD (Z=0.706, n.s.). Either the pilots didn't notice the wrong information (which supports the hypothesis of an automation-induced complacent behaviour) or they didn't have any observable physiological reaction in terms of FD.

On the other hand, the Mann-Whitney U test reveals a strong effect of the 'reliability' factor (Z=2.882, p<0.004); this test compares the Group A and Group B *within* Test 2. The analysis of FD confirms the increased complexity of processing MEDIUM reliability information.

4.4.3 E3: decision time (DT)

The descriptive statistics for DT are reported in Table 12.

Table 12. Decision time (in seconds) under the effect of 'correctness of information' (Test 1 vs Test 2) and 'reliability of information' (Group A vs Group B).

	Decision Time
Test 1	31.65 (s.d. 2.32)
Test 2	42.85 (s.d. 4.47)
Group A	32.57 (s.d. 1.8)
Group B	41.93 (s.d. 4.89)

Similar results to FD were found for DT. The Wilcoxon Signed-Rank test shows no statistical effect of 'correctness' of decision support information on pilots' DT (Z=1.883, n.s.).

The Mann-Whitney U test, instead, reveals a statistically significant effect of the 'reliability' factor ($Z=2.722$, $p<0.006$).

A correlation is found between FD and DT (Spearman's test: $\rho=0.509$, $p<0.011$), which contributes to the robustness of the conclusions.

4.5 Discussion

All the three claims have been confirmed by the experimental results. The main conclusions are that (a) MEDIUM reliability worsens ADR decision performance and (b) LOW reliability improves pilots' performance in discarding erroneous information.

In both cases, reliability information has proven to allow pilots to make a more informed decision, which is a determining element in the design of a safety-critical system.

This experiment showed that reliability information has an effect on pilots' decision performance. An improvement in decision accuracy is detected but it is not possible to draw robust conclusions from this experiment alone because of insufficient statistical power. More robust conclusions about the impact of SaIRA on pilots' decision accuracy can be obtained from other experiments of our empirical assessment campaign, which will be published in the upcoming future.

5. CONCLUSIONS

A major contribution of this study is demonstrating the effectiveness of NDM principles as a driver for the design of DSS technology capable of improving human decision making performance and accuracy in safety-critical contexts.

A novel decision support information framework for complex fault-management procedures on-board modern aircraft is proposed, which is specifically designed to favour human mental simulation. SaIRA, the DSS proposed, is found to improve human decision accuracy, decision performance, and situation awareness during dynamic reconfiguration decisions. Other ancillary results also cooperate to attest the effectiveness of the framework, e.g. reduced cognitive workload and reduced frustration in situation of heightened stress and time pressure.

The positive results obtained in this study make the decision support framework proposed a promising approach which we plan to study also in other decisional contexts, different from aviation, e.g. nuclear power plant control.

6. REFERENCES

Beach, L. (1998). *Image theory: Theoretical and empirical foundations*. Mahwah, New Jersey London: Lawrence Erlbaum.

Caires, G., & Stout, J. (2002). Newest Advanced Integrated Avionics Software Package Flown for First Time Aboard the F-22 Raptor Air Dominance Fighter. Retrieved February 13, 2010, from www.lockheedmartin.com.

Conmy, P., & McDermid, J. (2001). High level failure analysis for Integrated Modular Avionics. *Sixth Australian workshop on Safety critical systems and software* (Vol. 3, pp. 13-21). St Lucia, Queensland: ACM Press.

Fox, C., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*(3), 585-603. Oxford University Press.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and research. *Human mental workload*, *1*, 139-183.

Hayashi, M., Huemer, V., & Lachter, J. (2006). Evaluation of an Advanced Fault Management System Display for Next Generation Crewed Space Vehicles. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *50*(1), 136--140. San Francisco, USA: Human Factors and Ergonomics Society.

Itier, J.-B. (2007). A380 Integrated Modular Avionics. *Proceedings of the ARTIST2 meeting on Integrated Modular Avionics*. Rome, Italy: INRIA. Retrieved from http://www.artist-embedded.org/artist.

Klein, G. (1989). *Recognition-primed decisions*. Fairborn, OH, USA: Klein Associates Inc.

Miller, T., Wolf, S., & Thordsen, M. (1992). *A decision-centered approach to storyboarding anti-air warfare interfaces* (pp. 1-47). San Diego, CA: Contract No. N66001-90-C-6023 for the Naval Command, Control and Ocean Surveillance Center.

Montano, G., & McDermid, J. (2008). Human involvement in dynamic reconfiguration of Integrated Modular Avionics. *IEEE/AIAA 27th Digital Avionics Systems Conference, 2008. DASC 2008*.

Noble, D. (1993). A model to support development of situation assessment aids. *Decision making in action: Models and methods*, 287-305. Norwood, NJ, USA: Ablex Publishing.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 521--533.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

Vidulich, M., & Hughes, E. (1991). Testing a subjective metric of situation awareness. *Human Factors and Ergonomics Society Annual Meeting*, *35*(18), 1307-1311. Human Factors and Ergonomics Society.