Trust in and Use of Automation: Their Dependence on Occurrence Patterns of Malfunctions

M. Itoh*, G. Abe** & K. Tanaka*
* Graduate School of Information Systems, University of Electro-Communications,
Chofu, Tokyo 182-8585, Japan
{mako,tanaka}@is.uec.ac.jp
** Road Transportation Research Division, Japan Automobile Research Institute
Tsukuba, Ibaraki 305-0822, Japan
agenya@jari.or.jp

ABSTRACT

In the present paper, the dynamics of trust in and use of automation are shown to depend on the occurrence patterns of malfunctions. Two experiments have been conducted as part of the present study. The first experiment compared the effects of two typical occurrence patterns, continuous and discrete. The second experiment analyzed the effects of the combination of the two occurrence patterns of malfunctions. Several analyses were performed and the following findings were obtained. If continuous malfunctions occur, operator trust is decreased significantly and eventually the operator does not rely on the automation, even under circumstances that are easy for the automation to handle. The longer the continuity, the longer this effect lasts. In contrast, discrete malfunctioning does not cause a significant decline in the operator's level of trust. A single malfunction or a small number of discrete malfunctions causes that some operators not to rely on the automation when the operation is difficult for the automation. However, the level of trust returned quickly under normal operating conditions. As operators gradually experience more individual malfunctions, they appear to become less sensitive to the malfunctioning.

1. INTRODUCTION

In modern technological systems, such as aircraft or nuclear power plants, computers can and often do bear authority and take responsibility for decision making and control. Various jobs which had previously been performed by human operators are now performed by automated systems. As the term supervisory control [15] indicates, human operators assure the role of a supervisor and automated systems operate as subordinates in modern systems. Thus, how effectively humans and automated systems can cooperate is a problem.

Since today's automated systems are highly complex, operators may fail to understand the mechanism of the automation. The complexity of the system can induce uncertainty concerning whether the automation works well in a particular situation. The operator's level of trust in an automated system is important under such circumstances. If the operator trusts the automation too much, the automation may be relied upon even when the circumstances are inappropriate for the automation. The overreliance on automation can be one cause of catastrophic accidents [17,18,19]. For example, one of the leading causes of A320 accident at Mulhouse-Habsheim appears to be the overconfidence of pilots with respect to the automation which was not designed to work in the situation [17]. If an operator distrusts an automated system, the system can be disabled. To make human-machine interaction effective, we should clarify how operators come to trust or rely on automation. This is important for establishing "human-centered automation" [2,16], because one of the aims of human-centered design is to foster user acceptance [12].

Several studies have examined on trust in and use of automation. Lee & Moray [4] and Muir & Moray [9] showed that automation reliability and the operator's trust in automation are major factors in the usage of automation. Riley developed a complex model of reliance on automation [13]. In his model, risk and perceived risk are other factors of automation use. Parasuraman & Riley have given a detailed survey on the usage of automation [11]. These papers show that trust is one of the most major factors that affects the use of automation.

However, the nature of trust has not been clarified completely. Muir [8] developed a theory of trust in automation and a method of measuring the subjective level of trust. Conducting an experiment using a simulated pasteurization plant, Lee & Moray [4] showed that the dynamics of trust can be modeled by a first-order lag model. One of the authors of this paper has developed a similar first-order lag model of the dynamics of trust in a previous experiment [3,7]. Although these models are valid for fitting the data, the reason why these models can be obtained remains unclear. What is the cognitive mechanism that drives trust? This mechanism should be made clear to realize the human-centered automation.

In order to construct the mechanism of the dynamics of trust, the present study focuses on how a malfunction or fault affects trust. Since subjective level of trust is often measured by a 10-point

0-7803-5731-0/99/$10.00 ©1999 IEEE

scale, one method which can be used to describe a mechanism of the dynamics of trust is a probabilistic approach which uses Bayes' theorem for belief updating. In this method, the total effect of the malfunctions is not dependent upon the occurrence patterns of the malfunctions. However, several studies have reported a tendency for the effect on trust of a malfunction to be situation dependent [3,7,10]. In other words, the effect depends on the occurrence patterns of the malfunctions. For example, in the first-order lag model of trust, the occurrences of faults or malfunctions at the n-th trial and the (n - 1)-th trial both affect trust. In two reports, [3,7], the decreasing effect on trust of the n-th malfunctioning can be strengthen by the (n - 1)-th malfunctioning. Itoh et al. [3] show that the decrease in the rate of trust tends to be lower after the subject has observed too much malfunctions. Parasuraman et al.[10] found that vigilance is high if the reliability of automation is changing, but that vigilance is low if the reliability remains constant.

Two experiments have been conducted in order to determine whether the effect of a malfunction depends on the occurrence patterns of the malfunctions. From a macroscopic point of view, Abe et al. [1] analyzed the data, which supported the dependence of the dynamics of trust on the patterns of the occurrences of malfunctions.

However, the reasons of differences in effects among the occurrence patterns of malfunctions remain unclear. This paper clarifies the manner in which the operator trusts the automation and how the trust shifts due to malfunctions. A model-based approach is used for the analyses.

2. RESEARCH VEHICLE

Process Control Task

The experiments conducted in the present study are applied to computer-controlled simulation of a plant (Fig. 1). The simulated plant produces mixed fruit juice.

Mixed fruit juice is produced to order. A desired quantity of mixed juice and the corresponding time required for pasteurization are specified in each order sheet. The process for producing mixed fruit juice in the simulated plant is as follows:

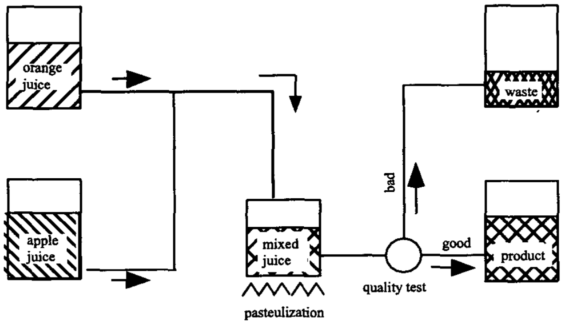

Fig. 1 Pasteurizing Plant for Mixed Fruit Juice

1. An order sheet arrives.
2. The orange juice and apple juice are taken into the mixture tank.
3. The time required for pasteurization is determined.
4. The mixed raw juice is pasteurized.
5. The quality of the mixed and pasteurized juice is tested, and if the quality of the mixed juice is not high enough, the juice will be discarded.

Automation

The production process of the mixed juice is automated in order to ensure that the numerous orders are filled quickly and accurately. Human operators, however, can intervene in the process in order to set the pasteurization time. This automated process is not always successful due to the following two reasons.

Error: The quantity of raw juice that flows into the mixture vat does not always equal exactly that specified in the order sheet. In the present paper, error (E) is referred to as the difference between the desired mass (D) and the actual mass (A) in the mixture vat, i.e., E=D-A. The automatic pasteurization is assumed to be successful in most cases if the absolute value of the error $|E|$ is within five percent of D (E^*). However, if $|E| > E^*$, the pasteurization time should be recalculated according to D, otherwise the pasteurization by the automation will fail.

Malfunction: The automation can set the pasteurization time incorrectly for various reasons. If a malfunction occurs, the pasteurized mixed juice will not pass the quality test even when $|E|<E^*\approx 0.05D$. An operator can only determine that a malfunction has occurred when the automatically pasteurized juice fails the test, and the juice is wasted. A malfunction can occur at the intermediate or high level of error (3 - 5 %), which means that the greater the error is, the difficult the operation is for the automation.

Operator's task

The operator is responsible for the production of mixed juice. The task imposed on the operator is the supervision of the automation. If the operator believes that the automation has not set the pasteurization time properly, the operator should intervene and set an appropriate pasteurization time.

The operator is encouraged to rely on the automatic system as much as possible, because orders must be filled as fast as possible and automatic pasteurization is faster than manual pasteurization.

3. SUMMARY OF PREVIOUS WORK

This section summarizes the experiments conducted in the present study and the results of analyses in a previous work [1].

Experiment 1

Purpose: The purpose of the present experiments is to determine whether the effect of a malfunction on levels of trust depends on the malfunction occurrence pattern. We begin this study by comparing two typical cases. The first is continuous pattern, in which all malfunctions occur one after another as a series of back to back malfunction. The second is discrete pattern, in which malfunctions are distributed uniformly.

Method: Subjects observe five malfunctions out of a total of 100 trials. Thus, the reliability of the automation is 95%. Fig. 2 illustrates the distribution of the malfunctions. Case A is continuous and case B is discrete. Twenty participants, graduate and undergraduate students, were asked to act as an operator for one of the two cases. The error (E) between the desired mass and the supplied mass is randomized for each trial. Three different measures are used to analyze the dynamics of trust.

1. Mode selection of setting pasteurization time.
   Each subject selects between manual mode or auto mode for each trial. In the manual mode, the operator intervenes and sets the time manually. In the auto mode, the operator allows the automation to set the pasteurization time. Since subjects are encouraged to operate in the auto mode, selection of manual mode indicates that the automation is distrusted. The term intervention rate is used here to refer to the proportion of subjects in a group who chose the manual mode for each trial. The intervention rate ranges from zero to one, and all operators rely on (intervene in) the automation if the intervention rate is zero (one). Thus, it can be assumed that the value of *1 - the intervention rate* corresponds to the level of trust.
2. Time till decision (TTD) on the mode selection.
   If the operator trusts the automation very much or not at all, TTD should be short. On the other hand, TTD should be longer if the operator's trust in the automation is vague or variable. Thus, it can be measured by TTD how much an operator is sure whether he/she trusts or distrusts the automation.
3. Subjective rating of trust.
   The operator's subjective level of trust in the automatic system is measured using a 10-point rating scale, which is normally used in studies on trust [3,4,5,7,9]. Each subject is requested to answer subjective feeling of trust for every 20 trials.

Results: In the previous study [1], analyses were performed in order to clarify major trends of trust. Fig. 3 shows an example of the data. This figure shows how the intervention rate shifts during trials. According to these analyses, the following results were obtained with respect to mode selection.

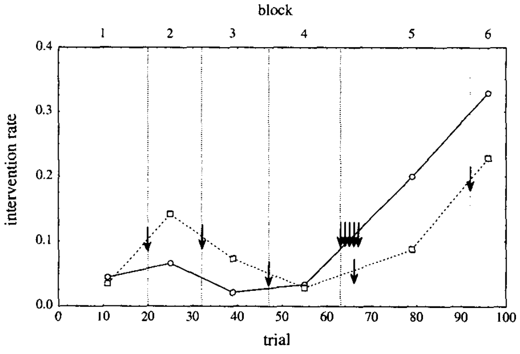

Fig. 3 Shifts of Intervention Rate (after Abe et al.[1])

1. The intervention rate increases significantly after experiencing the five successive malfunctions.
2. No strong evidence was found which indicated that the intervention rate increases if the operator observes a single malfunction.
3. The total effect of the five successive malfunctions on trust is not the same as the total effect of the discrete malfunctions.

Analyses of TTD and the subjective rating of trust showed similar results.

The above results support that the dynamics of trust depends on the occurrence patterns of malfunctions.

Experiment 2

Purpose: The purpose of the second experiment is to examine the effects of the occurrence patterns of malfunctions in more detail. This experiment is an attempt to answer two questions. First, how many successive malfunctions are sufficient to cause a significant and macroscopic decrease in trust? In other words, are five trials necessary required in order to cause a significant decrease in trust? Second, does the reestablishment of trust after observing continuous malfunctions depend on the preceding experience?

Method: Subjects observe five malfunctions out of a total of 100 trials. In order to compare the effects of number of continuous malfunctions on trust with the results obtained in the first experi-

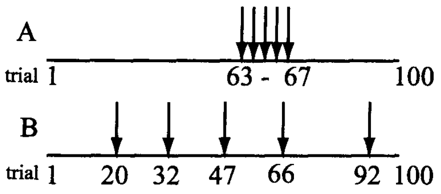

Fig. 2 Schedule of First Experiment

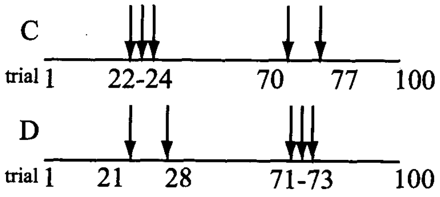

Fig. 4 Schedule of Second Experiment

ment, three successive malfunctions are provided in the present experiment. Two groups are prepared in order to compare the effect of experience, as shown in Fig. 4. Case C is continuous-discrete and case D is discrete-continuous. Sixteen graduate and undergraduate students participated in this experiment. The subjects are divided into two groups randomly. Other conditions, such as the error (E), are configured in the same manner as the first experiment. Since mode selection and TTD can accurately reflect the operator's subjective feelings toward the automation, the two objective measurements are used.

Results: Abe et al. [1] have analyzed how trust shifts after experiencing three successive malfunctions. For mode selection, by plotting the moving average and developing a linear regression model as a function of trial number and the error, the following results have been obtained.

1. The three successive malfunctions cause an increase in the intervention rate, similar to the case for the five successive malfunctions.
2. If the operators do not have much experience, the increased intervention rate due to the successive malfunctions does not decrease immediately after the end of the malfunctions.
3. If the operators are very experienced, the increased intervention rate decreases quickly to the previous level after the end of the malfunctions.
4. After observing successive malfunctions, the experienced group became more sensitive to the error than the less-experienced group did.

These findings can be explained as follows.

- After the occurrence of the successive malfunctions, the less-experienced operators tend to be confused as to how best to interact with the automation.
- The experienced operators, on the other hand, are not confused.

One possible explanation for this is that the subjects have changed their perception of the limits within which the automation works. Similar results have been obtained for TTD.

Thus, the answer to the first question is that the significant decrease in trust occurs even when the number of successive malfunctioning is less than five. The answer to the second question is yes. The reestablishment of trust after observing continuous malfunctions is dependent on the preceding experiences.

4. MODEL-BASED ANALYSES

In the previous work, the mean value for each block of trials was used for the analyses of intervention rate and TTD, where a block is referred to a number of successive trials. This approach is useful for examining the major trends of trust since the relationship between trust and the error can be ignored. However, it would not be useful to clarify how the tendency to rely on automation depends on the value of the error or how this tendency shifts. A (linear) multiple regression analysis was used to clarify the relationship between the measures and the error explicitly. Strictly speaking, however, the relationship between the intervention rate and the error can not be regarded as linear, as described in the following subsection. In order to understand exactly how operators trust in and rely on the automation, analyses based on models of the relationship between the measures and the errors should be performed.

Models of measures

Trustworthiness of automation: A system does not work well in every situation even if the system is highly reliable. Rather, the system does work within a prescribed range of conditions. Thus, the system is not reliable if the prescribed conditions are not met. For example, some stall protection systems in commercial aircraft do not work below the prescribed height. However, in most cases, the border between the condition in which the automation should work and the condition in which the automation is not supposed to work is somewhat unclear for operators. The range over which the automation might be applied can be divided into three zones: a reliable zone, a gray zone, and an unreliable zone, as shown in Fig. 5.

Mode selection: In these experiments, subjects were informed that the automation is reliable if the absolute value of the error (E) is less than five percent (E^*). Thus, the intervention rate might be modeled as shown in Fig. 6, and can be described as a sigmoid function.

TTD: If an operator trusts or distrusts the automation strongly, their decision will be made quickly. However, if the operator is not sure whether the automation is trustworthy or not, then TTD tends to be large. TTD might be modeled as shown in Fig. 7, since the operator tends to hesitate to judge the situation if the error is around the limit of the functioning of the automation (E^*).

Analyses of the dynamics of trust

This paper focuses on the intervention rate.

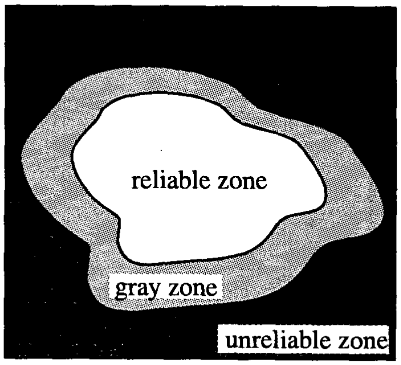

Fig. 5 Trustworthiness of Automation

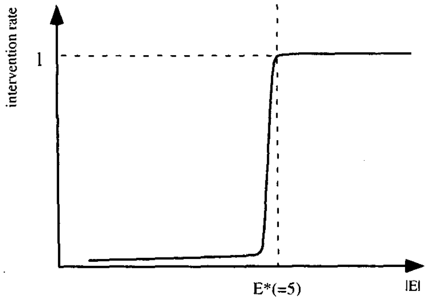

Fig. 6 Model of Intervention Rate

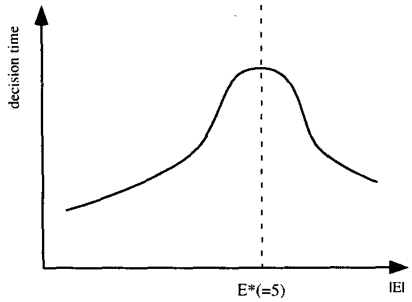

Fig. 7 Model of TTD

Intervention rate: For each group, the intervention rates for a number of successive trials, such as trials 1-30, trials 2-40, etc. Figures 8 and 9 are examples of the obtained plots. Three typical plots are shown in each figure. Solid circles indicate the trials after experiencing successive malfunctions. Analyses based on these plots revealed the following:

1. Appropriateness of the modeling of the intervention rate
   1a. Intervention rate can be modeled by a sigmoid function.
2. Effects of the occurrence of the five continuous malfunctions
   2a. Just after the malfunctions, the intervention rate at the intermediate level of error (around 3 - 4%) increases considerably.
   2b. Just after the malfunction, the intervention rate at the low level of error (less than 2%) increases slightly.
   2c. The increased intervention rate remains at a high level for some time.
3. Comparison between the effects of the five and three continuous malfunctions
   3a. In both cases, a few operators tend to choose the manual mode even when the error is small.
   3b. Although the intervention rate also increases for a moment due to the three malfunctions, it decreases gradually. The decrease in the intervention rate after the three malfunctions is faster than that after the five malfunctions.
4. Effects of occurrence of a single malfunction or a small number of discrete malfunctions
   4a. If the subject observes the malfunction(s) for the first time, the intervention rate at the high level of error increases.
   4b. The intervention rate does not always increase, even when subjects receive another malfunction, if they have already experienced malfunctions several times.
5. Comparison between levels of experience for three continuous malfunctions
   5a. In both cases, the intervention rate becomes greater at the intermediate or high level of error.
   5b. If subjects have a lot of experiences, they tend to rely on the automation when the error is small. On the other hand, the subjects tend to choose the manual setting if they have little experience. The major factor of this observation would be the difference in the number of experiences of malfunctioning.

The above findings indicate the following:

1. If continuous malfunctions occur, the intervention rate increases not only around the high level of error, but also at low level of that. The greater the number of continuous malfunctions, the longer the effect lasts.
2. A single malfunction or a small number of discrete malfunctions may increase the intervention rate around the high level of error, but this effect may not last long. As operators gradually experience more individual malfunctions, they appear to become less sensitive to the malfunctioning.

5. CONCLUDING REMARKS

The model-based analyses have reconfirmed that the dynamics of trust in automation depends on the occurrence pattern of the malfunctions. The results of the present study suggest that the reliability of the automation or the frequency of malfunction alone is not always sufficient to determine the level of trust. Thus, if the subjective feeling of trust in automation is described by a probabilistic measure, a belief updating method such as Bayes' theorem is inadequate for modeling of a mechanism of the trust dynamics. With Bayes' theorem, the total effect of the malfunctions is not dependent on the occurrence pattern. Thus, another method is required for describing the cognitive mechanism of trust. The theory of evidence [14] is one possibility.

As shown in the present study, the model-based analyses can be a useful tool for clarifying when operators tend to choose the manual mode. Continuous malfunctioning, which has a strong impact on subjective feelings, caused operators to intervene even when the condition were appropriate for auto mode. This can be regarded as distrust. Thus, this model-based approach can link trust and distrust or overtrust. For example, if an operator relies on the automation in a situation for which the automation is not designed to work, this is deemed overtrust. Further research should be performed in order to clarify how operators begin to overtrust the automation.

6. ACKNOWLEDGMENTS

This work has been supported in part by Grants-in-Aid for Scientific Research 09558047 and 11780299 of the Japanese Ministry

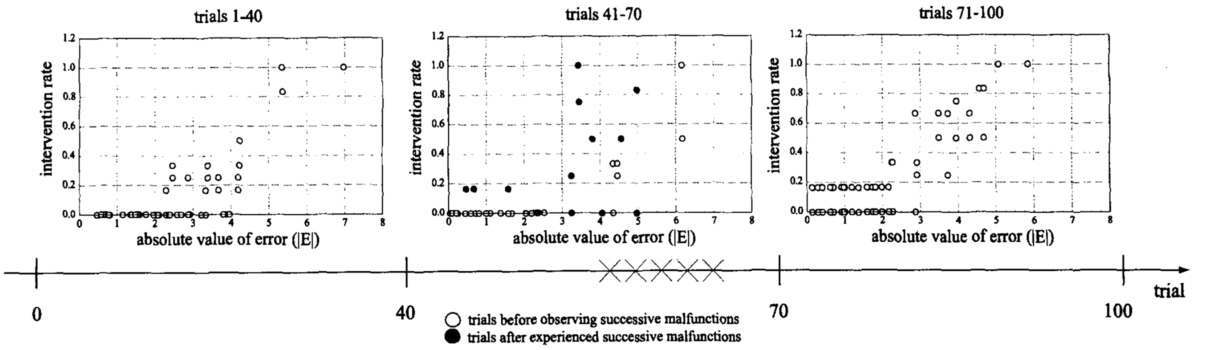

Fig. 8 Dynamic Shift of Intervention Rate (Continuous Group)

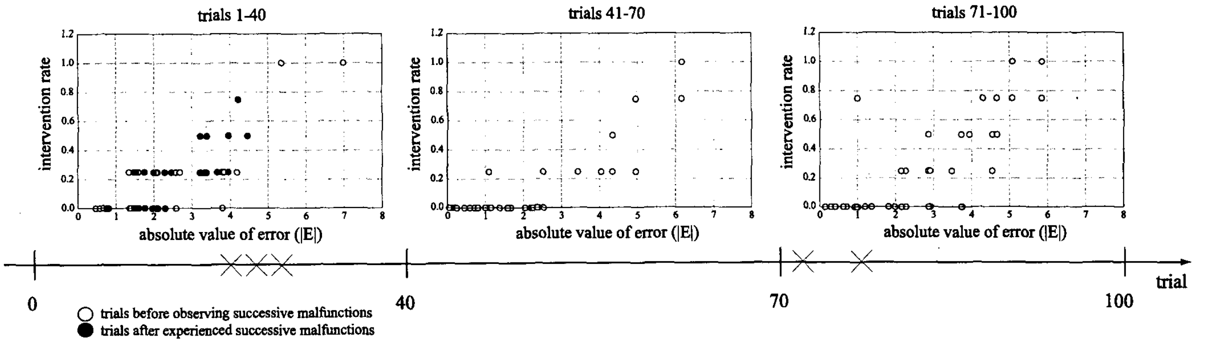

Fig. 9 Dynamic Shift of Intervention Rate (Continuous-Discrete Group)

of Education, Science, Sports and Culture, and the Center for TARA (Tsukuba Advanced Research Alliance) at the University of Tsukuba.

7. REFERENCES

[1] Abe, G., M. Itoh, K. Tanaka (to appear). Occurrence Patterns of Malfunctions and Dynamics of Operator's trust in Automatic Systems, *Proc. 38th SICE Annual Conference* (in Japanese).

[2] Billings, C. (1997). *Aviation Automation*, LEA.

[3] Itoh, M., T. Inagaki, N. Moray (to appear). Trust in Situation-Adaptive Automation for Systems Safety, *Trans. SICE* (in Japanese).

[4] Lee, J., N. Moray (1992). Trust, Control Strategies and Allocation of Function in Human-Machine Systems, *Ergonomics*, 35(10), 1243-1270.

[5] Lee, J., N. Moray (1994). Trust, Self-Confidence, and Operator's Adaptation to Automation, *International Journal of Human Computer Studies*, 40, 153-184.

[6] Molloy, R., R. Pasaruraman (1996). Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects, *Human Factors*, 38(2), 311-322.

[7] Moray, N., T. Inagaki, M. Itoh (to appear). Situation Adaptive Automation, Trust, and Self-Confidence in Fault Management of Time-Critical Tasks, *Journal of Experimental Psychology: Applied.*

[8] Muir, B. (1994). Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automation Systems, *Ergonomics*, 37(11), 1905-1922.

[9] Muir, B., N. Moray (1996). Trust in Automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation, *Ergonomics*, 39(3), 429-460.

[10] Parasuraman, R., R. Molloy, I. Singh (1993). Performance Consequences of Automation Induced "Complacency," *The International Journal of Aviation Psychology*, 3(1), 1-23.

[11] Parasuraman, R., V. Riley (1997). Humans and Automation: Use, Misuse, Disuse, Abuse, *Human Factors*, 39(2), 230-252.

[12] Rouse, W. (1991). *Design for Success*, Wiley.

[13] Riley, V. (1996). Operator Reliance on Automation: Theory and Data, in R. Parasuraman and M. Mouloua (eds.), *Automation and Human Performance*, 19-35.

[14] Shafer, G. (1976). *The Mathematical Theory of Evidence*, Princeton University Press.

[15] Sheridan, T. (1992). *Telerobotics, Automation, and Human Supervisory Control*, MIT Press.

[16] Woods, D. (1989). The Effects of Automation on the Human's Role: Experience from Non-Aviation Industries, *In Norman & Orlady (eds.), Flight Deck Automation: Promises and Realities*, NASA CP-10036, 61-85.

[17] http://flightdeck.ie.orst.edu/scripts/dbsqldev/eviaccident.idc?Study_ID=1#summary.

[18] http://www.rvs.uni-bielefeld.de/publications/Incidents/DOCS/ComAndRep/A330-Toulouse/Rapport.html.

[19] ICAO (1990). Airbus A320-231, VT-EPN, accident at Bagalore, India, *ICAO Circular 263-AN/157*, 111-163.