Automation surprise looked at from a Demands-Resources Perspective

Karel Hurts & Robert Jan de Boer
Amsterdam University of Applied Sciences,
The Netherlands

Automation surprise (AS) is usually seen as a sign of the breakdown of pilot-aircraft interaction. In attempt to resolve several conflicting findings with respect to the precise relationship between pilot workload, degree of automation (DoA), and the frequency of experiencing AS, it was hypothesized that the average AS-rate (number of AS-occurrences per flight – or per unit time – and per pilot) depends on the specific way in which Elapsed Flight Duty Period (seen as a type of “demands”) combines with DoA (seen as a type of “resources”), rather than on each of these two factors considered on their own. Specifically, the average AS-rate was expected to be higher for non-matching than for matching combinations (both being high or both being low) of DoA and Elapsed FDP. This hypothesis was based on psychological arousal theory, signal-detection theory, and general research findings pertaining to the development of automation trust during human interaction with automated systems. Data collected in a survey held among 200 airline pilots just failed to confirm the hypothesis. However, the average AS-rates that were observed were in the expected direction. In the discussion, the theoretical implications of this finding will be addressed.

Introduction

Automation surprise is a phrase that first appeared in the aviation literature in the 1990’s (Woods et al., 1994; Sarter et al., 1997). Dekker (2009) defines automation surprises as those cases where:

a) “The automation does something …
b) … without immediately preceding crew input ...
c) … related to the automation’s action, …
d) … and in which that automation action is inconsistent with crew expectations.”

Note that the discrepancy to which this definition refers may have been present already for a while before the pilot becomes aware of it. This is similar to the phenomena of inattentional blindness, automation-related complacency, and automation bias (De Boer, 2012; De Boer et al., 2014; Parasuraman & Manzey, 2010).

In the existing literature, automation surprise is often associated with loss of situation awareness under conditions of high cockpit automation (Operator's Guide to Human Factors in Aviation, 2014; Optimum Use of Automation, 2006). From this point of view, automation surprise is considered an undesirable phenomenon because of the risk of losing aircraft and flight control and, ultimately, the risk of operational safety hazards.

However, available research shows that the phenomenon of automation surprise (AS) cannot be explained or functionally understood in a simple way, involving only a single or a few factors. The following list of research findings illustrates the ambiguity that surrounds attempts to understand the relationship between amount of workload, degree of cockpit automation, and behavioural phenomena such as automation surprise, complacent pilot behaviour, and pilot situation awareness.

a) Complacent pilot behaviour (i.e., missing important signals from the environment and from the cockpit instruments due to inattention) may be associated with high workload (Parasuraman & Manzey 2010), but also with low workload (Sarter, 2008; Norman, 1990; Matthews & Desmond, 2001).

b) With higher degrees of automation, often poorer situation awareness is observed, but superior situation awareness has also been observed, compared to lower degrees of automation (Kaber & Endsley, 2004; Onnasch et al., 2014).

c) In a previously conducted AS-study (Hurts & De Boer, 2014) it was found that higher amounts of external workload are sometimes associated with lower (i.e., not-expected) frequencies of experiencing AS.

d) In the same study, it was found that degree of cockpit automation was not significantly correlated with the frequency of experiencing AS, despite the fact that higher degrees of automation seem to offer more opportunities for experiencing AS.

In an attempt to understand the seemingly conflicting findings regarding the relationship between degree of automation, amount of external workload, and the frequency of experiencing AS (see points c and d above), a different perspective on the nature and function of AS was developed. As will be seen below, this perspective is based on psychological arousal theory, as well as on signal detection theory. It is also based on existing research concerning the way in which pilot trust and mistrust in automation develops.

**Problem statement and hypothesis**

*Step 1: the goal of optimizing psychological arousal*

One theory that combines the notions of amount of external workload and degree of automation in a single construct is psychological arousal theory. From the research that has been devoted to this theory, it follows that the pilot does not just attempt to minimize his effective workload (or arousal level), but rather tries to *optimize* it
automation surprise seen from a demands-resources perspective

(Young & Stanton, 1997; Wilson & Rajan, 1995; Matthews & Desmond, 1997). In the present study, pilot arousal level is seen as being determined by the combination of current degree of cockpit automation - seen as a type of “resources” - and Elapsed Flight Duty Period (FDP) – seen as a type of “demands”. (In this article, degree of automation – or DoA - will be defined as the complexity of the flight control mode, see Table 1 for further details.) Specifically, if the determining factors are both high or both low (are matching), the arousal level can be considered to be optimal. Otherwise (if these factors are not matching), it can be said that there is overarousal or underarousal.

Usually, the pilot has no direct control over Elapsed FDP (i.e., the number of hours he/she has been working without interruption). Therefore, under conditions of over- or underarousal he/she can influence his current arousal level only by adjusting the current DoA (see step 3 below for the details).

**Step 2: detecting an automation-pilot conflict as trigger for testing automation trust**

On a different note, it is likely that DoA is also used by the pilot to calibrate his/her current level of trust in the cockpit automation. From the literature on the importance of trust in semi-automated working environments (see, e.g., Bass & Pritchett, 2008), it can be expected that automation trust must occasionally be tested in order to build and maintain it, or, if unavoidable and necessary, to (temporarily) reduce it. For example, automation distrust may arise due to the automation being intransparent to the pilot. This may, in turn, cause him/her to (temporarily) reduce DoA. It is proposed that an obvious trigger for conducting such tests is formed by the detection and conscious experience of a conflict between expected and actual automation behaviour. Specifically, during a test phase the pilot attempts to identify the cause of the conflict, and, if necessary, adjusts the current DoA accordingly (i.e., choose more automation or less automation, depending on the relative amount of trust the pilot has in him-/herself as pilot and the automation).

**Step 3: increasing the importance of arousal considerations during the test phase**

It is at this point during the test phase that arousal considerations come into play. Obviously, these considerations have to be somehow reconciled with performance-related and safety-related considerations. It is proposed that a natural way for the pilot to increase the importance of arousal considerations under conditions of over- or underarousal is to *lower the threshold for detecting a conflict between expected and actual automation behaviour*. This proposition is based on the general logic of

---

2 “Detecting a conflict” should be compared to detecting a signal, as described by signal detection theory (SDT). As is the case for signals in SDT, it is assumed that conflicts occur in a noisy environment, containing many other types of events that may suggest that there is a conflict. Though the pilot cannot perceive a conflict directly, he/she can statistically weigh the evidence supporting the existence of a “true” conflict. As in SDT, the pilot can make two types of error regarding the detection of a conflict. A false alarm occurs if the pilot only believes that there is a conflict, whereas, in fact, there is none. For example, the pilot mistakes some piece of innocent automation status information for an alert
signal-detection theory as follows: as a result of lowering this threshold, more conflicts will be detected (on the average) in a fixed time period, compared to when the threshold remains unchanged. This, in turn, has the effect that more tests will be conducted in the same time period, and, eventually, that more opportunities will be created for changing the current DoA.

Implications of the three-step process

Note that there is no guarantee that this strategy for influencing DoA will always result in an improvement of the pilot’s arousal level. Nonetheless, in cases of over- or underarousal lowering the conflict detection threshold seems to be an effective strategy for influencing the probability that pilot arousal level will shift in the direction of optimal arousal. This expectation is supported by studies that show that problematic pilot-automation interactions may occur if DoA remains too high during an extended period of flight time, as illustrated by the phenomenon of automation overreliance (also referred to as automation bias or complacent pilot behaviour). In terms of our model, overreliance becomes a risk if a high DoA is combined with a low Elapsed FDP (signaling underarousal). The reversed combination of a high Elapsed FDP with a low DoA (signaling overarousal) is also known to be associated with problematic interactions, as illustrated by the phenomena of automation underreliance, automation disuse or non-conforming pilot behaviour (Parasuraman, 1997).

Finally, note that the detection of a conflict between expected and actual automation behaviour may also have other causes and implications than those discussed above. For example, learning by the pilot from previously resolved conflicts is likely to affect future pilot-automation interaction. Also, it is likely that the detection of a conflict is influenced by the size of the discrepancy, as well as by the frequency with which conflicts have occurred in the past (see also De Boer, 2012). However, in this article, the role played by these additional factors will not be further discussed.

Assuming that each detected conflict results in an automation surprise (AS), the following hypothesis can be derived from the previous discussion:

Hypothesis

If DoA at the time of the last AS and the Elapsed FDP at the same time do not match, the frequency of experiencing AS is higher compared to when they match (interaction between DoA and and Elapsed FDP with respect to AS-frequency).

In order to test this hypothesis, the survey data that were described in the study of Hurts and De Boer (2014) were re-analysed. It was assumed that Elapsed FDP signaling unexpected danger. A miss occurs if the pilot ignores or somehow fails to detect a “true” conflict. Such errors might be due to inattentional blindness.
would (partly) reflect the build-up of pilot fatigue, which, according to the literature (Stanton & Young, 2000), can be considered to be one aspect of mental workload.

Method

Participants and procedures

For this study, the data were used that were collected in the 22-question survey described by Hurts and De Boer (2014). Two hundred pilots participated in the survey, most of whom were recruited through Crew Center of KLM, the VNV (Dutch Association of Airline Pilots), and National Aerospace Laboratory NLR. Most respondents filled in the web version of the survey. It took them from 20 to 30 minutes. Though the survey was filled in anonymously, a few questions were included in order to verify that the respondents were really airline pilots (e.g., respondents were asked about the various aircraft they had been flying and they had to indicate how they had been approached for their participation). An automation surprise (AS) was briefly explained to the participants in terms of a few typical pilot reactions to automation behaviour. Participants were required to describe their last AS, as well as provide information about several (predefined) accompanying circumstances. Only a subset of the twenty-two questions were of direct interest to the present study, as will be explained below.

Design

Dependent variable

The frequency with which an AS occurred for any participant was measured in two different ways:

AS-frequency score 1 (flight-based frequency measure): this score (one per participant) was defined as the fraction of the total number of operated flights on which an AS was experienced by a participant. It was estimated on the basis of the answers given by the participants to the survey question “How many flights ago was your last automation surprise?”, as follows:

\[ \text{Number of flights since last AS} \]

\[ \text{can be seen as somehow estimating the period of the frequency with \ which AS is experienced during a flight (here, period is defined as the number of consecutive non-AS-flights separating two AS-flights). Let’s call this estimator } \text{#NASF}. \]

However, because the time at which the survey was filled in was assumedly chosen at random by the pilot and/or researcher, the period must, on the average, have passed only for 50% at the time of the measurement of #NASF. Therefore, and because frequency is the inverse of the period, the term \[ 2 \times \text{#NASF} \] appears in the denominator of the formula. The constant 0.5 is added to this term as a means to correct for the fact that the participants most likely included the last AS-flight in their count of the number of flights since their last AS-flight. This has resulted in overestimations of the values for period. This “counting error” can occur only once in a period: hence, the value of \[ (2 \times (0.5)) \] in the denominator of the formula.

---

3 Number of flights since last AS can be seen as somehow estimating the period of the frequency with which AS is experienced during a flight (here, period is defined as the number of consecutive non-AS-flights separating two AS-flights). Let’s call this estimator #NASF. However, because the time at which the survey was filled in was assumedly chosen at random by the pilot and/or researcher, the period must, on the average, have passed only for 50% at the time of the measurement of #NASF. Therefore, and because frequency is the inverse of the period, the term \( 2 \times \text{#NASF} \) appears in the denominator of the formula. The constant 0.5 is added to this term as a means to correct for the fact that the participants most likely included the last AS-flight in their count of the number of flights since their last AS-flight. This has resulted in overestimations of the values for period. This “counting error” can occur only once in a period: hence, the value of \( (2 \times (0.5)) \) in the denominator of the formula.
\[
    f(AS_{-1}) = \frac{1}{2 \times ((\text{Number of flights since last AS}) - 0.5)}
\]

**AS-frequency score 2 (time-based frequency measure):** this score (again, one per participant) was defined as the average number of AS-flights experienced by a participant per month. It was computed by multiplying the outcome of the above-mentioned formula by the answer given by the pilot to the question: “How many flight do you operate in a month, on the average?”.

It should be noted that both AS-frequency scores yield slight overestimations of the “real” AS-rate because no scores could be computed for participants who never had experienced an AS (see under Results for more details).

**Independent variables**

**Degree of automation (DoA)** was assessed using a seven-point scale, with scoring categories ranging from “No automation” to “Full automation” (see Table 1). The categories were designed in a post-hoc fashion by an experienced flight instructor based on the open-answers given by the participant to the question “What flight mode were you in at the moment of the last automation surprise?” (one score was computed for each participant).

**Elapsed Flight Duty Period (FDP)** refers to the number of hours the participant had been working without interruption at the time of his/her last AS. (Participants were asked to choose one among several numerical time intervals.)

**Analyses**

Data were analysed using multiple regression analyses in which DoA and Elapsed FDP were entered as predictors, and frequency of experiencing AS was used as dependent variable.

The interaction between DoA and Elapsed FDP was entered into the regression analyses as a third predictor - and was used for testing the hypothesis. This was done as follows:

a) Elapsed FDP was first dichotomized, giving the values high (1) and low (-1), depending on whether the participant’s score was above or below the average Elapsed FDP of 5.5 hours. This was done in order to end up with an easy-to-interpret interaction.

b) For each participant, the product between DoA (in mean-centered form) and Elapsed FDP was computed and this term was entered as a separate predictor in the regression model. Note that the product term was an ordinal scale variable with both negative values (corresponding to non-matching combinations) and positive values (corresponding to matching combinations).
c) The hypothesis would be confirmed if the effect of DoA × Elapsed FDP on the dependent variable was statistically significant \((p < 0.05)\) and if higher product values (corresponding to matching combinations of DoA and Elapsed FDP) were associated with lower AS-frequency values (on the average) than lower product values (non-matching combinations).

Table 1. Seven scoring categories, and their associated frequencies, for measuring complexity of flight control mode (degree of automation). Based on answers to open survey question and measured on an ordinal scale. 1 = lowest, 7 = highest degree of automation.

<table>
<thead>
<tr>
<th>Complexity of flight control mode (degree of automation)</th>
<th>% of valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FD ON, MANUAL FLIGHT</td>
<td>4.8</td>
</tr>
<tr>
<td>2 AP OFF, AT ON, FD ON MANUAL FLIGHT</td>
<td>0.6</td>
</tr>
<tr>
<td>3 AP ON, AT OFF, MANUAL SELECT</td>
<td>0.6</td>
</tr>
<tr>
<td>4 AP/AT ON, MANUAL SELECT (HDG, VOR/LOC, VS)</td>
<td>18.7</td>
</tr>
<tr>
<td>5 AP/AT ON, FMS GUIDANCE SINGLE (HOR./VERT.)</td>
<td>6.6</td>
</tr>
<tr>
<td>6 AP/AT ON, FMS GUIDANCE DUAL, APPR. MODE</td>
<td>66.9</td>
</tr>
<tr>
<td>7 AUTOLAND</td>
<td>1.8</td>
</tr>
<tr>
<td>Total valid</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Results

Some demographics

Of all respondents, 96% was male, 54% was in the rank of captain, and 42% was in the rank of first officer (the balance is in the rank of second officer). With regard to aircraft type currently operated, respondents mentioned Boeing 737NG, Airbus A330, Boeing 777, Embraer 170/190, and Fokker 70/100 as the aircraft types flown most frequently. This reflects the fact that most respondents were employed by KLM, the fleet of which is primarily composed of the above-mentioned planes.

The average age of the respondents was 38 years, with a range from 23 to 58 years, \(sd = 9.63\) years. Moreover, the mean value for amount of flying experience was 8867 hr, \(sd = 5480\) hr, with a range from 750 hr to 27500 hr. Finally, the average number of flights per month was 22.8, with a range from 3 to 43 flights, \(sd = 15.09\) flights.
In the analyses mentioned below, DoA was treated as an interval-level variable, even though, strictly speaking, it was only an ordinal scale variable.

**Frequency of experiencing AS**

The frequency of experiencing AS could only be computed for 186 (93%) respondents. These were the respondents who had indicated that they had at least one AS-experience. Therefore, both frequency scores provided slight overestimations of the “real” frequency with which AS’s occurred. The average value for AS-frequency score 1 was 0.08 - or 8% AS-flights - , median = 0.03, sd = 0.13. This was based on an average value for number of flights since the last AS of 71, median = 20, sd = 170. The average value for AS-frequency score 2 was 1.44 flights per month, median = 0.40, sd = 2.95.

It turned out that neither AS-frequency score was normally distributed (both were skewed to the right). Therefore, in the analyses mentioned below, both scores were first subjected to a log10-transformation. After transformation, both transformed scores passed the K-S-normality test at a 0.05 significance level.

**Interaction between DoA and Elapsed FDP**

Figure 1 shows the average values of AS-frequency score 1, broken down by Elapsed FDP (high versus low) and by DoA. Regression analysis showed that the interaction between DoA and Elapsed FDP just failed to reach the level of significance, $t(159) = 1.84, p = 0.07$, but was in the expected direction. The two main effects (one for DoA, the other for Elapsed FDP) were not significant either, $p > 0.10$.

Figure 1 suggests that the expected interaction was stronger for low degrees of automation. A post-hoc analysis revealed that, for the lowest four degrees of automation (less than or equal to the median rank of 4), the difference between the low and high Elapsed FDP-groups was significant, $F(1,38) = 5.00, p < 0.05$, partial $\eta^2 = 0.12$ (one-way analysis of variance).

The regression analysis belonging to AS-frequency score 2 revealed similar results: the interaction between DoA and Elapsed FDP was again almost significant, $t(158) = 1.93, p = 0.06$. The mean frequency scores generally followed the same (expected) pattern as in Figure 1.

---

4 With regard to the main effect of Elapsed FDP, it should be noted that in Figure 1 the FDP-base rate was not taken into account (this is the probability with which the various original FDP-values occur, irrespective of whether or not an AS is observed). Follow-up analyses show that, after being corrected for this base rate, the average probability of experiencing an AS for the two highest FDP-intervals (8-11 hours, that is) becomes significantly higher than that for the remaining FDP-intervals (Hurts & De Boer, 2014).
It is concluded that, though the hypothesis could not be confirmed, the average values for AS-frequency scores 1 and 2 showed a trend in the expected direction: scores were lower if DoA and Elapsed FDP were matching than if they were not matching. The difference between matching and non-matching combinations was particularly salient for the four lowest degrees of automation.

Conclusions and Discussion

In this study, the hypothesis was tested that the frequency of automation surprises depends on the extent to which the current degree of cockpit automation (DoA), seen as a type of “resources”, matches the current value for Elapsed FDP, seen as a type of “demands”. It turned out that the average frequency of experiencing an automation surprise (AS) was lower if DoA (assessed at the time of the last AS) matched Elapsed FDP (assessed at the same time), compared to a non-matching combination. Though this effect was expected, it just failed to reach the level of statistical significance.

Figure 1. Average values for AS-frequency score 1, broken down by DoA and Elapsed FDP. Dashed lines represent best-fitting regression lines. Y-axis values deliberately shown on a logarithmic (base 10) scale.

The absence of any difference in average AS-frequency between low Elapsed FDP-pilots and high Elapsed FDP-pilots under conditions of high cockpit automation needs explanation. Perhaps, on some short-haul flights (i.e., flights with a duration of less than 6 hours – precisely the duration that corresponds to shorter-than-average values for Elapsed FDP), there are operational constraints that require the pilot to
continuously fly with high cockpit automation. In that case, and following the rationale of this study’s hypothesis, it is not likely that the pilot will detect and resolve more automation-pilot discrepancies than under conditions of low automation. This would explain the pattern of results observed in Figure 1. However, this post-hoc explanation should be treated with care and further research is needed to investigate it.

Severity of AS-consequences

In terms of signal-detection theory, lowering the threshold for detecting a conflict can be expected to result (in the long run) in more false alarms: cases where it is incorrectly assumed by the pilot that a conflict has been detected (more liberal response bias). Therefore, the strategy of lowering the conflict detection threshold is expected to generate less severe AS-consequences (non-matching combinations of DoA and Elapsed FDP), on the average, compared to the situation where the threshold is higher (matching combinations, more conservative response bias). This latter expectation was tested in a post-hoc analysis of the statistical interaction between DoA and Elapsed FDP with respect to the self-reported severity of the consequences of the last AS. Severity of AS-consequences was assessed using a six-point scale, with scoring categories ranging from “No consequences” to “Damaged aircraft”. In other respects, the expectation was tested in a way similar to that used for testing the main hypothesis of this article.

Figure 2 shows the average values of Severity of AS-consequences, broken down by Elapsed FDP (high versus low) and by DoA. It turned out that the interaction between DoA and Elapsed FDP was not significant, $p > 0.10$. As can be seen, the contrast between high Elapsed FDP and low Elapsed FDP was only in the expected direction for the highest degree of automation (higher Severity of AS-consequences-scores for the high Elapsed FDP-group). The two main effects (one for DoA, the other for Elapsed FDP) were not statistically significant either, $p > 0.10$.

It is concluded that the data on Severity of AS-consequences do not provide evidence for the expectation that the effects of DoA and Elapsed FDP can be understood in terms of more or less incorrectly detected conflicts (false alarms) - depending on whether the response bias for detecting conflicts took on a more liberal or a more conservative value, respectively.

Implications

Regardless of the fact that the findings of this study do not allow very strong conclusions, future researchers should not discard more traditional ways of looking at AS. For example, it is still likely (as others have suggested) that experiencing AS is a sign of a vulnerability in automation-pilot interaction. At the same time, researchers are encouraged to investigate more fully the possibility of additional purposes being served by the experience of AS – an issue that has only been touched upon in this study.
In addition to attempting to validate the central ideas of this study under different and better controlled circumstances, future research should address the following questions:

1. Will the main hypothesis be confirmed if the “demands” affecting the pilot (now assessed by means of Elapsed FDP) are assessed in different ways?
2. How precisely does automation surprise affect the current mode of cooperation between pilot and automation?
3. How does learning from previously explained and resolved conflicts affect future pilot-automation interaction?
4. Are there other ways (besides lowering the conflict detection threshold under conditions of over- or underarousal) in which the pilot can improve his arousal level?
5. What role will automation surprise play during pilot-automation interaction if it is considered as an experience with variable intensity (i.e., a pilot can be more or less surprised about automation behaviour)?
References


