



Towards an explanatory, predictive account of awareness

Rachid Hourizi*, Peter Johnson

Department of Computer Science, University of Bath, Claverton Down, Bath BA2 7AY, UK

Abstract

In this paper we investigate the utility of an explanatory and predictive account of awareness in the (re)design of complex human–computer interaction. We start by describing both the wider field of human computer awareness and the specific sub-field of situation awareness in the aviation domain. We also discuss the wide range of observations and descriptions, which detail both success and failure in the acquisition, maintenance and repair of such awareness across a range of domains. Next, we assert the need for an explanatory and predictive model of the phenomenon in the search for a reduction in the breakdowns reported and propose such a model. We then go on to investigate the utility of our model as a guide for design through the discussion of a recent experiment involving manipulations of an animated warning signal on a simulated cockpit control panel. Our results show initial support both for the model and for our earlier assertion of its utility. We conclude that our composite view of awareness yields practical benefit in the design of human computer awareness support and provides a basis for future research.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Awareness; Situation awareness; Predictive model; Design

1. Problem review

As computer systems have become embedded in complex, dynamic environments with increasing numbers of users, objectives and potential information sources, human–computer interaction (HCI) designers have been forced to confront a number of challenges, whose effects were, in many cases more muted in the more restricted domain of desktop computing. Prominent amongst these challenges is that of human computer awareness (HCA)—a growing problem in a number of complex domains, where groups of people are often involved with multiple automated systems in the pursuit of multiple objectives. Medical practitioners, for example, are noticing a growing group of failures, which stem, at least in part, from a breakdown in practitioners' awareness of the state of their own computerised tools [1]. Beyond the medical domain, similar problems have

been observed in both process control [2] and naval [3] settings.

In fact, this growing problem of breakdowns in the people's awareness of their increasingly powerful, computerised tools goes beyond the safety critical domain, through administrative and office settings [4] to home and leisure activities such as visiting art galleries [5], using public transport [6] and even driving a car. Canadian research, for example, shows that drivers are often unaware of freight trains approaching a rail/road crossing [7].

This breadth of domain and context in which awareness failures occur is reflected in a growing level of interest in a widespread research community. Various threads of investigation are emerging, each addressing a particular sub-strand of the issues surrounding HCA and the design support it requires. A growing body of research, for example, is now going into the study of *peripheral* awareness [8,9], where the task being supported may be secondary to a more important work objective and, as a consequence, the intrusive, dominant alerts used in a safety critical setting may be less suitable. In a related but largely separate initiative, CSCW researchers have moved beyond the design of office

*Corresponding author. Tel.: +44-1225-384427; fax: +44-1224-383493.

E-mail addresses: r.hourizi@bath.ac.uk (R. Hourizi), p.johnson@bath.ac.uk (P. Johnson).

tools which promote colleagues' awareness of each other, through the *mediation* of automated tools [10] to investigate and prototype "contextually aware" automation where the objective is to imbue the automation itself with some model of worker activity and availability [11].

In this paper, we will focus on a specific sub-category of the wider problem space—the highly specialised domain of the commercial aircraft cockpit. The acquisition, maintenance and repair of awareness in this context (often referred to as "situation awareness") appear particularly difficult to achieve and can be devastating when lost. A large number of well documented examples exist in which competent, experienced pilots seem to have been unaware of crucial flight details, such as proximity to mountains [12], impending weather conditions or the poor condition of their target runway [13]. In many cases, this lack of awareness has resulted in serious incidents or crashes, with resultant loss of mission, aircraft and human life. Examples of this error class can be found in the official reports on the Air Inter Strasbourg crash in 1992 [12], the 1995 loss of an American Airlines aircraft near Cali in Colombia [14] and the Qantas accident at Bangkok [13].

As we investigate this complex area, it becomes clear that, as with the other domains described at the start of this section, an awareness of computerised system state and activity is particularly problematic. Numerous reports and studies spell out the ongoing concern of active pilots that they are frequently unaware of both current and future system activity or often the logic behind either (e.g. [15]).

In sum, therefore, the phenomenon of system awareness in the cockpit exists within a wider (and growing) context of awareness *of, through* and perhaps even *by* automation. Across this wider context, a series of higher-level questions crop up time and again: What is awareness? Why does it break down? How can we design interaction such that the most severe forms of breakdown become less likely? This paper will address domain and context specific awareness failures occurring between pilots and autopilot in common commercial aircraft. Behind these specific questions and investigations, however, we will both draw from and, wherever possible, feed back to the wider community interested in the design of computer systems, which support the acquisition, maintenance and repair of user awareness in its various forms, and in the fast moving multi-agent, multi-task, environments mentioned at the start of this section.

2. Current literature and proposed extension

For many of the relatively new strands of research, described above, the current state of their art consists of

detailed descriptions of both successful and failed interaction between human and computerized tool (e.g. [16]). Beyond these observational studies, a second body of research findings are provided by experimental prototypes (e.g. [18]) designed with one eye on execution of a particular task or activity and a second on broadening the community's understanding of awareness and the nature of the support required if it is to be maintained. For many of these research groupings, however, Olson and Olsen's comments [20] on the current state of affairs in CSCW can be appropriated for a summary of current awareness research:

"The field of CSCW has mostly been at the stage of building point systems. Many more systems have been built than have been evaluated. Some attempts at understanding the dimensions by which the systems and impacts vary have been proposed, but the effort to turn these dimensionalizations into deeper understanding are scattered and inconclusive"

Our own chosen sub-field of (aviation) situation awareness, however, benefits both from a huge body of reported accidents and incidents (many of which involve HCA breakdowns) and from a well-developed research program dating back several years (e.g. [21]). It has, therefore been developed further than many of the areas on our list to include descriptive frameworks, within which the problem of HCA can be discussed. These frameworks are often quite detailed in their description both of the information, which must form part of the ultimate *state* of awareness, and of the *processes* by which it can be achieved. Sarter and Woods [22] for example, draw upon their own observational studies to assert the importance of notions including perception, attention and knowledge to the acquisition, maintenance and repair of awareness. In parallel research Endsley [23] adds anticipation to the mix, citing the pilots' need to be "ahead of the plane" in the complex, rapidly changing environment of the cockpit, whilst Gaba et al. [1] describe the subtle cues on which awareness can often rest in the field of anaesthesiology.

In fact, in the specialized area of mode error, situation researchers have even gone so far as to generate a sufficiently deep explanatory model of system behaviour that they have been able to predict one category of future awareness failures (or more specifically identify system states which are prone to such errors). Rushby for, example, has developed and tested a formal method, which identifies gaps in the system-information provided to pilots' and, therefore potential breakdowns in their awareness of automation state [24].

This formal approach is not, however, (and never claimed to be) a universal solution to the wide range of situation awareness failures reported. We have demonstrated in previous work [25], for example, the intuitive notion that the availability of pertinent information in an environment is not, in itself, sufficient to guarantee

that people will become aware of it. Work remains then to continue to build explanatory and, more importantly predictive models of awareness in the areas not caused by a lack of information. In other words, we still need an explanatory and usefully predictive account of breakdowns where information is available but overlooked.

With this in mind, we propose such a model for HCA supporting design, drawing on and extending the findings discussed above. Our starting point for this requirements model is the very ambiguity inherent in the term “awareness” itself. In everyday speech, for example, the seemingly straightforward question “Were you aware that the kitchen light was on?” can vary, according to context. In the simplest case, this can mean simply “Did you see that the kitchen light was on?” In a context both questioner and questionee are clearly looking at a scene in which the light is obviously visible, however, the meaning of the question mutates to become “Did you pick out the presence of the kitchen light from amongst the many elements of the scene that you undoubtedly saw?”—a question which is roughly analogous to the question “did you attend to the fact that the kitchen light was on?” This is not yet, however, the end of the story, in yet another context, the question could carry implications for the recipient of the question in the form of a required action or response. In this context, then the question becomes equivalent to “Did you understand the implicational meaning of the fact that the light was on?”

Building from the threads of previous research, therefore, we believe a composite model is needed, in which the information in the environment must pass through a series of cognitive processes, before it can be considered to form part of a person’s awareness—a notion also inspired by, if not directly built upon the work of Barnard and May in their work on interacting cognitive subsystems [26]. We propose, therefore a definition of the state of awareness in which raw data from the environment has been (1) available (in line with Rushby’s models), (2) perceived, (3) attended to in some manner and (4) subject to further, higher level cognitive processing.

We can immediately use this model to provide insight to the variations in our kitchen light example. The first question “Did you see that the kitchen light was on?” can be thought of as a question about perceptual or level (2) awareness. The second question “Did you pick out the presence of the kitchen light from amongst the many elements of the scene that you undoubtedly saw?” becomes a question of whether level (3) or attentional awareness has been achieved and the last question “Did you understand the implicational meaning of the fact that the light was on?” refers to the higher level cognitive (often semantic) processing inherent in level (4) awareness.

More importantly, however, we are also in a position to describe cockpit awareness breakdowns not only in terms of the final awareness desired, but also in terms of

the particular sub-process, which failed. We could, for example, imagine separate breakdowns in which information was available but not seen (a level (2) failure), picked up visually (or aurally, etc.), but overlooked, because the pilots attention was elsewhere (a level (3) failure). We could extend this notion of process related levels of awareness to a situation in which information was available, seen, attended to but not understood in terms of its meaning or implication (a level (4) failure). In other words, in the place of the single, rather clumsy question “Why was the pilot not aware of that issue, we are now able to ask *multiple* focused questions each aimed at a separate potential failure point in the acquisition, maintenance and repair of awareness. By extension, we can use similar questions in the early stages of design in order to predict potential problems as Rushby [24] was able to do at the level of information availability—e.g. Is the users’ attention likely (or in some safety critical cases certain) to be drawn to the relevant information source? Is the information presented in such a way that it can be easily processed by the receiver in terms of the implication for action it is intended to convey?

It seems sensible, for example that a problem at the availability level might require simply the inclusion of supplementary information (as suggested by Rushby [24]), whereas this very action may actually increase the chances of breakdown at the attentional level (the supplementary information could simply distract the pilot from other important information). More subtle errors could, for example, arise in the case that clumsy interface design increased the perceptual strength of a particular indicator (i.e. made it bigger, brighter, louder or more central), without making it relevant to the underlying notion being indicated. In the terms of our levels, the potential error would be likely to occur at the semantic processing level, despite the extra support of the perceptual.

Whilst we could imagine that such an approach, if shown to be practical, could be of benefit across a range of environments and contexts, we must start by establishing a proof of concept both for the validity of the model and for its asserted utility. With this in mind, we will use the rest of this paper to investigate two questions; (1) Can we provide empirical evidence to support our intuitive description of awareness and (2) If so, can we manipulate design elements in an authentic interface to demonstrate its utility?

3. Experiment

Our exploration of these overlapping questions involved the investigation of the well-documented interaction between pilots and autopilot in the Airbus A320, a large commercial passenger aircraft. This

interface (described in more detail in the “Equipment” section below) was chosen exactly because it has been implicated in a number of high profile HCA breakdowns, such as the crash of an Air Inter aircraft near Strasbourg, France in 1992. Investigation of the interfaces involved, therefore, gives us both an authentic arena in which we can ultimately make a direct contribution to a real world problem and an excellent context in which to test our model.

The particular design enhancement we have chosen to manipulate on the other hand is drawn from an entirely different domain—traditional cinematic animation. Specifically, we will attempt to use the notion of anticipatory or predictive movement, to draw user attention to autopilot activity. This technique, described in detail in the seminal book “The Illusion of Life” [27] and considerably more briefly below involves a small counter movement being used as a precursor to an important action in many animated features.

Our first interest in the technique stems from the apparent similarities between (1) drawing viewer attention to a specific area of a complex, rapidly changing cinema screen, shortly before a character performs an action of great importance to the plot and (2) drawing pilot attention to a specific area of a complex, rapidly changing cockpit interface, shortly before the autopilot performs an action of great importance to the ensuing flightpath. We have a second reason for selecting this technique, however, hence the selection of this potential enhancement from the relatively wide category of changes, which could feasibly have affected pilots’ automation awareness; Along with the intriguing possibility that this technique could be usefully transferred to a safety critical domain, we believed that it also had the characteristic that it could be manipulated in line with our suggested awareness model, a notion we will expand below.

Before we go into further detail about this experiment, however, it is, perhaps sensible to recap the higher level questions the empirical research will address.

The first such question surrounds the utility of our approach with regard to the particular awareness failures in our chosen domain and context. “Does our approach yield demonstrable benefit in our chosen domain and context?” Clearly, our chosen methodology, which starts with an explanatory account of the processes of perception, attention and semantic processing is very different from the prototypes described above, driven primarily by an understanding of the domain itself and the “subtle cues” it provides [1]. We see no value in any assertion that our deeper but more general approach can replace domain specific understanding but must, at least, be able to demonstrate that it can add value to or supplement those described earlier. In other words, we must show the value of an approach based primarily on an explanatory

account of awareness rather than domain specific activities.

This leads us to a second, more specific question, “Are we really able to individually target the processes listed and, in each case, affect the level of awareness achieved by our participants?” If we cannot do so, e.g. if we find no way to affect awareness by better supporting the participants’ perception or semantic processing, we will have added little to previous treatments of the subject. Whilst we cannot demonstrate generalisable utility in this, our initial proof of concept, therefore, we must at least provide evidence that our localised approach yields tangible utility.

Finally, we must demonstrate one further ability—the specification of design elements, which support the sub-processes in question. In the context of the work already completed in this area, it would not be sufficient for us to re-categorise the breakdowns and failures observed in our example, nor even to fit them within a strongly plausible causal structure. If we are to claim a real contribution in the area, we must be able to bridge the gap between the deeper understanding constructed and specific design elements utilised in the (re) design of our interaction.

3.1. Overview

The scenario chosen for this investigation was an aircraft descent scenario from a starting altitude of 10,000 ft (a context similar to the one in which the original Strasbourg HCA breakdown occurred). In the course of the experiment, our participants were asked to execute a series of instructions to effect this descent, all the while ensuring that the aircraft was travelling towards an airport, which was below them (!) and slightly to their right. In the course of this scenario, the automation would (unbeknown to the participants) make alterations to the course of the flight, such that the aircraft would start to move *away from* the fictional airport. We will refer to these independent actions as *interventions* on the part of the autopilot.

Next, we added components to the display (described in more detail below) warning the participants that such an intervention had taken place. We also varied the nature of the signal in different conditions in order to examine the different reactions occasioned by each warning type, i.e. we used each condition (aside from the control) to target a different level of awareness from our model.

Finally, we measured the participants’ awareness of the ongoing flight (i.e. “situation awareness”), recording both reported observations that “something unexpected was happening” and subsequent participant activity (if any) to correct the problem. In this way, we started to separate those interventions, which had been seen but not fully understood (indication only that our

“perceptual” level of awareness had been achieved) from those, which had been further processed (indication that our “semantically processed” level of awareness had been reached).

This experiment, then, addressed the three high level questions posed in the previous section since (1) it was based upon a manipulation of the levels of awareness defined in our model, (2) it addressed a well-reported HCA breakdown from the aviation domain and (3) it provided evidence of the utility (or otherwise) of a specific design solution.

3.2. Hypotheses

In order to move towards specific hypotheses, we need to define in a little more detail exactly those manipulations, which we chose to perform. In order to do this, we need to dig a little deeper into the nature of the warning signal, which indicated the autopilot activity. From the brief description given in the previous section, it is clear that the warning involved the addition of a small movement to the interface, shortly before each intervention took place.

If we look closely at this warning signal, however, it seems likely that at least two characteristics are important. The first is the perceptual strength of the signal—the characteristic, which makes it likely that the signal will be picked out from amongst all the competing information available in the display presented to our participants. Intuitively, this characteristic could be manipulated by making the signal larger, brighter or subject to a larger degree of movement, providing us with ample opportunity to manipulate it in the course of our experiment.

The second characteristic of interest was the semantic relevance of the signal to the underlying autopilot activity. It seemed sensible to assume that we could manipulate this characteristic by reducing (or increasing) the cognitive load on the viewer through a reduction (or increase) in the amount of mental work required to map from the incoming signal to its underlying meaning—a notion explored by Johnson and Johnson [19] in the field of task performance, but often overlooked in HCA support. We believed that this reduction in cognitive load would reduce lead to the higher levels of awareness being achieved with greater regularity and, ultimately, reduce the number of breakdowns observed.

This was the point at which our choice, of warning signal—predictive movement—became important. We believed that the craft skill inherent in its development masked an optimisation in viewer cognitive load, since the link between the first movement and the second was easier to process than that between a flashing light or ringing bell and a subsequent action. In other words, an efficient semantic link existed in the connection between

movement in the warning signal and movement in the underlying system (e.g., a change in altitude or heading).

We could therefore develop conditions in our experiment in which we took performance with the existing interface as a first (control) condition (C1) and added two further conditions involving warning signals with different signal characteristics emphasised. In the first of these additional conditions (C2), we would add a warning signal of high signal strength (large, extensive movement) and in the second (C3), we could reduce the raw perceptual strength but increase the semantic relevance of the warning in the manner described above. In this way, we hoped to affect awareness at different levels in each condition. C1 would measure the probability of awareness with the participant armed only with raw environmental data (i.e. they would get no special warning that an intervention was taking place). In C2 and C3, however, we could measure both the frequency with which our participants noticed (a) that any intervention was taking place (i.e. the probability of perceptual awareness only) and the frequency with which they achieved sufficient understanding of the intervention to correct the erroneous flight path in the middle of a busy, fast moving environment (i.e. the probability of higher level, cognitively processed awareness occurring).

If we were to gain support for our model, then, we would need to show that explicit support provided at different levels of awareness would provide tangibly different results, ultimately affecting the extent to which people saw and/or understood the context of the flight. With this in mind, our first hypothesis was that the provision of any warning signal (i.e. an attempt to draw our participants attention towards the interventions) would increase the likelihood that these interventions would be reported.

More clearly stated then, this first hypothesis (H1) becomes:

H1: The inclusion of *explicit* information about autopilot activity will increase the number of reported observations that such activity had occurred, i.e. a significantly greater number of such reports will occur in those conditions where such information is given (C2, C3) than in the condition where it is not, (C1).

Beyond this, however, we were able to make predictions about the likelihood of our participants moving from simply seeing that something had occurred to understanding what it was. Again we can phrase this second hypothesis (H2) more clearly in terms of our experiment:

H2: The inclusion of a specific semantic link between the warning signal given to participants and the underlying autopilot activity will increase the participants’ understanding (cognitively processed awareness) of such activity, leading to a significantly higher rate of correction of those undesirable interventions reported.

I.e. the ratio of interventions corrected to interventions reported will be significantly higher in the condition where explicit semantic support is included (C3) than in those where it is not (C1, C2).

For good orders sake, we should also include a third, weaker hypothesis that would hold only if we were able to balance the signal strength in C2 and C3. If no significant difference existed in the number of observed interventions, we would expect to see not only a rise in the ratio of corrections to observations, but also a significant higher number of corrections per intervention in C3. In other words, if the only important difference between the characteristics of two interaction designs is that one better supports the cognitive processing of information which is available, perceived and attended to, then that design will result in a higher incidence of true awareness (understanding) than its competitors. In terms of our experiment, this third hypothesis (H3) could be phrased as follows:

H3: If the number of reported interventions is similar in the two conditions involving warning signals (C2, C3), then the absolute number of corrections observed in C3 will be significantly higher than in C2.

3.3. Subjects

In order to test these hypotheses we asked 30 postgraduate students to participate in our between-subjects experiments (separated into three groups of 10, one for each of our three conditions). Clearly, the use of non-professional participants reduces the ecological validity of our experiment, but the resource of commercial pilots' time is extremely limited and we felt that our, interface literate replacements would be sufficient for this initial empirical study.

3.4. Equipment

Having recruited our participants, we set up a simple working simulation of the panel and displays in question

on a Pentium-4 PC with a 19" screen. We then programmed our control and extended interfaces using the Java programming language, relying heavily on the swing graphical interface packages to produce the simulated interfaces, described below.

First, we constructed an input interface, a faithful replica of the flight control unit (FCU) (Fig. 1) used in the A320. The FCU consists of four dials, six buttons and three switches. The dials allow targets for speed (SPD), lateral heading (HDG), altitude (ALT) and vertical speed (VS) to be given to the autopilot and the three buttons pertinent to this scenario allow the selection either of the first/second autopilot or the execution of an expedited instruction (simply put, the execution of an instruction without the delay associated with arming an autopilot). The switches allow alternate entry parameters to be used in place of those described above but were not used in this experiment and will therefore be omitted from this description.

For those interested in a more complete description of the FCU or of the other A320 panels described here, one can be found in our previous work on the subject [17] or in the official accident report of the Strasbourg crash [12].

When not entering a set of target parameters to the autopilot, the participants were asked to monitor the progress on the display provided, which provided the only indication that the flight was/was not following its intended trajectory. The display concerned, appropriately named the primary flight display or PFD (see Fig. 2), showed an animated representation of the aircraft's flight relative to an artificial horizon, along with a series of moving bars indicating current speed, altitude and heading.

Having simulated both input panel and display, we added a further software module, which played a pre-recorded audio track, containing a series of instructions from a fictional air traffic controller (ATC). The instructions recorded were drawn from an a voice

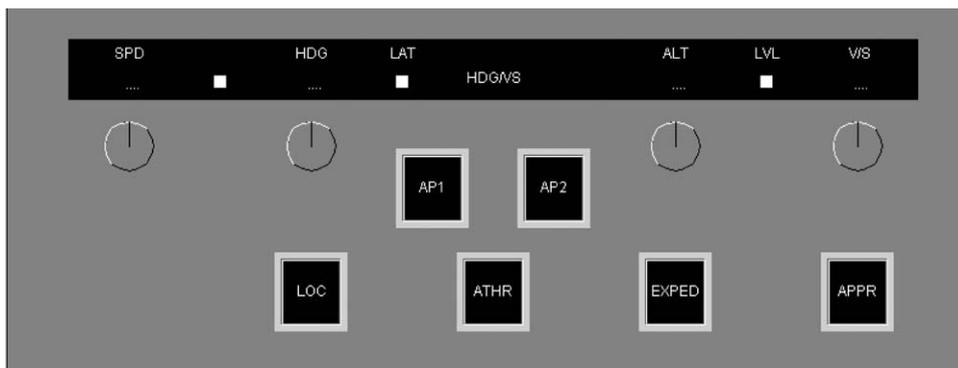


Fig. 1. Airbus A320 flight control unit (FCU).

transcript drawn from an authentic descent sequence and the use of pre-recorded, rather than live speech ensured that each instruction would occur at an identical time for each participant in each condition.

Finally, we added a software module, which recorded the participants' input to the FCU. Each parameter entry, button press and autopilot selection was recorded, along with a timestamp, which allowed us to specify both the point at which each instruction was finally entered, along with the execution time, missteps and corrections. This automated recording procedure was supplemented by manual recording on a carefully standardised form both during each the experiment and in a carefully controlled post-experiment debrief.

3.5. Design

Having laid the nature of the equipment we would use and the hypotheses we wished to test, our next task was the design of the specific warning signals to be used by our participants. The control condition, of course,

needed no such design but the implementation of the warning signals in C2 and C3 would be central to our experiment.

We started by selecting the location of the interventions, considering a number of options, before settling upon our final choice. Our first candidate was to introduce a warning signal across the entire display (e.g. install a flashing colour change across every inch of the background). This was rejected on the basis that such a strong signal would be potentially diverting or distracting, swamping the very awareness or understanding we wished to examine. We felt that this approach would be analogous to playing a very loud sound in the middle of a delicate piece of music—the listener would have no choice but to divert their full attention to the sound, destroying any semblance of awareness of events in the underlying music.

Next, we considered mixing modalities, an approach, which is often used in industrial alarm systems where monitoring of a visual display is often interrupted by a loud bell or siren, in the case that an important (and usually undesirable) event had occurred. Again, this seemed to be inappropriate for our purposes, firstly on the basis that it was likely to entirely distract the participants from the central objective of flying the plane and secondly because it would draw us into the complications of cross modal (semantic) mapping. We needed a warning signal, which was strong enough to draw attention and (in condition C3, at least) give a very clear semantic signal to the participants, but not so strong that it would prevent any other work being carried out until it had been understood.

Finally, we settled upon a solution, which involved (a) just one mode of signal (animated visual display) and a single icon within the diverse range of instrumentation on our display. The icon we chosen was the one representing the aircraft in the centre of our PFD (displayed as two rectangles representing wings and a circle representing the body in Fig. 2).

We then moved on to define separate signals for C1 and C2, choosing a directionless pulsing or throbbing of the icon for C2 (see Fig. 3) and a small predictive movement (described above and detailed in Fig. 4) for

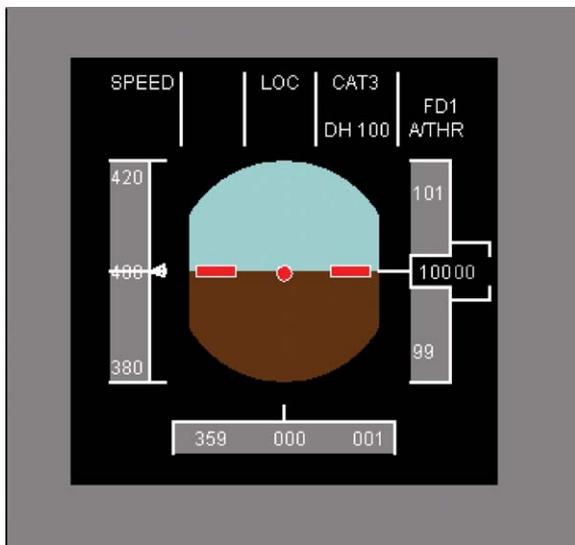


Fig. 2. Airbus A320 primary flight display (PFD).

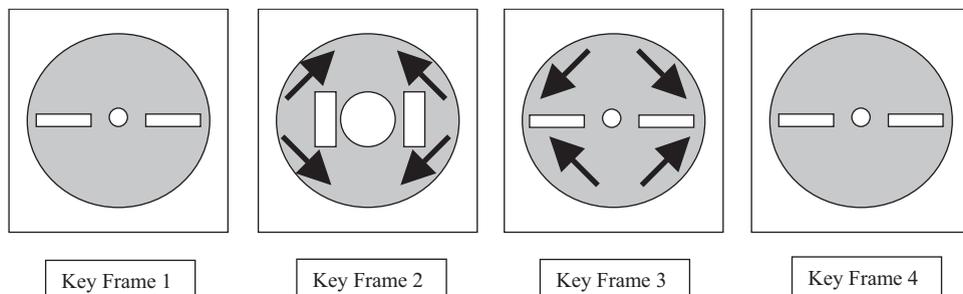


Fig. 3. Animated sequence of warning signal for all interventions in condition C2.

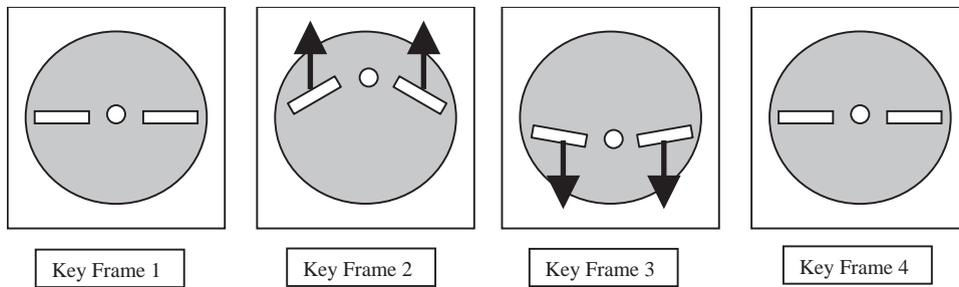


Fig. 4. Example animated sequence, warning of impending autopilot instruction to descend in condition C3.

C3. Importantly, we ensured that the movement of the icon edges and the resulting maximum size of the icon in condition C2 was greater than that in C3 (i.e. the warning signal in C2 was larger and subject to greater overall movement than that in C3), but portrayed no inherent direction. By contrast, the warning animation chosen for C3 was smaller and moved less, but contained a definite “direction (up, down, left or right), corresponding to the direction in which the autopilot was altering the underlying path of the aircraft. As a consequence, therefore, we were able to assert a strong signal but small semantic relevance in C2 and the reverse in C3, the exact conditions necessary for our hypotheses, detailed above.

The final animations chosen for the warning signals in condition C2 is detailed in Fig. 3. The four key frames (i.e. snapshots in the course of an animated sequence which show the extreme point of a particular movement or action) show the initial state of the icon (key frame 1), a squashing effect (Icon gets taller and larger, before reverting to its original size—key frames 2 and 3) and the return to its original state at the end of the animation. The effect of these movements, when run together in a repeated sequence is that the icon appears to pulse or throb in position—a warning signal used before each intervention regardless of direction in this condition.

The corresponding warning signal for condition C3 is shown in Fig. 4. In this case, the animated sequence was based upon the notion of predictive motion, described above, with a different animation being used to anticipate movements in each direction. As a consequence, these key frames describe only the specific warning used as the autopilot was about to affect a *descent*. In this sequence, we see the original state of the icon (key frame 1), the initial “anticipatory movement”—in this case upwards—before the semantically relevant downward movement (used to indicate an imminent autopilot instruction to move the system downwards) in key frames 2 and 3 and the return to the original icon state at the end of the animation (key frame 4). Similar sequences were designed for an ascent (in which the icon drew down slightly before leaping

upwards), left turn (small anticipatory movement to the right before a leap to the left) and right turn (small move to the left before a leap to the right), each providing a semantic link between the animation and the underlying autopilot activity.

3.6. Procedure

Having completed this preparation, our next task was the implementation of our experiment:

Each participant was first given training in the use of controls and displays provided in the simulation and required to successfully complete a practice run before undertaking the final, recorded flight scenario. Importantly, they were given the context within which the test scenario would take place (a descent to an airport which was below them and to their right) and briefed on the vocabulary to be used throughout. They were also told that the automated systems might “assist” in their flight, that this might involve independent interventions, and that an ongoing awareness both of the interventions and the flight path itself would be crucial to a successful completion of the mission. They were not, however given any further details about the interventions themselves.

The participants were then asked to enter appropriate settings to the input panel, such that the autopilot would execute the instructions given by ATC, e.g. in response to the ATC request “go to 5000 ft for now and await further instruction”, the participant would instruct the autopilot to execute a descent to 5000 ft and watch the display to ensure that this was carried out. The final recorded scenario was based upon flight transcripts of an authentic A320 descent, with instructions, vocabulary and timing being almost identical to those reported.

Next, distractions and diversions were added through the use of checklists and verbal confirmations, based upon those observed during flight. Examples of these distractions included requests to confirm flight parameters, check weather reports and set speed bugs. In all, every attempt was made to recreate a complex, busy environment in which sufficiently competent

participants could attempt authentic scenarios armed only with the electronic flight instruments described above.

The participants were then asked to complete a portion of descent from 10,000 ft altitude to a point shortly before the final landing sequence. ATC instructions and interaction with a fictional co-pilot were recorded on tape and our subjects were asked to complete them. Before commencing the flight, each participant was fully briefed on the flight context (goals, objectives and priorities) and given skeletal information about other environmental factors.

During the course of the flight, the autopilot was programmed to make exactly four interventions of the kind described above, i.e. on four occasions (at fixed times from the start of the scenario), each participant was subjected to a deviation in the flight path independently of their own actions. Unbeknown to the participants, however, each intervention involved an undesirable alteration to the flightpath, i.e. the autopilot moved the plane *away from* the airport (in direct conflict with the participants’ original instructions). The participants’ instructions were clear on this point, deviations from a downward and right-moving trajectory were to be corrected immediately. Complete awareness of the flight environment across all conditions would, therefore have resulted in each participant making four rapid but appropriate corrections (one after each intervention).

Given the difficulty of proving that participants had or had not noticed an intervention, we made three kinds of observations during the flight scenarios. Firstly, each interaction between pilot and FCU was recorded with dial (or button/switch), timestamp and resulting parameter target all noted. This allowed us to extract corrections entered by the participants when undesirable interventions were noticed. Secondly, we asked the participants to verbally inform the experimenters in the case that they saw either changes in flight path or any other unexpected behaviour on the part of the aircraft during the scenario (concurrent protocol).

Finally, we debriefed the participants at the end of each run, asking them again whether the aircraft and controls had behaved as they expected throughout (post-hoc protocol).

4. Results

The results in each of our three conditions are reported in Figs. 5–7.

The tables are divided into two sections. The first (the left-hand side) left hand describes the reaction of each participant (numbered P1–P10), in each condition to each individual intervention (described, in order of occurrence as Int1, Int2, Int3 and Int4). Each reaction is recorded as follows:

- No Entry: Participant neither reported nor corrected the undesirable intervention.
- R: Participant reported that “something was wrong” after an intervention, but was not sufficiently aware of the details of the intervention to be able to correct it.
- C + Time (in seconds): Participant corrected the undesirable flight path (i.e. moved the aircraft back onto an appropriate Downward and right turning trajectory). The time taken to effect this correction (in seconds) is also given.

The right-hand side of each table then summarizes the results for each participant, showing (from left to right), the total number of interventions and/or corrections, the total number of reported (but not corrected) interventions and the total number of corrections.

Following this coding scheme, then, we can see from Fig. 5 that the third participant in the first condition missed the first intervention (Int1), noticed but was not able to correct the second intervention (Int2), noticed, understood and corrected the third intervention (Int3),

Intervention:	Int1	Int2	Int3	Int4	Total (Reported +/-or Corrected)	Total (Reported)	Total (Corrected)
Participant							
P1	R			R	2	2	0
P2	R		R	C=11 secs	3	3	1
P3		R	C=25secs		2	2	1
P4					0	0	0
P5			R	R	2	2	0
P6					0	0	0
P7			R		1	1	0
P8		R			1	1	0
P9			R		1	1	0
P10					0	0	0
Total	2	2	5	3	12	12	2
Mean					1.2	1.2	0.2

Fig. 5. Results for control condition (C1)—no warning signal on autopilot intervention.

Intervention:	Int1	Int2	Int3	Int4	Total (Reported +/or Corrected)	Total (Reported)	Total (Corrected)
Participant							
P11		R	C=18 secs	C=12secs	3	2	2
P12	R			R	2	2	0
P13			R	R	2	2	0
P14	R			R	2	2	0
P15	R	R			2	2	0
P16	R	R	C=6 secs	R	4	3	1
P17	R	R			2	2	0
P18				R	1	1	0
P19	R		R	R	3	3	0
P20	R	R	R	R	4	4	0
Total	7	5	5	8	25	23	3
Mean					2.5	2.3	0.3

Fig. 6. Results for second condition (C2)—high signal strength, low semantic salience.

Intervention:	Int1	Int2	Int3	Int4	Total (Reported +/or Corrected)	Total (Reported)	Total (Corrected)
Participant							
P21			R	C=4secs	2	2	1
P22	R		R	R	3	3	0
P23					0	0	0
P24	C=8secs	C=5secs	C=5 secs	C=10 secs	4	3	4
P25				R	1	1	0
P26			R		1	1	0
P27			R	C=5 secs	2	2	1
P28	C=4 secs	C=8 secs	C=6 secs		3	3	3
P29		R			1	1	0
P30	R	R			2	2	0
Total	4	4	6	5	19	18	9
Mean					1.9	1.8	0.9

Fig. 7. Results for third condition (C3)—low signal strength, high semantic salience.

taking 25s to do so and then failed to even notice the fourth intervention (Int4). These figures sum to give two interventions either reported or corrected of which only one was, in fact, corrected.

4.1. Reported interventions (H1)

At this early stage in our results, a number of interesting trends were apparent:

Our first hypothesis (H1) involved the frequency with which our participants would notice (report) the undesirable autopilot interventions. Specifically, we believed that the inclusion of some anticipatory warning signal would significantly improve the likelihood that interventions would be reported. The baseline for comparison was the control condition in which only twelve of the 40 interventions were reported (i.e. $12/40 = 30\%$ of interventions were recognised using only the animation of the animated display). This result is particularly interesting since it is very close to the results obtained by Johnson and Pritchett in their reconstruction of the Strasbourg accident [28], suggesting that we had some success calibrating the complexity of experimental task to a realistic level of difficulty. The results of

our second condition, C2 (i.e. the condition in which the interventions were indicated by additional animation of high signal strength but low semantic relevance) saw observations rise to 62% (i.e. 25 of a possible 40 interventions) and the third, C3 (in which we used a smaller but semantically more relevant signal) returned 47% (19 of a possible 40).

Here at least, we were able to perform one-way ANOVAs on the number of reported interventions (regardless of ultimate corrections). In individual tests, we found that the results in C1 were significantly different from the much higher numbers found in C2 ($p < 0.05$) and C3 ($p < 0.05$), supporting H1 and, with it, our belief that an understanding of independent autopilot activity was important to the participants understanding of the flight. Importantly, we also found no significant differences between C2 and C3, leading us to the conclusion that we had assured approximate parity between the signal strengths in each condition.

4.2. Corrected interventions (H2, H3)

As we described earlier however, alerting our participants that an intervention was taking place was only

one part of our objective in this experiment. We were also interested in the frequency with which they would go beyond the level of awareness at which they *knew that something was happening* to the level at which they *understood the intervention* sufficiently to correct its consequences. Here again, our results were encouraging with only 2/40 (5%) of interventions being corrected in the control condition (C1), rising to 3/40 (10%) in C2 and 9/40 (22%) in C3, the condition offering most semantic connection between the signal chosen and the underlying autopilot action.

In this case, our parametric tests (one-way ANOVA) showed no significant difference between either C1 and C3 or C2 and C3, although as we have shown above (and will discuss further below) there is a clear trend towards improvement in C3. One explanation could be that our relatively small number of participants contributed to these non-significant results, but we must also consider the notion that individual differences between participants played a role. Nonetheless, we still draw some encouragement from these results. A glance at Fig. 8, for example suggests that a strong trend emerged both in the number of reports actually leading to corrections and the absolute number of corrections in C3 were dramatically different from the corresponding number in C2, providing support, at least for the spirit of hypotheses H2 and H3.

Equally interesting was the fact that these results were more widely spread between participants in C3 (in which 4 participants registered at least one correction) than in either C1 or C2 (2 participants registered at least one correction in each condition). Our results could not, therefore be explained simply by a higher ability to follow complex data on the part of the successful participants in condition C3.

Beyond the support for our hypothesis, however, it is also worth noting that the figures show a small and unexpected discrepancy between reported interventions and corrections—a discrepancy that exists both in

condition C2 and C3 (though not in the control condition, C1). We had expected some interventions to be reported but not sufficiently well understood for correction to be possible (hypothesis H2). We had not, however, expected unreported interventions to be corrected since it seemed extremely unlikely that participants would have stumbled across a solution to an undesirable event without realizing that anything was wrong.

We can take either a defensive or a more positive approach to these unexpected results. The defensive option is relatively straightforward; In each case, the participants reported in the post scenario debrief that they (falsely) believed the interventions concerned to have been reported—in other words, they reported the instructions and training as having been clearly delivered and had merely missed an assignment. For this reason, all corrections will be considered a sign both of noticing an intervention and of understanding it for the balance of this paper.

The positive approach by contrast is not quite so straightforward. The reporting omissions may again be indications of an interesting phenomenon outside the immediate scope of this experiment. One candidate explanation for this phenomenon is that the participants did notice the intervention (as discussed above) but were sufficiently busy with other tasks that they simply dropped the lower priority objective of reporting. This notion of low priority task “shedding” could (again), if properly understood, provide further insight into awareness breakdowns in the exactly those complex, dynamic environments which form the backdrop to this work. The omissions observed and the reasons they occur are, therefore worthy of inclusion in our future work.

4.3. Potential confounds

Despite these pockets of success, however our experiment is not without flaws. Since these flaws also constitute a form of feed-back to the community, we will discuss them here, before summarising our findings and drawing up plans for the future.

Our first problem was that some priming seems to have crept into our results, despite our efforts to the contrary. If we revisit the tables of results in Figs. 5–7, it becomes clear that the frequency with which participants reported interventions rose dramatically in the second half of each scenario. If we compare the total number of reports in the first half of each condition (C1=4, C2=12 and C3=8) with those in the second half (C1=8, C2=13, C3=11), we find large rises in the first and third conditions (though little difference in the second). We find it probable, therefore, that the likelihood of an intervention rose as the elapsed time in the scenario (and with it the number of previously

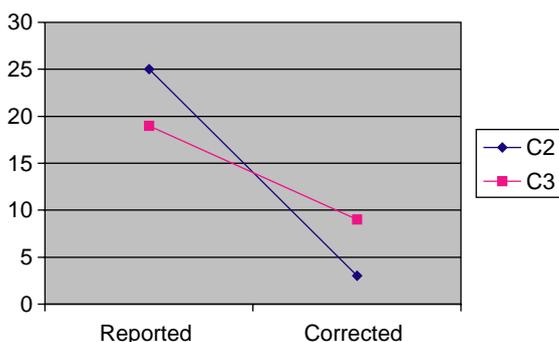


Fig. 8. Total number of (1) reported and (2) corrected interventions across all participants, grouped by condition (Conditions C2, C3).

reported interventions) increased. Intuitively, it seems that each participant who had noticed at least one intervention paid closer attention to subsequent developments in the scenario i.e. noticing a first intervention primed our subjects to notice the next one.

Interestingly, however, this presumed priming effect was not so noticeable in our high signal strength condition (C2). In this condition, seven participant reported the first intervention, whilst only five reported the second and third (see Fig. 6). To some extent, this remains an unsolved puzzle, since we do not have a very strong explanation for it, though the high signal strength in that condition, combined with a relatively low workload at the beginning of our scenario may go some way to explaining the unusually high reporting frequency for Int1.

Encouragingly, however, even ignoring the high reporting and correction frequency in the second half of our scenario, it seems likely that we would still find support for our hypotheses. If we consider only those results obtained in the first and second interventions (i.e. we ignore the second half of each scenario), in order to remove this priming effect, we can still see our expected trend towards higher reports in C2 (C1=4, C2=12, C3=8) and towards higher corrections in C3 (C1=0, C2=0, C3=4). Whilst our original diagnosis of a priming effect stands, therefore, we find it not to have been fatal to the underlying themes of the paper.

More aggressively still, we could take the opposite approach to these unexpected results and, in the place of excluding them from our findings, treat them as a pointer towards further useful research in the area. One version of this approach would be to assert this priming effect as a largely *positive* (though still unanticipated) outcome of this study, which if found to be repeatable and controllable, such an effect could lead us to an even more efficient support of high level awareness through the use of regular animated updates to pilots—sadly such findings are far beyond the scope of this current report and must be left for the future.

Returning to our immediate concerns, we could also find fault in the granularity of measurement, used to determine a “correction” in this experiment. Previous researchers [28] have shown in previous experiments that professional pilots are just as likely to pick up unintended autopilot activity from an unexpected flight path or aircraft position as from the primary flight display. In our experiment, then, we are prone to results, caused not by participant awareness of our warning signals, but rather by a comparison of subsequent deviation from the expected flight path (a result which lies outside the scope of our current hypotheses). The discrepancy in the mean delay between corrections being achieved in C2 and C3 (C2=12s, C3=6s), could, therefore, in part, be explained by this phenomenon.

Once again, however, whilst we concede the fact that future research may be needed to fully explain this phenomenon, it seems unlikely that the thrust of our argument need be greatly altered. There is no reason to believe that a significantly greater number of such “subsequent corrections” would have occurred in our semantically relevant condition (C3) than in its strong signal counterpart (C2). If any bias exists, it seems likely that C2, with its extended delay times would be its most likely location. By extension, therefore, the risk to our results is that the ratio of corrections in C3 against C2 would actually have risen, strengthening our claims. Once again, the minimum we can assert is that no damage was done to our central thesis.

5. Discussion and conclusion

So what, then, have we learned from this exercise?

At the level of our specific hypotheses, we can report encouraging results. We were able to successfully demonstrate an ability to affect the number of undesirable events observed by our participants (i.e. hypothesis H1 returned a significant result) and a noticeable trend emerged in the number of observations, which led to *corrections*. Whilst our ability to statistically support these findings is currently limited, a glance at the distribution of our findings strongly suggest a relationship between our design manipulations and the final results in our experiment.

If we move to the level of the particular domain and context chosen, then, we can claim a limited advance in our search for specific design solution, which deals with the problems observed in the Airbus A320. On one hand, at our chosen level of authenticity (moderate), we seem to have found an interface design, which increases the likelihood that people (at least in the age range provided by our participant population) notice unexpected automation activity. In this sense, we have demonstrated a proof of concept for a specific enhancement, which can be fed back into our chosen domain for further field-testing.

Both the non-professional participants and the practical limitations in our simulation, however, keep us from making the assertion that we have a ready made solution, which could be used “As is” in a full-scale commercial cockpit. In this domain specific sense then, a number of supplementary questions need to be asked about the scalability of our solution at least in terms of the physical size of the environment (the displays, though relatively authentic in size and detail would have been distributed over a wider area in a real aircraft), the complexity of the task (we are far from having tested the applicability of this particular design across a sufficiently wide range of contexts to make meaningful conclusions about its affect on the overall safety of the

cockpit environment) and the specialist nature of the subject population (trained pilots). While these are significant issues to address, we believe that they are empirical, rather than theoretical. That is to say, we believe that our theoretical ideas are scaleable but that our empirical work has yet to investigate this scalability.

Importantly, our results do indicate that a simple addition of raw signal strength may not be enough to deal with the awareness breakdowns observed. If, for example, we were to amend condition C2 to provide a warning signal of sufficient strength that every participant was certain to notice it (i.e. if we could produce a condition, in which 40/40 interventions were reported), our ratio of corrections to interventions (3 of a possible 25 in our experiment) suggests that we might expect only $3/25 \times 40 = 5$ corrections to result. This is still not as efficient as the design included in condition C3, in which we already have 9 corrections from only 19 reports. An optimal solution, then, is likely to involve a subtle blend of signal strength and semantic relevance. Again we believe that the nature of this blend is also an empirical issue, which can easily be assessed.

It is at this theoretical level, with some encouraging empirical results, that we make a contribution back to the wider community of HCA researchers and practitioners. For those involved in the development and prototyping of awareness supporting systems, for example, we have proposed that an approach based on a deeper, explanatory account of the processes by which awareness is achieved brings tangible utility to the design process in at least one authentic and problematic area. For practitioners in safety critical fields, we have also provided initial evidence of a category of breakdowns, which cannot be solved simply through an increase in the signal strength of appropriate alarms or alerts. We believe this theoretical model can extend to imply further challenges for those interested in peripheral awareness. Could we, for example manipulate the semantic relevance of our information sources, rather than their raw signal strength, to produce the kinds of interaction and awareness required by peripheral displays?

In conclusion, we believe that this work does make a direct contribution to the wider field of awareness research. We have demonstrated, for example, that our model, involving availability, perception, attention and semantic processing is important not only in description of the space as a whole but also in the identification of local design solutions which allow us to identify individual, if overlapping elements which must included in any comprehensive support for higher level awareness. As yet, we can say little about the difference between particular design elements which draw attention to pertinent events and states (whether system or otherwise) and those which facilitate the further (semantic) processing of the information attended to,

but we have clearly shown that the raw perceptual strength of a given signal is far from the end of the story in the creation and support of awareness.

Acknowledgements

This work is being funded by the EPSRC (grant number GR/R40739/01) and supported by QinetiQ and Westland Helicopters.

References

- [1] Gaba DM, Howard SK, Small SD. Situation awareness in anaesthesiology. *Human Factors* 1995;37:20–31.
- [2] Hogg DN, Torralba B, Volden FS. A situation awareness methodology for the evaluation of process control systems: studies of feasibility and implication of use. Storfjell Norway Oecd Halden Reactor Project, 1993.
- [3] Hutchins E. *Cognition in the wild*. Cambridge, MA: MIT Press; 1995.
- [4] Pedersen E. People presence or room activity: supporting peripheral awareness over distance. In: *ACM CHI Companion Proceedings of the Conference on Human Factors in Computing Systems*, Los Angeles, CA, 1998. p. 283–4.
- [5] Holmquist LE, Skog T. Informative art: information visualization in everyday environments. In: *Proceedings of Graphite*, Melbourne, Australia, 2003, 2003.
- [6] Mankoff J, Dey AK, Hsieh Kientz G, Lederer J, Ames SM. Heuristic evaluation of ambient displays. In: *Proceedings of ACM CHI Conference on Human Factors in Computing Systems*, New York: ACM Press. p. 169–76.
- [7] TSB of Canada. Crossing accident canadian national train No. 313-1M-14 Mile 47.48, Mont-Joli Subdivision, Caupascal, Quebec. Report Number R94m0100, 1994.
- [8] Tan DS, Czerwinski M, Robertson GG. Women go with the (optical) flow. In: *Proceedings of ACM CHI Conference on Human Factors in Computing Systems*, New York: ACM Press, 2003. p. 209–15.
- [9] Grudin J. Partitioning digital worlds: focal and peripheral awareness in multiple monitor use. In: *Proceedings of ACM CHI Conference on Human Factors in Computing Systems*, Seattle, WA, 2001. p. 458–65.
- [10] McCarthy JF. Event manager: support for the peripheral awareness of events. In: *Proceedings of the Second International Symposium on Handheld and Ubiquitous Computing (Huc 2000)*, Berlin: Springer, 2000.
- [11] Horvitz E, Jacobs A, Hovel D. Attention-sensitive alerting. *Proceedings of the Conference on Uncertainty and Artificial Intelligence (UAI 1999)*, Stockholm, 1999.
- [12] Bureau Enquetes Accidents. Rapport De La Commission D'Enquête Sur L'accident Survenu Le 20 Janvier 1992 Près Du Mont Sainte-Odile (Bas Rhin) À L'airbus A 320 Immatriculé F-GGED, 1992.
- [13] ATSB. Investigation Report 1999045538. Boeing 747-438, VH-OJH, Bangkok, Thailand, 23 September 1999, 2001.

- [14] Aeronautica Civil of the Republic of Colombia. AA965 Cali Accident Report, Near Buga, Colombia, December 20 1995, 1996.
- [15] Wiener E. Human factors of advanced technology (“Glass cockpit”) transport aircraft. Nasa Contractor Report 1 77528, Nasa Ames Research Center, Moffett Field, CA, 1989.
- [16] Federal Aviation Administration. Federal Aviation Administration Human Factors Team Report on the Interfaces Between Flight Crews and Modern Flightdeck Systems. Washington, DC, 1996.
- [17] Hourizi R, Johnson P. Beyond mode error: supporting strategic knowledge structures to enhance cockpit safety. In: Blandford A, Vanderdonk J, Gray P, editor. People and computers XV—interaction without frontiers, Joint Proceedings of HCI 2001 and ICM 2001. Berlin: Springer; 2001.
- [18] Cadiz J, Gupta A, Jancke G, Venolia G. Sideshow: providing peripheral awareness of important information. Microsoft Research Tech Report Msr-Tr-200183, 2001.
- [19] Johnson P, Johnson H, Hamilton F. Getting the knowledge into HCI-theoretical+practical aspects of task knowledge structures. In: Schraagen JM, Chirman SF, Shaun VL, editors. Cognitive Task Analysis. Mahwah, NJ: Lawrence Erlbaum Associates. p. 201–14.
- [20] Olsen GM, Olson JS. Research on computer supported collaborative work. In: Helander M, Landauer T, Prasad P, editors. Handbook of human computer interaction, 2nd Completely Revised ed. Amsterdam: Elsevier; 1997.
- [21] Endsley MR. Design and evaluation for situation awareness enhancement. In: Proceedings of the Human Factors Society 32nd Annual Meeting, Human Factors and Ergonomics Society. Santa Monica, CA: HFES, 1988. p. 97–101.
- [22] Sarter NB, Woods DD. How in the World did we ever get into that mode? Mode error and awareness. In: Supervisory Control. Human Factors 1995;37:5–19.
- [23] Endsley MR. Situational awareness in aviation systems. In: Garland DJ, Wise JA, Hopkin VD, editors. Handbook of aviation human factors. London: Lawrence Erlbaum Associates; 1999.
- [24] Rushby J. Using model checking to help discover mode confusions and other surprises. In: Javaux D, editor. Proceedings of the Third Workshop On Human Error, Safety and System Development. Liege, Belgium: HESSD. 1999.
- [25] Hourizi R, Johnson P. Unmasking mode error: a new application of task knowledge principles to the knowledge gaps in cockpit design. In: Michitaka Hirose, editor. Proceedings of Interact 2001, Eighth IFIP TC. 13 Conference on Human–Computer Interaction. Tokyo, Japan: IOS Press; 2001. p. 229–46.
- [26] Barnard PJ, May J. Towards a theory based form of cognitive task analysis of broad scope and applicability. In: Schraagen MC, Chipman SF, Shalin VL, editors. Cognitive task analysis. London: Lawrence Erlbaum Associates Inc; 2000. p. 147–63.
- [27] Johnson T, Johnson O. The illusion of life: Disney animation. New York: Little, Brown and Company; 1981.
- [28] Johnson A, Pritchett AR. Experimental study of vertical flight path mode awareness. Presented at the Sixth IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design and Evaluation of Man–Machine Systems, Cambridge, MA, 1995.