Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments
Lecture Notes in Computer Science Volume 8004, 2013, pp 193-202

Linking Context to Evaluation in the Design of Safety Critical Interfaces

Michael Feary[1], Dorrit Billman[2], Xiuli Chen[3], Andrew Howes[3], Richard Lewis[4], Lance Sherry[5], and Satinder Singh[4]

[1]NASA Ames Research Center, Moffett Field, California {michael.s.feary@nasa.gov}, [2]San Jose State University Foundation at NASA Ames, Moffett Field, CA, USA {dorrit.billman@nasa.gov} [3]University of Birmingham, Birmingham, England, UK {HowesA@bham.ac.uk}, [4]University of Michigan, Ann Arbor, MI, USA {rickl, baveja@umich.edu}, [5]George Mason University, Fairfax, VA, USA {lsherry@gmu.edu},

Abstract. The rate of introduction of new technology into safety critical domains continues to increase. Improvements in evaluation methods are needed to keep pace with the rapid development of these technologies. A significant challenge in improving evaluation is developing efficient methods for collecting and characterizing knowledge of the domain and context of the work being performed. Traditional methods of incorporating domain and context knowledge into an evaluation rely upon expert user testing, but these methods are expensive and resource intensive. This paper will describe three new methods for evaluating the applicability of a user interface within a safety-critical domain (specifically aerospace work domains), and consider how these methods may be incorporated into current evaluation processes.

Keywords: Work Analysis, Evaluation, Human Performance Modeling, Human – Automation Interaction

1 Introduction: A pressing challenge for new methods for technology evaluation

In many work domains, technology is designed to support users in carrying out functions and goals needed to do the work. Developing good user interfaces requires knowledge of what the work requires, knowledge of the environment in which the technology may be used and knowledge of human performance constraints. In this paper, we refer to this knowledge collectively as *context*.

The need to incorporate contextual information in evaluation of new technology in safety-critical domains is evident in incident and accident reports. There is recognition of this need in aviation and space domains as shown by changes to regulations and guidance material in aviation. Traditionally, these regulations have been written with the intent of removing as much context information as possible to allow for wide applicability; however the aviation regulatory community has recognized the increasing need for context information to be included in the evaluation. A good example of the

.

introduction of a requirement for context information is illustrated by European Aviation Safety Agency (EASA) Certification Specification 25.1302, below:

"This installed equipment must be shown, *individually and in combination with other such equipment,* to be designed so that qualified flight-crew members trained in its use can safely perform *their tasks associated with its intended function* by meeting the following requirements:

(a) Flight deck controls must be installed to allow accomplishment of these tasks and information necessary to accomplish these tasks must be provided.

(b) Flight deck controls and information intended for flight crew use must:

(1) Be presented in a clear and unambiguous form, at resolution and precision *appropriate to the task*.

(2) Be accessible and usable by the flight crew in a manner consistent with the *urgency, frequency, and duration of their tasks*, and

(3) Enable flight crew awareness, if awareness is required for safe operation, of the effects on the aeroplane or systems resulting from flight crew actions.

(c) Operationally-relevant behaviour of the installed equipment must be:

(1) Predictable and unambiguous, and

(2) Designed to enable the flight crew to intervene in a manner *appropriate to the task*.

(d) To the extent practicable, installed equipment must enable the flight crew to manage errors resulting from the kinds of flight crew interactions with the equipment that can be reasonably expected in service, assuming the flight crew is acting in good faith." (EASA CS25.1302)

This new regulation presents challenges to the state of the art evaluation methods, by explicitly requiring context information (stated as task characteristics) to be included in the certification and approval process. In addition, this regulation calls for methods that can be used to demonstrate alignment between the task and the intended function of the technology under evaluation.

Given these requirements, this paper will provide a brief background of usability evaluation in safety-critical interface evaluation. We will then describe three candidate methods for evaluating the applicability of a user interface within a work context, and consider how these methods may be incorporated into current evaluation processes. We will also show how the three methods can be used independently or linked to provide different levels of resolution for the different evaluation requirements.

2 Human-Automation Interaction Evaluation and Safety Critical Domains

We begin by examining the characteristics of the aviation and space domains we have been involved in. We have noticed five characteristics that we believe place significant constraints on effective methods for evaluating Human-Automation Interaction methods.

0.1 ***The work is performed by experts, and access to experts may be limited.*** Analysts and designers may share a part of the domain expert's knowledge, but operational expertise is required for evaluation. Limited access can be a key constraint, particularly when simple observation is insufficient. In addition, the population of experts evaluators may be reduced to the point that that the overall evaluation has limited utility (Faulkner, 2003; Macefield, 2009). Access to static expertise in documentation or training materials may be of limited value for many reasons, including reliance on specific procedures that may change, be obsolete, or be operationally invalid. Further, even when usability assessments are conducted by usability experts, assessment done by different experts may vary considerably in both the nature and severity of problems identified (Molich et al 2010).

0.2 ***The systems are increasingly complex and interactive***. The dynamics and complexity of interactive systems may make it difficult to identify and explicitly define and present situations for evaluation. (Feary, 2005)

0.3 ***The cognitive activity may be difficult to understand from observation.*** There may be few external cues about internal processes, although these hard-to-observe activities may be very critical parts of work. (Caulton, 2001)

0.4 ***It is often difficult to clearly separate the work to be done and the functionality used to accomplish it with new technology.*** Use of automation can change the nature of the work being accomplished to the extent that it is difficult to separate this work from the functionality provided by the new technology. For example, navigation is a critical work activity in aviation and space domains, and there are many different means available for accurate navigation. If the work, in this case finding one's way from point A to point B, can be separated from the functionality used to accomplish it (e.g. using a GPS navigation system), it is possible to generate evaluation methods which are more broadly applicable to new technologies. This characteristic is true in many information work domains.

0.5 ***There is a need for methods usable early in the development process.*** This is problem that is not unique to safety critical domains. In the aviation community, the FAA has recognized this need in its' 2012 workplan, by stating that "Consideration of the safety aspects must be embedded within the initial concept development – otherwise, whole aspects of the technology or operational concept may need revision in order to ensure safety." It can be difficult to provide functionality that behave enough like the final product early in the design process to be valid for user testing, and it may be too expensive to resolve issues discovered late. Therefore, user interface evaluation in safety-critical domains requires an increased emphasis on methods beyond user testing.

The need for improved evaluation methods is becoming more apparent when these characteristics are combined with the increasing pace of development of technologies. Specifically the rate of development of information automation being proposed in

safety-critical domains has increased dramatically in the last decade. The volume of candidate automation concepts being presented also highlights the need for more efficient methods to meet the schedule requirements of the often large, expensive and complex safety critical design projects that typically allow limited time for evaluation. The need for methods that can respond to the increase in the number of technologies requiring evaluation—and their combinations in particular contexts—has been recognized by the safety-critical organizations, (US FAA, 2012; EASA, 2007).

3 Three Methods for Integrating Context into the Evaluation of Safety-Critical Interfaces

In the previous section, we discussed the need for new Human – Automation Interaction evaluation methods. In this section we will discuss three candidate methods, Work Technology Alignment Analysis, Task Specification Language and Optimal Control Modeling.

3.1 Work - Technology Alignment (WTA) Analysis

The first method, *Work-Technology-Alignment*, evaluates how well technology aligns with the structure of the work it is intended to support (Billman, et al., 2010, 2011, in preparation). Technology that is better aligned with a domain of work activity should support more effective performance, in that domain. Assessment of alignment depends on discovering the elements and organization of the work domain, and on assessing how well the entities and organization of the technology corresponds with that needed for the work domain. The method uses Needs Analysis to identify the elements and structure of the work and integrates proposals from several research traditions in HAI, Human Computer Interaction (HCI), Work Domain Analysis (WDA), and related disciplines to form the analysis. The goal of the analysis is to help identify where work and the functionality used to accomplish the work are not aligned, and to help provide insight into how to provide better alignment, and therefore improve Human-Technology performance.

High technology - work structural alignment means that there is a strong match at the level of particular elements (entities, relations, and operations), and that the organization of elements in the technological system (the "system") aligns with the organization of elements in the work domain. Conversely, a design might have weak WTA alignment for several reasons:

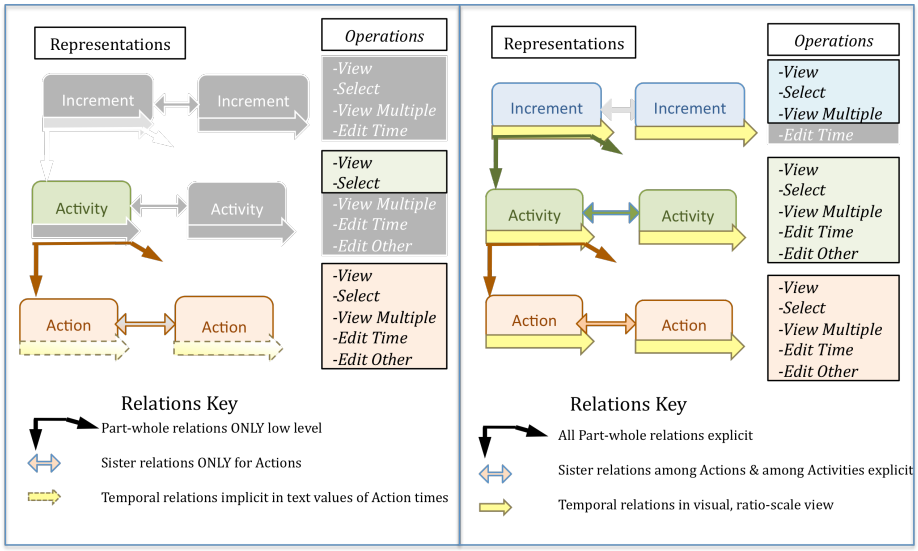
- Elements of the work are not represented in the system (missing functions);
- Elements of the system are unrelated to the work (system overhead or irrelevant "features"); or
- Elements in the work domain and elements of the system are organized differently.

We predict that systems with high alignment will provide multiple benefits: faster and more accurate performance; less training; better skill retention; and successful

operation over a wider range of goals or situations, including novel, infrequent, or emergency conditions.

An initial study assessed alignment to evaluate the technology used in a space flight control work, specifically, software for planning flight activities of the International Space Station. This included a needs analysis of the work structure, an analysis of legacy technology, and redesign of software guided by the alignment to the work structure. Figure 1 illustrates the improved alignment of the redesigned system. Based on these analyses, performance differences were predicted using legacy versus redesigned systems. Predictions were tested in a comparative experiment using tasks and material closely matching a subset of real operator work. Performance using the revised prototype had half the errors and took half the time, for critical tasks revising the scheduled time of events.

This case study suggests the Work-Technology Alignment evaluation method should be further developed as a means for incorporating context information in evaluations of complex, safety-critical technology. Research in progress is developing more structured methods of representing work, representing the technology, and comparing these representations.



Panel A: Legacy software Panel B: Redesigned software

Figure 1. Shaded representations and operations indicate aspects of the domain structure that are not expressed in the software. Differences in relations are noted in the key. The redesigned prototype aligns much better with the domain structure. Performance was better with the redesign, and particularly in tasks that tapped into

3.2 Task Specification Language (TSL)

The second method, analysis based on *Task Specification Language* (TSL) (Sherry et al., 2009) provides a task structure that can be applied to a wide variety of domains for detailed evaluation. This method maps traditional task analysis information into a more usable format, integrates contextual information, and responds to the need for methods and tools that do not require extensive expertise to implement and interpret. The goal is to provide a framework for developers and evaluators to think about the work activity (task), how the task is triggered, and the cues provided to the user to enable task completion and monitoring of task completion. The method may be used independently to identify issues in development, or used to provide input to computational models.

TSL is an approach to documenting the cognitive operations required by the users to perform mission tasks providing a framework for a more structured Cognitive Walkthrough (Wharton et al., 1992) and can be enhanced to use recently available "affordable models of human performance," to emulate Simulated User Testing.

The TSL specifically categorizes operator actions into the following categories:

1. Identify mission task and objectives
2. Select appropriate automation function to perform the mission task
3. Access the appropriate display, panel, page for the automation function
4. Enter the appropriate information
5. Confirm and Verify entries
6. Monitor progress and Initiate Intervention or New Tasks of the automation relative to the objectives of the mission task and initiate intervention if required.

A key contribution of TSL is the focus on failures to identify the correct mission task, or failures to select the appropriate automation function, and failures to monitor progress. These operator actions are exclusively decision-making actions that rely heavily on cues in the cockpit and recall of memorization items. When cues are ambiguous or are not sufficiently salient, human operators have been documented to exhibit poor reliability. Every operator action category has its own set of unique cues to guide operator actions and their own set of pitfalls. For example, the reliability of the Enter actions are affected by the ergonomics of the input devices. The reliability of the Access actions are determined by the location and user-interface navigation design.

The inclusion of all 6 steps of TSL allow for the method to be used independently, but the intention of future work is to enable steps 3, 4 and 5 to be automated, and the information collected in steps in 1 and 2 to be used in computational methods to enable steps 3, 4 and 5 to be automated.

3.3 Optimal Control Modeling (OCM)

The third method makes use of *optimal control modeling* to predict the strategies that people will adopt given specifications of (1) human information processing architecture, (2) the subjective utility functions that people adopt, and (3) the person's experi-

ence of the task environment. This approach uses cognitive architecture in context, and generates strategies for interaction with automation (Howes, et al., 2009; Lewis et al., (2012); Payne et al., in press; Eng et al (2006)). This work utilizes the contextual information in the form of cognitive architecture constraints, and fits well with the specific characteristics that safety-critical domains tend to provide, such as a population of expert users as the basis for evaluation.

The approach is based on a theoretical framework for the behavioral sciences that is designed to tackle the adaptive, ecological and bounded nature of human behavior (Lewis et al., 2004; Howes et al., 2009). It is designed to help scientists and practitioners reason about why people choose to behave as they do and to explain which strategies people choose in response to utility, ecological context, and cognitive information processing mechanisms. A key idea is that people choose strategies so as to maximize utility given constraints. In this way, the method provides an analytic means to predict and understand behavior as a function of the elements of context identified at the outset of this paper: the goals of the work are represented in explicit utility functions; the environments of training and performance are represented in the ecological context, and the human performance constraints are represented in the explicit assumptions about cognitive mechanisms. Payne and Howes (in press) and Lewis et al. (2004) illustrate the framework with a number of examples including pointing, multitasking, visual attention, and diagnosis. Importantly, these examples span from perceptual/motor coordination, through cognition to collaborative interaction.

Lewis et al's (2012) model of simple word reading brings together three threads that are critical to understanding cognition in the cockpit: (1) mathematical models of eye movement control (Engbert et al., 2005; Reichle, Rayner, & Pollatsek, 2003); (2) work on how higher-level task goals shape eye movement strategies (Rothkopf, Ballard, Hayhoe, & Regan, 2007; Ballard & Hayhoe, 2009; Salverda, Brown, & Tanenhaus, 2011); and (3) Bayesian sequential sampling models of lexical processing and perception (Norris, 2009; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). The model is an instantiation of a more general architecture for the control of active perception and motor output in service of dynamic task goals. The model decomposes the problem into *optimal state estimation* and *optimal control*, mediated by an information processing architecture with independently justified bounds.

Eng et al. (2006) report a model of the time taken and working memory loads required to perform simple tasks with Boeing Flight Deck of the Future (FDF) and existing 777 interfaces. Critically, optimal control modeling was used to select the strategies for both interfaces. The FDF performed better than the 777 for both time and working memory conditions. Across both tasks, the FDF consistently supported a strategy that allowed for a lower working memory load compared to the best case working memory load in the 777 (in one task by 175 milliseconds and in another by 1375 milliseconds). The FDF also performed better on time on task than the 777 (in one task by 100 milliseconds and in another by 500 milliseconds). These results validated the explicit design objectives behind the FDF interface. The interface comes at no cost to the time required to complete tasks while enabling a better distribution of working memory load. The success of the modeling was critically dependent on optimal control modeling to determine the predicted strategies because without this crucial constraint it is possible for the models to use almost any strategy on each of the

two interfaces. Model fitting without such constraints, which is a more common means of modeling human behavior, could not make any predictive discriminations between the two interfaces.

In new work we are developing models and empirically investigating how pilots switch attention between aviation and navigation tasks. In the model Bayesian state estimation is used to maintain a representation of two variables: (1) the aircraft attitude and (2) the body-centric location of FMS buttons. The scheduling of eye-movements between attitude indicators and the FMS are determined by the utility associated with the accuracy of these state estimates. Inaccurate estimates of button locations leads to data entry errors. Inaccurate estimates of attitude lead to poor situation awareness. Depending on whether the pilot wishes to prioritize data entry, using the FMS, or awareness of attitude then an optimal schedule of eye-movements is selected by the control model. Future work will be focused on providing tools to support the use of the modeling approach in evaluation processes.

4 Summary

In this paper we presented the case for why context information is important in evaluation of Human-Automation interaction, why new evaluation methods are needed for safety-critical systems and the characteristics of the problem that make it challenging. We then briefly described three candidate evaluation methods that have some promise of meeting this challenge in different ways. Work Technology Alignment (WTA) analysis provides a mechanism and metric for overall assessment of technology against the work context. The intention of the WTA assessment process is to produce a body of structured information that enables comparing the technology, training and procedures to a representation of the work to assess fitness-for-purpose.

Task Specification Language (TSL) emphasizes the need to explicitly define the mission tasks, and the accompanying functionality to complete the tasks, with the means by which the human operator can monitor the task completion. TSL is designed to be used as an independent assessment tool, or in cooperation with a computational method, such as an OCM model.

Optimal Control Modeling (OCM) provides machinery to enable a more thorough evaluation of safety-critical Human-Automation Interaction in the time limited evaluation process. The approach is to provide the model with functionality and context information—including assumptions about human information processing constraints—and then computationally generate rational strategies that could be used to achieve the work goals given those constraints. The strategy information generated by the analyses could then be used within existing evaluation processes to help identify Human-Automation Interaction vulnerabilities. These methods collectively provide a path forward for including context information into safety-critical work domain evaluations.

5 References

1. Ballard, D. H., & Hayhoe, M. M. (2009). Modeling the role of task in the control of gaze. *Visual Cognition*, *17*(6-7), 1185-1204.
2. Dorrit Billman, Michael Feary, Debra Schreckengost, and Lance Sherry. 2010. Needs analysis: the case of flexible constraints and mutable boundaries. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '10). ACM, New York, NY, USA, 4597-4612. DOI=10.1145/1753846.1754201 http://doi.acm.org/10.1145/1753846.1754201
3. Billman, D., Arsintescucu, Lucia , Feary, M., Lee, J., Smith, A., and Tiwary, R. Benefits of matching domain structure for planning software: the right stuff. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 2521-2530. DOI=10.1145/1978942.1979311 http://doi.acm.org/10.1145/1978942.1979311 (2011)
4. Billman,D., Arsintescu, L., Feary, M., Lee, J., Schreckenghost, D., & Tiwary, R. (in preparation as NASA Technical Memorandum). Product-Based Needs Analysis and Case Study of Attitude Determination and Control Operator (ADCO) Planning Work.
5. Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, *20*(1), 1-7 (2001)
6. Eng, K., Lewis, R. L., Tollinger, I., Chu, A., Howes, A., & Vera, A. (2006). Generating automated predictions of behavior strategically adapted to specific performance objectives. In Proceedings of the sigchi conference on human factors in computing systems (pp. 621–630).
7. Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. SWIFT: a dynamical model of saccade generation during reading. *Psychological review*,*112*(4), 777 (2005)
8. European Aviation Safety Agency, *Certification Specification 25.1302* and *Acceptable Means of Compliance 25.1302* Installed Systems and Equipment for Use by the Flight Crew (2007)
9. Faulkner, L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods*, *35*(3), 379-383, (2003)
10. Feary, M. Formal Identification of Automation Surprise Vulnerabilities in Design, Doctorate Dissertation, Cranfield University (2005).
11. Federal Aviation Administration AVS Workplan for Nextgen. Federal Aviation Administration , USA. (2012)
12. Howes, A., Lewis, R.L. and Vera, A.. Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action Psychological Review (2009)
13. Lewis, R., Shvartsman, M., and Singh, S. The adaptive nature of eye-movements in linguistic tasks: How payoff and architecture shape speed-accuracy tradeoffs, Topics in Cognitive Science (to appear).
14. Macefield, R. How To Specify the Participant Group Size for Usability Studies: A Practitioner's Guide. *Journal of Usability Studies*, *5*(1), 34-45 (2009)
15. Molich, R., Chattratichart, J., Hinkle, V., Jensen, J. J., Kirakowski, J., Sauro, J., ... & Traynor, B. Rent a Car in Just 0, 60, 240 or 1,217 Seconds?-Comparative Usability Measurement, CUE-8. *Journal of Usability Studies*, *6*(1), 8-24, (2010).

16. Norris, D. Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, *116*(1), 207, (2009)
17. Payne, S. and Howes, A. Adaptive Interaction: A utility maximisation approach to understanding human interaction with technology, Morgan Claypool lecture (In Press).
18. Reichle, E. D., Rayner, K., & Pollatsek, A. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, *26*(4), 445-476, (2003)
19. Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. Task and context determine where you look. *Journal of Vision*, *7*(14), (2007).
20. Salverda, A. P., Brown, M., & Tanenhaus, M. K. A goal-based perspective on eye movements in visual world studies. *Acta psychologica*,*137*(2), 172-180, (2011).
21. Sherry, L., Medina-Mora, M., John, B., Teo, L., Polson, P., Blackmon, M., ... & Feary, M. System Design and Analysis: Tools for Automation Interaction Design and Evaluation Methods. Final Report NASA NRA NNX07AO67A (2010)
22. Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140-159, (2008).
23. Wharton, C. Bradford, J. Jeffries, J. Franzke, M. Applying Cognitive Walkthroughs to more Complex User Interfaces: Experiences, Issues and Recommendations CHI '92 pp381–388. (1992)