Frazer Nash / ONR – The Assurance of Human Computer Interfaces and Interactions

**Literature Search and Derived Principles**

|  |  |  |
| --- | --- | --- |
| Title | *Frazer Nash / ONR – The Assurance of Human Computer Interfaces and Interactions* | |
|  | *Literature Search and Derived Principles* | |
| Version | *P3* | *16th April 2020* |
| Previous Version | *P2* | *21st October 2020* |
| Client Reference | 62639-001 | |
| Author |  | |
| Reviewer(s) |  | |

**Contents**

1. [General Introduction 8](#_bookmark0)
   1. [Background 8](#_bookmark1)
   2. [Scope and Purpose 8](#_bookmark2)
   3. [Report and findings structure 9](#_bookmark3)
2. [Overall Method and Process 9](#_bookmark4)
   1. [Literature review 9](#_bookmark5)
   2. [Development of principles 11](#_bookmark6)
      1. [Independent development of draft principles and concerns 11](#_bookmark7)
      2. [Consolidation of draft principles and guidance 12](#_bookmark8)
3. [Literature review 13](#_bookmark9)
   1. [HCI Design Processes and Methods 13](#_bookmark10)
      1. [Summary of Key Literature on HCI Design 13](#_bookmark11)
      2. [Conclusions drawn from key literature on HCI Design 19](#_bookmark12)
   2. [Generic HRA Methodology Limitations and Current Development 20](#_bookmark13)
      1. [Summary of Key HRA Methodology Sources 20](#_bookmark14)
      2. [Conclusions drawn from Key HRA Methodology Sources 23](#_bookmark15)
   3. [Methodology to Incorporate HRA within PSA 24](#_bookmark16)
      1. [Key Sources about HRA Incorporation 25](#_bookmark17)
      2. [Conclusions drawn from Key Literature on Incorporation of HRA in PSA 27](#_bookmark18)
   4. [Potential Sources of HCI Data 27](#_bookmark19)
      1. [Found Literature Containing HCI Data 27](#_bookmark20)
      2. [Conclusions from Found Literature Containing HCI Data 33](#_bookmark21)
4. [The System of Assessment Principles 35](#_bookmark22)
   1. [The Problem Facing ONR 35](#_bookmark23)
   2. [The Way Forward 36](#_bookmark24)
   3. [Principles: their Scope and Definition 38](#_bookmark25)
   4. [Qualitative and quantitative HCI Safety Assessment Principles 39](#_bookmark26)
      1. [Qualitative Human –Computer Interaction and Interface Design Safety Assessment 41](#_bookmark27)
   5. [Safety Assessment Principles for Human Computer Interaction and Human Error Probability](#_bookmark28) [Estimation 49](#_bookmark28)
5. [Discussion of HSBC Submissions involving HCI and Human Error 53](#_bookmark29)
   1. [Introduction and general considerations 53](#_bookmark30)
   2. [Derivation of Indicative HRA Data 56](#_bookmark31)
      1. [Complexity 57](#_bookmark32)
      2. [Available Time 59](#_bookmark33)
      3. [Team Performance and Team Communications 60](#_bookmark34)
      4. [Procedures 61](#_bookmark35)
      5. [HCI Navigation 62](#_bookmark36)
      6. [Data Reading 63](#_bookmark37)
      7. [Control Operating 64](#_bookmark38)
      8. [All Groups Combined 65](#_bookmark39)
   3. [Discussion of the Groups and Data Combined 68](#_bookmark40)
      1. [Further Data Exploration 70](#_bookmark41)
      2. [Re-grouping 71](#_bookmark42)
   4. [Uncertainty and Importance Analysis 73](#_bookmark43)
      1. [HBSC Modelling and Granularity 73](#_bookmark44)
      2. [HCI Improvement in HCI Assessment 76](#_bookmark45)
6. [Discussion 77](#_bookmark46)
   1. [The Strategic Picture 77](#_bookmark47)
   2. [The Tactical Picture 79](#_bookmark48)
7. [Recommendations 81](#_bookmark49)
8. [References 82](#_bookmark50)

Figure 1: A Functional Display for Reactor Coolant System (RCS) Inventory Control (Wu et al. 2012) Figure 2: Generic Cognitive Framework Structure (Whaley et al., 2012)

Figure 3: Example analogue and digital displays for Emergency Core Cooling System (ECCS) and RCS (Massaiu and Fernandes, 2017)

Figure 4: average error rates and error bands by display type and THERP task type (Massaiu and Fernandez, 2017)

Figure 5: Combined Data for the Grouping ‘Complexity‘ Figure 6: Combined Data for the Grouping ‘Available Time’

Figure 7: The Combined Data for the Grouping ‘Team Performance and Communication’

Figure 8: The Combined Data for the Grouping ‘Procedures’ Figure 9: The Combined Data for the Grouping ‘HCI Navigation’ Figure 10: The Combined Data for the Grouping ‘Data Reading’

Figure 11: The Combined Data for the Grouping ‘Control Operation’

Figure 12: All groups log transformed with trend line and Pearson Correlation R= 0.96

Figure 13: Smoothed and Normalised Distribution of the Probability in Log10 for All Error Data

Table 1: Issues of human performance with new technology from the International Nuclear Energy Research Initiative (INERI, 2015).

Table 2: High-level Summary of AI Design Guidelines (after Amershi et al., 2019) Table 3: Descriptive Summary Statistics for Each Group and All Groups Combined

Table 4: Comparison of broadly team cognitive and broadly single person action groups Table 5: Recommendations arising from research study and development of principles

**Abbreviations and Acronyms**

|  |  |
| --- | --- |
| 5C | 5th Centile of a Distribution |
| 95C | 95th Centile of a Distribution |
| ABWR | Advanced Boiling Water Reactor |
| AI | Artificial Intelligence |
| ALARP | As Low As Reasonably Practicable |
| ANS | American Nuclear Society |
| ASEP | Accident Sequence Evaluation Program |
| ASME | American Society of Mechanical Engineers |
| ATHEANA | A Technique for Human Event Analysis |
| CREAM | Cognitive Reliability and Error Analysis Method |
| ECCS | Emergency Core Cooling System |
| EOP | Emergency Operating Procedure |
| EPRI | Electric Power Research Institute |
| ESREL | European Safety and Reliability Conference |
| FAA | Federal Aviation Administration (US) |
| GDA | (ONR) Generic Design Assessment |
| HBSC | Human Based Safety Claim |
| HCI | Human Computer Interface / Interaction |
| HEART | Human Error Assessment and Reduction Technique |
| HEP | Human Error Probability |
| HF | Human Factors |
| HMI | Human Machine Interface |
| HRA | Human Reliability Analysis |
| HSE | Health and Safety Executive (UK) |
| HTA | Hierarchical Task Analysis |
| HuREX | Human Reliability data Extraction framework |
| IDF | Interaction Design Foundation |
| IDHEAS | IntegrateD Human Event Analysis System |
| INERI | International Nuclear Energy Research Initiative |

|  |  |
| --- | --- |
| NARA | Nuclear Action Reliability Assessment (EdF) |
| ND | UK Health and Safety Executive Nuclear Directorate (previous title for ONR) |
| NPP | Nuclear Power Plant |
| NRC | Nuclear Regulatory Commission (US) |
| NTSB | National Transport Safety Board (US) |
| NUREG | Nuclear Regulation (US NRC) |
| OAT | Operator Action Tree |
| ONR | Office for Nuclear Regulation (UK) |
| PSA | Probabilistic Safety Assessment |
| PSAM | Probabilistic Safety Assessment and Management Conference |
| PSF | Performance Shaping Factor |
| PWR | Pressurised Water Reactor |
| RCS | Reactor Coolant System |
| RHR | Residual Heat Removal |
| RP | Requesting Party (ONR GDA) |
| SA | Situational Awareness |
| SACADA | Scenario Authoring Characterisation and Debriefing Application |
| SPAR-H | Standardized Plant Analysis Risk-Human Reliability Analysis |
| TAFEI | Task Analysis for Error Identification |
| THERP | Technique for Human Error Rate Prediction |
| TLX | (NASA) Task Load Index |
| US NRC | United States Nuclear Regulatory Commission |

**Executive Summary**

This document reports on work that shows the Office for Nuclear Regulation a way forward in the handling of human-based safety claims that involve human–computer interactions. Currently, Licensee and Generic Design Assessment Requesting Party claims are typically made based upon existing human reliability assessment methods that predate the use of computers for user interfaces. Work undertaken a decade ago on behalf of the UK Health and Safety Executive’s Nuclear Directorate (which became the Office for Nuclear Regulation) showed that such methods can perform poorly for quantifying human error for human – computer interactions.

A high-level literature search showed international concern and ongoing attempts to quantify human error probabilities with Human Computer Interaction. The review suggests that acceptance of a quantitative case will be based upon methods capable of identifying and addressing cognitive error. However, such approaches are currently not fully developed both in terms of cognitive error identification and subsequent quantification.

Therefore, substantiation of human-based safety claims places a much greater onus on the need for a convincing demonstration that qualitative design has been informed by thorough qualitative identification and minimisation of error when interacting with computers. Such a demonstration will require evidence of the suitability of design processes intended to identify and minimise error and the application of qualitative task and error assessment to safety-related design to demonstrate that error potential is actually minimised.

This report sets out a hierarchical framework of principles derived from the literature search and the study team’s experience of human-computer interaction and error quantification. The principles have been prepared to be of use to an Office for Nuclear Regulation inspector in identifying aspects to consider when assessing safety case submissions.

However, the body of knowledge required for excellent qualitative design, quantitatively proven is complex. Consequently, the extent of the framework found to be necessary suggests that further work is necessary to evaluate whether these principles can be integrated into existing Office for Nuclear Regulation guidance materials or whether the important interrelationships between them might be lost if they do not stand alone. Recommendations are made both strategically and tactically for the way forward.

# General Introduction

# Background

The reliability of human interaction with advanced digital systems has long been an area of interest and concern within Human Factors (HF). As long ago as 1983, Bainbridge identified key issues and challenges that impact operator and system performance and reliability when automation is applied. Many aspects of the “ironies” identified by Bainbridge relate to the management, presentation and understanding of information by the system to the operator.

As technology has progressed the means and complexity of human-system interfaces has developed. Nowadays, the extent of information potentially available to system operators is enormous and the means by which this is organised, presented and manipulated by operators is equally large.

This presents challenges within the field of HF both in terms of design but also in assuring safety. Nowhere is this issue more apparent than in the design and safety assurance of new Nuclear Power Plants (NPP). New NPP seek to employ advanced, automated digital systems to manage their systems; a distinct departure from the hard wired, analogue control systems of earlier plants.

To more effectively regulate the development and implementation of such systems, the Office for Nuclear Regulation (ONR) wishes to better understand the “HF implications of deploying advanced Human Machine Interfaces in new reactor designs”. This need has arisen from work undertaken during the three (to date) Generic Design Assessments (GDA) of proposed new reactor designs for application within the UK. In each GDA to date the issue has challenged both ONR and the relevant Requesting Party (RP).

To ensure appropriate regulation, with a consistency of purpose and approach, ONR wishes to develop advice and guidance for Inspectors on how such systems and interfaces can be substantiated within a modern standards safety case from an HF perspective so that they may make informed judgements on RP and Licensee submissions. In doing so, it will be necessary to consider the implications to both design development and also Human Reliability Analysis (HRA). The advice produced will inform an interim regulatory position that provides sensible conservatism while not imposing excessive and unnecessary regulatory burdens.

# Scope and Purpose

The scope of this report is to present advice and guidance on substantiation of advanced Human Machine Interfaces (HMI) (specifically within modern NPP safety cases). The advice and guidance offered is presented in the form of assessment principles (superordinate and subordinate) intended to be applied by ONR inspectors along with particular guidance for the consideration of the use of such interfaces within quantified HRA.

In addition to the principles themselves, supporting and guiding text is provided to justify and clarify the extent, purpose and focus of the principles.

In support of the identified principles, and to inform of their evidence base, the report also presents the findings of a literature review. The literature review was undertaken to explore and clarify the areas of concern and to inform the aspects to be considered.

The principles have been written in a manner commensurate with ONR’s existing Safety Assessment Principles (SAPs) although the work undertaken here does not suggest how they may be integrated into or alongside the existing set of SAPs.

# Report and findings structure

The structure of the remainder of the report is as follows.

Section 2 defines the research and development methods used. As well as the basic process flow from literature review to development of principles the section identifies aspects of our approach including collaborative yet independent working to make best use of the experience of research team members.

Section 3 presents the literature review. It provides discussion of the particular areas covered by the literature review, how these were identified and the resulting evidence base identified. Alongside the narrative presentation of the reviewed literature information is also presented on the author’s considerations of the findings and their implications.

Section 4 offers proposed assessment principles and associated guidance. Both superordinate and subordinate principles are identified and defined, along with the explanatory text to further inform the reader and to indicate the particular research evidence that has driven the principle.

Section 5 provides wider discussion on the findings of the project and what general difficulties and challenges are apparent and are likely to remain apparent for those assessing safety case submissions involving advanced HMI.

Section 6 supports the discussion presented in Section 5 in identifying a set of recommendations for further work that may be required.

Section 7 presents references employed during the project.

# Overall Method and Process

# Literature review

A high level search of literature pertinent to their concerns on Human Computer Interaction (HCI) and HRA was undertaken. A two layered search strategy was chosen to maximise the possibility of useful information returns to inform further work.

The first layer of search focused on known sources for useful information. These sources are those known to the lead author of this report that represent ongoing activity in the HCI and HRA domains. In light of previous poor information returns, the literature review was confined to the following classes of journal and conference proceedings:

* Systems engineering
* HF / ergonomics
* HCI

In addition, other members of the team and three external HF specialists were consulted for potential sources. The external specialists were Brian Sherwood Jones and Jerry Williams in the UK and Andreus Bye at the Halden Project. There was strong concurrence of sources with a favourable likelihood of returns.

In particular, the Probabilistic Safety Assessment and Management (PSAM) proceedings for conferences held since 2010 were chosen and have been examined. The lead author attended two of these conferences and, in addition, a special workshop on HCI and HRA also run under by PSAM. These are sources closely connected to the nuclear sector and predominantly focus upon quantitative probability estimation.

Information sources were also sought from outside the nuclear sector. Very little was found of relevance beyond that related to glass cockpits in aircraft.

Key search terms for the primary search were formulated and used as follows:

* Human performance
* Human reliability
* Human error
* Human error identification
* Human error prevention
* Human-computer interfaces
* Human-computer interactions
* Transparent interface design
* User interface design
* Software development processes
* Probabilistic safety assessment
* Probabilistic risk assessment

The second layer of search has been to seek out references of relevance to the issue at hand from the first layer. The secondary sources were predominantly supportive of the primary with only occasional refuting or paradoxical results.

The results from this strategy provided a strong volume of returns on HRA methods in development, the applicability of HRA to HCI and the role of HCI in nuclear operator task performance. However, the identified sources did not focus upon the effectiveness of HCI design processes or guidance and provided very little on optimising HCI design content for human performance and reliability.

An internet based supplementary search for further sources concerning HCI design and design processes was undertaken. This identified the extremely comprehensive online source provided by the Interaction Design Foundation (IDF).

The IDF represents a large community of HCI/human interaction designers. Its sources are scientific in character, comprising over 2000 pages of peer reviewed information in total. The overall character of the information reflects a strong nexus between software design development and HCI design development intended to deliver efficient human performance. However, there was no detectable overlap with concerns of probabilistic HRA, beyond the general construct of human performance.

The secondary searches used additional search terms as suggested by references given in the primary search returns. The number and diversity of the secondary search terms was not recorded.

Five main themes have emerged and these are discussed in turn in section 3. The order of presentation for these themes has been chosen to minimise forward cross-referencing. The sequence approximately matches the order in which the issues would be encountered in a new nuclear design project. These themes are:

* Challenges in HCI design processes and methods;
* Generic HRA methodology limitations and current development endeavours/improvements;
* Methodology to Incorporate HRA within Probabilistic Safety Assessment (PSA);
* Published sources of HCI data (after 2010);
* Potential sources of HCI data.

Following completion of the literature review and initial consideration of the findings by the study team a discussion was held. The aim of the discussion was threefold:

* To share further details of the findings and the evidence that supports them;
* To reach a broadly consistent, shared position on the findings;
* To agree an outline structure for the guidance / principles to be developed (see Section 2.2).

The discussion was led by the lead author of the literature review who presented the key findings and the sources which contributed to them. The group discussed the findings with the rest of the study team able to question their detail and to offer their own positions and thoughts.

In addition to the team discussion a discussion was held with the ONR specialist inspector to share the findings and therefore the likely direction of intent with any principles or guidance developed. During this discussion, the intent to prepare principles akin to ONR SAPs with supporting text was also confirmed.

# Development of principles

# Independent development of draft principles and concerns

Following completion of the literature review and associated discussion the study team divided to create four independent sets of draft principles.

Aside from the intent to create a set of principles akin to ONR SAPs which address the aspects identified by the literature review the team were at liberty to develop principles (both super and sub- ordinate) along with explanatory text and justification and associated concerns.

# Consolidation of draft principles and guidance

Following development of four sets of principles these were consolidated into a single, consistent and comprehensive set. Consolidation activity removed repetition while capturing the full scope of principles developed by the team.

Consolidation also developed the structure and hierarchy within the principles and in particular the super and sub-ordinate relationships. Following initial development these were reviewed and refined by the study team.

As a component of this activity and in line with the identified principles, outline guidance was prepared on the probabilistic aspects of assessing safety case submissions where quantification of human error potential is provided.

# Literature review

# HCI Design Processes and Methods

This literature group addresses design processes and methods used to produce an HCI design that minimises the likelihood of human interaction error occurring. The literature falls into two broad categories:

* Concerns about HF methods and processes that may be used to inform ‘good’ HCI design;
* Prescriptions and guidelines to inform the design of a given interface, for example, a page displayed upon a screen.

This subsection first outlines the content of key literature sources that have influenced the review about HCI processes and methods and their potential contribution to error minimised human-computer interactions. Subsequently, conclusions are drawn about the adequacy of HCI processes and methods to achieve human error minimisation.

# Summary of Key Literature on HCI Design

Diaper and Stanton (2004) published a Handbook of Task Analysis for HCI. The chapters of this book were drawn from the HF and HCI development communities at that time. However, the passage of time has not led to a settled maturity on preferred methods.

For example, Thimbleby (2013) discusses the application of action graphs and user performance analysis. Action graphs map state transitions within the software/interface and user state changes – in the form of actions. The paper asserts that “Action graphs can be used to directly apply user performance metrics and hence formal evaluations of interactive systems”. In particular, it would appear to be appropriate for desk-based evaluations where empirical ones cannot be undertaken. The paper illustrates the application of the method to a consumer device, a digital multimeter and an infusion pump.

An earlier paper by Thimbleby (2009) targets the contribution that this and related methods can make to ‘safety-critical interactive systems development’. Further, he asserts in Thimbleby (2007) that even simple interactive systems are too complex for user evaluation (e.g. by the polling of opinion, user trials etc) to be able to evaluate and check reliably.

Instead, he proposes state transition diagrams as an effective, formal method for undertaking such work. However, it is very important to note that his illustrations of formal methods are only applied to simple devices such as calculators. Accordingly, the resource demands of such a formal method would appear to be very onerous for the design of HCI for a NPP.

It is also worth noting that Thimbleby’s conclusion about user evaluation is implied in work undertaken by Baber and Stanton (1994) which also uses state transition diagrams to map human – computer interactions and identify error potential. As with Thimbleby’s work, the interfaces are at the level of consumer products such as tram ticket machines.

The free online textbook ‘Encyclopaedia of Human – Computer Interaction’, published by the Interaction Design Foundation (2020) now in its second edition has over 4000 pages. It puts forward methods claimed to be applicable to a range of equipment from household objects to aircraft cockpits. It identifies three distinct types of formal methods of HCI analysis; these are:

* Users – with the predominant focus being on the formal documentation of one user and their interactions but with no recognised formal methods for dealing with groups of users at a higher level;
* Systems – the documenting of what happens within the system “out of sight of the user” when interaction occurs;
* World – in effect the circumstances that surround the human-computer interaction otherwise described as contextual or environmental.

Roger’s (2012) identifies that the most cited method on the web for HCI design is activity analysis. This method is described (Kaptelinin) at length in Chapter 16 of the online human – computer interaction encyclopaedia (2020).

Like state space diagrams, the origins of activity theory reach back quite some time and were first published in English by Leontiev (1978). The character of the method closely relates to hierarchical task analysis but also addresses cognition. The method does this within a three layer hierarchy which is, from top to bottom:

* an activity, supported by;
* a number of actions with each action in turn supported by;
* a number of operations

Fundamental to this theory was the notion that there is a motive for an activity, each action has goals and each operation has conditions.

The method closely aligns with one published as an extension of Hierarchical Task Analysis (HTA) by Hickling et al. (1992). In this latter method the depth of hierarchy is not constrained and motives, actions and goals are collectively known as goals. These can be the objective of an individual or a group. Because of its application this method specifically recognised that there should be a 1:1 alignment between a human goal and a required safety function or subfunction.

Irrespective of the particular presentation rules, both methods focus upon the very important point to HCI design that any consideration of tasks must also consider cognition. The cognition of importance is the pursuit of the required goals. (The method of Hickling et al. (op. cit.) also has a generative grammar or syntax for developing tree structures and checking for completeness and coherence of goals and their related subgoals.)

Hallbert et al. (2017) takes a broad view of the HF methods and techniques that can be used for characterising operator ‘in-the-loop’ performance and error. Therefore, it has a specific focus on processes and methods that might be used to evaluate performance and error. This paper is directed

towards application of digital technologies in modern NPP control. It summarises an earlier literature search about 57 identified human performance issues. (INERI, 2015). That search identified 37 key source documents from general industry, fossil power, health systems, manufacturing, aviation, mineral processing refinery and unmanned aircraft systems and another 35 from the nuclear sector. Those 57 issues were reduced to 9 broader categories as shown in Table 1below.

The issues identified in Table 1are matters that have concerned writers within the relevant 72 nuclear and non-nuclear source documents. These issues can range from ‘obvious things to worry about’ to things that have been found during the introduction of digital technology to a particular application sector or plant.

*Table 1: Issues of human performance with new technology from the International Nuclear Energy Research Initiative (INERI, 2015)*

|  |  |
| --- | --- |
| **Activity** | **Indicative Schedule** |
| Human-system interface complexity | Misplaced salience |
| Situation assessment (situation awareness) | Keyhole effect |
| Out-of-the-loop with the level of automation |
| Lack of early detection support |
| Missing task critical information |
| Requisite memory trap |
| Cognitive Workload | Cognitive workload due to alarm overload |
| Cognitive workload due to excessive nuisance alarms |
| Cognitive workload due to data overload |
| Physical Workload | Physical workload |
| Crew performance | Coping with complex disturbances |
| Opacity in a digital system | Complexity creep |
| Dealing with diverse information across different sources | Concurrent use of analog and digital systems |
| Fatigue due to digital environment | Anxiety, time pressure, work criticality, and other stressors |
| Confirmation/trust in a (digital) system | Low trust in sensor readings |

The remainder of the INERI paper considers, at a high-end generic level, HF methods that can consequently be applied to evaluate or assess human performance/human error issues. It also puts forward a particular human error classification scheme of six cognitively-based stages. This begins with the observation of system states and ends with the execution of a procedure in the face of the

observed system states. This, together with the identified issues is used to discuss, again at a general level, how performance measures should be selected.

The literature outlined so far is not sensitive towards the form that the digital technology might take. One of the latest developments in the nuclear sector is the application of digital technology to provide some level of automated diagnosis. A paper describing guidelines for the design of Human – Artificial Intelligence has been published by Amershi et al. (2019). This work stems from the Microsoft Research organisation and puts forward 18 guidelines for Artificial Intelligence (AI) interface design. These 18 guidelines are, in turn, derived from 168 candidate AI design guidelines drawn from (unsighted) academic literature.

It would be unhelpful to include the full details of the table given within that paper, because the intricacies of the table entirely depend upon the full reading the detailed paper. However, a summarisation of the grouping and headline title for each guideline is given in Table 2 below.

*Table 2: High-level Summary of AI Design Guidelines (after Amershi et al., 2019)*

|  |  |  |
| --- | --- | --- |
| **Group** | **No.** | **Guideline Title** |
| Initially | G1 | Make clear what the system can do |
| G2 | Make clear how well the system can do what it can do (i.e. how might the system make mistakes) |
| During interaction | G3 | Time when to act or interrupted based on the user’s task context |
| G4 | Show information relevant to the user’s current task and  environment |
| G5 | Match relevant social norms (cultural and social contexts and expectations) |
| G6 | Mitigate social biases (ensure language used avoids undesirable stereotypes and biases) |
| When wrong | G7 | Support efficient invocation of AI when needed |
| G8 | Support efficient dismissal of AI |
| G9 | Support efficient correction of the AI system when it is wrong |
| G10 | Scope services when in doubt (disambiguate or gracefully degrade provision of services to users) |
| G11 | Make clear why the system did what it did |
| Over time | G12 | Remember recent interactions |
| G13 | Learn from user behaviour |
| G14 | Update and adapt cautiously |

|  |  |  |
| --- | --- | --- |
| **Group** | **No.** | **Guideline Title** |
|  | G15 | Encourage granular feedback (user preferences during regular interaction) |
| G16 | Convey the consequences of user actions |
| G17 | Provide global controls (to adjust AI system monitoring scope and behaviours) |
| G18 | Notify users about changes (on updates or capabilities) |

Earlier work at Microsoft Research was reported by Horvitz (1999). This sets out fundamentals for so- called mixed initiative user interfaces which he defines as those where users and intelligent agents collaborate efficiently. These would appear not only to set foundations for subsequent AI research such as that discussed immediately above, but also to be equally appropriate where the user and the technology can choose or negotiate which fulfils a given function: that is, flexible allocation of function.

Thus far, the reviewed papers and sources focus on particular methodological aspects that are specific to HCI design and development. However, Endsley, (2004) provides a comprehensive exposition of the underlying principles for user centred design, a comprehensive description of the design development life-cycle and appropriate HF methods within it for design development and design evaluation.

Further, the book also provides a comprehensive exposition of relevant psychological factors including error: the so-called Situational Awareness (SA) Demons. Whether or not one particularly favours the perspective of situational awareness, the book nevertheless provides a comprehensive map of psychological issues and design processes and methods that can be applied that could achieve a good verified and validated user interface.

All sources outlined so far have a methodological focus on overall processes and methods within them. However, a body of literature exists that purports to be guidance for the design of particular interfaces suitable for process monitoring and control; such as NPPs. When one considers the methodological complexity involved it is troubling that grossly oversimplified design guidelines are readily apparent that may trivialise the complexity of the issue.

However, much more comprehensive guidelines for high integrity process control are published such as ‘Effective Operator Display Design’ published the Abnormal Situation Management Consortium (2015). Whilst these clearly written guidelines are promoted by Honeywell, they are available to all, appear to be without commercial bias, and address guidelines relevant to a hierarchical display system for process monitoring and control.

Wu et al. (2012) have provided the only significant experimental paper which evaluates the effectiveness of a design based upon what they call Function-Based Task Analysis. This produces a display design that is equivalent in form and content to so-called ecological ergonomic displays. An illustration of such display taken from their paper is shown in Figure 1 below.



*Figure 1: A Functional Display for Reactor Coolant System (RCS) Inventory Control (Wu et al. 2012)*

Wu et al. also provide an illustration of their task analysis. It is clearly hierarchical and shows functions that are synonymous with operator goals when controlling RCS. The notion, structure and form are very closely related to that of Leontiev and Hickling et al.

Wu et al. compare operator performance using a ‘Traditional display of RCS inventory control’ with their functional display. The traditional display is a hybrid of a mimic layout with some embedded mini-mimics which are themselves essentially functional.

The results of their experiment show no statistically significant results in the accuracy of abnormality diagnosis between the two display types. However, statistically significant results are shown for false diagnosis rate which diminishes by an order of magnitude when using the functional display. Both situational awareness and the subjective National Aeronautics and Space Administration (NASA) Task Load Index (TLX) workload assessment are statistically significant but only show small but practical differences in favour of the functional display.

Section 3.4.1 considers human reliability error data in relation to glass cockpit displays. It is of note that the United States Federal Aviation Authority (2001) has published ‘Human Factors Design Guidelines for Multifunctional Displays’. This well-researched document is comprehensive in its scope and provides guidelines on design processes and displays for: air traffic control, weather, navigation, automation, and general aircraft display principles. It is interesting to note that it draws guidelines

from consumer computing, business computing and nuclear displays. It also provides HF guidelines derived from research into the different application areas that the document intends to cover.

# Conclusions drawn from key literature on HCI Design

Thus far, some key material on HCI design methodology has been outlined but no significant conclusions have been drawn. As noted previously the literature review was brief and addressed subjective issues of importance and therefore may be framed by preconceptions within the study team.

Within the process control industry there appears to be an oversimplification of the issues involved in producing effective HMI through computerised displays. This is perhaps best illustrated by the prevalence of what might be termed ‘animated piping and instrumentation diagrams’.

The reviewed information provides strong suggestions that this simplified approach, and the implicit assumptions it contains, ignores psychological issues and display techniques, such as the so-called ecological (or task-based) displays that are more akin to traditional non-mimic based control panels.

In contrast, even the small number of sources outlined in the preceding subsection show the following:

* + - * Comprehensive design and development processes for process control HCI are set out in available literature;
      * Display development with a strong link to task analysed cognition can be effective;
      * These development processes include references to methods for analysing tasks, including methods that can address actions and activities as well as goals thereby linking cognition with action;
      * The sources provide methods for identifying potential or actual human errors and provide qualitative frameworks for their classification and description;
      * Literature exists which provides strong links between the psychological issues involved in interacting with HCI and the designer of that HCI;
      * Process control specific literature is available for translating task and functional requirements into actual display designs;
      * The literature encompasses displays ranging from a simple on-screen manifestation of classical

‘knobs and dials’ through to reconfigurable displays and artificial intelligence.

The lead author is struck by the majority of discovered information being at the level of describing or appealing to underlying HF principles and ‘conceptual steers’ rather than a more mechanistic approach of ensuring a display is well laid out, legible, clearly understood etc.

The above conclusions are positive and the meta conclusion is, therefore, that provided a period of well-directed and considered literature searching is undertaken, it is possible to arrive at a library of guidance for HCI design and development.

However, there is continuing debate about which methods are best for arriving at a designed HCI:

* + - * Online evidence in the form of an encyclopaedia which exceeds 4000 pages (Interaction Design Foundation, 2020), illustrates that developing effective HCI design demands a comprehensive and specialised knowledge of the HF involved.
      * This same resource clearly shows that disagreement exists between HCI specialists on the preferred methods to be chosen to address users and usability.
      * From the early 1990s (Baber and Stanton, 1993) a theme has endured (Thimbleby, 2009) which suggests that a move away from relying on usability trials for identifying and rectifying human error is required and that formal desk-based methods using finite state/state change models is required. Such an approach is resource intensive with no obvious economising shortcut. However, it may be feasible to undertake less structured human error identification processes and early user trials within development process to identify more obvious errors so that the design approach changed so reducing the burden on a later formal process. However, and crucially, it can be argued, as does Thimbleby, that if a formal approach, such as finite state models, is not applied at some stage then one cannot know what the error artefacts of the user interface might be (Thimbleby, 2020).

The most striking overall conclusion that comes from looking at HCI design methodology is that the great majority of literature coming from HCI professionals explicitly addresses the issue of cognition within the user interacting with the computer. This is encouraging as not only is it asserted by HCI professionals that cognition is integral to human-computer interaction; but that there are also tractable frameworks and methods to address it. In turn, this is an essential insight which gives strong hope that error potential interactions can be identified during qualitative HCI design.

This last conclusion is very important for two reasons:

* + - * there is a sufficient body of knowledge and methodology to enable HCI to be designed that reduces human error to as low as reasonably practicable and subsequently qualitatively demonstrate;
      * by the application of those processes, that the qualitative design meets the ALARP criterion.

# Generic HRA Methodology Limitations and Current Development

# Summary of Key HRA Methodology Sources

Accepting that cognition is significant in the assessment of HCI means that it is necessary to obtain a general appreciation of how cognition is currently being addressed in HRA methods, or is likely to be, in the foreseeable future. Literature on current developments in HRA has been sought to understand what is, or might become, available for quantifying human error probability (HEP).

Boring (2012) points out that THERP was a seminal HRA method and that subsequent methods have built upon it by addressing error identification, quantification and PSA integration. As part of that review he points out that whilst THERP had its origins within HF there has been a subsequent divergences between HF and HRA. Whilst the extent to which this has happened may differ within different countries there is little doubt that this is generally true.

This divergence has perhaps led to HRA methods following the THERP tradition; that focus on error as an artefact of technology rather than error as an artefact of psychology. This statement is made in the full knowledge that THERP contains references to cognition in the particular context of handling alarms and making diagnoses. However the experience suggests that the relevant aspects of the THERP methodology are rarely invoked.

Whilst it is true that THERP and associated approaches such as ASEP have been oriented to the attributes of non-computer based user interface design, SPAR H has a stronger focus on the attributes of tasks and psychology. Nevertheless, it is fair to say that the HEART method from 1986 has the strongest focus on psychology and task attributes of any published HRA quantification method in regular use.

Chen et al. (2014), as part of work looking into cognitive HRA methods undertook a meta-review based upon the work of Everdij and Blom (2010) and that of Bell and Holroyd (2009). Taking these two sources together, Chen et al. (op. cit.) had access to 188 reviews of human performance and HRA methods. As others have done, they sought to develop a taxonomy of HRA methods. They propose that HRA methods should be classified as three types: task driven, context driven and cognition driven.

Based upon their description of the taxonomy THERP would fall into a context driven method and HEART would be a task driven method. The main thrust of their work was a comparison of CREAM (Hollnagel, 1988) and IDHEAS (Whaley et al., 2012). Chen et al. term both of these methods as cognitive driven.

The US Nuclear Regulatory Commission (2000) sought to address the recognised deficit in addressing cognition in HRA in their development of ATHEANA. It was intended that the method should be used following incidents in order to obtain a better understanding of the Performance Shaping Factors (PSF)

/ Performance Influencing Factors (PIF) present that led to incidents and, therefore, should be included in any revision of HRA techniques. In practice, the method was found to be complicated and was ultimately considered as a corpus of expertise on human error rather than providing a usable technique to identify causal relationships between identified factors; or to identify root causes.

In the meantime in the UK the HEART technique has been modified by EDF to create the Nuclear Action Reliability Assessment (NARA) (Kirwan et al. 2005) technique. Somewhat ironically the application of this technique focuses more upon artefacts of technology and interfaces than upon the psychological dimensions of tasks being undertaken.

In contrast, and in the last decade, the US NRC has been promoting research and development on two different approaches to HRA quantification that have a much stronger cognitive basis.

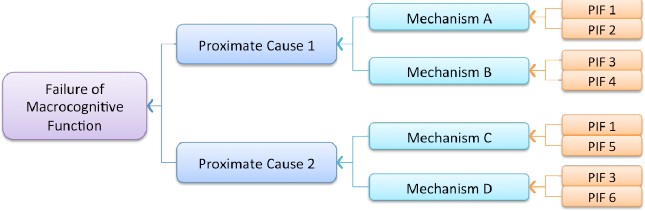
Whaley et al. (2012) put forward a cognitive framework for “bridging human reliability analysis

psychology”. This came up with an overall framework as follows:

* Detecting/noticing
* Sense making/understanding
* Decision-making
* Action implementation
* Team coordination

These five categories of cognitive performance are referred to as macro cognitive functions.

For those familiar with the Rasmussen stepladder model these are unsurprising ”macro cognitive functions”. However, the real significance of the framework comes within the logical structure intended to allow the quantification of cognitive error causing phenomena. Figure 2 below is reproduced from their paper.



*Figure 2: Generic Cognitive Framework Structure (Whaley et al., 2012)*

The overall method is provided with:

* Explanation about why particular PIFs are important
* Information about how particular PIFs influence human cognition into errors
* One easy-to-use tool that can inform HRA and other applications

The IDHEAS (IntegrateD Human Event Analysis System) is said to have “a strong foundation in human performance and cognitive psychology theory and employs a cause-based quantification model” (Huafei Liao, 2015). From pilot testing, Huafei Liao (op. cit.) considers the technique to have advantages in the treatment of cognition and approaches that handle modelling complexity. Further, the method is formal and self-consistent in the quantification derived. However, the author also concludes that there is insufficient guidance on how to undertake underpinning task analysis, that analyst judgement is required to adjust PSF’s (here called PIF’s) and that “fairly extensive resources are needed for qualitative analysis”. Other conclusions are given but these are an artefact of the methodology rather than fundamental of the characteristics in applying the HRA method to identify and quantify human error.

An alternative approach which continues in development is Phoenix (Ekanem et al., 2015). The Phoenix approach uses a Bayesian belief network originally populated with likelihoods and probabilities given by ‘human reliability experts ‘. However, in practice the method is not worked as intended and a hybrid algorithm is now being developed which uses multiple data sources to populate the model (Groth et al., 2019) [A hybrid algorithm for developing third-generation HRA methods using simulated

data, causal models and cognitive science, Reliability Engineering & System Safety, 191 November 2019 Elsevier] in their paper the authors state that HRA methods have been developed based upon cognitive science and separately new sources of HRA data but little research has been done on how to fuse the scientific advances together. This is still very much a work in progress.

Dang and Stempfel (2012) have shown that there are discrepancies between actual performance conditions and an analyst’s PSF ratings. At around the same time, Groth and Swiler (2012) have noted that “methods like SPAR-H require the analyst to assign values for all the PSF’s regardless of the PSF observability [and] this introduces subjectivity into the human error probability calculation”. Their proposal for dealing with this is to apply Bayesian techniques so that *a priori* probabilities can be used to modify an analyst’s judgement.

Both of these papers are important because they illustrate the vulnerability of an analyst making PSF judgements and both refer to the significant effect that this can have on overall estimates of HEP.

In other work, Liu and Li (2014) collected human error data and compared results with predictions by SPAR H (Gertman et al., 2005). Experimental subjects were students who undertook tasks using simplified Emergency Operating Procedures (EOP) using a ‘Microworld’® simulator. The simulator comprised three screens showing plant status, EOP procedural displays and a questionnaire. The plant status display had graphical and alphanumeric elements in the form of an animated P&ID “following current HCI examples in the NPP industry”.

Different experimental conditions of time and complexity were used in order to explore PSFs in practice with those declared within SPAR H. The reliabilities obtained cannot be considered representative of experienced NPP operators. Therefore, this study does not constitute a useful source of reliability data. Throughout the reported results the authors consider central, 5th and 95th percentile estimates.

There are, nevertheless, some important findings from Liu and Li’s 2014 work. They conclude (perhaps unsurprisingly, given the use of students), that SPAR H produces optimistic predictions of HEP relative to the empirical data. However, if the SPAR H PSFs are to be considered valid then one would expect a lawful relationship between them and empirically obtained data. They found that the relative effect of SPAR H PSFs for experience/training and stress/stressors empirically agreed with SPAR H. However, a more complicated relationship existed for time availability and experience which were found to be significantly and negatively related to the empirical HEP. However, task complexity was found to be significantly and positively related to the HEP but that this relationship, in turn, also depended on time availability and experience.

Liu and Li (op. cit.) point out that the findings of the International HRA Empirical study and the US Empirical HRA Study also showed relatively optimistic predictions being made by SPAR H.

# Conclusions drawn from Key HRA Methodology Sources

Established HRA practitioners are well aware that the inclusion of human error within a PSA can become a complex affair. It is clear from current HRA methodological development that, if those endeavours succeed, then incorporation within a PSA will become more demanding still. This subsection briefly outlines efforts that are being made to improve incorporation of assessed human reliabilities within a PSA.

The fundamental conflict within HRA, that has been endured for many years, appears to continue. A need exists for better phenomenological inclusion (such as cognition and organisational factors) yet complaints persist that HRA methods are not easy to apply and are time consuming.

American developments stimulated by the US NRC since the mid-1990s have shown that the incorporation of cognition into HRA generative methods is complex while within the industry there have been persistent complaints that HRA fails to consider cognition, organisational factors in general and safety culture in particular. This too is somewhat reflected in literature reviewed here.

Work in the last decade on analysts’ uses of PSFs has formally shown that resulting estimates of effect can be poor. Practitioners within the field who are critical are well aware of differences they might ascribe to a PSF relative to their colleagues arising from a desire to give the right answer or other biases such as undue optimism or absence of sufficient appreciation of tasks or systems of work.

Further, these mis-estimations significantly affect HRA outcome probabilities.

When it is considered that HCI interaction is complex not only in terms of the human phenomena involved but in terms of the number of interactions that may need to be mapped and analysed; and further, when this is coupled with the difficulties in the correct application of PSF’s, this raises a fundamental question about whether PSFs can be validly applied by analysts in the assessing of HCI.

Put another way the question becomes “Is the power of an analyst’s brain sufficient to subjectively encompass that complexity and turn the HCI attributes into a baseline HEP and relevant, properly gauged, PSFs?” Given that the qualitative identification of error in HCI design can, it is asserted, only be undertaken using highly structured formal methods the clear answer would seem to be ‘No’.

Where attempts to produce new HRA methods have been undertaken in the USA the work has shown that expert judgements are insufficient to populate a method and real data from simulators, incident reports or other sources is necessary. As with the genesis of HEART it would appear that formal data collection means that multiple types of data sources are now being entertained, at least in the USA.

Any truly meaningful method for HRA prediction must involve a considerable degree of analyst training in psychology and the modelling and handling of probability to be able to adequately address all the potential factors that may impact the reliability of a task. The alternative is to oversimplify and exclude the consideration of factors which could, potentially, significantly influence the estimation of risk. Any such oversimplification naturally should be distinctly conservative in any resulting assessments of reliability.

# Methodology to Incorporate HRA within PSA

# Key Sources about HRA Incorporation

Within the last 20 years, the US NRC and EPRI have jointly developed PRA methodology for PSA assessment of fires in nuclear power plant facilities (US NRC, 2005). Whilst, obviously, aimed at fires this is widely held within the USA to be a significant improvement on the way HRA should be treated within a PSA.

Contrary to early meta frameworks for PSA execution this approach makes clear reference to the importance of qualitative analysis, where possible, and examining and then evaluating error probability in human-based safety claims (HBSC).

It also sets out specific prescriptions for the structural incorporation of human error events within fault tree and event tree models. In addition, it addresses pre-initiating, initiating and post initiation human error events. Further, it explicitly requires the consideration of cognitive error.

Nevertheless, a number of criticisms have been levelled at the method in application and these are listed, in no particular order, below.

* + - * Reliance for qualitative task and error analysis is placed upon ATHEANA (NRC, op. cit.). Therefore it inherits the difficulties and complexities of using that approach. Accordingly, the application of the framework requires considerable human factors expertise and knowledge of human behaviour in fires if it is to be used effectively.
      * The method can produce more optimistic HRA probability estimates than other US-prescribed HRA approaches and this is notwithstanding their application for fire scenarios.
      * Where high uncertainty exists, screening values are provided that the criteria for determining screening value selection, whilst closely prescribed, do not take account of all the epistemic uncertainties that might exist, whether in a fire or not.

Notwithstanding these criticisms, proper recognition and acknowledgement should be made of the fact that the method does provide improved frameworks for incorporating HRA into PSA. It is salutary to note the points made by Boring (op. cit.) that THERP has provided the template for HRA assessment and incorporation within PSA, for all subsequent HRA methods for 50 years.

The American Society of Mechanical Engineers (ASME) and the American Nuclear Society (ANS) (2009) jointly put forward addenda to the 2008 ASME/ANS Standard for Level 1/Large Early Release Frequency Probabilistic Risk Assessment For Nuclear Power Plant Applications. At around the same time the US NRC (2009) published a regulatory guide on determining the technical adequacy of PSA results. As a result of these interventions, Electric Power Research Institute (EPRI) produced their HRA calculator (EPRI, 2013).

Anecdotal evidence from users of the calculator suggests that it is considered to be beneficial in speeding up the selection and calculation of an ‘appropriate’ HEP. Nevertheless, the same criticism can be applied as with fire PSA; that the application of the calculator can only produce credible results if thorough qualitative HF insights have been obtained that allow the character of the task to be properly understood; thereby enabling an informed calculation of HEP.

There can be little question that the onset of ‘digital technology’ is now internationally perceived to be a challenge for HRA. Julius et al. (2014) have specifically set out a wish list for new HRA techniques by scrutiny of existing HRA techniques. This work has been informed by the EPRI HRA Users Group and would appear to have been strongly influenced by the experience in using the EPRI HRA Calculator (op. cit.).

The work of Julius et al. (op. cit.) has identified a number of issues about the incorporation of digital control systems into plant operations. These are summarised below.

* + - * Operators develop overreliance on the digital controls;
      * Common failures may affect both instrumentation and prompts to which procedure should be executed;
      * It may be harder for the operator to recognise error and thus recovery of initial operator failures may be limited;
      * Operator errors could be induced by software errors or system failures that are not readily apparent;
      * New human error modes or failure rates can arise relating to selection of controls or indications when presented digitally;
      * Standard soft interface objects may not distinguish between safety and non-safety interface elements;
      * Operator errors in digital system test or maintenance may introduce additional forms of latent error to system operation.

Recognising the gap between existing HRA methods and digital interfaces Julius et al. offer three principal insights concerning digital interface systems:

1. These issues (summarised above) underscore the importance of “undertaking good qualitative analysis conducted on the first principles of understanding the PRA scenario and associated plant technical context”.
2. They pragmatically (and maybe, perforce) assert that current HRA methods can be used to

assess HEP’s while additional research and more data insights are obtained.

1. Probabilistic risk assessments of plants with digital control systems should conduct a range of uncertainty and sensitivity evaluations in order to assess the potential impact of changes to operator failure rate or new failure modes. For example, by varying the amount a recovery credit in varying the assessed level of dependence between operator actions.

Notwithstanding the suggestions put forward by Julius et al. in 2014, no subsequent publications had been found that indicate how the range of uncertainty and sensitivity evaluations should be conducted and what factors could influence data decisions within them.

On the matter of dependence, which in this instance focuses upon that dependence induced by digital systems between HBSCs, no references have been found to address that declared requirement.

# Conclusions drawn from Key Literature on Incorporation of HRA in PSA

Significant progress has been made in recent years n better incorporating HRA into PSA. This initiative has arisen within the USA and provides a clear prescription of how HF and HRA should influence PSA modelling development. However, deficiencies are still seen.

In particular, the habitual US practice of separating qualitative HF and quantitative HEP assessment appears to be perpetuating. Notwithstanding, as discussed in the previous subsection, vigorous research and development efforts are continuing to better incorporate cognitive factors into HRA.

Julius et al. (op. cit.) have identified seven factors where digital systems may alter human reliabilities. In the six years since these factors were identified, no evidence has been found that they are being addressed.

# Potential Sources of HCI Data

This last subsection of the literature review focuses upon data sources that may offer potential for the derivation of HCI data. It is not the objective of this section to seek to analyse and compare that data in any great detail but rather to outline the sources and report any methodological insights contained within them.

# Found Literature Containing HCI Data

It must be emphasised, once again, that the literature review undertaken has, by the instruction of the client, been rapid in nature. Therefore, any data contained within the sources and any subsequent analysis of it in this report must be considered as indicative rather than definitive.

In 2010 the UK Health and Safety Executive (HSE) GDA of the Westinghouse AP1000 and EdF/Areva UK EPR reactor designs considered whether THERP, Accident Sequence Evaluation Program (ASEP) and the Standardised Plant Analysis Risk-Human Reliability Analysis (SPAR H) methods were suitable for evaluating human reliability where HCI was involved.

It was concluded that the methods were unsuitable because assessment showed a considerable difference, of approximately one order of magnitude more optimistic, between the distribution probabilities within those methods than published HCI data would suggest. However, a paradoxical result was noted, albeit with sparse data in that mission-based reliability (e.g. warm start-up other reactor) was an order of magnitude more reliable than elemental reliabilities of humans interacting with screen-based icons or pushbuttons et cetera. This work was considered of sufficient significance to be placed in the public domain in academic literature by Hickling and Bowie (2012) [applicability of human reliability assessment methods to human-computer interfaces, cognitive technology and work, 15,1], and directed at the safety and reliability international community by Hickling (2012) [A comparison of published human computer interaction reliability data with established HRA methods]. Hickling and Bowie (op. cit.) stated “In our view, the typical, and unsubstantiated belief in the improved human reliability in HCI relative to consent conventional interfaces, requires considerably more

systematic study, before the apparent paradoxes in the little available data can be satisfactorily reconciled and an objective demonstration provided to support this belief.”

ONR can take credit for these publications influencing research that has followed.

Liu and Li (2016, 1) and (2016, 2) undertook work directly stimulated by Hickling and Bowie’s statement. They compared conventional and digital power plant main control rooms in terms of task complexity using 69 NPP operator subjects. Their taxonomy of complexity, drawn from nine reference sources, consisted of the following:

* Deficiency,
* Overabundance,
* Variety,
* Ambiguity,
* Relationship,
* Variability,
* Unreliability,
* Novelty,
* Incongruity,
* Action complexity.

Each factor within their taxonomy is subjectively rated (in Chinese) using Likert scales. Their comparison was based upon performance in normal and abnormal/emergency situations with corresponding scale ratings reported by subjects. The studies were undertaken in the Daya Bay and Ling-Ao Phase II simulators. These two papers are complex in their information and content. However, the overall results consistently show greater complexity for digital versus conventional control rooms.

In addition, as both highly experienced and less experienced NPP operators are involved in the study they were able to examine the effects of experience upon complexity. For normal operation they found no difference as a function of experience. However, in abnormal/emergency situations they found that more experienced operators encountered less complexity relative to the inexperienced.

These two papers tend to reinforce the notion that digital control rooms may be more difficult for operators than traditional interfaces that comprise of physical controls. The summary provided here does not do justice to the high density of information given in these two papers and they are strongly commended to any reader of this report for detailed reading and understanding. Because of the impossibility of having two simulated room versions interfacing the same plant it is possible to argue at a superficial level that the differences seen are a function of design approaches and conventions applied within the two control room design projects. However, the consistency of the difference seen irrespective of scenarios and operator experience is so consistent it seems unlikely to be an artefact of control room design but of digitally-based interfaces.

This particular study constitutes a source of qualitative data from a sound methodological base. It maintains the concern about differences between traditional and digital interface schemes.

In follow-on work, Liu and Li (2014) collected human error data and compared results with predictions by SPAR H. Experimental subjects were students who undertook tasks using simplified Emergency Operating Procedures (EOP) using a ‘Microworld’® simulator. The simulator comprised three screens showing plant status, EOP procedural displays and a questionnaire. The plant status display had graphical and alphanumeric elements in the form of an animated P&ID “following current HCI examples in the NPP industry”.

Different experimental conditions of time and complexity were used in order to explore PSFs in practice with those declared within SPAR H. The reliabilities so obtained cannot be considered representative of experienced NPP operators. Therefore, this study does not constitute a useful source of reliability data. Throughout the reported results the authors consider central, 5th and 95th percentile estimates.

There are, nevertheless, some important findings from this work. They conclude (perhaps unsurprisingly given the use of students), that SPAR H produce relatively optimistic predictions of HEP relative to the empirical data. However, if the SPAR H PSFs are to be considered valid then one would expect a lawful relationship between them and empirically obtained data. They found that the relative effect of SPAR H PSFs for experience/training and stress/stressors empirically agreed with SPAR H.

However, a more complicated relationship existed for time availability and experience which were found to be significantly and negatively related to the empirical HEP. However, task complexity was found to be significantly and positively related to the HEP that this relationship, in turn, also depended on time availability and experience.

Liu and Li (op. cit.) point out that the findings of the International HRA Empirical study and the US Empirical HRA Study (Forester et al. (2016) also showed relatively optimistic predictions being made by SPAR H.

Jung et al. (2016) continue in the process they started in 2015 in developing a data collection framework to build an HRA database. This is called the Human Reliability data Extraction framework (HuREX). This data is being collected on a simulator with experienced nuclear power plant operators. Two different nuclear plant simulations have been used: one for a Westinghouse three loop Pressurised Water Reactor (PWR) and one relevant to OPR 1000. An illustration accompanying their paper shows that all operator displays are screen-based. Within the body of the paper they publish data on errors in procedural tasks. Following definitions for error of omission and commission given within THERP they have declared 22 preliminary error probabilities and cognitive activity headings of: Information gathering and reporting; Response planning and instruction, Situation interpretation and Execution.

The given detailed descriptions of task types show that the collected data is functionally analogous to that contained within the THERP method. However, it does not appear to include any additional errors that might be associated with navigation or other page selection mechanisms.

Further work is reported using the HuREX method by Park et al. (2017). In this work, nominal human error probabilities provided by the Korean HRA method and the CREAM method are compared with empirical data.

In their comparison of obtained data with those given in CREAM most of the data is less than an order of magnitude difference with CREAM published data.

Their published comparisons with their Korean HRA method are more sparse, providing only two examples. Again, in both cases their predicted result is, at worst, within a factor of four with that predicted by their method.

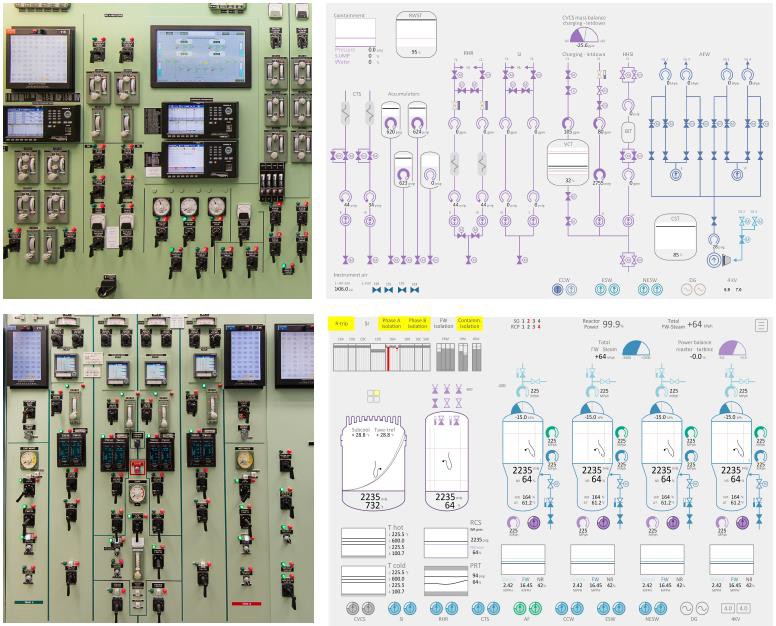
It is important to note that both Yung and Park emphasise their data to be preliminary being a feasibility demonstration of collecting simulator-based data.

Zhao et al. (2012) have also studied computerised EOP’s and execution reliability relative to time available. This study had 60 subjects all of whom majored in engineering. Data for error rate, operation time and subjected workload were analysed In this case they investigated the influence of absolute time availability (i.e. pre-prescribed at the same level for all participants) on procedures pursuit and relative time (i.e. how long operators took to user procedures).

For absolute times, a standard operating time was derived from pre-trials and time pressures applied by multipliers of: 0.8, 1, 1.2, 1.4, 1.6. Under the five of time pressure conditions error rates were found to improve but, nevertheless, the average error rate range was within the range 0.6 to 0.1. They also showed that operation time increased as time pressure levels were diminished. NASA TLX measures showed corresponding subjected workload and time pressures perceptions diminished as time pressure diminished. The high error rates obtained are conspicuous and, I would suggest, show the rates error that are likely to occur when time pressures are very close to those obtained from normative performance studies for execution time. Accordingly, this would suggest that much longer timescales than those obtained in timed normative performance studies should be claimed for PSA. Clearly, establishing where error rates became tolerable for typical PSA claims has not been the subject of this study.

Nuclear human reliability performance data must be obtained on simulators, particularly for abnormal and emergency scenarios. However, the aviation sector has much higher rates of task repetition and therefore some data on ‘Glass cockpits’ has been obtained.

Massaiu and Fernandes (2017) have compared human reliability in the execution of, what they term, identification tasks on analogue and digital systems. This work has been undertaken on the Halden Loviisa nuclear power plant simulator using 16 licensed US operators answering thirty six questions. All operators were questioned concurrently. By ‘identification tasks’ they mean the identification of plant/parameters statuses on simulator plant panels in one trial and on tablet-based digital overview- displays on a second trial. The two forms of interface are characterised by illustration in their paper. This is reproduced below in Figure 3.



*Figure 3: Example analogue and digital displays for Emergency Core Cooling System (ECCS) and RCS (Massaiu and Fernandes, 2017)*

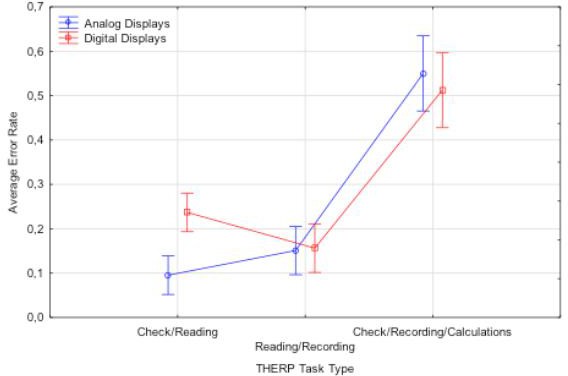
A typical question would be “Is the Residual Heat Removal (RHR) system line up correct for this condition?”. The experimental, so-called micro tasks, represent part of a real task that would be undertaken in the execution of emergency procedures. Tasks were chosen to include different forms of information processing: check/read, compare, calculate, and synthesise. All questions involved, as a minimum checking or reading information. Participants’ response time and accuracy were recorded. The study did not explore error effects in control actions. Their results show that the task type being undertaken is a stronger determinant of operator error rate than the human –system interface. Indeed, the effect of the two interface classes on error was statistically insignificant. This finding is important and is, therefore, worthy of further exploration.

Average accuracy rates and proportion of “Don’t know” answers, as well as response time, for each type showed some benefit of analogue displays over the digital displays. The participants were statistically significantly slower with the digital panels by around 20%. The tasks excluded navigation tasks between digital pages.

Noting the criticisms and concerns of Hickling and Bowie (op. cit.) Massaiu and Fernandes (op. cit.) compared their experimentally obtained error rates with comparable THERP classifications (but not the estimated THERP error rates). This revealed that digital displays had the lowest error rate for reading/recording tasks while for analogue displays the check/reading tasks showed fewer errors.

Both interface types had higher error rates for check/reading/calculations tasks.

The average error rates and error bands on those error rates are shown in their paper and reproduced below.



*Figure 4: Average error rates and error bands by display type and THERP task type (Massaiu and Fernandez, 2017)*

In their paper, Massaiu and Fernandez (op. cit.) also compared the outcome errors with the South Korean HRA method. As this method is not applied in the UK it is not discussed further here and, in any event provides more insights about the South Korean method than about analogue versus digital displays.

No direct comparisons offered between the THERP error rates and those outturn error rates obtained in the study. However, the authors suggest that unfamiliarity with the particular plant, the contextualisation of tasks and into subject competition as an artefact of the experiment may have had some bearing on this.

Massaiu and Fernandez (op. cit.) whilst noting the significant difference in response times for analogue and digital displays do not discuss the potential significance of this when multiple tasks, as in real procedures, might be undertaken. Setting aside the additional time for navigation between pages this may result in task time dilation for digital displays. However, this supposition must be treated with care because comparable studies had not been undertaken.

Moving from nuclear to aviation, the US NTSB published a study in 2010 (NTSB, 2010). They compared accident rates, fatal and otherwise between conventional cockpit and glass cockpit aircraft and studied the occurrence of accidents by, severity of outcome, time of day whether conditions, purpose of flight, phase of flight as well as pilot certification level and instrumentation rating (i.e. instrument flying certified or not). Their report is both succinct and detailed.

NTSB (op. cit.) concluded a decrease in total accident rates but an increase in fatal accident rates for the selected group of glass cockpit light aircraft when compared to similarly typed aircraft equipped with conventional/traditional displays. Overall, they concluded that the analysis did not show any significant improvement in safety for the glass cockpit study group.

They conclude for this class of aircraft and the types of pilots flying them that their levels of knowledge may not be sufficient or specific to Glass cockpits. And so, generalised training and guidance is no longer sufficient for pilots to operate glass cockpit avionics (within the selected group of light aircraft). They further conclude that simulator procedural trainers must be the way forward for training pilots to identify and respond to glass cockpit avionics failures or malfunctions – because these and the reaction to these cannot be safely replicated in like aircraft.

The NTSB conclusion is not without its critics however. Zimmerman (2018) suggests that the NTSB study did not control for exposure i.e. distance travelled. He provides a graph showing accident data normalised for distances travelled and concludes that the glass cockpit is as safe as the ‘steam gauge cockpit’. This data spans the period from the beginning of 2004 to the beginning of 2017. The accident data over that period can be clearly seen to be on a downward trend. Nevertheless, the NTSB investigations of accidents did identify things that could be done to improve procedures and training for the use of Glass cockpits and they certainly acknowledged the greater complexity of such interfaces systems.

In the light of the NTSB concerns of 2010 I have searched to see whether there has been an audit of Glass cockpits against the Federal Aviation Administration (FAA) HF design guidelines for multifunction displays (Mejdal et al., 2001). Such a review would be useful in providing insights not only on the design quality of Glass cockpits but on the suitability of the design guidelines. No such review has been found.

# Conclusions from Found Literature Containing HCI Data

It is clear that there is international recognition amongst nuclear regulators and operators that a necessity for HCI HRA data exists. Work, cited above, is being undertaken in South Korea to try and

et al. (op. cit.). Julius (op. cit.) and others have pointed up the importance of an internationally developed in recognised taxonomy. The work of Massaiu and Fernandez (op. cit.) illustrates the masterful analysis of data using different taxonomies which purport to not only classify data but provide some psychological characterisation of human error.

Therefore, in part, the problem of the taxonomy arises because no standardised accepted frameworks for the psychological phenomena involved with human error currently exist. The US NRC are actively and internationally seeking collaborators for the Scenario Authoring Characterisation and Debriefing Application (SACADA). For example, see Chang et al. (2014). The SACADA database for human reliability and human performance, allows for such multiple taxonomies.

In the opinion of the lead author for this report, the additional data seen, relative to that used by Hickling and Bowie (2013) does not provide sufficient substantive data from which an HCI oriented method might be created. The data of Massaiu and Fernandez (op. cit.) is not at the same level of decomposition in terms of HCI elements as the majority of that found (see Hickling and Bowie, op. cit.). However, it encompasses a number of interface elements within the micro-tasks as defined by Massaiu and Fernandez. Interestingly, and notwithstanding their methodological reservations, the order of magnitude for error would appear to be broadly similar to much of the more elemental data previously gathered for the 2010 HSE ND work.

However, the most significant finding from the work of Massaiu and Fernandez is quite clearly that error would appear to be a function of task characteristics rather than HCI designed elements. This is a very important finding and strongly indicates that the attempts within the USA to produce task characterisation methods for HRA is well-founded.

The work that follows on from the methodological trend started by Hickling and Bowie (op. cit.) has undertaken comparisons of outturn human reliability data obtained from simulators with THERP, SPAR-H and a South Korean HRA method. This data taken collectively renders challenges on the suitability of those established methods but, as already said, does not provide sufficient data for the development of an alternative method.

In the absence of highly developed theory and the corresponding taxonomy for human error phenomena, the 2016 work of Liu and Li (op. cit.) is considered to be of considerable importance. Their two-simulator-based study demonstrated a difference in perceived complexity between analogue and digital control rooms, with their results favouring the former. In addition, they, like Massaiu and Fernandez (op. cit.) have on a much smaller task span, shown that task times are longer in a digital control room.

The issue of complexity has also been acknowledged in the aviation sector by the NTSB (op. cit.) and, as was pointed out in the review of literature for HCI design it is implicitly acknowledged throughout the literature and websites that have been reviewed in that regard.

# The System of Assessment Principles

In order to arrive at a system of assessment principles that covers the issues identified by the literature review and from our own experience, it is necessary to revisit the fundamentals of the regulatory problem that ONR faces in the assessment of HCI and to consider what insights the searched literature provide on this problem.

# The Problem Facing ONR

ONR is faced with the problem of assessing claims about the qualitative HCI design that is involved in HBSCs. Further, those HBSCs can be probabilistically estimated by one or more human error probabilities alone or in combination. However, as in 2010, when earlier work human reliability assessment where HCI is involved was undertaken for HSE ND as part of a GDA, there are still no commonly agreed methods for quantifying human reliability where HCI is involved. The following quotation, found from the Internet, incisively summarises the situation that prevails.

*One of the abiding problems of safety critical ‘first of’ systems is that you face, as David Collingridge observed, a double bind dilemma:*

1. *Initially an information problem because ‘real’ safety issues (hazards) and their risk cannot be*

*easily identified or quantified until the system is deployed, but*

1. *By the time the system is deployed you now face a power (inertia) problem, that is control or change is difficult once the system is deployed or delivered. Eliminating a hazard is usually very difficult and we can only mitigate them in some fashion.*

*This is, as Collingridge noted, a paradox and it’s a paradox that lies at the heart of our problems in dealing with a lot of new technologies, from drone strikes to genetic screening. By the time we figure out what we need to regulate, there’s always an incumbency; happy, nay eager, to argue as to why we can live with the downside. Collingridge was pretty gloomy about our ability to work on the ‘prediction’ side of the dilemma using a Bayesian approach, due to the high level of ignorance (read ontological risk) which in his view invalidated Bayesian risk assessments.*

*When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult and time consuming.*

*David Collingridge, The Control of Technology (NY, St. Martin’s Press, 1980) [Downloaded from https://criticaluncertainties.com/2013/10/28/collingridges-dilemma/]*

The ontological problem for HCI is there is no clear definition of the factors that will affect human reliability in the use of that HCI. This notion is supported by the literature review covered here which shows there is no systematic evidence on what those might be. The continued efforts (largely from the USA) to identify such an approach, with or without HCI still continue.

Correspondingly, without those elusive insights, it is impossible to collect data of established relevance to HEP quantification. Therefore, a clear philosophy and strategy is required if ONR are to make their assessment of HCI and its role in HBSCs tractable, rational and credible.

# The Way Forward

Fortunately, the literature does indicate a way forward that can effectively sidestep this dearth of useful information and methodological bind. A number of findings are relevant to inform a way forward. These findings are set out below within the logical reasoning that has led us to our underlying philosophy for structuring HCI assessment principles.

*HCI design quality*

Massaiu and Fernandes (2017), of the Halden Project, have clearly shown that errors observed in NPP task simulations arise not from the characteristics of the HCI but from the complexities uncertainties and task requirements controlling the plant. Their work confirms a viewpoint that has been believed in Halden for quite some time now according to (Bye, 2017), who heads the Halden HF research and practitioner group.

This key finding must be carefully placed in context to properly understand its implications. The Halden Project has been designing HCI for NPP control since 1972. Throughout the intervening period they have had HF expertise contributing to the designs and experimental testing of those user interfaces.

Therefore, it is reasonable to assume that this finding comes about because of very high quality of qualitative HCI design. It is conservative, because there can be no clear measurement of the relative impact of their expertise against, say, that of a team embarking upon an HCI-based interface for process control for the first time. Such interfaces as those produced by the Halden Project are often described as ‘transparent ‘. That is, the interface itself does not significantly impinge upon the pursuit of the primary tasks of plant diagnosis monitoring control, reconfiguration testing isolation, etc.

It would be logically wrong to suppose that a lower quality of interface design could not lead to human reliability problems. Indeed, a catalogue of problems can be readily postulated including:

* Loss of train of thought by having to think about reconfiguring the HCI system to show different displays;
* Constraints in HCI design meaning that information needing coincident assimilation or comparison cannot be seen thereby placing a burden on human memory;
* Rarely used or inconsistent or incompatible dialogue/syntaxes which puzzle operators and give pause for thought;
* Generally poor quality of interaction design which results in errors without consequence for nuclear safety but which, nevertheless, introduce time delays and reduce operator self- confidence and trust within the interface.

Indeed, and firstly, INERI (2015) has systematically identified concerns associated with the introduction of new technology in NPP and elsewhere. The four bullet points listed above fit firmly within their resulting framework. Hallbert et al. (2017) have proposed methods and measures for characterising ‘in the loop’ operator performance that should address these.

*HCI design process*

Secondly, the literature review has shown a wide range of accepted methods and HCI design community expertise that can be applied to achieve high-quality HCI design yet which aren’t readily apparent within NPP design; for a comprehensive example see Interaction Design Foundation (2020). Consideration of these methods suggests that a level of HCI design quality with a corresponding minimisation of human error in the use of that HCI, is readily possible and exceeds the processes and approaches which have underpinned recent GDA submissions.

*Qualitative error assessment during design*

Third, the literature review has shown that systematic and logical human error qualitative identification methods can be applied to an HCI design at any stage in its development. For example, Thimbleby, (2013) or Baber & Stanton (1994) or (1996) and also Stanton & Baber (2005). Further, such a process has been proven by experience for safety critical systems in medical and aviation. Ideally, of course, this should be done proactively at the commencement of design. However, it is clear that the methods can be applied at any stage and will effectively reduce human error potential provided that design changes actually occur from the insights so gained. In practice, ONR can be confronted with HCI designs of some maturity but there is nothing in these formal qualitative human error identification and analysis methods which mean they cannot be applied retrospectively right through from paper-based design two simulator based exemplar.

*Proposal*

Taking the three key findings above together leads us to the following viewpoint.

We consider it would be unsafe to assess a safety case using HEPs that in some way purport to be related to HCI design characteristics. There is still an absence of structured, applicable HRA data relating HCI design features to human error. Such an analytical focus would miss the essential and now very widely endorsed finding in HRA research, that the human reliability problem arises in plant and task complexity and not the HCI design, *provided that design is of good quality that follows HF principles and findings.* However and very importantly, such an analytical focus for HBSC assessment can only be assumed to be valid if the construct of identifying and minimising human error potential in the HCI has been applied, by design development.

If the above statement is taken to be an absolute requirement then it may appear to ignore the principle of ALARP for design where design change can be deemed unnecessary if the risk or the consequences of that risk are considered insignificant to safety.

Unfortunately, when explored a little deeper a fundamental philosophical and methodological

First, if there is no numeric metric by which risk of error, as a function of HCI design, can be measured then there is no basis, beyond the qualitative opinion formed by the initial process by which the potential error was identified, for establishing whether, that error risk is likely to be high or low.

Accordingly there may be no logical process, open to independent (especially regulator) scrutiny that can be applied to sentence that risk. This perspective underscores the central importance of obtaining simulated tasks error data where and when possible.

Second, if it is accepted that an error in a task which is non-consequential for nuclear safety can, by HCI – induced task interdependencies, affect consequential tasks then this means that all interaction tasks on that HCI need to be treated with equal weight. Julius et al. (op. cit.) specifically draw attention to the potential for new actions and dependences to affect human error in their work and the immediately preceding discussion illustrates exactly that point.

These perspectives show there is a methodological trap in decomposing HCI design to the consideration of the tasks directly and demonstrably related to nuclear safety. The above logic shows that the human error reduction problem for HCIs a holistic one and over-decomposition of the problem space to single claims ignores the principal level at which there error induction problems occur.

One must therefore conclude that, unless attention has been given to human error minimisation in the qualitative design throughout the interfaces then there can be no logical demonstration that the HCI design reduces the risk of human error to an ALARP position.

The corollary of this conclusion is that, provided the overall HCI design has been consistently designed and shown to minimise error then, reliance in safety case assessment can be placed upon a clear demonstration by the organisation making the safety case that the numerical HEP estimates are based upon the characteristics of the primary plant control tasks including the complexity of those tasks and the complexity of the physical phenomena involved within them. Fortunately, this leads a HRA assessor into much more familiar HF territory and reliability quantification experience which is underpinned by understanding and literature going back as far as 50 years.

# Principles: their Scope and Definition

The following principles are based upon the argumentation given above, and are derived from both searched literature and the experience of the authors in the design and assessment of computer based user interface systems.

We have couched the principles as being applicable to Human Computer Interactions or Interfaces. This redefinition recognises that human error can arise separately or in combination from the design of interactions or the interfaces upon which those interactions occur. Because the developed principles are applicable both to interactions and interfaces we have coined this acronym to avoid the ambiguity in the more widely used acronym ‘HCI’ which can be used or read as human computer interaction or human computer interfaces.

The assessment principles for qualitative design focus upon the, predominantly psychological, HF issues associated with such systems that could affect human performance (the timeliness of undertaking and completing a task) and the human reliability (failing to undertake the task or undertaking it incorrectly).

Such human failures can be induced by the design of the system. Design failures that induce poor performance or error can arise in the design of user interfaces: that is what can be seen or done on a screen. They can also arise from interaction failures: that is the dynamic aspects where the interface changes of itself or through the sequence of user manipulations with system based cues or feedback– sometimes referred to as syntax.

The assessment principles for quantitative human reliability prediction focus upon the necessary reliance upon high quality qualitative design and the principles underpinning processes and methods by which ontological uncertainty concerning human error and HCI can be addressed in a probabilistic safety assessment.

In 2020, at the time of writing, the anticipated role for HCI in nuclear systems continues to grow. Accordingly, these principles are applicable to systems having roles that might be described as fulfilling one or more of the following functions and have been derived where possible, from literature addressing such functions can include:

* The display of information to indicate the need for human control actions;
* The provision of control affordances to implement human actions changing the state of the plant or equipment within it;
* The display of information informing users of the occurrence of automated decisions or their success, failure, resulting plant state and any indicated human actions resulting from such failures of automated actions;
* The display of information to inform human diagnosis and subsequent decision making.

The above list is one of distinguishable functions and should not be read to imply that a particular system must only be intended to carry out only one of those functions.

# Qualitative and quantitative HCI Safety Assessment Principles

The principles given in the following subsections are structured at two levels; Superordinate principles and subordinate principles. The superordinate principles are, we believe, fundamental to the assessment of HCI safety. However, whilst logically sound, they stand the danger of being too general to inform and guide specific assessment of the complex problem space that is HCI. Accordingly, the subordinate principles provide more detail and relate the subordinate goals needed to be achieved if the superordinate principle is to be complied with.

Nevertheless, these remain principles and necessarily imply the application of a considerable level of HF design guidance and applicable methods if they are to be met in practice.

In respect of the subordinate principles we believe them to be written in a sufficiently clear fashion that they could be readily converted from the language of principle to the language of factors that an ONR inspector might wish to consider.

As a primer document, we believe this approach to be appropriate for this report because it provides more detail around the superordinate principles. This not only helps to clarify the intent of the superordinate principles but by the overall structure helps inform an ONR inspector of an overall approach and safety concerns for HCI.

Before setting out our proposed principles an illustration is provided of how a subordinate principle can be converted into an amplifying paragraph for a superordinate principle. However, the volume of information and the relationships between that information lead us to recommend that ONR consider the development of a technical assessment guide supporting the superordinate principles.

An example subordinate principle:

### “Project plans and development processes should recognise that the capture of HF specifications and requirements is complex and is unlikely to be complete at the outset of a software design and development project. Accordingly plans and resources should be sufficient to address emergent human factors.”

An example of equivalent inspector guidance text:

*“At an early stage in an HCI development project, an inspector may wish to consider how thorough the capture processes for HF specifications and requirements might be. In addition, they may further wish to consider how such specifications and requirements can and will be amplified in subsequent project stages. It may be appropriate to consider whether emergent HF requirements, especially those occurring as an artefact of HCI design characteristics, are identified in more detailed design development processes and, where applicable, reapplied to other parts of that same design.”*

Each superordinate principle is presented in bold italics and subordinate principles in italics. Where appropriate, justification and commentary is also given. In some instances, a justification stands alone for super and related subordinate principles. In other cases it has been appropriate to provide further justification for particular subordinate principles. In some instances where we consider that a particular technical pointer might be required text has been added following the form of “an inspector may wish to consider”. For economy of overall content, we have used the structures of presentation in a flexible way.

One further critical point must be made in order to explain the language in the principles that follow. ONR, quite properly, consider both qualitative and quantitative human reliability in making their assessments of a particular case. We have very consciously added the phrase ‘human performance and human reliability’ in some instances. We have done this because an important measure of task disruption by an HCI is the time of the interactions in meeting the demands that HCI impinge upon the primary task of dealing with the NPP. Accordingly, it is important that measures of human performance, i.e. efficiency in successful task performance are seen by the regulator to have been applied in addition to design and assessment measures that demonstrate the avoidance of failure.

# Qualitative Human Computer Interaction and Interface Design Safety Assessment

### HCI Design 1: Dutyholders shall follow recognised systematic project plans and design processes for the design and development of Human-Computer Interfaces (HCI) throughout all phases of the design lifecycle. Further they must demonstrate that those design processes incorporate analytical activities to achieve understanding of the operator tasks with potential to affect nuclear safety. They shall as an integral part of those processes involve personnel who are competent to predict and measure human performance and error. Dutyholders will also ensure that end user interests are accommodated within their designs.

Justification: the literature review has shown that a wide range of HF design and analysis methods are available for addressing software driven user interfaces. The search has also shown a general agreement that software driven user interfaces have strong cognitive components and by their breadth and potential complexity require the application of specific methods that ensure proper integration of HF within the broader software development project.

It will be seen from the literature review that there is no immediate prospect of human reliability quantification methods to inform and drive qualitative design processes. Accordingly, any HBSCs must depend upon the overall high quality of software driven user interface designs that directly or indirectly affect them.

The principle specifically places an emphasis on an ‘even-handed’ approach to design for nuclear safety and non-safety related interactions and interfaces contained in the system. This is done because the complexity and breadth of many such interface systems carries with it a correspondingly greater potential to induce negative transfers of training and expectations from non-safety-related interfaces to safety-related interfaces. In addition, many such systems will be used ‘day in day out’ and so give strong reinforcement to expectations about human computer dialogues. Accordingly, to minimise the risk of a user being distracted in safety-related actions design consistency of interaction objects and interaction dialogues should be required throughout an HCI design.

The breadth and depth of extant HF knowledge on methods and design principles means that high quality design will only be achieved if HF is properly deployed and thorough involvement of end users is incorporated into the development and assessment of HCI designs.

The following subordinate principles support the main principle and address specific aspects where the achievement of acceptable levels of human reliability could be compromised by inadequate methods or insufficient analytical attention.

*HCI Design 1.1: Dutyholders shall implement initial project capture processes shall result in comprehensive specifications and requirements for human factors that will influence the usability and efficiency of the finished interface designs to achieve the diagnostic, monitoring and control functions required.*

Justification: Repeated studies over a period now spanning more than 30 years have shown that software project failures stem from inadequate specification and requirements capture processes. In HF it is also widely recognised that failures in the capture of user requirements or specifications to meet

human needs also result in inadequate software. Such inadequacies can be error inducing and fundamental features required to support human tasks may otherwise be absent entirely.

Expectations on processes or methods: An inspector would expect to see explicit HF involvement at the commencement of the software project. Specifications at commencement should ensure that fundamental HF requirements will be met. These would include interaction dialogue and interface aspects such as visual legibility or differentiation, control actuation feedback, speed of feedback provision and so forth. Initial requirements would include the capture of process monitoring and control task requirements when using the HCI.

*HCI Design 1.2: Project plans and development processes shall recognise that the capture of HF specifications and requirements is complex and is unlikely to be complete at the outset of a software design and development project. Accordingly, plans and resources shall be sufficient to capture additional specifications and requirements as the design progresses and to address emergent HF identified by earlier HF processes.*

An inspector would further expect to see the more detailed development and iteration of those specifications and requirements as the project continues. Where the precise requirements cannot be determined by standards, design norms or design guidelines in task analytical or other appropriate methods should be seen to be applied to resolve those difficulties.

*HCI Design 1.3: Systematic HF processes should be in place throughout the design development to ensure that those specifications and requirements are met and can be demonstrated to be met.*

An inspector would also expect to see these initial project definition processes being later tested to verify their suitability and validate their implementation by showing the minimisation of human error and support for project required levels of human performance (for example, in task execution time).

### HCI Design 2: Dutyholders shall identify and consider the error potential for those user interactions with HCI that are relevant to nuclear safety in all relevant operational contexts. Such consideration shall include the cognitive elements of interaction and how these might lead to error. Cognitive error prediction methods must be justified and demonstrated as suitable for use. Given the reliance that must, of necessity, be placed upon the insight of the error analyst, their analyses should be informed by a structured and justifiable list of qualitative factors that invoke the consideration of task complexity and other relevant cognitive error causal factors.

Justification: The breadth and depth of HCI systems means there is great potential for inconsistent interface design or interaction dialogues within an NPP control room. Such inconsistency needs the well- recognised phenomenon of negative transfer of training. Thus, if non-safety-related tasks are not closely scrutinised and ensured to be consistent with safety-related tasks then greater error potential exists.

To be assured of success from the outset of a project will require the close collaborative working of software development specialists with HF professionals specialising in the identification and mitigation of human error.

Inspectors may wish to consider the mechanisms by which the required consistency will be arrived at when a new design is at an early stage. They may further wish to consider how HF design principles have been invoked to influence interface or interaction designs to be error minimised from an early stage.

For a mature design, inspectors may wish to consider how inconsistency between different interface and interaction elements can be identified and what steps have been taken to ensure that error induction is minimised.

### HCI Design 3: Dutyholders shall ensure that the design intent for HCI systems and their relationships to other user interface systems is fully developed at a very early project stage This intent shall consider, throughout design, safety assessment and commissioning, the reliability of HCI systems relative to other user interface systems of higher integrity and the impact this will have on HBSCs. Dutyholders shall demonstrate that the intended design will lead to the reliable transfer by users of plant control or monitoring tasks from one interface system to another when required. Such a demonstration will require confirmation that HCI system failure modes and effects that could compromise human functions required for nuclear safety will be promptly detected by users and that human factors have been properly addressed to support a prompt transfer to another system, when required.

***Such human factors could, amongst others, include user unfamiliarity, information paucity, increases of workload, changed human resource demands or other potential sources of additional task difficulty.***

Justification: HCI systems are increasingly used for day-to-day use and provide comprehensive information which is considered powerful by users. However such systems, due to their architectural and software complexity can only support very low technical and human claims, or no claim at all for nuclear safety. The day-to-day reliance upon such systems can result in reluctance to transfer their task focus and attention to other sources of information. Such behaviour has been seen by humans using information systems ranging from the simple, such as a ‘car sat nav’ to the complex, such as aviation integrated cockpit information systems. This may be particularly prevalent in the nuclear case, it is likely that the other ‘second or third line system’ will provide a much more restricted set of information to users. Where an HCI system also provides automated diagnosis and decision-making or provides significant support to decision making by users there may be significant reluctance to abandon that system in the event of its failure due to the perceived or real additional demands placed on the operator by the back-up system.

Expectations on processes or methods: An inspector would expect to see a suite of interface systems being designed as a whole rather than as separated interface systems brought together at a later date and with continual attention to the requirements for reliable task transfer.

An inspector would expect to see clear demonstrations early in the design process that the designers were mindful of the challenges to human performance and reliability that these issues raise. For example, by considering the design and performance evaluation of different interface systems, especially, but not confined to centralised control facilities, as an integrated whole. Demonstrations of acceptable reliability would show that users can make a smooth and unchallenging transfer from one

system to another and that any decrement in human performance resulting from such a transfer will be acceptable to the safety case.

As the design progresses, an inspector should expect to see early simulations of such transfers to ensure that they will occur with the levels of human reliability anticipated by the design and, if not to make design adjustments well before the design of the suite of interface systems is finalised.

An inspector should expect that the assessment and evaluation of such task transfers for nuclear safety, will include the progressive application of design principles, design assessment and task simulations of those transfer processes. These would seek to establish the clarity with which the need to abandon one system and use another, and the likely user confidence in so doing, are clearly demonstrated. Such demonstration should be continued throughout the design, commissioning, operator training and commencement of system(s) application by using increasingly high levels of scenario and task fidelity.

Where early stage data about inter-interface transfers is held by system architects, designers, human factors specialists or others, that data should be freely passed to those tasked with training, commissioning or running the facility so that effective training and user performance evaluation is possible.

### HCI Design 4: Where users place some reliance upon the availability of a non-claimed HCI, or one having a level of integrity broadly comparable to humans performing those same tasks for nuclear safety, dutyholders shall ensure that they are authorised and quickly able to disable the system from making: automated diagnoses intended to inform users safety-related actions, or to automatically make those actions. This accounts for cases where users have good cause to believe that the system is no longer reliable or is otherwise not relevant to the situation in hand. Because of low integrity or high complexity of HCI systems, such user decisions may need to be made based upon suspicions arising from an absence of data or system behaviour rather than by positive confirming evidence.

Justification: Irrespective of the elegance of design and the integration of different user interfaces, experience suggests there may well be a reluctance to make necessary transfers from low integrity information-rich HCI deemed to be highly informative, if operators are not clearly and very positively empowered by training and procedures to do so. Even so, such reluctance may still persist to abandon a system of questionable dependability. Further, it is reasonable to suppose that such potential reluctance may be further reinforced if the available cues for such a transfer arise from an absence of system behaviour or information.

Expectations on processes or methods: Therefore, inspectors should expect to see clear operating principles endorsed by plant management, trainers and regulators alike that unambiguously set out the parameters for making a decision for transfer. Such decision making parameters need to be realistic bearing in mind the fault evolution timescales that may be involved. Confirmation for such decisions must not rely upon time consuming, detailed or complex fault finding and interrogation of the information system. They should also not rely upon the authority of senior plant personnel who may not be immediately available and are likely not to be fully informed by both prior knowledge and real- time information of the technical situation requiring the transfer decision to be made.

Inspectors should also expect to see that simulator-based and other training rewards conservative decision making to transfer to another system in the face of uncertainty, rather than that which might be deemed optimal for a particular fault scenario.

### HCI Design 5: Dutyholders shall use methods in the design to develop HCI dialogue or interface design that ensure the efficient achievement of the human thinking and action tasks that are required for achieving nuclear safety. They shall systematically and thoroughly address the possibilities of human error in those tasks and demonstrably eliminate or minimise the likelihood of such errors as far as is reasonably achievable. If human error likelihood is assessed to still be of potential significance for nuclear safety then the Dutyholder shall incorporate within the HCI design means for the prompt and effective recovery from such errors.

Justification: The breadth and complexity of software driven interfaces and the hardware that supports them, means that there are many potential sources of error induction. In addition, the lack of universally accepted design conventions for interactions and interfaces coupled with the current rapidity of creativity in change in software platforms means that such conventions are unlikely to emerge for some time to come.

As the principal concerns for nuclear safety are the support of required human performance and the avoidance of human error with undesirable consequences for safety, then it becomes necessary to design such interfaces and dialogues for success so that actions required for nuclear safety are completed in a timely manner and at the same time designed so that error potential is minimised.

In traditional/conventional user interface design, there has been a strong tendency to examine the possibility of human error only after the design has been finalised. If this same approach is taken within software development then the late discovery of error induction may require many and widespread changes elsewhere in the software in order to ensure consistent interface or dialogue design across one or more information system. The costs of such widespread changes have, historically, proved inhibitors to making such changes: so compromising human performance or error likelihood.

Accordingly, the possibility of human error and its minimisation or elimination by design must be an integral ongoing part of a software project from the outset. In addition, the consideration of error from the outset provides longer time for understanding the mechanisms for error induction and providing demonstrable defences against such errors.

*HCI Design 5.1: Reliance must not be placed alone on compliance of interaction dialogues or user interfaces on published design standards to minimise the likelihood of human error. HF task and error analytical methods shall be used to systematically and objectively demonstrate that tasks in general and especially those required for nuclear safety will be efficient and that error possibilities have been eliminated or the likelihood of their occurrence minimised. Such analyses shall be done in a manner that considers the facets intrinsic to the use of advanced HCI such as system navigation, processing and management of information, cognitive activity leading to interface input.*

*HCI Design 5.2: Dialogues requiring interactions between two or more screens and interactions within screens/pages shall be designed to be efficient in terms of minimising the time required and human*

*resources needed. They shall be designed with a clear and demonstrable intent to minimise the level of time and attention needed to reconfigure the interface system and hence minimise the level of distraction from the primary tasks those interface support and that focus on the status of the plant’s nuclear safety.*

Justification: nuclear operating history, especially that within real and simulated faults, has emphasised the importance of operator attention being directed to the correct diagnosis and handling of plant faults. Accordingly, the aim of the above principle is to ensure that the time and attention taken from the primary task are minimised in order to, correspondingly, minimise the risk of losing a train of thought, especially at a time when stress levels may be higher than desirable for best human performance and minimised error.

*HCI Design 5.3: The navigation map between screens/pages/ hidden information layers should be readily understood and remembered by system users with minimal training and available pathways should be explicitly and clearly indicated.*

Justification: Efficient navigation not only minimises time off the primary task of plant control or monitoring but having sufficient simplicity that it is memorable helps to ensure that users will not waste time pursuing incorrect routes or, alternatively, undertake the correct actions but on the wrong quadrant/system/subsystem/equipment train and by so doing reduce the level of nuclear safety intended by design.

*HCI Design 5.4: Consistent design should be sought between the syntax and sequences of interaction dialogues for nuclear safety and those required for other purposes. This is necessary to ensure that uniform user expectations can be achieved and any negative transfers of training with consequent error potential avoided.*

Where user interactions relate to fulfilling requirements for nuclear safety, and where there may be significant levels of user uncertainty, short timescales or other psychological stressors, at the time of fulfilling those requirements, then the need to reconfigure the user interface should be minimised in order for users to undertake actions for confirmation of diagnosis, making a diagnosis, situation monitoring or control.

*HCI Design 5.5: Dutyholders shall ensure that the design of supervisory checking by one person upon another is effectively supported by the HCI systems in use.*

Justification: The introduction of HCI can alleviate the need for operators to walk from panel to panel to undertake claimed safety actions. This removes an explicit cue for a control room supervisor that indicates the significance of the plant about to be inspected or manipulated, thus making supervision harder.

*HCI Design 5.6: Specific provisions should be in place within an HCI to support supervision and overcome this potential human performance decrement and the overseer and task performer should be able to have a clear understanding of why and when such checks are required.*

Expectations on processes or methods: The inspector should expect to see a design approach that allows the person who is to be occasionally checked to know, through explicit cues that this is the case

so that the checking need can be called to supervisory attention if necessary. It should be clear to the task performer when sanction to continue can only be given when such a check has been satisfactorily completed. Where there are sequences of actions that constitute one or more HBSCs, then the overseer should be able track the actions of the other without requiring continual interruption for verbal or other forms of interpersonal communication.

The inspector should expect to see a procedural process (paper or electronic) that is focused on HBSCs and other important tasks and is not made over-onerous by avoidable task interruptions*.*

### HCI Design 6: Dutyholders shall ensure that the scope, capabilities, limitations and failure modes of HCI systems that inform nuclear safety related human decisions or action shall be formally identified. The limitations and failure modes shall be systematically considered to establish how they could compromise nuclear safety were system or subsystem use to continue beyond the system’s scope or when in failed states. Further, Dutyholders shall ensure by integrated design that the scope, capabilities and the failure potential of those systems, are made clear to users by system training and operating documentation comprehensively and where reasonably practicable effectively supported by real-time indications.

***Moreover, there shall be explicitly pre-defined absences of indications, system behaviours or data unavailability where real time indication is not practicable. These shall be used to define when a system might be in error states or has failed in a way that could adversely influence user acceptance of its decisions, advice or indications affecting nuclear safety. This should all be done irrespective of whether the HCI is intended for automated diagnosis, diagnostic, support, fault detection, human monitoring or human control implementation.***

Justification: Users of such systems must be able to clearly recognise the continued usefulness of any such system at any given moment and, therefore, to confidently abandon it when it manifests error or failure. That abandonment must be clearly defined so that sanction for abandonment is emphatic. To minimise time of the primary task of attending to the plant, it should be possible for the user to disambiguate the situation easily and quickly, using a pre-structured task, by comparing their own situation interpretation or intended decision with that put forward by the machine.

*HCI Design 6.1: Where HCI decisions control plant state directly or by human implementation, the basis for that decision should be clear and the impact of that decision upon plant state easily reversible by the user if they deem that decision to be incorrect. This principle shall not be implemented by the system user or supervisor being obliged to act a gatekeeper for each decision and endorse it before it is enacted. Further, the user interface must confirm that the machine decision or human implementation has, or has not had the intended impact on the plant. Reversal of plant state shall not be required for nuclear safety on timescales shorter than 30 minutes.*

*HCI Design 6.2: Where reasonably practicable, the system capabilities should gracefully degrade. For example, automated diagnostics or alarm system processing should not be set to a failed state if only one of a number of parameters on which the decision is made is unavailable or of questionable validity.*

*Instead, the system should make clear via the page affecting their acceptance, or rejection of that decision that the automated determination has more limited validity than the norm.*

*HCI Design 6.3: Users of automated diagnostic or decision making engines shall only be required to arbitrate upon or endorse diagnoses or decisions before they are implemented as manual or automated control actions in circumstances where the calculated level of uncertainty or confidence index in that decision threatens to challenge the limit of risk acceptable to the safety case. The bases for that calculation of uncertainty should be made clear to the user. The machine must make a clear distinction for the user between uncertainty arising from the absence of one or more information inputs and uncertainty where all inputs are present but the machine cannot disambiguate the situation with acceptable levels of confidence. Where a missing information input can be resolved manually it shall be possible for the user to readily input that missing information at the time when it is required should they deem it appropriate so to do to re-run the decision process.*

*HCI Design 6.4: Where an automated diagnostic or decision making system that can influence nuclear safety is able to learn by experience, then that learning should not be allowed to automatically extend the scope of the safety case scenarios or fault sequences that it can address without a formal revision and acceptance process. That process must include a clear description of the extended scope of applicability that is demonstrably conveyed to system users and is shown to be easily understood and reliably remembered by them. To ensure that trust in the system by end–users is maintained it shall be possible for the user to easily scrutinise the scope of system applicability to fault scenarios. Learning that improves the sensitivity to indications or the resulting accuracy of decision making within existing scenarios can be permitted. However, changes resulting from such learning must be regularly reviewed and checked and shown to be acceptable to the safety case. In the event they are not then it must be possible to undo that learning and revert to an earlier state of machine ‘understanding’*.

### HCI Design 7: Dutyholders shall ensure that systematic task and error analytical methods are used to consider cognitive factors and explicitly document any cognitive issues that are so identified. A complete record of such identified issues shall be maintained and a record made of how they have been resolved by dialogue, interface design or other measures.

Justification: Cognitive error prediction methods must be justified and demonstrated as suitable for use. Given the reliance that must, of necessity, be placed upon the error analysts insights and expertise, their analyses should be informed by a structured and justifiable list of qualitative factors that invoke the consideration of task complexity or other cognitive error causal factors.

Inspectors should expect to see awareness of the cognitive issues associated with such systems (for example, as tabulated in INERI (2015). Amongst other things these include: information complexity, loss of plant status awareness through automation, physical and cognitive workload, system opacity, the reconciliation of different diverse information sources, levels of trust in digital systems

### HCI Design 8: Duty holders shall ensure that information within the HCI system which is not relevant to the defined duties of an HCI user shall not require their time or attention. This minimises the potential of the system to induce avoidable distractions from the performance of tasks involving nuclear safety involving that or any other interface system. In particular, forms of manifestation that particularly demand attention such as alarms, flashing or screen overlays must not be permitted.

Justification: Many HCI systems have the capability to compile and transmit large quantities of data. There is a repeatedly seen tendency to instrument process variables and to provide them with indications and alarms to operators in centralised locations whether or not these are relevant to their tasks. Past incidents have repeatedly shown that an overabundance of information can camouflage that which is important and blur an operator’s priorities and objectives.

Expectations on processes or methods: This principle has a close relationship to the principle concerning HCI scope, limitations and failure modes. It requires a judicious balance, by design and analysis, between them in order to ensure that the requirement for HCI system or subsystem abandonment occurs quickly and cleanly.

Inspectors should expect to see clear principles for information ownership. For example, data that is only relevant to technicians maintaining the HCI system should be directed to other user interfaces, possibly in other places and not to those user interfaces directly involved with the monitoring and control of nuclear safety.

Inspectors should expect to see systematic analysis to identify information that is relevant and irrelevant to those directly involved nuclear safety.

Inspectors should expect to see that information requiring immediate attention but not readily interpretable by those monitoring and controlling nuclear safety is directed elsewhere. However, should that information be relevant to those monitoring and controlling nuclear safety then duty holders should be able to show analysis and design solutions which make it readily interpretable to such users.

# Safety Assessment Principles for Human Computer Interaction and Human Error Probability Estimation

### Human Error Estimation in HCI 1: Dutyholders shall ensure that the incorporation within probabilistic assessment models of HBSCs involving the use or abandonment of HCI, whether a claimed system or not, in favour of another should be conservatively made by taking full account of the wide levels of both probabilistic epistemic and aleatory uncertainties, that exist around the making of such claims. Where such claims are potentially important contributors to the probabilistically assessed risk for nuclear safety, then the impacts of that uncertainty on the overall risk should be evaluated and its acceptability upon the assessed level of risk fully considered.

Justification: given the number of human factors that can be identified that might inhibit the transfer from one user interface system to another, there is likely to be a probability for failure to transfer which could be of significance to a large number of HBSCs than would otherwise be made.

Accordingly, the probability of such a failure must be explicitly factored in to risk models so that the significance to risk of such a transfer can be properly understood.

Expectations on processes or methods: An inspector should expect to see risk modelling activity being undertaken at an early project stage and in parallel with the formulation of the design intent for a suite of interface systems that might be used together or sequentially in support of HBSCs. Without the

early stage. However, as one failure to transfer may potentially impact upon a number of HBSCs it is likely that risk model logic alone will indicate the importance of such a dependent failure upon subsequent claims within task sequences.

### Human Error Estimation in HCI 2: Operator interactions with HCI and their direct effects upon the reliability of HBSCs must be adequately and demonstrably addressed. Direct effects arise where human-computer interactions are integral parts of tasks undertaken to fulfil a HBSC.

Justification: arguably, this principle is already contained within existing ONR SAPs. However, it provides a basis for setting out some subordinate principles that are given below which we believe are of particular importance to HCI.

*Human Error Estimation in HCI 2.1: HBSCs involving tasks with integral human computer interaction must rely upon demonstration of the systems transparency in use. Such a demonstration must rely on the suitability of the qualitative HF design processes and methods applied to that system and also a clear demonstration that potential human errors arising as an artefact of that system design or as a function of the tasks performed upon it, have been identified, then minimised or eliminated.*

*Human Error Estimation in HCI 2.2: Proportionate to the stage of system development, interaction error data that characterises the propensity of the system as a whole to induce user errors must be collected. Initially, this will be done to provide a generic index of system error induction potential. Later it will be focused by the developing safety case on the specific HBSCs. The likely tolerability of those error likelihoods to the final safety case must be analysed. Irrespective of the design stage, analyses and projections of tolerability must take account of the currently unanalysed fraction of human error contribution. Where those levels of error might not be tolerable then there must be convincing design processes and methods in place to arrive at human error likelihood that will be as low as reasonably achievable.*

*Human Error Estimation in HCI 2.3: Given that human error arising as an artefact of the system design has been convincingly demonstrated to be as low as reasonably achievable, then the quantification of human reliability supporting HBSCs should focus upon errors in the specific cognitive task demands to fulfil the claim made. In these circumstances, the fractional contribution of human computer interaction to errors in those tasks can be generically estimated based upon error data gathered either during design development or based upon simulator operator data in an already completed system.*

### Human Error Estimation in HCI 3: Operator interactions with HCI and their indirect effects upon HBSCs must be adequately and demonstrably addressed by HCI design. Indirect effects arise where human computer interactions with a system inform of the need for users to interact with another, user interface system.

*Human Error Estimation in HCI 3.1: Where a human computer system as a whole or information contained within it must not be relied upon, due to its designed or assessed level of reliability or integrity, for the execution of tasks involved in HBSCs for nuclear safety, then any information that could be used within those claimed tasks should be clearly demarcated as unsuitable for that purpose on the system interfaces.*

*Human Error Estimation in HCI 3.2: Where a human-computer system or a portion of such a system must be abandoned due to failures within it, then claims for the use of another system or system portion must be substantiated by a convincing demonstration that the required abandonment will be both unambiguously annunciated and that abandonment will also happen quickly and with sufficient reliability. It will not be sufficient for a claim to rely upon the clarity and salience of that annunciation alone but it must also relay upon the factors affecting the users’ states of mind when they receive that annunciation.*

*Human Error Estimation in HCI 3.3: Where a system is unsuitable or through faults, becomes unsuitable for use in tasks fulfilling HBSCs then the transfer of required tasks to another user interface, whether it be computer-based or not, should not result in an unacceptable decrement in human performance and reliability that could affect the consequent HBSCs.*

*Human Error Estimation in HCI 3.4: Claims for taking over the decisions intended to be made by automated decision making systems, upon faults in those systems, should only be made where a clear and convincing task-based demonstration has been made that users do not place excessive trust or over- reliance upon such a system. Such systems must make abundantly clear where and when human-based decisions are required either because they are always beyond the scope of that system or fall outside the capabilities of that system due to faults or failures. Where such take overs are claimed, consideration must be given whether the information on which such decisions are made should be simplified in order to assure the level of human reliability required, or likely to be required, by the safety case.*

*Human Error Estimation in HCI 3.5:There are no established methods for predicting the probability of cognitive errors based upon published data because the psychological and task-factors affecting such errors are complex and there is corresponding diversity in psychological theories and models. Therefore, task simulations or actual task execution in practice must be relied upon to provide suitable data. In addition, the levels of demanded human reliability may require a considerable body of data to be collected in the probability of errors relevant to human based safety claims are to be predicted with statistically evaluated confidence limits. Practically, such data is most unlikely to be available until a user interface design is mature and stable. Premature attempts to provide estimates may result in premature ‘fixing’ of interaction and interface design that, with the befit of experience, do not achieve human error rates that are as low as reasonably achievable.*

*Human Error Estimation in HCI 3.6: Given that HBSCs with corresponding error probability estimates need to be made at a stage where a mature user interface design may not be available, then a catalogue should made of epistemic uncertainties that cannot reasonably yet to be addressed within claims and those where incomplete information to inform the analysis of such claims, remains unavailable. To the extent reasonably practicable, analytical processes for such claims must seek to minimise epistemic uncertainties for the completed safety case so that the uncertainty which remains is aleatory (confined to known unknowns).*

*Human Error Estimation in HCI 3.7: Where a user-computer interface system is deemed to be ready for system commissioning, simulator application, or use in practice, then data collected in system development and continuity uncertainties should be passed on to those undertaking the subsequent life cycle stages. Data collection should continue so that sources of epistemic uncertainty that remain incompletely addressed are*

*resolved in the light of actual or simulated task operations. Such sources will, as a minimum, include the usability of procedures and the impacts of training and knowledge. The collection of such data must be used to inform the revision of the safety case but also used to make design changes to the cognitive task environment that will improve human reliability in the execution of claimed tasks.*

Expectations on processes or methods: The gathering and promulgating of data for human computer interactions and cognitive adjuncts is immature. Accordingly, there may be ontological uncertainties (unknown unknowns) associated with the phenomena involved. Therefore an inspector should expect to see an inwards (experimentally based) and outwards looking (based on literature and project experience elsewhere) watching brief. This should be maintained throughout the system lifecycle to establish whether there are new discoveries or factors not captured in their extant data and which might affect the HEP estimates associated with HBSC.

# Discussion of HSBC Submissions involving HCI and Human Error

* 1. **An Essential Caution**

This entire report places considerable weight on the difficulties in attempting to use HCI derived data for HRA purposes in the context of a PSA. Nevertheless, the authors recognise that a temptation might exist or attempts might be made to apply the data given here. This section restates and reinforces the reasons why this should not be done.

* The limited literature search on which this work is based has not comprehensively and exhaustively gathered all relevant data that may be within the public domain. Nor has any experimental work specific to this study been undertaken. Only anchors of best and worst human reliability using HCI estimates have been sought.
* Sufficient data has been gathered to suggest that the following statements might apply:
  + Human reliability using HCI may be as bad as unity
  + Human reliability using HCI might be as good as a 1/1000 error rate however available data appears to be truncated.
* The published papers are consistently unable to provide qualitative descriptions of task difficulty and user interface complexity, therefore, published human error data cannot be calibrated against qualitative statements describing the task, personnel or context. Accordingly, it is not possible to selectively apply data for central estimate and error-bounded HRA.
* Data has been obtained largely from NPP simulators which may not reflect ‘state-of-the-art‘ HCI. However, it cannot be supposed that ‘state-of-the-art’ HCI affords better performance as HCI capability now provides greater potential task complexity than hitherto.
* A taxonomy for task complexity/difficulty/ease has yet to be derived for HCI-based tasks. Experts assert that HCI-based tasks involve more cognition than NPP gen 1 and gen 2 ‘knobs and dials’ interfaces. Therefore, the credible gauging of where human reliability might sit within the postulated distribution is not yet feasible.
* Data has been obtained from a number of different countries each potentially having different working practices and methods for team and task management as well as likely differing population stereotypes with respect to human machine interfaces.
* The taxonomic frameworks used for the gathering of data that referenced authors have published differ. Therefore, the grouping of data given in this report as ‘themes’ can only provide a broad impression of reliabilities not focused indications.
* Whilst data from a study may be linked to a theme such as procedures, the error probability derived may well be determined by some other factor such as task complexity that was not assessed nor considered as a variable of interest in the study. This a function of the data not being collected via carefully designed experiments in which variables that might affect error rate are controlled in order to examine the effect of an independent variable of interest.
* Formal experiments examining the magnitude and direction in the influence of variables upon human reliability and task performance remain very much in their infancy. In part, this arises from commercial protection issues, but also from the multitude of tasks that can be supported and implemented using HCI and the many different and creative ways in which interfaces might be designed to satisfy the necessary functional requirements.
* Researchers have gathered their data informed by different research objectives e.g. verifying the suitability or otherwise of an existing method to address HCI-based tasks or to seek to establish a descriptive task taxonomy which might be used in the future for categorising and estimating human reliabilities.
* Some tasks which are functionally and logically expected to affect human reliability and which can be central to the making of a safety case are only weakly supported by published data. In particular, the following of procedures whether on or off-line is not well supported. In addition, screen-based systems centrally depend upon successful navigation between different pages. Navigational data is only available for 1980s based systems and may only be marginally applicable with modern sophistication involving hyperlinks, three-dimensional navigation as well as point-and-click tools that were not available when that data was gathered. Nevertheless, the constraints of systems at that time, arguably, may make navigation tasks simpler than on some modern systems.
* The increasing sophistication in the processing and display of alarms has potentially changed human reliabilities associated with tasks that are invariably central to the making of a sound safety case.
* The main body of this report repeatedly emphasises that assurance of human reliability using HCI can only be obtained through comprehensive design reflecting HF principles and, for safety critical applications, the need for the exhaustive testing by scrutiny and trialling to identify and eliminate human errors.
* This report has emphasised the importance of HCI end users understanding the capabilities and limitations, normal scope and scope conditional upon failure of a system if they are to interact with it successfully. This is a key point which cannot be ignored when it is recognised that HCI- based systems can only achieve modest levels of integrity and, ultimately, cannot be exhaustively tested and proved.
* The intermingling of claimed tasks requiring high reliability and tasks where reliability is desirable only for economic reasons means that interaction syntaxes and interface design

principles must be consistently applied. It is no longer acceptable or sufficient to attempt to ‘cherry pick’ high integrity tasks and ensure that they have good levels of HF analysis and human engineering design. This is because unanalysed and unclaimed tasks, within the same interface system, can adversely affect the performance of analysed tasks (for example through the negative transfer of training or by time consumption and distraction when difficulties are experienced with non-essential tasks). Design consistency in static display configuration and dynamic interaction syntaxes requires a significant application of design, testing, simulator and training resource to ensure that human error risks, across the piece, are driven to a level as low as reasonably practicable.

* In many instances, the best available HCI for NPP still relies heavily upon providing mimic diagrams derived from plant process and instrumentation diagrams that display parametric and nonparametric status of systems and equipment. There are some notable exceptions, but NPP HCI design has yet to comprehensively grasp the valuable contribution that can be made by using displays that are variously called: functional, mixed mode, ecological or task-based. Accordingly, there is some prospect that gathered data might be bettered but the demonstration of improvement relative to published data can only be achieved by statistically representative user trials.

Taking all of the above together and from the preceding report sections, this report has clearly concluded that it is not possible to cross-compare between the application of HRA to relatively simple ‘knobs and dials’ interfaces with their attendant psychological considerations and modern ones with different and sometimes more complex psychological considerations when users interact with HCI.

Accordingly, a central tenet of this report is that the wide uncertainties about the psychological factors affecting the error potential that exists in the limited available data and the difficulties in understanding the context of that data’s derivation mean that, at best, it can only be used to inform uncertainty and sensitivity analyses.

As already and repeatedly stated in this report, this means that the quality of HCI interfaces and consequent assessed human reliabilities must rely very heavily upon comprehensive demonstration of error elimination by design and thorough qualitative error identification. Without such a demonstration, even uncertainty estimates must be called into question because published data appears to have been gathered from organisations who have already implemented comprehensive qualitative design efforts specifically aimed at minimising human error.

# Introduction and general considerations

The following discussion provides guidance to ONR inspectors on the aspects to consider and the credible levels of reliability that can be associated with the use of HCI.

However, as explained below, the situation regarding HRA of HCI remains complex without absolute and definitive “correct” answers. Therefore, it is imperative that inspectors apply informed judgement on a case by case basis with consideration of the broader merits and failings of any particular safety

case. This situation is driven both by the general complexity of the issue and the relative lack of definitive and applicable data.

The guidance offered here is informed by data found during the literature search undertaken during this study; yet this only involves 67 data points of which 21 are reused from Hickling and Berman (op. cit.), the remaining 46 are new. Of the 67 data points 53 (79%) are specific to the nuclear power sector.

Considerable uncertainty remains in how that data should be applied, thus limiting the possibility of definitive guidance. The key difficulties are:

* Classifying and grouping data;
* Establishing the lower tails for distributions of error probability (‘the better end’), and therefore establishing central estimate;
* Uncertainty in the precise context and granularity of data points;
* The literature search rendering the data was, as specified by ONR ‘high-level ‘and there may

be other data and further information available that has not been sought or seen.

The degrees of uncertainty are such that there is currently no justification for providing human-based safety claims with central probability estimates for tasks involving HCI without context specific and system specific data being available (notwithstanding the enduring need for adequate qualitative justification).

Accordingly, the guidance here is based upon the premise that probabilistic human-based safety claims and related human errors associated with HCI should be handled by means of uncertainty analysis and importance analysis in the context of a PSA.

Further it will be seen that human-based safety claims must rely upon a demonstrated good quality of HCI design including formal analysis of error potential in subsequent design rectification where appropriate.

# Derivation of Indicative HRA Data

The literature search undertaken as part of the study that has developed this guidance did identify HRA data in the use of HCI. However, it is limited in two significant ways:

* The taxonomies used for classifying data differ between different studies
* Each study emphasises that their datasets are small and, accordingly are described in terms such as ‘indicative’, ‘draft’, etc..

In addition, as specified by ONR, the literature search was high level and not an exhaustive trawl of human error data relating to HCI. However, the data has been obtained using HCI, contexts and scenarios relevant to nuclear power.

A re-examination of the datasets identified in the literature search suggests seven themes. These are:

* Complexity,
* Available Time,
* Team Performance and Team Communications,
* Procedures,
* HCI Navigation,
* Data Reading,
* Control Operating.

Some of the available data fits more than one of the themes listed above. Thus, the themes are not considered to be mutually exclusive and, therefore, under no circumstances should they be used in any fashion that postulates additive or multiplicative interactions between two or more of them.

Accordingly, it is suggested that their use is only for application to support a postulated range for uncertainty within scenarios or tasks that are dominated by that kind of theme.

For each of the themes, the data have been combined to derive a mean, 5th centile and 95th centile error probabilities on a lognormal scale.

The following subsections of this section provide:

* A description of the construct associated with each theme;
* A listing of each data source together with a brief description;
* A depiction of the data on a lognormal distribution, together with the best fit polynomial regression line. The regression value R in every case shows that a large proportion of the data variation is accounted for by the best fit line.

Following the description and depiction of each grouping, a summary table of descriptive statistics is included.

# Complexity

Complexity data includes data directly described as concerned with task or scenario complexity, data for workload and data from sources concerned with tasks described as relatively holistic level including post-fault operations, high workload or diagnosis.

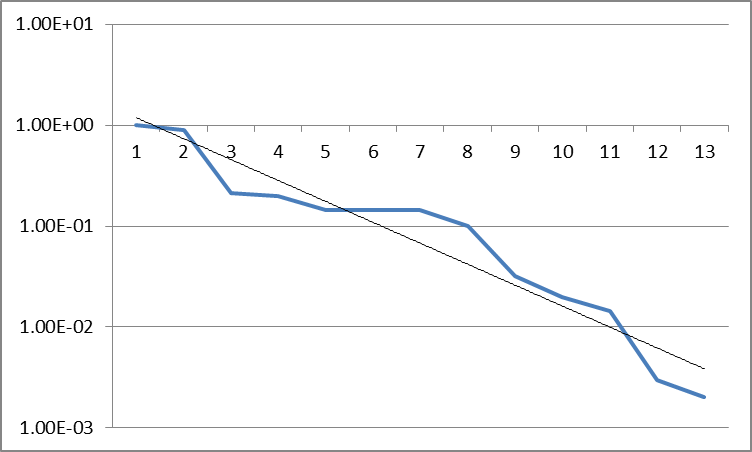
Seven sources provided 13 data points for this distribution.

* + - * Wu et al. (2012) obtained comparative data of HCI process monitoring and decision-making using a traditional mimic-based HCI and a functionally or task-based HCI.
      * Jung et al. (2017) (undertook a study to examine how simulated data might be collected to generate human error probabilities.
      * Jang et al. (2013) undertook a study of HRA in soft control rooms within which diagnostic data was obtained.

The remainder of the studies are reused from the estimation of diagnostic error made by Hickling and Berman (2010) as now follows and were all described as diagnostic error.

* + - * Roth et al. (1994) undertook an empirical investigation of operator performance in simulated emergencies deemed to be cognitively demanding.
      * Hannamann et al. (1985) in developing a time-based model for cognitive human reliability also provide data that was transformed into diagnostic failure data by Hickling and Berman (2010) for the previous ND work on HCI and HRA.
      * Broberg et al. (2010) published diagnostic data for Hammlab simulator-based loss of feedwater scenarios derive for the International HRA Empirical study.
      * Bye et al. (2010) published similar diagnostic data for the same international study for other post-fault scenarios.

The data extracted from the studies has been sorted into order of descending probability on a log scale and the polynomial regression fitted to that data as shown in Figure 5 below.



*Figure 5: Combined Data for the Grouping ‘Complexity‘*

This assumed lognormal distribution has a mean error probability of 0.07 with 5th centile (near the best) of 0.003 and a 95th centile of 0.9.

**OFFICIAL**

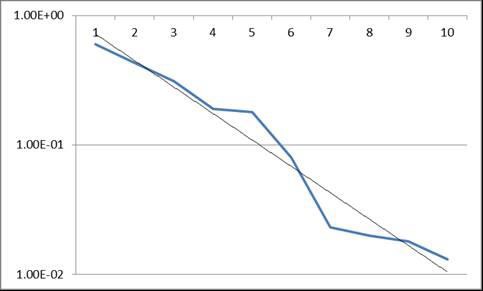
# Available Time

The theme of available time concerns whether or not, in a particular fault scenario, there is sufficient time, as shown by simulator studies, to undertake the actions required.

Two sources of data provided 11 data points for analysis.

Zhao et al. (2012) and also Liu Peng & Zhizong (2014) examined performance of post-fault tasks in circumstances where a previously empirically derived normative time was available or time availability in ratios to the normative of 0.8, 0.9, 1.2, 1.4, or 1.6.

Again, a lognormal distribution of resulting error probabilities is assumed and is depicted in Figure 6 below.



*Figure 6: Combined Data for the Grouping ‘Available Time’*

For this distribution a mean error probability of 0.09 was obtained with a 5th centile of 0.02 and 95th centile of 0.5.

# Team Performance and Team Communications

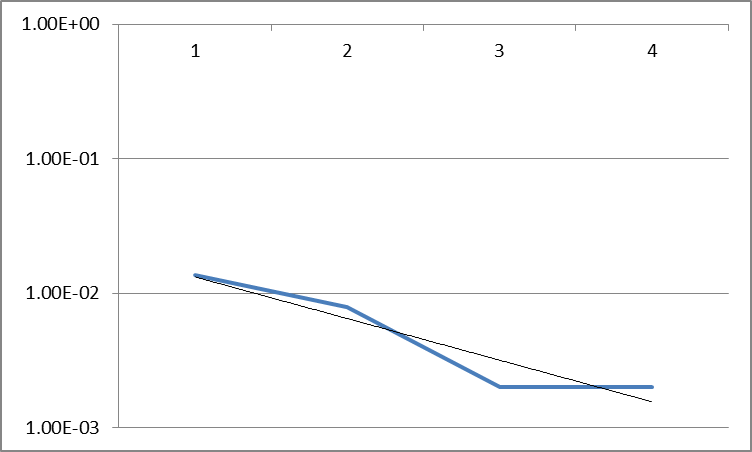
This theme is concerned with how a team interacts together to fulfil the required work demands. Team communications has been included because it can be considered central to team performance.

Three data sources have provided six data points supporting this particular construct.

Hwang et al. (2009) looked at error rates in team performance during an NPP start-up supported by automated procedures.

Yung et al. (2017) in examining errors associated with procedure handling obtained data on transferring from one procedure to another which is an important post-fault and team-based undertaking. They also obtained data on communications between a centralised control and other locations. In a very different study Workman (2007) examined the performance of what he described as a virtual team i.e. one working together collaboratively over networks but in geographically different locations. Workman studied a form of teamwork creating high task demands and potential difficulties. However, unlike control room operations this particular task and opportunity for recovery stretching over hours to days. Hence, this study provided the most reliable human error data.

The logarithmic distribution of error probabilities is depicted in Figure 7 below.



*Figure 7: The Combined Data for the Grouping ‘Team Performance and Communication’*

The mean error probability is 0.005 with a 5th centile of 0.002 and 95th centile of 0.01.

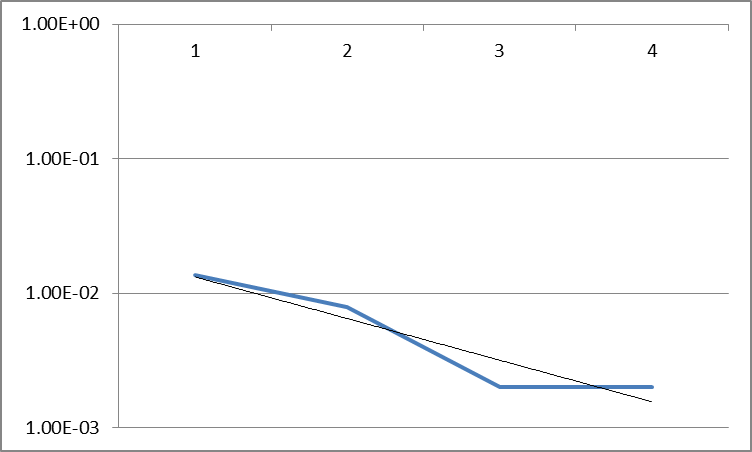
# Procedures

The ‘procedures’ theme concerns errors occurring during the management, handling and following of HCI-based procedures.

Jung et al. (2017) obtained error data on several different aspects of procedure use including transfer between procedures, execution of a step and entry to a step. Their data also included the directing of information gathering.

Yang (2013) identified error data concerned with the omission of an operation, operation on the wrong object, performing the wrong operation and mode confusion. All of these happened under the influence of procedures and, therefore, might be argued to be procedural in character because they occur when following procedures, whether or not they occur as a direct consequence of procedure structural content.

Figure 8 shows the distribution of error data obtained from these two sources.



*Figure 8: The Combined Data for the Grouping ‘Procedures’*

The mean error is 0.01, the 5th centile error 0.003 and the 95th centile 0.1.

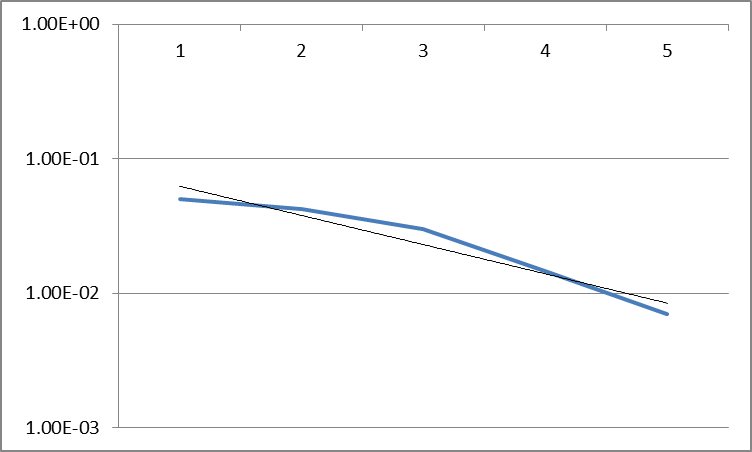
# HCI Navigation

Navigation concerns the action of moving between one display surface/pages/format and another.

Musseller et al. (1996) examined errors associated with icon selection using clicks and double clicks. Wiedenbeck (1999) also studied icon selection with different degrees of available information to describe that icon. Icons are widely used for navigation to other pages.

Fischer and Doherty (2008) examined errors when selections are made using capped and uncapped menu structures. Snowberry (1985) considered errors associated with shallow wide menus. Finally, Jang et al. (2013) identified errors involving operations on the wrong object. These could occur as a result of navigation errors.

Figure 9 below shows the error distribution for data concerning or related to navigation.



*Figure 9: The Combined Data for the Grouping ‘HCI Navigation’*

The mean error is 0.02, the 5th centile 0.008 and the 95th centile 0.05.

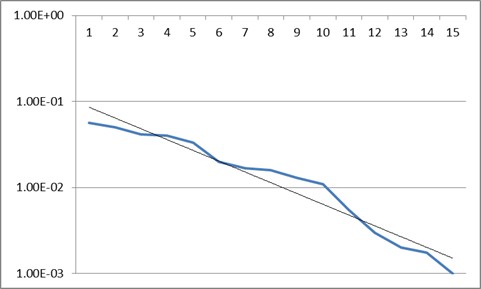
# Data Reading

This construct involves any errors associated with the reading of displays, described as a task core, or the induction of error identified as a result of a display.

For this construct data has been identified from 6 sources that have provided 19 data points.

* + - * Wu et al. (2012) have examined error rates on conventional screen-based depictions of nuclear power plant against functional or ecological/task-based displays of nuclear plant.
      * Yung et al. (2017) have looked errors associated with alarm indications, indicator status reading of simple values, parameter verification, parameter comparison, detection of anomalies associated with e.g. a boot curve, detection of a parameter abnormality and errors associated with trends.
      * Massaiu and Fernandes (2017) compared operator reliability using analogue or digital HMI and have obtained error data for HCI associated errors.
      * Musseller et al. (1996) examined errors associated with icon selection using clicks and double clicks.
      * Wiedenbeck (1999) also considered icon selection with different degrees of available information to describe that icon. Icons are widely used for navigation to other pages.
      * Westerman et al. (1995) examined the retrieval of data from a linear database. Whilst having control elements this is largely a task reliant upon display.

Figure10 below shows the distribution of data reading errors on a logarithmic scale.



*Figure 10: The Combined Data for the Grouping ‘Data Reading’*

For this distribution, the mean error is 0.01 with a 5th centile at 0.002 and 95th centile at 0.05.

# Control Operating

This theme concerns errors associated with operation of an HCI control device , such devices could include the likes of pseudo-analogue control knobs, pushbuttons, drag sliders, etc.. In one sense, it might be expected that control data might represent a subset of display data because each control devices displayed upon a screen but requires additional active selection in order to interact with it as a control.

Five data sources were identified to provide data on errors in HCI-based control. This provided 7 data points.

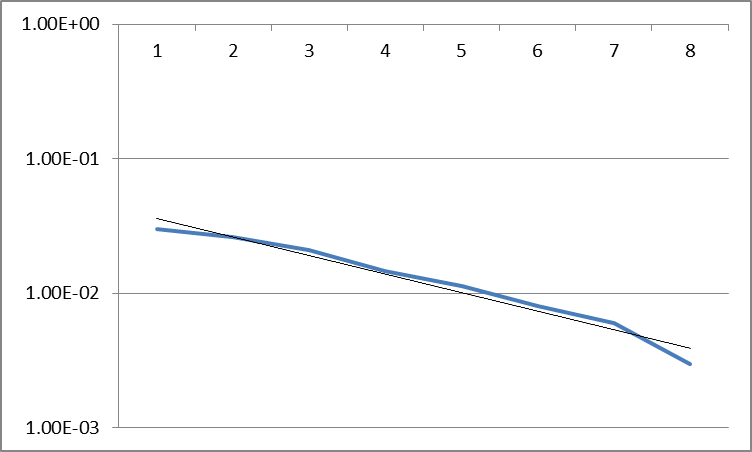
Jung et al. (2017), as part of the study to compare NPP simulator data with the South Korean HRA method, obtained data for the manipulation of a simple control operation continuous control and manipulation of control in the dynamic situation.

Jang et al. (2013) obtained NPP simulator data on the probability of operating the wrong object and performing the wrong operation.

McKenzie and Zhang (2001) examined keying error rates where an on-screen keyboard appeared at random positions – a task that most would consider difficult.

Trewin and Pain (1999) examined the probability of perseverated or ‘stuttered’ keystroke operation.

Figure11 shown below depicts the distribution of the resulting data for control operating.



*Figure 11: The Combined Data for the Grouping ‘Control Operation’*

This logarithmic distribution has a mean error probability of 0.01 with a 5th centile of 0.004 and a 95th centile of 0.04.

# All Groups Combined

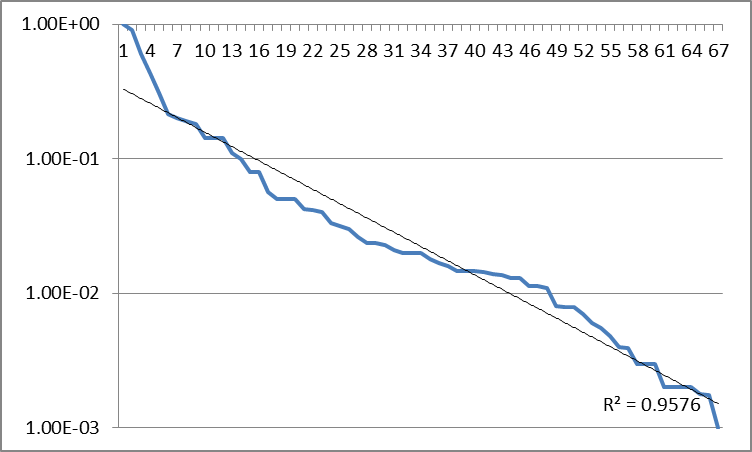
To better inform consideration of the above distributions all of the data have been combined to produce one general distribution. In addition, this allowed further data points that were unjustified for inclusion in the above distributions to be added. These are now described.

Jang et al. (op. cit.) obtained simulator-based error data on mode confusion and delayed operations. No additional sources were included.

This provided a grand total of 67 data points giving a distribution with a geometric mean of 0.02 with 5th centile 0.002 and a 95th centile of 0.4.

The resulting log transformed distribution is shown in Figure 12 below.

The regression line intercept is calculated to be at 0.4. However, readers should note the sharp upward turn of data from the best fit exponential regression line which accounts for around 10% of all data points in the distribution.



*Figure 12: All groups log transformed with trend line and Pearson Correlation R= 0.96*

It was not considered worthwhile to undertake statistical analyses of the earlier distributions, beyond descriptive statistics, due to the relatively small number of data points they contained. However, this all-data combined distribution containing 67 points justifies more detailed statistical attention.

To investigate that the assumption that a lognormal distribution is justified for the combined data, the log10 data was assigned their appropriate Z scores in Microsoft Excel. This resulted in the distribution shown below in Figure 13.

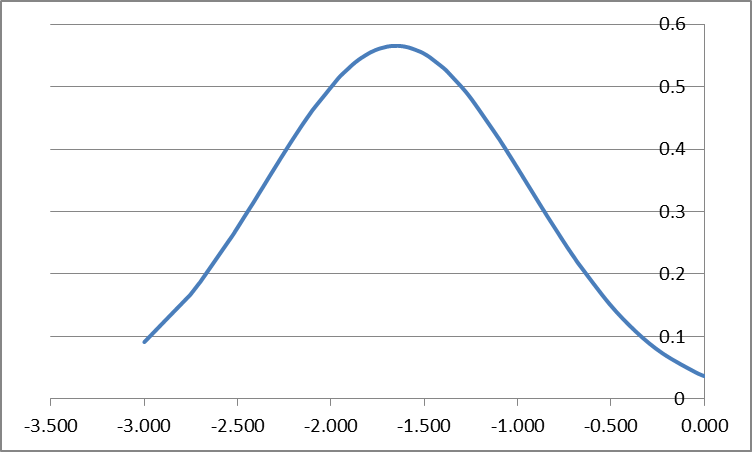


Figure 13: Smoothed and Normalised Distribution of the Probability in Log10 for All Error Data

The distribution does not show the tails of a characteristic bell-shaped normal distribution. At the upper end this is because probabilities are truncated at unity. Nevertheless, some hint of a characteristic distribution tail can be seen at the upper end. At the lower end no such corresponding tail can be seen. This may well represent a truncation of the distribution because the establishment of lower error probabilities would require, at the very least, an order of magnitude greater number of trials than those which have been published.

Notwithstanding this potential constraint on further statistical analysis, it is important for ONR decision- making to seek confidence intervals on the distribution of consolidated error data.

An attempt was made to calculate confidence intervals for the mean of these log-normally distributed error probabilities. This was done by applying calculation to the natural logarithms of the error probabilities. The difficulties in doing this and potential statistical methods for overcoming them are discussed by Parkin et al. (1990). Three methods were attempted to derive 5th and 95th confidence intervals. These were:

* The so-called ‘naïve method’ which applies the formula usually used for a normal distribution not described by logarithms;
* Cox’s method as described by Land (1971) and cited in Parkin (op. cit.), which assumes a Student t distribution; and
* The method of Zhou and Gao (1997) set out in Olsson (2017) which assumes a normal distribution but modified parameters in the equation.

In brief, the methods displayed one or both of the following difficulties:

* Upper confidence intervals above unity,
* Lower confidence intervals above unity,
* Confidence intervals not spanning the calculated mean.

These difficulties and outcomes are described by Land (op cit.) and cited in Parkin (op cit.). Accordingly, it is only possible to draw statistical conclusions that are partially substantiated by scrutiny of the available data but that cannot be substantiated by formal statistical tests.

# Discussion of the Groups and Data Combined

Table 3 below provides summary descriptive statistics for each of the different themes and for all data combined.

*Table 3: Descriptive Summary Statistics for Each Group and All Groups Combined*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Group** | **Geometric Mean** | **Log 5th Centile** | **Log 95th Centile** | **5C-mean EF** | **mean- 95CEF** | **Asymmetry** |
| Complexity | 0.07 | 0.003 | 0.9 | 23 | 13 | Tail To Lower |
| Available Time | 0.09 | 0.02 | 0.5 | 5 | 6 | Symmetrical |
| Team Performance & Communication | 0.005 | 0.002 | 0.01 | 3 | 2 | Tail To Lower |
| Procedures | 0.01 | 0.003 | 0.1 | 3 | 10 |  |
| HCI navigation | 0.02 | 0.008 | 0.05 | 3 | 3 | Symmetrical |
| Data Reading | 0.01 | 0.002 | 0.05 | 5 | 5 | Symmetrical |
| Control Operation | 0.01 | 0.004 | 0.04 | 3 | 4 | Tail To High |
| All Groups | 0.02 | 0.002 | 0.4 | 10 | 20 | Tail To Higher |

Table 3 shows that there is currently no justification for applying central estimates from methods such as THERP, ASEP, SPAR-H etc. to obtain central estimates. It is arguable that methods which focus more upon cognition and task complexity such as CREAM or HEART, might be more applicable. Nevertheless, such an attempt would still be hampered by the fundamental sparsity in available and applicable data.

Without specific investigation and for similar reasons, it is unclear whether NARA would be suitable or not. On the one hand, it is derived strongly from HEART, but on the other the generic task descriptors focus more heavily on user interface characteristics rather than issues of task complexity. However,

irrespective of the method and as the literature search has found, for example Dang and Stempfel (op. cit.) there are difficulties in HRA analysts being ‘calibrated’ to the phenomena and human error probabilities at in general they are seeking to estimate. As extensively discussed in the main body of the report that informs this guidance there are reasons to suppose this problem becomes even more acute where HCI is concerned.

Put another way, there is nothing in Table 3 that is inconsistent with the work of Hickling and Berman (op. cit.) as published in an augmented form by Bowie and Hickling (op. cit.).

From Table 3 a number of features can be noted. The central estimate, geometric mean, for: procedures, HCI navigation, data reading and control operation are all at the lower end of the x/100 probability range. Only team performance and communication betters that in the x/1000 range at

0.005. As might be expected from task character, the themes of complexity and available time (the latter which may really be a causal factor of complexity) are at the upper end of the x/100 error probability range. With the exception of available time and HCI navigation, the 5th centile values are at the lower end of the y/1000 error probability range.

Because of the inability to calculate meaningful confidence limits for any data group distributions, the distance between the geometric mean and the 5th and 95th centile error probabilities has been calculated as an Error Factor (EF), as used in THERP. This can only constitute a crude indication of whether the distribution of data is symmetrical or asymmetrical. If it is asymmetrical the EFs differ and the greater EF possibly indicates which way the tail of the distribution lies. Based upon the limited data available, this provides some hint as to whether the distribution could be extended or not. It is important to note that such an indication can only be relevant if the assumption of a lognormal distribution is valid. The EF is the multiplier required for the 5th centile estimate to match the geometric mean and the multiplier required for the geometric mean to reach the 95th centile. (It is important to note that the centiles are calculated on the logarithm distribution not the untransformed error data.)

The EF is noted to be approximately symmetrical only for the themes available time, HCI navigation and data reading. This provides a suggestion, albeit weak, that these distributions cannot be extended either in the upward or downward directions.

The themes of complexity and of team performance & communication have a greater EF between the 5th centile and the mean than the EF between the mean and the 95th centile. This suggests these distributions may extend further in the direction of lower error probabilities.

In respect of complexity this is a particularly interesting finding. The ratio between the lower and upper EFs is 1.75 and the 5th centile is 0.003. At face value, they suggest that the construct complexity encompasses other constructs whose distributions fall within the range of distribution but may range from, say, 0.0005 up to unity.

In principle, this inference raises the prospect that if complexity could be subjectively rated over that range then this might be a useful human reliability dimension. However both theoretically and pragmatically this may be a misleading conclusion for 2 reasons. First, this is because, subjectively, the construct of complexity is itself complex and therefore likely to be very difficult for human reliability

analysts to rate. Secondly, the very fact that it encompasses such a wide range of other ‘discovered probabilities’ suggest that complexity is a resultant of the other more readily defined factors that the theme encompasses. Indeed, in this study the complexity theme has within it diagnosis. This theme carries with it the methodological complexity of rigorously defining cognitive behaviour ‘in writing’.

Accordingly it is suggested that whilst useful for grouping data within this study it might be less helpful if it were assumed to be a promising way forward as part of a future HRA method.

On the other hand, it should be noted that existing HRA methods that focus upon diagnosis do not necessarily also include other cognitive functions such as recognition, attention, decision-making etc.. Accordingly, whilst it might seem to go against some previous findings when attempting to develop HRA methods, it would be an interesting study to examine whether, in practice, the broader construct such as complexity which encompasses more cognitive elements might actually be more easily and reliably subjectively rated by HRA practitioners. As pointed out elsewhere in the report Dang and Stempfel (op. cit.) have noted the unreliability of assessors in rating or scaling performance shaping factors. This is central to the success of any HRA method and therefore merits further study.

In the group ‘Team Performance & Communication’ the ratio of upper and lower EFs suggests that the distribution may actually encompass lower error probabilities. However, there must be very high uncertainty that this is so as the distribution relies upon only 4 error probability data.

The EFs for available time and control operation hint that their distributions may be more likely to extend towards higher probabilities.

The grouping of all data together suggests, by a ratio of 2 between the upper and lower EFs that there is more data extending in the upward direction. However as the 95th centile is already 0.4 and there is registered data for inevitable error (unity) then it is reasonable to suggest that the upper limit for error rate in any uncertainty analysis, for example, should be set to unity.

Perhaps the most important conclusion from the above data is that there is no general justification for entertaining human error probabilities that are better/more reliable than 0.001 where HCI is involved. Further, where HCI and cognition are involved, the methodologically safe option is to assume an error probability approaching unity. (The practical implications of this and their impact upon the application of HRA to design improvement are discussed further below)

# Further Data Exploration

From the results obtained, and notwithstanding the difficulty in obtaining a full suite of useful statistics, it might appear that the safest distribution that could be used to meaningfully encompass HCI error probabilities in a PSA is that of all data combined. However, such an approach would blunt the discrimination view, which this author believes to be widely agreed, that tasks and actions with significant cognitive content are generally less likely to be reliable than those without. This becomes an important issue if changes to human factors are to be informed by HRA in the context of a PSA.

Notwithstanding the limited set of data available some further exploration of the data is therefore merited.

The data for the themes called complexity, time available and team performance & communication have been combined. This has been done on the basis that all of these reliabilities generally represent, at least in a control room, the combined performance of a team where the task demands and the task methods demand cognition and generally possess some degree of complexity. In addition, the 2 categories of error recorded by Chang et al. (op. cit.) of mode confusion and delayed operation have also been added into that group. This has been done on the basis that mode confusion requires a collective error and delayed operation even if by an individual operator implies some failure of supervision or wider situation awareness.

# Re-grouping

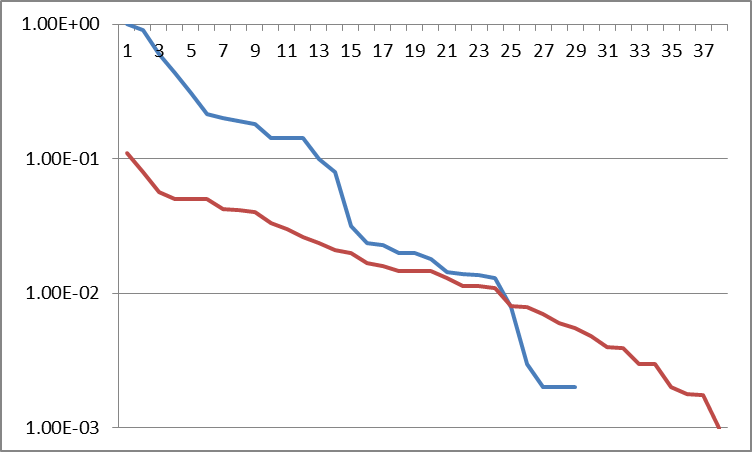
The groups Procedures, display reading and control operation have also been combined. Table 4 below provides summary descriptive statistics for the two resulting distributions.

*Table 4: Comparison of broadly team cognitive and broadly single person action groups.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Group** | **Geometric Mean** | **5th Centile** | **95th Centile** |
| Complexity, time and team | 0.05 | 0.002 | 0.8 |
| Procedures, navigation, reading and control | 0.01 | 0.002 | 0.06 |

As the intention is to inform the possibility of using this distribution in the context of uncertainty analysis then the use of the error factor to investigate potential tails extension has not been undertaken.

A graphical comparison is also provided in Figure 14 below.



*Figure 14: A Graphical Comparison of Broadly Cognitive and Single Person Action Groups*

Further scrutiny of the cognitive data points below 0.01 has been undertaken. These data points can be described as follows:

* Performance with a task-related or ecological ergonomic display rather than a ‘dynamic P&ID or mimic’,
* Transfer from one procedure to another,
* A verbal request from a central control room to an outside location,
* A geographically distributed virtual team with days for error recovery,
* Team interaction with an automated prompting system.

These data points can be characterised by being marginal for inclusion in this group because they involve:

* Transfer from one nuclear procedure to another,
* Having features of automated alarm and action support for a nuclear task,
* Days of grace time and multiple layered independent scrutiny for error recovery - non nuclear,
* State-of-the-art interface design supporting enhanced performance.

Whether or not the lower limits of error probability found reflect a genuine situation or are an artefact of data truncation, there appears to be no current justification within published literature to suggest

that an overall uncertainty range for HCI human error should not be any less than unity at its highest and 0.001 at the very best. This conclusion is consistent with that arrived at in the earlier work by Hickling & Berman (op. cit.) performed in 2010 and published by Hickling and Bowie (op. cit.) with a handful of new data sources in 2012.

Obviously, should the submitting organisation have data specific to their interface and interaction design to justify better probabilities then this could be used. This is discussed further below.

# Uncertainty and Importance Analysis

Taking all of the above together regulatory expectations for all HCI-involved human reliabilities in a PSA/PRA could be set to be estimated as being within the range seen for all data combined and this to be handled within a PSA as an uncertainty analysis. This would have a central estimate of 0.02 and upper bound of unity and a lower bound of 0.002.

Such an approach means that the inclusion of HRA data within the PSA can still be relied upon for directing HF concerns and this would need to be done by means of an importance listing. Irrespective of applied probabilities the qualitative task analytical phases of HCI evaluation, when undertaken by competent HF professionals, are more diagnostic of HCI issues when not driven by the narrow assumptions and taxonomies of HRA. However as pointed out earlier in this report, this only remains true if the person undertaking those task analyses not only understands HF but understands the HCI and how it might initiate error.

Recognising that expectations for uncertainty and importance analysis would then be informed by data exploration as described above, places a greater onus on the HCI design being ALARP in respect of error. This is because the character of the data means that only exploratory scrutiny can be undertaken rather than statistical analysis giving confidence intervals.

The conclusions already given from the literature search that informs this guidance make abundantly clear that qualitative excellence of HCI design needs to be demonstrable and, particularly for the nuclear and safety case context there must also be readily scrutinised demonstration that analysis for qualitative error has been used to inform design change. Further, given the problems of data illustrated in the exploration given above in this section it is most definitely preferred that a broad range of uncertainty is included for HCI related human error.

This conclusion is not only supported by relevant good practice in qualitative design but its importance becomes more heavily justified when the paucity and difficulties of data analysis are apparent as seen above.

# HBSC Modelling and Granularity

For importance analysis it might be expected that a submitting organisation would use a simple approach of setting all human errors to unity – as indicated by the all data distribution If sensitivity analysis shows this to be acceptable in respect of the insights given from the assessed human impact upon nuclear safety risk then, it is further suggested that there might be no further expectation for a

‘finer grained’ analysis. However, a finer grain of analysis may be justified if the ‘set to unity’ approach suggested is a blunt instrument that could lead to disproportionate HF intervention. That is, it fails to differentiate or prioritise where human issues should be rectified to meet the requirements of the safety case.

Should the submitting organisation consider that a finer grained assessment is justified because an ‘improvement’ in the claimed probability allows better exploration of the magnitude of the gravity of a human error issue, for nuclear safety and hence design, then, as appropriate, the risk analyst might choose either ‘HCI & cognitive involvement’ or ‘Action involving HCI’.

If this were to be done and noting the generally-held assumption that cognitive tasks are less reliable than action tasks and further noting that the most favourable cognitive data points are derived from specific favourable features that reduce cognitive error probability, then it is suggested that a pragmatic application which differentiates between cognitive and action data would be as follows:

* + - * Cognitive uncertainty would range from unity to 0.1;
      * Cognitive uncertainty would range from unity to 0.01 where enhanced measures to support cognition have been demonstrated by practical trials to be effective in and also having with levels of design comparable with or demonstrably better than those sources from which the derived data has been obtained, and formal error identification and reduction studies..
      * Action uncertainty would range from 0.1 to 0.001.

The suggestion that the cognitive uncertainty range should have a lower limit of 0.01 stems from 2 perspectives:

* + - * There would be wide agreement in the HRA community that human reliability involving cognition (or indeed complexity, available time or team performance) tends to be less reliable than the performance of action tasks informed by correct cognition.
      * There is evidence of considerable uncertainty about whether HCI measures that are supposed to enhance cognitive performance are effective or not. For example see the three papers by Liu and Li (op cit.).

Implicit in the descriptions of the above uncertainty ranges is a notion about the scope of application or granularity. This level of granularity is now made explicit.

UK HRA practice in NPP safety cases has, in recent times, embraced more sophisticated modelling. For example safety cases have been developed using Operator Action Trees after the manner of THERP. These trees are characterised by outputs that probabilistically estimate human-based safety claims at a functional level. Generally this functional level is decomposed to the level of human error description that is comparable with fallible inherent physical characteristics or automated responses implemented by protection or control. For example the claim might look like “the operator can successfully start the high head safety injection pump following automatic failure to do so”. This will consist of a number of actions for example, involving isolating automation, starting motors, opening valves and checking

indications. Thus, a number of human error probabilities will be included in the OAT typically at a ‘THERPian’ level of individual control operation or display scrutiny. In thest examples cognitive steps of detection or situation recognition are also included. Further, the OATs can contain one or more error recovery pathways and acknowledgements of human error dependency. In the last for example, a failure to scrutinise one display may lead to a number or all claimed control actions not being taken.

It is suggested that inherent in the cognitive uncertainty range derived above is the notion of recognition and diagnosis potentially informed by alarms and displays guided by procedures involving a control room team as a whole. Accordingly, it is suggested that ONR might expect such claims to be made holistically by means of a single quantified claim. However, whilst detailed modelling would not appear to be justified, unless the HRA analysts look for dependencies within the suite of actions involved and the cognitive stage, and then their mitigation or removal there can be no demonstration that such dependencies will be minimised in practice.

It follows that once an assessment has been carried out for cognition for initial diagnosis then, as per convention following assessments can address the matter of procedure use, controls and displays.

Nevertheless, there may be a need for analysts to consider a separated cognitive claim if it is considered there is sufficient complexity inherent in the situation that the human error probabilities for subsequent actions involving procedures, displays and controls are influenced in some way by further cognitive attention or decision making. An example of this might be a transient with the potential to start slowly and increase in speed so necessitating a different suite of actions form those initially decided upon or procedurally directed.

In contrast to the claims for cognitive content, claims for action should match the level of granularity contained in the data points contributing to the uncertainty range for action HRAs involving HCI. Therefore, that uncertainty should be applied to each control and display separately as if performed by a single person. This is also logically sensible as it follows that the more controls or displays on HCI with which interaction is required the greater the probability of some error occurring that can influence the claim.

The descriptions for the 2 complementary approaches given above do not contain any notions of error recovery. However, the notion of recovered error – within the context of a control room team and a credible claim by an independent person or persons – implies a further cognitive activity of a kind generally described as going on in parallel with procedure, control and display actions. However, as is already the HRA convention, such claims can only be credibly made if the dependency between the initial cognitive claim and the subsequent one supposed to implement recovery is properly considered. As already stated, we might expect to see explicit consideration and estimation of levels of dependency involved given that it is like the same overall team and the same individuals who may or may not be already very busy doing other things.

In considering all of the above it is crucial to emphasise that the fundamental qualitative steps of considering whether or not a claim is credible still need to be made. The approach suggested above in no way diminishes or excuses the need for thorough, risk informed and HCI informed qualitative as well as quantitative analysis.

# HCI Improvement in HCI Assessment

Notwithstanding all the foregoing, there remains an undoubted benefit for HCI in undertaking HRA, when it is performed properly. For HCI this must be done by taking account of the prevailing human factors independent and irrespective of the taxonomies and considerations contained in any particular HRA quantification method. This is because currently, those methods do not provide a qualitative focus upon HCI design with descriptive frameworks that actually identify induced errors. Also, they do not contain data to evaluate such errors or the design and usage factors leading to them.

The need for such an approach is currently acutely necessary because it is the only plausible means to provide a strong and logical connection between specific HCI design features and their relationship to human error and thus to risk. Therefore, whilst the uncertainty and sensitivity approaches outlined above will be appropriate for estimating probabilistic risk, such an approach, if used alone, will detract from the absolutely necessary focus on the actual relationships between HCI design and human error potential. Therefore, credible and complementary steps must be undertaken in the overall HCI design process and its HF assessment. There must be design and design assessment processes which incorporates specific focuses on minimising, by design, error induction and providing by HCI design, where practicable, the means for the rectification or mitigation of such error. That approach must be clearly documented and the selection of the methods used properly justified.

As with many other HF requirements relating to nuclear safety, this objective is easier to achieve when an HCI design commences from a greenfield. However, where HCI design already exists as a supposedly finished entity or, indeed, where that design has evolved to match evolution in the technology or of usage then this can be more challenging. As a minimum, in that situation, ONR might expect to see a systematic and thorough approach to the identification of error potential in HCI and a demonstration either that the error is already minimised or has been minimised through the intervention of considering qualitative error and changing design when required to minimise risk.

# Discussion

This section provides discussion and overview of our findings. This discussion leads to recommendations to ONR for future steps and the way forward.

# The Strategic Picture

The literature review has clearly shown that there is still a paucity of data reflecting human reliability when interacting with computer-based systems. We can postulate to fundamental reasons why this might be so.

Firstly, there is an absence of cognitive theory that is both widely accepted and that provides a tractable framework for classifying HCI based errors. Secondly, the strong USA influence over HRA and the long tradition of using THERP invites those seeking such data to use similar approaches.

The difficulties of cognitive classification are also strongly reflected in the attempts made in the last 15 years from the USA to develop a cognitively-based HRA method. The ongoing sparsity of data and the impossibility of reconciling data obtained through different classificatory frameworks means that it is not currently possible to rely on probabilistic quantification of HCI-induced human errors. Rather, the application of sensitivity and uncertainty analyses are required to understand the human contribution to risk where HCI is involved.

The situation is neatly summarised by Porthin et al. (2020) (Ref. 59) who state “It cannot be generally concluded that either analogue or digital control rooms would be always better than the other advanced control rooms have the potential to offer error reducing and supportive features to the user but may also introduce increased complexity and interface management tasks as well as potentially error prone team working practices”. Based upon their empirical work they go on to say “The main PSF categories presented in HRA methods and international guidelines seem still relevant in Advanced Control Rooms. Quantitative empirical studies show that especially training and experience, availability and quality of procedures as well as task type or complexity are important factors. However, digitalization changes the way in which the PSFs should be defined and measured, and the effects of PSFs on the error estimates may be different.”

Further difficulties are apparent; if one assumes that suitable HRA data did exist, then the application of that data to provide acceptably accurate human reliability estimates, would require a clear understanding of the error induction mechanisms present within the HCI and implicated in the tasks involved in any particular HBSC. Unless those tasks have been performed with the HCI in question and errors observed then any informational basis, for the HRA estimator, is absent.

The literature review also showed a considerable strength in the HCI design community who have an arsenal of qualitative design assessment methods. It is significant to note that those methods accommodate a lack of cognitive theory by placing far greater emphasis upon the use of task analysis during HCI system development. As in classical ergonomics, such analyses may include error postulation

at early stages in the design but progressively increasing levels of fidelity in task simulation as the design progresses within the context of a systematic software design development project.

The literature on qualitative methods has also shown that systematic error identification methods do exist and can be applied to identify where error might arise in the use of HCI, and to consider its nuclear significance, and thereafter to identify the forms of error that might occur in practice. Such methods are being used for safety critical medical devices that appear to be conspicuous by their absence within the nuclear domain.

Within British regulatory law in general and nuclear law in particular, there exists a requirement to show that a design reduces risks in accordance with the ALARP principle. However, if such a design is accepted on the grounds of use alone then a major assumption must also follow that any errors of potential significance for nuclear safety have already been ‘flushed out ‘by that usage.

Such an assumption does not appear to be justifiable on two grounds. First, the support for such an assertion and assumption would require operational experience feedback at a level of granularity and with an error classificatory system that would identify interaction errors with HCI. Secondly, as has been repeatedly pointed out amongst those undertaking qualitative HCI design, the number of interactions of potential significance to error induction is very large and therefore unlikely to be captured in OPEX. This is true even ignoring the potential for non-critical interactions to destabilise operator performance: either through reduced self-confidence or reduced available task time due to error recovery. Indeed, in some parts of the HCI design community the magnitude of the work required to examine user performance by user trials is so great that they deem user trials at all, or statistically sound user trials, to no longer be a feasible way for predicting error. For example, see Thimbleby (2007) (Ref. 49).

## Recommendation 1: ONR may consider how they will make clear to licensees and RPs that a greater onus lies upon qualitative design being shown to be ALARP.

Fundamentally then, even an ‘as given’ HCI system should not be accepted as meeting the ALARP principle unless there has been a systematic process of error identification and elimination or mitigation by design: whether that design be achieved via interface change, procedural adjustment or training. It is fortunate that the methods put forward by Stanton, Baber and, separately, Thimbleby (Ref. 49) are scalable for different levels granularity. This is because they each rely on the identification of state changes as the focus for error analysis. Of course, a state change can be defined to involve many or a few human interactions depending on the granularity of description. Of course, if a coarse level of granularity is chosen for error analysis then it behoves the analyst to properly understand the detail of HCI system and user behaviours within that particular informational bucket.

Within the nuclear sector there appears to be a sentiment amongst designers that the use of HCIs straightforward and, therefore, not prone to reducing human error. The apparent lack of connectivity between the nuclear industry and the HCI design community reinforces our impression that designers think interface design straightforward because there is little problem, for them, in producing instrumented and animated P &IDs. Perhaps the ‘obviousness’ of this solution avoids the generation of

concerns about the actual complexities of using such an approach. As the literature search has shown there is continued experimental endorsement of the superiority of displays variously called: task- based, ecological ergonomic, or functional displays relative to animated P&IDs. To this author’s knowledge, this replicates findings of other studies going back some 20 years.

Taking all of the above together, it appears inevitable that, for the immediate and foreseeable future, ONR will have to accept assumptions about good human performance and reliability based upon qualitative error interventions in the design of HCI systems and HCI-related tasks. It can be argued that such a requirement is already reflected in the existing HF SAPs. As a goal-based regulator it cannot be ONR’s role to prescribe particular methods for undertaking error identification and qualitative analysis the expectation needs to be signalled very clearly at an early stage in any nuclear project that such an approach should be thoroughgoing. (As an historical note, it is worth reflecting the fact that Sizewell B Main Control Room computer interfaces were designed with direct HF involvement in all objects displayed on screen mindful of necessary success and the avoidance of failure).

# The Tactical Picture

In undertaking this work we have been mindful of the need for ONR to maintain SAPs at a balanced level of granularity. That is: sufficiently detailed to be understood but also sufficiently general to be applied in all cases.

From the list of proposed principles it could be implied that significant changes to existing SAPs could be required. This is not the case and, as already shown, principles can be readily tailored as guidance for inspectors or descriptions of reasonable expectations that they might have. However, any work to tailor principles as guidance or to aggregate them should accommodate the complexity of the issues at hand and their interactions.

## Recommendation 2: Further work could be undertaken in order to integrate the proposed principles into the existing ONR SAPs.

**Recommendation 3: Any guidance developed from the findings of this work should preserve and support the complex interrelationships between them.**

To minimise the need for change we have carefully considered whether and how principles might be prioritised and have concluded that there is no easy way to achieve this. We have, however, avoided the trap of putting forward principles that are for design or design processes rather than for regulatory assessment.

We recognise there are important interfaces between the principles put forward here and other philosophies and principles contained elsewhere that have a bearing on the design of HCI systems. In particular, an elegant design which supports human diagnosis and decision-making is unlikely to be characterised as particularly reliable in the context of standards such as IEC 61508 where many such systems would be classified as SIL 1. If well-designed HCI-based user interfaces are to support high levels of human reliability then it is entirely possible that the human reliability issue becomes bounded by the level of reliability that can be claimed for supporting system.

Considering the reliabilities alone, however, may be unhelpful because it might be concluded that such HCI systems have no role whatsoever to play in a positive contribution to nuclear safety. If the promise of better diagnosis and decision-making through the use of advanced HCI can be fulfilled then such systems may become a very significant contributor to the avoidance of poor diagnoses or decisions which, by their very character, can systematically lead to a number of systematically related errors of omission arising from a diagnostic/decision error of commission.

## Recommendation 4: ONR may consider whether a multidisciplinary group could address this issue of a balance between HCI and human reliability more systematically. Appropriate disciplines to include would include: HF, C&I, PSA and deterministic safety assessment specialists.

The principles and issues reported here have focused upon HCI in general and have not considered specific issues such as artificial intelligence, Bayesian belief systems or alarm systems. In addition, we have not been able to address the problem of data superfluity where HCI systems design provides centralised control facilities with unnecessary data. Such data may be unnecessary because fewer critical indications are necessary to undertake correct diagnoses and decisions or, put another way, the granularity of the data is too fine. Alternatively, data may be superfluous because those receiving it are not responsible for its processing or application. Nevertheless, we suggest that data superfluity arises directly as an artefact of the capacity and capability of HCI systems and its impact cannot be ignored.

However, we also note that the existing HF SAPs do address this issue in terms of resultant design but not necessarily at the level of design process. Unless the issue is tackled within design user interfaces will be unnecessarily complex and this could then undoubtedly affect user performance and reliability.

## Recommendation 5: The issue of data superfluity, its causes and effects, and the adequacy of the pre-existing ONR SAPs could be re-examined.

In undertaking this work we have been repeatedly impressed by the qualitative design processes and methods used by the HCI design community and concerned at an apparent corresponding absence of their application to HCI in the nuclear sector. We are also struck by the acknowledged importance of applying task analytical methods and user trials within that community. It is apparent that HCI design and evaluation is now an established and mature discipline in its own right. Given the likely volume of qualitative information arising from a well assessed HCI we are left to wonder whether the simplified and heavily distilled HF knowledge contained within HRA methods, and the analyst to exercise those methods can sensibly encompass the necessary body of knowledge to make well-informed HEP judgements. We do not feel in a position to make any categorical recommendation on this issue beyond the following.

**Recommendation 6: ONR should maintain a careful watching brief on the magnitude and complexity of qualitative HCI evaluation needed to demonstrate an ALARP position and consider whether that complexity merits a clear discipline separation between the HRA probabilistic analyst and the qualitative HRA designer, together with their corresponding regulatory assessors.**

# Recommendations

The recommendations arising from the research activity follow in Table 5.

*Table 5: Recommendations arising from research study and development of principles*

|  |  |
| --- | --- |
| **Recommendation number** | **Recommendation** |
| 1 | ONR may consider how they will make clear to licensees and RPs that a greater onus lies upon qualitative design being shown to be ALARP. |
| 2 | Further work could be undertaken in order to integrate the proposed principles into the existing ONR SAPs. |
| 3 | Any guidance developed from the findings of this work should preserve and support the complex interrelationships between them. |
| 4 | ONR may consider whether a multidisciplinary group could address this issue of a balance between HCI and human reliability more systematically. Appropriate disciplines to include would include: HF, C&I, PSA and deterministic safety assessment specialists. |
| 5 | The issue of data superfluity, its causes and effects, and the adequacy of the pre-existing ONR SAPs could be re-examined. |
| 6 | ONR should maintain a careful watching brief on the magnitude and complexity of qualitative HCI evaluation needed to demonstrate an ALARP position and consider whether that complexity merits a clear discipline separation between the HRA probabilistic analyst and the qualitative HRA designer, together with their corresponding regulatory assessors. |

# References

1. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil,R. and Horvitz, E., (2019). "Guidelines for Human-AI Interaction". Proc: CHI Conference on Human Factors in Computing Systems (CHI 2019), May 4-9, 2019, Glasgow, Scotland UK. Pub: ACM, New York, NY, USA, https://doi.org/10.1145/3290605.3300233.
2. ASEP: Swain, A. D., (1987). “Accident Sequence Evaluation Program Human Reliability Analysis Procedure. NUREG/CR – 4772, US Nuclear Regulatory Commission, Washington DC.
3. ASM Consortium, (2015). "Effective Console Operator HMI Design, Second Edition". ASM (Accident Scenario Management) Consortium, Houston, Texas.
4. ASME & ANS (2019) "ASME/ANS RA-Sa-09, Addenda to ASME/ANS RA-S-2008 Standard

for Level 1/Large Early Release Frequency Probabilistic Risk Assessment for Nuclear Power Plant Applications", February 2009.

1. Baber, C. and Stanton, N. A., (1994). Task analysis for error identification, Ergonomics, 37, 1923 - 1942.]
2. Baber, C. and Stanton, N. A., (1996). A systems approach to human error identification, Safety Science, February-April 1996, Pages 215-228.]
3. Stanton, N. A. and Baber, C., (2005). Validating task analysis for error identification: reliability and validity of a human error prediction technique, Ergonomics 48 (9) pp 1097- 1113.
4. Bainbridge, L. (1983). “The Ironies of Automation”. Automatica, Vol 19 pp 775-779. Elsevier B.V., Netherlands.
5. Bell, J. and Holroyd, J., (2009). "Review of human reliability assessment methods". Report RR 679. Health and Safety Laboratory , Buxton.
6. Boehm, B. (1986). "A Spiral Model of Software Development and Enhancement", ACM SIGSOFT Software Engineering Notes, ACM, 11(4):14-24, August 1986.
7. Boehm, B. (2000). "Spiral Development: Experience, Principles, and Refinements, Spiral Development Workshop, February 9, 2000. Pub: SEI Joint Program Office, HQ ESC/DIB, 5 Eglin Street, Hanscom AFB, MA 01731-2116, USA.
8. Boring, R. L., (2012) 50 Years of THERP and Human Reliability Analysis, PSAM 11, ESREL2012 Probabilistic Safety Assessment & Management and the Annual European Safety and Reliability Conference, 25-29June, Helsinki, Finland.
9. Broberg, H., Massaiua, Julius, J and Bertil Johansson, B (2010). “The International HRA Empirical Study: Simulator Results From the Loss of Feedwater Scenarios”. Proc. of PSAM 10, 7-11 June 2010, Renaissance Seattle Hotel, Seattle, Washington USA.
10. Bye,A., Lois, E., Dang, V.N., Parry, G., Forester, J., Massaiu, S., Boring, R., Braarud, P. Ø., Broberg, H., Julius, J., Männistö, I, Pamela Nelson, P. (2010) “The International HRA Empirical Study - Phase 2 Report Results From Comparing HRA Methods Predictions To Hammlab Simulator Data On SGTR Scenarios”. HWR-915, Institutt for energiteknikk, NO- 1751 Halden, Norway.
11. Chang, Y.J., Bley, D., Criscione, L., Kirwan, B., Mosleh, A., Madary, T., Nowell, R., Richards, R., Roth, E.M., Sieben, S., Zoulis, A. , (2014). "The SACADA database for human reliability and human performance". Reliability Engineering & System Safety, 125, May 2014, pp 117 - 133, Elsevier.
12. Chen et al. (2014) [Chen, K., Lia, Z., Qiu, Y., and He, J., (2014). " A Comparison of Two Cognition-driven Human Reliability Analysis Processes - CREAM and IDHEAS". Proc: PSAM 12, Probabilistic Safety Assessment & Management Conference, Honolulu, Hawaii.
13. CREAM: Hollnagel, E. (1998). “Cognitive Reliability and Error Analysis Method (CREAM)”.

Elsevier Science.

1. Dang, V.N, and Stempfel, Y (2012) Evaluating the Bayesian Belief Network as a Human Reliability Model - The Effect of Unreliable Data". PSAM 11, ESREL2012 Probabilistic Safety Assessment & Management and the Annual European Safety and Reliability Conference, 25- 29June, Helsinki, Finland.
2. Diaper, D. and Stanton, N. A. (2004). “The Handbook of Task Analysis for Human-Computer Interaction”. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, USA.
3. Ekanem, N. J., Mosleh, A. and Shen, S-H. (2015) [Phoenix - A model-based Human Reliability Analysis methodology: qualitative Analysis Procedure". Reliability Engineering and System Safety. 145 pp 301-315, Elsevier.
4. Endsley, M. (2004). "Designing for Situation Awareness: an approach to user centred design". Taylor & Francis Group, Boca Raton
5. Engineering Interactive Computer Systems, EICS'09,July 15-27, 2009, Pittsburgh, Pennsylvania, USA. Copyright 2009 Association for Computing Machinery, P.O. Box 30777,New York, USA.
6. EPRI, (2013) "EPRI HRA Calculator® Version 5.0", 2013, EPRI, Palo Alto, California, USA.
7. Everdij, M.H.C and Blom, H.A.P., (2010) "Safety Methods Database (Version 0.9)". Downloaded in 2010 from [HTTP://www.nlr.nl/documentsflyerssatdb.pdf](http://www.nlr.nl/documentsflyerssatdb.pdf)].
8. FAA (2001). "Human Factors Design Guidelines for Multifunctional Displays", DOT/FAA/AM - 01/17, Office of Aerospace Medicine, Washington DC, USA.
9. Forester, J.,Huafei Liao, H., Dang, V.N., Bye, A., Lois, E., Presley, M., Marble, J., Nowell, R.,and Broberg, H. (2016). “The U.S. HRA Empirical Study – Assessment of HRA Method Predictions against Operating Crew Performance on a U.S. Nuclear Power Plant Simulator”. NUREG 2156, US Nuclear Regulatory Commission, Washington DC, USA.
10. Michael Hildenbrandt3, Bruce Hallbert7, and Tommy Morgan7
11. Groth, K., Smith, R. and Moradi, R. et al. (2019). "A hybrid algorithm for developing third- generation HRA methods using simulated data, causal models and cognitive science". Reliability Engineering & System Safety, 191 November 2019, Elsevier.
12. Groth, K. and Swiler, L. P., (2012) "Use of a SPAR H Bayesian Network for predicting Human Error probabilities with missing observations". PSAM 11, ESREL2012 Probabilistic Safety Assessment & Management and the Annual European Safety and Reliability Conference, 25- 29June, Helsinki, Finland.
13. Hallbert, B.P., Boring,R. L., Park, J. and Jung, W. (2017) "Methods and Measures for Characterizing Operator Performance for Operator-in-the-Loop Studies". PSAM Topical Conference 2017 Human Reliability, Quantitative Human Factors, and Risk Management, 7 - 9 June 2017 in Munich, Germany. Note: The conference proceedings were not published but are held by ONR and the lead author (Ned Hickling).
14. HEART: Williams J. C. (1988) A Data-based Method for Assessing and Reducing Human Error to Improve Operational Performance. Proc. 4th IEEE Conference on Human Factors in Power Plants, Monterey, California.
15. Hickling, E.M., Harvey, J.G. and Hollywell, P.D. (1992) The Consistent and Structured Definition of Process Plant HCI-based Tasks Using the “PAGODA” Method”. Contemporary Ergonomics 1992, Taylor & Francis, London.
16. Hickling, E.M., and Berman, J. (2010). “EPR and AP1000 Workstream 5: The Applicability to Human-Computer Interfaces of Data in HRA Methods. GSB2152EPRWS1R1V1. Joint publication: Greenstreet Berman & Synergy Consultants Ltd.
17. Hickling, E.M., and Bowie, J (2012) Applicability of human reliability assessment methods to human-computer interfaces". Cognitive Technology and Work, 15, pages19-27. Springer
18. Hollnagel, E., (1988)."CREAM: Cognitive Reliability and Error Analysis Method". Elsevier, Oxford, England.
19. Horvitz, E. (1999). "Principles of Mixed-initiative User Interfaces". Proc: CHI' 99, ACM SIGCHI Conference on Human Factors in Computing Systems, Pittsburgh, May 1999.
20. Huafei Liao, H., (2015). "Insights from Pilot Testing of the IDHEAS HRA Method". Procedia Manufacturing, 3, pp 1350 - 1357, Elsevier.
21. Hwang S-L, Liang G-F, Lin J-T, Yau Y-J, Yenn T-C, Hsu C–C, and Chuang C-F (2009) “A real- time warning model for teamwork performance and system safety in nuclear power plants”. Saf Sci, 47, pp425–435.
22. Interaction Design Foundation, (2020). "The Encyclopaedia of Human-Computer Interaction". Downloaded on 14/03/2020 from https[://w](http://www.interaction-design.org/literature/book/the-)ww[.i](http://www.interaction-design.org/literature/book/the-)n[teraction-design.org/literature/book/the-](http://www.interaction-design.org/literature/book/the-) encyclopedia-of-human-computer-interaction-2nd-ed.
23. International Nuclear Energy Research Initiative, INERI, (2015). "Human performance issues in analog, hybrid, and digital control rooms"., Eds., Park, J., W. Jung, Boring, R. L., Ulrich, T., Kovesdi, C., Saychenko, K., and Hallbert, B. P., INL/EXT-15-37277, Idaho National Laboratory, Idaho Falls, Idaho, USA. Also published as: KAERI/TR-6194/2015, Korea Atomic Energy Research Institute, Daejeon, Republic of Korea.
24. Jang, I., Kim, A.R., Al Harbi, M. A-S., Lee, S.J., Seung, J.L., Kang, H.G. and Seong, P.H. (2013) “An empirical study on the basic human error probabilities for NPP advanced main control room operation using soft control”. Nuclear Engineering and Design, 257, pp 79– 87, Elsevier, Netherlands.
25. Julius, J.A., Moieni, P., Grobbelaar, J.,and & Kohlhepp,K., (2014) . "Next Generation Human Reliability Analysis - Addressing Future Needs Today for Digital Control Systems". Proc: PSAM 12, Probabilistic Safety Assessment & Management Conference, Honolulu, HawaiHickling, E.M., Harvey, J.G., and Hollywell, P.D., (1992) "The Consistent and Structured Definition of Process Plant HCI-Based Tasks Using The 'Pagoda' Method".Proc: Ergonomics Society Annual Conference, in Contemporary Ergonomics, Taylor and Francis, London.
26. Jung, W., Park,J., Kim,Y., Choi, S-Y., and Kim, S. (2016). "An Approach of Simulated Data Collection and Analysis to Generate Human Error Probability" Proc: PSAM13, October 2 - 7, 2016, Seoul, South Korea.
27. Kaptelinin, V., (2020) "16 Activity Theory". Accessed on 25/03/2020 from : https[://w](http://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-)ww[.i](http://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-)n[teraction-design.org/literature/book/the-encyclopedia-of-human-computer-](http://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-) interaction-2nd-ed/activity-theory.
28. Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G. and Umbers, I. (2005), "Nuclear Action Reliability Assessment (NARA): a data-based HRA tool". Safety & Reliability, vol. 25. no. 2 pp 38 - 45. Safety & Reliability Society, Oldham, Greater Manchester, England.
29. Kontogiannis, T. and Moustakis, V. (2002). “An experimental evaluation of comprehensibility aspects of knowledge structures derived through induction technique: a case study of industrial false diagnosis”. Behaviour & Information Technology, 21, 2 pp117-135.
30. Land, C. E., 1972. “And Evaluation of Approximate Confidence Interval Estimation Methods for

Lognormal Means”. Techno metrics 14: PP 145 – 158.

1. Leontiev, A. N., (1978). "Activity, consciousness, and personality". Pergamon Press, Oxford, UK.
2. Lind, M. Seipel, S and Mattiason, C (2001). “Displaying meta-information in context”. Behav

Inf Technol, 20,6, pp427-432.

1. Liu,P. and Li, Z., (2016, 1). "Comparison between conventional and digital power plant main control rooms: A task complexity perspective, part 1: Overall results and analysis, International Journal of Industrial Ergonomics, 51 pp2 - 9, Elseivier.
2. Liu,P. and Li, Z., (2016, 2). " Comparison between conventional and digital nuclear power plant main control rooms: A task complexity perspective, Part II: Detailed results and analysis", International Journal of Industrial Ergonomics, 51 pages 2 - 9, Elseivier.
3. Liu,P. and Li, Z., (2014). "Human Error Data Collection and Comparison with Predictions by SPAR H, Risk Analysis", Risk Analysis, 34(9): pp 1706-19.
4. Mackenzie IS, Zhang SX (2001) An empirical investigation of the novice experience with soft keyboards. Behav Inf Technol, 20(6), pp 411–418.
5. Massaiu, S. and Fernandes, A. (2017) "Comparing Operator Reliability in Analog vs. Digital Human-System Interfaces: An Experimental Study on Identification Tasks". PSAM Topical Conference on Human Reliability, Qualitative Human Factors, and Risk Management 7 - 9 June 2017, Munich, Germany.
6. Mejdal, and McCauley, M.E. (2001)."Human Factors Design Guidelines for Multifunction Displays" DOT/FAA/AM-01/17. Office of Aerospace Medicine Washington, DC, USA.
7. Müsseler, J Meiunecke, C Döbler , J (1996). “Complexity of user interfaces: Can it be reduced by a mode key?” Behav Inf Technol, 15,5, pp 291-300.
8. Nordby K, Raanas RK, Magnussen S (2002) “Keying briefly presented multiple-digit numbers”.

Behav Inf Technol 21(1): pp27–38.

1. NRC (2005) "The SPAR-H Human Reliability Analysis Method". (NUREG/CR-6883, INL/EXT- 05-00509, US Nuclear Regulatory Commission, Washington, DC, USA.
2. Nuclear Regulatory Commission (2000). "Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA). Rev. 1, NUREG-1624, 2000, NRC, Washington DC. , USA.
3. NRC, (2005). "EPRI/NRC-RES Fire PRA Methodology for Nuclear Power Facilities Volume 2: Detailed Methodology". US NRC NUREG/C R - 6850, US Nuclear Regulatory Commission, Washington, DC, USA.
4. NRC (2009) "Regulatory Guide 1.200, An Approach For Determining The Technical Adequacy Of Probabilistic Risk Assessment Results For Risk-Informed Activities", Rev. 2. US Nuclear Regulatory Commission, Washington, DC, USA.
5. NTSB, (2010). "Introduction of Glass Cockpit Avionics into Light Aircraft", NTSB/SS - 01/10 PB 2010 - 917001, National Transportation Safety Board, Washington DC, USA.
6. Olsson, U. (2017) “Confidence Intervals for the mean of a Log-Normal Distribution”.

Downloaded from: <http://jse.amstat.org/v13n1/olsson.html> on 18/08/2020.

1. Park, J., Kim, Y., and Jung,W. (2017) Application of the HuREX Framework to Support Existing HRA Methods: Case Studies". Proc: PSAM Topical Conference on Human Reliability, Qualitative Human Factors, and Risk Management 7 - 9 June 2017, Munich, Germany.
2. Parkin, T. B., Chester, S. T. And Robinson, J. A. (1990). “Calculating Confidence Intervals for the Mean of a Log normally Distributed Variable”. Journal of Soil Science Society of America, 54: pp 321 – 326.
3. Pirzadeh, L. (2010). Human Factors in Software Development: A Systematic Literature Review, MSc Thesis for Chalmers University of Technology, Gothenberg, Sweden. Downloaded on 17/03/2020 from [http://publications.lib.chalmers.se/records/fulltext/126748.pdf.](http://publications.lib.chalmers.se/records/fulltext/126748.pdf)
4. Porthin, M., Liinasuo, M, and Kling, T. (2020). “Effects of Digitalization of Nuclear Power Plant Control Rooms on Human Reliability Analysis: A Review”. Reliability Engineering and System Safety, 194, [106415].
5. Rogers, Y., (2012) "HCI Theory: Classical, Modern, and Contemporary", Synthesis Lectures on Human-Centered Informatics, May 2012, Vol. 5, No. 2 , Pages 1-129, Morgan & Claypool, Online Publishers at https:/[/w](http://www.morganclaypool.com/)w[w.morganclaypool.com/](http://www.morganclaypool.com/) , <https://doi.org/10.2200/S00418ED1V01Y201205HCI014>).
6. Roth, E. M. , Mumaw, R. J., Lewis, P.M. (1994) “An Empirical Investigation of Operator Performance in Cognitively Demanding Simulated Emergencies”. NUREG/CR-6208 U.S Nuclear Regulatory Commission, Washington, D. C., USA.
7. Snowberry K, Parkinson S, Sisson N (1985) “Effects of Help Fields on Navigating Through a

Hierarchical Menu Structure”. Int J Man-Mach Stud 22, pp479–49.

1. SPAR-H: Gertman, D., Blackman, H., Marble, J., Byers, J. and Smith, C. (2005). “The SPAR-H Human Reliability Analysis Method”. NUREG/CR-6883, US Nuclear Regulatory Commission, Washington DC.
2. SPAR-H: Gertman, D., Blackman, H., Marble, J., Byers, J. And Smith, C. (2005). “The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883, US Nuclear Regulatory Commission, Washington DC.
3. Thimbleby, H., (2007). "User-centred Methods are insufficient for safety-critical Systems", Proc: USAB'07-Usability & HCI for Medicine and HealthCare, 3rd Symposium of the Austrian Computer Society, WG HCI&UE NOVEMBER 22, 2007, Medical University of Graz, Austria.
4. Thimbleby, H., (2009). "Contributing to Safety and Due Diligence in Safety-critical Interactive Systems Development by Generating and Analyzing Finite State Models". Proc: First Symposium on
5. Thimbleby, H.,(2013). "Action Graphs and User Performance Analysis," International Journal of Human-Computer Studies, 71(3), pp.276-302, Elsevier BV, RELX Group, London.
6. Thimbleby, H., (2020). Personal Communication to Ned Hickling 18/03/2020. Record held.
7. Trewin, S. and, Pain, H. (1999). “A model of keyboard configuration requirements”. Behav Inf

Technol 18, 1:27–35.

1. Whaley, A. M, , Stacey M. L. Hendrickson, S.M.L, Boring, R.L., Joe, J.C., Le Blanc, K,L., and Xing,J. "Bridging Human Reliability Analysis and Psychology, Part 1: The Psychological

Literature Review for the IDHEAS Method". PSAM 11, ESREL2012 Probabilistic Safety Assessment & Management and the Annual European Safety and Reliability Conference, 25- 29June, Helsinki, Finland.

1. Westerman, S.J., Davies, D.R., Glendon, A.I., Stammers, .RB. and Matthews, G. (1995) “Age and cognitive ability as predictors of computerized information retrieval”. Behav Inf Technol, 14, pp 313–326.
2. Wiedenbeck, S. (1999). “The use of icons and labels in an end-user application program: an empirical study of learning and retention”. Behaviour & Information Technology, 18.2, pp 68- 82.
3. Wikipedia, 2020. "Software development process" Downloaded on 21/04/2020from: https://en.wikipedia.org/wiki/Software\_development\_process]Wu, X., Gao, Q., Song, F., Liu, P., Li, Z., Dong, X., (2012) "Evaluating FBTA-Based User Interface Design for digital Nuclear Power Plants". Proc: PSAM 11, ESREL2012, Probabilistic Safety Assessment & Management and the Annual European Safety and Reliability Joint Conferences, 25-29June, Helsinki, Finland.
4. Zhao, F., Dong, X., Li, Z. (2012). "How does time availability influence the execution of computerized emergency operating procedures", PSAM 11, ESREL2012 Probabilistic Safety Assessment & Management and the Annual European Safety and Reliability Conference, 25- 29June, Helsinki, Finland.
5. Zhou, X-H., and Gao, S. (1997), “Confidence Intervals for the Log-normal Mean” .Statistics in

Medicine, 16, 783–790.

1. Zimmerman, J. (2018). "Glass cockpits - don't make it harder than it really is". Downloaded on 18/03/2020from: https://airfactsjournal.com/2018/03/glass-cockpits-dont-make-it-harder- than-it-really-is/ .