

Novel strategies to assess sparse data in human reliability analysis

Thesis submitted in accordance with the requirements of the
University of Liverpool for the degree of Doctor in Philosophy

by

Caroline Pinheiro Maurieli de Morais

March 2021



Supervised by

Prof Edoardo Patelli (primary)

Prof Michael Beer (secondary)

Dr Raphael Moura (industrial)

Abstract

Novel strategies to assess sparse data in human reliability analysis

Caroline (Pinheiro Maurieli de) Morais

Major industrial accidents are usually attributed to problems in the interaction of human, technological and organisational factors. Although many of these accidents are almost impossible to predict, some can be predicted and prevented by techniques dealing with human error assessment. This is the reason why it is expected that comprehensive risk analyses use a technique known as human reliability analysis. It usually relies on data elicited from experts with operational knowledge, or empirically collected from simulated scenarios, records of near-misses, and major accident reports. The present research proposes the use of MATA-D (Multi-attribute Technological Accidents Dataset), which is based on major accident investigation reports, thus potentially capturing a more realistic relationship between human erroneous actions and technological and organizational factors that shape human performance.

Currently, the most recommended probabilistic tool to model human reliability data is the Bayesian network. However, the assessment of its conditional probability tables (CPTs) requires enough data to describe all possible conditions dictated by the model. However, despite increasing collection efforts of empirical human reliability data, the available databases are still insufficient to fulfil conditional probability tables, especially for models where each variable is conditioned on many others. In these cases, the most common solution relies on the adoption of expert elicitation to fill in the missing combinations.

This research has been focused on developing strategies to enable empirical data-driven human reliability analysis, such as precise and imprecise probability tools to tackle epistemic uncertainty inherent to such databases. The used probabilistic tools are Bayesian and credal network, the latter to tackle missing data in conditional probability tables. Using credal networks the prediction analysis depicts results with interval probabilities rather than point values measuring the effect of missing-data variables. As taking decisions is more difficult when comparing intervals than point-values, a decision-making strategy is suggested to unveil the most relevant variables for risk reduction in presence of imprecision. The results support the hypothesis that realistic uncertainty depiction implies less conservative human reliability analysis and improves risk communication between assessors and decision-makers.

Finally, a natural language processing technique based on machine-learning has been developed to extract and classify new accident reports in order to collect new data for MATA-D. This aims to decrease the number of missing combinations in CPTs. A constant collection of new data for this dataset aims not only to decrease epistemic uncertainty in human reliability data but also to timely update models, reflecting changes in human behaviour due to evolving technology and organisational arrangements. The automated approach, called the virtual *human factors classifier*, is able to classify a new report more than one thousand times faster than a human being.

Future developments are discussed, such as the strategy to compute reliability analysis with confidence, by using credal networks and c-boxes to tackle different and sometimes very small sample sizes in a database.

Acknowledgements

I would like to acknowledge my supervisors – Professor Edoardo Patelli, Prof Michael Beer and Dr Raphael Moura for not always providing me with the answers that I expected, but with those that I needed. Special thanks to my primary supervisor, Edoardo, for promoting his regular research group meetings – a unique moment to share our ideas.

My employer, the Brazilian Oil & Gas regulator (ANP), for supporting my research. Especially those who endorsed my temporary absence from work to focus on my PhD in 2017 (Marcelo Mafra, Waldyr Barroso and Magda Chambriard) and those who are supportive of applying the lessons learned in my PhD now that I am back to work (Nayara Ferreira, Thiago Pires, Mariana Franca and Raphael Moura).

My husband and my daughter, Leandro Maurieli and Camila Morais, not only for inspiring me and filling my life with purpose but also for sharing the adventure of leaving everything behind to pursue a PhD in another country. Our limits have been tested by high costs and high uncertainties – I am glad we have found extra energy to remain happily united. Thanks also to Mike, who has complemented our family. Also, not forgetting our dog, Sushi, for giving us an extra reason to laugh!

My parents, Zaida and José, for supporting me in all life dimensions. My sister and brother, Clarissa and José, for being my role models. My mother-in-law and all my in-laws, for understanding our absence and dreams.

Institute director Scott Ferson for his inspirational leadership which has fuelled our minds with challenging as well as amusing discussions about risk communication and imprecise probability. Together with his extensive knowledge covered by his intriguing humbleness, our Institute has been paved with a solid base of true respect for women, other nationalities and minorities which has made us feel safe in knowing that we would be judged only for our ideas.

All the Risk institute colleagues, specially Noemie and Raneesha, for sharing both life and academic ideas. To my inspirational research group, especially Diego and Adolphus for always supporting my research as well as finding time to chat. The Risk Institute managers Marco de Angelis and Dominic Calleja for providing a great atmosphere, interesting lectures and discussions.

The HSE colleagues who have supported my research, especially Ed Corbett who had found time in his agenda to present his team and premises in Buxton.

LDC Development Team, with their support for PGR students. Especially Dr Shirley Cooper who shares the importance of non-technical skills for academics, and Dr Eli Saetnan for providing training on how to teach adults. Also, in the LDC team, Lynne Elliot, who helped to manage the writing retreats during the COVID-19 crisis, and for sharing PhD Goals almost every day of this pandemic.

The University's English Language Centre, for the support provided to international students, especially to Dr Jeni Driscoll who runs the Thesis writing for international research students.

Carole Rhodes from the Faculty library and Jack Carter-Hallam from the School of Engineering.

Alexandre Glitz for being an inspirational colleague and always supportive of my research.

Suzanne, my dear friend from the English department who knows exactly the pains and gains of being a mature and self-funded PhD student.

To all the scouse friends that made my life full of good memories outside the office: my friends from Diana's Society (Diana, Amp, Ana, Lilith and Yijun), Ceri, my favourite artists and neighbours in Liverpool (Alex and William) and my Capoeira mates (especially Mestre Parente and Tequila).

My ANP colleagues from whom I learn every day. I know that the absence of only one person can impact a small team – thanks for supporting my PhD anyway.

Finally, to all offshore oil & gas workers, from operators to offshore installation managers, and the families and work colleagues from the nine lost souls in the explosion at FPSO Cidade de Sao Mateus: I will always use the best of my knowledge and skills to help prevent such an accident happening again in Brazil.

A post-VIVA note to my friend and sister-in-law, Luciene: one of my dreams was to go to your graduation day. Sadly, you left us during pandemic. I promise your son will graduate in a timely manner to enjoy life as it should be.

Table of Contents

Abstract	3
List of abbreviations.....	6
Chapter I: Introduction	1
1. Aims and objectives of this research.....	5
2. Original contribution.....	6
3. Thesis structure	8
Chapter II: Using a dataset of accident reports to model human behaviour – the first contact with sparse data issues	10
Analysis and estimation of human errors from major accident investigation reports	13
1 Introduction	13
2 Methodology background	14
3 Proposed approach: using datasets of major accidents reports	20
4 Case study.....	23
5 Conclusions	39
Chapter III: Using credal networks to assess sparse empirical data	41
Robust data-driven human reliability analysis using credal networks	44
1. Introduction.....	44
2. Theoretical background	45
3. Proposed methodology.....	57
4. Case study	68
5. Conclusions.....	98
Chapter IV: Minimising epistemic uncertainty by collecting new data	100
Identification of human errors and influencing factors: a machine learning approach	102
1. Introduction.....	102
2. Theoretical background	104
3. Methodology.....	112
4. Case studies.....	121
5. Discussion.....	130
6. Conclusions.....	133
Chapter V: Modelling human reliability with confidence	135
Chapter VI: Conclusion	138
List of publications	141
Bibliography	144
Appendices	155

List of abbreviations

ANP: National Agency for Petroleum, Natural Gas and Biofuels (*Agência nacional do petróleo, gás natural e biocombustíveis*, the Brazilian oil, gas and biofuels regulator)

CPD: conditional probability distribution

CPT: conditional probability table

CREAM: cognitive reliability and error analysis method

FPSO: floating production storage and offloading

FSO: floating storage and offloading unit

FSU: floating storage unit

HE: human error

HF: human factors

HEP: human error probability

HRA: human reliability analysis

MATA-D: multi-attribute technological accidents dataset

MLE: maximum likelihood estimator

NRC: nuclear regulatory commission

PSF: performance shaping factor

UK : United Kingdom

Chapter I: Introduction

In February 2015, an explosion at an offshore oil & gas installation in Brazil killed 9 workers and left 26 seriously injured. Since the beginning of the investigation into the accident, two human errors had been evident among the causes: one from the operator controlling the system who had started the emergency scenario that would last for one hour before the explosion, and one from the installation manager, who as an emergency commander had not been able to correctly mitigate the situation and permitted people to access an area where the gas sensors and alarms had already shown the potential to cause an explosion (ANP, 2015). If the investigation had stopped at the most immediate causes, the conclusion could be that human errors alone had led to the accident. However, the accident investigation report has shown that the human errors were rooted in causes more related to technological and organizational factors, such as flaws in the commission (i.e. implementation) of the human-machine interface of the control room, as well as flaws in the emergency card instructions that the manager had followed (ANP, 2015, Morais et al., 2016, Moura et al., 2017c). This confirms what Trevor Kletz said about major accident investigation practice: if an accident could be compared to an onion, a human error would be its outer layer. To understand why those human errors have occurred the investigators have to peel the onion down to its deepest layers, which will very often end up unveiling organizational and technological failures (Kletz, 2001). This affirmation has been empirically demonstrated by a recent study that, based on an analysis of major accident reports, has revealed that less than 1% reportedly have human errors and person-related factors alone (Moura et al., 2016), and that 48% are a combination of human-technology-organization factors.

After acknowledging that problems in this interaction are the cause of a large part of major industrial accidents, it is important for society to understand in which extension they are predictable, to make decisions such as choosing a factory location (e.g. Seveso Directive 2012/18/EU). Some risks are usually not desired by society due to their potential to cause the loss of life of workers and nearby communities (e.g. the aftermath of the Bhopal accident) (Broughton, 2005), the environmental impact (e.g. the blowout at the Macondo well, leading to approximately 5 million barrels of oil spilt into the Gulf of Mexico in 2010) (CSB, 2014), the loss of assets and sustainability of local economies (e.g. the above-mentioned explosion at the offshore installation in Brazil which also damaged the installation thereby impairing the production continuity and impacting the natural gas supply in the region) (Morais et al., 2016).

Predictions are always difficult and affected by huge uncertainty. However, techniques dealing with human error assessment can be used to improve our knowledge and prediction (Kirwan, 1994). Acknowledged techniques that improve the interaction of humans with technology are known as human factors engineering (HFE), and techniques that predict and prevent human error risks are known as human reliability analysis (HRA).

Although human-machine interactions started to be studied during the second world war to decrease plane crashes (Wickens et al., 2015), the interaction between workers, tools, technologies and techniques started to be systematically studied in 1951 in the coal mining industry, when it was defined as a sociotechnical system (Trist and Bamforth, 1951) – a project initiated by the Human Factors Panel of the Committee on Industrial Productivity in the United Kingdom (UK). Since then, Human factors (HF) have been more often acknowledged as the study of systems' design to ensure that the demands to humans do not exceed their natural capabilities (both sensory-motor capabilities, and cognitive functions such as decision making and problem solving) (Hollnagel, 1998). The UK Chartered Institute of Ergonomics and Human Factors (CIEHF) states that human factors can be used interchangeably with the term ergonomics, although this is more often used for physical aspects of the workplace while human factors often encompass the wider system such as organisational factors. Human factors engineering (HFE) is the application of human factors knowledge to the design and construction of socio-technical systems (EI and IOGP, 2020). It is composed of prescriptive guidelines of recommended practices in the industry to fulfil such objectives, e.g. placement of valves to facilitate safe and efficient access, and analysis and review of human-machine interface (HMI), control rooms and alarms (EI and IOGP, 2020). However, although human factors engineering gives guidance on improvements in such systems, it does not assess their risks.

To assess a system's overall risk, a human reliability analysis (HRA) is necessary. The technique consists of a systematic process of analysing the risks arising from the human-technology-organization interactions and has started to be more seriously studied after the incident at the Three Mile Island nuclear power plant in 1979 (Kirwan, 1994). Human reliability analysis methods (some of them described in Chapter 2) rely on models of how human performance depends on the conditions in which the tasks are carried out – these conditions usually referred to as *performance shaping factors* (PSFs, Hollnagel, 1998). HRA can be used as a stand-alone analysis or combined with a component reliability analysis into a risk analysis (Hollnagel, 1998, Kirwan, 1994), as depicted in Figure 1-1.

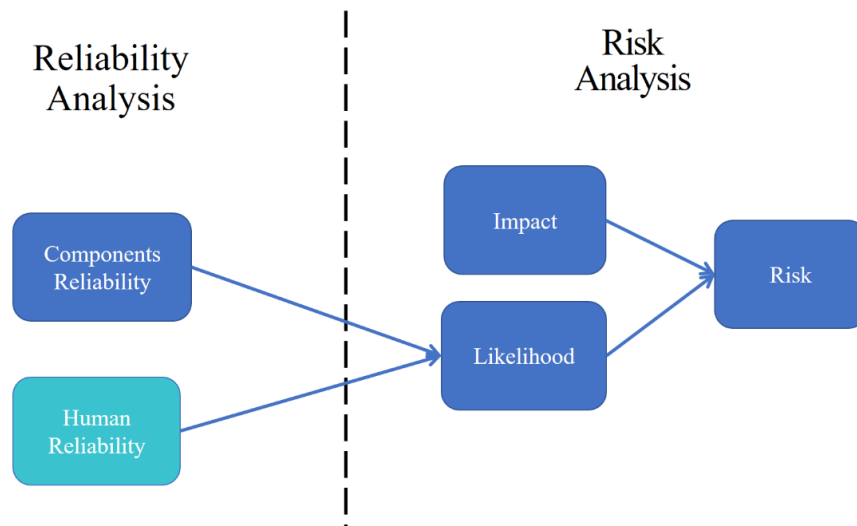


Figure 1-1. Human reliability analysis within risk analysis

From the available literature regarding existing HRA methods, it could be noticed that human reliability analysis use depends on the industry sector, e.g. there are more methods and described applications in nuclear power plant industries than in oil & gas exploration and production (Bell and Holroyd, 2009). On the other hand, human factors engineering has been growing at a large pace in the oil & gas industry (EI and IOGP, 2020).

It has been argued that human factors and human reliability analysis communities are artificially separated and could be contributing more to each other. One contribution would be to share empirical data collection efforts. Their different historical origins have resulted in different research approaches: although human factors studies need to collect data to generate design recommendations, HRA has been relying heavily on expert estimation and HRA methods to make probabilistic predictions (Boring and Bye, 2008). In industry practice, HRA relies on empirical data only to validate methods and studies.

One previous study has investigated a method to adjust HRA data using human factors data. The study has compared error probabilities in test and control conditions, and the resulting probability ratios have been suggested to inform the selection of performance shaping factors multipliers in HRA methods (Griffith and Mahadevan, 2015). However, it is not an easy task to fully convert human factors data into HRA data: although the HF community is concerned with the existence of a phenomenon (qualitative), HRA is concerned with its frequency, translated on the *human error probability* (quantitative). Therefore, human factor data usually misses the denominator *number of opportunities for error* that is essential for HRA (see Equation 1-1).

Equation 1-1

$$\text{human error probability} = \frac{\text{number of observed errors}}{\text{number of opportunities for error}}$$

The use of empirical data to directly inform a human reliability analysis is still a research topic (Groth et al., 2014), as industry practice mostly relies on empirical databases to validate expert elicitation results (Kirwan, 1997a). Some attempts of empirical data collection efforts had been found lacking due to the users' perception that they are unacceptably variable compared to components reliability (Kirwan, 1994), and that the observation of some human errors is so small that is not "statistically significant" (Kim, 2020). At the same time, there are complaints that the HRA traditional approaches are very time-consuming processes (Kirwan, 1994), and provide overly conservative results (French et al., 2011). The cost of conservatism might ultimately lead to over-designed plants (Kirwan, 1997a), consequently to inadequate plans of resource allocation. Over-conservative results might be due to the lack of knowledge on realistic performance shaping factors and on how those factors actually impact human performance (Liu and Liu, 2020). The majority of HRA methods consider that performance shaping factors degrade human performance, when human factors research shows that some of them indeed improve (CA Authority, 2016). Also, it is difficult to elicit the combined effect of more than three performance shaping factors in human performance (Liu and Liu, 2020), due to the inability of experts to reason under more than three conditions (Evans et al., 2003).

These issues might be preventing new industry sectors to adopt HRA as the primary approach to systematically account for the interactions between humans, technological and organisational factors in risk assessments (Zio, 2018). Therefore, this research attempts to find solutions to these issues unlocking three gaps in the HRA niche: the lack of experiments that use empirical data that better depict the influence of realistic performance shaping factors, the lack of use of probabilistic methods that embrace and depict variation, and lack of methodologies that avoid the use of expert judgements or strong assumptions in the quantification step in cases of missing empirical data.

From the assumption that major accident reports have the potential to provide realistic data about the interactions of performance shaping factors and human performance, this research has experimented using the Multi-Attribute Technological Accidents Dataset (MATA-D), a dataset derived from a collection of major accident reports classified into a human reliability classification scheme (Moura et al., 2020, Moura et al., 2016). Before this research,

MATA-D has never been used on human reliability analysis. Previous analysis of this dataset has suggested that the human-technological-organisation interactions do have a pattern (Moura et al., 2017a), and the dependency between variables makes it suitable to feed not only classic probabilistic models such as Fault Trees but also Bayesian networks.

Choosing the right method to model might be also impeding the use of empirical data. Although research has recognised that Bayesian networks better describes socio-technical systems (Mkrtchyan et al., 2015), HRA industrial practice heavily relies on Fault Trees (Kirwan, 1994). The problem of using classic fault trees is that they might be missing combinations that are very rare – as they provide only an approximate method, called cut sets, that drops the smallest term (Fenton and Neil, 2012). On the other hand, discrete Bayesian networks are exact, as they account for every combination.

However, accounting for every combination also has a cost: it needs more data to describe the combinations (Mkrtchyan et al., 2016). This generates a problem that can be even worse than the inaccuracy of fault trees: by assessing expert judgements to inform the missing combinations in Bayesian networks, a full HRA might be exposed to expert's bias (Mosleh et al., 1988) and will be more time consuming (Wisse et al.). In other words, by relying solely on Bayesian networks, there is a chance that the HRA field might always consider the HRA data-sparse - even with so many new collection efforts being conducted. For this reason, this research investigates other causal probabilistic tools as well as other methods that can describe some missing combinations without requiring expert's probability judgments or any other strong assumptions.

1. Aims and objectives of this research

The present research aims to help make human reliability analysis more acceptable to the engineering community by: making it faster, less conservative and more transparent.

(1) making the quantitative step faster by using empirical data;

(2) making it less conservative by using a probabilistic tool, and by using a dataset that provides the relations between performance shaping factors and human errors that have led to incidents;

(3) making it more transparent to decision-makers by showing how missing data on interactions of individual, organisational and technological factors impact the results.

To achieve this aim, two objectives have been defined: (i) to model human reliability with a probabilistic tool and existing dataset; (ii) to propose solutions to tackle sparse data on conditional probability distributions without eliciting experts.

Figure 1-2 summarises the research aims and objectives, followed by the respective research solutions and the conference and journal papers where they have been discussed and published.

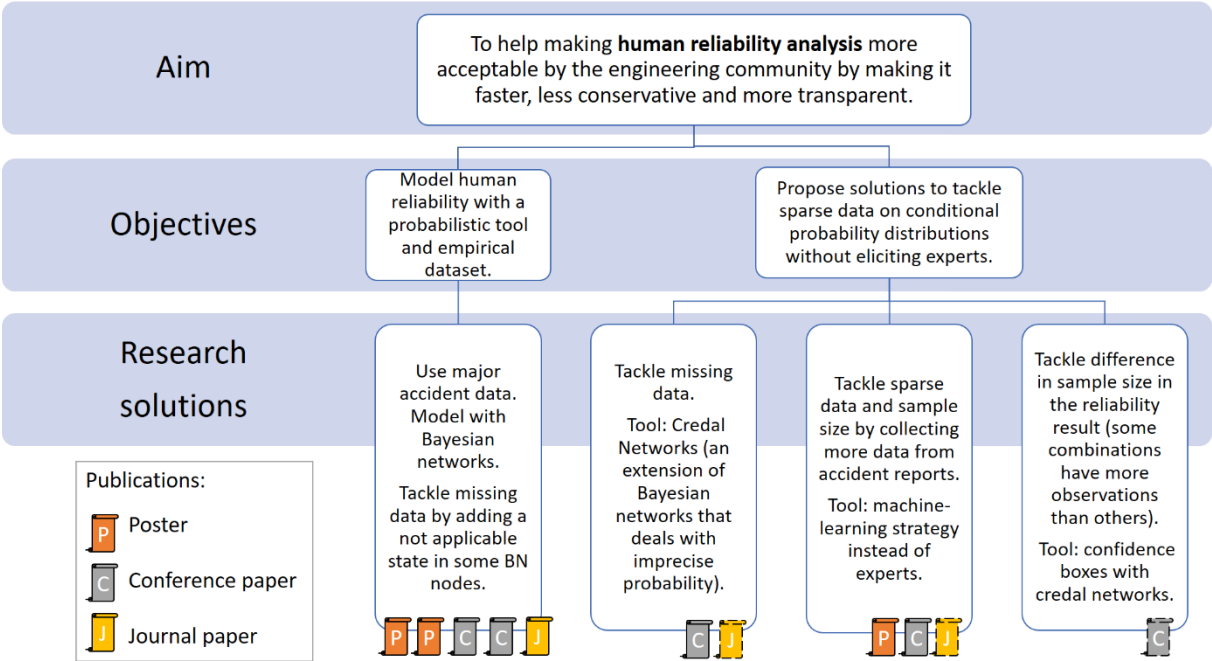


Figure 1-2. Aims, objectives and original contribution

2. Original contribution

The thesis argues that the human reliability community might have enough data to rely on data-driven analysis if the right imprecise probability tools are used. The main original contributions of this study are represented in Figure 1-2 as *research solutions*.

The original contributions are the developed models, tools and techniques for addressing the problem of lack of data in human reliability, all able to depict the uncertainty of empirical data. In summary, the developed models have been based on:

Research solution 1) The use of empirical data from major accidents (MATA-D). Modelling tool: Bayesian networks. Tackle missing data by adding a not applicable state in some BN nodes.

In this research, a model of human behaviour has been developed, and a novel methodology to estimate human error probabilities (HEPs) using data from major accident investigation reports. The approach is based on Bayesian Networks used to model the

relationship between performance shaping factors and human errors. The proposed methodology allows minimizing the expert judgment of HEPs, by using a pragmatic strategy that is able to accommodate the possibility of having no information to represent some conditional dependencies within some variables. Therefore, the approach increases the transparency about the uncertainties of the human error probability estimations. The approach also allows to identify the most influential performance shaping factors, supporting assessors to recommend improvements or extra controls in risk assessments. Formal verification and validation processes are also presented.

Research solution 2) Tackle missing data in conditional probabilities. Modelling tool: credal networks (an extension of Bayesian networks that deals with imprecise probability)

A human reliability model of a real operation in the oil & gas industry has been developed, after carefully selecting an operation that is safety-critical and that comprises interactions between humans (from different teams), technological and organizational factors. The manuscript describes all the qualitative and quantitative steps taken to translate textual and numerical data from documents such as operational procedures, hazard analysis and description of previous related incidents. A methodology has been developed to admit (and depict) missing data without strong assumptions. However, to use this methodology, it has been necessary to shift from Bayesian to the credal network as the probabilistic tool. In order to increase the acceptance of the credal network in the oil & gas community, a methodology known as *bow-tie assessment* is explored to shape the model, as its structure is well accepted in this industry sector.

Research solution 3) To tackle sparse data and sample size by collecting more data from accident reports (expand MATA-D). Tool: machine-learning strategy instead of experts (bag-of-words to extract and SVM to train against MATA-D and classify new reports).

A novel strategy to extend and update the empirical dataset has been proposed: the use of natural language processing with machine-learning to extract and classify information from new major accident reports. The strategy is proposed not only to reduce epistemic uncertainty but also to continuously update the dataset with information from the impacts of newly developed technology on humans. Besides training a model with MATA-D, and showing the model prediction metrics for a test set of major accident reports, the results from two case studies of new reports from the aviation and oil & gas industries have been comprehensively analysed. The analysis is useful to understand how the model works for the false negatives and

false positives, together with a discussion of which of them might have the most impact on human reliability analysis.

Research solution 4) Strategy under development: calculating the reliability with confidence, to tackle differences in sample size in the reliability result (some combinations have more observations than others). Tool: confidence boxes (c-boxes) with credal networks.

Most of the attempts aimed at substituting expert-driven human reliability assessment methods with empirical data-driven techniques have failed due to the high uncertainty of human reliability databases and limitations of traditional probabilistic tools to deal with it. Although the previously proposed models show that Bayesian and credal networks could be a more suitable approach to model human reliability data, such analyses usually apply some modelling procedures such as normalisation, which have the potential to implicitly affect the degree of information regarding the unevenness of sample sizes. In this last research solution, we propose to tackle these limitations by using confidence boxes (c-boxes) with credal networks, aiming at providing risk assessors with a rigorous framework for data uncertainty leading towards more efficient and robust modelling solutions.

The small icons on the bottom of each research solution in Figure 1.2 correspond to the related publications which describe the models and methodologies developed during the PhD. The dashed line icons represent the papers under peer review (on the date of the thesis submission), and the solid line icons are the publications already issued.

3. Thesis structure

This thesis is structured as a *collection of papers*, a compilation of research manuscripts written during the PhD. The research solutions have been presented in three manuscripts submitted to peer-reviewed academic journals. All of them have been preceded by conference papers and discussed in conference presentations with peers, which have proved to be a good way of testing their usefulness and originality within the research community.

The second chapter is based on the first journal paper, where the first research solution presented in Figure 1-2 has been applied. This research paper is published in the [Special Issue of Human Performance and Decision Making in Complex Industrial Environments \(SI034B\) ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems – Part B: Mechanical Engineering](#). Besides me as the leading author, it has three co-authors, all my PhD supervisors who have contributed to the conceptualization and mainly with reviewing and suggesting improvements to my first manuscript.

The third chapter starts with an overview of the challenges and shortcomings of the first journal paper, along with the motivations for the second research solution presented in Figure 1-2. The chapter continues by presenting the second article, in its version accepted for publication in the *Reliability Engineering & System Safety Journal*. Besides me as the leading author, this paper has three other co-authors in addition to my supervisors: two colleagues from my research group that have mainly contributed with software code and part of the methodology, and a colleague from my teamwork in Brazil who has contributed with the case study selection and data.

The fourth chapter describes additional challenges and motivations for the research solution presented in Figure 1-2. The chapter is based on the third article in its version accepted for publication in the *Safety Science Journal*. Besides me as the leading author, this paper has two other co-authors in addition to my supervisors, who have mainly contributed to software code and part of the methodology.

The fifth chapter describes the final challenges faced when dealing with sparse data in a human reliability database. This chapter focuses on the research question and a summary of the fourth research solution presented in Figure 1-2, instead of a full description of the methodology. It has been decided not to include the preliminary analysis of this piece of research in this thesis, as it has been submitted to a conference rather than an academic journal. It is the intention that in the near future it will lead to a fourth manuscript.

The sixth and last chapter wraps up the main conclusions of the thesis and presents possible future research directions.

Other conference papers and posters have been produced during the PhD, as presented by the publication icons on Figure 1-2. The dotted icons are the papers still under peer review. Other contributions during PhD are white papers, a co-creation of a new research project, reviews in journals, publications and a book revision. They are listed in the *List of publications* on page 142, that also presents details on the submission dates of the above-mentioned manuscripts.

Finally, the appendices contain some of the developed models coded in MATLAB, as well as more detailed input and output data (e.g. conditional probability tables, precise and imprecise probability results).

Chapter II: Using a dataset of accident reports to model human behaviour – the first contact with sparse data issues

Overview

The objective of the first part of the research has been to assess an existing empirical dataset to model how human performance is impacted by performance shaping factors. Due to the importance of learning from previous major accidents, the dataset chosen has been MATA-D (Moura et al., 2020). This importance can be explained by using again the example of the Brazil offshore installation accident case cited in the *Chapter 1 Introduction*. Comparing it with previous major accidents, it can be noticed that root causes are similar not only to accident reports from the oil and gas industry but also to other industry sectors with a similar level of complexity regarding human-technology-organisation interactions. Nonetheless, few practical risk assessments do consider databases from different industries in their risk quantifications.

To embed the learning from the accidents process in human reliability analysis this research proposes to use a human reliability dataset based on major accidents from different industry sectors: the MATA-D (Moura et al., 2020). In the past, other work has been done to use accident database data for HRA models. Although this work also investigated events in different industry sectors (nuclear, aviation, maritime and occupational safety), it had not differentiated between near miss or accidents but focus on the psychological mechanisms behind each event (Sträter, 2000). From a preliminary study to understand MATA-D's potential use, a comparison between the Brazil offshore installation accident in 2015 with the other accidents in this dataset has been made, and it has been observed that the combination of the human erroneous action and the organizational factors are exactly the same as the Bayer CropScience Pesticide Waste Tank Explosion in 2008 (CSB, 2011). Furthermore, those human errors and performance shaping factors are similar to a set of previous accidents that occurred in many industry sectors (Moura et al., 2017c). Plus, previous study has shown that some combinations are recurrent and do have a pattern (Moura et al., 2017a), being possible to conclude that the combinations are not governed by aleatory uncertainty alone (true random or uncontrollable processes), but also by epistemic uncertainty (which can be reduced, at least theoretically by collecting new data or using more detailed models) (Patelli et al., 2016).

Moura et al.'s study to obtain MATA-D has focused on retrospective HRA (i.e. assessing the risk of accidents that have already happened), while the present study objectives are to use their data on prospective HRA (i.e. assessing the risk of something that hasn't actually happened, such as the alignment of irregular working hours with design failures) (Hollnagel,

1998, Boring and Bye, 2008). In fact, this was one of the strongest reasons which have led Moura et al. to use the human reliability classification scheme from the Cognitive Reliability and Error Analysis Method (CREAM) (Hollnagel, 1998) – for its capability of being applied in both ways (Moura et al., 2016).

To use it in a prospective analysis, a probabilistic tool able to model causality had to be chosen among many other data analysis tools. The way the MATA-D has preserved the dependency between all variables (human errors and performance shaping factors) for each event made it possible to be used in methods that rely on conditional probability tables, such as Bayesian networks. Besides that, the literature review has pointed to Bayesian networks as the strongest candidate to HRA, e.g. for its capability of accounting for the dependency between performance shaping factors (Mkrtchyan et al., 2015), and for the possibility of having its results explainable and traceable (Arrieta et al., 2020).

The biggest challenge of this research has been unveiled at the moment the variables of the network – especially those with many dependencies – started to be assessed: when the MATA-D frequencies are translated to conditional probability tables, many of the combinations are empty for all states of a variable. The feeling that it could be a barrier to this study has turned into its most important research question when it has been realised, through the literature review, that this was one of the most important practical issues in quantifying HRA (Mkrtchyan et al., 2016): the issue of sparse data.

It might sound strange to mention the sparse data problem, given that the human reliability and human factors communities have been generating more empirical data than ever (see session 2.1 *Data*). However, sparse data will continue to be an issue for those modelling with Bayesian networks or any other method that accounts for the conditional dependencies among variables. The use of Bayesian networks requires much more data than, for example, fault trees.

In summary, the following section of this thesis proposes a realistic and innovative approach for estimating human error probabilities using data from major accident investigation reports. The approach is based on Bayesian Networks used to model the relationship between performance shaping factors and human errors. The proposed methodology allows to minimize the expert judgment of human error probabilities, by using a strategy that is able to accommodate the possibility of having no information to represent some conditional dependencies within some variables. Therefore, the approach increases the transparency about the uncertainties of the human error probability estimations. The approach also allows to

identify the most influential performance shaping factors, supporting assessors to recommend improvements or extra controls in risk assessments. Formal verification and validation processes are also presented.

The next pages of this chapter are based on the first manuscript originated from this first phase of the research. I have been the leading author, and responsible for the conceptualization, data analysis, methodology and writing the first draft. The article has been co-authored by Dr Raphael Moura¹, Prof Michael Beer², and Prof Edoardo Patelli³.

¹ National Agency for Petroleum, Natural Gas and Biofuels (ANP), Av. Rio Branco, 65, CEP 20090-004, Centro, Rio de Janeiro, RJ, Brazil, and Institute for Risk and Uncertainty, University of Liverpool, Chadwick Building, Peach Street, Liverpool L69 7ZF, United Kingdom

² Institute for Risk and Reliability, Leibniz Universität Hannover, Callinstr. 34, 30167 Hannover, Germany, to Tongji University, Shanghai, China, and to the Institute for Risk and Uncertainty at University of Liverpool

³ Centre for Intelligent Infrastructure, University of Strathclyde, James Weir Building, 75 Montrose St, Glasgow G1 1XJ, United Kingdom, and to the Institute for Risk and Uncertainty at University of Liverpool

Analysis and estimation of human errors from major accident investigation reports⁴

1 Introduction

Despite the increasing level of automation and the advent of artificial intelligence (Ramos et al., 2018), realistic risk assessments of high-hazard industries should ideally be performed through the analysis of the complex interaction between human, machine, and organizational systems (Zio, 2018).

Human reliability analysis defines a collection of qualitative and quantitative methods used to account for human factors in social-complex industries in a systematic way (Henderson and Embrey, 2012). Their main aims are to identify the possible human errors in a task (i.e., task analysis) (Kirwan and Ainsworth, 1992), to quantify them (when needed), and to propose solutions to prevent or mitigate human errors (Kirwan, 1997a). The analysis uses the assumption that human errors are triggered by the interaction among individual, technological, and organizational factors, the so-called performance-shaping factors.

Qualitative methods for human reliability provide only the identification of human errors and possible preventive or mitigation solutions. Quantitative human reliability methods provide the same functions as the qualitative methods, plus an estimation (or an adjustment) of the human error probabilities (HEPs) according to the defined performance shaping factors in a specific scenario. Different quantitative human reliability methods exist, including technique for human error rate prediction (THERP) (Swain and Guttman, 1983), standardized plant analysis risk-human reliability analysis (SPAR-H) (Gertman et al., 2005), human error assessment and reduction technique (HEART) (Williams, 1988), cognitive reliability and error analysis method (CREAM) (Hollnagel, 1998) and a technique for human event analysis (ATHEANA) (Cooper et al., 1996). These quantitative methods allow to find or adjust human error probabilities according to the performance shaping factors in the specific industrial context being assessed (organizational, technological, and individual factors). However, human error probabilities obtained with quantitative methods are often affected by imprecision, sparse, and/or incomplete human error data (Bye, 2018, Kirwan, 1997b) leading to under-estimated or over-estimated probabilities (Kirwan, 1997a). This uncertainty may be one of the causes that are preventing industries from adopting risk assessments that account for human errors (Zio, 2009). Although some safety regulators do accept qualitative analysis on human errors (e.g.,

⁴ <https://doi.org/10.1115/1.4044796>

see Ref. (Bell and Holroyd, 2009)), human error probabilities are required by probabilistic safety (risk) assessments.

Ideally, a human error probability should be obtained by observing operators performing specific tasks and quantifying the frequency of their errors

Equation 2-1

$$\text{human error probability} = \frac{\text{number of observed errors}}{\text{number of opportunities for error}}$$

However, this is often an impractical task due to the variability of human behaviour, industrial installations, and tasks performed. The current research presents a novel methodology to estimate human error probabilities by collecting data from major accident reports. Bayesian networks are proposed to estimate human error probabilities to exploit information about the conditional dependencies among human errors and performance shaping factors. The present methodology also addresses the problem of working with sparse data, which eventually leads to incomplete conditional probability distributions for some nodes of the Bayesian networks. The approach consists of creating an additional state for those variables, in order to accommodate and account for the lack of information. It is believed that this strategy increases the transparency about the uncertainties of the human error probability estimation without the need of additional assumptions. This approach has the potential to better capture the interaction between human, machine, and organizational systems, providing additional contexts and scenarios not fully achieved by simulators, near-miss reports, and expert elicitation.

2 Methodology background

This section presents the proposed approach and theoretical background for the estimation of human error probabilities, including data collection, data analysis, verification, and validation.

2.1 Data collection

Data collected from real operations are considered the most credible human error data, followed by data derived from real operations (i.e., incidents, near-misses, and accidents), simulators and expert judgment (Figure 1-2) (Kirwan, 1997a).

A summarized description of the strengths and pitfalls of each type of data are described in the following.

Expert judgment: Experts are individuals with recognized knowledge or skill in a specific domain. Sometimes expert elicitation is the only available data source (Mosleh et al., 1988); thus, their opinions are aggregated by adopting methods to reduce expert elicitation variability (Mkrтчhyan et al., 2016, Shirazi, 2009). However, expert elicitation is considered the least credible source of data. This is because experts can be oriented by different sources of bias (Mosleh et al., 1988), be systematically overconfident about the accuracy of their judgments (Lin and Bier, 2008) and be experienced in the taxonomy used (Kirwan, 1997a). Ultimately, it is improbable to have a human reliability analysis that does not rely on expert judgment to some extent, as all methods start with a qualitative analysis of possible scenarios (Laumann et al., 2018).

Simulators: Data from simulators are collected at mimicked control rooms or other workspaces where real operators perform specific tasks under normal or emergency scenarios. Data collected from simulators is often restricted to human-machine interfaces in control rooms. Often collected data needs to be calibrated by expert judgment adopting well known approaches, e.g., scenario authoring, characterization, and debriefing application (SACADA) (Chang et al., 2014), Halden Man–Machine Laboratory (HAMMLab) (Gertman et al., 2005, Lois, 2009), human reliability data extraction (HuREX) (Kim et al., 2017), and operator performance and reliability analysis (OPERA) (Park and Jung, 2007). This approach is strong on detecting human errors, but weak on detecting all the performance shaping factors. This is due to the decontextualization of the studied tasks (Gertman et al., 2005), for instance, operators know that their actions will not have any consequence and often know that their actions are being observed (Kirwan, 1997a).

Derived data from real operation: Data from real operations come from direct task monitoring, near-miss events, and major accidents.

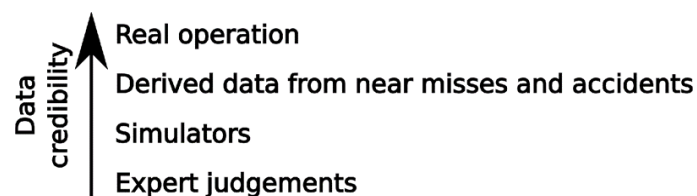


Figure 2-1. Data credibility for human error probability assessment (adapted from Ref. (Kirwan, 1997a))

The direct task monitoring is the method where a real operational task is observed at the moment it is performed by an assessor or recorded and analysed after the event. It is considered one of the best data sources but it lacks data for tasks rarely performed. For instance, the database computerized operator reliability and error database (CORE-DATA) has been partially generated with data derived from real operations (Gibson and Megaw, 1999).

Data from near-miss events are those that collect human errors and performance shaping factors from events that had the potential to cause considerable damage to assets and people but they had no relevant consequence (Park et al., 2017, Preischl and Hellmich, 2013, Preischl and Hellmich, 2016). This kind of data has the benefit of describing more errors related to hardware (such as manually operated valves) and relating human errors to performance shaping factors. However, near-miss reports are generally restricted to what needs to be communicated to the regulator (Preischl and Hellmich, 2013, Preischl and Hellmich, 2016); thus, relevant factors may not always be reported (Kletz, 2011).

Data from major accident reports have the potential to deliver stronger relation between performance shaping factors and human errors (Moura et al., 2016, Kyriakidis et al., 2015). This is because detailed analyses of the causes that led to the accidents are required and performed (API, 2010). Despite the potential benefits, the strategy of using major accident data to estimate performance shaping factors and human error probabilities is not yet fully explored.

2.2 Bayesian networks

Bayesian network (BN) is a powerful graphical tool that has received an increasing interest due to their capability of providing efficient factorization of joint probability distributions, exploiting information about the conditional dependencies among variables (Tolo et al., 2018). Bayesian networks have also been used for the estimation of Human Error Probability on different industrial sectors, as described by the thorough review of (Mkrtchyan et al., 2015).

Let consider a simplified Bayesian network for modelling human error as shown in Figure 2-2. Each ellipse called “node” represents variables such as “organizational factors,” “technological factors,” “person-related factors,” “cognitive errors,” and “execution errors.” The arrows represent the direction of the causal relationship between variables. In the model presented, the organizational factors are defined as the parent node of cognitive errors and, likewise, cognitive errors as the child node of organizational factors. The organizational factors are denoted a root node of the network, as it does not have parents. The causal relationships

between variables are defined by conditional probability distributions. These distributions are usually represented by crisp values numerically coded in conditional probability tables (CPTs) (Tolo et al., 2015).

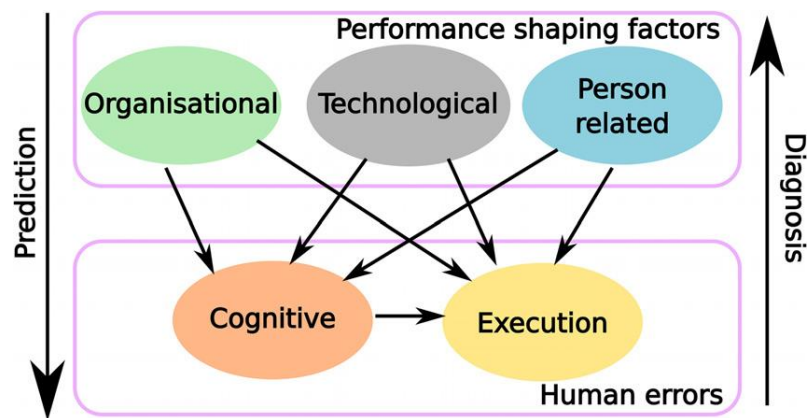


Figure 2-2. Simplified Bayesian network for human error probability

The main advantages of using Bayesian networks for human reliability analysis are as follows (Mkrtchyan et al., 2015):

- Deal with lack or incomplete data of human errors in complex industries by integrating expert judgment and other different sources of information in the model.
- Allow to consider dependencies among factors by using joint probabilities, to combat the frequent (and possibly mistaken) assumption of independencies between performance shaping factors and human errors.
- The acyclic graphs are easy to understand and potentially facilitate the communication between engineers, psychologists, and social scientists in multidisciplinary risk analysis.
- The possibility to update the marginal probabilities of the variables, when new information becomes available.
- Provide reasons for the results by allowing to identify which performance shaping factors are affecting individual human errors (Chen and Pollino, 2012).
- The capability of performing “what if” scenarios analysis by fixing the state of variables, as well as to propagate the information in the direction of interest (Tolo et al., 2017).

2.3 Identifying conditional dependencies from sparse data

Data for human error are usually sparse or missing. Although data can be collected from an increasing number and variability of accident reports (e.g., collecting reports from different safety regulators or from different industry sectors), some conditional dependencies might

continuously fail to appear in the available data. Therefore, inferences of the human error probabilities are generally performed based on expert elicitation. Experts can contribute by providing direct probability values (direct elicitation) or give their opinion through qualitative scales, questionnaires, relative judgments (indirect elicitation) (Mkrtchyan et al., 2015). Alternative approaches are based on data derived from underlying method relationships (Groth and Mosleh, 2012, Yang et al., 2013), or from specifically designed simulators (Groth and Mosleh, 2012, Sundaramurthi and Smidts, 2013). The discussion of the mathematical theory behind these approaches is beyond the scope of the present paper; however, the interested reader can refer to Refs. (Tolo et al., 2017, Nielsen and Jensen, 2009, Fenton and Neil, 2012). Some basic background about conditional probability distributions is provided in the *Appendix A*.

2.4 Verification and validation

Once the human error probabilities are obtained, they should be verified to test if the model works as it is supposed to work (Mkrtchyan et al., 2015). If the correct inputs are given, the appropriate outputs are seen (Kirwan, 1997a). In Jentsch words, we should ask ourselves: “Did we build the system right?” (Jentsch, 1993). Verification can also be referred as “internal validation,” when used as a test to measure the variation between assessors, so the result can be repeated no matter the team or the day when the analysis is conducted (Kirwan, 1997a).

Few published researches based on Bayesian network to infer human error probabilities have presented a verification process (Mkrtchyan et al., 2015). Truco et al. have presented their verification results, after creating a set of hypothetical profiles at the extreme points, varying from the highest to the lowest level of each factor (Trucco et al., 2008). Yang et al. have conducted a sensitivity analysis focused on the “context control modes” of the method CREAM, using expert judgment (Yang et al., 2013). They have suggested that in a successful model a slight change toward the negative effects of a context control mode would result in the increment of the human error probability.

The literature suggests that higher levels of performance shaping factors would result in higher levels of human error probability and that combinations of performance shaping factors would result on greater adverse impact on human error probability (Henderson and Embrey, 2012). That means that human reliability should reflect the features of a coherent system with multistate components, where the performance of a system improves whenever

any component or subset of component improves, and vice-versa (Samaniego, 1985, Barlow and Wu, 1978).

To validate a model, one should test if the system does what is supposed to do in the real world: if the outputs have a good correlation to “real world data” (Kirwan, 1997a). In Jentsch words, we should ask ourselves: “Did we built the right system?” (Jentsch, 1993).

A common method to validate a model is to conduct cross validation, splitting available data sets into training and test sets. However, this approach is adopted in data-rich applications, which are not the case presented in rare events such as human errors in major accidents (Mkrтчyan et al., 2015). For these events, Kirwan suggests the comparison of the new results with existing human error data of better or similar credibility level (Kirwan, 1997a). The measurable criteria used are correlation, accuracy, the degree of optimism/pessimism, and precision (Kirwan, 1997a).

2.4.1 Correlation

The degree of the predictive relationship is usually presented via a scatterplot of predicted versus actual human error probability. Although validations usually try to express parametric correlation (with the square of the correlation coefficient), the majority of validation research conducted by the human reliability community have been expressed via nonparametric correlation (Kirwan, 1997a, Williams, 1988, Kirwan et al., 1997), assuming that human behaviour does not rely on any assumption of the distribution function or the joint distribution of performance shaping factors.

The nonparametric correlation tests are Spearman’s rank correlation coefficient (Pirie, 2004) and Kendal’s coefficient of concordance (Kendal’s s) (Abdi, 2007). Although both tests are different, the interpretation of both coefficients is similar: the correlation between the two variables will be high when observations have a similar (or equal) correlation of one. Likewise, if the coefficient value is next to zero, the correlation between the results from the model and the reference is small.

2.4.2 Accuracy

In risk assessment, an ideal accuracy level is when estimates lie within a factor of three of the “true” values, but it is acceptable if falls within a factor of ten (Kirwan, 1997a). Model accuracies are often represented graphically in a scatterplot of the results against reference data using logarithms scale.

2.4.3 Precision

An aspect of precision is the degree to which the estimate or technique, when not accurate, is pessimistic rather than optimistic (Kirwan, 1997a). A pessimistic estimate is a prediction whose bias makes the value worse (riskier, costlier, etc.) which makes the estimate more conservative. Conservative estimates lead to safer but at the same time more expensive designs. Therefore, it is important to find strategies that provide more realistic HEPs for the industry. Histograms are also plotted to present how human error estimates were distributed into accuracy bands within pessimistic and optimistic factors of 3, 10, and 100.

3 Proposed approach: using datasets of major accidents reports

3.1 Bayesian network definition

All the steps required to build a Bayesian network from major accident reports are described in the following:

Definition of the nodes: Bayesian network nodes represent the variables obtained from any taxonomy able to classify performance shaping factors and human errors. The chosen taxonomy must be able to classify the performance shaping factors and human errors at a level that is common for all the sectors.

States of the nodes: Root nodes have only two states: the state “0” and state “1” representing the logical entries of the accident dataset during data collection, i.e., 0 when a variable (e.g., performance shaping factor or human error) is absent or not observed on the accident by the investigator, and 1 when the variable has been observed.

Child nodes have been augmented with an additional state called “no data.” This state is used to handle cases where specific combinations of events (i.e., the conditional probabilities) are not observed in the dataset. This strategy not only permits the assessment of the conditional probability tables without expert judgment but also increases the transparency on the uncertainties of the result (i.e., human error probability).

Definition of the structure: The Bayesian network structure (Figure 2-3) has the objective of capturing the dependencies between performance shaping factors and human errors but also among performance shaping factors and human errors, and explicitly modelling their multilevel, hierarchical influences on each other. Experts with psychology and sociology knowledge might be elicited to obtain this type of structure (e.g., to identify the causal relationships of cognitive errors and organizational factors). Although one of the aims of this

research was to avoid expert biases, it is acknowledgeable that at some level of the assessment the experts are essential – if not for eliciting the prior probabilities, they will be for the model structure or for the taxonomy used.

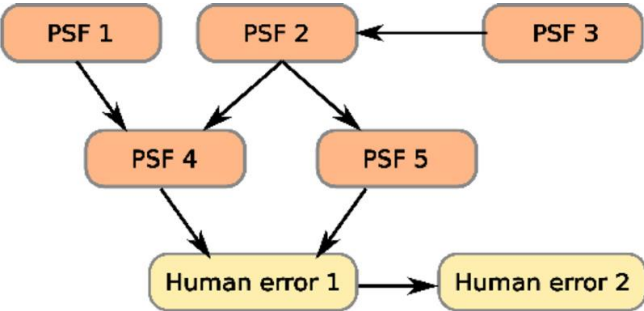


Figure 2-3. Example of a structure reflecting the causal relationships within variables

3.2 Assessment of the conditional probability tables

In order to avoid experts’ biases on eliciting probabilities, the present work uses solely the information from dataset in order to obtain the conditional probability distributions. Let consider a dataset from accident reports able to classify human errors and corresponding performance shaping factors (PSFs) as shown in Table 2-1. Conditional probability tables for root nodes are defined as the frequencies for each performance shaping factor obtained in the data collection and presented in

Table 2-2 and Table 2-3.

Table 2-1. Example of a dataset with human errors and PSFs identified for each accident

Accident	Human error 1	Human error 2	PSF 1	PSF 2	PSF 3	PSF 4	PSF 5
Accident #1	1	0	0	1	1	1	0
Accident #2	0	1	0	0	0	0	0
Accident #3	0	0	0	0	0	0	0
Accident #4	1	0	0	0	0	0	0
Accident #5	1	0	0	1	1	1	0
Accident #6	1	1	0	0	0	0	0
Accident #7	1	0	0	1	0	1	0
Accident #8	1	0	0	0	0	0	0
Accident #9	1	0	0	0	0	1	1

Table 2-2. Example of prior probabilities of root node PSF 1

PSF 1	
State 0	1
State 1	0

Table 2-3. Example of prior probabilities of root node PSF 2

PSF 2	
State 0	$6/9 = 0.7$
State 1	$3/9 = 0.3$

The conditional probabilities for child nodes depend on the structure defined on the network and are obtained by counting the frequency of all the possible combinations of the parent node's states in the dataset. The frequency is the number of accidents that present a specific combination divided by the number of accidents in the dataset. The frequencies obtained (Table 2-4) are then normalized, as the prior probabilities of the set of states of the child node must sum to one (Table 2-5). The same process is repeated for each combination of the conditional probability table. When this process is complete it is possible to compute the posterior probabilities for each node. The posterior probabilities of the state 1 of the child nodes designated to human errors will be the human error probabilities.

Table 2-4. Example of the CPT for node 'human error 1'

PSF 1		State 0				(...)
PSF 2		State 0				(...)
PSF 3		State 0				(...)
PSF 4		State 0		State 1		(...)
PSF 5		State 0	State 1	State 0	State 1	(...)
Human error 1	State 0	$2/9 = 0.2$	0	0	0	(...)
	State 1	$3/9 = 0.3$	0	0	$1/9 = 0.1$	(...)
	No data	0	1	1	0	(...)

The boldface value '1' is added by the analyst when there is no information about certain combination from the data.

Table 2-5. Normalized CPT

PSF 1		State 0				(...)
PSF 2		State 0				(...)
PSF 3		State 0				(...)
PSF 4		State 0		State 1		(...)
PSF 5		State 0	State 1	State 0	State 1	(...)
Human error 1	State 0	0.4	0	0	0	(...)
	State 1	0.6	0	0	1	(...)
	No data	0	1	1	0	(...)

When the dataset used does not provide information for defining conditional distributions within certain variables states, the variable state no data is set to 1. If this strategy were not used, the prior probabilities of states 0 and state 1 of the child node for that given combination would have both probabilities set equal to zero, making it impossible to compute the conditional probability table. In Ref. (Fenton and Neil, 2012), it is suggested to assigning equal probability to all the unknown combination of events. However, using the latter strategy, a researcher loses the information of what combinations do not lead to human errors according to the dataset, which can be potentially used in the future.

3.3 Validation and verification

The verification of the models is performed through what-if analysis, to test how the model behaved when analysing well-known scenarios (Tolo et al., 2017). To achieve that, some hypothetical scenarios have been created by fixing each state of each performance shaping factor node of the Bayesian network, and observing how the changes affected the human error probabilities.

Results from the what-if analysis are used to verify the model but also to obtain the maximum and minimum bounds of human error probabilities after varying each performance shaping factor to its maximum and minimum values. The validation process is performed by comparing the results obtained by the constructed model against data provided by references using the same taxonomy.

4 Case study

4.1 MATA-D dataset

For the present research, the multi-attribute technological accidents dataset (MATA-D) is adopted (Moura et al., 2016). The dataset contains 238 major accident reports classified under the CREAM taxonomy (Hollnagel, 1998). A single taxonomy is used to describe both human errors and performance shaping factors for a variety of industrial sectors. Only trusted investigation boards have been used to build the dataset. Logical values, i.e., binary code of 1s or 0s, are used to designate whether or not a human error or factor was observed. This resulted in a matrix of zeros and ones with 238 rows (the number of accidents) by 53 columns formed by 39 performance shaping factors (Table 2-6) and 14 different types of human errors (Table 2-7).

Table 2-6. Performance shaping factors used in MATA-D dataset

Organisational Factors	Technological Factors	Person related factors
Communication failure	Equipment failure	Permanent related
Missing information	Software fault	Functional impairment
Maintenance failure	Inadequate procedure	Cognitive style
Inadequate quality control	Access limitations	Cognitive bias
Management problem	Ambiguous information	Temporary
Design failure	Incomplete information	Temporary related
Inadequate task allocation	Access problems	Memory failure
Social pressure	Mislabelling	Fear
Insufficient skills		Distraction
Insufficient knowledge		Fatigue
Adverse ambient conditions		Performance Variability
Excessive demand		Inattention
Inadequate work place layout		Physiological stress
Irregular working hours		Psychological stress

Table 2-7. Human erroneous actions and cognitive functions used in the MATA-D dataset

Cognitive function failures⁵			Execution Errors
Observation errors	Interpretation errors	Planning errors	Wrong time
Observation missed	Faulty diagnosis	Inadequate plan	Wrong type
False Observation	Wrong reasoning	Priority error	Wrong Object
Wrong Identification	Decision error		Wrong place
	Delayed interpretation		
	Incorrect prediction		

4.2 Bayesian network

The methodology presented in Section 3 has been used to construct a Bayesian Network model from the MATA-D dataset and summarized in *Table 2-8*. The resulting structure of the Bayesian network is shown in *Figure 2-4*.

⁵ In the original research article, table 2.7 has been named as *Human errors used in the MATA-D dataset*

Table 2-8. Summary of the methodology to build the Bayesian Network model

Nodes and states	Structure	Conditional probability table	Verification and Validation
<p>The nodes are variables defined in CREAM taxonomy (Hollnagel, 1998).</p>	<p>The connections between the nodes were defined according to relations based on expert judgement, from the same author of the taxonomy used to define the nodes (Hollnagel, 1998). He has named it the ‘antecedent-consequence relation’.</p>	<p>The conditional probability tables for the root nodes were obtained directly from the frequencies of each performance shaping factor according to (Moura et al., 2016), e.g. design failure is equal to 66%, so at the conditional probability table the state ‘1’ of the root node ‘design failure’ is 0.66 and the state ‘0’ is the complement to one: 0.34.</p>	<p>To verify any incoherence in the model, a what-if analysis was conducted by fixing the states of the variables.</p>
<p>From 39 possible performance shaping factors and 14 possible human errors, only six were not used, due to their absence on the accident reports.</p>	<p>A different structure less reliant on expert judgement was proposed at (Morais et al., 2018), by using common patterns of PSFs and human errors identified on (Moura et al., 2017b).</p>	<p>The frequency for combinations between factors and errors for the child nodes have been extracted from the dataset entries.</p>	<p>To validate the model, the estimates were tested against reference data published on (Hollnagel, 1998) according to correlation, accuracy and precision.</p>
<p>The root nodes have two states: ‘0’ and ‘1’ (following the logical entries of the MATA-D dataset) and child nodes have ‘0’, ‘1’ and ‘no data’.</p>	<p>The structure depicts the influence between performance shaping factors, which means that some performance shaping factors are also child nodes.</p>	<p>Due to the high number of combinations between the states of the parent nodes that a child node has reached, obtaining the frequencies per combination from the dataset was not a trivial task. A code was used to read the table and extract the probability for each combination. For more information on the code and on how to use it, please contact the authors.</p>	<p>To verify any incoherence in the model, a what-if analysis is conducted by fixing the states of the variables.</p>
<p>The root nodes have two states: ‘0’ and ‘1’, to designate whether or not an evidence was encountered on an accident report.</p>	<p>The structure represents the influence that performance shaping factors have upon each other.</p>	<p>The conditional probability tables for the root nodes are obtained directly from the frequencies of each performance shaping factor according to the dataset.</p>	<p>To validate the model, the estimates are tested against reference data according to correlation, accuracy and precision. If possible, the reference data should be obtained from operational experience.</p>
<p>The child nodes have the states ‘0’, ‘1’ and ‘no data’. The latter state is used when the dataset does not provide a specific combination between the parent nodes.</p>	<p>Eventually, this means that some performance shaping factors are also child nodes.</p>	<p>The frequency for combinations between factors and errors are obtained also from the dataset inputs for each accident.</p>	
	<p>All human errors are child nodes of the performance shaping factors.</p>		

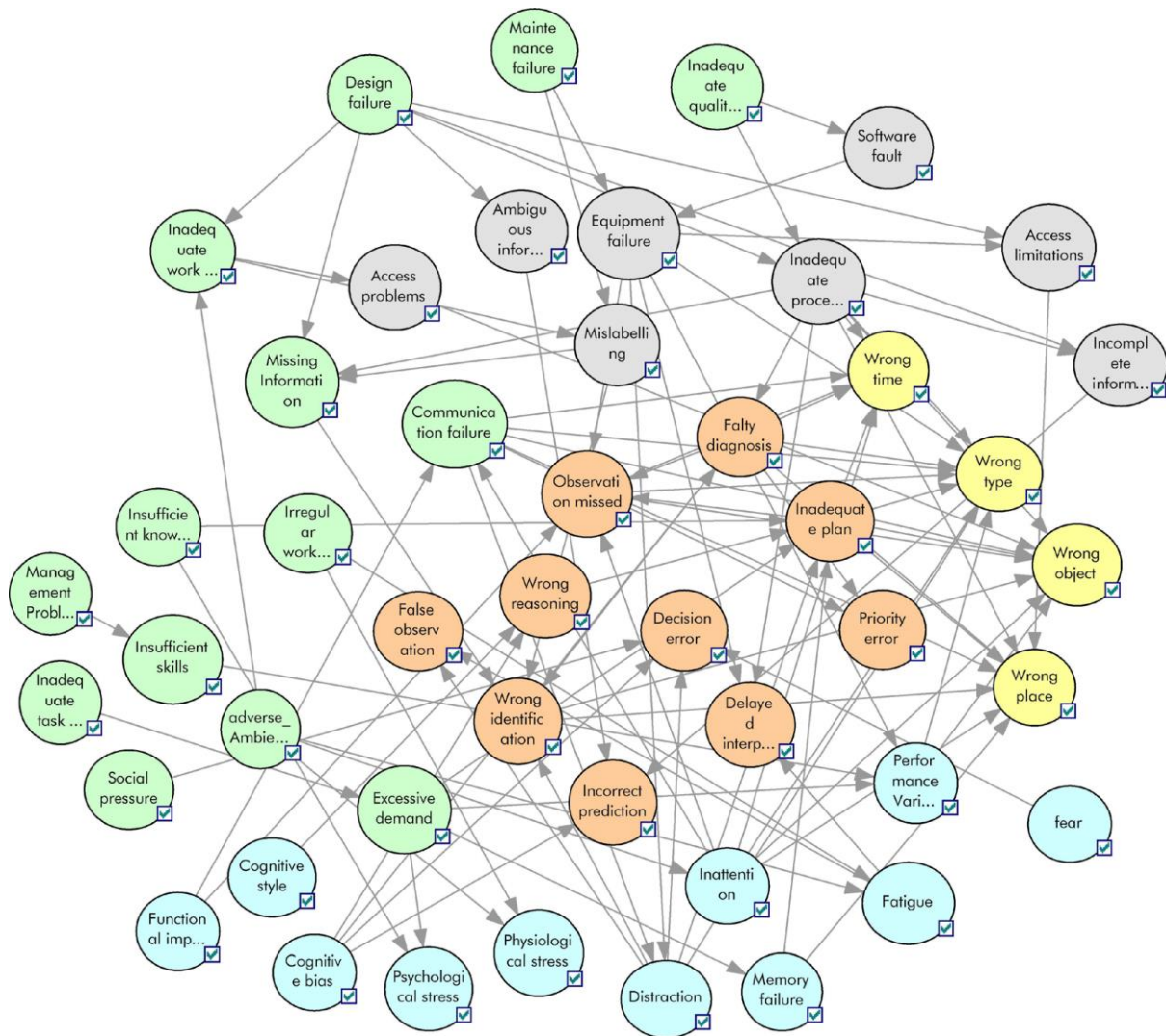


Figure 2-4. Model for predicting human error probabilities adapted from Ref. (Hollnagel, 1998)

4.3 Human error probabilities

The HEP obtained analysing the MATA-D dataset are presented in Table 2-9 and graphically represented in a scatter plot in Figure 2-5. The state 0 indicates the probability of a specific human error not being triggered by a specific combination of performance shaping factors. The state no data indicates the number of times a combination of those factors has not occurred in the dataset.

Table 2-9. HEPs from model compared with data reference (Hollnagel, 1998)

	Human cognitive and execution errors	Lower bound from reference	Basic value from reference	Upper bound from reference	Human error probability
Observation	Observation missed	2.00×10^{-2}	7.00×10^{-2}	$*1.70 \times 10^{-1}$	1.57×10^{-1}
	False Observation	3.00×10^{-4}	1.00×10^{-3}	3.00×10^{-3}	3.54×10^{-2}
	Wrong Identification	2.00×10^{-2}	7.00×10^{-2}	$*1.70 \times 10^{-1}$	1.54×10^{-2}
Interpretation	Faulty diagnosis	9.00×10^{-2}	2.00×10^{-1}	6.00×10^{-1}	1.30×10^{-1}
	Wrong reasoning	Not provided	Not provided	Not provided	1.13×10^{-1}
	Decision error	1.00×10^{-3}	1.00×10^{-2}	1.00×10^{-1}	9.14×10^{-2}
	Delayed interpretation	1.00×10^{-3}	1.00×10^{-2}	1.00×10^{-1}	5.19×10^{-2}
	Incorrect prediction	Not provided	Not provided	Not provided	3.90×10^{-2}
Planning	Inadequate plan	1.00×10^{-3}	1.00×10^{-2}	1.00×10^{-1}	9.89×10^{-2}
	Priority error	1.00×10^{-3}	1.00×10^{-2}	1.00×10^{-1}	6.55×10^{-2}
Execution	Action at wrong time	1.00×10^{-3}	3.00×10^{-3}	9.00×10^{-3}	1.24×10^{-1}
	Action of wrong type	1.00×10^{-3}	3.00×10^{-3}	9.00×10^{-3}	1.02×10^{-1}
	Action on wrong object	5.00×10^{-5}	5.00×10^{-4}	5.00×10^{-3}	2.34×10^{-2}
	Action of wrong place	1.00×10^{-3}	3.00×10^{-3}	9.00×10^{-3}	3.01×10^{-1}

* The literature provides 1.7×10^{-2} . However, this value is lower than the lower bound. In this paper, the authors decided to replace this value to 1.7×10^{-1} .

For the purpose of verification, the obtained probabilities have been compared against data from Ref. (Hollnagel, 1998). The interval of the reference is described by the lower and upper bounds for each human error.

Figure 2-5 shows higher human error probabilities than the reference data. A possible interpretation of this trend might be attributed to the methods used to collect reference data (Hollnagel, 1998), where all human errors were accounted for, including those that have not produced an accident. Thus, more opportunities of errors were accounted on the denominator of Equation 2-1, making the resulting probabilities lower than those obtained with the present approach. The human error estimates are the values obtained for the probabilities of the state 1 of each child node. The results of state 0 and the state no data are presented in Table 2-10. A comparison of the results obtained for each state is also presented in Figure 2-6.

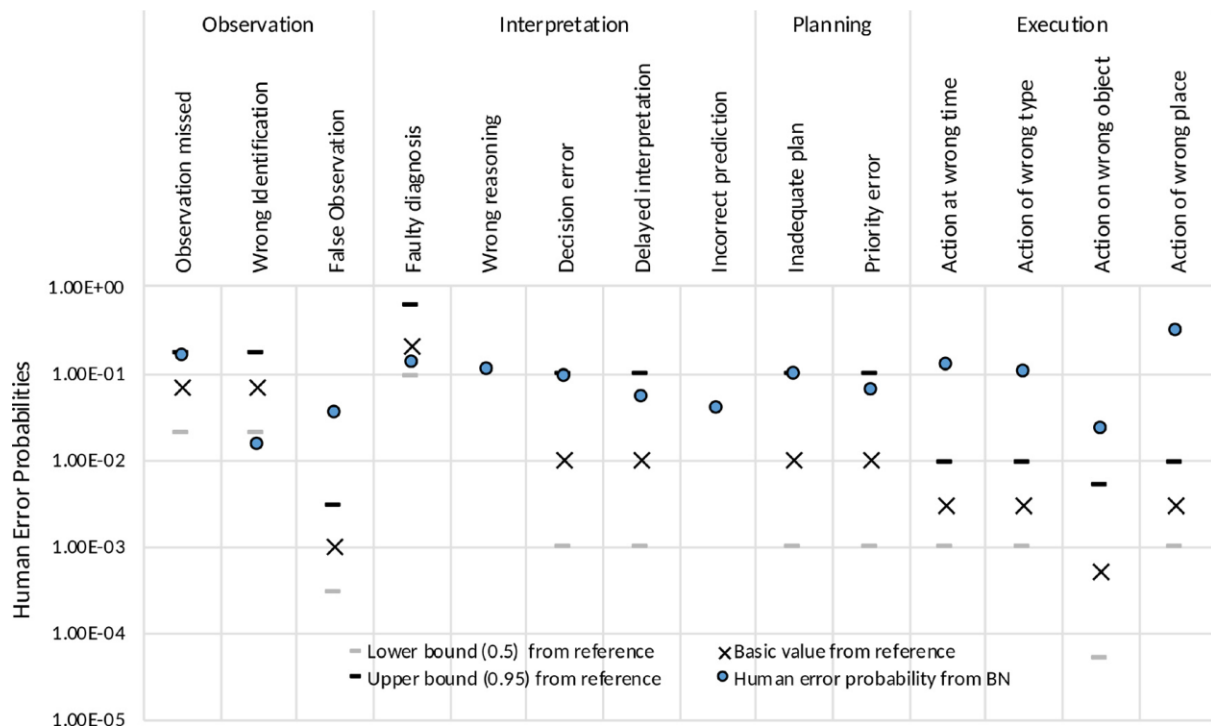


Figure 2-5. Human error probabilities from the proposed approach and from Ref. (Hollnagel, 1998) plotted in a logarithmic scale

Table 2-10. Results of all states of human error probability nodes on the proposed model

Cognitive and execution errors	State '0'	State '1' (HEP)	State 'no data'
Observation missed	8.22×10^{-1}	1.57×10^{-1}	2.07×10^{-2}
Wrong Identification	9.58×10^{-1}	3.54×10^{-2}	6.62×10^{-3}
False Observation	9.71×10^{-1}	1.54×10^{-2}	1.38×10^{-2}
Faulty diagnosis	8.70×10^{-1}	1.30×10^{-1}	0.00
Wrong reasoning	8.87×10^{-1}	1.13×10^{-1}	0.00
Decision error	8.96×10^{-1}	9.14×10^{-2}	1.24×10^{-2}
Delayed interpretation	9.45×10^{-1}	5.19×10^{-2}	2.71×10^{-3}
Incorrect prediction	9.61×10^{-1}	3.90×10^{-2}	0.00
Inadequate plan	8.85×10^{-1}	9.89×10^{-2}	1.65×10^{-2}
Priority error	9.31×10^{-1}	6.55×10^{-2}	3.92×10^{-3}
Action at wrong time	8.27×10^{-1}	1.24×10^{-1}	4.89×10^{-2}
Action of wrong type	7.68×10^{-1}	1.02×10^{-1}	1.30×10^{-1}
Action on wrong object	9.05×10^{-1}	2.34×10^{-2}	7.16×10^{-2}
Action of wrong place	6.49×10^{-1}	3.01×10^{-1}	5.06×10^{-2}

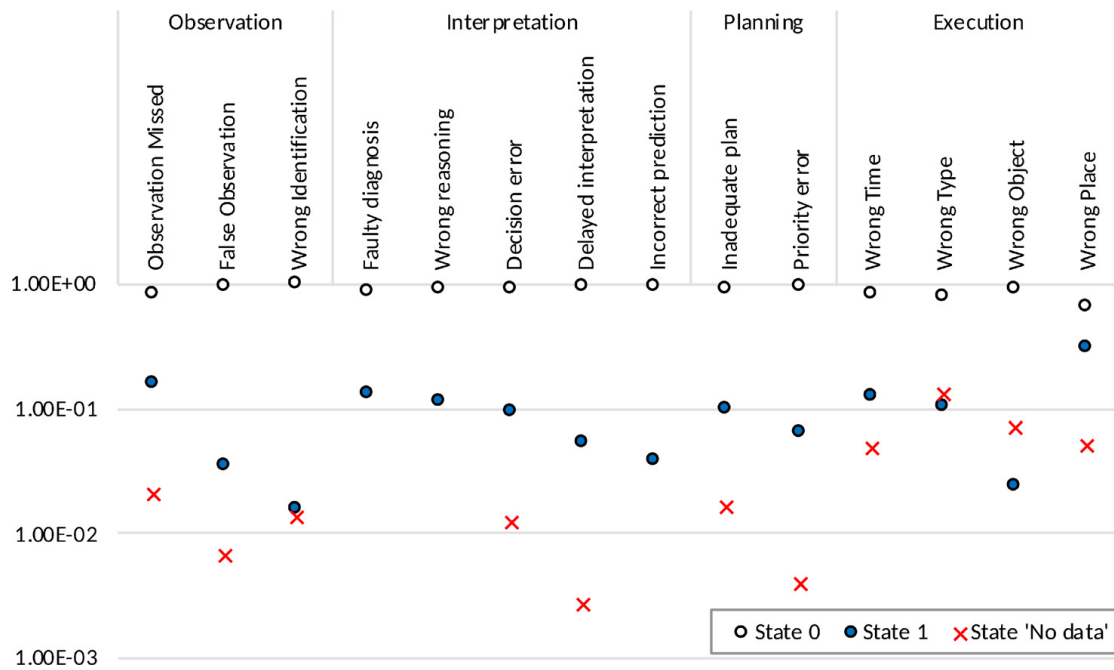
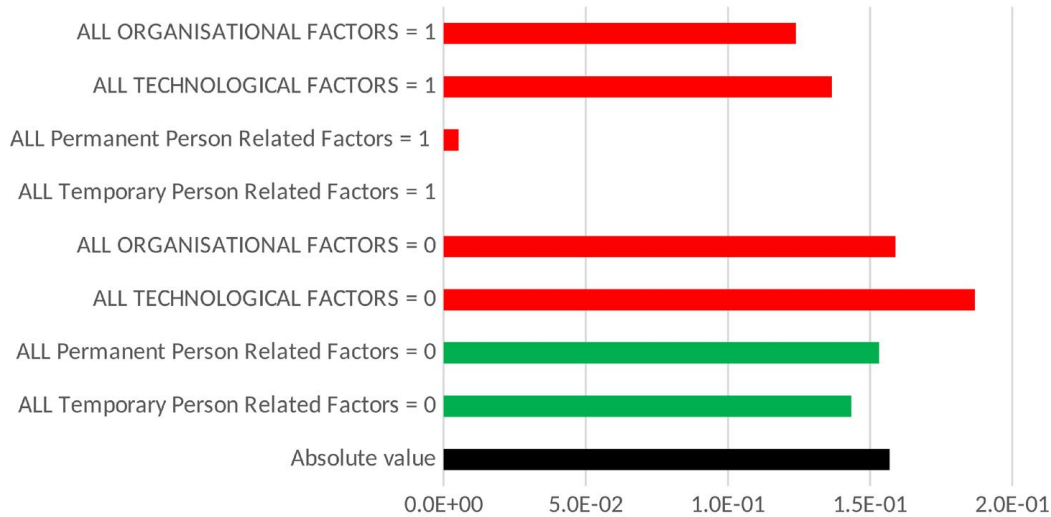


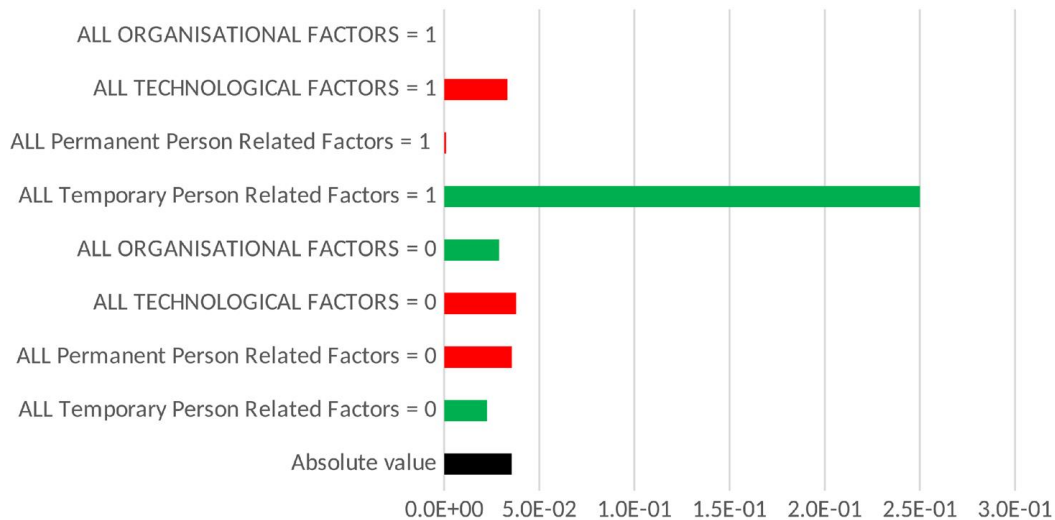
Figure 2-6. States estimates for the proposed model

To test if the model outputs work as they were supposed to work, a what-if analysis was conducted, by fixing the states of sets of performance shaping factors one-at-the-time. In Figure 2-7, the black bars represent the values estimated for the model; the green and red bars can be interpreted as a spectrum of human error probabilities after varying all performance shaping factors to their best and worst-case scenarios. The green bars represent the expected results for a specific variation, whereas the red bars represent the unexpected results. The expected results represent those values that are expected from a coherent system according to the formal definition used for reliability technological systems. For instance, in a coherent system, the probability of having a human error decreases if a set of performance shaping factors are set to zero (best-case scenario) and increases in case of performance shaping factors increased to 100% (worst-case scenario). The obtained figures show that in the scenario of having all the organizational factors failing to work, the cognitive error of missing an observation (i.e., “observation missed”) would in fact decrease. This is possibly be explained by an increase in performance that humans might be using to compensate organizational errors. This reinforces the theory that humans are not only probable initiators of an event but also the last chance to recover a problem initiated by organizational and technological factors (Hollnagel, 1998).

(a) **Probability of Observation missed**



(b) **Probability of False Observation**



(c) **Probability of Wrong Identification**

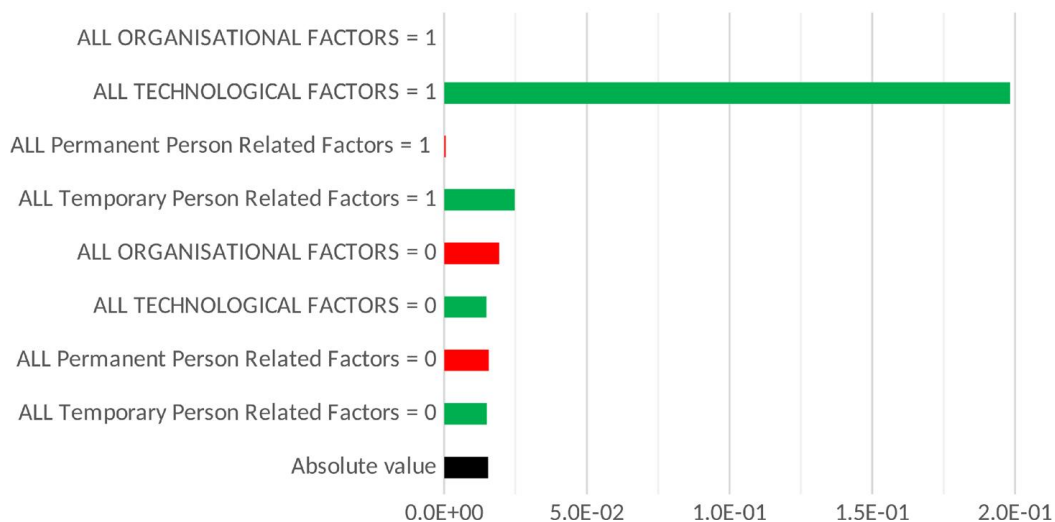
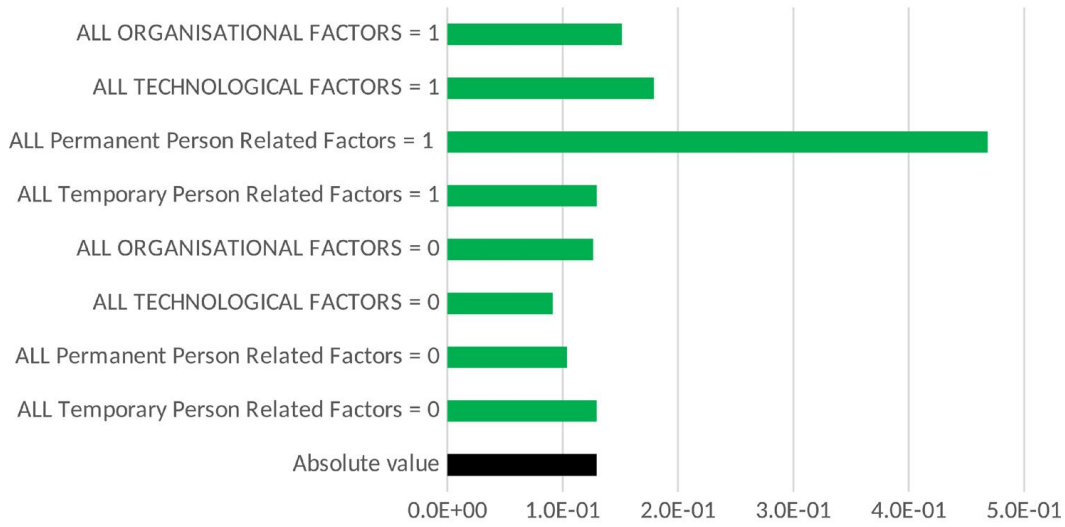
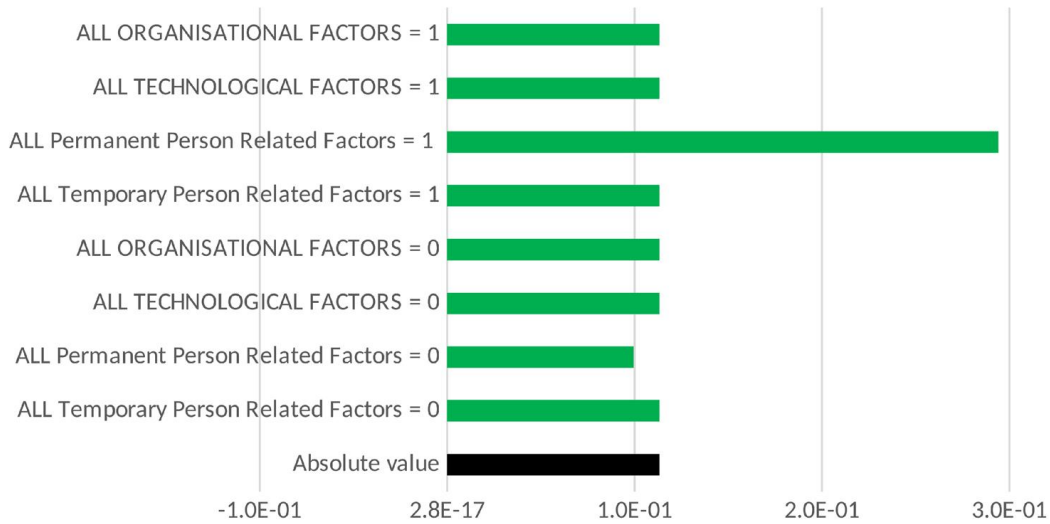


Figure 2-7. (a)–(n) Human error probabilities estimated by fixing the performance shaping factors one-at-the-time

(d) Probability of Faulty diagnosis



(e) Probability of Wrong reasoning



(f) Probability of Decision error

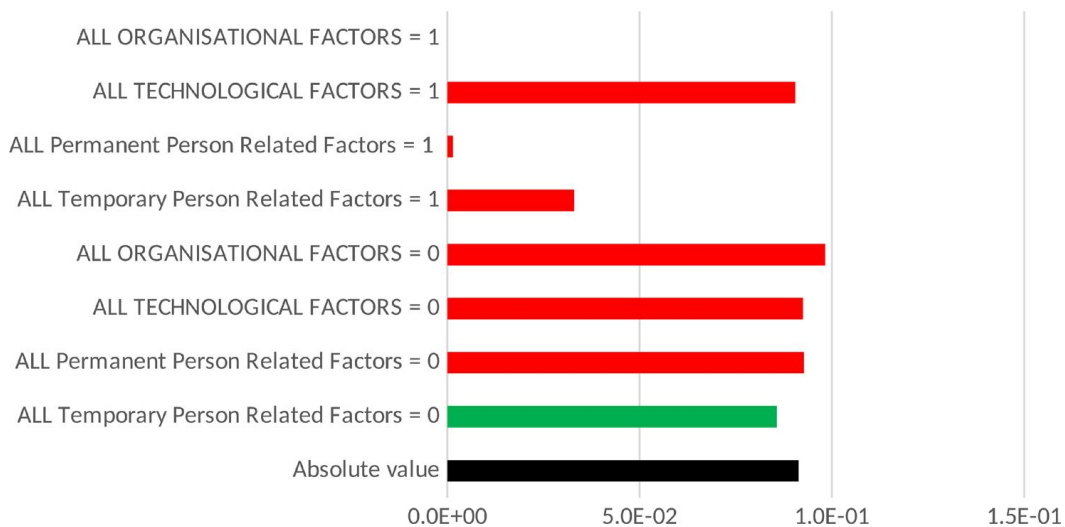
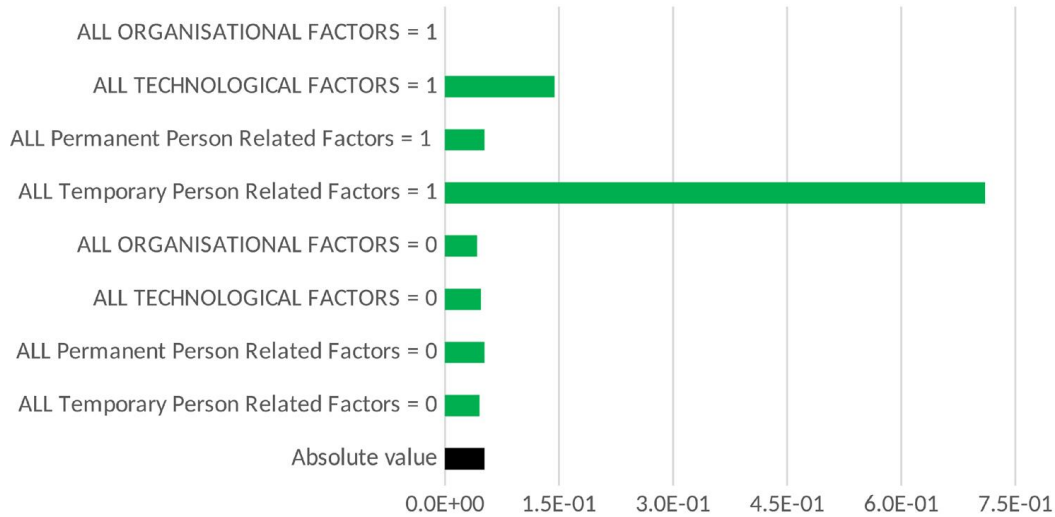
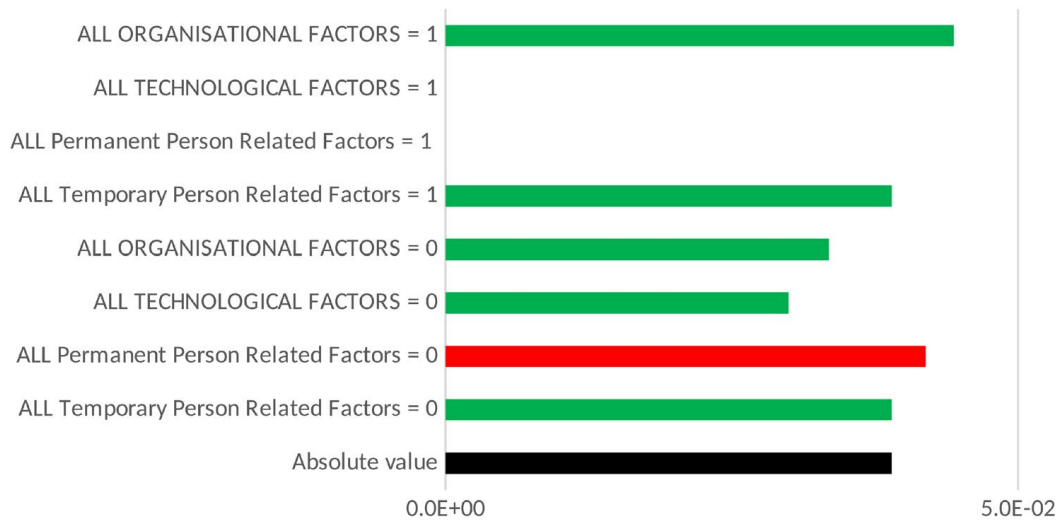


Figure 2.7 (continued)

(g) Probability of Delayed interpretation



(h) Probability of Incorrect prediction



(i) Probability of Inadequate plan

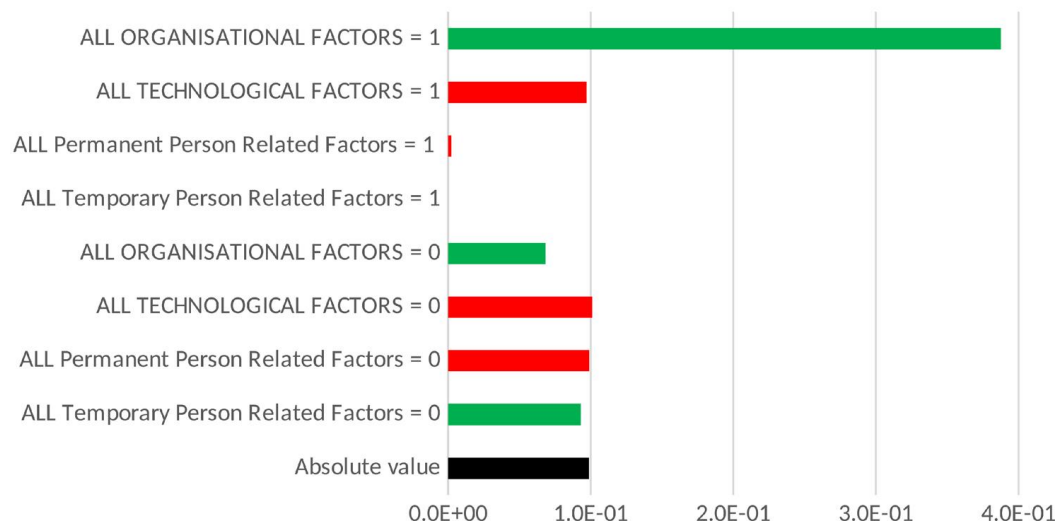
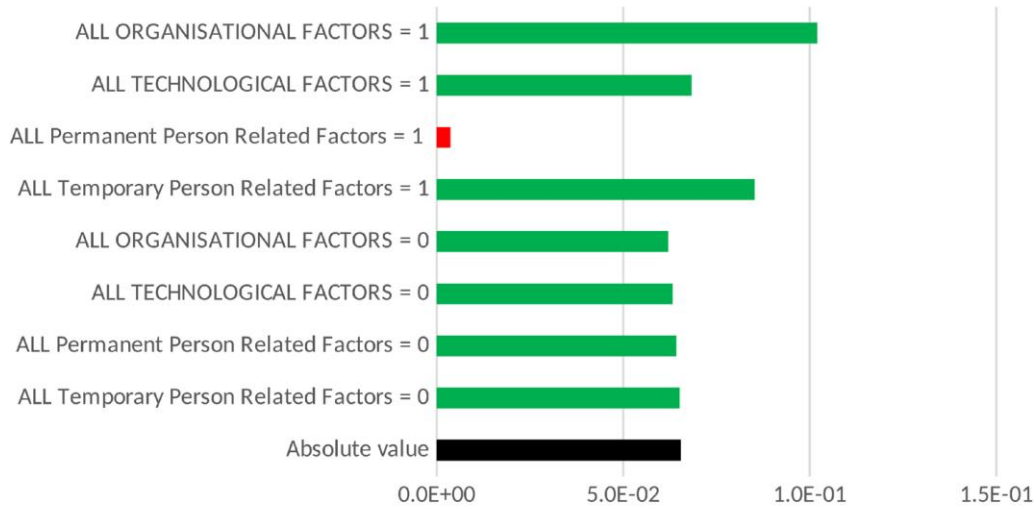
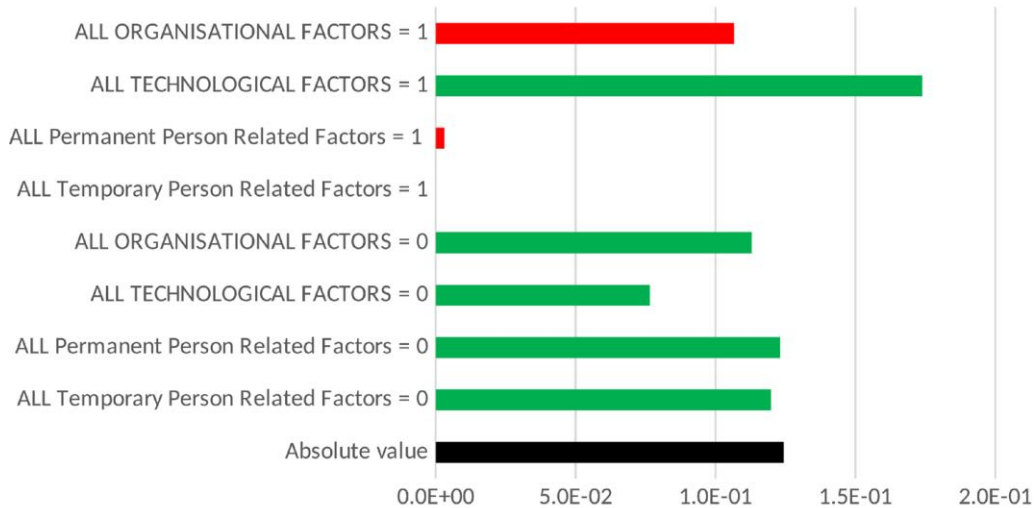


Figure 2.7 (continued)

(j) Probability of Priority error



(k) Probability of Wrong Time



(l) Probability of Wrong Type

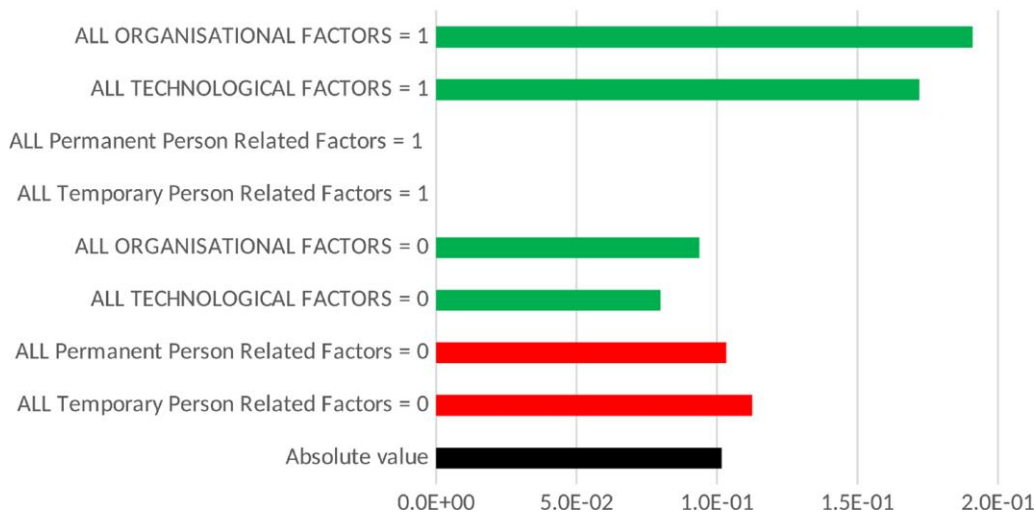
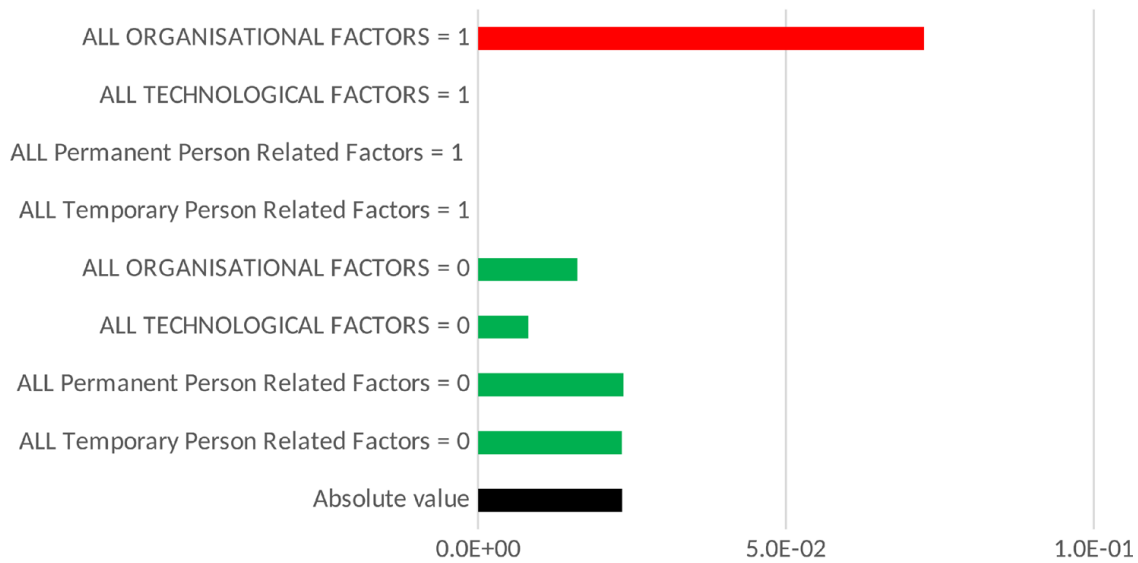


Figure 2.7 (continued)

(m) Probability of Wrong Object



(n) Probability of Wrong Place

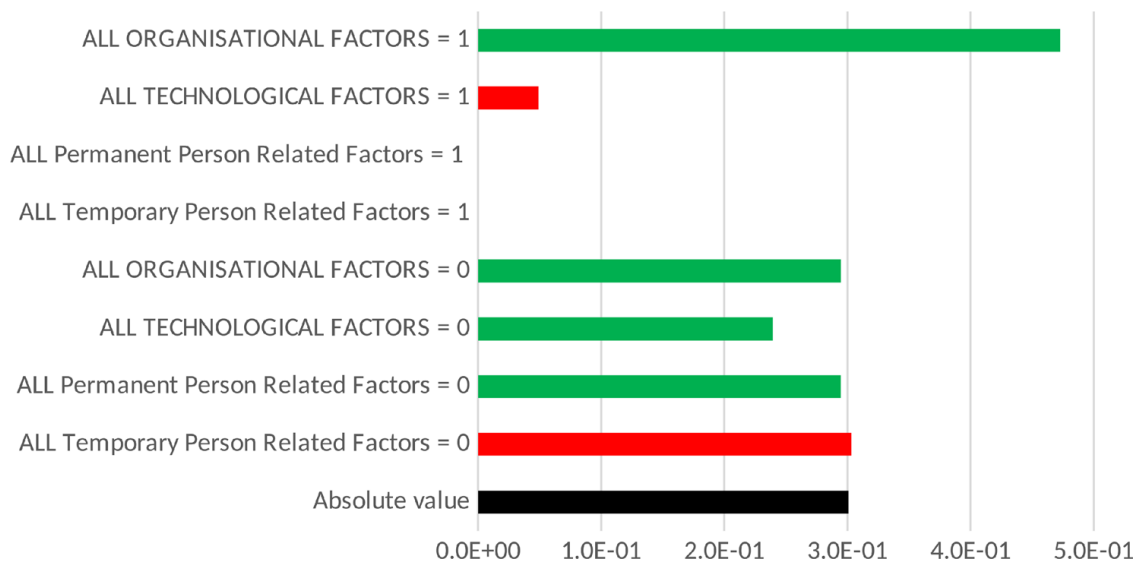


Figure 2.7 (continued)

Variations of some sets of performance shaping factors also resulted in zero probability human errors, as presented in Table 2-11. This shows that some human errors are impossible to occur under the specific conditions of performance shaping factors present in the MATA-D database.

Table 2-11. Sets of performance shaping factors variations producing zero human errors probability

Human error Probability = 0	Simulated Scenarios (sets of PSFs at their worst-case scenarios)
Observation missed	When All Temporary Person Related Factors = 1
False Observation	All organisational factors = 1
Wrong Identification	Functional impairment (a permanent person related factor) = 1 All organisational factors = 1 Missing information (an organisational factor) = 1
Faulty diagnosis	--
Wrong reasoning	--
Decision error	All organisational factors = 1 Social pressure (an organisational factor) = 1
Delayed interpretation	All organisational factors = 1
Incorrect prediction	All Permanent Person Related Factors = 1 Cognitive bias (a permanent person related factor) = 1 All technological factors = 1 Ambiguous information (a technological factor) = 1
Inadequate plan	ALL Temporary Person Related Factors = 1 Memory failure (a Temporary Person Related Factor) = 1
Priority error	--
Wrong time	ALL Temporary Person Related Factors = 1
Wrong type	ALL Temporary Person Related Factors = 1 Performance Variability (a Temporary Person Related Factor)= 1 ALL Permanent Person Related Factors = 1 Functional impairment (a Permanent Person Related Factor) = 1
Wrong Object	ALL Temporary Person Related Factors = 1 Inattention (a Temporary Person Related Factor) = 1 ALL Permanent Person Related Factors = 1 Functional impairment (a Permanent Person Related Factor) = 1 All technological factors = 1 Access problems (a technological factor) = 1
Wrong place	ALL Temporary Person Related Factors = 1 ALL Permanent Person Related Factors = 1

To validate the model, its outputs had been tested on the correlation, accuracy and precision to existing data obtained at (Hollnagel, 1998). The reference data were collected from simulators, expert judgment, laboratory controlled cognitive experiments, and simulation studies of inspection tasks (from simulated process plant and training schools). According to Hollnagel (Hollnagel, 1998), data sources for human errors such as observation and execution were relatively well established at the time they were collected (approximately 1998). On the other hand, the author declared that interpretation and planning behaviours were mostly based on expert judgments. In addition, Ref. (Hollnagel, 1998) does not provide probabilities for “wrong reasoning” and “incorrect prediction.” To validate the model only the basic values provided in Ref. (Hollnagel, 1998) are used. Figure 2-8 shows a scatter plot of human error probability predicted from the model versus human error probability from the Ref. (Hollnagel, 1998).

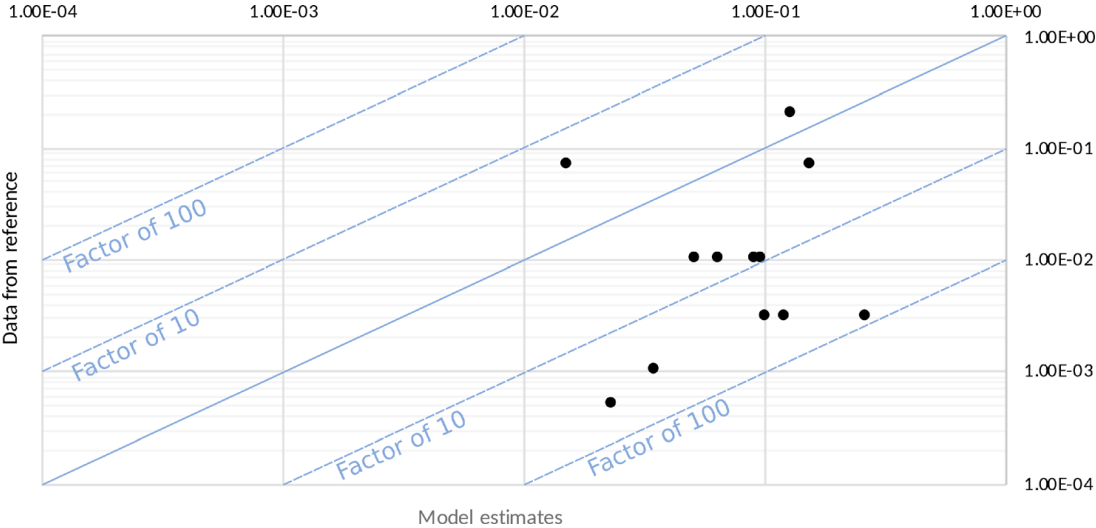


Figure 2-8. Human error probabilities (HEPs) from model versus HEPs from the reference in a logarithmic scale

The present research has also tested nonparametric correlation, as human behaviour does not rely on any assumptions on the distribution function. The nonparametric correlation tests of Spearman’s correlation coefficient and Kendal’s coefficient of concordance are both presented in Table 2-12. Both correlation coefficients are very small and not statistically significant. As shown on the scatterplot in Figure 2-8, seven of the human error probabilities estimated lied within a factor of 10 and five within a factor of 100 of the reference.

Table 2-12. Nonparametric correlation results

Spearman’s correlation coefficient (ρ_s)= 0.20115
--

Correlation between model outputs and values in the reference

Kendal's coefficient of concordance (Kendal's τ) = 0.3333

To evaluate their accuracy within a factor of 3, the results were also plotted in a histogram (Figure 2-9). When not accurate, the histograms also illustrate if the estimates are pessimistic or optimistic if compared to the reference. The model outputs had presented more pessimistic estimates rather than optimistic ones, meaning that the majority of HEPs estimated through both models tend to be higher than the reference. The histogram provided in Figure 2-9 shows how spread the results are.

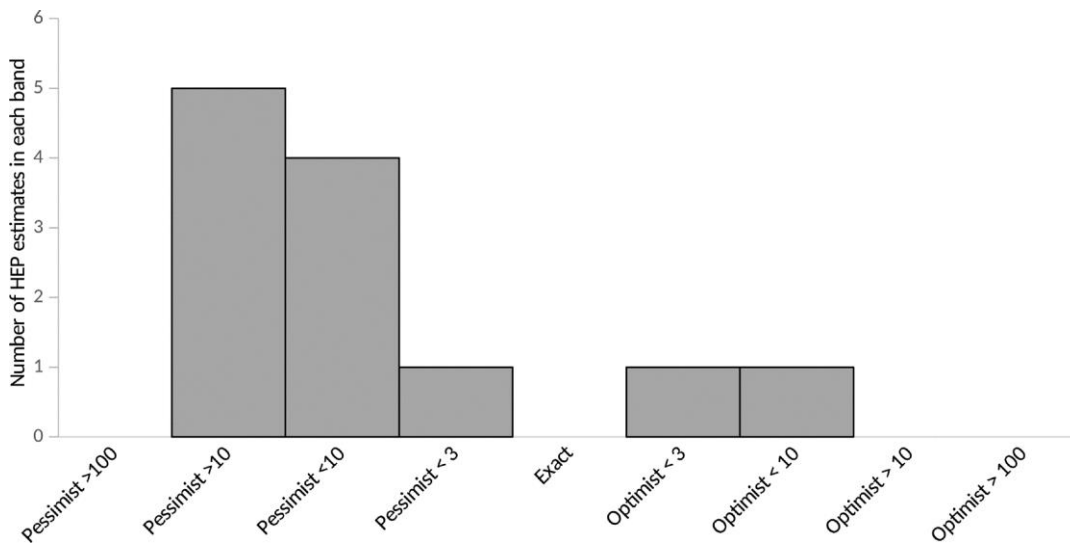


Figure 2-9. Histogram with accuracy bands

Table 2-13 presents the lower and upper bounds of human error probabilities after varying all performance shaping factors to their minimum and maximum values.

Table 2-13. Human error probability uncertainty after varying performance shaping factors

Cognitive and execution errors	Lower bound	Human error probability	Upper bound
Observation missed	5.30×10^{-3}	1.57×10^{-1}	7.75×10^{-1}
False Observation	1.00×10^{-3}	3.54×10^{-2}	3.27×10^{-1}
Wrong Identification	5.00×10^{-4}	1.54×10^{-2}	1.98×10^{-1}
Faulty diagnosis	9.15×10^{-2}	1.30×10^{-1}	4.69×10^{-1}
Wrong reasoning	9.95×10^{-2}	1.13×10^{-1}	2.94×10^{-1}
Decision error	1.40×10^{-3}	9.14×10^{-2}	2.72×10^{-1}
Delayed interpretation	2.10×10^{-2}	5.19×10^{-2}	7.10×10^{-1}

Incorrect prediction	2.30×10^{-3}	3.90×10^{-2}	8.49×10^{-2}
Inadequate plan	2.20×10^{-3}	9.89×10^{-2}	3.88×10^{-1}
Priority error	2.03×10^{-3}	6.55×10^{-2}	1.02×10^{-1}
Action at wrong time	3.20×10^{-3}	1.24×10^{-1}	3.52×10^{-1}
Action of wrong type	1.00×10^{-4}	1.02×10^{-1}	1.91×10^{-1}
Wrong Object	7.10×10^{-3}	2.34×10^{-2}	7.65×10^{-2}
Action of wrong place	1.30×10^{-6}	3.01×10^{-1}	4.73×10^{-1}

4.4 Discussion

The case study shows the applicability of the approach for the available datasets of major accidents. These databases are capable to describe the interaction between human, machine and organizational systems and that the human error probabilities obtained have a similar order of magnitude of those used by industry to feed real risk assessments. However, some aspects brought by the verification and validation steps have to be better understood before considering the probabilities ready to be used to feed risk assessments.

The verification applied to the case study shows some human errors increasing if one or a set of performance shaping factor are decreased (and vice-versa). This may suggest an inadequacy of the used model or may also indicate that complex socio-technical systems do not necessarily behave as a coherent system. If right, the results of the case study suggest that some degraded performance shaping factors (or the combination of them) may cause also positive effects on human behaviour. Similar behaviour has been described by psychology research, which described that vigilance (the ability to maintain concentrated attention over prolonged periods of time) can actually decrease due to low levels of workload, an organizational shaping factor (CA Authority, 2016). The verification step also has demonstrated that some human errors are unlikely to happen for specific states of performance shaping factors, as can be observed from some null human error probabilities.

The validation step has exposed a low correlation between the results obtained with the Bayesian network and the reference, as the model do not provide a predictive relationship with data from the reference used (Hollnagel, 1998). However, a new validation process must be conducted with data with similar source quality as the dataset (i.e., operational experience), as the data used as reference was partially obtained from simulators and expert elicitation. The human error probabilities obtained from the model tend to be higher than the reference, meaning that if they are used to feed risk assessments they will lead to a safer design. The majority of

results falls within a factor of 3 and 10 than within a factor of 100, which is normally acceptable to feed risk assessments. This validation aspect is important to develop because although higher than the real probabilities lead to safer design, they are not desirable as it can direct resources to the wrong risks.

The what-if analysis undergone in the verification and validation steps has also provided a spectrum of human error probability variations that can be seen as the uncertainty of estimates from different scenarios. In other words, varying the performance shaping factors in the Bayesian networks provides a distribution of human error probabilities, where uncertainty boundaries can be obtained. To better capture the uncertainty associated with the dataset, two aspects of the data collection are suggested for future research. First, the data collection should be conducted by at least three experts, to improve the quality of the measure (Shirazi, 2009). Second, the number of publicly available reports should be increased, allowing more experts to improve and test the dataset.

5 Conclusions

This research has presented a robust approach based on Bayesian network to obtain human error probabilities by using data from major accident reports. As major accidents attract the attention of the media, society, governments and regulators – generating prosecutions that demand more investigation time and larger teams of skilled and (ideally) independent and dedicated investigators. The proposed approach allows to:

- Provide human error probabilities with a deeper understanding of the performance shaping factors involved.
- Use data from different tasks (e.g., inspection and maintenance), rather than focusing on control room operations' tasks.
- Use data from all human-machine interfaces, including hardware (e.g., manually operated valves) and not only focused on control-room screens.
- Analyse human errors and performance shaping factors in different sectors of complex social-technical industries, if the same taxonomy is used.

The probabilistic method proposed allows not only to deal with scarce data but also to quickly update the values when a specific set of performance shaping factors is observed during the operational phase (e.g., through safety audits or equipment inspection). By introducing an additional state in the node of the Bayesian Network, the proposed approach allows to address

the problem of lack of information about specific conditional probability thus increasing the transparency about the uncertainties of the human error probability estimation. Verification and validation steps are provided to assess the accuracy of the estimated human error probabilities and the uncertainty related to the model or dataset used. The approach presented in this paper have the potential to minimize the use of human reliability analysis methods to quantify and calibrate human error probabilities, thus minimizing the need of expert elicitation – leaving for them the important mission of identifying critical tasks and the possible types of human errors associated, discussing possible controls and developing mitigation actions.

Chapter III: Using credal networks to assess sparse empirical data

Overview

This chapter's introductory overview presents a discussion on the challenges and limitations of the results obtained on the first objective of the research (concatenated in the research paper presented in Chapter 2), and motivation for starting the second objective (concatenated in the manuscript in this chapter).

A limitation of the first paper is relative to the qualitative model used, which has been an attempt to make a generic model of human behaviour in a complex industry, that could be adapted according to specific cases (see Figure 2-4). However, for the second manuscript, it has been comprehended that in industry practice it is more useful to model the operation which is conducted by humans (see Figure 3.11). For this reason, the oil & gas regulator safety auditors in Brazil (ANP) have been contacted to help to select a case study: any real operation of their interest, which contains the interaction of humans, technology and organisational factors, and which they have enough data to understand its frequency and consequence. A simple model that could demonstrate whether (and how) imprecise probabilities could help missing data problems in such models. The ANP safety advisor who has gratefully helped with the case selection and data has been included as a co-author of the manuscript.

Also, comparing CREAM human error probability (HEP) data (Hollnagel, 1998) to the HEPs generated by the developed Bayesian network, as suggested by Figures 2.5 and 2.8, might not be valid, as some of the CREAM HEP data might not have been conditioned on PSFs. Thus, the validation step was not conducted for the case study of the second manuscript – although future validation is advised if human error data is collected from the selected operation. A challenge faced in the first research paper, which has motivated the second phase of the research, has been to develop a better strategy to tackle the missing data problem. The issue arises when attempting to quantify the human reliability in the case study model using the MATA-Dataset: many of the conditional dependencies between human errors and performance shaping factors are not found in the database. For these cases, the conditional probability tables (CPT) presented some missing combinations, with both states being null. Although this incomplete information could suggest that certain human errors are impossible under certain organisational and technological conditions, it is more reasonable to interpret them as uncertain information about an event rather than considering it an impossible event (Fenton and Neil, 2012). Although the first paper has proposed a practical way of dealing with this problem

without needing experts' judgements on the probability estimate of such events, the strategy is simplistic and carries hard assumptions (such as explained in the approach labelled as '*not applicable*' state in Section 2.2.4. Common approaches to deal with missing data in HRA).

A more appropriate strategy has been proposed in the second manuscript: the use of credal networks to model also the lack of information without making strong assumptions (Cozman, 2000). According to Cozman, the credal network allows the representation of incomplete beliefs through a set of measures.

For this reason, this chapter investigates a novel methodology for dealing with missing data using intervals comprising the lowest and highest possible probability values, shifting the probabilistic tool from Bayesian to credal networks.

The first attempt and idea have been presented at the ESREL conference, receiving relevant feedback from the human reliability community (Morais et al., 2019a). The Figure 3A depicts the difference on applying the '*not applicable*' state and the *credal network approach* in a simpler model used in (Morais et al., 2019a), where the nodes *observation missed*, *inadequate plan*, *wrong time*, *wrong place* and *wrong type* had missing data combinations on their conditional probability tables.

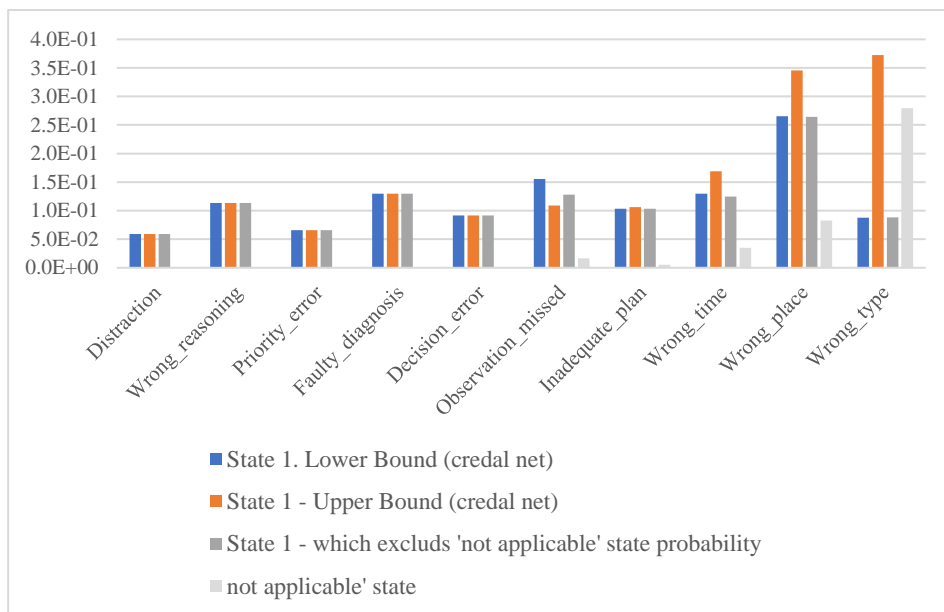


Figure 3A. Difference on applying the '*not applicable*' state and the *credal network approach*

Figure 3A shows that for the majority of these nodes the probabilities obtained with the '*not applicable*' states are equal to the lower bound of results obtained by the credal network. These results indicate that the '*not applicable*' state strategy might miss worse scenarios for human error probabilities – which is not desirable for risk assessments.

The credal network combined with the novel methodology has been applied in this chapter to quantify the risks associated with a critical task in offshore oil & gas installations. The last challenge has been deciding which is the most relevant variable in the presence of imprecision – the reason why a novel decision-making strategy for diagnostic analysis has also been suggested.

This research aims to provide less conservative human reliability analysis by providing realistic uncertainty depiction, ultimately improving risk communication between risk assessors and decision-makers.

The next pages of this chapter are based on the second manuscript aligned with the second objective of the research. I have been the leading author, and responsible for the conceptualization, data analysis, methodology and writing the first draft. The article has been co-authored by Mr Hector Diego Estrada-Lugo⁶, Dr Silvia Tolo⁷, Mr Tiago Jacques⁸, Dr Raphael Moura, Prof Michael Beer, and Prof Edoardo Patelli.

⁶ Institute for Risk and Uncertainty, University of Liverpool, United Kingdom

⁷ University of Nottingham, United Kingdom

⁸ National Agency for Petroleum, Natural Gas and Biofuels (ANP), Av. Rio Branco, 65, CEP 20090-004, Centro, Rio de Janeiro, RJ, Brazil

Robust data-driven human reliability analysis using credal networks

1. Introduction

The risks arising from the interaction of workers, tools, technologies and techniques can be assessed in industry through a systematic process known as human reliability analysis (HRA) (Kirwan, 1994). HRA aims to identify the possible types of human errors for each task, to understand which factors might trigger them, and to propose solutions to reduce human errors. In the early stages of human reliability practice, engineers have started to collect data on human errors using the same concepts of component reliability – focusing on errors occurred in function of tasks and time. More recently, engineers have started to work together with psychologists and sociologists, moving the empirical focus to measure errors under certain context (i.e. performance shaping factors, also known as performance influencing factors and human factors, which includes organisational and technological factors) (Kirwan, 1994, Hollnagel, 1998). Unfortunately, many of those databases had been discredited due to their large variability, especially if compared against the components reliability estimates (Kirwan, 1994). Overall, many data collection projects have been mostly used to validate methods based on expert judgement rather than serving a data-driven human reliability analysis (Kirwan, 1997a). This might be one of the reasons why some authors consider the state of the art in quantitative human reliability analysis too poor to make the summative assessments of risk and reliability required by regulators (French et al., 2011). This highlights the urgent need for novel tools and methodology able to tackle such limitations (French et al., 2011).

The starting point of this work is the research question whether imprecise probability theory might help to capture and adequately uncertainty about model human, ensuring its credibility. This could potentially translate in numbers the *soft barriers concept* already used in safety analysis. *Soft barriers (or soft defences)* consist of risk reduction measures that rely on human decisions or actions (i.e. administrative systems or procedures), acknowledged to be more variable than *hard barriers* which rely on hardware such as physical or technical components (Reason, 2016, Sklet, 2006). Thus, soft barriers are already recognised as carrying a higher degree of variability, and safety analysts would potentially benefit from the depiction of soft barriers variability.

As the very name suggests, the reliability of soft barriers is considered more uncertain than that associated with hard barriers. Variability is inherent to human behaviour. (Kirwan, 1994) Recent research suggests that Bayesian network, a graphical probabilistic tool developed

in the late 1980s, could be a more suitable solution to model the uncertainty associated with human reliability analysis (Mkrtchyan et al., 2015). However, its use implies the need to characterise the conditional probability distribution associated with each model variable, requiring a larger amount of data than is usually required by other traditional tools, such as fault and event trees (Mkrtchyan et al., 2016). This implies that despite increasing empirical data collection efforts, the problem of missing human reliability data would persist, as many of the conditional dependencies between human errors and performance shaping factors are not found in the available databases. Although in theory this would suggest the impossibility of certain human errors under certain organisational and technological conditions, it is more reasonable to interpret such information as the result of a lack of knowledge rather than a reliable depiction of reality, as uncertain information rather than impossible events (Fenton and Neil, 2012). Hence, many of the human error probabilities proposed in existing human reliability methods are based on experts' opinions rather than on the incomplete available information (Mkrtchyan et al., 2016, Cozman, 2000).

This paper proposes an alternative strategy that captures the inherent imprecision of human behaviour within soft safety barriers and accounts for typical missing data in conditional probability tables, bypassing the need for strong and often unjustified assumptions (see examples in section 2.2.4). The strategy relies on the use of credal networks, an extension of Bayesian networks characterised by the capability of representing imprecision (Cozman, 2000). The approach proposed in this study expands on strategies developed by some of the authors in a former study (Morais et al., 2019a).

The current paper is organised as follows: the theoretical background in section 2 focuses on the nature of empirical data and the qualitative and quantitative tools to model them, including the approaches used so far to tackle missing human reliability data. Section 3 describes the proposed alternative approach based on credal networks to tackle the problem of sparse data, and their mathematical background. The developed methodology is then applied to a case study in section 4, where the human reliability of depressurising oil tanks in an offshore oil & gas installation has been evaluated. Finally, the advantages, possible applications and limitations of the approach are discussed in section 5.

2. Theoretical background

2.1 Human reliability empirical data

Empirical data are obtained by observation and experimentation. The definition of human reliability data entail information able to provide a *human error probability* (HEP) for

each operational task in function of time or context (performance shaping factors), i.e. number of observed errors by number of opportunities for error (Kirwan, 1994, Hollnagel, 1998). It is common practice in human reliability analysis to fill gaps within the data with expert opinions: the provision of probability measures by experts is known as *expert elicitation*. Although largely adopted in practice, it is widely recognised that expert elicitation is affected by bias (Mosleh et al., 1988) and overconfidence (Lin and Bier, 2008). It might also be unfeasible if experts need to elicit a variable under many simultaneous conditions (Evans et al., 2003). Therefore, research efforts have been directed at collecting empirical human reliability data. The latter may be essentially divided into four major categories: laboratory-based studies (Griffith and Mahadevan, 2015, Di Flumeri et al., 2019), simulators (e.g. HuREX, SACADA, HAMMLab, and ongoing efforts to develop a data framework to quantify the IDHEAS method) (NRC, 2014, Jung et al., 2020, Chang et al., 2014, Xing et al., 2016), derived from near-misses, i.e., incident events that could have resulted in severe consequences (Park et al., 2017, Reason, 2016, Preischl and Hellmich, 2013), and finally analysis derived from major accidents (Moura et al., 2016, Kyriakidis et al., 2015). They all have their strengths and pitfalls in relation to volume of generated data, insights of cognitive mechanisms, correlation with performance shaping factors, and availability to the public (Morais et al., 2020). Previous studies have offered suggestions on how to generate meaningful HRA empirical data, regarding preparation, collection, analysis, and application (Kim, 2020).

In the human reliability field, data collection and classification are usually done by other humans (experts), but further research is addressing the need for computer support. For example, simulators data can be observed and debriefed by experts as in the worksheets described by (Groth and Mosleh, 2012), but also can be recorded by specifically designed simulators (Sundaramurthi and Smidts, 2013). In incidents databases, the data might be collected through extensive reading of investigation reports (Moura et al., 2016) or by using a machine-learning strategy of text recognition and classification (Morais et al., 2019b). However, collecting more data is usually expensive and is not an assurance of decreasing the uncertainty but on the contrary, it may result in an increase of uncertainty due to poor sample quality (Siegrist, 2011).

The characteristics of the generated database can impact the choice of the quantification tool used (e.g. if each variable is recorded per event and is clear about variables dependencies, or if overall results are aggregated). Sometimes, the results from data collection efforts are aggregated for the purpose of publishing an article, but the authors maintain a copy of the full database in a public data repository. For example, the study in (Moura et al., 2016) provides

human errors and influencing factors as aggregated results, serving well the purpose of fault and event tree analysis. Nevertheless, the complete database behind the study allows to identify whether a variable (factor) have occurred or not for each event (Moura et al., 2020). This allows the use of tools that require explicit relationships between all variables, such as Bayesian and credal networks.

2.2 Tools to model human reliability data

For risk-informed decision making, causal or explanatory models are widely regarded as preferable to traditional statistical approaches (Fenton and Neil, 2012). This makes graphical probabilistic tools particular appealing for the task, since they are able not only to provide a good and intuitive representation of operation but also to quantify the associated risk and uncertainty (Kirwan, 1994). In HRA, the most reportedly used tools are fault trees (FT), event trees (ET) and, more recently and mainly in research, Bayesian networks (BN) and credal networks (CN) (Morais et al., 2019a). For all graphical probabilistic tools, the model structure (also known as topology) plays an important role on the numerical outputs. Thus, most human reliability methods suggest a qualitative analysis that result in a graphical structure of an operational task before the quantification of its human error probabilities. An exception to this practice would happen if the model structure were also driven by data, as investigated by (Groth and Mosleh, 2012). However, the application of such tools to real-world operations would imply the need for (very) large amount of data, a need not met by current human reliability databases for most industries and operations (Mkrtchyan et al., 2016).

2.2.1. *Qualitative analysis: model structure*

Critical tasks, potential human errors and performance shaping factors are identified by qualitative analysis, resulting in a structure for the model and preferably establishing causality. Meticulous conduction and clear description of the qualitative analysis improves the consistency of quantification results (Kirwan, 1997a, NRC, 2014). For this reason, *critical task analysis* is used here to identify the relevant model variables and *bow-tie diagrams* to define the relationships between variables.

Critical task analysis entails the identification and examination of tasks performed by humans as they interact with systems. For assessing human reliability, only the critical tasks need to be selected, i.e., the key tasks that prevent (or recover from) an incident event. One of the most popular methods is the hierarchical task analysis (HTA) (Smith et al., 2011), which starts by describing the work as imagined (e.g., written information such as operational

procedures, equipment's manuals and risk analysis) and, if possible, comparing it with the work as done (e.g. using interviews and walking through the task at site with workers involved in the operation). The basic steps to a HTA are: identification of main hazards, which tasks contribute to hazards, who performs each task, when and in what sequence; the representation of tasks in tables or diagrams in sufficient detail, and finally the identification of potential human errors and performance shaping factors (Smith et al., 2011). A risk or hazard identification analysis is an important aid to identify which tasks are critical (Hollnagel, 1998, Smith et al., 2011). For the identification of potential types of human errors and performance shaping factors, it is recommended that assessors follow guidelines of an existing human reliability method (e.g. HEART, THERP, CREAM), as each has a different set of taxonomies and cognitive models. An example of HTA is provided in the case-study analysed in the following sections. The structure resulting from the hierarchical task analysis can be converted into graphical probabilistic models (e.g. fault tree, Bayesian network), where the operation chronological-sequence would determine the direction of links between human actions, according to some traditional human reliability approaches (Hollnagel, 1998). However, results of such sequential model could fail to deliver meaningful results, making it difficult for the assessors to diagnose the actions and PSFs that are more relevant to the overall risk. To overcome this, the outputs provided by HTA can be structured as a causal analysis, by selecting which tasks correspond to the risk event, and its trigger, control, mitigation and consequent events. This modelling approach, proposed as the *causal taxonomy of risk* by (Fenton and Neil, 2012), resembles the *bow-tie approach*, a popular qualitative risk analysis in Oil & Gas industry.

(CGE, 2017).(Reason, 2016, Trbojevic, 2008) This can be seen in Figure 3-1 where the nodes in the Bayesian Network represent the main component of the Bow-tie diagram. The risk event node in the 'causal taxonomy' diagram represents the hazard (top event) in the middle of the 'bow-tie diagram', which is triggered by the events on the left and produces the consequence on the right. The blocks between triggers and hazard are the measures to prevent hazards (control node), while the blocks between hazard and consequence are the mitigation barriers (mitigation nodes) (CGE, 2017, Salvi and Debray, 2006). Bow-tie diagrams have been already used to model and quantify human factors by using a combination of fault and event trees (Salvi and Debray, 2006, Targoutzidis, 2010) and Bayesian networks (Léger et al., 2009).

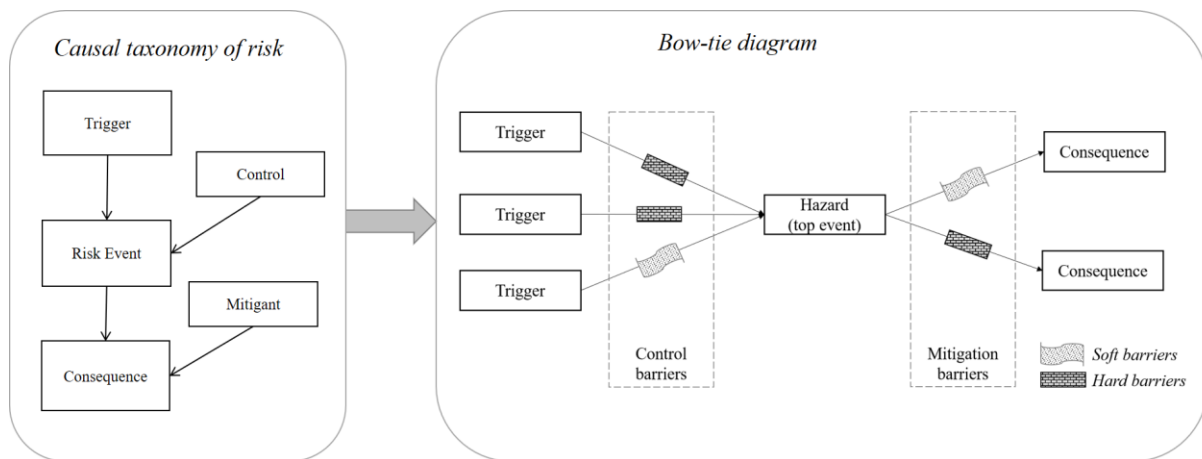


Figure 3-1. Similarity of the 'causal taxonomy of risk' between a Bayesian network and a 'bow-tie diagram'

2.2.2. Quantitative analysis with Bayesian networks: data inputs and outputs

The quantitative analysis aims at finding the probability of human errors initiating an accident event under different scenarios of performance shaping factors, ideally based on the model resulted from the qualitative part. For many years, fault and event trees have been the most used tools in human reliability quantification techniques (Kirwan, 1994). Previous studies have been demonstrating that Bayesian networks (BNs) might be a better choice than more traditional probabilistic tools (such as fault and event trees) to model and extract all information from human reliability data, many of them explored in a comprehensive review in (Mkrtchyan et al., 2015). Indeed, Bayesian networks are potentially more intuitive than fault trees, as modellers do not need to understand logical gates, just the existence of relations between variables. Variables are represented by *nodes* in the network, and their instantiation is defined by at least two *states* independent from each other (e.g. Boolean states: true or false, success or failure). Variables are known as *parent nodes* if they influence others, the *children nodes*. *Root nodes* are variables without parents. This relationship is represented as directed edges or arrows, whose direction defines the influence of parents on their child node, thus a link cannot point in both directions.

For instance, in the example in Figure 3-2, nodes PSF1, PSF2 and PSF3 represent different performance shaping factors (PSF) that trigger human error (HE) – as it is often assumed in HRA. PSF1 represents the *organisational factor*, PSF2 the *technological factor* and PSF3 the *individual factor* and they are parents of the node HE. PSF2 is a parent node of PSF3 while only PSF1 and PSF2 are root nodes.

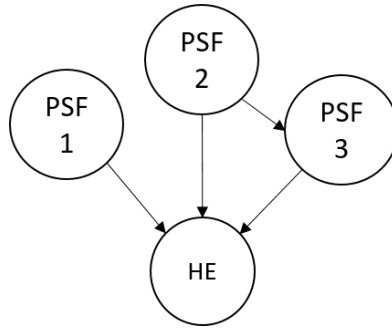


Figure 3-2. Example of a simple Bayesian network used for modelling human error.

The *conditional probability tables* (CPTs) specify the strength of the relationships represented by the network links. Root nodes require the estimation of unconditional probabilities as they are not conditioned by other nodes. Children nodes require the estimation of conditional probabilities as they are conditioned on the state of the parent nodes. The size of the resulting CPT dictates the amount of data needed. For instance, considering 2 states per node (e.g., True, False), a child with one parent requires the estimation of 4 conditional probabilities in a 2x2 table; if a child node has 2 parents, the CPT contains 8 conditional probabilities (a 2x4 table) and so on by following the rule $s^{(n_p+1)}$ where s represents the number of states and n_p the number of parent nodes (Nielsen and Jensen, 2009).

The structure of a Bayesian network for a set of n random variables (X_1, \dots, X_n) induces a unique joint probability density that can be written as a product of the individual density functions, conditional on their parent variables π_i :

Equation 3-1

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | \pi_i)$$

where, x_i represents the status of random variable X_i , π_i represent the status of all variables that are parents of the variable X_i .

For the case of HE shown in Figure 2, we use $P(HE=T)$ to indicate the probability of HE to be *true* and $P(HE=F)$ the probability that HE is *false*. We might also be interested in calculating the probability of the HE when all the PSFs are *true*. Then, the Equation 3-1 becomes:

Equation 3-2

$$P(HE = T, PSF1 = T, PSF2 = T, PSF3 = T) = P(HE = T | PSF1 = T, PSF2 = T, PSF3 = T)P(PSF3 = T | PSF2 = T)$$

Instead, the overall probability that the Human Error is true (HE=True) is obtained via marginalisation. This means that all the 8 combinations of conditional probabilities involved in the states of PSF producing the desired state of the node HE need to be added as follows:

Equation 3-3

$$\begin{aligned}
P(HE = T) = & P(HE = T | PSF1 = T, PSF2 = T, PSF3 = T)P(PSF3 = T | PSF2 = T) + \\
& P(HE = T | PSF1 = T, PSF2 = T, PSF3 = F)P(PSF3 = F | PSF2 = T) + \\
& P(HE = T | PSF1 = T, PSF2 = F, PSF3 = T)P(PSF3 = T | PSF2 = F) + \\
& P(HE = T | PSF1 = T, PSF2 = F, PSF3 = F)P(PSF3 = F | PSF2 = F) + \\
& P(HE = T | PSF1 = F, PSF2 = T, PSF3 = T)P(PSF3 = T | PSF2 = T) + \\
& P(HE = T | PSF1 = F, PSF2 = T, PSF3 = F)P(PSF3 = F | PSF2 = T) + \\
& P(HE = T | PSF1 = F, PSF2 = F, PSF3 = T)P(PSF3 = T | PSF2 = F) + \\
& P(HE = T | PSF1 = F, PSF2 = F, PSF3 = F)P(PSF3 = F | PSF2 = F).
\end{aligned}$$

The calculation of the joint probability of a Bayesian network becomes an impossible task to be carried on manually since the number of combinations quickly explodes with the number of nodes present in the network. For instance, with binary discrete variables and 10 nodes, it requires the calculation of $2^{(10+1)} = 2048$ combinations. The computation of the posterior probabilities of the queried nodes, from prior probabilities and evidence can be carried out adopting different inference methods. Exact inference algorithms based on analytical approaches provide the value of the interval probability such as by computation tree (Nielsen and Jensen, 2009), while approximation algorithms provide probabilities near the true value (Tolo et al., 2018). Usually, end users do not need to fully understand the applied inference algorithm, however they must have in mind that the complexity of the model and their need for reproducibility of results might impact their choice. Although exact inferences result in the computation of exact probability interval, they are computationally expensive and unfeasible for large sized systems. Consequently, for large networks approximation algorithms are necessary, although usually associated to unknown rate of convergence which can compromise the robustness and reproducibility of the analysis (Estrada-Lugo et al., 2019b, Tolo et al., 2018).

Bayesian networks are also used for diagnosis. They allow to identify the input with the higher impact on the output. For instance, an analyst would like to identify which PSF is the most likely trigger for the HE. Using the Bayes' rule the conditional probability of PSF1 knowing that HE has occurred (that represents the evidence) can be computed:

Equation 3-4

$$P(PSF1 = T | HE = T) = \frac{P(HE = T | PSF1 = T) \times P(PSF1 = T)}{P(HE = T)}$$

Similarly, the conditional probability for PSF2 and PSF3 can be computed. The above Equation can also be used to calculate the probability of PSF1 knowing that HE has not occurred, i.e., $P(PSF1 = T|HE = F)$ and any other combination of events. This method is known as Bayesian inference. Diagnosis is particularly useful in HRA to investigate which factors affect human error the most, which helps risk analysts in proposing risk reduction measures. Additional benefits of using Bayesian networks for HRA are that different sources of information can be combined, and parent nodes can be dependent on each other – important features considering the mutual influence of performance shaping factors. There are different strategies to define the Bayesian networks graphical structure. Domain knowledge engineers usually prefer to follow a library of patterns, known as *idioms*. Each idiom represents a type of uncertain reasoning, being the four more common the cause-consequence idiom, measurement idiom, definitional/synthesis idiom and induction idiom (Fenton and Neil, 2012). It is also possible to learn Bayesian network structure from data (Groth and Mosleh, 2012, Groth et al., 2019), although this feature is considered more useful for data-rich applications. Usually this is not the case for human reliability data (Mkrtchyan et al., 2016). Instead of choosing between Bayesian networks or fault trees to model human reliability data, one can opt to transform Fault Trees into Bayesian networks (Bobbio et al., 2001) or even to combine both, as demonstrated by previous studies that have integrated human reliability Bayesian networks into systems' Fault Tree analysis (Martins and Maturana, 2013, Trucco et al., 2008, Ramos et al., 2020). Besides supporting the evaluation of reduction measures at the organisational level (Trucco et al., 2008), or to complement an existing system reliability analysis with human reliability elements, the Bayesian network - Fault Tree integration might provide a better acceptance of Bayesian networks in sectors already familiar with Fault Trees.

2.2.3. Missing data in Bayesian networks' conditional probability tables (a recurrent problem in HRA)

Missing data is a main problem for the application of Bayesian networks to model and quantify human reliability analysis. Describing all possible combinations within variables comes at a cost: a huge amount of data needed. For instance, with respect to the conditional probability table in Table 3-1 representing the model in Figure 3-2, all states of a combination must sum to one, as defined by a probability axiom (Fenton and Neil, 2012, Nielsen and Jensen, 2009).

Table 3-1. Conditional Probability Distribution of node 'Human Error' (HE).

PSF1: Organisational factor	TRUE				FALSE			
PSF2: Technological factor	TRUE		FALSE		TRUE		FALSE	
PSF3: Person related factor	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
HE: Human error = FALSE	0	0.1	0.0	0	0.1	0.7	0.5	0.4
HE: Human error = TRUE	1	0.9	0.0	1	0.9	0.3	0.5	0.6

However, note Table 3-1 has a column which both states have zero probability (showed with bold font), because that combination of factors has never being recorded (i.e. there is no available data). This results into a computational (the missing combination does not comply with a probability axiom and preventing the use of the inference algorithms) as well as conceptual problem preventing the use of Bayesian networks. The conceptual problem is that, although this particular missing data set has been previously defined as impossible path (Fenton and Neil, 2012), treating it as an impossible event is equal of assuming that this combination of states is impossible to occur. However, there is no evidence to corroborate such hypothesis. It seems more reasonable to assume that the lack of data is an indication of an uncertain event, due to past events with incomplete information (Fenton and Neil, 2012). For this reason, it is assumed that missing data in HRA may be due to lack of observations rather than due to the impossibility of the associated event. This is tantamount to acknowledging that a combination of events that have not been observed in past events and collected into a database might actually occur. This concept is present in almost all human reliability data collection efforts: for simulators, debriefing does not always clarify which PSFs have triggered a human error (Kim, 2020); for near-miss reports, events might be underreported to regulators (Preischi and Hellmich, 2013); for accident reports (Moura et al., 2020), even after scrutinised investigations (Moura et al., 2016), some factors might not be observed or reported due to investigators' time, knowledge and bias constraints (Kletz, 2011). On the basis of such observations, the next paragraphs review how previous studies have dealt with the uncertainty caused by missing data, especially when using Bayesian networks.

2.2.4. Common approaches to deal with missing data in HRA

When observations are not available to fully define conditional probability distributions (CPDs), a standard approach adopted in practice is to *assign equal probability for both states* (Fenton and Neil, 2012). This is also the standard approach used by some Bayesian networks software (Bencomo and Blair). However, such strategy implicitly relies an extremely strong assumption and it might introduce significant bias in favour of a state that is actually rare.

Linear interpolation algorithms have been also used to fill data gaps in CPTs, by extracting information on the factor effects from known CPDs using anchors, i.e. positions in CPTs which the filling method will be based on, and extrapolate for the unknown CPDs. An ordinary linear interpolation procedure is then adopted to generate data searches for the maximum and minimum parameters (known prior probabilities) and interpolate the values in-between (Martins and Maturana, 2013). The functional interpolation (Podofillini et al., 2014) and the Cain calculator (Cain, 2001) are methods to build CPTs from limited expert judgement, and they seem to be adaptable to work solely based on empirical data – provided that the database fulfils the anchors, instead of prompting them from experts. The *functional interpolation* method consists of approximating CPD anchors with functions, interpolating among available CPDs to obtain full set of approximating functions, and discretizing them back to obtain the full set of CPTs (Mkrtchyan et al., 2016, Podofillini et al., 2014). *Cain calculator* differs not only on the position of anchors, but also on further calculating interpolation factors for parent nodes, and missing relationships in CPDs by using interpolation factors (Mkrtchyan et al., 2016, Cain, 2001). The method directly exploits monotonicity, as interpolation factors to determine the proportion of change in the child states probabilities from parent nodes and missing relationships in CPTs (Mkrtchyan et al., 2016, Cain, 2001). Monotonicity might be an unjustified assumption as it implies that parents’ effect on children state has a constant direction, with monotonic and positive influence. However, contextual factors effects on human could be also affected by the model structure (Martins and Maturana, 2013), or by socio-technical systems not necessarily behaving as coherent systems with multistate components (Morais et al., 2020). Indeed, this has been also pointed by a validation study of HRA methods with empirical data, which has concluded that significant improvement in the treatment of dependence is needed for all methods assessed (NRC, 2014).

Expert elicitation is the most common strategy for filling gaps on data (Lin and Bier, 2008). Using *expert judgement* to elicit data means asking one or more experts in a field what probability they would assume for a specific set of conditions. Many approaches exist in HRA to tackle issues related to expert opinions, e.g., bias (Mosleh et al., 1988), disagreement (Mkrtchyan et al., 2015) and overconfidence (Lin and Bier, 2008). Experts can contribute with direct probability values (i.e., direct elicitation) or via relative judgements (i.e., indirect elicitation), e.g., give their opinion through qualitative scales, questionnaires (Ramos et al., 2020). There are approaches to aggregate human error probabilities estimated by multiple experts, and some are able to distinguish the variability of HEPs from the variability between the experts (Podofillini and Dang, 2013). *Expert elicitation* are limited to the estimation of small

CPTs due to humans' inability to estimate the influence of more than three factors simultaneously (Evans et al., 2003) or the impracticable large number of combinations leading to excessive elicitation burden (Wisse et al.).

Noisy-OR method is the most used model to populate CPTs from partial information, supporting both *expert elicitation* and empirical *data mining* (Mkrtchyan et al., 2016, Xiang and Jia, 2007). The approach assumes that parents are independent, and each parent node combination of binary states produces an effect on a child node. Finally, their interaction is expressed by a logic OR gate. For HRA these are undesired assumptions (Mkrtchyan et al., 2016). To tackle these impediments, extensions have been proposed. The *noisy-MAX model* enabling multi-states nodes (Henrion); the *recursive noisy-OR* (RNOR) model allows multiple causes as input (Lemmer and Gossink, 2004) and inhibition when multiple causes are present to allow the impact of each factor (Kuter et al.). The *non-impeding noisy-AND tree* allow both reinforcement and undermining effects (Xiang and Jia, 2007). However, these Noisy-OR extensions generally address either dependent influences or multi-state nodes rather than both issues simultaneously (Mkrtchyan et al., 2016).

A pragmatical solution consists of adding an extra state to child node with missing combination in its CPT. This extra state is often labelled '*not applicable*' state: the states without data remain with zero probability and the '*not applicable*' state is assigned with the number one (Fenton and Neil, 2012). If the new state propagates to other children nodes, all new combinations generated from this state have to be also assigned to '*not applicable*' states. In HRA field, it has been observed that this strategy strongly assumes that the missing combinations are impossible to occur, although its use increases the transparency about uncertainties, and helps to maintain track of missing combinations in CPTs (Morais et al., 2020).

Artificial data implies the generation of data with known properties by an algorithm rather than expert opinion. The *maximum likelihood estimator* (MLE) identify the missing values as the probability that makes observed data the most likely to occur (Myung, 2003). MLE was used in human reliability research to test a modelling approach where performance shaping factors have a joint effect on human error probability (Stempfel and Dang, 2012). The study was not aimed at filling missing data, but to test the boundaries of Bayesian networks for HRA by using artificial data, e.g. testing the effect of different sample sizes. Although the approach seems promising to estimate missing data in an unbiased manner, there are two potential weaknesses to address. Firstly, the assumption underlying the randomly generated data is an inherent limitation of the approach (Stempfel and Dang, 2012). Secondly, while

interpreting an MLE-based analysis the user should not jump to conclusions if one model fits the data better than another. This is because achieving a superior fit might be unrelated to the model's fidelity to the underlying process, but merely because the more parameters a model have the higher the chance of fitting all data – sometimes performing even better than the real models that generated the data (Myung, 2003).

The approach of *deriving data from underlying method relationships* is based on the principle that the model structure is what ultimately defines the conditional probability distributions. If the empirical database does not provide information for a certain combination, the assessors can go back to the qualitative analysis and merge some factors until the full CPT can be assessed. This assumption is based on causal information that can be learned from theories underlying HRA methods, patterns in the data or expert judgement (Groth and Mosleh, 2012, Groth et al., 2019). The approach is also known as *synthesis idiom* (determining synthetic nodes from parents by using a combination rule) (Fenton and Neil, 2012). Merging data from factors *communication failure* and *missing information in CREAM methodology*, as they both relate to communication, is a good example of *synthesis idiom* (Hollnagel, 1998). In a marine engineering application, CREAM (Hollnagel, 1998) has been synthesised by incorporating fuzzy evidential reasoning and Bayesian inference logic to model dependency among common performance conditions (Yang et al., 2013). In (Groth and Mosleh, 2012), a structure simplification has been conducted by identifying *error contexts*, after a preliminary analysis of data using correlation and factor analysis. Error contexts can be also obtained with self-organising maps to analyse patterns from major accident reports (Moura et al., 2017a). *Deriving data from underlying method relationships* reaffirms the importance of the qualitative assessment as changing the structure also changes the amount of information needed (NRC, 2014).

Although data generated in simulators has been traditionally used to validate probabilities obtained by experts (Kirwan, 1997a, NRC, 2014), recent research investigates its use to fill missing data. In (Groth and Mosleh, 2012), recorded events from multiple simulator data collection efforts have been merged by a structured set of performance shaping factors guided by a theoretical model that aggregates their information from over a dozen HRA methods. In (Groth et al., 2014), a Bayesian updating process was conducted on HEPs generated by simulator data – the prior distribution being based on an HRA method, and the likelihood function specified to match simulator data. Yet, simulators have their limitations. A summary of important changes in simulators code to account for the human performance uncertainty has been listed after reviewing HRA methods, options of probabilistic models, and interface

(Sundaramurthi and Smidts, 2013). A summary of lessons learned from challenges in data collection from simulators has been suggested by (Kim, 2020), which considerations might assist on the use of simulator as a unique data source to HRA models or to complete missing information.

All approaches described here make *assumptions*, some more than others. The issue underlying the adoption of *unjustified assumptions* is that they can lead to significant deviations from reality, resulting in risk underestimation or wrong resource allocation. Furthermore, no characterization of uncertainty is provided by the presented approaches making impossible for the decision-makers to associate output uncertainties with missing data.

3. Proposed methodology

3.1 Credal networks

This paper proposes a methodology of replacing missing combinations in CPTs with probability intervals. This requires a shift from Bayesian network to credal networks. There are a few examples of applications of credal nets in literature, e.g. elicitation of experts with different opinions in military field (Antonucci et al., 2009), risk of fire in residential buildings (Estrada-Lugo et al., 2019a) and railway (Estrada-Lugo et al., 2019b). To the best of the authors knowledge, credal network has not been previously adopted in the context of HRA with the exception of a preliminary research on a conference proceedings by some of the authors of this work (Morais et al., 2019a).

Credal networks are a generalisation of Bayesian networks sharing an identical graphical structure but being characterised by different probability values *Figure 3-3*. Credal networks rely on imprecise probability theory to deal with the lack of data and to avoid the use of expert judgement or unjustified assumptions. Thus, a credal network is a directed acyclic graph with random variables described in terms of sets of probabilities (credal sets) instead of crisp values as in a Bayesian network (Estrada-Lugo et al., 2020). This results in higher flexibility, allowing probabilities to be expressed also in the form of inequalities (Cozman, 2000). *Figure 3-3* provides a graphical representation of a credal network, where each Bayesian network represents a *local combination of the network*, i.e. a set of probability values complying with theoretical constraints.

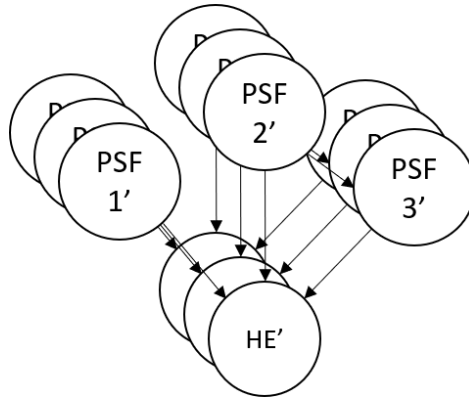


Figure 3-3. Credal network - a set of Bayesian networks characterised by different probability values.

A credal set, $K(X_i)$, consists of a group with a finite number of probability distributions $P(X_i)$. More rigorously, according to the theory of imprecise probability, *the credal set* is a closed and convex set of probability mass functions (Walley, 1991). Likewise, the conditional credal set, $K(x_i|\pi_i)$, represents the set of conditional probability distributions $P(x_i|\pi_i)$ where similarly to the case of Bayesian network π_i represent the status of all the parents' nodes of the variable X_i . When defining the probability of each state $P(X_i = x_i)$ of a variable X_i , the credal set can be expressed as an interval probability with the bounds defined by the extreme of the set of probability: $\underline{P}(X_i = x_i) = \min_{K(X_i=x_i)} (P(X_i = x_i))$ and a upper bound $\bar{P}(X_i = x_i) = \max_{K(X_i=x_i)} (P(X_i = x_i))$.

There are several sets of probability measures that can be used to represent a credal network depending on the notion of independence for imprecise probability. The present study uses the *strong extension* of a credal network that allows having extreme points represented by standard Bayesian networks (Cozman, 2000). In other words, the smallest set of local Bayesian networks that contain combinations of extreme points (i.e., the convex hull, CH) corresponds to the definition of a credal network:

Equation 3-5

$$K(X_1 = x_1, \dots, X_n = x_n) := CH \left\{ P(X_i) | P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | \pi_i) \right\}$$

When working with credal networks, the posterior probabilities are expressed in the form of intervals. The lower and upper bounds must be real numbers and they must be complementary, as shown in the equations below:

Equation 3-6

$$\bar{P}(X_i = x_i) + \sum_{j \neq i} \underline{P}(X_i = x_j) \leq 1$$

and Equation 3-7

$$\underline{P}(X_i = x_i) + \sum_{j \neq i} \bar{P}(X_i = x_j) \geq 1$$

where the summation in Equation 3-5 and Equation 3-6 is over all the states of the variable x different than x_j .

3.2. Inference methods for credal networks

A credal network, like a Bayesian network, can be computed for predictive as well as diagnostic purposes when imprecise data sets are present. To compute the inference of strong extension of credal networks, the lower and upper bounds of an event of interest referred to a query node (x_q) are given as the marginalised probability (Estrada-Lugo et al., 2019b):

Equation 3-8

$$\underline{P}(X_q = x_q) = \min_{P(x_q) \in K(x)} P(X_q = x_q) = \min_{P(x_q) \in K(x)} \sum_{x_1, \dots, x_n \setminus x_q} \prod_{i=1}^n P(X_i = x_i | \pi_i)$$

Equation 3-9

$$\bar{P}(X_q = x_q) = \max_{P(x_q) \in K(x)} P(X_q = x_q) = \max_{P(x_q) \in K(x)} \sum_{x_1, \dots, x_n \setminus x_q} \prod_{i=1}^n P(X_i = x_i | \pi_i)$$

The model outputs are obtained by computing the lower and upper bounds of the posterior probability of the queried variable $P(x_q)$, when we insert the evidence (x_e):

Equation 3-10

$$\underline{P}(X_q = x_q | X_e = x_e) = \min_{P(x_q) \in K(x)} \frac{\sum_{x_1, \dots, x_n, x_q} \prod_{i=1}^n P(X_i = x_i | \pi_i)}{\sum_{x_1, \dots, x_n \setminus x_q} \prod_{i=1}^n P(X_i = x_i | \pi_i)}$$

Equation 3-11

$$\bar{P}(X_q = x_q | X_e = x_e) = \max_{P(x_q) \in K(x)} \frac{\sum_{x_1, \dots, x_n, x_q} \prod_{i=1}^n P(X_i = x_i | \pi_i)}{\sum_{x_1, \dots, x_n \setminus x_q} \prod_{i=1}^n P(X_i = x_i | \pi_i)}$$

In the above equations, the summation operator in the nominator acts over all the variables, including the queried variable in state $x_q(x_1, \dots, x_n, x_q)$, while in the denominator, the summation is done only on the variables that are different from the queried variable ($x_1, \dots, x_n \setminus x_q$).

In credal networks the computation of the posterior probabilities of the queried nodes requires dedicated inference methods and often the approximate approaches are inevitable if using continuous variables (Estrada-Lugo et al., 2019b, Tolo et al., 2018). (Cozman, 2000)The approximation algorithms used in credal networks can be divided in inner approximation (e.g. linear programming, Hill-climbing (Cano et al., 2007)) and outer approximation (e.g., branch and bound (Cano et al., 2007), pseudo-network (Estrada-Lugo et al., 2019b)). The inner and the outer approximations provide probability bounds which enclose the exact probability interval (see Figure 3-4).

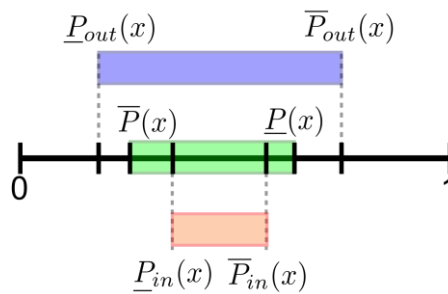


Figure 3-4. Inference methods for credal networks

An approximate inference algorithm combined with an exact method is used here. It adopts linear programming as an optimization method to find the extreme points of the credal set and then the variable elimination method is used to obtain the posterior of each local combination. The combination providing the minimum value is considered as an approximation to the lower bound. The upper bound is obtained from the combination yielding the maximum value. More details on mathematical background and inference methods applied to credal networks can be found in (Cozman, 2000, Estrada-Lugo et al., 2019b). Freely available packages that implement algorithms to compute credal networks can be found in (Cozman, 2000, Tolo et al., 2018, Antonucci et al.).

3.3. Defining the intervals to replace missing data combinations

Credal networks are used for handling imprecise and incomplete beliefs of standard Bayesian models where the missing CPT combinations are replaced by intervals comprising the lowest and highest possible probabilities, i.e., zero and one [0,1]. Therefore following the example in **Error! Reference source not found.** the replace missing CPT combinations

become: $P(\text{HE}=\text{T}|\text{PSF1}=\text{T},\text{PSF2}=\text{F},\text{PSF3}=\text{T})=[0,1]$ and
 $P(\text{HE}=\text{F}|\text{PSF1}=\text{T},\text{PSF2}=\text{F},\text{PSF3}=\text{T})=[0,1]$.

Due to strong extension properties, it was possible to replace missing CPT combinations (e.as in Table 6) with probability intervals comprising the lowest and highest possible probabilities, i.e. zero and one [0,1]. It is possible to use intervals with upper bounds less than 1 (e.g., [0, 0.5]), and the impact is a reduction on the widths of the posterior probabilities' intervals. However, as both states have to sum up to one, assuming 0.5 of one state is assuming 0.5 for the complementary state – and that would mean observations on both conditions. As the missing combinations in MATA-D mean the total lack of observations for both states, the present methodology considers that the probability interval [0,1] would be the option that best indicate the total lack of data: the number zero expresses the minimum and the number one the maximum probability of occurrence of the associated event.

Credal networks can model non-monotonic behaviour (thus more realistic human factors effects on human performance might be captured) and allows more than two states per node (enabling its application to HRA methods describing many states of human performance). Replacing missing combinations in CPTs with [0,1] intervals is a straightforward process if the table contains only one missing combination. However, in CPTs with more than two missing combinations (e.g.

Table 3-6 describes the CPT of subtask 3.3.A, where the assessors defined *incorrect prediction* as the potential cognition failure for the task, in a context where the main PSFs were *cognitive bias*, *management problem*, *insufficient knowledge*, and *adverse ambient conditions*. Table 3-6 shows the frequency this same context occurred in accidents recorded in MATA-D. Differently from CPTs shown in Table 3-4 and Table 3-5, some combinations of states of these variables do not have any reported event within all 238 accidents in the dataset (e.g. combinations #8, #10, #12 , #14 and #16). Therefore, as the lack of possible combinations events in MATA-D is interpreted as missing data rather than impossible events, the incomplete combinations were replaced by zero-to-one intervals [0,1]. As this node contains intervals, it was defined as a credal node. For this model, the majority of children nodes with more than four parent nodes had to be defined as credal nodes.

Table 3-6), the process is cumbersome, since the introduction of probability intervals in a CPT implies the review of all other probability values in order to verify the strong extension condition expressed in Equation 3-8 and Equation 3-9 (i.e. the summation of the lower/upper bound of one of variable state and the upper/lower bounds of the other states must equal to one).

The process of replacing missing data with intervals has been automatized and available in the developed tools.

3.3. Overview of how the proposed methodology works

The methodology is composed by four main modules and summarised in *Figure 3-5*. Part A converts MATA-D to prior probabilities in conditional probability tables (detailed procedure is described in a previous study (Morais et al., 2020), but also in the case study section 4.3). Part B adds intervals $[0,1]$ to combinations with no data in the conditional probability tables, transforming the nodes into credal nodes. The theory is detailed in section 3.3, and the algorithm is named switch to upper extreme in OpenCossan (Patelli et al., 2018). Part C performs the inference of the credal network with both discrete and credal nodes (theory detailed in section 3.2). Part D uses variable elimination to obtain the outputs of the model, where the posterior probabilities are expressed as intervals for credal nodes.

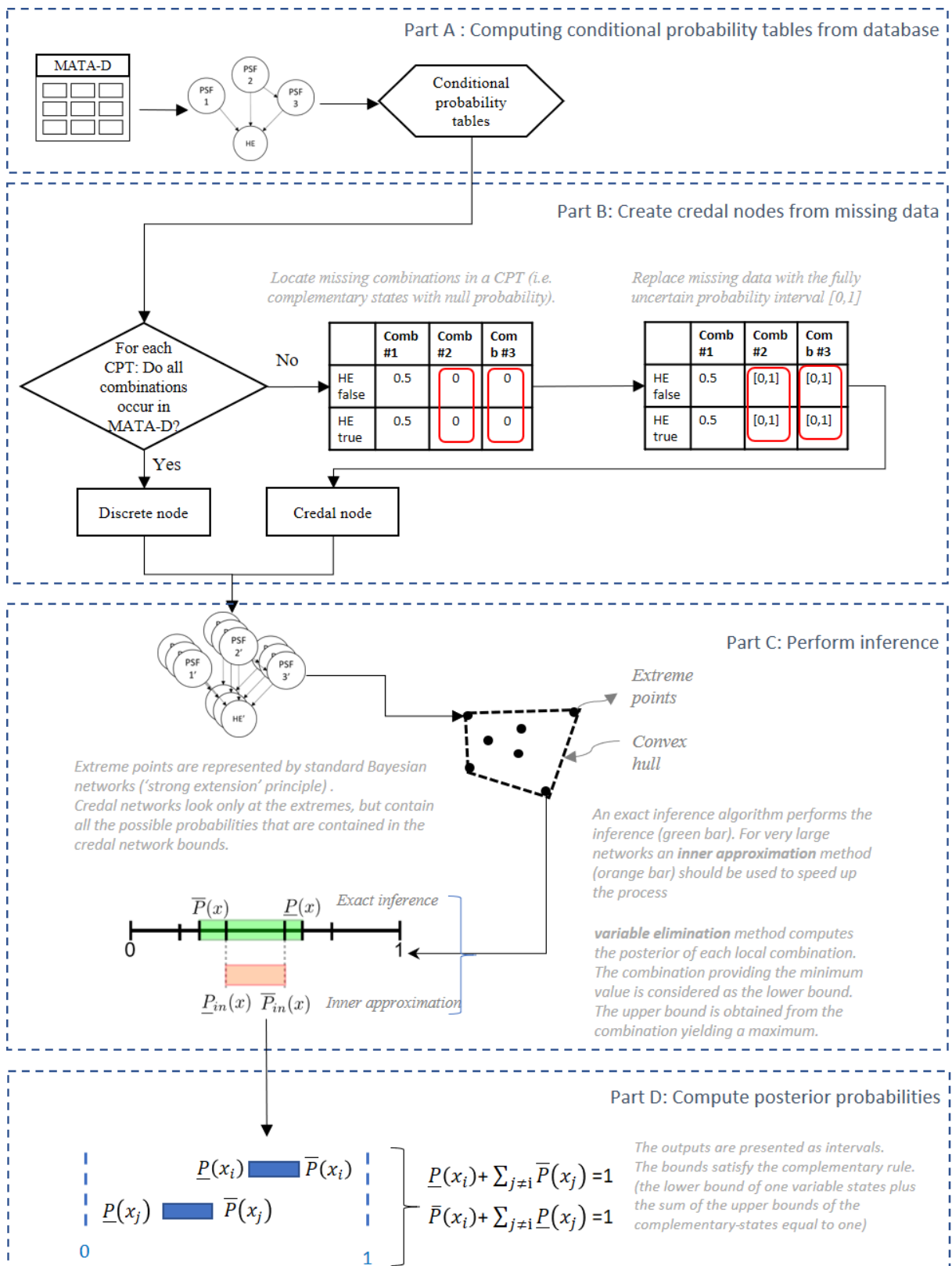


Figure 3-5. Flowchart of methodology highlighting how the mechanisms of credal network algorithm works

3.5. Decision making and criteria selection with imprecise results

In the case all the CPT combinations of a specific node are unknown, $[0,1]$ intervals represent the complete ignorance about that specific event. As a consequence, the results also become intervals, and wider intervals are often associated to more data missing. Therefore, credal networks with imprecise probability support the decision-makers to take more informed decisions by presenting the results with their associated accuracy (Patelli et al.). In addition, the diagnostic analysis provides the sensitivity analysis for HRA models, helping to allocate resources to the most influencing factors of a specific human error. Despite previous attempts to rank the variables in presence of imprecision (see e.g. (Antonucci et al., Troffaes, 2007)), challenges remain and the comparison of two or more variables affected by imprecision is not straightforward.

Let consider the simple example shown in Figure 3-2. If decision-makers want to reduce $P(HE=T)$, then they might ask if $P(PSF1=T)$ has to be reduced or $P(PSF2=T)$. This is different than reducing the imprecision of the conditional probability of the event, e.g. $P(HE=T/PSF1=T)$. In human reliability analysis, a decision-maker can interpret the lower bound of the HE probability as the best-case scenario and the upper bound as the worst-case scenario. Following this reasoning the upper bound will contain information about the highest possible probability of error under the conditions defined in the model. Criteria might vary between decision-makers, i.e. risk-prone versus risk averse. Thus, a general strategy is suggested:

- $[0,1]$ interval for the posterior probability cannot support decisions, thus more data should be collected, or a penalty should be applied;
- Wider intervals suggest insufficient of data to support the importance of a factor (and more evidence is needed to answer the question with confidence);
- Small intervals suggest that there is enough evidence to support a statement;
- Collecting more data is not an assurance that wide intervals would decrease, as it might represent state combinations that are indeed rare to happen – for these cases, it would be interesting to measure the confidence in the analysis before taking decisions, by computing the reliability with a tool such as confidence-boxes (Ferson et al., 2014).
- Different factors might have overlapping intervals and the most impacting factor might also be the most uncertain one. The *interval dominance* criteria (Troffaes, 2007) is used in this study for selecting the most important factor. Interval dominance criteria is a method for classification accuracy usually taken as heuristic, where an interval is called dominant if

might have a higher probability than a probability of the variable valued on another node (Troffaes, 2007).

The suggested criteria are summarised in the workflow shown in Figure 3-6.

. To explain the identified criteria, the pairwise comparison of hypothetical factors shown in Figure 3-7 is performed. The factors represent conditional probabilities, i.e. probability that a PSF is true knowing that a HE has occurred. In the first case the interval for the factor A is contained in the interval of the factor B, thus B is selected as the most impacting factor due to interval dominance as B has a highest upper bound. In the second case, the two factors C and D have the same lower bounds, but D has a larger interval. Therefore, it seems logic to select D because it might be possible that the factor D has a larger influence but certainly has at least the same influence of the factor C. In the third case, the factor E has the lower bound larger than the upper bound of the factor F. Hence, we have the guarantee that the factor E is more important than F. The fourth case G has the lowest lower bound, but H has the highest upper bound. Again, we select H exactly based on its highest upper bound probability – as in this case, both intervals have the same width. The fifth case shows the two factors I and J with the same upper bounds but with J having a higher lower bound. Therefore, it is logic to select J.

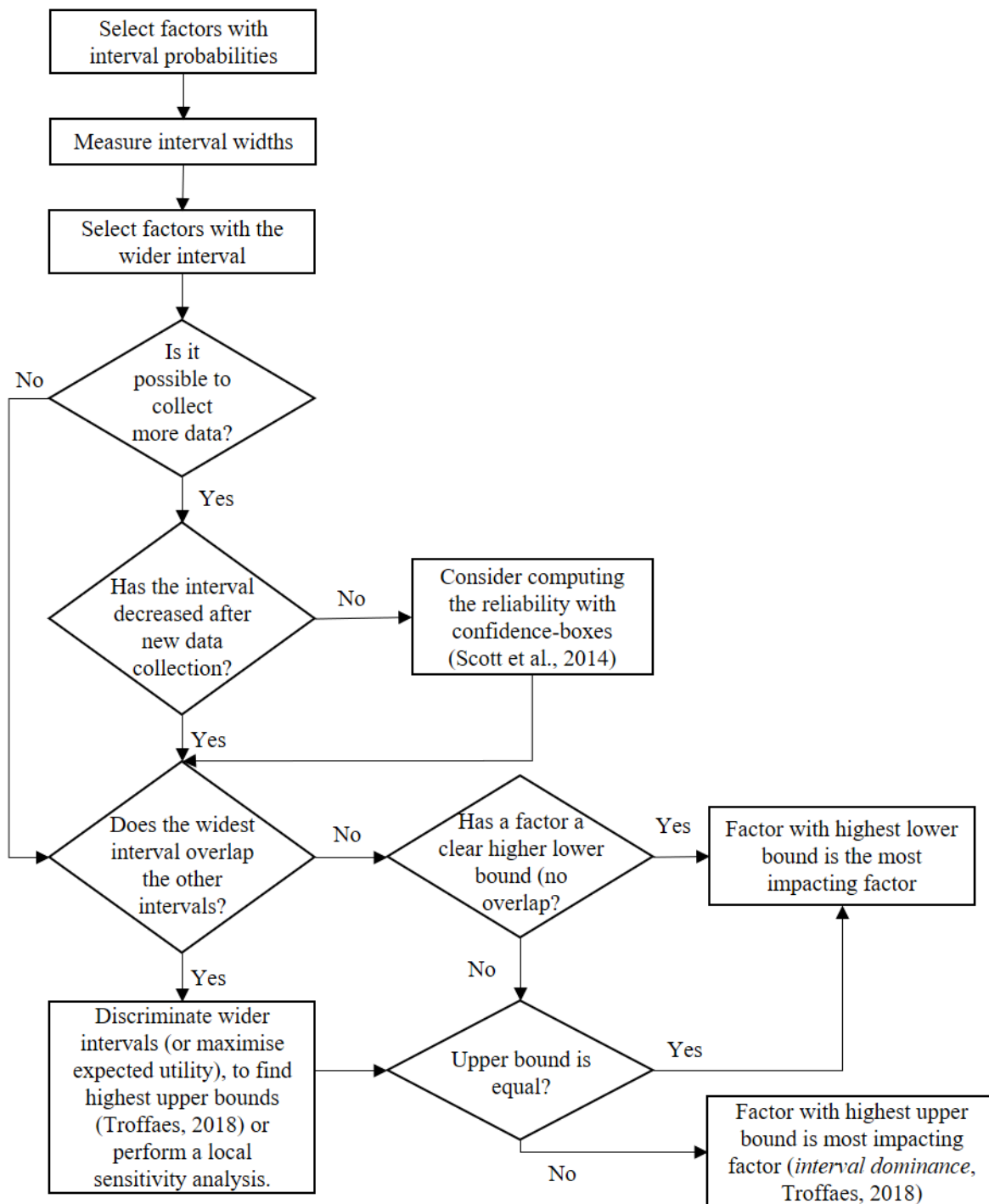


Figure 3-6. Suggested criteria for decision-making in sensitivity analysis of HRA

A more rigorous criteria could be developed if there are dependencies between parent nodes as for PSF2 and PSF3 in Figure 3-2. For instance, reducing $P(PSF2=T)$ might also reduce $P(PSF3=T)$. Therefore, a dependency analysis is required (e.g. including evidence in node PSF2 and PSF3 to calculate $P(HE)$ and then including evidence in $P(PSF3)$ and $P(HE)$ to calculate $P(PSF2)$). For instance, the imprecision of PSF3 could derive entirely from the imprecision of PSF2.

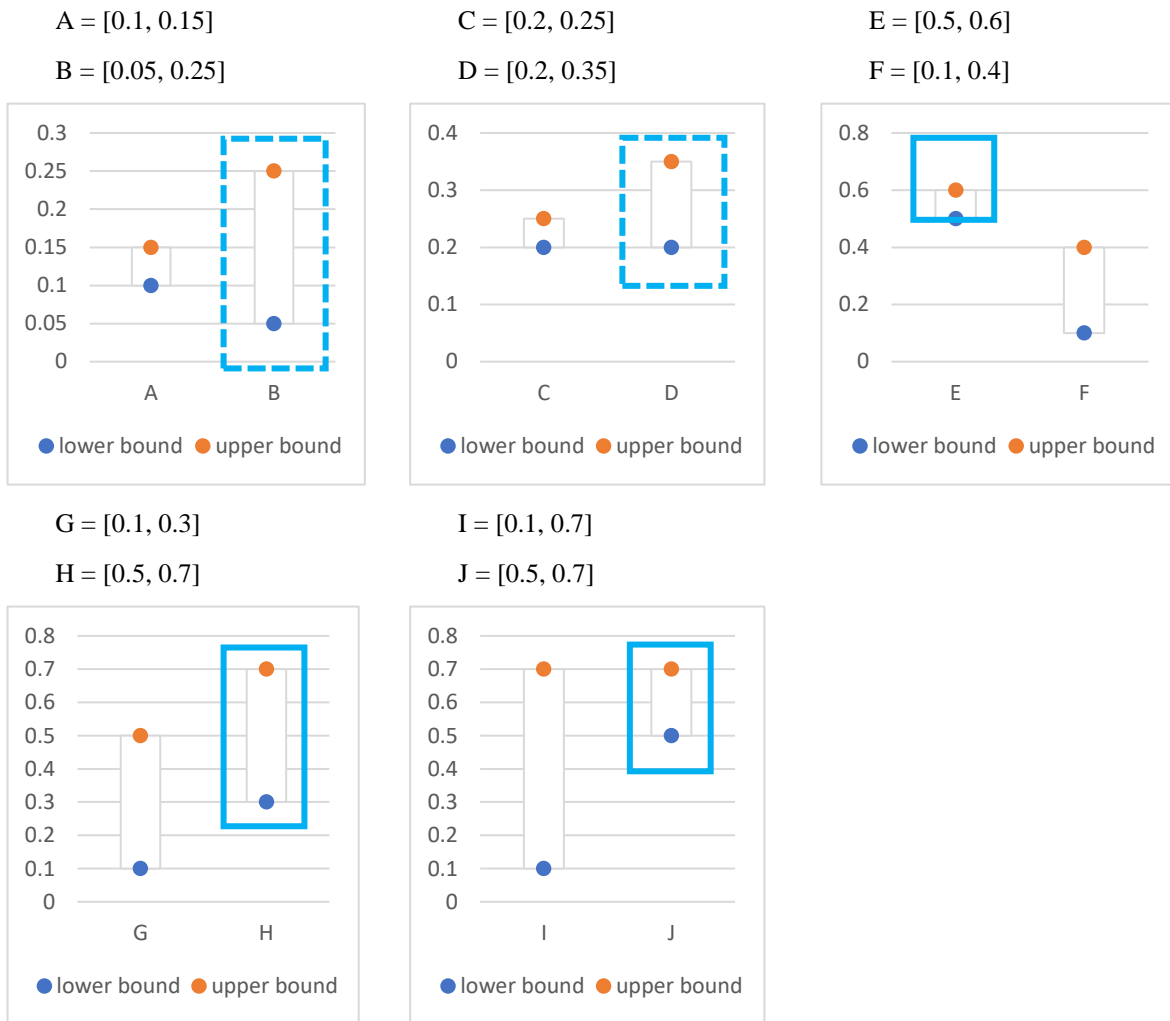


Figure 3-7. Pairwise comparison of hypothetical factors – highlighted by dashed lines are the results that could depend on the decision-making style; by solid lines: results where there is no doubt.

Results highlighted by dashed lines in Figure 3-7 are those that could have easily led to a different interpretation if the suggested criteria were not strictly followed, as they might depend on the decision-making style (many people would rather prefer allocating resources in more certain probabilities). Results highlighted by solid lines are those where there is no doubt (both lower and upper bound are higher).

3.6. Software

The credal networks methodology and the associated inference and diagnostic algorithms are implemented in the OpenCossan Bayesian network toolbox (Tolo et al., 2018), part of the OpenCossan software (Patelli et al., 2018, Patelli et al., 2016). OpenCossan is an open-source and object-oriented software for uncertainty quantification purposes based on Matlab.

The Bayesian network toolbox is used for reduction, inference computation and sensitivity analysis of credal networks (Patelli et al., 2016, Tolo et al., 2018). The object-oriented code of the toolbox allows flexibility. It automatically selects the required algorithms according to the type of node defined in the network. For instance, if the CPTs are complete and include only crisp probability values, *discrete nodes* are used. Otherwise, if the CPTs have missing combinations, *credal nodes* are used.

The toolbox allows to automatically substitute missing data with intervals and calculating the corresponding bounds.

4. Case study

This case study aims to quantify the human reliability of operator during the storage tank depressurisation on static offshore oil & gas installations known as FPSO (floating production storage and offloading system) and FSO (floating and offloading system – also known as FSUs, floating storage units). The operation is necessary for safety reasons, to avoid explosion of storage tanks due to overpressure (Vinnem, 2001). However, under certain wind conditions the vapours released might reach a source of ignition (e.g. other equipment, operations and maintenance works) with the potential to cause fire, explosion or financial loss due to emergency production shutdown (de Vos et al., 2006, Alan Keith et al., 2012). The operators are the main barriers to prevent an incident event, with little or no support from automatic systems/technology. The human reliability analysis provides a risk-informed support tool for engineers/project managers to evaluate the eventual need for design changes.

4.1. Description of the case study: FPSO's and FSO's storage tank venting

FPSOs are offshore installations that process oil & gas and store oil. Their system has production facilities on deck and storage tanks in the hull (Figure 3-8). In a generic design, a FPSO receives crude oil from an undersea reservoir via flexible risers. The incoming flow is then separated into oil, gas, and water (and sometimes salt) by process equipment on deck. The separated oil is stored in the vessel's tanks for periodic offloading to a shuttle tanker (Figure 3-10) using a floating hose, or to an FSO via fixed pipelines (Shimamura, 2002). Thus, FSOs do not have the production and process facilities (Figure 3-9).



Figure 3-8. FPSO⁹



Figure 3-9. FSO²



Figure 3-10. Shuttle tanker¹⁰

During FPSO/FSO operations, inert gas (nitrogen) is usually injected in the storage tanks, to blanket their ullage spaces and avoid an explosive mixture of oxygen and hydrocarbon vapours. In a safe design concept, when tanks are over-pressured their vents are opened (automatically or manually) to allow inert gas to escape (Figure 3.11) and avoid overpressure (Vinnem, 2001). This depressurisation of oil cargo tanks is known as *cargo venting operation* (HSE, 2010). During the operation, a small amount of hydrocarbons vapours, associated with the inert gas, escapes. This adds some risk of flammable vapours meeting a spark at the deck, resulting in a fire and/or explosion (Alan Keith et al., 2012, HSE, 2010).

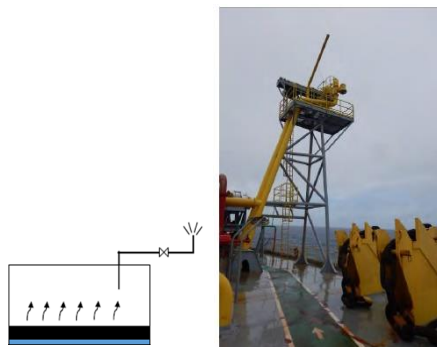


Figure 3-11. Scheme of a tank with its vent outlet and a photo of a vent outlet on a FPSO¹¹

FPSOs/FSOs and shuttle tankers have similar storage tanks venting systems, but the risk is higher for FPSOs/FSOs because they do not navigate during operation, as they are moored. Therefore, the vapours are not easily dispersed by wind as in shuttle tankers (HSE, 2010). In addition, FPSOs/FSOs have their deck space more packed with equipment than tankers (as can be noted by comparing Figure 3-8 to Figure 3-10), impeding flammable vapour to dissipate. The operational risk increases in case of low wind speed prevents vapours to dissipate, and in case of wind blowing vapor towards the process plant increases the chance of encountering

⁹ FPSO and FSO figure source: https://www.modec.com/fps/fps_o_lineup/index.html

¹⁰ Shuttle tanker figure source: <https://www.hellenicshippingnews.com/oil-tanker-demand-solid-but-trade-tensions-could-change-that/>

¹¹ Cargo vent outlet figure and scheme source: http://www.anp.gov.br/images/EXPLORACAO_E_PRODUCAO_DE_OLEO_E_GAS/Seguranca_Operacional/Relat_incidentes/Sao_Mateus/anp-final-report-fps-o-cdsm-accident.pdf

ignition sources – generated by maintenance tasks, nearby support vessels and helicopters, droplets falling from flare, and equipment. Even explosion proof equipment (i.e. Ex equipment) can be a source of hazard if their electrical installations are not correctly maintained (Rangel and Sanguedo, 2018).

Accidents related to venting operation have the potentiality to create significant financial losses due to the loss or delay of production (de Vos et al., 2006). For instance, in Brazil, whilst duty holders are increasing their production of lighter crude oil (ANP, 2020b), they have been challenged with increasing number of cases of emergency shutdowns (ESD) triggered by gas detectors been activated by flammable vapours originated during cargo venting operation (ANP, 2020a). Past related incidents have been investigated on relation to the vapour content (Alan Keith et al., 2012) and possible sources of ignition (Pursel et al., 2016a, Pursel et al., 2016b), triggering the UK safety regulator to require duty holders to take appropriate measures to prevent fire and explosion (HSE, 2010).

After the risk assessment, it comes the decision on what is the more appropriate safeguard to implement: a design modification of the system or operational measures performed by workers (de Vos et al., 2006, HSE, 2010). Even in installations where this operation is partially automatized, human decisions are still part of the process as imposed by weather conditions and concomitant operations with other nearby installations. The human reliability analysis proposed in this work attempts to support this decision. The risk evaluated is the chance of a human error triggered by different performance shaping factors of initiating an incident event.

4.2. Qualitative analysis: Model qualitative part: defining the structure

The qualitative part of the study defines the model structure. It was based on the operation's hierarchical task analysis: a structured way of condensing large amount of written information into a sequence of critical actions, screening potential human errors modes, performance shaping factors, and flagging tasks performed by different teams. The definition and criticality of individual tasks were based on information from: a safety bulletin from the UK health and safety regulator (HSE, 2010), related incidents (Alan Keith et al., 2012, Pursel et al., 2016a, Pursel et al., 2016b), different design and operational measures (de Vos et al., 2006) and written operational procedures and risk analysis (including computational fluid dynamics model) from two different duty holders operating in Brazil (not referenced here for

confidentiality reasons). All the evaluated documents had not yet considered human reliability analysis.

Figure 3-12 presents the identified hierarchical task analysis where ‘A’ refers to tasks performed by team A cargo/marine team, ‘B’ to radio-operator, ‘C’ to production team, and ‘D’ to maintenance team. Starting at the top, the first box specifies the overall task, i.e. cargo venting operation. The next layer of boxes describes the complete tasks in eight steps. Some steps consist of straightforward tasks such as taking a reading from a control panel; other steps are complex and described in more detail in the next layer of boxes. Each layer provides a complete description of the task, but each level provides more detail in a hierarchy way. After critical tasks were selected, their potential human errors and respective performance shaping factors were identified using the authors’ expertise and knowledge. The *antecedent-consequent model* (i.e. a CREAM human reliability methodology) was used as a supporting tool as it provides the correlation between human errors and performance shaping factors. Appendix B provides a detailed description of tasks, their potential human errors and PSFs and the full correlation table adapted from (Hollnagel, 1998). Note that a more realistic model would have required the use of interviews and walking through the task at site with workers involved in the operation.

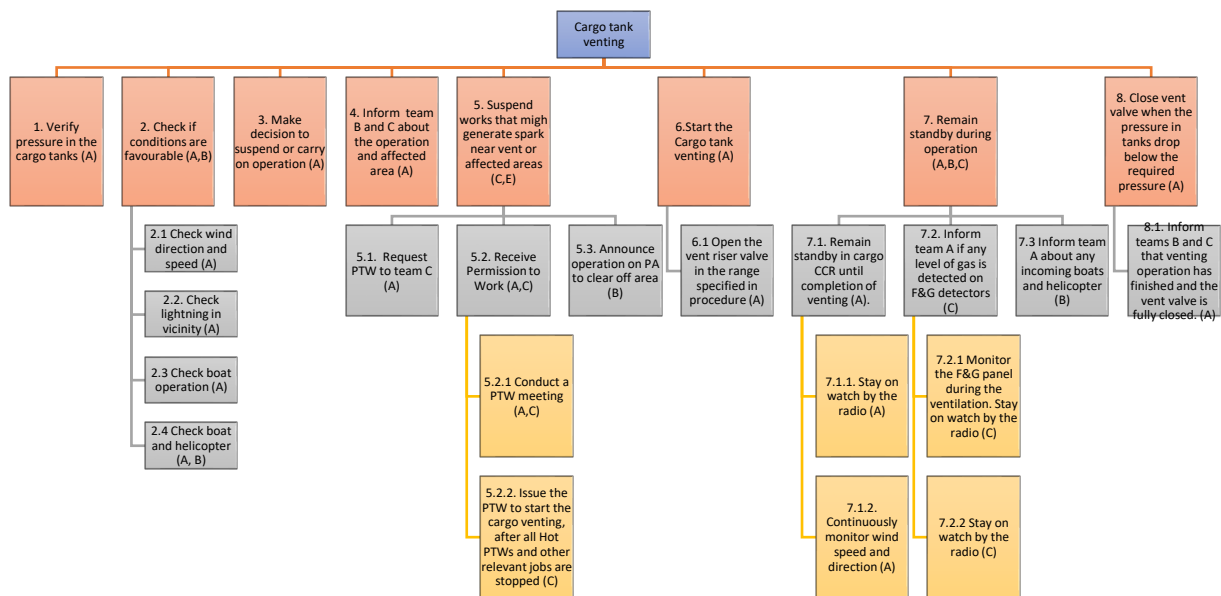


Figure 3-12. Diagram of critical tasks analysis (using methodology of hierarchical task analysis)

After defining the nodes with critical task analysis, the links between nodes were defined (the model structure). Instead of having a model based merely on the chronological task sequence, the *cause-consequence idiom* (Fenton and Neil, 2012) was used, which resembles

the logic of a bow-tie diagram. Using this idiom, each node receives a function in the model: risk or consequence event, risk trigger, risk control, or consequence mitigation. The task of actually opening the cargo tank valve (or failing to close it if the conditions change) was selected as the *risk event* node. The tasks and PSFs that would trigger the risk event are the *trigger nodes*. The tasks and PSFs that would prevent human error in the risk event or prevent the gas spreading to undesired directions were defined as the control nodes (regarding the task analysis sequence, the tasks that would finish just before the valve is opened). The consequence node is not a task nor a PSF, but the representation of possible outcomes in case the risk event actually happens, such as emergency shutdown or fire. The mitigation nodes are tasks and PSFs that would help to prevent or mitigate the consequence (e.g. tasks that would prevent spark, and tasks or systems conditions that have to be working concomitantly with the venting, from the moment the valve is opened until it is closed). The resulting model structure (model #1) is presented in Figure 3-13 where discrete nodes are represented by rectangles (child nodes in green, root nodes in blue), and credal nodes by grey ellipses.

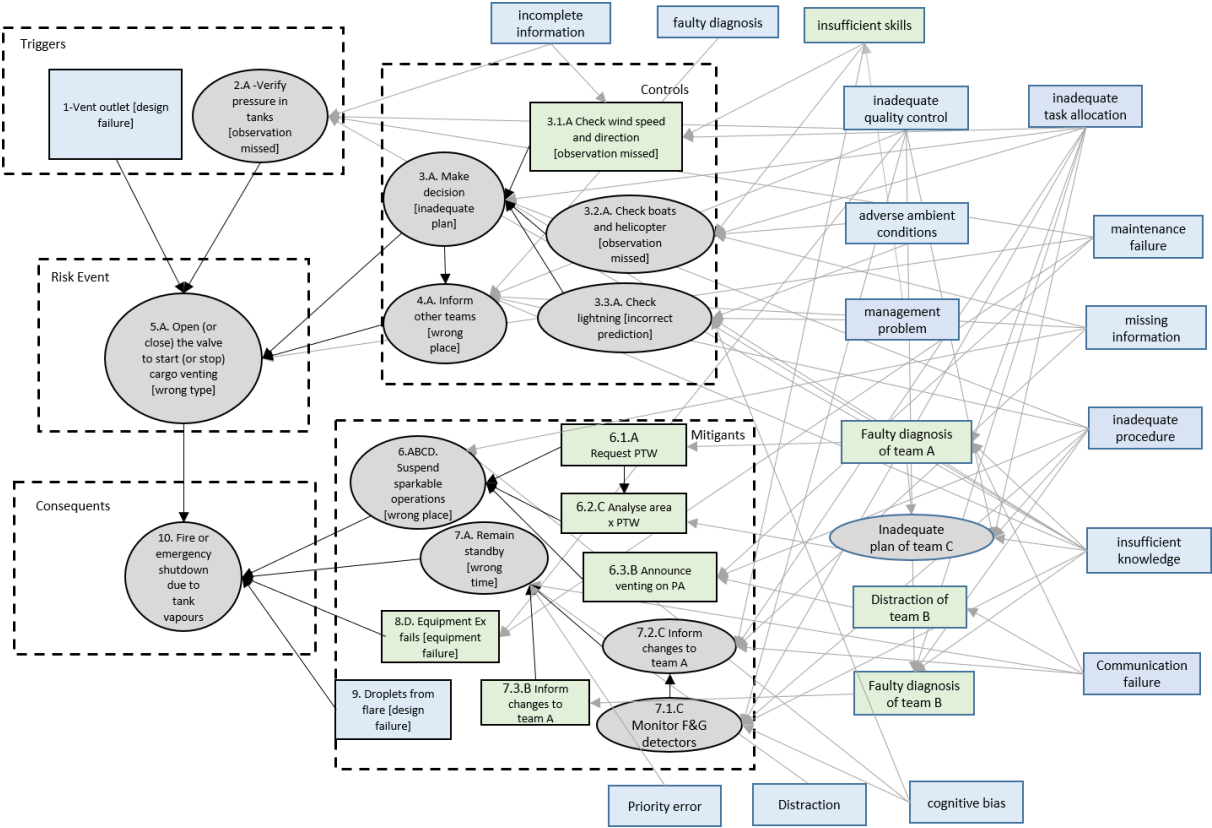


Figure 3-13. Proposed human reliability model structure for the tank venting operation (model #1)

An alternative model #2 has been created and shown in Figure 3-14. It differs from model #1 in the classification given for subtasks of tasks 3, 6 and 7, and consequently their PSFs. This is because each node of model #1 corresponds to a task in the hierarchical task

analysis, while in model #2 some nodes have been merged by using underlying CREAM method relationships.

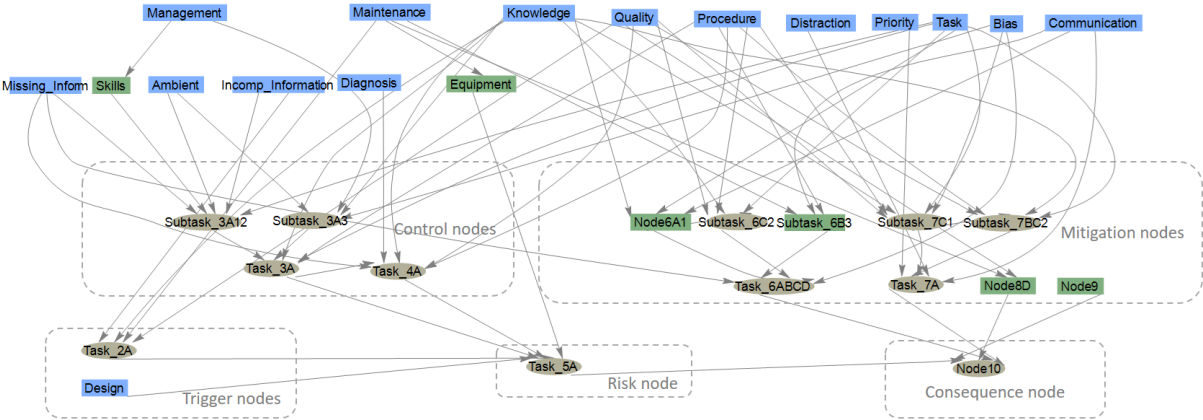


Figure 3-14. Model #2, some nodes were merged by using underlying CREAM method relationships

The decision to create a second model has been made to compare the impact of the structure simplification in the quantification results, and to measure the impact of a potential limitation of the database used, which did not account for recurrent error modes in the same event. In model #1 there are some combinations of parents and children nodes with the same error mode classification – which results in many missing combinations in the quantification phase. In contrast, due to the merged nodes, model #2 does not contain children nodes with the same classification as their parents (e.g. if child and parent nodes had the same human error, the parent was replaced by the next performance shaping factor in the structure, provided that the logic of the HRA method was maintained). Although model #2 resulted in less uncertain model (due to the less number of missing combinations), the simplification is not required for the use of the methodology proposed – thus model #2 and its results are found on Appendix C, while a brief comparison of both models are presented in results session.

Table 3-2 presents a summarised description of nodes and links of model #1, while model #2 description is presented at Table 3-3.

In Model #2, the model simplification strategy of synthetizing or collapsing nodes by applying ‘underlying method relationships’ has been used to avoid the same human error mode in consecutive nodes (as a strategy to minimise incomplete paths in the conditional probability tables).

The performance shaping factors of CREAM classification scheme, and their links to different tasks reflect the overarching influence of organisational and technological factors on performance of different teams (e.g. the root node *inadequate procedure* is the parent of six

children nodes in model #1: task 3.A, task 4.A, subtask 6.3B, inadequate plan of team C in task 6, subtask 7.1.C, and faulty diagnosis of team B in task 7). Finally, cognitive functions have been modelled separately if they were underlying tasks performed by different teams (e.g. in model #1, faulty diagnosis of team A in task 6 and faulty diagnosis of team B task 7 have been kept separated in two different nodes).

Table 3-2. Nodes' details in model #1

Trigger nodes						
Node (task number and their classification in CREAM taxonomy)	Task description	Team performing the task	Parent nodes (subtasks or PSFs, and their classification in CREAM taxonomy)	States	Node type	Data source
PSF 1 (Design failure, an organisational factor)	Tank vent outlet incorrectly designed and in unsafe location.	Not applicable (in operational phase)	None	two (true/false)	Discrete	MATA-D (Moura et al., 2020, Moura et al., 2016)
Task 2A (Observation missed, cognitive function failure)	Verify pressure in cargo tanks	Cargo team (A)	PSFs: maintenance failure, incomplete information, inadequate quality control, insufficient knowledge.	two (true/false)	Credal	MATA-D
Control nodes						
Task 3A (Inadequate plan, a cognitive function failure))	Decide between suspending or continuing operation	Cargo team (A)	Subtask 3.1.A; subtask 3.2.A; subtask 3.3.A. PSFs: inadequate procedure; inadequate task allocation; insufficient knowledge	two (true/false)	Credal	MATA-D
Subtask 3.1A (Observation missed) <i>Note (1)</i>	Check wind speed and direction	Cargo team (A)	PSFs: incomplete information; inadequate task allocation; insufficient skills	two (true/false)	Discrete	MATA-D
Subtask 3.2.A (Observation missed) <i>Note (1)</i>	Check boats and helicopter	Cargo team (A)	PSFs: inadequate task allocation, insufficient skills, missing information, adverse ambient conditions	two (true/false)	Credal	MATA-D
Subtask 3.3.A (Incorrect prediction, a cognitive function failure)	Check lightning	Cargo team (A)	PSFs: adverse ambient conditions, cognitive bias, insufficient knowledge, management problem	two (true/false)	Credal	MATA-D

Task 4A (Action in wrong place, also known as action out of sequence, execution error)	Inform other teams of upcoming operation	Cargo team (A)	PSFs: inadequate procedure, inadequate quality control, insufficient knowledge, missing information, faulty diagnosis	two (true/false)	Credal	MATA-D
Risk event node						
Task 5A (Execution of wrong type performed, execution error, e.g. action performed too fast, too slow or in wrong direction (Hollnagel, 1998))	Start tank venting by opening a valve (or failing to stop the venting operation by closing a valve)	Cargo team (A)	PSF 1 (design failure); task 2A; task 3A, task 4A, PSF equipment failure	two (true/false)	Credal	MATA-D
Mitigation nodes						
Task 6ABCD (Action in wrong place)	Suspend operations that generate spark	Cargo team (A), radio-operator (B), production team (C), maintenance team (D)	Subtask 6.1A, subtask 6.2.C, subtask 6.3.B, cognitive bias, missing information	two (true/false)	Credal	MATA-D
Subtask 6.1.A (Action in wrong place) <i>Note (2)</i>	Request permission to work (PTW) to suspend operations that generate spark	Cargo team (A)	Faulty diagnosis of team A Parent nodes of faulty diagnosis of team A: PSFs inadequate task allocation, communication failure, insufficient knowledge	two (true/false)	Discrete	MATA-D
Subtask 6.2.C (Action in wrong place) <i>Note (2)</i>	Analyse affected area and issue permission to work (PTW)	Production team (C)	Subtask 6.1.A, inadequate plan of team C Parent nodes of inadequate plan of team C: faulty diagnosis of team A, inadequate task allocation, insufficient knowledge, inadequate quality control, inadequate procedure	two (true/false)	Discrete	MATA-D
Subtask 6.3.B (Action in wrong place) <i>Note (2)</i>	Announce tank venting will start on public address system (PA, i.e. speakers)	Radio-operator (team B)	PSFs: distraction (of team B), maintenance failure, inadequate procedure	two (true/false)	Discrete	MATA-D

			Parent node of distraction of team B: communication failure			
Task 7A (Action performed at wrong time (execution error))	Remain standby in marine control room until venting completion	Cargo team (A)	Subtask 7.2.C, subtask 7.3.B, PSFs: priority error, distraction, communication failure)	two (true/false)	Credal	MATA-D
Subtask 7.1.C (Observation missed)	Monitor level of gas detection	Production team (C)	PSFs: cognitive bias, inadequate procedure, inadequate quality control, inadequate task allocation, insufficient knowledge	two (true/false)	Credal	MATA-D
Subtask 7.2.C (Action performed at wrong time) <i>Note (3)</i>	Inform changes of system state to team A (if flammable gas is detected by sensors in production modules)	Production team (C)	Subtask 7.1.C , PSFs: communication failure, inadequate task allocation, insufficient skills, missing information	two (true/false)	Credal	MATA-D
Subtask 7.3.B (Action performed at wrong time) <i>Note (3)</i>	Inform changes of system state to team A (unplanned helicopter or boat approaching)	Radio-operator (team B)	Faulty diagnosis of team B Parent nodes of faulty diagnosis of team B: PSFs inadequate procedure, inadequate quality control, inadequate task allocation, insufficient knowledge	two (true/false)	Discrete	MATA-D
PSF 8.D (Equipment failure, a technological factor)	Failure of explosion proof equipment (i.e. Ex equipment), generating spark	Maintenance team (D)	PSFs: maintenance failure, inadequate quality control	two (true/false)	Discrete	MATA-D
PSF 9 (Design failure)	Droplets from flare	Not applicable	None	two (true/false)	Discrete	UK offshore hydrocarb on releases database (HSE, 2020b)
Consequence node						
10 (consequence, not classified in CREAM taxonomy)	Fire or emergency shutdown due to tank vapours	Not applicable	Task 5A, task 6.ABCD , task 7.A , PSF 8.D (equipment failure), PSF 9 (droplets from flare)	Three (No consequence; ESD; Fire)	Credal	Brazilian incident system and regulator reports (ANP, 2020a);

						UK FPSOs (Pursel et al., 2016a, Pursel et al., 2016b); UK offshore hydrocarbon releases database (HSE, 2020b)
--	--	--	--	--	--	---

Note (1): In this model#1, tasks 3.1.A and 3.2.A have been represented separately. In the alternative model#2 these nodes have been merged (as they have same cognitive function and are in the same team).

Note (2): In model #1, task 6.ABCD and subtasks 6.1.A, 6.2.C and 6.3.B have the same human error mode. In model #2, using the underlying HRA method relationships, human error of subtasks 6.1.A, 6.2.C and 6.3.C was replaced by the next cognition function described in the model structure.

Note (3): In model #1, tasks 7.A, and subtasks 7.2.C and 7.3.B have the same human error mode. In model #2, the subtasks 7.2.C and 7.3.C were merged and the human error was replaced by the next cognition function described in the model.

Table 3-3. Nodes' details in model #2 (only nodes that differ from model #1 are shown)

Node (task or PSF, and their classification in CREAM taxonomy)	Description	Team performing the task	Parent nodes (task or PSF, and their classification in CREAM taxonomy)	States	Source
Control nodes					
Task 3A (Inadequate plan) <i>(different from Model #1, due to subtasks)</i>	Decide between suspending or carrying on operation	Cargo team (A)	Subtask 3.1.A & 3.2.A merged (observation missed), subtask 3.3.A (incorrect prediction), PSFs inadequate procedure, inadequate task allocation, insufficient knowledge	two (true/false)	MATA-D
Subtask 3.1.2A (Observation missed) <i>(different from Model #1)</i>	Check wind speed and direction and Check boats and helicopter	Cargo team (A)	PSFs: incomplete information, inadequate task allocation, insufficient skills, missing information, adverse ambient conditions	two (true/false)	MATA-D
Note: In model #2, nodes 3.1.A and 3.2.A have been merged, as they represent the same cognitive failure and are potentially performed by the same person in the same team)					
Mitigation nodes					
subtask 6.1.A (faulty diagnosis, cognitive function failure)	Request permission to work (PTW) to suspend operations that generate spark	Cargo team (A)	PSFs: inadequate task allocation, communication failure, insufficient knowledge	two (true/false)	MATA-D

(different from Model #1)

Note: In this model, instead of repeating 'action in wrong place' as the human error mode in 6.1.A it has been used the cognitive function pointed by the risk assessor as underlying that specific action (in this case, 'faulty diagnosis').

subtask 6.2.C (inadequate plan,cognitive function failure)	Analyse affected area and issue permission to work (PTW)	Production team (C)	Subtask 6.1.A (faulty diagnosis), PSFs inadequate procedure, inadequate quality control, inadequate task allocation, insufficient knowledge	two (true/false)	MATA-D
---	--	---------------------	---	------------------	--------

(different from Model #1)

Note: In this model, instead of repeating 'action in wrong place' as the human error mode in 6.2.C it has been used the cognitive function pointed by the risk assessor as underlying that specific action (in this case, 'inadequate plan').

Node subtask 6.3.B (Distraction, a temporary individual factor)	Announce tank venting will start on public address system (PA, i.e. speakers)	Radio-operator (team B)	PSFs: communication failure, maintenance failure, inadequate procedure	two (true/false)	MATA-D
--	---	-------------------------	--	------------------	--------

(different from Model #1)

Note: In this model, instead of repeating 'action in wrong place' as the human error mode in 6.3.B it has been used the cognitive function pointed by the risk assessor as underlying that specific action (in this case, 'distraction').

Node task 7A (Action performed at wrong time, execution error)	Remain standby in marine control room until venting completion	Cargo team (A)	Subtask 7.1.C (observation missed), subtask 7.2.BC (faulty diagnosis), PSFs priority error, distraction, communication failure	two (true/false)	MATA-D
---	--	----------------	--	------------------	--------

(different from model #1, due to some different PSFs)

Node subtasks 7.2.BC (faulty diagnosis, cognitive function failure)	Inform changes of system state to team A (flammable gas is detected by sensors in production modules)	Radio-operator (Team B), production (Team C)	Node 7.1.C (observation missed), PSFs inadequate procedure, inadequate quality control, inadequate task allocation, insufficient knowledge	two (true/false)	MATA-D
--	---	--	--	------------------	--------

(different from model #1)

Note: merged subtasks 7.2C and 7.3B

4.3. Quantitative analysis part: feeding data to the probabilistic tool

The strategy to quantify and predict human performance used in this study diverges from the original CREAM method (Hollnagel, 1998), which suggests the evaluation of worker control level on performing an operation (i.e. scrambled, opportunistic, tactical, strategic) by adjusting the human error probabilities according to common performance conditions. In this study, the control level and common performance conditions were not evaluated: instead, the assessors selected the PSFs for each task but the HEP was solely adjusted by empirical data. This was possible as the model of the task was made with the same taxonomy (i.e., classification scheme) described in CREAM and used in MATA-D: a set of 53 variables including performance shaping factors, cognitive functions and human execution errors.

Therefore, the quantitative analysis required the definition of the CPT for the network structure defined in Section 4.2. The conditional probability tables of children nodes were computed as relative frequencies gathered from empirical data found from the MATA-D (Multi-Attribute Technological Accidents Dataset (MATA-D) (Moura et al., 2020, Moura et al., 2016). This relies on the interpretation that the relationship between human errors and their influencing factors in FPSO/FSOs operations are equivalent to those observed in the industrial accidents included in the dataset. MATA-D was selected as the main empirical source of data for three main reasons:

1. it provides dependency between human errors and performance shaping factors;
2. it contains data from industries with equivalent level of socio-technical complexity as FPSOs/FSOs;
3. it allows to incorporate lessons from different industries rather than waiting for the reoccurrence of similar accident patterns (Morais et al., 2020).

Two nodes had different data sources. Node 9 (droplets from flare) relates to a specific design failure that leads to droplets falling from flare (a potential ignition source). Although design failure data from MATA-D could have been used, it was decided to use more specific information regarding flares from the UK offshore hydrocarbon releases database (HSE, 2020b). Node 10 (consequence node), which represents the possible consequences of having flammable gas above safe limits in installations have variable states (*fire*, *emergency shut-down* and *no-consequence*) that cannot be related to any variable available in the MATA-D. Thus, specific data from similar offshore installations was used. The data for emergency shut-downs due to gas detectors activation during tank venting in FPSOs was obtained from near-misses investigations (obtained during safety audits) and incident reported to the Brazilian regulator (ANP, 2020a). The information about frequency of droplets from flare in FPSOs was obtained

from (HSE, 2020b), and ignition followed by fire in FPSO during tank venting was obtained from conference papers describing investigations of similar occurrences in UK North Sea FPSOs (Alan Keith et al., 2012, Pursel et al., 2016a, Pursel et al., 2016b).

Root nodes prior probabilities are obtained straightforward from the MATA-D, as they are not conditioned by any other nodes. However, the calculation of conditional probability tables for children nodes is more complex and nodes with many parents require an impracticable time to be assessed manually. Thus, a dedicated script code was developed to automatize the procedure of collecting the combination of events from the database (see *data collection code* in Appendix D).

The procedure of how the data in MATA-D translates into number in conditional probability tables is based on the fact that prior probabilities are expressed in terms of K events out of N trials. For example, in Table 3-4, the PSF *design failure* was observed (i.e., true) in 157 events out of 238 accidents, thus the resulting relative frequency of 0.66 was translated into prior probability distribution of design failure being true (0.66) and false (1 – 0.66). As the distribution of this root node does not lack data, it is defined in the model as a discrete node.

Table 3-4. Prior probabilities of nodes PSF 1, 8D and 9, all discrete root nodes

Design failure from MATA-D	FALSE	0.34
	TRUE	0.66
Node PSF 8D (equipment failure) from MATA-D (Moura et al., 2020)	FALSE	0.44
	TRUE	0.56
Node PSF 9 (Droplets from flare) from (HSE, 2020b)	FALSE	9.97×10^{-1}
	TRUE	3.0×10^{-3}

Table 3-5 shows the conditional probability table of subtask 3.1.A – where the assessors of the qualitative analysis identified that the operator could miss an observation, triggered by the PSFs *incomplete information*, *inadequate task allocation*, and *insufficient skills*. For instance, the combination #1 in the CPT represents the events in MATA-D where none of the PSFs was observed (i.e., false). According to MATA-D this context combined with the cognition failure *observation missed* occurred in only 8 out of 238 accidents, while the same context without *observation missed* occurred in 59 out of 238 accidents. The respective relative frequencies in MATA-D are 0.03 and 0.25, but in terms of prior probabilities these numbers are expressed as 0.12 and 0.88 as probabilities range from 0 to 1 (in other words the numbers

0.03 and 0.25 were normalised within the range 0 to 1, thus the probability of combination #1 when *observation missed* is *false* is equal to 0.88 and the probability of combination #1 when *observation missed* is *true* is equal to 0.12). As all the combinations are complete for this specific CPT, this node is defined as a discrete node in the model.

Table 3-5. Prior probabilities in CPT for subtask 3.1.A (variable: *observation missed*), a discrete child node

	Combination #1	Combination #2	Combination #3	Combination #4	Combination #5	Combination #6	Combination #7	Combination #8
Incomplete information	false	false	false	false	true	true	true	True
Inadequate task allocation	false	false	true	true	false	false	true	True
Insufficient skills	false	true	false	true	false	true	false	True
Observation Missed - FALSE	0.88	0.84	0.91	0.87	0.60	0.50	0.73	0.67
Observation Missed - TRUE	0.12	0.16	0.092	0.13	0.40	0.50	0.28	0.33

Table 3-6 describes the CPT of subtask 3.3.A, where the assessors defined *incorrect prediction* as the potential cognition failure for the task, in a context where the main PSFs were *cognitive bias*, *management problem*, *insufficient knowledge*, and *adverse ambient conditions*. Table 3-6 shows the frequency this same context occurred in accidents recorded in MATA-D. Differently from CPTs shown in Table 3-4 and Table 3-5, some combinations of states of these variables do not have any reported event within all 238 accidents in the dataset (e.g. combinations #8, #10, #12, #14 and #16). Therefore, as the lack of possible combinations events in MATA-D is interpreted as missing data rather than impossible events, the incomplete combinations were replaced by zero-to-one intervals [0,1]. As this node contains intervals, it was defined as a credal node. For this model, the majority of children nodes with more than four parent nodes had to be defined as credal nodes.

Table 3-6. Prior probabilities in CPT for subtask 3.3A (variable: *incorrect prediction*), a credal child node

	Combination #1	Combination #2	Combination #3	Combination #4	Combination #5	Combination #6	Combination #7	Combination #8	Combination #9	Combination #10	Combination #11	Combination #12	Combination #13	Combination #14	Combination #15	Combination #16
Cognitive bias	false	false	false	false	false	false	false	false	true	true	true	true	true	true	true	true
Management problem	false	false	false	false	true	true	true	true	false	false	false	false	true	true	true	true

Insufficient knowledge	false	false	true	true	false	false	true	true	false	false	true	true	false	false	true	true
Adverse ambient conditions	false	true	false	true	false	True	false	true	false	true	false	true	false	true	false	true
Incorrect prediction FALSE	0.99	0.93	0.91	1.0	1.0	1.0	0.88	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
Incorrect prediction TRUE	0.01	0.07	0.09	0.0	0.0	0.0	0.12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The complete CPTs for all nodes can be found on the Appendix E. More details on how to convert the relative frequencies from MATA-D to the CPTs can be accessed on (Morais et al., 2020). OpenCossan software was used to evaluate the models. The analyses were performed on a machine with x16 Intel Xeon CPU ES-2679 v2 @2.50GHz and 252.4Gb RAM. For model #1, the computational time for the predictive analysis was in average 3.2 hours/node. The diagnostic analysis required 2.5 hours per queried node. For model #2, the computational time for predictive analysis and diagnostic analysis was in average 0.74 hours/node and 0.64 hours/node, respectively. If the same analysis is performed on a middle-range laptop it requires 20 and 11 hours/node to run predictive analysis of model #1 and for model #2, respectively. Diagnostic analysis would have required 9 and 5 hours per query of model#1 and for model #2, respectively. The algorithm of variable elimination has been used in all the analysis.

4.4. Results

4.4.1. Predictive analysis

The results of the predictive analysis are presented in Table 3-7 for model #1, Figure 3-15 and Figure 3-16 for the model #1 and Figure 3-17 and Figure 3-18 for model #2, while some possible diagnostic analysis are presented from Table 3-8 and from Figure 3-19. In Table 3-7 the posterior probabilities are presented for all variables' states, which are TRUE and FALSE for the nodes related to tasks and performance shaping factors, and states *no consequence*, *emergency shutdown* and *fire* for the node related to the consequence event. The posterior probabilities of discrete nodes are point values and those of credal nodes are intervals. For instance, the probability that *subtask 3.1.A (check wind speed and direction)* is *true* is a point value (a crisp probability), as the lower and upper bounds are the same. For the *subtask 3.3.A (check lightning)* the result in state *true* is represented by an interval. Another aspect about the binary credal nodes, is that the lower bound of the false state and the upper bound of the

true state sum up to one (as well as the lower bound of the true state and the upper bound of false state). In the credal node ‘consequence’, with three states, the unity is achieved if summing up two lowest states of the lower bound with the highest state of the upper bound, as well as summing up the two lowest states of the upper bound with the highest state of the lower bound.

The state TRUE of each binary node represents the probability of an error has been observed, and the state FALSE probability that an error has not been observed. Thus, for the subtask 3.1A probabilities can be interpreted as follows: for every thousand times operators read an instrument to check wind speed and direction, chances are that in 159 times they misread it. Similarly, for the subtask 3.3A: for every thousand times operators check the weather to predict if lightning is going to occur, between 34 and 42 times they incorrectly predict it. The distinction between results for discrete and credal nodes can be better visualised in Figure 3-15, which depicts the true states of trigger, control, mitigation and risk event nodes and Figure 3-16 which depicts all the three states of consequence node.

Comparing the results obtained from models #1 and #2 reveals smaller intervals in model #2 (especially tasks 3A, 6ABCD and 7A). The majority of model #2 results lie inside the intervals of model #1 (except for the subtasks assigned with different human error modes, such as subtasks 6.1A, 6.3B and 6.2C). Furthermore, it was noticed that the majority of probability intervals comprises the frequencies obtained directly from MATA-D (Moura et al., 2016). For instance, the ‘wrong type’ error mode has the relative frequency of 11.80% in MATA-D, while the posterior probability of task 5A (assigned with the same error mode) presents a probability interval between 10.08% to 17.82%. The predicted results might represent the interaction effect between human errors and PSFs, depicting the uncertainty of a certain type of human error occurring under a specific context (e.g. *wrong type* has a relative frequency of 11.80% in all 238 accident events in MATA-D, however, 10.08% – 17.82% would be the imprecise probability for it happening under the context of the PSFs *equipment failure, design failure, observation missed, inadequate plan* and *action in wrong place* occurring altogether). When inference is performed, the interval of posterior probabilities depicts the inputs you do not have enough data.

Table 3-7. Prediction of posterior probabilities in all variable states (model #1)

Event	State	Lower bound	Upper bound
TRIGGERS			
Task 2A (<i>observation missed</i>)	FALSE	0.83	0.84
	TRUE	0.16	0.17
CONTROL BARRIERS			

Task 3A <i>(inadequate plan)</i>	FALSE	0.66	0.92
	TRUE	0.08	0.34
Subtask 3.1A <i>(observation missed)</i>	FALSE	0.84	0.84
	TRUE	0.16	0.16
Subtask 3.2A <i>(observation missed)</i>	FALSE	0.82	0.83
	TRUE	0.17	0.18
Subtask 3.3A <i>(incorrect prediction)</i>	FALSE	0.96	0.97
	TRUE	0.034	0.04
Task 4A <i>(action in wrong place)</i>	FALSE	0.60	0.71
	TRUE	0.29	0.40
RISK EVENT			
Task 5A <i>(execution of wrong type)</i>	FALSE	0.82	0.90
	TRUE	0.10	0.18
MITIGATION BARRIERS			
Task 6 ABCD <i>(action in wrong place)</i>	FALSE	0.37	0.84
	TRUE	0.16	0.63
Subtask 6.1A <i>(action in wrong place)</i>	FALSE	0.62	0.62
	TRUE	0.38	0.38
Subtask 6.2C <i>(action in wrong place)</i>	FALSE	0.62	0.62
	TRUE	0.38	0.38
Subtask 6.3B <i>(action in wrong place)</i>	FALSE	0.58	0.58
	TRUE	0.42	0.42
Task 7A <i>(action performed at wrong time)</i>	FALSE	0.49	0.94
	TRUE	0.06	0.51
Task 7.1C <i>(observation missed)</i>	FALSE	0.83	0.86
	TRUE	0.14	0.17
Task 7.2C <i>(action performed at wrong time)</i>	FALSE	0.85	0.86
	TRUE	0.14	0.15
Task 7.3B <i>(action performed at wrong time)</i>	FALSE	0.58	0.58
	TRUE	0.42	0.42
CONSEQUENCE			
Node 10 <i>(consequence of hazard event)</i>	No consequence	0.8658	0.9999
	Emergency shut-down (ESD)	6.211×10^{-5}	0.1342
	Fire	7.908×10^{-8}	5.669×10^{-7}

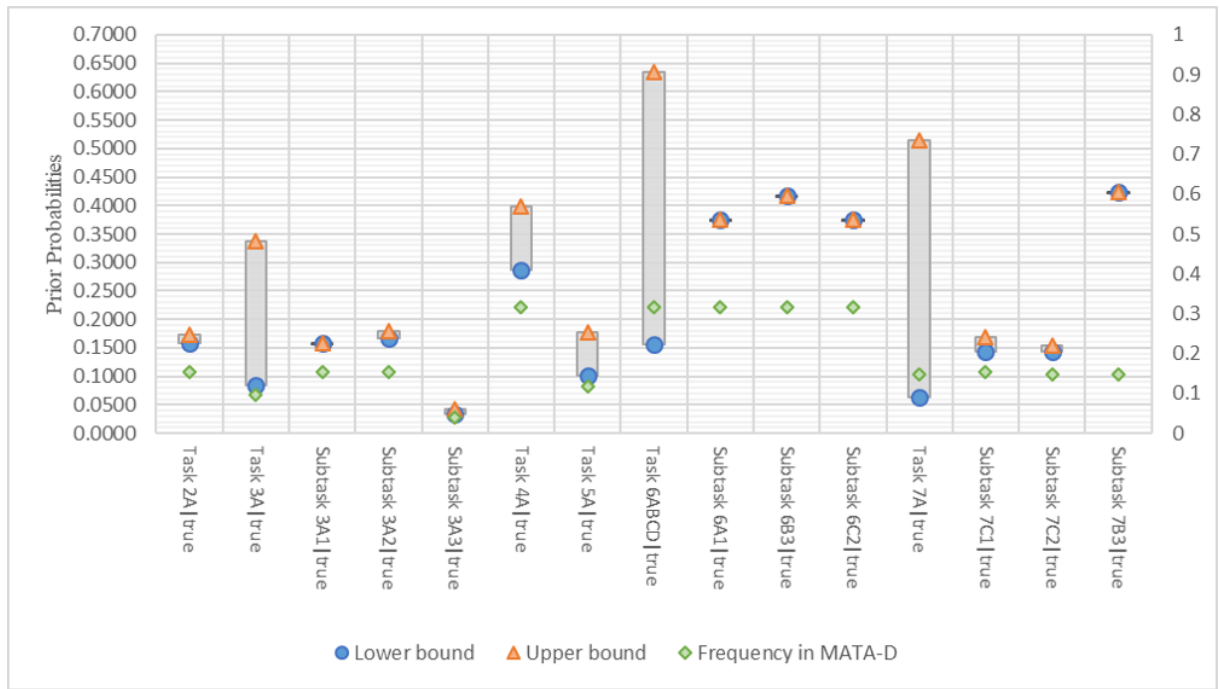


Figure 3-15. Point and interval posterior probabilities for the cargo venting human reliability model #1

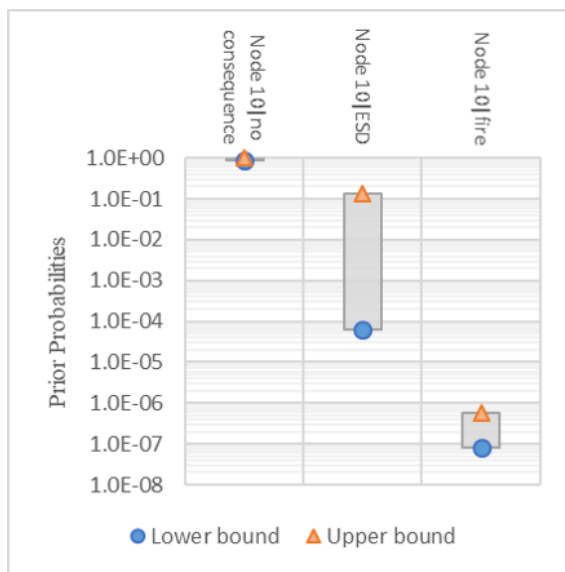


Figure 3-16. Posterior probabilities for the three states of the consequence node of model #1

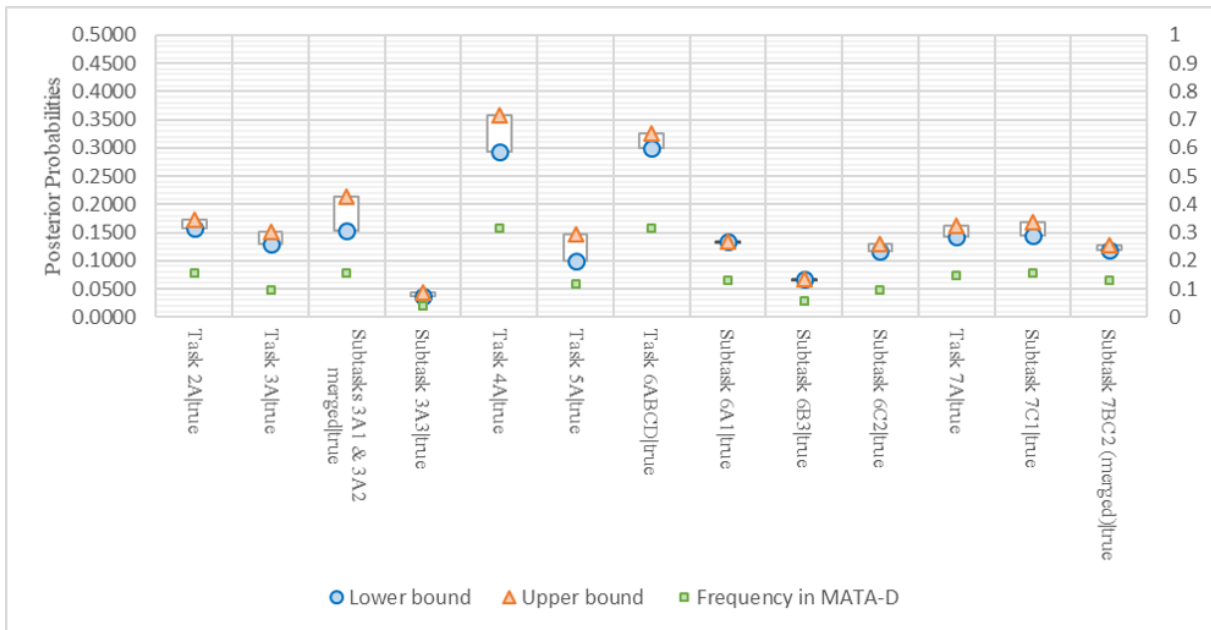


Figure 3-17. Point and interval posterior probabilities for the cargo venting human reliability model #2

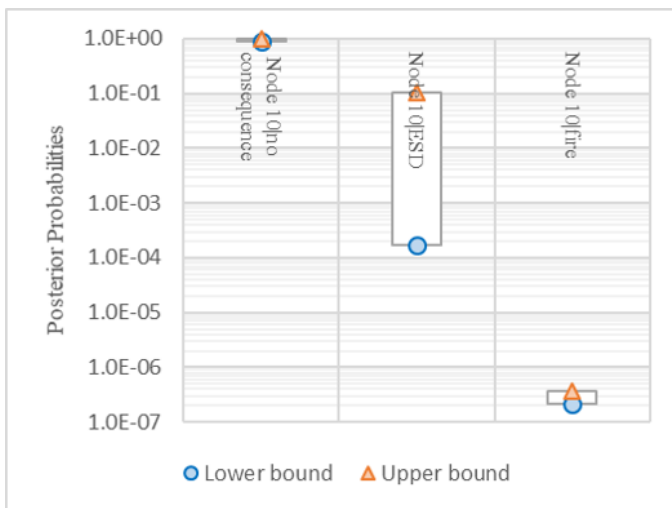


Figure 3-18. Posterior probabilities of three states of consequence node in model #2

4.4.2. Diagnostic analysis

The ability to provide diagnostic analysis is one of the key features of Credal Network allowing the simulation of many scenarios. This allows to track and quantify the most important relations for each node and assisting in the identification of efficient risk reduction measures. The diagnostic analysis – also known as *sensitivity analysis* – is performed by introducing evidence into a node (i.e. observation) and querying another node of interest. For briefly, only the results directed to the risk and consequence events of the human reliability model, and to other findings that help explaining the methodology are presented. The diagnostic analysis for all tasks can be assessed in Appendix G (model #1) and Appendix H (model #2).

The objective here is to assess which tasks and PSFs are more relevant in triggering an operator error in the critical task of opening the cargo venting valve (task 5A). Figure 3-19 shows the sensitivity analysis for *task 5A* of model #1 to preceding tasks while Figure 3-20 presents the sensitivity analysis with respect to the PSFs. The probability values of the sensitivity analysis of task 5A are reported in Table 3-8. Using the criteria proposed in the methodology section, the most impacting task is task 2A (verify pressure) and the most impacting PSF is incomplete information (technology factor).

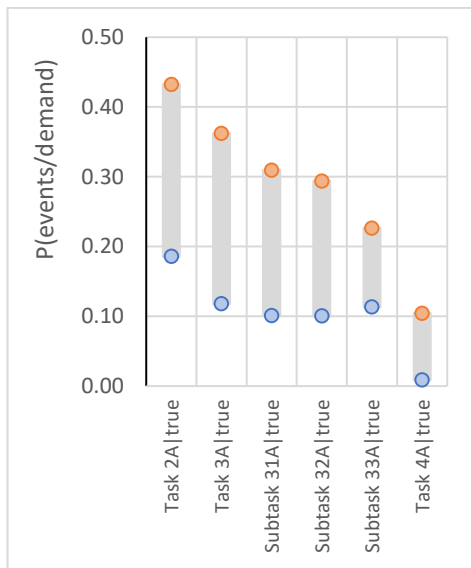


Figure 3-19. Task 5A|true - sensitivity to tasks (model #1)

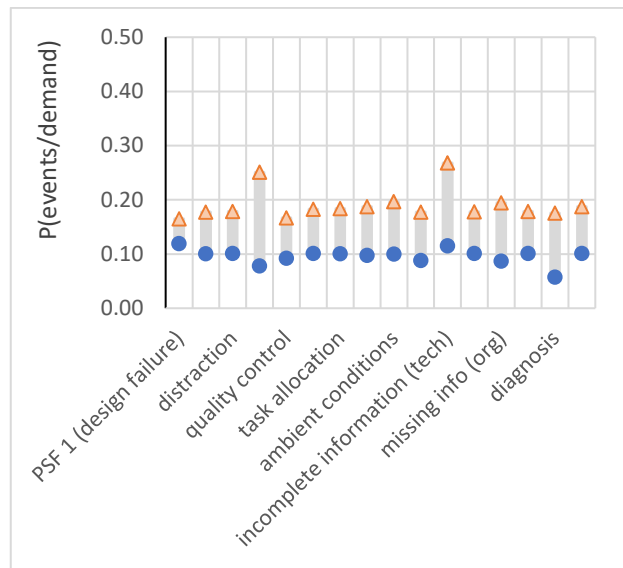


Figure 3-20. Task 5A|true - sensitivity to PSFs (model #1)

Table 3-8. Sensitivity analysis of task 5A to other tasks and PSFs in model #1.

Task 5A true (query)		
Evidence added to:	Lower bound	Upper bound
Tasks		
Task 2A true	0.1859	0.4322
Task 3A true	0.1182	0.3621
Subtask 31A true	0.1009	0.3092
Subtask 32A true	0.1006	0.2936
Subtask 33A true	0.1136	0.2264
Task 4A true	0.0090	0.1040
Performance shaping factors		
Node1(Design) True	0.1190	0.1649
Bias true	0.1005	0.1775
Distraction true	0.1008	0.1782
Maintenance True	0.0782	0.2506
Quality True	0.0921	0.1667
Management True	0.1010	0.1826
Task True	0.1003	0.1836
Knowledge True	0.0972	0.1871
Ambient True	0.0996	0.1962
Procedure True	0.0880	0.1769

Incomp Info (tec) True	0.1147	0.2677
Communication True	0.1009	0.1779
Missing Info (org) True	0.0871	0.1945
Priority True	0.1008	0.1782
Diagnosis True	0.0570	0.1754
Skills True	0.1009	0.1875

An interesting finding to showcase the impact of missing data and the choice of criteria to interpret the diagnostic analysis is presented in Figure 3-21, the sensitivity of subtask 3.2A to PSFs in model #1. The wider interval in PSF *ambient conditions* shows its high uncertainty due to incomplete data regarding its interactions with the human error mode of subtask 3.2A. The result suggests that if poor ambient conditions occur, it has the potential to be the most impacting factor to trigger human error. On the other hand, if other criteria were used to benefit more certain intervals, a possible candidate of most impacting PSF could be insufficient skills, as this factor has the highest lower bounds.

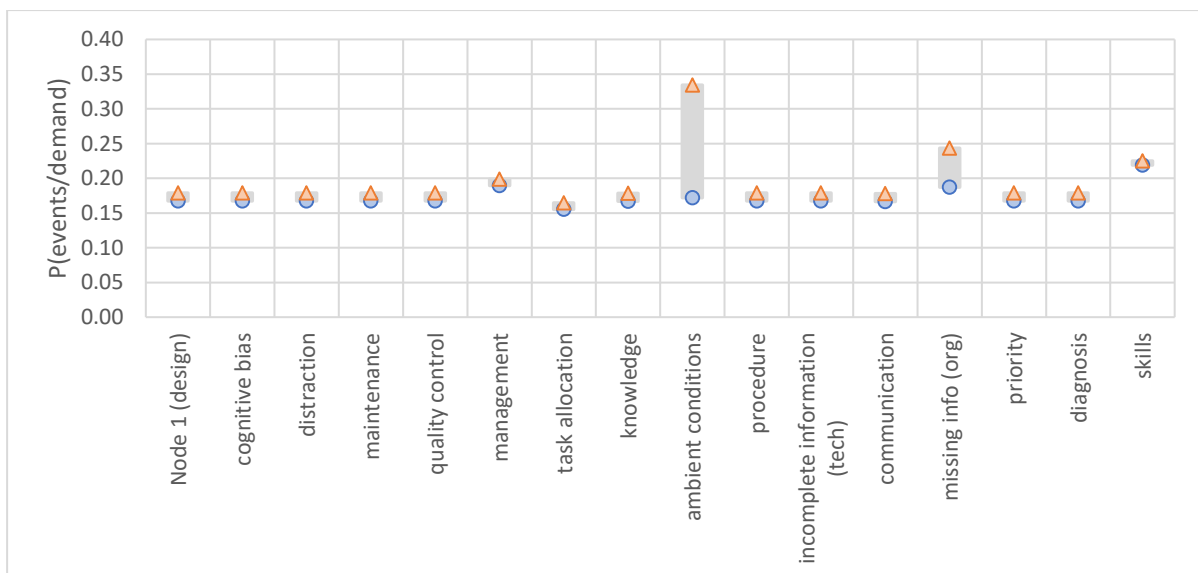


Figure 3-21. Node 3.2A/true - sensitivity to PSFs

Figure 3-22 to Figure 3-27 show diagnostic analysis for tasks 3A, 6ABCD and 7A, which are linked to subtasks, respectively. Their subtasks are the main difference between both models (i.e. assignment of different human error modes). What stands out in these figures is the difference in uncertainty between results from model #1 and #2.

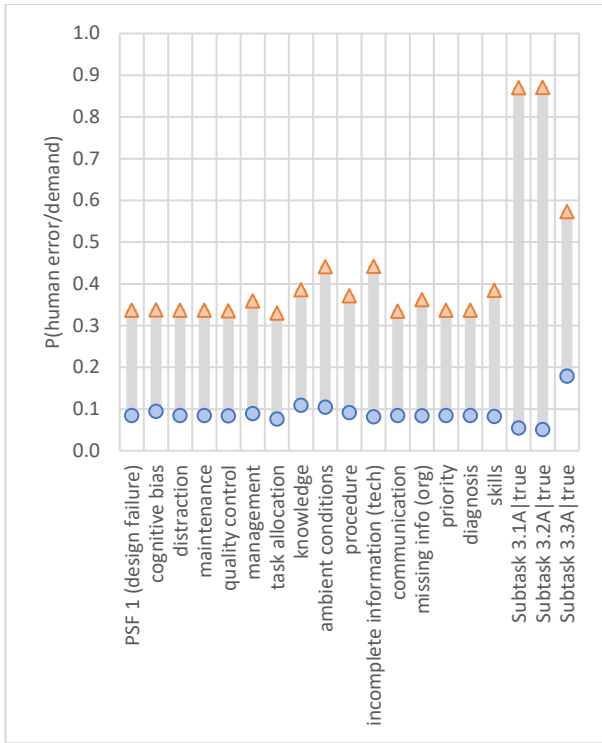


Figure 3-22. Node 3A/true - sensitivity to PSFs and subtasks 3.1A, 3.2A & 3A3 (model #1)

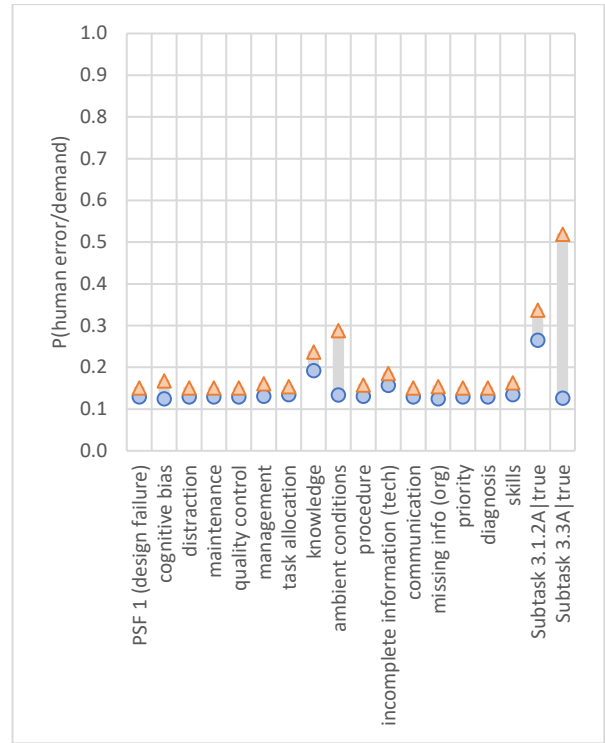


Figure 3-23. Task 3A/true sensitivity to PSFs and subtasks 3.1.2A and 3.3A (model #2)

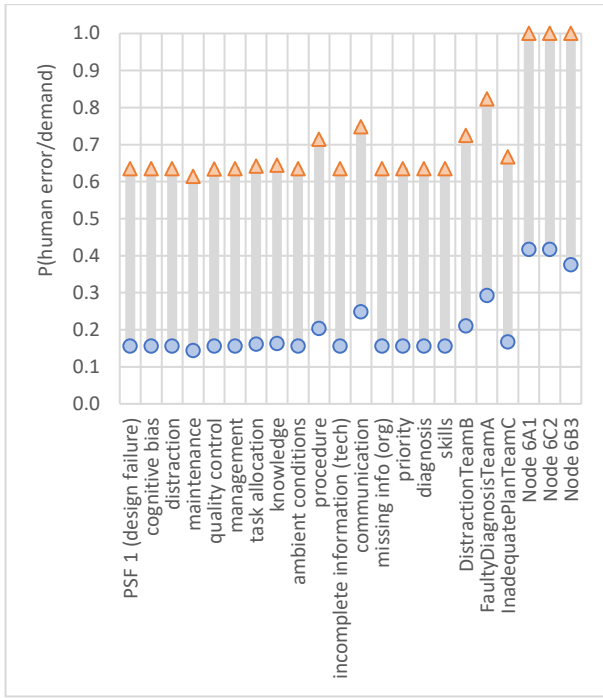


Figure 3-24. Task 6ABCD/true sensitivity to PSFs and subtasks 6.1A, 6.2C, 6.3B (model #1)

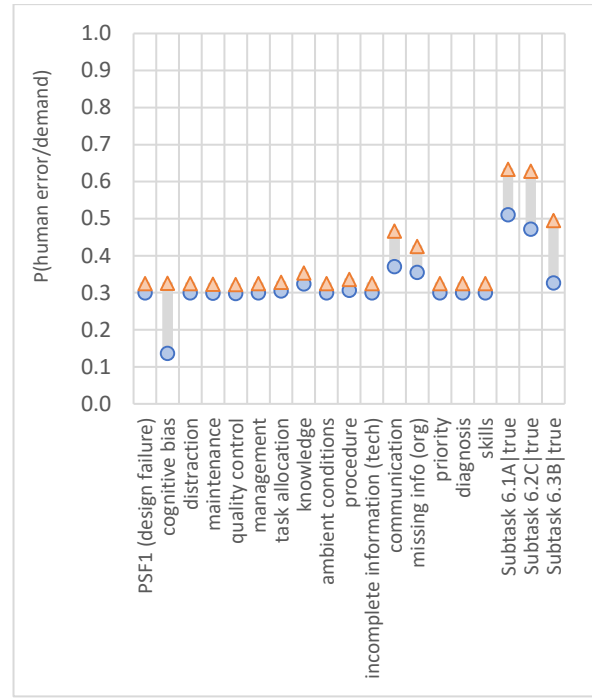


Figure 3-25. Task 6ABCD/true - sensitivity to PSFs and subtasks 6.1A, 6.2C & 6.3B (model #2)

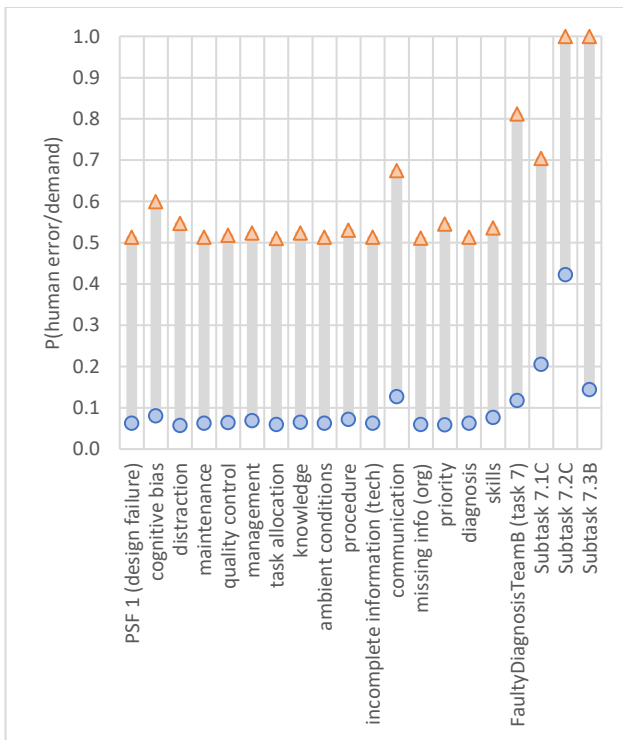


Figure 3-26. Task 7A/true sensitivity to PSFs and subtasks 7.1C, 7.2C and 7.3B (model #1)

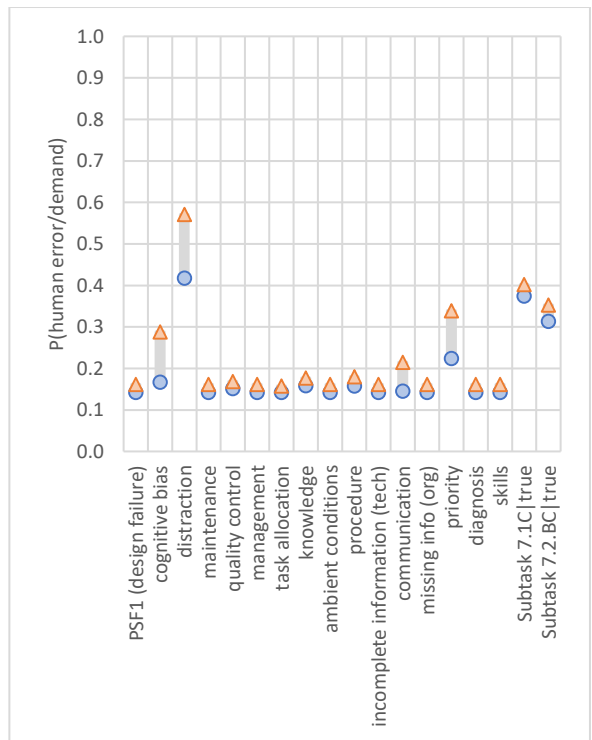


Figure 3-27. Task 7A/true - sensitivity to PSFs and subtasks 7.1C & 7.2BC (model #2)

Table 3-9 presents diagnostic analysis of the impact of tasks and PSFs in the consequence events of emergency shutdown (ESD) and fire during cargo venting operation in FPSOs/FSOs.

Figure 3-28 is the graphical representation of intervals for ESD sensitivity, represented in logarithmic scale to facilitate the analysis of lower bounds. Figure 3-29 shows the fire sensitivity to tasks and PSFs in log scale. By pairwise comparison of the two most impacting factors for fire to happen, task 5A (wrong action of opening the valve) and PSF 9 ('droplets from flare'), it is clear that 'droplets from flare' is the most impacting factor as, according to the criteria, the intervals do not overlap and 'droplets from flare' has the highest lower bound.

Table 3-9. Sensitivity analysis to tasks and PSFs of ESD and fire occurring as a consequence

Evidence on node	Node 10 ESD queried <i>P(event/days)</i>		Node 10 fire queried <i>P(event/days)</i>	
	Lower bound	Upper bound	Lower bound	Upper bound
Performance Shaping Factors				
Node PSF 1 (design failure)	6.97x10 ⁻⁵	0.13	8.67 x 10 ⁻⁸	5.50 x 10 ⁻⁷
Node PSF 9 (droplets from flare)	9.47x10 ⁻⁶	0.13	2.89 x 10 ⁻⁵	2.07 x 10 ⁻⁴
Node PSF 8D (equipment failure)	0	0.19	0	0
Cognitive bias	1.37x10 ⁻⁴	0.14	6.20 x 10 ⁻⁸	5.71 x 10 ⁻⁷
Distraction	9.56x10 ⁻⁵	0.13	7.00 x 10 ⁻⁸	5.75 x 10 ⁻⁷
Maintenance failure	5.63x10 ⁻⁵	0.19	4.37 x 10 ⁻⁸	6.20 x 10 ⁻⁷
Inadequate quality control	5.75x10 ⁻⁵	0.13	6.74 x 10 ⁻⁸	4.93 x 10 ⁻⁷
Management problem	6.77x10 ⁻⁵	0.14	7.77 x 10 ⁻⁸	5.78 x 10 ⁻⁷
Inadequate task allocation	5.64x10 ⁻⁵	0.15	7.41 x 10 ⁻⁸	5.70 x 10 ⁻⁷
Insufficient knowledge	6.02x10 ⁻⁵	0.14	7.67 x 10 ⁻⁸	5.95 x 10 ⁻⁷
Adverse ambient conditions	6.60x10 ⁻⁵	0.14	7.95 x 10 ⁻⁸	6.17 x 10 ⁻⁷
Inadequate procedure	6.08x10 ⁻⁵	0.13	5.48 x 10 ⁻⁸	5.24 x 10 ⁻⁷
Incomplete information (technology)	8.68x10 ⁻⁵	0.20	8.58 x 10 ⁻⁸	9.43 x 10 ⁻⁷
Communication failure	1.40x10 ⁻⁴	0.14	4.15 x 10 ⁻⁸	4.80 x 10 ⁻⁷
Missing information (organisation)	8.96x10 ⁻⁵	0.14	6.27 x 10 ⁻⁸	6.83 x 10 ⁻⁷
Priority error	7.45x10 ⁻⁵	0.13	7.43 x 10 ⁻⁸	5.78 x 10 ⁻⁷
Faulty diagnosis	4.80x10 ⁻⁵	0.12	5.18 x 10 ⁻⁸	5.47 x 10 ⁻⁷
Insufficient skills	7.40x10 ⁻⁵	0.14	7.57 x 10 ⁻⁸	5.92 x 10 ⁻⁷
Distraction of team B	5.17x10 ⁻⁵	0.14	5.69 x 10 ⁻⁸	5.24 x 10 ⁻⁷
Faulty diagnosis of team A	3.69x10 ⁻⁵	0.15	3.65 x 10 ⁻⁸	4.88 x 10 ⁻⁷
Faulty diagnosis of team B	1.01x10 ⁻⁴	0.14	2.73 x 10 ⁻⁸	4.94 x 10 ⁻⁷
Inadequate plan of team C	6.29x10 ⁻⁵	0.14	7.18 x 10 ⁻⁸	5.60 x 10 ⁻⁷
Tasks and subtasks				

Task 2A true	1.34x10 ⁻⁴	0.31	1.19 x 10 ⁻⁷	1.60 x 10 ⁻⁶
Task 3A true	1.26x10 ⁻⁴	0.27	1.06 x 10 ⁻⁷	1.38 x 10 ⁻⁶
Subtask 31A true	8.79x10 ⁻⁵	0.20	7.86 x 10 ⁻⁸	1.1178 x 10 ⁻⁶
Subtask 32A true	7.02x10 ⁻⁵	0.20	7.44 x 10 ⁻⁸	9.95 x 10 ⁻⁷
Subtask 33A true	1.14x10 ⁻⁴	0.16	8.14 x 10 ⁻⁸	7.55 x 10 ⁻⁷
Task 4A true	1.61x10 ⁻⁵	0.07	6.89 x 10 ⁻⁹	2.65 x 10 ⁻⁷
Task 5A true	5.10x10 ⁻⁴	0.84	3.90 x 10 ⁻⁷	1.98 x 10 ⁻⁵
Task 6ABCD true	0	0.17	0	0
Subtask 6.1A true	0	0.16	0	4.31 x 10 ⁻⁷
Subtask 6.2C true	0	0.16	0	4.31 x 10 ⁻⁷
Subtask 6.3B true	0	0.16	0	4.31 x 10 ⁻⁷
Task 7A true	6.72x10 ⁻⁴	0.14	0	0
Subtask 7.1C true	1.92x10 ⁻⁴	0.14	4.66 x 10 ⁻⁸	4.91 x 10 ⁻⁷
Subtask 7.2C true	3.79x10 ⁻⁴	0.14	0	3.70 x 10 ⁻⁷
Subtask 7.3B true	1.32x10 ⁻⁴	0.14	0	5.17 x 10 ⁻⁷

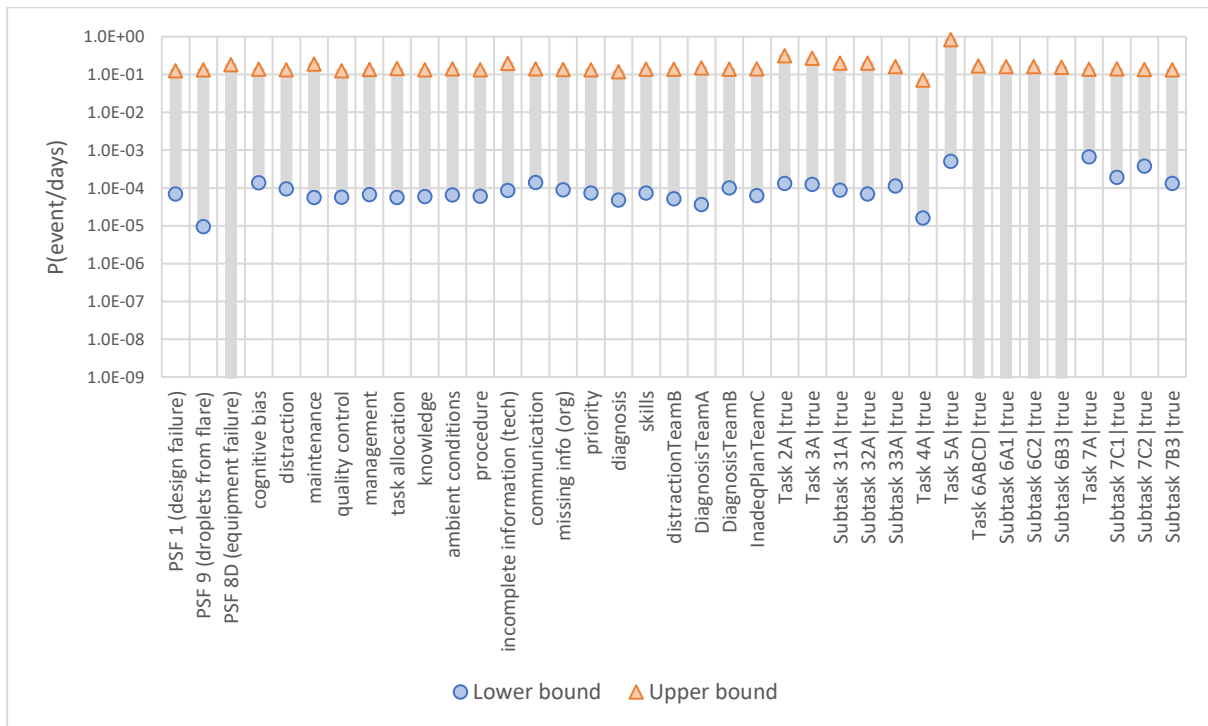


Figure 3-28. Sensitivity Node 10/ESD (in log scale).

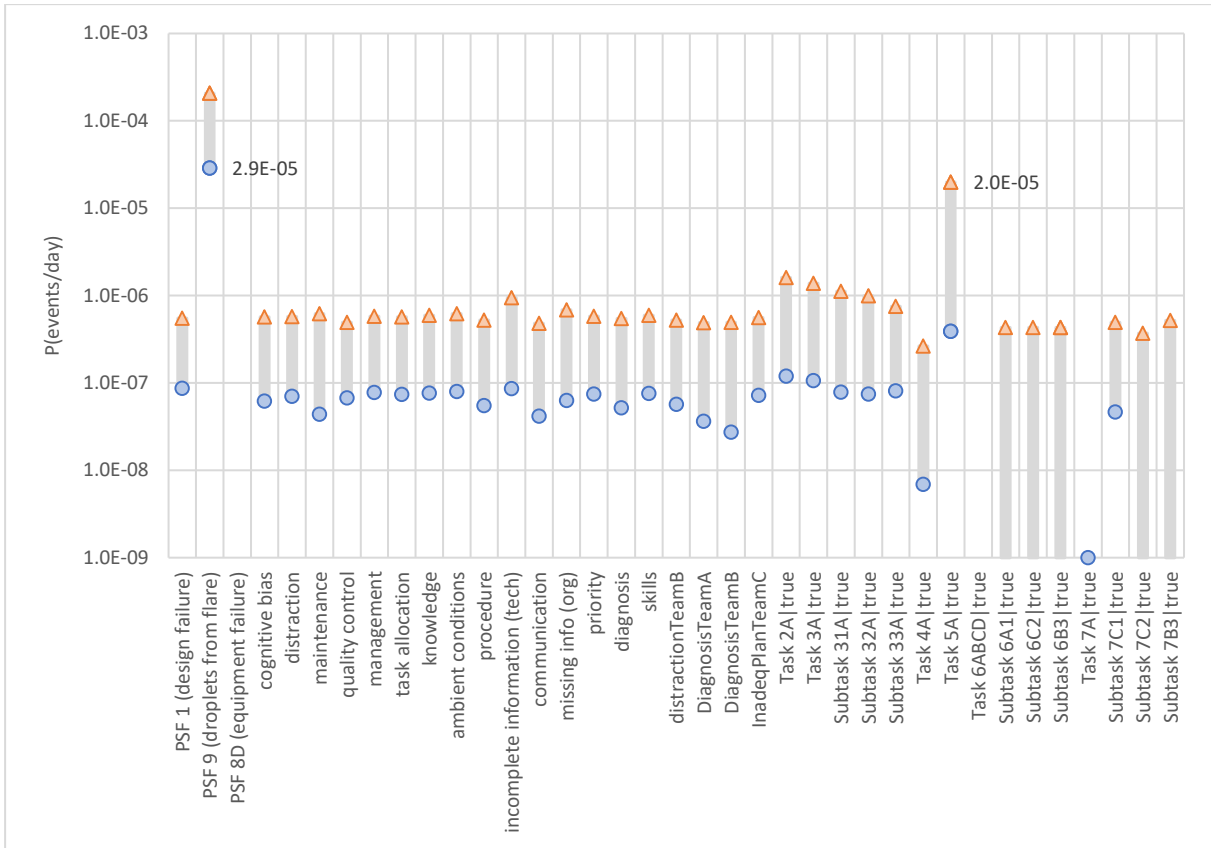


Figure 3-29. Node 10/Fire - sensitivity to tasks and PSFs (log scale)

Table 3-10 presents a summary of the most impacting factors for each task and subtask in model #1, where the factors in bold are those that are also the most impacting factors in model #2. The used criteria to select the most critical factors for each task, in order to either control the effect on a specific node or to reduce its uncertainty was presented in the methodology section.

Table 3-10. Summary of most influencing factors in tasks of model #1 and #2 (in bold where both models agree)

Node	Most influencing tasks or performance shaping factors for model #1	Most influencing tasks or performance shaping factors for model #2
Task 2A true	PSF incomplete information (tech factor)	PSF incomplete information (tech factor)
Task 3A true	Subtask 3.1A	Subtask 3.3A
Subtask 3.1A <i>(equals to 3.1.2A in model #2)</i>	PSF incomplete information (tech factor)	PSF ambient conditions, followed by incomplete information (tech factor)
Subtask 3.2A <i>(equals to 3.1.2A in model #2)</i>	PSF adverse ambient conditions (org factor)	
Subtask 3.3A	PSF adverse ambient conditions	PSF adverse ambient conditions

Task 4A	PSF faulty diagnosis	PSF faulty diagnosis
Task 5A true	Task 2A (verifying pressure, cognitive failure of missing an observation), followed by PSF of incomplete information (technological factor)	Task 2A
Task 6ABCD	Subtask 6.1A (request PTW, tied up with subtask 6.2C, analyse area to issue PTW). Both are actions out of sequence, but in different teams.	Subtask 6.1A
Subtask 6.1A	Faulty diagnosis of team A	Communication failure
Subtask 6.2C	Subtask 6.1A , followed by the PSF of faulty diagnosis of team A	Subtask 6.1A
Subtask 6.3B	Distraction of team B, closely followed by the PSF inadequate procedure	Communication failure
Task 7A	Subtask 7.2C (inform changes in gas detection to team A)	Distraction
Subtask 7.1C	Cognitive bias of team C	Cognitive bias
Subtask 7.2C (=subtask 7.2BC in model#2)	Communication failure	Cognitive bias
Subtask 7.3B (=subtask 7.2BC in model#2)	Faulty diagnosis of team B	
PSF 8D	Maintenance failure	Maintenance failure
PSF 9	<i>N.B. no impact from other PSFs of CREAM taxonomy</i>	
Node 10 ESD	Task 5A (opening or closing the cargo venting valve, wrong type execution error)	Task 5A
Node 10 fire	PSF 9 (droplets from flare)	PSF 9 (droplets from flare)

4.6. Discussion

The case study has shown the applicability of credal networks to analyse the human reliability by performing predictive and diagnostic studies in presence of missing data. It was noted that besides the fact that the cargo venting task occurs in an error prone context, the model also shows that even if the human failure events occur the risk to safety and financial loss is very low (see Figure 3-16).

It has been observed that, the majority of relative frequencies from MATA-D (Moura et al., 2016) lies inside the posterior probabilities' intervals obtained using credal networks. This can be interpreted as nominal HEPs being adjusted by their empirical relations with the selected PSFs, in a different methodology than proposed by previous studies (Kim et al., 2018).

Nominal HEPs would be the relative frequencies in MATA-D and empirical relations with PSFs provided by credal network. In practice, this would mean that while an expert is still needed for the qualitative task of selecting the PSFs, the proposed methodology has the potential to replace or at least complement the contribution from experts on the quantitative analysis of traditional HRA methods, as they would no more be needed to define the strength of PSF influence. The proposed methodology also provides the adjustment of upper and lower bound empirically.

A possible explanation for the quantified human error probabilities (HEP) associated to the model#1 tasks 4A, 6ABCD, 7A, and subtasks 6A1, 6B3, 6C2, and 7B3 being higher than typical HRA method's numbers (e.g. 10^{-4} to 10^{-2}) is because these HEPs do not refer to nominal HEPs. In traditional HRA methods such as THERP, all of the estimated HEPs in the data tables provided are nominal HEPs, which are usually modified upward after being adjusted by the effects of PSFs (Swain and Guttman, 1983). Conversely, the results of this study refer to HEPs already adjusted by the PSFs solely driven by empirical data (i.e., the relations between PSFs and human errors in MATA-D). Another possible explanation for higher HEP is that this model have accounted for the PSFs directly related in the context, without further propagating the antecedent-consequent model proposed by Hollnagel in CREAM (see the antecedent-consequents' table provided in the supplementary material). For example, according to the antecedent-consequent model, the PSF *Incomplete Information* has *inadequate procedure* and *design failure* as its antecedents. If the full antecedent-consequent links between PSFs are added, the HEPs decrease, as the more parent nodes we have connected to a child, the smaller its probability (this had happened on a previous model used, with standard Bayesian network and MATA-D (Morais et al., 2020)).

It was noted that the confidence in our results is often to the second digit, while the nominal HEPs of traditional HRA methods (e.g. HEART, THERP) provide estimates with larger error bounds (e.g., one order of magnitude between the 5th and the 95th percentiles in some cases). This fact might be explained for two main reasons. Firstly, because the results obtained in this study are related to the final HEP estimates after task-specific PSFs have been considered, while traditional HRA methods estimates are nominal HEPs where the uncertainty bounds include not only the random variability of individuals but also the presumed uncertainty of the analyst in the HRA process (Swain and Guttman, 1983). In our study we are proposing a methodology that does not need to account for the uncertainty of the analyst, which is one of the reasons why the estimates have skinner uncertainty bounds. Secondly, the uncertainty bounds of the nominal HEPs in the other methods were designed to predict many

different contexts, while in this study few specific PSFs were selected as the modellers knew the context from the documents used in task analysis.

This study has also shown how credal networks can be used to identify risk reduction measures of the human reliability model, by investigating the effect of each factor over each task. This may support reduction measures to decrease the risk of human error, fire and emergency shutdown during the cargo venting operation.

The proposed criteria for selecting the most impacting factors aims to support comparison between different interval probabilities, identifying which variable is most important. For instance, to decrease the chances of having a human error of ‘*wrong type*’ during the event of opening the cargo venting valve (task 5A), reduction measures should focus mainly on the verification of cargo tank pressure (task 2A). The most important technological factor is *incomplete information* (i.e. temporary interface failure where the information provided by the interface is incomplete, e.g. error messages, directions, warnings (Hollnagel, 1998)). The most important organisational factors is *maintenance failure* (i.e. missing or inappropriate management of maintenance leading to equipment not operational or indicators not working (Hollnagel, 1998)), although this factor would clearly benefit of further data collection to minimise its uncertainty. To decrease the chances of emergency shutdown due to cargo venting, the critical task to be improved is task 5A (opening or closing the cargo venting valve, execution error of wrong type). To reduce the chances of having fire as a consequence, the most important organisational factor to tackle according to this model are ‘droplets falling from flare’, possibly caused by design failure. The dependencies among variables should also be considered. For instance, in Figure 3-26 and Figure 3-27, it is possible that the imprecision of 7.2C derives entirely from the imprecision of 7.1C. Thus, further analysis would be required to fully understand the effect of both subtasks in task 7A.

Although it was clear that the criteria can be refined to reflect other decision-making style (for instance, some decision-makers might feel more comfortable to give higher value to more precise intervals), it is also recommended that a unique criterion is used by all decision-makers of the same organisation.

Consistent with the literature, this research found that different model structures – obtained in the qualitative part of the analysis – impact the quantification. The significant decrease of uncertainty in model #2 nodes is evidenced by the smaller intervals obtained. This is a consequence of the reduced number of unknown combinations in CPTs following the adoption of the synthetic idiom strategy, avoiding children nodes with the same CREAM taxonomy as their parent nodes. Furthermore, the analysis of the most impacting factors in

Table 3-10 have identified 63% of agreement between both models. Although model #1 can be used without such simplification, using underlying method relationship provides a strategy to reduce the uncertainty and computational time of the model without significantly impairing the accuracy of the results.

A final reminder about the model is that the probabilities of occurrence refer to the type of error mode and not directly to the task – for instance, task 2A results relates to the statistics of the variable ‘observation missed’ in MATA-D, and not to specific statistics of cargo operators failing to verify the cargo tanks pressure. This seems to be the main source of difference in models #1 and #2 (due to subtasks assigned with different human error modes). More importantly it means that the assessor’s opinion during the safety critical task analysis directly influences the results (as they assign human error and PSFs to tasks), and that it is possible to validate or update the model if human performance data is collected from cargo venting operation in FPSOs and FSOs.

4.6. Further developments

This paper used human reliability analysis as an aid to investigate the risks between operational change and design change options. However, further studies could be undertaken, such as further comparing the risk result to the company’s risk matrix, or estimating the societal risk by projecting the risk found on the model on a F-N curve (fatal events frequency x number of fatalities per year).

Although the approach of modelling empirical data with credal network is a much-needed shift from conservative to realistic modelling, it is important to note that the methodology presented only considers interval probabilities for the nodes with missing data. However, input data with intervals can be used for all nodes if data are imprecise due to other reasons rather than sparse data, such as human subjects’ variability. Thus, it is suggested that credal networks and the methodology suggested in this paper is further applied to other types of HRA datasets, such as those obtained in a laboratory-based study or in a simulated control-room. The code is available in Open Cossan website, therefore other research groups can test their own data.

5. Conclusions

A novel methodology for assessing human reliability under uncertainty and lack of data has been presented. The proposed methodology accepts and embraces the variability of human reliability databases – including their missing data – as an intrinsic aspect of any science that relies on human behaviour. Credal networks as an extension of Bayesian networks have been proposed to characterise the available data without making unjustified assumptions. It is a necessary tool for data-driven human reliability methods and avoid expert opinion to fill incomplete information. This is not a statement to stop using methods that rely on expert judgement. Experts should still be needed to structure the qualitative part of the human reliability analysis, such as modelling the tasks and establishing a framework to classify human errors and performance shaping factors for each task.

Traditional human error reliability methods usually suggest human error nominal probabilities that are adjusted according to the selected performance shaping factors. Thus, depending on these factors and the strength of their influence defined by experts' judgement, the estimated human error probabilities have large variability (and as credible as the expert selected). The methodology proposed removes the need of experts' judgment for this quantification step of the human reliability analysis and therefore reducing the associated bias and variability. The methodology might be of interest to both risk assessors and decision-makers. To risk assessors because credal networks provide a rigorous framework to deal with sparse data and imprecision avoiding strong assumptions, resulting in a much-needed shift from conservative to realistic modelling. To decision-makers (e.g. manager, regulator) because it provides a more accurate and realistic decision-making tool (e.g. bounds of the estimations can be interpreted as the best and worst-case scenarios), and because they can decide if the quality of the results (given by the intervals) is satisfactory or more resources in collecting additional data are needed. In summary, the risk communication between risk assessors and managers has the potential to be improved by the transparency provided by using imprecise probability, being fairer to compare the risks between components and human reliability analysis and to allocate resources accordingly. The proposed approach allows to describe a variable with more than two states allowing the adaptation to other existing HRA methods with multiple states. In addition, model reduction using intuitive application of underlying relations based on the human reliability method such as CREAM is an effective approach for reducing the uncertain in the results and the computational costs.

The approach has been successfully applied to a real case from oil & gas offshore industry, where a human reliability model could provide support to decision-makers and depict the uncertainties inherent to human behaviour. The credal network model has been created by translating the critical task analysis sequential structure into a cause-consequence structure that depicts also control and mitigation barriers, well known in the oil & gas industry as a bow-tie structure. The methodology permits to analyse non-monotonic behaviour, allowing to capture more realistic performance shaping factors effects on human performance and detecting the features of the scenario most likely to contribute to initiate (or fail to recover from) an incident event. This study also demonstrates that human reliability analysis is able to support design and operational decisions. Oil & gas operations can be assessed through scientific methodologies – with the possibility to borrow empirical evidence from industries with similar task complexity.

Continued efforts are needed to make reliable tools more accessible to the human reliability community and accepted by industrial partners and regulators. This study has shown the importance of using probabilistic tools that accept and depict uncertainty and imprecision supporting the fully data-driven human reliability analysis.

Chapter IV: Minimising epistemic uncertainty by collecting new data

Overview

In this chapter, the problem of missing data is tackled by complementing the existing dataset, the MATA-D. As stated in *Figure 3-6. Suggested criteria for decision-making in sensitivity analysis of HRA* and at the *Conclusions* of Chapter 3, although imprecise probabilities might help decision-makers make decisions without all the necessary data, they might prefer to invest in collecting more data. The process of collecting new data for this dataset should be constant not only to decrease epistemic uncertainty in human reliability data but also to update models and reflect changes in human behaviour due to evolving technology and organisational arrangements.

Obtaining new accident reports to expand MATA-D is quite an easy task, as a large portion of major accidents are publicly available on the internet. However, the difficulty resides in finding trained experts available to read and classify the reports against the CREAM taxonomy. Reading and classifying one whole accident report is a time-consuming process, which delays the learning-from-accident process.

For this reason, the third part of this research proposes an automated approach as a new collection methodology. The machine-learning approach developed is able to train the computer on a predefined classification scheme (taxonomy), which will be called the *virtual human factors classifier*. The machine is trained according to previously labelled accident reports by human experts.

The natural language processing (NLP) approach used has been tested as soon as the first preliminary report from the accident with the Boeing airplane model 737 MAX accident has been issued, and preliminary results have been presented in a conference (Morais et al., 2019b). After the conference, discussion with peers and supervisors have led to improvements and an extension to the machine-learning approach. This extension is presented in the present chapter, which also includes two case studies – used to demonstrate how data from different sectors can be used to train the machine, providing an efficient cross-discipline knowledge transfer. Accuracy, precision, recall and F1 score metrics have been used to measure the performance of the machine-learning model by comparing it to the classifications provided by the same human experts of MATA-D.

It is worth reinforcing that the focus of this study was to expand the dataset to decrease epistemic uncertainty in human reliability analysis. Therefore, this work has focused on testing

largely accepted and validated machine-learning techniques. In the future, other NLP approaches might be investigated such as (Ribeiro et al., 2020).

The next pages of this chapter are based on the third manuscript aligned, also aligned with the second objective of the research – to tackle sparse data. I have been the leading author, and responsible for the conceptualization, data analysis, methodology and writing the first draft. The article has been co-authored by Ka Lai Yung¹², Karl Johnson¹³, Dr Raphael Moura, Prof Michael Beer, and Prof Edoardo Patelli.

¹² Faculty of Applied Science & Engineering, University of Toronto 35 St. George Street, Room 157, Toronto, ON M5S 1A4, Canada

¹³ Centre for Intelligent Infrastructure, University of Strathclyde, James Weir Building, 75 Montrose St, Glasgow G1 1XJ, United Kingdom

Identification of human errors and influencing factors: a machine learning approach

1. Introduction

One of the most acknowledged ways to prevent design errors in complex industries is to conduct risk assessment, where multi-disciplinary teams revise a design according to information from past accidents, components and human reliability. There are industrial recommended practices on how companies should use lessons learnt from past accidents (CCPS, 2010), research on how they are actually using it (Drupsteen et al., 2013) or how it could be used (Moura et al., 2017b, Moura et al., 2017a). The lessons learnt encompass not only hazards but also their frequency of occurrence, which are used to quantify risks in probabilistic risk analysis, or to estimate order of magnitude in semi-quantitative analysis (e.g. LOPA) and qualitative analysis when risk ranking is required (Baybutt, 2016).

Regarding frequency, component failure databases play a central role in quantitative risk analysis, where data is majorly provided by components manufacturers and sometimes shared within groups of industry operators, such as the Maintenance Steering Group (MSG-3) in aviation (EASA, Gonçalves and Trabasso, 2018) and the Offshore and Onshore Reliability Data (OREDA) in upstream oil & gas (OREDA, Lima et al., 2019). However, there is still plenty of space for the development of databases to support *system safety*, which should be able to include systems and installations rather than only components' parts, as well as the interaction between human, organizational and technological factors (Leveson).

To fill this information gap, a human reliability database has been created comprising major accidents from different industry sectors (with the same level of complexity), all classified with an established human reliability taxonomy (Moura et al., 2016). The database, known as MATA-D, has currently 238 accident events classified into 53 variables, including human erroneous actions and their influencing factors (Moura et al., 2020). Although it is already possible to use it for human reliability analysis (Morais et al., 2020, Morais et al., 2021 (in press)), it would be desirable to reduce its uncertainty, leading to more precise risk estimates. To understand how to decrease its uncertainty, it is important to understand the different representation of the uncertainties within the dataset: aleatoric to model uncontrollable events, e.g. impairments and cognitive bias, or epistemic/reducible uncertainty due to missing data and theoretically reducible (Patelli et al., 2016). It is acknowledged in the human reliability field that human behaviour is dependent on the context, varying according to organizational and

technological factors (Hollnagel, 1998). The lack of information on these factors' interactions (seldomly observed and reported) is the major contribution to the epistemic uncertainty. Thus, to reduce epistemic uncertainty it would be desirable to expand the database, by collecting more accident reports and classifying them in order to increase the chance of describing more human-machine-organisation interactions.

However, collecting empirical data is time-consuming and expensive, especially in human reliability field, where data collection and classification are usually done by other humans (experts in their fields). MATA-D database have been constructed through extensive reading and classifying 238 accident investigation reports (Moura et al., 2016), a task that have taken around one year to be completed. The classification also required specialised knowledge, as the assessors had to be minimally trained on the taxonomy used to pursue the classification.

The present study proposes to enlarge a human reliability dataset by replacing (or supporting) human coding by automated classification of accident reports from any industrial sector using a pre-defined human factor's taxonomy. In order to absorb lessons learnt from different industry sectors – , the objective is to continually add to the dataset reports only from industries with the same level of complexity regarding the interaction of organisational structure, technology and humans (Moura et al., 2016). The work hereby presented is a substantial improvement and extension of the strategy proposed by some of the authors of this paper in a conference (Morais et al., 2019b). Therefore, the aim of the present research is not only to expand MATA-D, but to do it faster and timely. The use of a machine-learning strategy for text recognition and classification is herein proposed because an experienced expert takes around 3 days to read and classify one accident report, which contains about two hundred pages. A machine-learning algorithm takes less than one minute. Thus, it would be interesting to develop a computer support, that could support risk specialists, or directly collect and update the database for every new accident report of interest. Caution would be needed on the acceptance criteria of this new data, as depending on the sample quality the uncertainty might increase (Siegrist, 2011). Therefore, a central research question of this study is whether a machine learning approach is capable of both accelerating the expansion of a human reliability database and maintaining the same data quality offered by human experts.

The approach, here called as *virtual human factors classifier* might be useful in other ways. For instance, it may be used to improve human reliability Bayesian and credal networks (Morais et al., 2020, Morais et al., 2021 (in press)), or to support cross-learning from different industry sectors. It can also support incident investigators in an unbiased fashion to consider possible performance shaping factors, which might have triggered human errors (instead of

focusing only on human errors). On the original aim of expanding MATA-D, risk assessors should benefit for the provision of more data thus more possible combinations between performance shaping factors and human errors, minimising missing data problem in the use of data for probabilistic approaches.

This paper has been divided into four parts. The first part gives a brief overview of the recent history of major accident data. The second section of this paper will examine the options of machine-learning strategies and performance metrics. The third section is concerned with the dataset, taxonomy and the methodology used for this study. The fourth section presents the findings of the research, focusing on the case study of including the analysis of two accident reports from aviation (Boeing 737 MAX) and oil & gas industry (FPSO CDSM, Cidade de Sao Mateus floating production storage and offloading unit).

2. Theoretical background

This section explores the literature regarding previous similar research regarding the investigation of accidents in different industry sectors, the selection of the most used machine-learning algorithms, and most appropriate performance metrics.

2.1. Related work in similar industry sectors

The present research has focused on previous studies that have used machine-learning strategies to classify textual narratives into safety and risk features. The sample also focused in industries with similar level of organisational and technological complexity as found in MATA-D, as well as those that have investigated at least one human factor as one of the features, such as aviation (Robinson et al., 2015), railway (Hughes et al., Heidarysafa et al.), oil & gas (Ribeiro et al., 2020), civil construction (Goh and Ubeynarayana, 2017) and maritime industries (Grech et al.). A comprehensive review of the application of machine-learning techniques in occupational accident analysis, however, mixing many industries with lower level of complexity is provided in (Sarkar and Maiti, 2020).

Despite large research and application of machine-learning approaches, gaps and needs for risk and reliability analysis remains. Previous studies have not classified full accident reports into a human reliability taxonomy – nor any attempts have been identified to expand databases of human reliability with the support of machine-learning, or within multiple industry sectors. For instance, only one specific human factor (situation awareness) has been analysed in maritime accident reports (Grech et al., 2002) while often the aim was to analyse near-miss or close call reports (daily basis reports that consist of small narratives of from workers (Hughes

et al.)) to support safety managers on having timely decisions upon risk controls (Robinson et al., 2015, Hughes et al., Heidarysafa et al., Ribeiro et al., 2020, Goh and Ubeynarayana, 2017).

The highest performance obtained are from the studies with texts sizes of around 200 words, and which have collapsed many classes into a few more frequent ones. However, the need to expand the MATA-D to support better risk analysis is to classify full major accident reports (with text sizes of around 200 pages) and to not discard nor collapse classes that are less labelled (sparse data).

2.2. Human-categorized text

Readers can easily categorize a document into its topic if they have the classification scheme in mind, an action that can be described as *manual coding* (Grech et al., 2002) and *human-categorization* (Goldberg, 2017). In cases where more than one *coder* or *rater* classifies the same documents, it is good practice to measure the interrater agreement with a coefficient, such as Cohen’s kappa (Kim et al., 2020).

Although human categorization is considered the standard approach, it is time-consuming and resource demanding. It is also prone to error, in particular when involving large databases (Robinson et al., 2015). The manual assessment of accident reports has been used by Moura et al. to create the Mata-D, after reading 238 accident reports and classifying them as Boolean values according to factors described in Table 4-2 (0 if a feature was not reported, 1 if a feature was reported), as represented in Figure 4-1. A step-by-step description of how the information has been classified is shown in (Moura et al., 2016) and the resulting dataset can be assessed in (Moura et al., 2020).

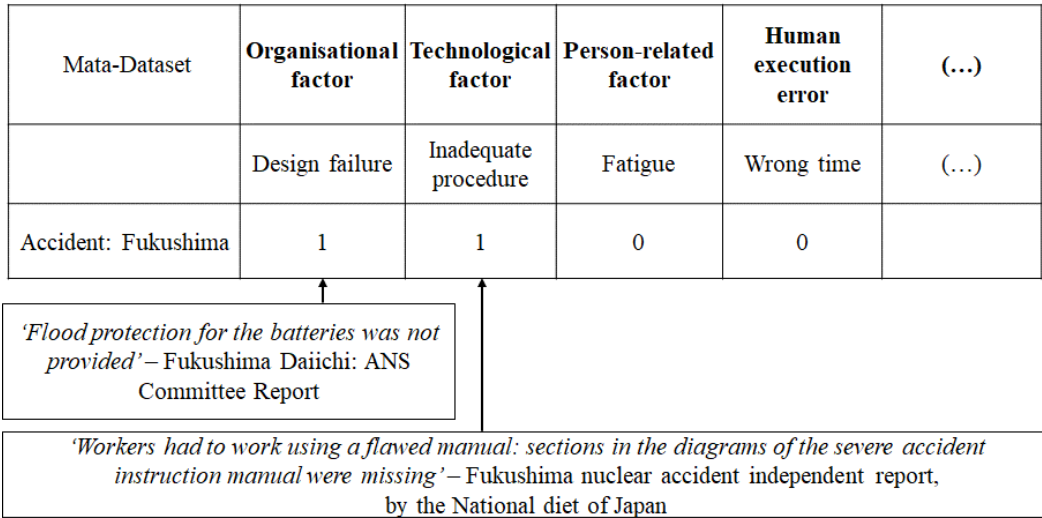


Figure 4-1. Human categorization analysis of accident reports issued for Fukushima nuclear accident (Daiichi, 2012, Fukushima Nuclear Accident Independent Investigation, 2012)

2.3. Automated text analysis algorithms

2.3.1. *Extracting and representing text features*

Before classifying a document, the text features need to be extracted to generate a *representation* of the document, capturing the properties that are important for further classification (Goldberg, 2017). There are many feature extraction methods available, but the methods that can be used to extract features from text data are mainly bag-of-words (BoW), TF-IDF and word2vec (Waykole and Thakare, 2018).

A bag-of-words model extracts features from the text, specifically the vocabulary of known words and their frequency of occurrence. The reason the model is called a ‘bag’ of words is that it does not consider any information about the order or structure of words. To use it on a set of documents, data is collected from text files and organised into a list, forming a vocabulary. To improve results and save computational time and memory the model ignores case, punctuation, and other frequent words that do not contain relevant information, such as stop words (e.g. ‘a’, ‘the’, ‘of’). To score the known words in each file (i.e. document), their presence is marked as Boolean values (0 and 1) – thus, using the list of words previously prepared, each new file is analysed and converted into a binary vector. To extract features from files, the order of words is discarded (Brownlee, 2017). *Bag-of-bigrams* is a special case of feature combinations that counts consecutive word sequences of a given length, which proves to be more powerful than bag-of-words, as word-bigrams are more informative than individual words. However, it is difficult to know a-priori which bigrams will be useful for a specific task, thus the modeller should assign the less important combinations previously with low weights. Bag of trigrams are also common, differently from 4-grams and 5-grams that are sometimes used for letters, but rarely for words due to sparsity issues (Goldberg, 2017).

TF-IDF (Term Frequency – Inverse Document Frequency) accounts for the frequency of each word in a set of documents and its useful to give higher scores to domain specific words, something that is considered a drawback for bag-of-words (as domain specific words which does not have higher frequency within a document may be ignored). TF-IDF reduces the score of frequent words in a document that are also frequent among all the documents, highlighting the words that are unique (Hughes et al., Waykole and Thakare, 2018).

Word2vec assumes that words that occur in the same contexts tend to have similar meanings (Goldberg, 2017), thus models constructed by word2vec algorithms will place words with common contexts next to each other in a vector space (Heidarysafa et al., Waykole and

Thakare, 2018). Word2vec models are two-layer neural networks, and depending on their architecture they are able to consider nearby context words more heavily than words with distant context (i.e. continuous skip gram), or to not account for context at all (i.e. continuous bag-of-words) (Waykole and Thakare, 2018).

2.3.2. *Classifying text features*

After the text relevant features are captured from the document and represented in a model, they are ready to be classified by a machine-learning technique. The most known and broadly tested techniques for automated text classification are the dictionary method, Naïve Bayes, support vector machines (SVM), latent Dirichlet allocation (LDA), latent semantic analysis (SMA), structural topic model (STM) (Kim et al., 2020). Aside from the dictionary method, they can be mostly divided into supervised and unsupervised learning methods (some authors further distinguish semi-supervised approaches, in which the training set contains a small amount of data with known categories and a large amount of data with unknown categories (Ratsaby and Venkatesh)). The method selection might be based on how texts are going to be classified, and if some documents have been previously classified by humans (allowing their use as examples to train the machine) (Goldberg, 2017, Kim et al., 2020). Figure 4-2 shows the main techniques for cases where the classification category is known and pre-defined, whereas Figure 4-3 shows techniques which classification category is unknown.

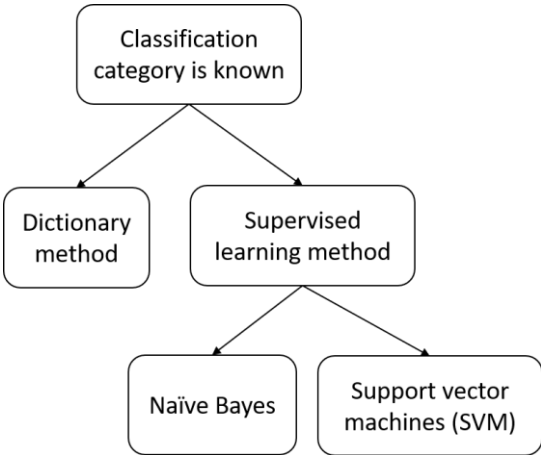


Figure 4-2. Most common automated text analysis techniques available when classification is known

In dictionary-based methods, the machine uses predefined sets of words to infer particular features of a text, relying on a user-defined dictionary. In such methods, the categories of interest are represented by single words, which are searched for by an algorithm through large bodies of text (Kim et al., 2020, Iliev et al., 2015). In the classification of

organisational factors in accidents, it would be equivalent to define into the algorithm that every time any of the words or the expressions *work shift, jetlag, lack of sleep, circadian rhythm*, is identified in the text, the algorithm classifies the feature as the organisational factor of *irregular working hours*.

Naïve Bayes and support vector machines (SVM) are popular supervised learning methods for text classification. Naïve Bayes is a simple Bayesian classifier which assumes that all attributes are independent of each other, thus independent of the word context and position in the document (Žubrinić et al., 2013, McCallum and Nigam). Naïve Bayes classifiers is reported to have better resilience to missing data than SVM classifiers (Shi and Liu), what potentially makes Naïve Bayes better to analyse fragments of texts (e.g. few paragraphs) and SVM to classify whole documents (Goh and Ubeynarayana, 2017, Wang and Manning).

Support Vector Machine (SVM) is one of the most popular supervised machine-learning algorithms, due to its little need for adjustments (Matlab, 2019), and due to their excellent prediction and generalization capabilities (Goh and Ubeynarayana, 2017, Arrieta et al., 2020). They can be used for classification, regression, or other tasks such as outlier detection (Arrieta et al., 2020). The SVM algorithm constructs a hyper-plane (or a set of them) in a high-dimensional space, so that a good separation between classes is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (Arrieta et al., 2020). The simplest case, when data have only two classes, a SVM classifies data by finding the *maximum-margin hyperplane* which separates the data points of one class from those of the second class (Matlab, 2019).

The support vectors cross the data points that are closest to the hyperplane that separate the classes. As SVM is a supervised learning model, it has to be trained before it cross-validates the classifier. Only then, the trained machine can be used to predict or classify new data. SVM is usually suggested if features' interaction might be important for classification, similar to a semantic space, as learned hyperplane separates documents belonging to different topics in the input space (Žubrinić et al., 2013). Although it is usually suggested in literature that for more complex problems, other SVM kernel functions can be used to obtain more satisfactory predictive accuracy (Matlab, 2019), previous studies show that the classification performance is not always better when non-linear polynomial kernel is applied, e.g. linear kernel outperforms non-linear when applied for multi-word classification (i.e. when the context information of individual words is captured) (Zhang et al., 2008).

When the classification category is unknown, a situation represented in Figure 4-3, unsupervised learning methods are usually chosen to infer latent categories.

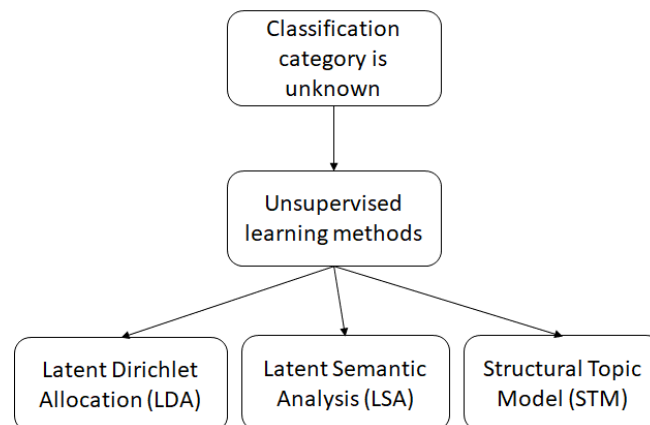


Figure 4-3. Most common automated text analysis techniques when classification is unknown

The latent semantic analysis (LSA) (Robinson et al., 2015, Deerwester et al., 1990) is a quantitative text-data analysis which employs singular value decomposition, which was a precursor of the latent Dirichlet allocation (LDA) (Blei et al., 2003), the first widely used topic model (Kim et al., 2020). LSA and LDA have similar methodologies, but LSA does not depend on rigorous statistical modelling. Statistical model estimates the categories or topics based on the pattern of word co-occurrences in the text. However, although unknown, the number of classes needs to be estimated before the analysis. Structural topic model (STM)(Roberts et al., 2016) is built upon LDA (Kim et al., 2020), thus both are topic models used to discover latent themes (i.e. thematic structures in documents), being able to reveal topic proportions in each document. STM has been designed to compensate LDA weaknesses, such as possibility of incorporating metadata (e.g. investigators' nationality and year a report was issued), and modelling direct correlations among topics (instead considering them independent) (Kim et al., 2020).

2.3.3. Measuring the performance of automatic text classification

The performance of a classifier is based on its capability to correctly assign new data to the correct class. This is often represented by the true and false positives, and true and false negatives. For a binary classifier, 1 is used to represent an observed variable in a dataset while 0 represents a non-observed variable:

- true positives occur when the true value is 1 and the model correctly predicts 1
- false negatives occur if the true value is 1 but the model wrongly predicts 0
- true negatives occur when true value is 0 and the model correctly predicts 0

- and false positives occur when true value should be 0 but the model predicts 1.

The selection of the best performance metrics to observe will vary according to how false positives and false negatives predictions will cost to the study’s objective. For example, the cost of false positive is higher if one is modelling how to identify spam emails (as someone can lose important information if an email is wrongly classified as spam). However, if the intention is to model the spread of a contagious disease, the cost of having a false negative is higher (as it is more impacting to public health if a person with a disease, an actual positive, does a test which wrongly classifies them as healthy, a false negative) (Ping Shun, 2018).

A confusion matrix is used to depict the four possible outcomes by comparing the true classes expected by the classes predicted (Google). On the confusion matrix plot depicted in Table 4-1, the rows correspond to the true class (also known as target Class), and the columns correspond to the predicted class (also known as output Class). The diagonal cells (in green) correspond to observations that are correctly classified, and the off-diagonal cells (in red) correspond to incorrectly classified observations. Some confusion matrices also show the percentage of the total number of observations in each cell, with additional columns and rows showing accuracy, prediction and recall measures (Matlab and Mathworks, 2018). In the example provided in Table 4-1 the confusion matrix indicates only the observations: 6 true positives, 2 false negatives, 1 false positive and 30 true negatives. Confusion matrices are even more useful if many variables are being classified, as it provides handy information on which classes are mostly misclassified to what other classes (Heidarysafa et al.).

Table 4-1. Confusion matrix example

True class	0	30	1
	1	2	5
		0	1
		Predicted class	

There are four main metrics to evaluate model performance according to true and false predictions: *accuracy*, *precision*, *recall* and *F-measures* score (Goh and Ubeynarayana, 2017). *Accuracy* is the fraction of correctly predicted data points out of all predictions and defined as follows.

Equation 4-1

$$Accuracy = \frac{(\text{true positives} + \text{true negatives})}{(\text{true positives} + \text{true negatives} + \text{false positives} + \text{false negatives})}$$

The potential problem of relying solely in accuracy is that it can be largely contributed by a large number of true negatives (Ping Shun, 2018), such as when dealing with imbalanced data (a dataset which has many more instances of certain classes than others) (Sun et al., 2009).

Precision is a good measure to indicate the proportion of positive identifications that are actually correct, or to monitor when the cost of a false positive is high (Ping Shun, 2018, Google, 2018). Precision is equal to 1.0 if the model produces no false positives and defined as follows:

Equation 4-2

$$Precision = \frac{(\text{true positives})}{(\text{true positives} + \text{false positives})}$$

When the cost of false negative is high, the *Recall* is a good measure to indicate if the proportion of actual positives are identified correctly (Ping Shun, 2018, Google, 2018). A model that produces no false negatives has a *recall* of 1.0. The recall metric is defined as:

Equation 4-3

$$Recall = \frac{(\text{true positives})}{(\text{true positives} + \text{false negatives})}$$

F-measures are useful if a balance between *precision* and *recall* is needed (Ping Shun, 2018), as empirical studies of retrieval performance have shown a tendency for *precision* to decline as *recall* increases (Buckland and Gey, 1994). It is also a good measure if the true classes present an uneven distribution such as a large number of true negatives (Ping Shun, 2018). If false negatives and false positives are equally costly, F_1 score represents the harmonic mean between recall and precision:

Equation 4-4

$$F_1 = 2 \cdot \frac{(\text{Precision} \cdot \text{Recall})}{(\text{Precision} + \text{Recall})}$$

However, if false negatives and false positives are not equally costly, F_β measure might be indicated as it is an abstraction of the *F-measure* where the balance of *precision* and *recall* are controlled by a coefficient called β . If false negatives cost more, $\beta > 1$; if false positives are more costly, $\beta < 1$ (He and Ma, 2013).

Equation 4-5

$$F_\beta = (1 + \beta^2) \cdot \frac{(\text{Precision} \cdot \text{Recall})}{(\beta^2 \cdot \text{Precision} + \text{Recall})}$$

Using the example given in the confusion matrix in Table 4-1, the accuracy of the model would be 92%, precision would be 86%, recall would be 75%, and using the results of precision and recall the F_1 score would be 80%.

Performance metrics may present different results depending on the size and on the randomised sample used for training and testing sets. To minimise the randomised sample effect, many studies present the metrics by variable (or sets of variables) instead of by overall indicators (Heidarysafa et al., Goh and Ubeynarayana, 2017, Grech et al., 2002, Zhang et al., 2019). The difference in performance metrics can be more transparently depicted by error estimates (Ribeiro et al., 2020). The need for smaller uncertainties between estimates can also define the size of training and testing sets. Some machine-learning practitioners even suggest to have larger testing sets than what is normally recommended, in order to increase the confidence in model predictions (not only because the error estimates of performance metrics decrease, but because the user can actually see how the model works for more samples) (Malato, 2015).

3. Methodology

Support vector machine was proposed to automatically read and classify accident reports into potential human factors, with the support of Bag-of-Words model for data collection. The model was trained and tested using data from MATA-D. This section better describes the dataset used and the procedures applied to train and test the models.

3.1. Dataset

The classification tool was trained using the data from Mata-D. The decision was based on the conceptual advantaged of potential cross-learning lessons from accidents in different sectors, but also brought two technical advantages regarding machine-learning techniques.

Firstly, the majority of accident reports were available to train and test the machine against the opinion classified by experts. Secondly, the dataset had a specific taxonomy, which simplified the decision on the automated text technique to choose.

The type of documents analysed were accident investigation reports, all in English, with an average size of two hundred pages. The accidents described in those reports had happened in different industry sectors with similar complexity regarding the interaction within humans, technology and organization, such as: aviation, chemicals factory, construction, food, oil & gas (exploration, refinery, petrochemical), metallurgical, nuclear, terminals and distribution and waste treatment plant. The documents chosen were the same used to construct a dataset of 238 reports classified into a human reliability taxonomy as described in (Moura et al., 2016), known as MATA-D which can be assessed in (Moura et al., 2020).

Table 4-2 shows the taxonomy used, the classification scheme developed for a human reliability method known as CREAM (cognitive reliability and error analysis method) (Hollnagel, 1998). This taxonomy comprises human errors and performance shaping factors (PSFs) such as organisational, technological and individual factors. CREAM’s taxonomy has the benefit of serving both accident analysis and risk analysis purposes. Thus, by continuously updating the dataset with new accident investigation reports, the dataset will provide risk and reliability analysis with better predictions of which combinations of factors mostly trigger accidents. Although MATA-D is the dataset which contains information on how 238 accident reports have been labelled against CREAM taxonomy, only the publicly available reports have been used to train and test the virtual classifier in the present study: a total of 106 reports.

Table 4-2. Taxonomy of human factors adopted in MATA-Dataset based on CREAM classification scheme

Organisational Factors	Technological Factors	Individual factors	Human Execution Errors
Communication failure	Equipment failure	Permanent related	Wrong time
Missing information	Software fault	Functional impairment	Wrong type
Maintenance failure	Inadequate procedure	Cognitive style	Wrong Object
Inadequate quality control	Access limitations	Cognitive bias	Wrong place
Management problem	Ambiguous information		
Design failure	Incomplete information	Temporary related	Cognitive function failures
Inadequate task allocation	Access problems	Memory failure	Observation missed
Social pressure	Mislabelling	Fear	False Observation
Insufficient skills		Distraction	Wrong Identification
Insufficient knowledge		Fatigue	Faulty diagnosis
Temperature		Performance Variability	Wrong reasoning

Sound	Inattention	Decision error
Humidity	Physiological stress	Delayed interpretation
Illumination	Psychological stress	Incorrect prediction
Other		Inadequate plan
Adverse ambient conditions		Priority error
Excessive demand		
Inadequate work place layout		
Inadequate team support		
Irregular working hours		

As the reports in the MATA-Dataset addressed different industry sectors, they presented different formats and vocabularies. The format changed not only in terms of number of pages, but also in terms of reproduceable sections in a corpus. The vocabularies varied not only on specificity of the different industrial sectors, but also in terms of taxonomy applied usually connected to the investigation methodology. This research used three different datasets: the first contained 106 publicly available reports (public at the time of the research), the second was a subset of the first dataset with 57 CSB reports (U.S. Chemical Safety and Hazard investigation board), and the third was another subset of the first dataset with 20 reports issued by NTSB (U.S. National Transport Safety Board). CSB is an U.S. independent government agency that investigates mainly industrial chemical accidents, covering accidents not only in chemical factories, but also in its branches (e.g. oil & gas, food, and metallurgical industries). NTSB is also an independent U.S. government agency, which investigates accidents in transportation, such as aviation, and including terminals and distribution. CSB and NTSB were chosen due to their larger number of reports in MATA-Dataset and due to their systematically organised and repetitive format (e.g. similar chapters titles and same order of chapters), which is potentially positive considering the training of a supervised learning technique.

The three datasets generated three different models: *all reports*, *CSB* and *NTSB* models. The reports were randomly split into a training-testing ratio of 80-20%, therefore generated a training set of 85 reports and a testing set of 21 reports for *all reports* model, 46 to train and 11 to test reports in CSB model, and 16 to train and 4 to test reports in NTSB model. The decision of choosing between an 80-20% split instead of a 90-10% was taken to increase the confidence in the results as suggested in (Malato, 2015)..

3.2. Machine-learning technique

As the classification of the category is known (i.e., predefined taxonomy), and the dataset was previously labelled by experts, a supervised learning method is the most adequate, short-listing the decision to Naïve Bayes or Support Vector Machine. It has been proven that Naïve Bayes classifiers perform better with missing data (Shi and Liu), and therefore it might be a good choice to identify human factors interactions in major accidents that are considered rare and uncertain events (Morais et al., 2020). However, SVM has the potentiality to better capture features interactions (Žubrinić et al., 2013) and better classify larger documents (Wang and Manning). Therefore, as interaction patterns have been observed between MATA-D factors in (Moura et al., 2017b) and the aim is to apply the tool to accident reports with 200 pages on average, an SVM model with a linear kernel has been chosen for classification. Bag-of-words was selected as the feature extraction tool to pre-process the features to be classified by SVM. The choice was not only due to its recognised simplicity and flexibility (Waykole and Thakare, 2018), but also because the intention to classify accident reports with no specific sector or domain suggested that it was better not to use models that capture too much the context from the training set into account – to avoid giving much higher importance to sector specific words or set of words (Goldberg, 2017).

The resulting automated text recognition and classification tool is referred to as the *human factors virtual classifier*. A simplified workflow of the proposed approach is shown in Figure 4-4.

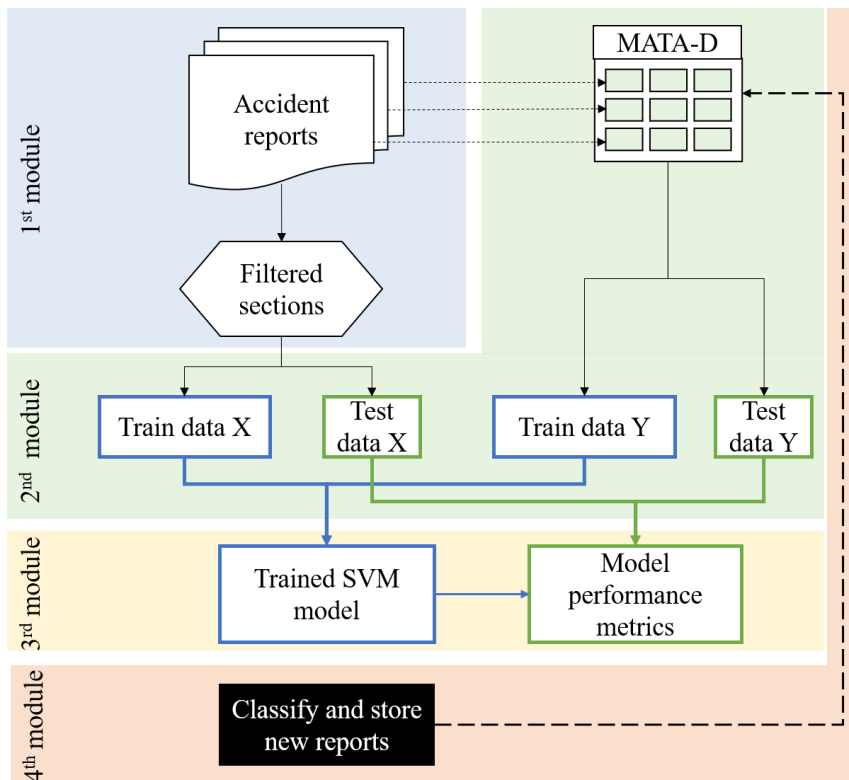


Figure 4-4. Simplified workflow of the human factor virtual classifier

In the first module, accident investigation reports were analysed. The documents in portable document format (i.e., files with PDF extension) were processed to check if the text in pdf files were recognised by the machine and, if not, an optical character recognition software (OCR) was used to convert them to text files – an important step for relatively old accident reports. After this pre-treatment, the tool scanned the accident reports, and their texts were sent to the next module. In the implemented version, the semi-supervised approach gave the users the option to manually identify relevant sections, which was the option used in this study. Otherwise, the most likely start and end of the targeted sections, *recommendation* and *lessons learned*, would be identified by a confidence scoring system (a basic algorithm, tailored for this project, which defines a dictionary of the most likely start and end target words in major accident reports), and these sections would be the output to the next module. Finally, the text was pre-processed to clean punctuation, stop words, and reduce words to their stem (e.g., ‘testing’ was reduced to ‘test’).

In the second module, using another confidence scoring system, the tool took each accident report’s file name and found the most likely corresponding entry in the MATA-D. For this reason, the accident reports had equally assigned names in dataset and correspondent PDF file. This gave the machine-learning component the desired output for each accident report, which was a combination of selected section texts and their known human factors. Then, the

selected text was converted into bag-of-words objects (X in *Figure 4-4*, forming the input of the model), and the factors extracted from the MATA-D (Y in *Figure 4-4*, served as the output of the model). The module partitioned the data into a training set (80% of total) and a testing set (20% of total).

In the third module, the model based on SVM was trained and tested using data input from the previous two modules. Finally, the parameters of the classifier were recorded and overall performance metrics (i.e., *accuracy*, *precision*, *recall* and F_1 score) were calculated based on test sets in all categories, as well as a confusion plot is generated. Only then, the tool was prepared to be used in the next module.

The fourth and final module of the tool allowed users to add a new report that was not yet part of the MATA-D. The result was a list of the human reliability factors identified by the tool (an array of the predicted positive factors), a small table with all positives and negatives predictions (the 53 factors of the chosen taxonomy), and a word cloud of the most relevant words in the report.

3.3. Implementation

All the computational work was carried out using MATLAB software, and supported by the *text analytics toolbox*, which used the bag-of-words model to extract text strings from files and prepare data for the machine-learning algorithm. The *MATLAB statistics and the machine-learning toolbox* was used to transform text inputs into binary classification adopting the Support Vector Machine. Data was extracted from the Excel based MATA-Dataset, while the accident report were in portable document format (i.e., PDF extension). The text recognition software embedded in Adobe Acrobat Pro was used to convert text-images to text-strings in cases where original reports had been saved as images (e.g. relatively old accident reports, such as the Public Inquiry into the Piper Alpha Disaster (Cullen, 1993)). Computational times to evaluate a new report, including the machine training time, took around 63 seconds (using all reports), 28 seconds (using CSB reports), and 19 seconds (using NTSB reports), using a laptop configured with Intel® Core™ i5-8265U CPU @ 1.60GHz and 16.0 GB of RAM.

The classification tool was implemented on a user-friendly web interface known as *Virtual Raphael* (after the name of the expert who conceptualized and co-created MATA-D), where the reader can classify their own accident report online, without the need to save it to the database. Together with the results a message is displayed to remind that the human factors outputs are just an indication to support the user, and that they potentially present a similar accuracy, precision, recall and F1 score of the test set shown in this study. The classifier tool is

freely available at the following web address:
https://cossan.co.uk/private/incident_classification/.

The web-interface, coded in JavaScript, links three main components: the MATA-D dataset, the public accident reports, and a collection of six Matlab scripts. The dataset and all the codes used in this work are also available to those readers and researchers that want to replicate the experiment or to do their own improvements:

- The dataset MATA-D with labelled classifications of each report is available at University of Liverpool's data repository, available at:
<https://doi.org/10.17638/datacat.liverpool.ac.uk/1018> (Moura et al., 2020).
- The links and references to the public accident reports classified in the MATA-D and used for the training and testing sets by the Virtual Raphael classifier are available from the Cossan website. However, due to property issues, they are not shared in their pdf formats.
- The source code of the methods is available from the GitHub repository of the Cossan software. The path is <https://github.com/cossan-working-group/VirtualRaphael/>.

3.4. Performance

To measure the performance of the human factors virtual classifier, the binary classifications available in MATA-D were used as target classes. Four performance metrics were selected: *accuracy*, *precision*, *recall* and F_1 score. The selection took into consideration that a typical accident in MATA-D is largely contributed by a large number of true negatives (an average of 46 negatives out of 53 categories were identified among all the reports), which might be classified as an imbalanced dataset. In those cases, F_1 score is considered a better metric than *accuracy* (if recall and precision are considered equally important).

The metrics were used to evaluate and compare the three trained classifiers using all reports, using the CSB reports and using the NTSB reports, respectively. To calculate them, ten randomly selected reports from the database were taken, maintaining constant the size of the samples and the training-test split. For each random sample generated, the training and testing sets were the same for the 53 category models created. The confusion matrices used to compare the true classes from MATA-D with the predicted classes are presented in Table 4-3 (all reports model),

Table 4-4 (CSB reports), Table 4-5 (NTSB reports). The green numbers represent the true positives and true negatives, while the red numbers are the false positives and false negatives – considering the cumulative sum of predicted results from 10 random training sets. The values in the tables indicate the counting of positive and negative classifications for all the reports.

Table 4-3. Confusion matrix of all reports' model predictions (cumulative sum of ten different samples)

True class	0	8720	565
	1	994	851
		0	1
		Predicted class	

Table 4-4. Confusion matrix of CSB reports' model predictions (cumulative sum of ten different samples)

True class	0	4730	198
	1	458	444
		0	1
		Predicted class	

Table 4-5. Confusion matrix of NTSB reports' model predictions (sum of ten different samples)

True class	0	1608	156
	1	194	162
		0	1
		Predicted class	

The performance metrics were calculated using Equations 1 to 4 and summarised in Table 4-6. The classifier model trained and tested with CSB reports obtained the best performance in all four metrics.

Table 4-6. Performance metrics according to confusion matrices cumulative sum of 10 randomly selected report from the database.

	all reports	CSB reports	NTSB reports
--	--------------------	--------------------	---------------------

Accuracy	86%	89%	83%
Precision	60%	69%	51%
Recall	46%	49%	46%
F1 score	52%	58%	48%

Instead of measuring the performance based in the predictions' cumulative sums, it was also useful to analyse how the performance metrics had varied according to different training and testing sets. Therefore, Table 4-7 shows the minimum and maximum results achieved by the performance metrics, as well as their mean and standard deviation (SD) if the ten random samples were considered separately. It was possible to observe that the results in Table 4-6 matched almost completely with the performance metrics mean values in Table 4-7.

Table 4-7. Performance metrics of ten randomly selected reports from the database considered separately

	All reports				CSB reports				NTSB reports			
	Min	max	mean	SD	min	max	mean	SD	min	max	mean	SD
Accuracy	83%	89%	86%	2%	87%	91%	89%	1%	80%	88%	83%	3%
Precision	51%	69%	60%	6%	64%	77%	69%	4%	37%	67%	51%	9%
Recall	40%	53%	46%	4%	42%	55%	49%	5%	36%	72%	47%	10%
F1 score	45%	60%	52%	4%	52%	61%	57%	3%	38%	57%	48%	5%

In this study, the linear SVM model trained with all public reports achieved a mean of 86% in the *accuracy*, 60% for the *precision*, 46% in the *recall* and 52% using the F_1 score. Table 4-7 shows a slightly higher performance when the model was trained using only the CSB reports, which might be explained by their similarity of format and industry sectors.

The results obtained had performed similarly to the benchmarked studies, as shown in the discussion section of this paper.

Another important type of performance is the training time required by the machine-learning algorithm. The elapsed time taken for the linear SVM to train and test with all reports was 63 seconds, with CSB reports was 28 seconds, and 20 seconds with the NTSB reports – all using the laptop configuration described in the methodology section.

Word clouds were used in this research on an attempt to inspect the bag-of-words contents from the training and testing sets in the different models, in order to better understand their performance. Figure 4-5, Figure 4-6, and Figure 4-7 provide visualisation to the more frequent words in training and testing sets bag-of-words for all reports, CSB reports and NTSB reports.



Figure 4-5. Word cloud for the trained model all reports



Figure 4-6. Word cloud for the CSB model



Figure 4-7. Word cloud for the NTSB model

4. Case studies

In order to test the model in new accident reports (i.e.y, not yet on Mata-D), two investigation reports from different industry sectors (aviation and oil & gas) were chosen to be analysed and classified by the same expert that originated the dataset. The results of the




automated classification were not shown to him before the task, to avoid him to get biased. Many tests were conducted prompting the automated tool to analyse different sections of each report, to see if the analysis of different chapters impacted the results in different ways. The results shown in Table 4-8 and Table 4-11 present the results when the tool analysed the full report.

4.1. Aviation case study – 2018 Boeing 737 MAX 8 AIRCRAFT final accident report

On October 2018, an accident with a Lion Air aircraft, led to 189 fatalities (KNKT, 2019). Five months later, in 2019, an Ethiopian Airlines plane crashed minutes after take-off, killing all 157 onboard (Marks and Dahir, 2020). The fact that both accidents involved the same aircraft model, a Boeing 737-8 MAX, had concerned civil society and safety regulators about the possible common flaws, which resulted in all 387 planes with same model grounded globally (BBC, 2019). The two events have been famously known by the potential design flaws of the Manoeuvring Characteristics Augmentation System (MCAS) which might have misled the pilots' actions (Chronopoulos and Guzman, 2020).

Differently from the first test of the tool performed on the preliminary accident report (Morais et al., 2019b), this research tested the machine-learning tool on the final accident report of the Lion Air Aircraft flight, issued on October 2019 (one year after the accident) (KNKT, 2019). Although the final accident report of Ethiopian airlines was reportedly issued (Marks and Dahir, 2020), the link was not accessible for unknown reasons until the date this paper was submitted to reviewers, thus not included in this research. For the classification of the Lion Air report, the three different training sets were also pursued (all publicly available reports, all CSB reports, and all NTSB reports). The final accident report was previously classified by the same experts which have classified MATA-Dataset within the CREAM human factors taxonomy, in order to compare their similarity in new reports. Table 4-8 shows the comparison between human factors classifications obtained with human coding and different training sets. The complete report was considered (from 'SYNOPSIS' to '6 APPENDICES').

The table was colour coded according to the legend below to help the reader understand how the model prediction metrics were calculated. It also helps to show what predictions the authors considered more important for this study (the darker the colour, the more important).

-  True positives: dark green (expert classified as '1' and machine predicted correctly as '1')
-  True negatives: light green (expert classified as '0' and machine predicted correctly as '0')
-  False negatives: dark red (expert classified as '1', but machine wrongly predicted as '0')

() False positives: red (expert classified as ‘0’, but machine wrongly predicted as ‘1’)

Table 4-8. Virtual classifier trained using different report sets vs. expert classification for Lion Airline accident report (Boeing 737-8MAX)

		Expert classification	Virtual classifier trained with all reports	Virtual classifier trained with CSB reports	Virtual classifier trained with NTSB reports		
HUMAN	Action	Execution (Error Modes)	Wrong Time	1	0	0	0
			Wrong Type	0	0	0	0
		Wrong Object	0	0	0	0	
		Wrong Place	1	1	0	1	
	Specific Cognitive Functions	Observation	Observation Missed	0	0	0	0
			False Observation	0	0	0	0
			Wrong Identification	0	0	0	0
		Interpretation	Faulty diagnosis	1	1	0	1
			Wrong reasoning	0	0	0	0
			Decision error	0	0	0	0
			Delayed interpretation	1	0	0	0
			Incorrect prediction	0	0	0	0
		Planning	Inadequate plan	1	0	0	0
			Priority error	1	0	0	0
		Temporary Person Related Functions	Memory failure	0	0	0	0
			Fear	0	0	0	0
			Distraction	1	0	0	1
	Fatigue		0	0	0	0	
	Performance Variability		0	0	0	0	
	Inattention		0	0	0	0	
	Physiological stress		0	0	0	0	
	Psychological stress		0	1	0	0	
	Permanent Person Related Functions	Functional impairment	0	0	0	0	
		Cognitive style	0	0	0	0	
		Cognitive bias	0	0	0	0	
	TECHNOLOGY	Equipment	Equipment failure	1	1	0	0
			Software fault	0	0	0	0
		Procedures	Inadequate procedure	1	1	1	1
		Temporary Interface	Access limitations	0	0	0	0

ORGANISATION		Ambiguous information	1	0	0	0
		Incomplete information	1	0	0	0
	Permanent Interface	Access problems	0	0	0	0
		Mislabelling	0	0	0	0
	Communication	Communication failure	1	0	0	0
		Missing information	1	1	0	0
	Organisation	Maintenance failure	1	1	1	0
		Inadequate quality control	1	1	1	1
		Management problem	1	0	0	0
		Design failure	1	1	1	1
		Inadequate task allocation	1	1	1	1
		Social pressure	0	0	0	0
	Training	Insufficient skills	1	1	1	1
		Insufficient knowledge	1	1	0	0
	Ambient Conditions	Temperature	0	0	0	0
		Sound	0	0	0	0
		Humidity	0	0	0	0
		Illumination	0	0	0	0
		Other	0	0	0	0
		Adverse ambient conditions	0	0	0	0
	Working Conditions	Excessive demand	1	0	0	0
		Inadequate work place layout	0	0	0	0
		Inadequate team support	1	0	0	0
		Irregular working hours	0	0	0	0
		Sum of true positives		11	6	8
		Sum of true negatives		30	31	31
		Sum of false positives		1	0	0
		Sum of false negatives		11	16	14
		Accuracy		77% ^(79%)	70%	74%
		Precision		92% ^(100%)	100%	100%
	Recall (or true positive rate)		50%	27%	36%	
	F1 Score		65% ^(67%)	43%	53%	

According to Table 4-8, the model trained with all reports retrieved the best accuracy, recall and F1 score. Only the precision was slightly lower than those obtained using the CSB and NTSB reports. When the classifier is trained with all reports the following factors were observed in the Lion Air accident operating with the Boeing 737 MAX: human error of execution of wrong place (i.e. action out of sequence); the cognitive function failure of faulty diagnosis; the technological factors of equipment failure and inadequate procedure; the organisational factors of missing information, maintenance failure, inadequate quality control, design failure, inadequate task allocation, insufficient skills, insufficient knowledge.

The confusion matrices for the three models are presented in Table 4-9.

Table 4-9. Confusion matrices for the Boeing 737 MAX accident report predictions

True class	All reports model		CSB reports		NTSB reports	
	0	1	0	1	0	1
0	30	1	31	0	31	0
1	11	11	16	6	14	8
	Predicted class		Predicted class		Predicted class	

The report was also classified after selecting its potentially more important sections, which carried more information about the accident causes (the report initial information was discarded, as it contained overall info about the plane and not about the accident). For all three models, the performance metrics obtained are mostly similar to the analysis of the whole report, with slight improvement only for all reports model in terms of accuracy (79%), precision (100%) and F1 score (67%). Table 4-10 shows the results after grouping the model outputs for all the 53 factors into 4 main groups (i.e., human errors, individual factors, technological factors, and organisational factors).

Table 4-10. Model performance by sets of human factors for the Boeing 737 MAX report.

All reports	Human errors and cognitive function failures	Individual factors	Technological factors	Organisational factors
<i>Accuracy</i>	71%	82%	75%	80%
<i>Precision</i>	100%	0%	100%	100%
<i>Recall</i>	33%	0%	50%	64%
<i>F1 Score</i>	50%	0%	67%	78%

The word cloud was included as it might serve as an additional support for the user to check if the information in the report is being correctly extracted or if there are problems that

deserve any intervention to improve the prediction performance. It might be also important to compare the word cloud obtained with the new report (Figure 4-8) with the word clouds of the training and testing sets (Figure 4-5, Figure 4-6, and Figure 4-7).



Figure 4-8. Word cloud for the Boeing 737 MAX accident report

To classify the Lion Air Accident report, the algorithm took 74 seconds with the model trained with all reports, 32 seconds with model trained with CSB reports, and 30 seconds with model trained with NTSB reports (considering the training time).

4.2. Oil & Gas case study: FPSO Cidade de São Mateus (CDSM) accident report

On February 2015, an explosion onboard FPSO Cidade de São Mateus killed nine, injured 26 workers, as well as caused damage to the installation, and production halt of two gas production fields up to this date (2021). The Brazilian Oil & Gas regulator (ANP) included in their investigation report root causes from the design phase to the emergency response. The FPSO (floating production, storage and offloading unit) was operated by BW Offshore in gas fields under concession to Petróleo Brasileiro S.A (Petrobras) in Brazilian waters (ANP, 2015).

The tool was also trained with the same training sets adopted to the aviation case study. The FPSO CDSM accident report was previously classified by the same experts as MATA-Dataset and Lion Airline report. Table 4-11 shows the comparison between human factors classifications obtained with human coding and the different training sets. The complete report was considered (from its title to ‘Conclusion’ chapter).

Table 4-11. Virtual classifier x expert classification for FPSO Cidade de Sao Mateus accident report classification

	Expert classification	Virtual classifier trained	Virtual classifier trained	Virtual classifier trained with
--	-----------------------	----------------------------	----------------------------	---------------------------------

				with all reports	with CSB reports	NTSB reports	
HUMAN	Action	Execution (Error Modes)	Wrong Time	0	0	0	0
			Wrong Type	0	0	1	0
			Wrong Object	0	0	0	0
			Wrong Place	1	0	0	1
	Specific Cognitive Functions	Observation	Observation Missed	1	0	0	1
			False Observation	0	0	0	0
			Wrong Identification	0	0	0	0
		Interpretation	Faulty diagnosis	1	0	0	1
			Wrong reasoning	1	0	1	0
			Decision error	0	0	0	0
			Delayed interpretation	0	0	0	0
		Planning	Incorrect prediction	0	0	0	0
			Inadequate plan	1	0	0	0
			Priority error	0	0	0	0
		Temporary Person Related Functions	Memory failure	0	0	0	0
			Fear	0	0	0	0
			Distraction	0	0	0	0
	Fatigue		0	0	0	0	
	Performance Variability		0	0	0	0	
	Inattention		0	0	0	0	
	Physiological stress		0	0	0	0	
	Psychological stress		0	0	0	0	
	Permanent Person Related Functions	Functional impairment	0	0	0	0	
		Cognitive style	0	0	0	0	
		Cognitive bias	1	0	1	0	
	TECHNOLOGY	Equipment	Equipment failure	0	0	0	0
			Software fault	0	0	0	0
Procedures		Inadequate procedure	1	1	1	1	
Temporary Interface		Access limitations	0	0	0	0	
		Ambiguous information	0	0	0	0	
		Incomplete information	1	0	0	1	
Permanent Interface	Access problems	0	0	0	0		
	Mislabelling	0	0	0	0		
ORGANISATION	Communication	Communication failure	1	0	0	1	
		Missing information	1	0	0	0	
	Organisation	Maintenance failure	1	1	1	0	
		Inadequate quality control	1	1	1	1	
		Management problem	0	0	0	0	
		Design failure	1	1	1	1	
		Inadequate task allocation	1	1	1	1	

Training	Social pressure	1	0	0	0
	Insufficient skills	1	0	0	1
	Insufficient knowledge	1	0	1	0
Ambient Conditions	Temperature	0	0	0	0
	Sound	0	0	0	0
	Humidity	0	0	0	0
	Illumination	0	0	0	0
	Other	0	0	0	0
	Adverse ambient conditions	0	0	0	0
	Working Conditions	Excessive demand	1	0	0
Inadequate work place layout		0	0	0	0
Inadequate team support		0	0	0	0
Irregular working hours		0	0	0	0
	Sum of true positives		5	8	10
	Sum of true negatives		35	34	35
	Sum of false positives		0	1	0
	Sum of false negatives		13	10	8
	Accuracy		75%	79%	85%
	Precision		100%	89%	100%
	Recall (true positive rate)		28%	44%	56%
	F1 Score		43%	59%	71%

The model trained with NTSB reports retrieved the best *accuracy*, *precision*, *recall* and *F₁* score. If trained with NTSB reports the following features are observed in the oil & gas installation, the FPSO Cidade de Sao Mateus: human errors of execution of wrong place (i.e. action out of sequence); the cognitive function failures of observation missed and faulty diagnosis; the technological factors of inadequate procedure and incomplete information (related to temporary interfaces); the organisational factors of communication failure, inadequate quality control, design failure, inadequate task allocation, and insufficient skills. For another visualisation of true and false predictions, the confusion matrices for the three models are presented in Table 4-12.

Table 4-12. Confusion matrices for FPSO Cidade de Sao Mateus accident report predictions

	All reports model		CSB reports		NTSB reports		
True class	0	35	0	34	1	35	0
	1	13	5	10	8	8	10
		0	1	0	1	0	1
		Predicted class		Predicted class		Predicted class	

To classify the Lion Air Accident report, the algorithm took 66 seconds with the model trained with all reports, 34 seconds with model trained with CSB reports, and 24 seconds with model trained with NTSB reports (considering the training time).

5. Discussion

MATA-D has the potential to incorporate the information of human reliability into risk assessments. It needs more data to increase its accuracy and reduce uncertainty. However, the data collection process of reading and classifying reports is a time consuming and challenging task, prone to errors. Therefore, this study aimed at demonstrating the capability of a machine learning tool trained using previously classified accident reports in MATA-D database to classify new accident reports with sufficient accuracy, precision and recall. In other words, this research investigates if machine learning is capable of accelerating the expansion of this database while maintaining the same data quality obtained with human experts. The results have shown that the automated classification of new accident reports can accelerate the data collection process, as it can reduce the time from around 3 days (when the report is classified by an expert) to around 1 minute.

5.1. Performance and accuracy of the automatic classifier tool

Four performance metrics were selected to investigate the differences between expert and machine-based classification. Table 4-14 benchmarks the performance metrics on this study against previous studies from literature. The results are summarised in Table 4-14 for the classifier trained using all reports. The classifier in this study and from previous studies were trained using all the human factors and the average performance among all the factors is reported in Table 4-14. Additionally, only the best results available from the literature were considered. For instance, in the study of (Grech et al., 2002), when more reports were tested the precision of the method dropped from 84% to 48%, and the recall dropped from 89% to “not possible to measure”.

Table 4-14. Average performance metrics for all the 53 factors versus results from literature

Metric	Test set	Aviation case study	Oil & gas case study	Previous studies
Accuracy	86% (SD = 2%)	77%	75%	44% (Robinson et al., 2015) 75% (Heidarysafa et al., 2018)

				90% (Ribeiro et al., 2020)
Precision	60% (SD = 6%)	92%	100%	22% (Robinson et al., 2015)
				73% (Goh and Ubeynarayana, 2017)
				84% (Grech et al., 2002)
Recall	46% (SD = 4%)	50%	28%	63% (Goh and Ubeynarayana, 2017)
				89% (Grech et al., 2002)
F₁ score	52% (SD = 4%)	65%	43%	53% (Ribeiro et al., 2020)
				67% (Goh and Ubeynarayana, 2017)
				71% (Heidarysafa et al., 2018)

The availability of an acceptable threshold for each performance metric, which could help to decide when the data collected by an automatic classification could be added to a database without corrupting its quality, is not available. The comparison in Table 4-14 shows that, from the four chosen metrics, only the *recall* is below the benchmark studies.

To understand how the *recall* impacts the quality assurance of this project, it is important to understand the objectives of the classification. At a first sight the *recall* metric seems to be the best candidate for human reliability classifier, because a performance shaping factor that goes undetected prevents the allocation of resources for the risk reduction. However, a good *precision* is also important for resource allocation— for a risk assessment purpose it might be more detrimental, as resources are allocated to prevent an event that might not really contribute to the risk. In other words, both false negatives and false positives are detrimental for the decision of partially replacing experts in the data collection. As it is not possible to achieve a *precision* and a *recall* of 100% at the same time (Buckland and Gey, 1994), it is suggested that a balance between both is achieved using the F_1 score. If at some part of the analysis, it is considered that the *recall* or the *precision* are not equally important, it is suggested to use F_β with $\beta > 1$ (*recall* more important) or $\beta < 1$ (*precision* more important).

Although the test set already provided the metrics needed to benchmark the performance of the proposed automatic classifier against previous studies, the presented case studies offered additional insights into how the classifier performed. The case studies have demonstrated the applicability of the approach for different sectors (i.e., aviation and oil & gas) although the performance achieved was slightly out of the bounds established by the test set standard deviation, especially regarding the *precision* and the *recall*. Literature suggests that this difference might be decreased by using domains specific training sets (Brownlee, 2018) and

this approach can be adopted to improve the *recall* for a specific industry sector. However, in this study the aim is to learn from accident occurred in different sectors and therefore training a generic classifier.

For the oil & gas case study trained with all reports, a perfect prediction (100%) has been obtained although with a low *recall* score (28%), meaning that only a few human errors and performance shaping factors were identified but no false positive.

It has also been tested whether grouping all the 53 factors into 4 main groups (i.e., human errors, individual factors, technological factors, and organisational factors) would have been able to improve the classification when the classifier is trained using all reports. For the aviation case study, as shown in Table 4-10, the F_1 score improved to 78% for organisational factors and only to 65% for technological factors (compared to the overall mean of 65% shown in Table 4-14). For the oil & gas case study, in Table 4-13, the F_1 score of organisational and technological factors improved to 78% and 67%, respectively due to the use of an enriched training dataset with higher frequency of organisational and technological factors. For both case studies, the F_1 score of human errors and individual factors performed worse when analysing the factors by groups.

Surprisingly, the oil & gas case study has showed better results when the classifier was trained using only NTSB reports. Although this set contains some reports related to oil & gas terminals and distributions, the majority of reports are from the aviation sector. The expectation was that CSB reports would have provided a better training set. For the Lion Air accident report, the classifier trained with all reports performed better than those trained only with NTSB reports, which contains more aviation specific language (as can be seen by the word cloud presented in Figure 4-7). This result might be due to the different formats used for the reports tested, as they are from different investigation bodies.

Observing the results of the case studies, it has been noted that the majority of categories detected by the machine-learning approach were inside the 26 most significant contributing factors per cluster identified in a previous research (Moura et al., 2017b). This might suggest training the classifier using only fewer frequent categories. However, tests were performed reducing the number of categories to the 13 most frequent ones, and the results did not present significant changes, e.g., an improvement of ~5% for *precision*, *recall* and F_1 score, but with a deterioration of the same level in the *accuracy*. Therefore, it has been decided to keep all categories in the training set, as the main goal of this research is to expand the current MATA-D dataset using the same categories already available and therefore decrease the uncertainty associated to rare combinations of human error and performance shaping factors.

This study had not found a significant difference between the automated classifications of full reports and of reports' selected sections. Word cloud figures were provided to visualise frequent words and aided the task of inspecting which sections of the accident reports provided more relevant information.

5.2. Future improvements and recommendations

One of the limitations of the current classifier is its moderate capability to identify infrequent classes. One solution is to enrich the training set with accident reports where those infrequent classes had occurred (according to an analysis provided by human factor experts) – by training the model on these classes it is expected that the overall *recall* metric will increase as more data is used. Different resampling strategies might also be used (e.g., targeting infrequent classes to resample rather than sampling the training data set randomly). Finally, algorithms that maximise the *recall* while using the *precision* metric as a constraint should also be investigated (see e.g. (Bennett et al., 2017)). Solutions to strengthen learning with regards to the small class might be applied (e.g., adjusting the SVM class boundary based on kernel-alignment). Further research might also assume higher misclassification costs applied to samples in the infrequent classes and seek to minimize high cost errors (Sun et al., 2009, Brownlee, 2021).

In addition, further development of the word cloud tool to inspect bags-of-words of each human factor category are suggested. This might also help to understand some infrequent classes. Additionally, adjustments or pre-processing on the format of accident investigation reports could potentially improve the predictions from automated classifiers. The availability of good quality accident reports will also improve the performance of automatic classifiers. For instance, accident reports should have consistent chapter enumeration, only repeated in the summary, or referred in the body text. Section titles should clearly state if the information explain the normal characteristics of the system and it should not mix important information about the accident within normal behaviour. Key information should also be provided in textual format and not only as image. Finally, the public availability of accident reports even if not in English (as translating tools are steadily getting better) would significantly contribute to the knowledge of human error.

6. Conclusions

A virtual human factors classifier based on machine learning has been presented to provide an automatic classification of accident reports involving human error. The approach represents an efficient way of expanding existing human reliability databases based on accident reports analysed by a machine-learning algorithm. The approach has the potential to substitute, or at least support, the classification task normally conducted by a human expert (a time-consuming process that could take weeks, depending on the complexity of the event and on the number of reports or inquiries available). The developed tool provides nearly real-time classification into a specific taxonomy able to classify a two hundred pages report in a minute (an insignificant time compared to the time required for a person to complete the same task).

The findings will be of interest for risk assessors of any industry sector that may need to learn more and faster from major accidents, as automated text analysis can help them to expand their datasets. The presented approach focused at collecting new data for the MATA-D, but the tool can easily be used with other human reliability taxonomy or to be applied to components' reliability data, as long as a labelled dataset is provided together with the text sources.

The case studies showed that the approach is robust and efficient. The performance metrics achieved are satisfactory when compared against human classification and previous studies. In addition, this is the only study which has been trained using reports from different industry sectors, and with a relatively large number of human reliability categories. The results have demonstrated the possibility of using machine-learning based approaches for helping the empirical data collection to improve human reliability analysis, and finally learning lessons from different industry sectors in an efficient and timely way.

Chapter V: Modelling human reliability with confidence

There is one last important issue in modelling data from MATA-D, which occurs in all the current human reliability datasets: the confidence which is dependent on the sample size. The diagram in Figure 3-6 had already posed this dilemma: even collecting more data, as proposed in Chapter 4, the size of the interval probability might be never small enough to make the decision-maker have the necessary confidence to make an informed decision. This is a problem flagged by some empirical HRA data collection projects: the need to understand if small sample sizes are statistically significant – it is important to statistically infer if a human erroneous action has occurred by chance or due to a combination of certain performance shaping factors (Kim, 2020).

The issue is recurrent when modelling safety critical tasks with Bayesian and credal networks using human reliability data: we have more confidence for some combinations in conditional probability tables (CPTs) than for others, due to their different sample sizes. In fact, this is an issue even for CPTs with no missing combination.

Table 5-1 is an example of a conditional probability table from a node in the model mentioned in Chapter 3. This CPT shows that, according to MATA-D, the combination of the variables *inadequate task allocation* and *insufficient knowledge* occurs more often with workers' *faulty diagnosis* than the combination of *communication failure* and *inadequate task allocation*.

Table 5-1. CPT showing samples sizes

Communication failure	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Inadequate task allocation	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
Insufficient knowledge	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
Faulty diagnosis - FALSE	71	15	62	44	1	1	7	6
Faulty diagnosis - TRUE	1	1	7	12	2	3	3	2

However, during the CPTs normalisation process necessary for assessing Bayesian and credal networks, the information regarding the sample size for each combination is lost, as shown in the normalised CPT (Table 5-2).

Table 5-2. Normalised CPT

Communication failure	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Inadequate task allocation	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
Insufficient knowledge	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
Faulty diagnosis – FALSE	0.99	0.94	0.90	0.79	0.33	0.25	0.70	0.75
Faulty diagnosis – TRUE	0.01	0.06	0.10	0.21	0.67	0.75	0.30	0.25

Just looking at the normalised CPT, it could be (wrongly) inferred that the odds of having the combination *communication failure – inadequate task allocation – faulty diagnosis* is slightly higher than the combination *inadequate task allocation – insufficient knowledge – faulty diagnosis*. However, if the decision-maker could see the sample sizes, she/he would infer the opposite. We suggest naming this CPT issue as *data disproportion*.

As small sample sizes and *data disproportion* are a typical problem in human reliability empirical data – for all data collection strategies – it has been decided to develop an approach where decision-makers could see the interval probability together with confidence. The approach suggested is to calculate the reliability by combining credal networks with confidence boxes (c-boxes).

Confidence boxes (c-boxes) have been selected due to their computability, which makes this tool suitable for reliability analysis purposes. The bases for c-boxes are the association of classical notions of confidence (Neyman, 1937), confidence distributions (Cox, 1958), imprecise probability concepts (Walley, 1991) and probability boxes (Ferson et al., 2003).

For instance, **Error! Reference source not found.** and Figure 5-2 below show how confidence can be depicted together with the probability. Our confidence about the probability of an event that has been observed in only one of ten trials (1/10) is not the same as that of an event observed to occur ten times in one hundred trials (10/100).

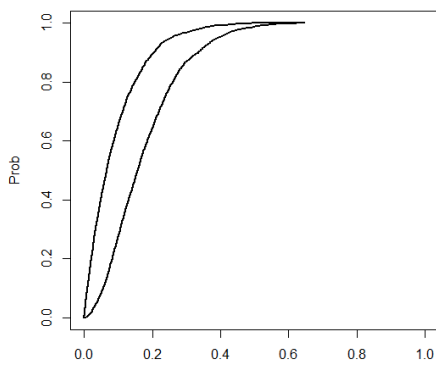


Figure 5-1. C-box of one out of ten trials (1/10)

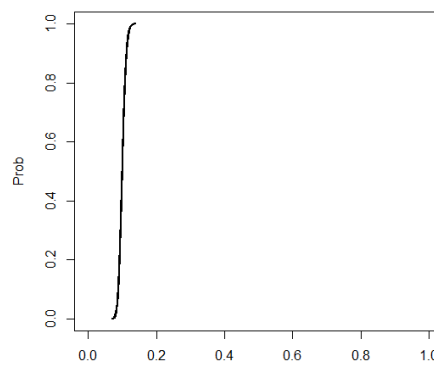


Figure 5-2. C-box of one hundred out of one thousand trials (100/1000)

The skinnier Figure 5-2 depicts that although one event out of ten trials has the same probability as an event that occurred 100 times out of a thousand trials, $P(1/10) = P(100/1000)$, our confidence about the probability $P(100/1000)$ is higher. The breadth between the left and

right bounds in each figure reflects the sample uncertainty, which is a function of sample size in the empirical data.

This research solution is still under development, and yet needs to be fully automatized in OpenCossan software, where credal networks are written. However, OpenCossan is Matlab based, whereas the code for c-boxes is implemented in R. Thus, the first case study to this new approach consists of a very small problem: a network of only three nodes. The ‘toy problem’ chosen is a fatigue model of a worker in offshore leadership: irregular working hours and excessive demand as the influencing factors for leadership fatigue.

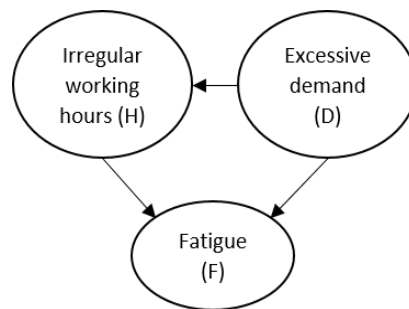


Figure 5-3. A simple model of fatigue.

The idea for the model has originated due to a workshop organised by the safety regulator in the UK, the Health and Safety Executive. In this workshop, HSE has suggested that more research should be addressed to understand how new work arrangements in place (e.g. different shifts and rotas) might increase workers’ fatigue, to help them develop an evidence based guidance for industry (HSE, 2020a).

This last piece of research has been aimed at a conference – and its abstract has been submitted and accepted by the 31st European Safety and Reliability Conference ([ESREL 2021](#)). This piece of work is not published in this thesis, but it is listed in the *List of publications*.

This research solution investigates mainly epistemic uncertainties, e.g. the relations within performance shaping factors, and between those and human errors. However, there are aleatoric uncertainties that might be not captured by this strategy, such as variance between crews’ performance in one specific plant. There are previous work where the estimates’ error bounds address a number of other sources of uncertainty and variability (e.g. model incompleteness, variability across plants and industries, incompleteness of the data), as HRA methods do not limit themselves to treat statistical uncertainty (Hallbert et al., 2006, Greco et al., 2021).

Chapter VI: Conclusion

Aims of the study

The present research has aimed to improve and reinforce the use of human reliability analysis as a tool to predict and prevent accidents in complex socio-technical systems. To achieve this aim, the research has explored a new dataset and some methods to overcome three problems: (i) the lack of confidence in HRA results due to variation and data sparsity, (ii) the over-conservatism of HRA methods due to lack of realistic relations between performance shaping factors and human errors, and (iii) the time-consuming and potentially biased process due to expert elicitation burden in the risk quantification.

Main research findings

This study has shown that MATA-D is a suitable empirical database to be used in human reliability analysis quantification and that modelling it with Bayesian or credal networks enables the assessor to directly infer the impact of performance shaping factors in human error probability. However, there are “implications of using an accident database as the basis for the developed HRA models. The data included in MATA-D is representative of situations that have resulted in accidents. Much of the probabilistic estimates that can be inferred from the database are conditional on the fact that the accident has occurred. Probabilistic risk assessment typically is not conditional on accidents, it’s aim is actually to assess the accident probability. Using conditional estimates in place of marginal ones may very well distort the numeric results”.

One of the most significant findings to emerge from this study is that credal networks might be an interesting alternative to Bayesian networks. Their ability to model imprecision better describes human reliability data which tends to be more uncertain and sparser than component reliability data. If used together with the developed methodology of filling missing combinations of conditional probability tables with intervals that vary from zero to one, credal networks eliminate the need for expert elicitation in the quantification step.

The research has also provided an alternative collection method to decrease the epistemic uncertainty of MATA-D and the time needed to extend and update it, based on natural language processing and machine-learning. Although the fact that a performance shaping factor observed during an accident does not necessarily mean that it is a driver, this assumption was based on the fact that previous study has shown that some combinations are recurrent and do have a pattern (Moura et al., 2017a), being possible to conclude that the combinations are not

governed by aleatory uncertainty alone (true random or uncontrollable processes), but also by epistemic uncertainty (which can be reduced, at least theoretically by collecting new data or using more detailed models) (Patelli et al., 2016).

Implications for the field of knowledge

The thesis argues that the human reliability community might have enough data to rely on data-driven analysis if the right imprecise probability tools are used. The evidence from this study provides insights needed to create the conditions for data-driven human reliability methods. It might be possible that the level of empirical data is already enough to conduct data-driven human reliability if such novel probabilistic tools that accommodate and reflect imprecision are used. The credal network strategy might not only contribute to find human error probabilities, but to adjust them according to the level of each performance shaping factors, in a different methodology than proposed by previous studies (Kim et al., 2018). This adjustment step is usually conducted by assessors according to the rules proposed in HRA methods.

This does not rule out using existing human reliability methods that rely on expert judgement, as they will still be needed to structure the qualitative part of the human reliability analysis, such as modelling the tasks and establishing a framework to classify human errors and performance shaping factors for each task.

The findings will be of interest to safety and risk assessors, as well as decision-makers who have to decide where to allocate resources based on risk assessments. The fact that credal networks provide results with interval rather than point probabilities might improve the transparency of the results and facilitate risk communication between risk assessors and decision-makers.

Although the present research does not include a full human reliability analysis that combines credal networks with c-boxes, the preliminary findings suggests that this combination has the potential to help the HRA community to fight off the fears of using empirical data that might lack statistical power. A natural progression of this work is to automatize an algorithm that combines credal networks with c-boxes, enabling the analysis of more complex human reliability models.

Recommendations for further research work

A possible shortcoming of the strategy of using MATA-D to generate human error probabilities (HEPs) is shown on the discussion following the results in Figure 2.5, where the HEPs from the proposed approach were higher than the CREAM's HEPs estimates provided

by (Hollnagel, 1998): in CREAM all human errors were accounted for, including those that have not produced an accident, thus increasing the HEP denominator ‘opportunities of errors’ (Equation 2-1) and decreasing the resulting HEPs. On the other hand, using MATA-D, we only have accounted for errors that resulted in accidents, consequently scrapping dozens or even hundreds of opportunities of error from the equation, those where the operator or the system have managed to recover the system. This discussion (and possible shortcoming) is also valid for the results obtained with the credal network, although the discussion was not raised in the second manuscript as the validation step was not undergone. Future work could investigate ways of turning such pessimistic denominator into a more realistic one, such as multiplying a correction factor.

Further research could usefully explore different algorithms to decrease time spent on computing credal networks, translations between different HRA taxonomies (thus MATA-D could be used to support other HRA methods), strategies to convert human factors to human reliability data, exploration of different text analytics tools to collect data from new major accident reports, methods to give higher weight to data classified by experts and lower weight to data classified by machine in new versions of MATA-D, collecting the same data from different assessors to understand uncertainties related to the collection step, and to expand the dataset to industry sectors not yet investigated such as the mining industry.

Regarding the automated data collection, it is possible that higher precision and recall might be achieved if using domains specific training sets, as suggested by literature (Brownlee, 2018). Although this clustering was avoided for the data collection, as the aim was to learn from accident occurred in different sectors and therefore training a generic classifier, it is a fact that MATA-D pools together industries and plants with different safety performance standards – such as oil & gas and nuclear industry (Sovacool et al., 2015, Burgherr and Hirschberg, 2008, Ritchie, 2020). Although the assumption used in the thesis was that the human behaviour and its relation with performance shaping factors are similar in industries with similar social-technical complexities, further investigation could be directed to understand if accident data from industries and plants with lower safety standards, are representative for industries and plants with higher safety standards.

Continued efforts are needed to make human reliability analysis part of the risk management toolkit of different industry sectors, enabling informed decisions under uncertainties arising from the complex human-technology-organization interactions.

List of publications

Below is a list of papers published or submitted during this PhD. In total, I have been the leading author of eight research manuscripts – five aimed at conference proceedings, and three aimed at journals. An additional conference paper is being finished and three posters have been presented during conferences at the University of Liverpool. At the end, I have also listed publications which I have worked on as a volunteer.

Papers published in conference proceedings:

Morais, C., Moura, R., Beer, M. and Patelli, E., 2018. Human reliability analysis—accounting for human actions and external factors through the project life cycle. *SAFETY AND RELIABILITY-SAFE SOCIETIES IN A CHANGING WORLD*, pp.329-338.

Morais, C., Moura, R., Beer, M. and Patelli, E., 2018. Attempt to predict human error probability in different industry sectors using data from major accidents and Bayesian networks. *14th Probabilistic Safety Assessment and Management, PSAM 2018*.

Morais, C., Yung, K. and Patelli, E., 2019. Machine-learning tool for human factors evaluation-application to lion air Boeing 737-8 max accident. *In Proceedings of the UNCECOMP 2019 and 3rd ECCOMAS Thematic Conference*.

Morais, C., Tolo, S., Moura, R., Beer, M. and Patelli, E., 2019. Tackling the lack of data for human error probability with Credal network. *In Proceedings of the ESREL*.

Morais, C., Moura, R., Beer, M. and Patelli, E., 2019, Estimativa da probabilidade de erro humano: uma análise da utilização e pesquisa dos métodos de confiabilidade humana, dados disponíveis e técnicas probabilísticas. *In Proceedings of ABRISCO*.

Manuscripts submitted to conference proceedings:

Morais, C., Ferson, S., Moura, R., Tolo, S., Beer, M. and Patelli, E., 2021 (*under review*). Handling the uncertainty with confidence in human reliability analysis . *31st European Safety and Reliability Conference*.

Papers published in academic journals:

Morais, C., Moura, R., Beer, M. and Patelli, E., 2020. Analysis and estimation of human errors from major accident investigation reports. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 6(1). <https://doi.org/10.1115/1.4044796>

Note: This research paper has been submitted on January 2019, published online as author's version on November 2019, and finally had its final version published on March 2020 in the [Special Issue of](#)

Human Performance and Decision Making in Complex Industrial Environments (SI034B) ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems – Part B: Mechanical Engineering.

Morais, C., Estrada-Lugo, D., Jacques, T., Tolo, S., Moura, R., Beer, M. and Patelli, E., 2021 (*in press*). Robust data-driven human reliability analysis using Credal Networks. *Reliability Engineering & System Safety Journal*. <https://doi.org/10.1016/j.ress.2021.107990>

Note: The first version of the second manuscript has been submitted to the [Reliability Engineering & System Safety Journal](#) on 11th December 2020. At the time this thesis has been submitted (before the VIVA), the manuscript was under peer review. However, this thesis was updated with the final version approved by the journal's peer reviewers (which occurred just after the VIVA).

Morais, C., Yung, K., Johnson, K., Moura, R., Beer, M. and Patelli, E., 2021 (*in press*). Identification of human errors and influencing factors: a machine learning approach. *Safety Science Journal*.

Note: The first version of the third manuscript has been submitted to the [Safety Science Journal](#) on 15th February 2020. At the time this thesis has been submitted (before the VIVA), the manuscript was under review. However, this thesis was updated with the final version approved by the journal's peer reviewers (which occurred just after the VIVA). <https://doi.org/10.1016/j.ssci.2021.105528>

White paper for the International Regulators' Forum (IRF) website:

Morais, C., Moura, R., Pires, T., França, M. Human Reliability in the Context of the Offshore Oil & Gas Industry. <https://irfoffshoresafety.com/wp-content/uploads/2020/06/IRF-Article-Human-Reliability.pdf>

Volunteer review of industry guidelines:

Energy Institute, [Report 454: Human factors engineering in projects](#), 2020, ISBN 9781787251991.

Energy Institute, Managing major accident hazard risks (people, plant and environment) during organisational change, 2020, ISBN: 9781787250826.

Center for Chemical Process Safety (CCPS), Project 281. Human Factors for Process Plant Operations: A Handbook, https://www.aiche.org/sites/default/files/docs/pages/2020-2021_ccps_annual_report.pdf

Volunteer Co-creation of research projects:

IM AWARE research project (Informed mining: risk reduction through enhanced public and institutional risk awareness). Research project funded by UKRI/ESRC. Available at: <https://gtr.ukri.org/projects?ref=ES%2FT003537%2F1>. (Awarded)

Risk analysis of plastic pollution in river and marine ecosystems of northeast Brazil (also known as Ocean Clean Up). Research project submitted to NERC/UKRI (not awarded).

Bibliography

- ABDI, H. 2007. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA, 508-510.
- ALAN KEITH, P., AUBREY MAURICE, T., HANS STEFAN LEDIN, H., SAFETY, E., OFFSHORE, D., REDGRAVE, C., BOOTLE, MERSEYSIDE, L. H. S. H. S. L., H, H. & D, S. J. Ignition Hazards and Area Classification of Hydrocarbon Cold Vents by the Offshore Oil and Gas Industry 2012
- ANP 2015. Investigation report of the explosion incident of the explosion incident occurred on 11/02/2015 in the FPSO Cidade de São Mateus Brazil: Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Available at: <https://www.gov.br/anp/pt-br/assuntos/exploracao-e-producao-de-oleo-e-gas/seguranca-operacional-e-meio-ambiente/fpso-cidade-de-sao-mateus> [Accessed 15 February 2021]
- ANP, A. N. D. P., GÁS NATURAL E BIOCOMBUSTÍVEIS. 2020a. RE: Incident Data from Oil and Gas Exploration and Production . Available at: <https://www.gov.br/anp/pt-br/assuntos/exploracao-e-producao-de-oleo-e-gas/seguranca-operacional-e-meio-ambiente/incidentes/dados-de-incidentes-de-exploracao-e-producao-de-petroleo-e-gas-natural> [Accessed 11 December 2020]
- ANP, A. N. D. P., GÁS NATURAL E BIOCOMBUSTÍVEIS. 2020b. RE: Monthly bulletin with data on oil and gas production in Brazil, information on producing states, basins, fields and wells produced. Available at: <https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/boletins-anp/boletim-mensal-da-producao-de-petroleo-e-gas-natural>. [Accessed 11 December 2020]
- ANTONUCCI, A., BRÜHLMANN, R., PIATTI, A. & ZAFFALON, M. 2009. Credal networks for military identification problems. *International Journal of Approximate Reasoning*, 50, 666-679.
- ANTONUCCI, A., DE CAMPOS, C. P., HUBER, D. & ZAFFALON, M. Approximating credal network inferences by linear programming. *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 2013 2013. Springer, 13-24.
- ANTONUCCI, A., HUBER, D., ZAFFALON, M., LUGINBÜHL, P., CHAPMAN, I. & LADOUCEUR, R. CREDO: a military decision-support system based on credal networks. *Proceedings of the 16th International Conference on Information Fusion*, 2013. IEEE, 1942-1949.
- API, A. 2010. API Recommended Practice 754. *Process Safety Performance Indicators for the Refining and Petrochemical Industries*, 1st Ed., American Petroleum Institute, Washington, DC.
- ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCÍA, S., GIL-LÓPEZ, S., MOLINA, D. & BENJAMINS, R. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
- BARLOW, R. E. & WU, A. S. 1978. Coherent systems with multi-state components. *Mathematics of operations research*, 3, 275-281.
- BAYBUTT, P. 2016. Designing risk matrices to avoid risk ranking reversal errors. *Process Safety Progress*, 35, 41-46.
- BBC. 2019. Boeing: Which airlines use the 737 Max 8? BBC, p.Newspaper Article. Available at: <https://www.bbc.co.uk/news/business-47523468> [Accessed 15 February 2021]
- BELL, J. & HOLROYD, J. 2009. Review of human reliability assessment methods RR679. *Health & Safety Laboratory*, 78. Available at: <https://www.hse.gov.uk/research/rrhtm/rr679.htm>. [Accessed 22 January 2019]
- BENCOMO, N. G. P. F. C. H. D. & BLAIR, G. RE: *GeNIe Modeler*. Available at: <https://www.bayesfusion.com/genie/> [Accessed 22 January 2019] BENNETT, P. N., CHICKERING, D. M., MEEK, C. & ZHU, X. Algorithms for active classifier selection: Maximizing recall with precision constraints. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017. 711-719.
- BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

- BOBBIO, A., PORTINALE, L., MINICHINO, M. & CIANCAMERLA, E. 2001. Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability Engineering & System Safety*, 71, 249-260.
- BORING, R. L. & BYE, A. Bridging human factors and human reliability analysis. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2008. SAGE Publications: Los Angeles, CA, 733-737.
- BROUGHTON, E. 2005. The Bhopal disaster and its aftermath: a review. *Environmental Health*, 4, 1-6.
- BROWNLEE, J. 2017. *A Gentle Introduction to the Bag-of-Words Model* [Online]. Available at: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/> [Accessed 24 June 2019].
- BROWNLEE, J. 2018. The Model Performance Mismatch Problem (and what to do about it) [Online]. Available at: <https://machinelearningmastery.com/the-model-performance-mismatch-problem/> [Accessed 27 June 2021].
- BROWNLEE, J. 2021. Cost-Sensitive Learning for Imbalanced Classification [Online]. Available: <https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/> [Accessed 15 February 2021].
- BUCKLAND, M. & GEY, F. 1994. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45, 12-19.
- BURGHERR, P. & HIRSCHBERG, S. 2008. A comparative analysis of accident risks in fossil, hydro, and nuclear energy chains. *Human and Ecological Risk Assessment*, 14, 947-973.
- BYE, A. 2018. Informing HRA by Empirical Data, Halden Reactor Project Lessons Learned and Future Direction. Proceedings of PSAM 14-*Probabilistic Safety Assessment and Management*, 16-21.
- CA AUTHORITY, C. A. 2016. CAP 737. Flight-crew human factors handbook. London: Civil Aviation Authority.
- CAIN, J. 2001. *Planning improvements in natural resource management. guidelines for using Bayesian networks to support the planning and management of development programmes in the water sector and beyond*. Centre for Ecology and Hydrology.
- CANO, A., GÓMEZ, M., MORAL, S. & ABELLÁN, J. 2007. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *International Journal of Approximate Reasoning*, 44, 261-280.
- CCPS, C. F. C. P. S. 2010. Guidelines for Risk Based Process Safety, John Wiley & Sons.
- CGE, R. M. S. 2017. The history of bowtie [Online]. Available at: https://www.cgerisk.com/knowledgebase/The_history_of_bowtie#Introduction [Accessed 11 December 2020]
- CHANG, Y. J., BLEY, D., CRISCIONE, L., KIRWAN, B., MOSLEH, A., MADARY, T., NOWELL, R., RICHARDS, R., ROTH, E. M. & SIEBEN, S. 2014. The SACADA database for human reliability and human performance. *Reliability Engineering & System Safety*, 125, 117-133.
- CHEN, S. H. & POLLINO, C. A. 2012. Good practice in Bayesian network modelling. *Environmental Modelling & Software*, 37, 134-145.
- CHRONOPOULOS, C. & GUZMAN, N. H. C. Is Smartness Risky? A Framework to Evaluate Smartness in Cyber-Physical Systems. 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference, 2020 2020.
- COOPER, S., RAMEY-SMITH, A., WREATHALL, J. & PARRY, G. 1996. A technique for human error analysis (ATHEANA). Nuclear Regulatory Commission.
- COX, D. R. 1958. Some problems connected with statistical inference. *Ann. Math. Statist*, 29, 357-372.
- COZMAN, F. G. 2000. Credal networks. *Artificial Intelligence*, 120, 199-233.

- CSB, U. S. C. S. A. H. I. B. 2011. Investigation Report of the Bayer CropScience Pesticide Waste Tank Explosion. Available at: <https://www.csb.gov/bayer-cropscience-pesticide-waste-tank-explosion/>. [Accessed 15 February 2021]
- CSB, U. S. C. S. A. H. I. B. 2014. Explosion and fire at the Macondo Well. Available at: <https://www.csb.gov/macondo-blowout-and-explosion/>. [Accessed 15 February 2021]
- CULLEN, L. W. D. 1993. The public inquiry into the Piper Alpha disaster. <https://www.hse.gov.uk/offshore/piper-alpha-disaster-public-inquiry.htm> [Accessed 15 February 2021]
- DE VOS, D., DUDDY, M. & BRONNEBURG, J. The problem of inert-gas venting on FPSOs and a straightforward solution. Offshore Technology Conference, 2006 2006. Offshore Technology Conference.
- DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.
- DI FLUMERI, G., DE CRESCENZIO, F., BERBERIAN, B., OHNEISER, O., KRAMER, J., ARICÒ, P., BORGHINI, G., BABILONI, F., BAGASSI, S. & PIASTRA, S. 2019. Brain–Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing With Highly Automated Systems. *Frontiers in human neuroscience*, 13.
- DRUPSTEEN, L., GROENEWEG, J. & ZWETSLOOT, G. I. J. M. 2013. Critical steps in learning from incidents: using learning potential in the process from reporting an incident to accident prevention. *International Journal of Occupational Safety and Ergonomics*, 19, 63-77.
- EASA, E. *International Maintenance Review Board Policy Board* [Online]. <https://www.easa.europa.eu/domains/aircraft-products/international-maintenance-review-board-policy-board-IMRBPB#group-easa-downloads>. Available: <https://www.easa.europa.eu/domains/aircraft-products/international-maintenance-review-board-policy-board-IMRBPB#group-easa-downloads> [Accessed 20th December 2020].
- EI, E. I. & IOGP 2020. Report 454: Human factors engineering in projects. Energy Institute.
- ESTRADA-LUGO, H. D., DE ANGELIS, M. & PATELLI, E. 2019a. Probabilistic risk assessment of fire occurrence in residential buildings: Application to the Grenfell Tower. 13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13. Seoul, South Korea. [10.22725/ICASP13.364](https://doi.org/10.22725/ICASP13.364)
- ESTRADA-LUGO, H. D., SANTHOSH, T. V., DE ANGELIS, M. & PATELLI, E. 2020. Resilience assessment of safety-critical systems with credal networks. Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference, Venice, Italy, 10.3850/978-981-14-8593-0_4192-cd
- ESTRADA-LUGO, H. D., TOLO, S., DE ANGELIS, M. & PATELLI, E. 2019b. Pseudo credal networks for inference with probability intervals. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 5.
- EVANS, J. S. B. T., HANDLEY, S. J. & OVER, D. E. 2003. Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 321.
- FENTON, N. & NEIL, M. 2012. *Risk assessment and decision analysis with Bayesian networks*, Crc Press.
- FERSON, S., KREINOVICH, V., GINZBURG, L. & SENTZ, F. 2003. Constructing Probability Boxes and Dempster-Shafer Structures. Sandia National Labs., Albuquerque, NM (US); Sandia National Labs. SAND2002-4015. Available at: <https://www.osti.gov/servlets/purl/1427258>. [Accessed 15 February 2021]
- FERSON, S., O'RAWE, J. & BALCH, M. 2014. Computing with confidence: imprecise posteriors and predictive distributions. *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*.
- FRENCH, S., BEDFORD, T., POLLARD, S. J. T. & SOANE, E. 2011. Human reliability analysis: A critique and review for managers. *Safety Science*, 49, 753-763.
- GERTMAN, D., BLACKMAN, H., MARBLE, J., BYERS, J. & SMITH, C. 2005. The SPAR-H human reliability analysis method. *US Nuclear Regulatory Commission*, 230, 35.

- GIBSON, W. H. & MEGAW, T. D. 1999. *The implementation of CORE-DATA, a computerised human error probability database*, HSE Books Norwich, UK.
- GOH, Y. M. & UBEYNARAYANA, C. U. 2017. Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, 108, 122-130.
- GOLDBERG, Y. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10, 1-309.
- GONÇALVES, F. C. C. & TRABASSO, L. G. 2018. Aircraft Preventive Maintenance Data Evaluation Applied in Integrated Product Development Process. *Journal of Aerospace Technology and Management*, 10.
- GOOGLE. 2018. Machine Learning Crash Course [Online]. Available at: <https://developers.google.com/machine-learning/crash-course/classification/> [Accessed 15 February 2021].
- GRECH, M. R., HORBERRY, T. & SMITH, A. Human error in maritime operations: Analyses of accident reports using the Leximancer tool. Proceedings of the human factors and ergonomics society annual meeting, 2002 2002. Sage Publications Los Angeles, CA, 1718-1721.
- GRECO, S. F., PODOFILLINI, L. & DANG, V. N. 2021. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliability Engineering & System Safety*, 206, 107309.
- GRIFFITH, C. D. & MAHADEVAN, S. 2015. Human reliability under sleep deprivation: Derivation of performance shaping factor multipliers from empirical data. *Reliability Engineering & System Safety*, 144, 23-34.
- GROTH, K. M. & MOSLEH, A. 2012. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model. Proceedings of the Institution of Mechanical Engineers, Part O: *Journal of Risk and Reliability*, 226, 361-379.
- GROTH, K. M., SMITH, C. L. & SWILER, L. P. 2014. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliability Engineering & System Safety*, 128, 32-40.
- GROTH, K. M., SMITH, R. & MORADI, R. 2019. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. *Reliability Engineering & System Safety*, 191, 106507.
- HALLBERT, B., BORING, R., GERTMAN, D., DUDENHOEFFER, D., WHALEY, A., MARBLE, J., JOE, J. & LOIS, E. 2006. Human event repository and analysis (HERA) system, overview. *US Nuclear Regulatory Commission*, Washington DC, Tech.Rep.NUREG/CR-6903.
- HE, H. & MA, Y. 2013. Imbalanced learning: foundations, algorithms, and applications. ISBN: 978-1-118-07462-6
- HEIDARYSAFA, M., KOWSARI, K., BARNES, L. & BROWN, D. Analysis of Railway Accidents' Narratives Using Deep Learning. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018 2018. IEEE, 1446-1453.
- HENDERSON, J. & EMBREY, D. 2012. Guidance on quantified human reliability analysis. Energy Institute, London. ISBN 978-0-852-93635-1. Available at: <https://publishing.energyinst.org/topics/human-and-organisational-factors/risk-management/guidance-on-quantified-human-reliability-analysis-qhra2> . [Accessed 24 June 2019].
- HENRION, M. Some Practical Issues in Constructing Belief Networks. UAI, 1987. 161-173.
- HOLLNAGEL, E. 1998. Cognitive reliability and error analysis method (CREAM), Elsevier.
- HSE, H. A. S. E. U. 2020a. Optimising Offshore Working Patterns – Shared Research Project [Online]. Available: <https://www.hse.gov.uk/aboutus/assets/docs/shared-research-offshore-working-patterns.pdf>

- [Accessed 15 February 2020]HSE, U. 2010. Assessment of the adequacy of venting arrangements for cargo oil tanks on FPSO and FSU installations. Available at: <https://www.hse.gov.uk/safetybulletins/cargooiltanks.htm> . [Accessed 11 December 2020]
- HSE, U. 2020b. HSE Offshore Statistics, Offshore Hydrocarbon Releases 1992 – 2016. Available at: <https://www.hse.gov.uk/offshore/statistics/index.htm>. [Accessed 11 December 2020]
- HUGHES, P., FIGUERES-ESTEBAN, M. & VAN GULIJK, C. From negative statements to positive safety. 26th European Safety and Reliability Conference, 2017 2017. CRC Press/Balkema, 307.
- ILIEV, R., DEHGHANI, M. & SAGI, E. 2015. Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, 7, 265-290.
- JENTSCH, F. G. 1993. Problems of systematic safety assessments: lessons learned from aircraft accidents. *Verification and Validation of Complex Systems: Human Factors Issues*. Springer.
- JUNG, W., PARK, J., KIM, Y., CHOI, S. Y. & KIM, S. 2020. HuREX—A framework of HRA data collection from simulators in nuclear power plants. *Reliability Engineering & System Safety*, 194, 106235.
- KIM, S. H., LEE, N. & KING, P. E. 2020. Dimensions of religion and spirituality: A longitudinal topic modeling approach. *Journal for the Scientific Study of Religion*, 59, 62-83.
- KIM, Y. 2020. Considerations for generating meaningful HRA data: Lessons learned from HuREX data collection. *Nuclear Engineering and Technology*, 52(8), pp.1697-1705. <https://doi.org/10.1016/j.net.2020.01.034>
- KIM, Y., PARK, J. & JUNG, W. 2017. A classification scheme of erroneous behaviors for human error probability estimations based on simulator data. *Reliability Engineering & System Safety*, 163, 1-13.
- KIM, Y., PARK, J., JUNG, W., CHOI, S. Y. & KIM, S. 2018. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. *Reliability Engineering & System Safety*, 173, 12-22.
- KIRWAN, B. 1994. *A guide to practical human reliability assessment*, CRC press.
- KIRWAN, B. 1997a. Validation of human reliability assessment techniques: part 1—validation issues. *Safety Science*, 27, 25-41.
- KIRWAN, B. 1997b. Validation of human reliability assessment techniques: Part 2—Validation results. *Safety Science*, 27, 43-75.
- KIRWAN, B. & AINSWORTH, L. K. 1992. *A guide to task analysis: the task analysis working group*, CRC press.
- KIRWAN, B., KENNEDY, R., TAYLOR-ADAMS, S. & LAMBERT, B. 1997. The validation of three Human Reliability Quantification techniques—THERP, HEART and JHEDI: Part II—Results of validation exercise. *Applied ergonomics*, 28, 17-25.
- KLETZ, T. Some Common Errors in Accident Investigations. *Safety and Reliability*, 2011 2011. Taylor & Francis, 4-13.
- KLETZ, T. A. 2001. *Learning from accidents*, Routledge.
- KNKT 2019. Aircraft Accident Investigation Final Report Boeing 737-8 (MAX) Lion Mentari Airlines KNKT.18.10.35.04. internet: KNKT.
- KUTER, U., NAU, D., GOSSINK, D. & LEMMER, J. F. Interactive course-of-action planning using causal models. Third International Conference on Knowledge Systems for Coalition Operations (KSCO-2004), 2004 2004. 37-52.
- KYRIAKIDIS, M., MAJUMDAR, A. & OCHIENG, W. Y. 2015. Data based framework to identify the most significant performance shaping factors in railway operations. *Safety Science*, 78, 60-76.
- LAUMANN, K., BLACKMAN, H. & RASMUSSEN, M. 2018. Challenges with data for human reliability analysis. Proceedings of ESREL 2018, 315-321.

- LÉGER, A., WEBER, P., LEVRAT, E., DUVAL, C., FARRET, R. & IUNG, B. 2009. Methodological developments for probabilistic risk analyses of socio-technical systems. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 223, 313-332.
- LEMMER, J. F. & GOSSINK, D. E. 2004. Recursive noisy OR-a rule for estimating complex probabilistic interactions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34, 2252-2261.
- LEVESON, N. 2020. Safety III: A Systems Approach to Safety and Resilience. Available at: sunnyday.mit.edu/safety-3.pdf. [Accessed 15 February 2021]
- LIMA, E. N., BENITES, R. D., MOSLEH, A. & MARTINS, M. R. 2019. A Methodology to Use Multi-Objective Optimization Criteria for an Offshore Topside Production System Since the Early Design Stages, and for The Unit Life Cycle. *Proceedings of the 29th European Safety and Reliability Conference*.
- LIN, S.-W. & BIER, V. M. 2008. A study of expert overconfidence. *Reliability Engineering & System Safety*, 93, 711-721.
- LIU, P. & LIU, J. 2020. Combined Effect of Multiple Performance Shaping Factors on Human Reliability: Multiplicative or Additive? *International Journal of Human-Computer Interaction*, 36, 828-838.
- LOIS, E. 2009. International HRA Empirical Study--phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Similar Performance Data, *Office of Nuclear Regulatory Research*, US Nuclear Regulatory Commission.
- MALATO, G. 2015. Why training set should always be smaller than test set [Online]. <https://towardsdatascience.com/why-training-set-should-always-be-smaller-than-test-set-61f087ed203c> [Accessed 15 February 2021]
- Towards Data Science. Available: <https://towardsdatascience.com/why-training-set-should-always-be-smaller-than-test-set-61f087ed203c> [Accessed 15 February 2021]
- MARKS, S. & DAHIR, A. L. 2020. Ethiopian Report on 737 Max Crash Blames Boeing. p. Newspaper Article. Available at: <https://www.nytimes.com/2020/03/09/world/africa/ethiopia-crash-boeing.html>. [Accessed 15 February 2021]
- MARTINS, M. R. & MATURANA, M. C. 2013. Application of Bayesian Belief networks to the human reliability analysis of an oil tanker operation focusing on collision accidents. *Reliability Engineering & System Safety*, 110, 89-109.
- MATLAB. 2019. Support Vector Machines for Binary Classification [Online]. Available at: <https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html> [Accessed 24 June 2019].
- MATLAB & MATHWORKS. 2018. Matlab documentation for confusion chart function [Online]. Available at: <https://uk.mathworks.com/help/stats/confusionchart.html> [Accessed 15 February 2021]
- MCCALLUM, A. & NIGAM, K. A comparison of event models for naive Bayes text classification. *AAAI-98 workshop on learning for text categorization*, 1998 1998. Citeseer, 41-48.
- MKRTCHYAN, L., PODOFILLINI, L. & DANG, V. N. 2015. Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliability Engineering & System Safety*, 139, 1-16.
- MKRTCHYAN, L., PODOFILLINI, L. & DANG, V. N. 2016. Methods for building conditional probability tables of Bayesian belief networks from limited judgment: an evaluation for human reliability application. *Reliability Engineering & System Safety*, 151, 93-112.
- MORAIS, C., ESTRADA-LUGO, D., JACQUES, T., TOLO, S., MOURA, R., BEER, M. & PATELLI, E. 2021 (*in press*). Robust data-driven human reliability analysis using Credal Networks. *Reliability Engineering & System Safety Journal*. <https://doi.org/10.1016/j.ress.2021.107990>
- MORAIS, C., GARCIA, A., SILVA, B., FERREIRA, N. & PIRES, T. Explaining the explosion onboard FPSO Cidade de São Mateus from Regulatory Point of View. *ESREL - Risk, Reliability and Safety: Innovating Theory and Practice* 25-29 September 2016 2016 Glasgow, Scotland.

- MORAIS, C., MOURA, R., BEER, M. & PATELLI, E. 2018. Attempt to predict human error probability in different industry sectors using data from major accidents and Bayesian networks. 14th Probabilistic Safety Assessment and Management, PSAM 2018.
- MORAIS, C., MOURA, R., BEER, M. & PATELLI, E. 2020. Analysis and estimation of human errors from major accident investigation reports. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 6 (1) : 011014. <https://doi.org/10.1115/1.4044796>
- MORAIS, C., TOLO, S., MOURA, R., BEER, M. & PATELLI, E. Tackling the lack of data for human error probability with Credal network. Proceedings of the ESREL, 2019 2019a.
- MORAIS, C., YUNG, K. & PATELLI, E. Machine-learning tool for human factors evaluation-application to lion air Boeing 737-8 max accident. UNCECOMP 2019 and 3rd ECCOMAS Thematic Conference, 2019 2019b. National Technical University of Athens.
- MOSLEH, A., BIER, V. M. & APOSTOLAKIS, G. 1988. A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety*, 20, 63-85.
- MOURA, R., BEER, M., PATELLI, E. & LEWIS, J. 2017a. Learning from major accidents: Graphical representation and analysis of multi-attribute events to enhance risk communication. *Safety Science*, 99, 58-70.
- MOURA, R., BEER, M., PATELLI, E., LEWIS, J. & KNOLL, F. 2016. Learning from major accidents to improve system design. *Safety Science*, 84, 37-45.
- MOURA, R., BEER, M., PATELLI, E., LEWIS, J. & KNOLL, F. 2017b. Learning from accidents: interactions between human factors, technology and organisations as a central element to validate risk studies. *Safety Science*, 99, 196-214.
- MOURA, R., M., B., E., P., J., L. & KNOLL, F. 2020. Multi-Attribute Technological Accidents Dataset (MATA-D). Available at: 10.17638/datacat.liverpool.ac.uk/1018 . [Accessed 4 February 2020]
- MOURA, R., PATELLI, E., LEWIS, J., MORAIS, C. & BEER, M. 2017c. Human factors influencing decision-making: tendencies from first-line management decisions and implications to reduce major accidents. *Safety and Reliability–Theory and Applications*.
- MURPHY, K. 2007. Software for graphical models: A review. *International Society for Bayesian Analysis Bulletin*, 14, 13-15.
- MYUNG, I. J. 2003. Tutorial on maximum likelihood estimation. *Journal of mathematical psychology*, 47, 90-100.
- NEYMAN, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333-380.
- NIELSEN, T. D. & JENSEN, F. V. 2009. *Bayesian networks and decision graphs*, Springer Science & Business Media.
- NRC, U. N. R. C. 2014. The international HRA empirical study: lessons learned from comparing HRA methods predictions to HAMMLAB simulator data, NUREG-2127. US Nuclear Regulatory Commission, Washington, DC.
- OREDA. Offshore and Onshore Reliability Data [Online]. <https://www.oreda.com/>. Available at: <https://www.oreda.com/> [Accessed 20th December 2020].
- PARK, J. & JUNG, W. 2007. OPERA—a human performance database under simulated emergencies of nuclear power plants. *Reliability Engineering & System Safety*, 92, 503-519.
- PARK, J., KIM, Y. & JUNG, W. Use of a Big Data Mining Technique to Extract Relative Importance of Performance Shaping Factors from Event Investigation Reports. International Conference on Applied Human Factors and Ergonomics, 2017 2017. Springer, 230-238.

- PATELLI, E., ALVAREZ, D. A., BROGGI, M. & DE ANGELIS, M. An integrated and efficient numerical framework for uncertainty quantification: application to the nasa langley multidisciplinary uncertainty quantification challenge. 16th AIAA Non-Deterministic Approaches Conference, 2014. 1501.
- PATELLI, E., GHANEM, R., HIGDON, D. & OWHADI, H. 2016. COSSAN: a multidisciplinary software suite for uncertainty quantification and risk management. *Handbook of uncertainty quantification*, 1-69.
- PATELLI, E., TOLO, S., GEORGE-WILLIAMS, H., SADEGHI, J., ROCCHETTA, R., DE ANGELIS, M. & BROGGI, M. 2018. OpenCossan 2.0: an efficient computational toolbox for risk, reliability and resilience analysis.- Proceedings of the joint ICVRAM ISUMA UNCERTAINTIES conference. Available at: <https://core.ac.uk/download/pdf/201001477.pdf>. [Accessed 20 December 2020].
- PING SHUN, K. 2018. Accuracy, Precision, Recall or F1? [Online]. Available at: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>: Towards Data Science. [Accessed 15 February 2021]
- PIRIE, W. 2004. Spearman rank correlation coefficient. *Encyclopedia of statistical sciences*, 12.
- PODOFILLINI, L. & DANG, V. N. 2013. A Bayesian approach to treat expert-elicited probabilities in human reliability analysis model construction. *Reliability Engineering & System Safety*, 117, 52-64.
- PODOFILLINI, L., MKRTCHYAN, L. & DANG, V. N. 2014. Aggregating expert-elicited error probabilities to build HRA models. *Safety and Reliability: Methodology and Applications*. CRC Press.
- PREISCHL, W. & HELLMICH, M. 2013. Human error probabilities from operational experience of German nuclear power plants. *Reliability Engineering & System Safety*, 109, 150-159.
- PREISCHL, W. & HELLMICH, M. 2016. Human error probabilities from operational experience of German nuclear power plants, Part II. *Reliability Engineering & System Safety*, 148, 44-56.
- PURSEL, M., GANT, S., NEWTON, A., BENNETT, D., O'SULLIVAN, L. & HOOK, P. Investigation of Cargo Tank Vent Fires on the GP3 FPSO, Part 1: Identification of Ignition Mechanisms and Analysis of Material Ejected from the Flare. Proceedings of Hazards 26 Conference, 2016a. Available at: <https://www.icheme.org/media/11721/hazards-26-paper-01-investigation-of-cargo-tank-vent-fires-on-the-gp3-fpso-part-1-identification-of-ignition-mechanisms-and-analysis-of-material-ejected-from-the-flare.pdf> . [Accessed 11 December 2020].
- PURSEL, M., GANT, S., NEWTON, A., BENNETT, D., O'SULLIVAN, L. & HOOK, P. Investigation of Cargo Tank Vent Fires on the GP3 FPSO, Part 2: Analysis of Vapour Dispersion. Proceedings of Hazards 26 Conference, 2016b. Available at: <https://www.icheme.org/media/11830/hazards-26-paper-02-investigation-of-cargo-tank-vent-fires-on-the-gp3-fpso-part-2-analysis-of-vapour-dispersion.pdf>. [Accessed 11 December 2020].
- RAMOS, M., UTNE, I. B., VINNEM, J. E. & MOSLEH, A. 2018. Accounting for human failure in autonomous ships operations. *Safety and Reliability-Safe Societies in a Changing World ESREL 2018*, 355-63.
- RAMOS, M. A., DROGUETT, E. L., MOSLEH, A. & MOURA, M. D. C. 2020. A human reliability analysis methodology for oil refineries and petrochemical plants operation: Phoenix-PRO qualitative framework. *Reliability Engineering & System Safety*, 193, 106672.
- RANGEL, E. & SANGUEDO, C. A. 2018. Considerations on the New Requirements for Electrical Installations in Hazardous Locations. *IEEE Transactions on Industry Applications*, 55 (1). pp.1030-1036.
- RATSABY, J. & VENKATESH, S. S. Learning from a mixture of labeled and unlabeled examples with parametric side information. Proceedings of the eighth annual conference on Computational learning theory, 1995. 412-417.
- REASON, J. 2016. *Managing the risks of organizational accidents*, Routledge.
- RIBEIRO, L. C. F., AFONSO, L. C. S., COLOMBO, D., GUILHERME, I. R. & PAPA, J. P. 2020. Evolving Neural Conditional Random Fields for drilling report classification. *Journal of Petroleum Science and Engineering*, 187, 106846.
- RITCHIE, H. 2020. What are the safest and cleanest sources of energy? [Online]. Available at: <https://ourworldindata.org/safest-sources-of-energy#licence> [Accessed 25 September 2021].

- ROBERTS, M. E., STEWART, B. M. & AIROLDI, E. M. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111, 988-1003.
- ROBINSON, S. D., IRWIN, W. J., KELLY, T. K. & WU, X. O. 2015. Application of machine learning to mapping primary causal factors in self reported safety narratives. *Safety Science*, 75, 118-129.
- SALVI, O. & DEBRAY, B. 2006. A global view on ARAMIS, a risk assessment methodology for industries in the framework of the SEVESO II directive. Elsevier.
- SAMANIEGO, F. J. 1985. On closure of the IFR class under formation of coherent systems. *IEEE Transactions on Reliability*, 34, 69-72.
- SARKAR, S. & MAITI, J. 2020. Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis. *Safety Science*, 131, 104900.
- SHI, H. & LIU, Y. Naïve Bayes vs. support vector machine: resilience to missing data. International Conference on Artificial Intelligence and Computational Intelligence, 2011 2011. Springer, 680-687.
- SHIMAMURA, Y. 2002. FPSO/FSO: State of the art. Springer. *Journal of Marine Science and Technology*, volume 7, pages 59–70
- SHIRAZI, C. H. 2009. *Data-informed calibration and aggregation of expert judgment in a Bayesian framework*. Doctoral dissertation. University of Maryland. <http://hdl.handle.net/1903/9883>
- SIEGRIST, J. 2011. Mixing good data with bad: how to do it and when you should not. *Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management*. Proceedings of the First International Symposium on Uncertainty Modeling and Analysis and Management (ICVRAM 2011) (pp. 368-373).
- SKLET, S. 2006. Safety barriers: Definition, classification, and performance. *Journal of Loss Prevention in the Process Industries*, 19, 494-506.
- SMITH, E., ANNE KOOP, D. N. V. & KING, U. K. S. Guidance on Human Factors Critical Task Analysis. In: ICHEME, ed. Hazards XXII Process Safety and Environmental Protection, 2011. Available at: <https://www.icheme.org/media/9267/xxii-paper-54.pdf>. [Accessed 11 December 2020].
- SOVACOO, B. K., KRYMAN, M. & LAINE, E. 2015. Profiling technological failure and disaster in the energy sector: A comparative analysis of historical energy accidents. *Energy*, 90, 2016-2027.
- STEMPFEL, Y. & DANG, V. N. 2012. Developing and evaluating the Bayesian Belief Network as a human reliability model using artificial data. In: *Proceedings of the European safety and reliability conference (ESREL 2011), September 18–22, Troyes, France; 2011.*
- STRÄTER, O. 2000. Evaluation of human reliability on the basis of operational experience. Doctoral dissertation. Gesellschaft für Anlagen und Reaktorsicherheit (GRS) mbH. ISBN 3-931995-37-2. Available at: <https://www.grs.de/sites/default/files/pdf/grs-170.pdf> [Accessed 5th October 2021]
- SUN, Y., WONG, A. K. & KAMEL, M. S. 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23, 687-719.
- SUNDARAMURTHI, R. & SMIDTS, C. 2013. Human reliability modelling for the next generation system code. *Annals of Nuclear Energy*, 52, 137-156.
- SWAIN, A. D. & GUTTMANN, H. E. 1983. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278, SAND80-0200. Final report. Sandia National Labs. Available at: <https://www.nrc.gov/docs/ML0712/ML071210299.pdf> . [Accessed 11 December 2020].
- TARGOUTZIDIS, A. 2010. Incorporating human factors into a simplified “bow-tie” approach for workplace risk assessment. *Safety Science*, 48, 145-156.
- TOLO, S., PATELLI, E. & BEER, M. Enhanced Bayesian network approach to sea wave overtopping hazard quantification. Proceedings of the 25th European safety and reliability conference, ESREL, Zurich, Switzerland, Sept, 2015. 7-10.

- TOLO, S., PATELLI, E. & BEER, M. 2017. Risk assessment of spent nuclear fuel facilities considering climate change. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3, G4016003.
- TOLO, S., PATELLI, E. & BEER, M. 2018. An open toolbox for the reduction, inference computation and sensitivity analysis of Credal Networks. *Advances in Engineering Software*, 115, 126-148.
- TRBOJEVIC, V. M. 2008. Optimising Hazard Management by Workforce Engagement and Supervision. RR637. Prepared by Risk Support Limited for the Health and Safety Executive, RR637.
- TRIST, E. L. & BAMFORTH, K. W. 1951. Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations*, 4, 3-38.
- TROFFAES, M. C. M. 2007. Decision making under uncertainty using imprecise probabilities. *International journal of approximate reasoning*, 45, 17-29.
- TRUCCO, P., CAGNO, E., RUGGERI, F. & GRANDE, O. 2008. A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliability Engineering & System Safety*, 93, 845-856.
- VINNEM, J. E. 2001. Operational safety of FPSOs: initial summary report, Great Britain, Health and Safety Executive. Available at: <https://www.hse.gov.uk/research/otopdf/2000/oto00086.pdf> . [Accessed 11 December 2020].
- WALLEY, P. 1991. Statistical Reasoning with Imprecise Probabilities. ISBN: 978-0412286605
- WANG, S. I. & MANNING, C. D. Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012 2012. 90-94.
- WAYKOLE, R. N. & THAKARE, A. D. 2018. A Review of feature extraction methods for text classification. *IJAERD*, 4, 351-354.
- WICKENS, C. D., HOLLANDS, J. G., BANBURY, S. & PARASURAMAN, R. 2015. Engineering psychology and human performance, Psychology Press.
- WILLIAMS, J. A data-based method for assessing and reducing human error to improve operational performance. Conference Record for 1988 IEEE Fourth Conference on Human Factors and Power Plants, 1988. IEEE, 436-450.
- WISSE, B. W., VAN GOSLIGA, S. P., VAN ELST, N. P. & BARROS, A. I. Relieving the elicitation burden of Bayesian Belief Networks. BMA, 2008.
- XIANG, Y. & JIA, N. 2007. Modeling causal reinforcement and undermining for efficient CPT elicitation. *IEEE Transactions on Knowledge and Data Engineering*, 19, 1708-1718.
- XING, J., PARRY, G., PRESLEY, M., FORESTER, J., HENDRICKSON, S. & DANG, V. 2016. An Integrated Human Event Analysis Systems (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application, NUREG-2199, Vol. 1. Washington, DC: US Nuclear Regulatory Commission.
- YANG, Z. L., BONSALE, S., WALL, A., WANG, J. & USMAN, M. 2013. A modified CREAM to human reliability quantification in marine engineering. *Ocean Engineering*, 58, 293-303.
- ZHANG, F., FLEYEH, H., WANG, X. & LU, M. 2019. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99, 238-248.
- ZHANG, W., YOSHIDA, T. & TANG, X. 2008. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21, 879-886.
- ZIO, E. 2009. Reliability engineering: Old problems and new challenges. *Reliability Engineering & System Safety*, 94, 125-141.
- ZIO, E. 2018. The future of risk assessment. *Reliability Engineering & System Safety*, 177, 176-190.

ŽUBRINIĆ, K., MILIČEVIĆ, M. & ZAKARIJA, I. 2013. Comparison of Naive Bayes and SVM classifiers in categorization of concept maps. *International journal of computers*, 7, 109-116.

Appendices

Appendix A

Bayesian networks can be represented by acyclic graphs, where nodes are connected to each other by arcs expressing dependencies among variables. The arcs directions must be coherent with the causal relationship of the connected variables. In the BN represented in Figure 0-1, the nodes A and B are called parent nodes of C, which is referred to as their child node. A and B are also called root nodes, as they do not have parents (Tolo et al., 2017). Figure 0-1 shows a graphic representation of the conditional probability expressed in the following equations:

Equation 0-1

$$P(C = c_1 | A = a_1, B = b_1)$$

Equation 0-2

$$P(C = c_2 | A = a_1, B = b_1) = 1 - P(C = c_1 | A = a_1, B = b_1)$$

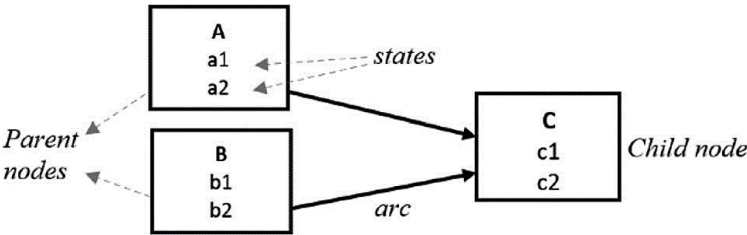


Figure 0-1. Directed acyclic graph of a Bayesian network

Bayes’ theorem expressed in Equation 0-2 provides the mathematical background for joint probabilities modelled by a generic BN with nodes X_1, X_2, \dots, X_n , where p_i refers to the outcomes assumed by the parents of the node X_i , which state is represented by X_i . The joint probability associated with this generic BN is represented by the following equation:

Equation 0-3

$$P(x_1, \dots, x_n) = \prod_i P(x_i | p_i)$$

If all nodes have a binary state, the number of combinations to consider in order to generate a child’s node conditional probability is two (a pair of combinations) to the power of the number of states of the parent nodes. These possible combinations are usually organized in conditional probability tables, as the one represented in Table 0-1.

Table 0-1. Example of CPT for the BN of Fig. 2.10

A	State 1		State 2	
B	State 1	State 2	State 1	State 2
State 1 of C	$P(C=c_1 A=a_1, B=b_1)$	$P(C=c_1 A=a_1, B=b_2)$	$P(C=c_1 A=a_2, B=b_1)$	$P(C=c_1 A=a_2, B=b_2)$
State 2 of C	$P(C=c_2 A=a_1, B=b_1)$ Or $1-P(C=c_1 A=a_1, B=b_1)$	$P(C=c_2 A=a_1, B=b_2)$ Or $1-P(C=c_1 A=a_1, B=b_2)$	$P(C=c_2 A=a_2, B=b_1)$ Or $1-P(C=c_1 A=a_2, B=b_1)$	$P(C=c_2 A=a_2, B=b_2)$ Or $1-P(C=c_1 A=a_2, B=b_2)$

The conditional probabilities represent the strength of the dependencies associated with each cluster of parent-child nodes and it will depend on the structure of the BN, specifically on how the nodes are connected to each other.

The inference computation in BNs can be obtained through some software packages, which allow the adoption of several algorithms, whether exact or approximate (Patelli et al., 2018, Murphy, 2007). Those algorithms and modelling techniques are used as a starting basis and supporting tool for our development, which extrapolates toward an enhanced approach with novel features.

Appendix B¹⁴

Detailed description of tasks, their potential human errors and PSFs and the full correlation table of consequence-antecedent adapted from (Hollnagel, 1998).

Safety Critical task analysis table

The table below provides information about the tasks involved in cargo venting operation, its criticality, the team responsible for each of them, potential human errors associated, performance shaping factors that trigger those human errors (from context and, in case of doubt, from Hollnagel's suggestion in consequent-antecedent links).

¹⁴ Appendices B to H are also available online at Supplementary material:

<https://datanywhere.liv.ac.uk/?linkid=KZi4zr6VWVVMqwMftD1IkpU7sApsZnJC8YDODS6ncAGbVD1eLp6wfg>

Cargo Venting Task	Subtask	Description	Team that performs the task	Human Execution Error	Cognitive function failure	Performance Shaping Factors
1. Verify pressure in the cargo tanks		Check operationality of tank pressure transmitters. Verify tank pressures at the vessel's control room.	Team A (cargo technician)	Not Applicable (NA)	Observation missed (because there was a measurement that was overlooked. No interpretation involved)	Organisational factors: inadequate quality control (check operationality of tank pressure transmitters), maintenance failure (if transmitters are in maintenance backlog), training, insufficient knowledge (no situational awareness/understanding of the context could be an issue) Technological factors: incomplete information (monitoring of tank pressures is performed by the vessel's control room)
2. Check if conditions (weather and simultaneous operations) are favourable for cargo venting operation	2.1 Check wind direction and speed	Prior venting, check the prevailing wind speed and direction in specific Instrument and compare with the tresholds specified in the procedure. Usually wind speed with less than 5knots have stricter operational measures. The direction may impact boat operations, maintenance being carried on the deck and process modules. (N.B. not clear in the procedures where instrumentation is located, thus it was assumed it is in the cargo control room or at the bridge. Something to be checked in a future walk-through).	Team A (cargo technician)		Observation missed	Organisational factors: inadequate task allocation, insufficient skills. Technological factors: incomplete information

<p>2.2 Check any lightning in the near vicinity</p>	<p>Check electrical storms (lightning) in the near vicinity. No instrument is used.</p>	<p>Team A (cargo technician)</p>	<p>Incorrect prediction (fail to anticipate side effects in some cases, speed of development or development of the event)</p>	<p>Organisational factors: insufficient knowledge, inadequate managerial rule, adverse ambient conditions. Potential Individual factors: cognitive bias (due to confirmation bias).</p>
<p>2.3 Check boat and helicopter operation in progress</p>	<p>Check if there are any nearby boat operation in progress. Usually, no instrument is used, but FPSO may have a planned support vessel operation. Check if there are any helicopter operation in progress or planned.</p>	<p>Team A (cargo technician): boats, support vessel Team B (radio-operator)</p>	<p>Observation missed</p>	<p>Potential Organisational factors: inadequate task allocation, adverse ambient conditions, missing information (a plan with helicopters and supply vessels is incomplete or was misunderstood), insufficient skills (in case of nearby boats, some of them are unplanned)</p>
<p>3 Make decision:</p>	<p>If wind speed and direction are not favourable and other simultaneous operation have priority, cargo technician and cargo superintendent have to decide if the tank pressure urgently demands venting or it can be delayed. /'If wind direction is clear off the any boat area, then venting during boat operations may be carried out. If wind is blowing in the same direction and if situation demands for venting then any boat operations will be suspended to facilitate venting'./To support the decision, they can check the procedure, which brings information from risk analysis of weather thresholds, and boat position to be avoided./ If the risk analysis have used the wrong assumptions (e.g. the computational fluid dynamics</p>	<p>Team A: cargo technician and cargo superintendent</p>	<p>Inadequate Plan</p>	<p>Inadequate Procedure, Inadequate Task Allocation, Managerial Rule, task planning and work procedure, Insufficient Knowledge</p>

		used assumes the vented gas is inert with no, the risk analysis and procedure will have the wrong tresholds./It will be up to the Cargo superintendent to evaluate and establish criteria for the periodicity of the operation's routines, including anticipation of the same when weather and simultaneous operations are favourable (plan ahead according to the operational planning and forecast of weather conditions.				
4. In case of unfavourable condition of wind direction and speed, inform teams B and C about the operation.		Inform the Production supervisor, control room operator and radio operator about the operation to be done and the area to be affected according to the wind condition. /The OIM and Production superintendent might decide whether to restrict loading flow./The communication is made by the radio.		Wrong Place (Omission in a sequence of actions)	Faulty Diagnosis (incomplete diagnosis), Inadequate Plan	Inadequate Procedure, Insufficient Knowledge, Missing Information, Inadequate quality control
5. Suspend all the Hot Work Permits, relevant spark potential permits, and any work near the vent riser (or areas affected by venting as reported by Cargo supervisor)	5.1. Request Permit to Work (PTW) to Operation team	When requesting the PTW, inform wind speed and direction and areas to be affected according to procedure.	Team A: cargo technician and cargo superintendent	Wrong Place , Omission in a sequence of actions	Faulty Diagnosis (incomplete diagnosis)	Insufficient Knowledge, communication failure, Inadequate Task Allocation
	5.2 Analyse affected area and affected ongoing and planned works	The Production superintendent usually coordinates the Permits to Work (PTW) and must decide which 'Relevant works' have to stop (those with potential to generate spark). A PTW meeting might be conducted with the affected teams.	Team C: PTW coordinator is usually team C.		Faulty Diagnosis, incomplete diagnosis, Inadequate Plan	Insufficient Knowledge, Inadequate Quality Control, Inadequate Task Allocation

	5.3 Stop relevant works	Maintenance teams have to follow the instructions to stop relevant work until further notice./ The cargo superintendent (team A) usually coordinate the suspension of cargo crane operation if the area is affected.	Team A (cargo superintendent) and D (maintenance teams in affected areas)	Wrong Place (Omission in a sequence of actions)	Cognitive Bias (Illusion of Control)	Missing Information
	5.3 Issue the PTW for cargo venting operation	PTW to cargo venting shall be issue only after all relevant works are stopped.	Team C: PTW coordinator is usually team C.	Wrong Place (Omission in a sequence of actions)		
	5.3 Announce the operation will start on 'PA'	Announce on public announcement (PA) system, on the languages specified on procedure, that cargo tank venting will start and all the personnel must clear off the vent riser area.	Team B: radio operator	Wrong Place (Omission in a sequence of actions)	Distraction	Communication Failure, Inadequate Procedure, Inadequate Quality Control, Maintenance Failure
6. After receiving the Permit to Work, start the Cargo tank venting.	6.1 Open the vent riser valve in the range specified in procedure	Start the operation opening the valve at range specified in the procedure, when the pressure in the tank is at a pressure specified in the procedure. These specifications varies in different wind conditions and under different tank pressures. The indications and actuation of the valves are in the cargo control room screen./To operate correctly the valves must be calibrated and fully functional.	Team A: Cargo technician	Wrong Type (Magnitude)	Faulty Diagnosis, Cognitive Bias	Inadequate Procedure, Inadequate Quality Control, Maintenance Failure, Insufficient Skills, Equipment Failure
7. Remain standby during operation	7.1. Remain in cargo control room until completion of venting.	Cargo technician can only leave if replaced by a colleague from the same team trained in the procedure. (N.B. Monitoring of tank pressures is performed by the cargo control room, that is usually different and	Team A: Cargo technician	Wrong Time	Priority Error, Distraction	Communication Failure, Management Problem

	far from the operation control room. To be checked in walk through.)					
7.1.1. Stay on watch by the radio	Cargo technician must keep attention to radio while performing the operation. If team C informs that level of gas detection is rising to certain threshold, the vent post valve must be closed.	Team A: Cargo technician				
7.1.2. Continuously monitor wind speed and direction	In case of wind change of direction and speed, inform the cargo technician	Team A: Pump man or another cargo technician (different from the one controlling the vent valve at the control room		Observation Missed, Cognitive Bias		Insufficient Knowledge, Inadequate Procedure, Inadequate Task allocation, Inadequate Quality control
7.2. Inform team A if any gas is detected on F&G detectors	If low levels of gas are detected, control room operator must inform on the radio to stop venting, if venting is still necessary, stop production.	Team C: Control room operator	Wrong Time	Observation Missed		Communication Failure, Inadequate Task Allocation, Insufficient Skills, Missing Information
7.2.1 Monitor the fire and gas (F&G) panel during the ventilation and stay on watch by the radio.	Control room operator shall monitor the lower explosive limit (LEL) via the gas detection system (screen at the production control room operation) during the whole cargo venting operation. If a high LEL is attained (usually around 40%), CRO should inform the cargo technician (team C) that venting must be stopped until the gas has been dispersed. Typically, the alarm is activated at 20% LEL and the	Team C: Control room operator		Observation Missed, Cognitive Bias		Insufficient Knowledge, Inadequate Procedure, Inadequate Task allocation, Inadequate Quality control

		executive action, in this case an ESD, is set at 60% LEL.' (de Vos D, Duddy M, Bronneburg J. The problem of inert-gas venting on FPSOs and a straightforward solution. 2006).				
7.3 Inform team A about any incoming boats and helicopter	In case of unexpected incoming helicopter or incoming boats in the affected area, inform the cargo technician. A decision should be made on stopping the cargo venting operation or requesting the boat or helicopter to change route.		Team B: Radio operator	Wrong Time	Faulty Diagnosis	Insufficient Knowledge, Inadequate Procedure, Inadequate Task allocation, Inadequate Quality control
8. Close vent valve when the pressure in tanks drop below the threshold established on procedure.	Close totally the valve when the tank pressure reaches the range specified in the procedure.		Team A: cargo technician	Wrong Type , Magnitude	Faulty Diagnosis, Cognitive Bias	Inadequate Procedure, Inadequate Quality Control, Maintenance Failure, Insufficient Skills, Equipment Failure
9. Inform teams B, C and nearby boats that venting operation has finished, and the vent valve is fully closed.			Team A: cargo technician	Wrong Place , Omission in a sequence of actions	Distraction	Communication Failure, Inadequate Task Allocation ,

Note 1: The most critical tasks, emphasised in bold, are the operational measures to prevent cargo venting related incidents.

Note 2: In doubt on a human error or another, it was assigned the one that would deliver the worst consequence (in a qualitative scale, as suggested by Hollnagel).

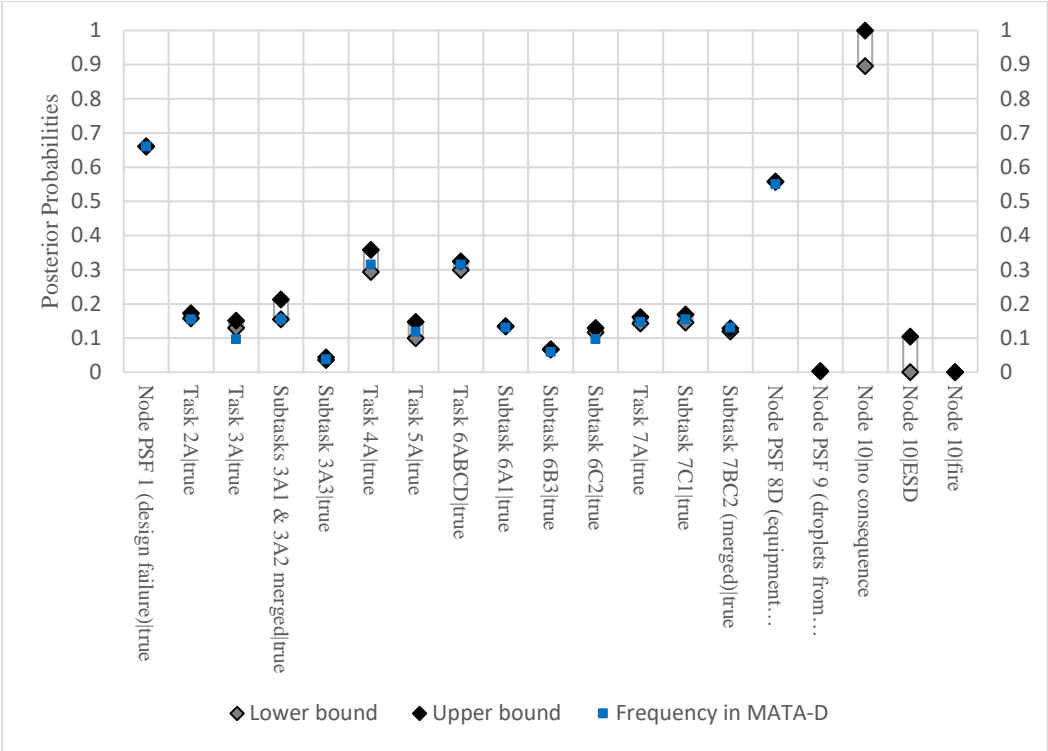
Note 3: Both procedures in Brazil duty holders contained instructions for the crew to inhibit (override) the gas detectors in the affected process area, in case the wind speed was below a threshold (e.g. 2knots) and with certain direction specified in the risk analysis that could reach some process modules. As it is not considered a best safety practice, as it maximises production continuity but minimises safety, it was not included in the task analysis nor in the model. Overriding the detectors means overriding the system automatic action to shut down the production and still leaves the vision and alarm of the gas level to the control room operator, so they can inform the cargo team to stop cargo venting. However, it increases the possibility of human error and accountability of the operator. Finally, to prevent reoccurrence of shutdown related incidents, a number of operational measures are proposed in references regarding this operation (cited in the paper case study session), and none of them proposing the override of a safety barrier such as gas detectors.

Note 4: Some references suggested the action 'select the vent outlet according to the wind direction, always with the objective of removing gases from the installation'. However, it was not considered a critical task analysis as 'downstream and upstream' vent outlets (in regard of the direction of a ship) are not an option for the installations analysed. For those with starboard/port vent outlets (right and left side of the ship), the outlets are so close to each other that does not make a difference for gas dispersion.

Appendix C

Model #2 and its prognostic results – comparing with frequencies obtained in MATA-D. Only state true is shown in the table and figure below.

Classification in CREAM	Event	Lower bound	Upper bound	Frequency in MATA-D
Design failure	Node PSF 1 (<i>design failure</i>) true	0.66	0.66	0.66
Observation missed	Task 2A true	0.1570	0.1727	0.155
Inadequate plan	Task 3A true	0.1293	0.1505	0.097
Observation missed	Subtasks 3A1 & 3A2 merged true	0.1544	0.2128	0.155
Incorrect prediction	Subtask 3A3 true	0.0361	0.0433	0.038
Action in wrong place	Task 4A true	0.2937	0.3581	0.315
Execution of wrong type	Task 5A true	0.0992	0.1469	0.118
Action in wrong place	Task 6ABCD true	0.2993	0.3245	0.315
faulty diagnosis	Subtask 6A1 true	0.1338	0.1338	0.13
distraction	Subtask 6B3 true	0.0668	0.0668	0.059
inadequate plan	Subtask 6C2 true	0.1163	0.1290	0.097
Action performed at wrong time	Task 7A true	0.1421	0.1611	0.147
Observation missed	Subtask 7C1 true	0.1449	0.1691	0.155
faulty diagnosis	Subtask 7BC2 (merged) true	0.1188	0.1281	0.13
Equipment failure	Node PSF 8D (equipment failure) true	0.5576	0.5576	0.55
	Node PSF 9 (<i>droplets from flare</i>) true	0.0027	0.0027	
	Node 10 no consequence	0.8960	0.9998	
	Node 10 ESD	1.72×10^{-4}	0.1040	
	Node 10 fire	2.17×10^{-7}	3.76×10^{-7}	



Appendix D

Data collection code from MATA-Dataset to Conditional Probability Table

This algorithm aims to facilitate the data collection from MATA-D to fulfil the CPT of a Bayesian network. This code imports data from Excel file, the original platform of MATA-D. It is not possible to make this process in Excel, because the CPT template matrix extrapolates the number of columns allowed

Data should be previously treated, which means that each child node must have a worksheet in the spreadsheet with information only from the accidents' lines for this variable and parent nodes. This pre-treated worksheet must have first column with variables' names, first cell must be the child (cell A1) and the following are the parents. The second column until the end (IE..n) will have 0s and 1s related to observations of MATA-Dataset. To do this, dataset might be transposed from original MATA-D.

Before running each child node, change the script according to instructions 'add' beside the commands. The expected results are inside the 'MgenieNormalised' file that will be generated after running this code.

Coded in MATLAB, thus it has to be saved as a matlab file (.m extension) before used.

```
%% Define parameters
Spath='M:\xxxx\xxxx\'; % add here the path where your Excel file is
SfileName='AccidentDatabase.xlsx'; % add here the name of Excel file
SsheetName='DecisionError'; % add here the name of worksheet
Srange='A1:IE2'; % add here the range where data is located in Excel, usually the first column contains
the names of the variables
Vorder=[2 3 4 1]; % add here the number of variables, including the child node. Number one is always
the last. In this case there are four variables, if there were only 2 variables, it would be [2 1]. The first
name in Excel table is the child node.

myTable=readtable(fullfile(Spath,SfileName),'Sheet',SsheetName,...
    'Range',Srange,'ReadVariableNames',false,'ReadRowNames',true);
myArray=table2array(myTable);

% Replace NaN with zeros
myArray(isnan(myArray))=0;

Nvariables=size(myTable,1);
Naccidents=size(myTable,2);

% Rearrange myArray
myArrayOrdered=zeros(size(myArray));
for n=1:Naccidents
    myArrayOrdered(:,n)= myArray(Vorder,n);
```

```

end

Cnames=myTable.Properties.RowNames;
Cnames=Cnames(Vorder);

%Number of combinations
Ncombinations=2^Nvariables;
Mcomparison=zeros(Ncombinations,1);

% Generate CPT
CPT_template=flipud(transpose(fullfact(ones(1,Nvariables)+1)-1));
for n=1:Ncombinations
    Mcomparison(n)=sum(all(myArrayOrdered==CPT_template(:,n)));
end

% Reshape Mcomparison for Genie (as this code has been initially designed to fit GenieModeller
CPTs, but can also be used in OpenCossan)
Mgenie=reshape(Mcomparison,2,length(Mcomparison)/2);
MgenieNormalised=zeros(size(Mgenie));
for n=1:size(Mgenie,2)
    MgenieNormalised(:,n)=Mgenie(:,n)/sum(Mgenie(:,n));
end

% Substitute nan (from normalization of 0 probability) with 0 (especially important step for datasets
with missing data)
MgenieNormalised(isnan(MgenieNormalised))=0;
disp('Variables:')
disp(Cnames)
disp('Normalised CPT')
disp(MgenieNormalised) %this result can be used directly in software like GenieModeller. To use in
OpenCossan, put it in the vertical direction

disp('Counts CPT')
disp(Mgenie) %use this result if you need to see the exact number of combinations, instead the
normalised combinations

```

Appendix E

The complete CPTs for all nodes of model #1.

Node 1 (design failure)

Design Failure - False	0.34
Design Failure - True	0.66

Node 2

Incomplete information	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T
Maintenance failure	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
Inadequate quality	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Insufficient knowledge	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Observation Missed - F	0.92	0.91	0.87	0.81	0.83	1	0.87	0.91	0.38	0.8	0.5	0.71	0	1	0.83	0.78
Observation Missed - T	0.08	0.09	0.13	0.19	0.17	0	0.13	0.09	0.63	0.2	0.5	0.29	0	0	0.17	0.22

Node 32A

Missing information	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T
Inadequate task allocation	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
Insufficient	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Adverse ambient	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Observation Missed - F	0.86	0.89	0.78	1	0.89	0.5	0.79	1	0.83	0	0.5	0	0.88	0	0.82	1
Observation Missed - T	0.14	0.11	0.22	0	0.11	0.5	0.21	0	0.17	0	0.5	0	0.12	0	0.18	0

Node 3.3A

Cognitive bias	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T
Management problem	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
Insufficient knowledge	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Adverse ambient conditions	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Incorrect prediction - FALSE	0.99	0.93	0.91	1.0	1.0	1.0	1.0	0.88	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0
Incorrect prediction - TRUE	0.01	0.07	0.09	0.0	0.0	0.0	0.0	0.13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Node 6ABCD

Wrong Place - 6.1.A	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T		
Wrong Place - 6.2.C	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	
Wrong Place - 6.3.B	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
Cognitive	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Missing information	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Wrong Place - 6.ABCD - False	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Wrong Place - 6.ABCD - True	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Node 61A

Faulty diagnosis of Team A in Task6	F	T
Wrong Place - False	0.73	0.39
Wrong Place - True	0.27	0.61

Node 62C

Wrong Place	F	F	T	T
Inadequate plan	F	T	F	T
Wrong Place - False	1.0	1.0	0.0	0.0
Wrong place - True	0.0	0.0	1.0	1.0

Node 63B

Distraction	F	F	F	F	T	T	T	T
Inadequate procedure	F	F	T	T	F	F	T	T
Maintenance failure	F	T	F	T	F	T	F	T
Wrong Place - False	0.78	0.77	0.65	0.52	0.33	1	0.43	0
Wrong Place - False	0.22	0.23	0.35	0.48	0.67	0	0.57	1

Node DistractionTeamB

Communication failure	F	T
Distraction - False	0.96	0.8
Distraction - True	0.04	0.2

Node FaultyDiagnosisTeamATask6

Communication failure	F	F	F	F	T	T	T	T
Inadequate task allocation	F	F	T	T	F	F	T	T
Insufficient knowledge	F	T	F	T	F	T	F	T
Faulty diagnosis - FALSE	0.99	0.94	0.90	0.79	0.33	0.25	0.70	0.75
Faulty diagnosis - TRUE	0.01	0.06	0.10	0.21	0.67	0.75	0.30	0.25

Node InadequatePlanTask6

Faulty diagnosis	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
Inadequate procedure	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
Inadequate quality control	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	T	T	T	T	T	T	
Inadequate task allocation	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	T	T
Insufficient knowledge	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	
Inadequate plan - false	1	1	1	1	1	0.92	0.88	0.82	0.85	0.8	0.5	0.8	0.71	1	0.67	1	1	1	1	1	0	1	0.5	1	1	0.67	0	0	0	1	1	0	0.5	1	0	0.5	1	0.78	
Inadequate plan - true	0	0	0	0	0.08	0.13	0.18	0.15	0.2	0.5	0.2	0.29	0	0.33	0	0	0	0	0	1	0	0.5	0	0	0.33	0	0	0	0	0	0	1	1	1	1	1	1	0	

Node 7A

Wrong Time (node 7.2.C)	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T		
Wrong Time (node 7.3.B)	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
Priority error	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	T	T	F	T	T
Distraction	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	T	T
Communication failure	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	
Wrong Time - F	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wrong Time - T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Node 71C

Cognitive bias	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
Inadequate procedure	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	
Inadequate quality control	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	T	T	
Inadequate task allocation	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	T	T
Insufficient knowledge	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	
Observation Missed - F	0.89	1	1	1	0.88	0.56	0.94	0.92	1	1	0.65	0.71	0.8	0.67	0.81	0.84	0	0	1	0	1	0	0	1	1	0	0	0	1	1	0	1	0.67	1	
Observation Missed - T	0.11	0	0	0	0.12	0.44	0.06	0.08	0	0	0.35	0.29	0.2	0.33	0.19	0.16	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0.33	0	1	

Node 72C

Observation Missed (node 7.1.C)	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T		
Communication failure	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	
Missing information	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	T	T	T	
Inadequate task allocation	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Insufficient skills	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Wrong Time - F	0.95	0.87	0.94	0.89	0.2	0.8	1	0.93	0.89	1	0	1	1	0.8	0	0	1	1	0	0.86	1	0.8	0.2	0.8	1	0	0.5	0.5	1	0	0.33	0.67	0.33	0.67	0	0
Wrong Time - T	0.05	0.13	0.06	0.11	0.2	0.8	0	0.07	0.11	0.89	1	0	0	0.3	0	0	0	1	0.14	1	0.8	0.2	0.2	0	1	0.5	0.5	1	0	0.33	0.67	0.33	0.67	0	0	

Node 73B

Faulty diagnosis	F	T
Wrong Time - F	0.89	0.61
Wrong Time - T	0.11	0.39

Node FaultyDiagnosisTeamB

Inadequate procedure	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T
Inadequate quality control	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
Inadequate task allocation	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Insufficient knowledge	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Faulty diagnosis - F	0.97	0.75	0.93	0.83	0.93	0.89	0.89	0.81	0	1	0.88	0.88	0	0.6	0.83	0.74
Faulty diagnosis - T	0.03	0.25	0.07	0.17	0.07	0.11	0.11	0.19	0	0	0.12	0.13	0	0.4	0.17	0.26

Node 8

Maintenance failure	F	F	T	T
Inadequate quality control	F	T	F	T
Node8D (Equipment failure) - F	0.56	0.47	0.31	0.33
Node8D (Equipment failure) - T	0.44	0.53	0.69	0.67

Node 9

9. Droplets from flare - FALSE	0.99726
9. Droplets from flare - TRUE	0.00274

Node 10 (consequence node, with 3 states)

5.A. Open (or close) the valve to start (or stop) cargo venting [wrong type]	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
6.ABCD. Suspend sparkable operations [wrong place]	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T
7.A. Remain standby [wrong time]	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	T	T	T	T
8.D. Equipment Ex fails [equipment failure]	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
9. Flare sytem failure [design failure]	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
no_consequence	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ShutDown	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fire OR Explosion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.003376	0.000844	0	0	0	0	0.0000	0	0	0	0	0	0	0	0	0	0	0	0

Appendix F

Find below the complete CPTs for all nodes of model #2 (only the nodes that are different from those in model #1)

Node3.1A&2A (merged)

Incomplete information	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
Missing information	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T		
Inadequate task allocation	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	T	T	T		
Insufficient skills	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	
Adverse ambient conditions	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	
Observation Missed - FALSE	0.87	1	0.81	1	0.94	0.5	0.82	0	0.83	0	0	0	0	0.85	0	0.93	1	0	0.75	1	0	0.5	0	0.64	0	0.7	1	0	0	0	0	0.5	0	1	0	0.57	0
Observation Missed - TRUE	0.13	0	0.19	0	0.06	0.5	0.18	0	0.17	0	0	0	0	0.15	0	0.07	0	0	0.25	1	0	0.5	0	0.36	0	0.3	0	0	0	0	0	0.5	0	0	0	0.43	0

Node6ABCD

Faulty diagnosis	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T		
Inadequate plan	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	
Distraction	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	F	F	T	T	F	F	T	T
Cognitive bias	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	T
Missing information	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T
Wrong Place - F	0.78	0.64	1	0.6	0.67	1	0	0	0	0.5	0.5	0	0	0	0	0	0.5	0.4	1	0	0	0	0.33	0	0.33	0	0.33	0	0	0	0	0	0	0	0.5	0.5	1	0
Wrong Place - T	0.22	0.36	0	0.4	0.33	0	0	0	0.5	0.5	0	0	1	0	0	0	0.5	0.6	1	0	0	0	0.67	0	0.67	0	0.67	0	0	0	0	0	0	0	0.5	0.5	1	0

Node61A

Communication failure	F	F	F	F	T	T	T	T
Inadequate task allocation	F	F	T	T	F	F	T	T
Insufficient knowledge	F	T	F	T	F	T	F	T
Faulty diagnosis - F	0.99	0.94	0.90	0.79	0.33	0.3	0.7	0.8
Faulty diagnosis - T	0.01	0.06	0.10	0.21	0.67	0.8	0.3	0.3

Node62C

Faulty diagnosis	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	
Inadequate procedure	F	F	F	F	F	F	F	F	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T
Inadequate quality control	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	F	F	F	F	T	T
Inadequate task allocation	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	
Insufficient knowledge	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	
Inadequate plan - F	1	1	1	1	0.92	0.88	0.82	0.85	0.2	0.8	0.5	0.8	0.29	0.71	0	1	0.67	0	1	0	1	0	0.5	0	1	1	0.67	0	0
Inadequate plan - T	0	0	0	0	0.08	0.13	0.18	0.15	0.8	0.5	0.2	0.29	0	0.33	0	0.33	0	0	1	1	0	0.5	0	0	0.5	0.33	0	0	

Node63B

Inadequate procedure	F	F	F	F	T	T	T	T
Communication failure	F	F	T	T	F	F	T	T
Maintenance failure	F	T	F	T	F	T	F	T
Distraction - F	0.99	1	0.71	0.5	0.89	0.95	0.8	1
Distraction - T	0.01	0	0.29	0.5	0.11	0.05	0.2	0

Node 7A

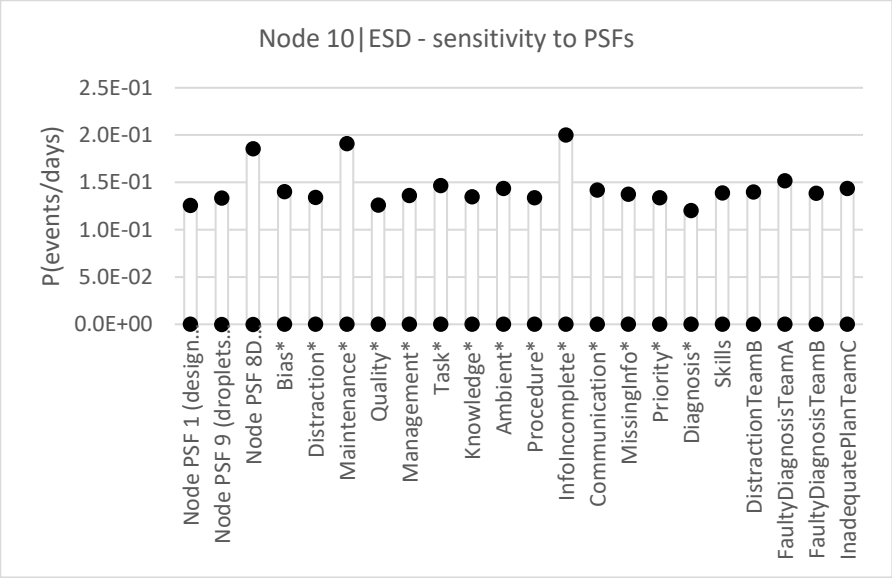
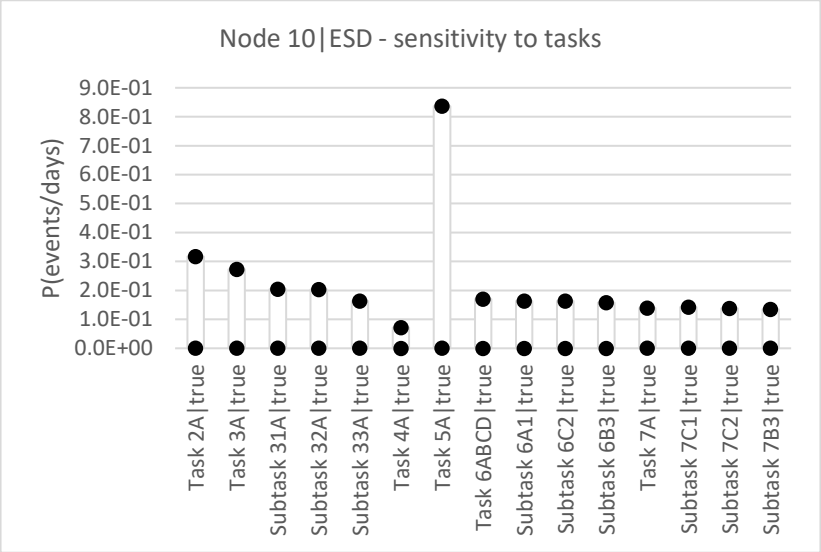
Observation Missed	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T			
Faulty diagnosis	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T			
Priority error	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T	T	T	T	T	T	T			
Distraction	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	T			
Communication failure	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T		
Wrong Time - F	0.94	1	0.5	0	0.13	0.88	1	0	0	0	0.3	0.8	0.67	0.33	0.67	0	0	1	0	0	0	0	0	1	0.33	0.67	0.33	0.67	0.5	0.5	1	0	0.5	0.5	0.67	0.33	0	0	0	0
Wrong Time - T	0.06	0	0.5	0	0.13	0.88	1	0	0	0	0.3	0.8	0.67	0.33	0.67	0	0	1	0	0	0	0	0	1	0.33	0.67	0.33	0.67	0.5	0.5	1	0	0.5	0.5	0.67	0.33	0	0	0	0

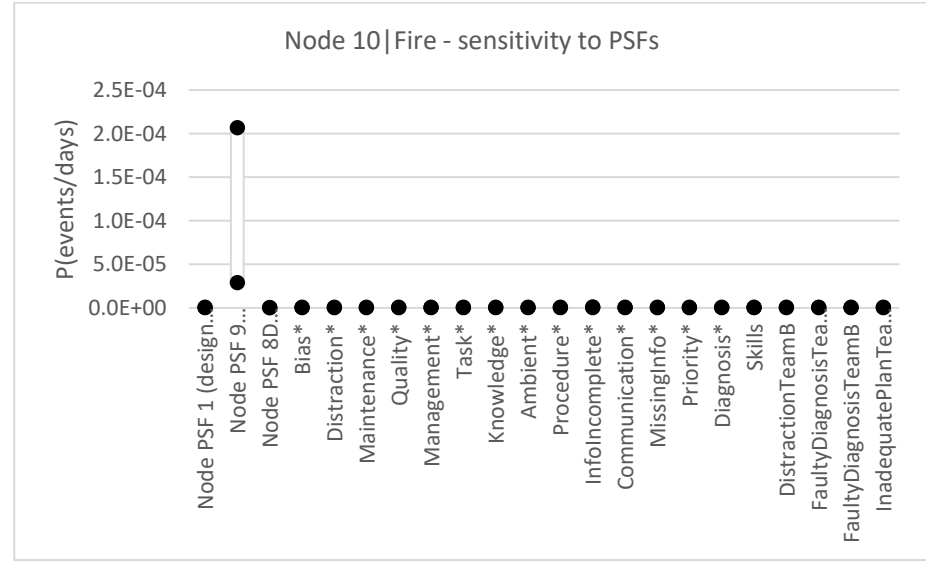
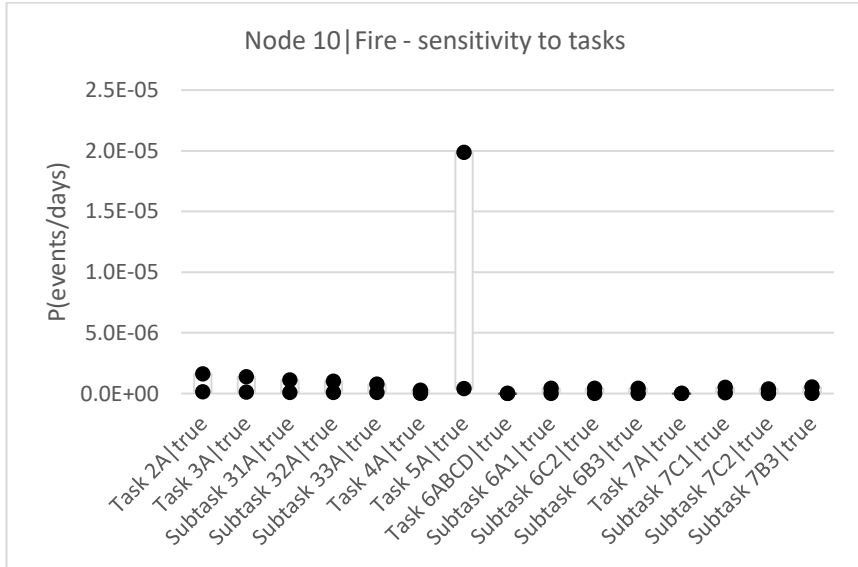
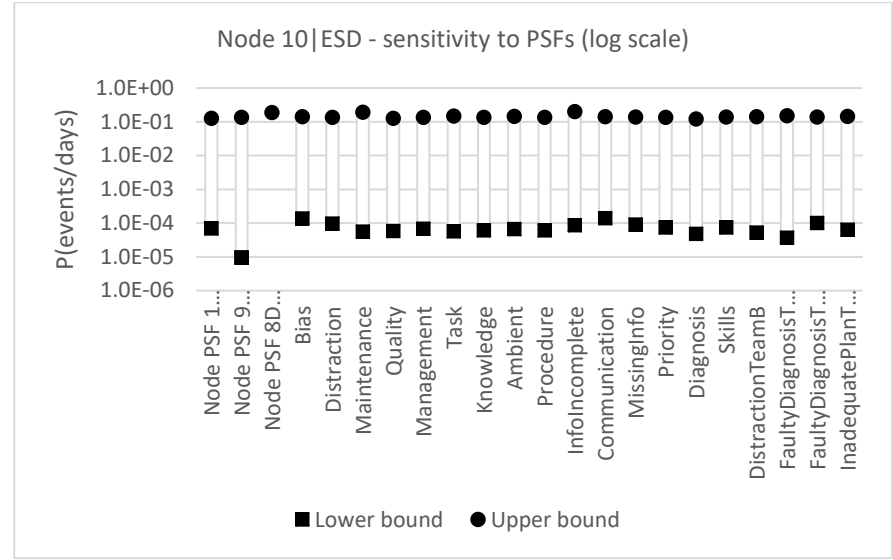
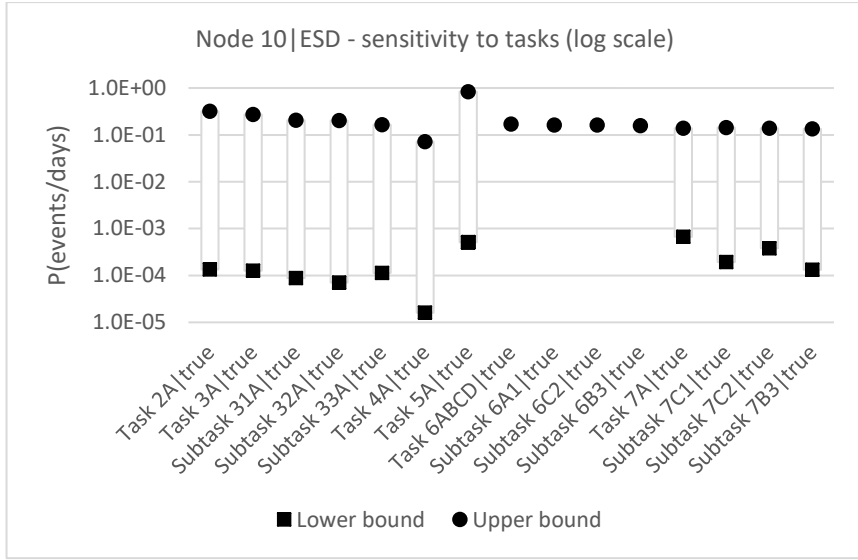
Node72BC

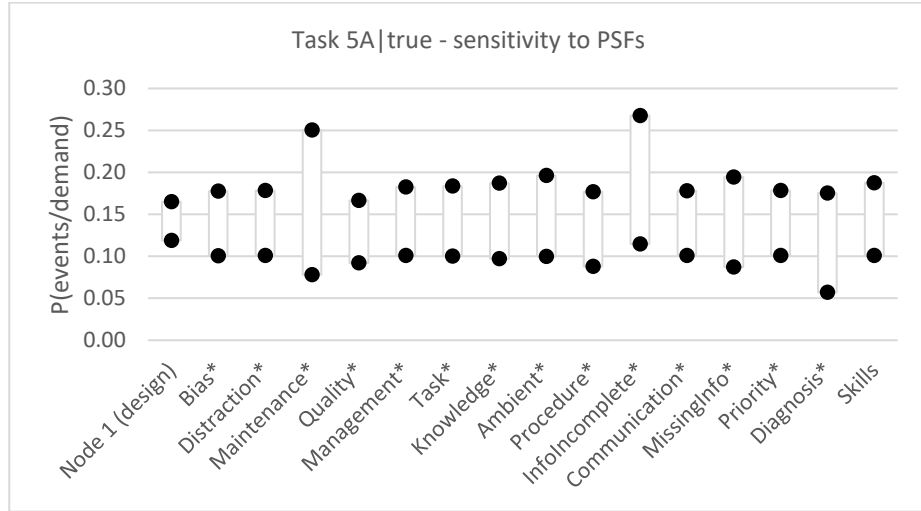
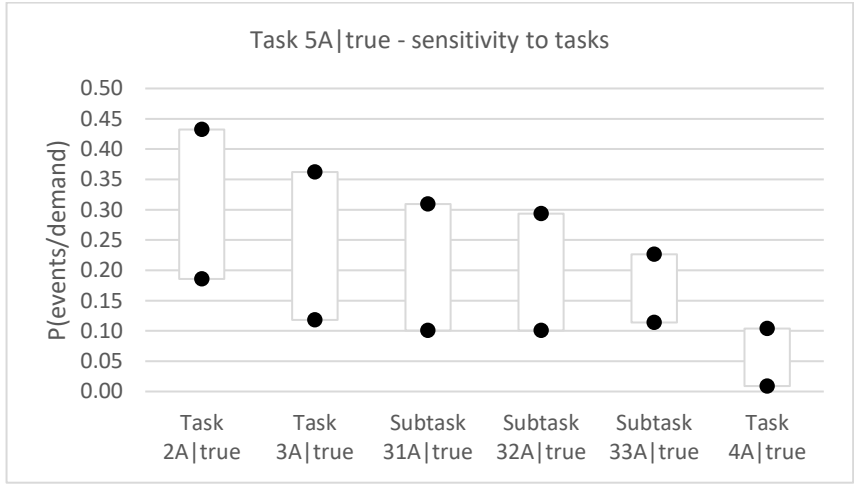
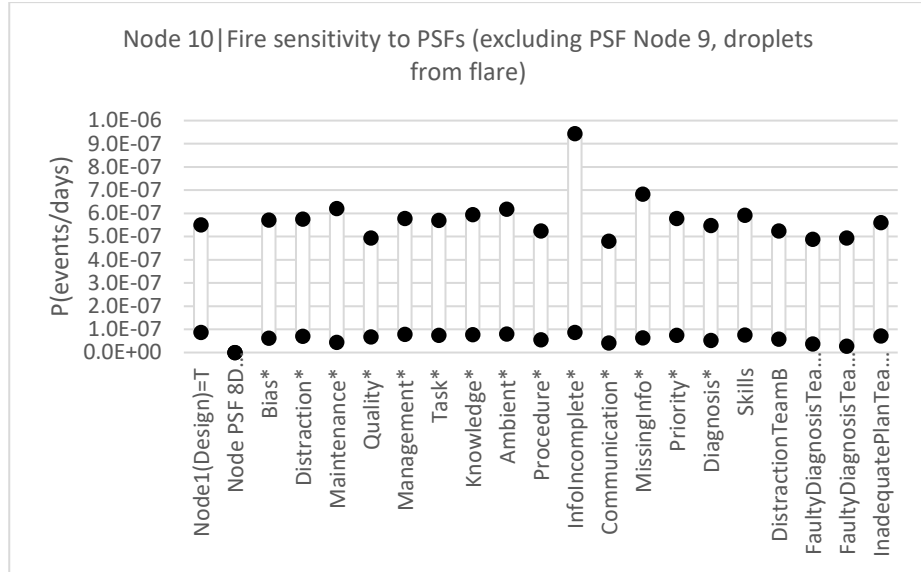
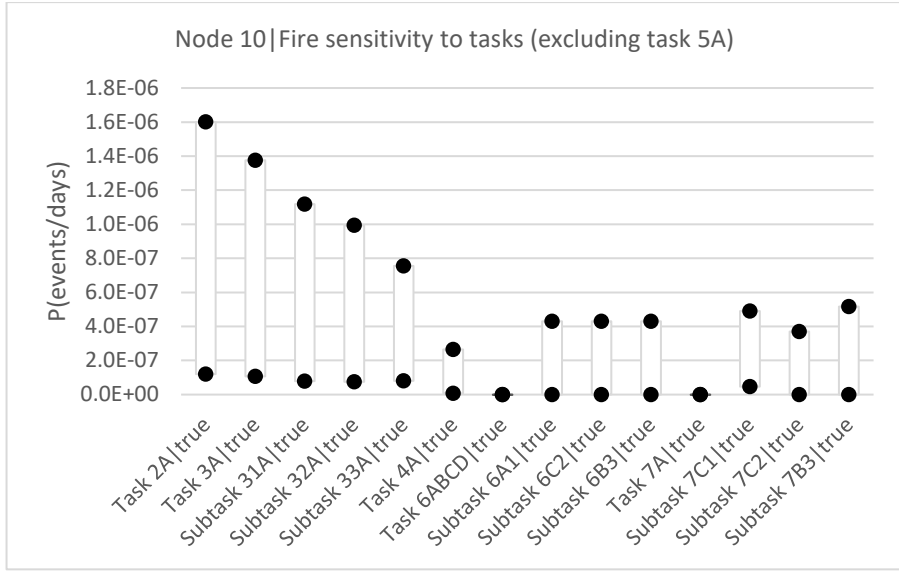
Observation Missed	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T				
Inadequate procedure	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T	T	T	T	T	T		
Inadequate quality control	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	F	F	F	F	T	T	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	T	T
Inadequate task allocation	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T	T	T	
Insufficient knowledge	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	
Faulty diagnosis - F	1	0.8	1	0.17	0.83	0.96	1	0.11	0.89	0.2	0.8	1	1	0.09	0.91	0.17	0.83	0	1	0	0	0	0.2	0.8	0.3	0.8	0	1	1	0	0	0	0	0.17	0.83	0	1	1	1	0.33	0.67	0.6	0.4
Faulty diagnosis - T	0	0.3	0	0.17	0.83	0.04	0	0.11	0.89	0.2	0.8	1	1	0.09	0.91	0.17	0.83	0	1	0	0	0	0.2	0.8	0.3	0.8	0	1	1	0	0	0	0.17	0.83	0	1	1	1	0.33	0.67	0.6	0.4	

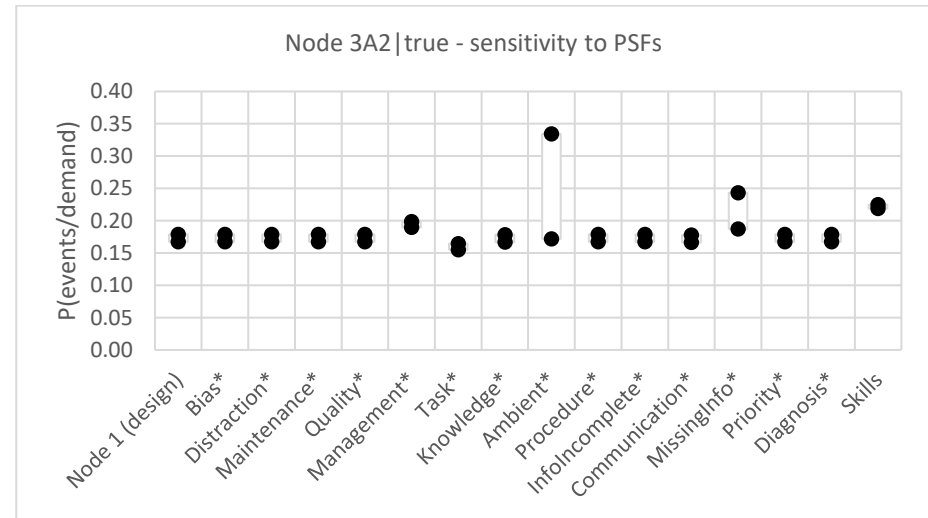
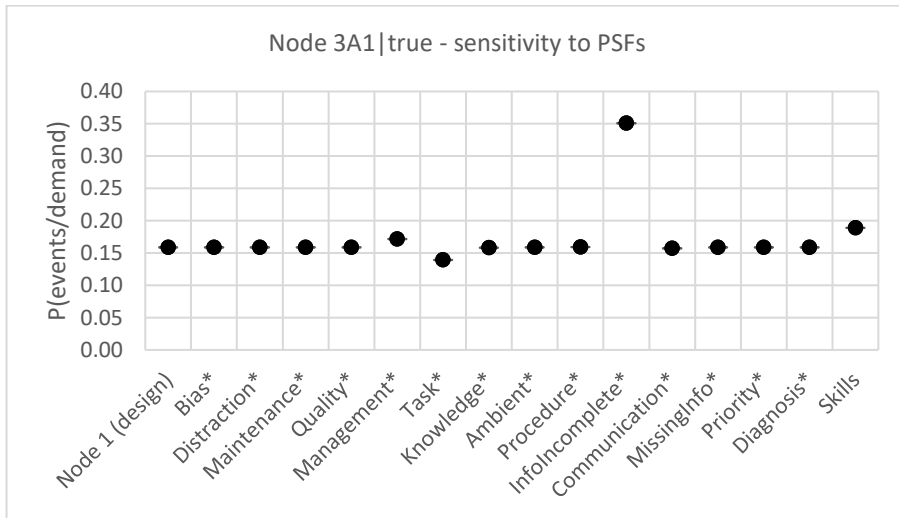
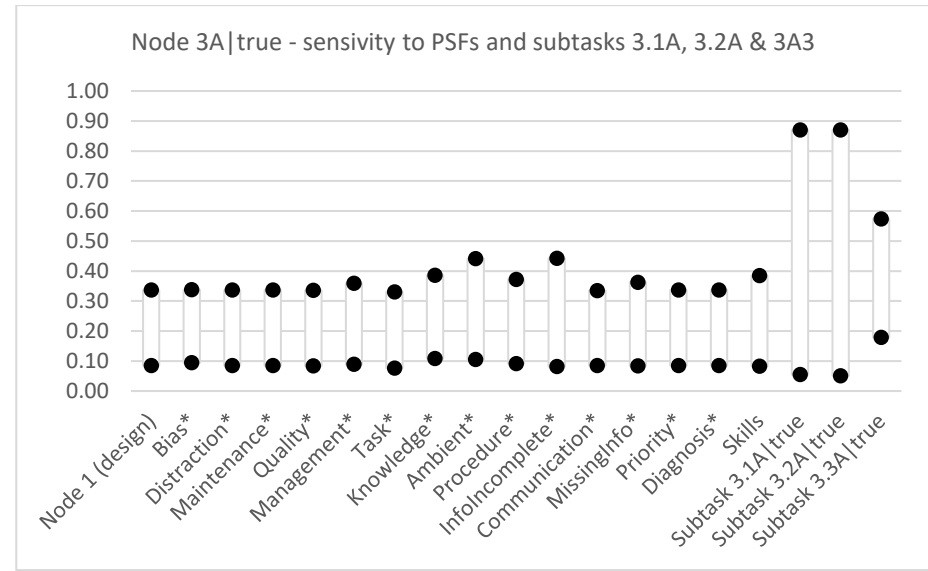
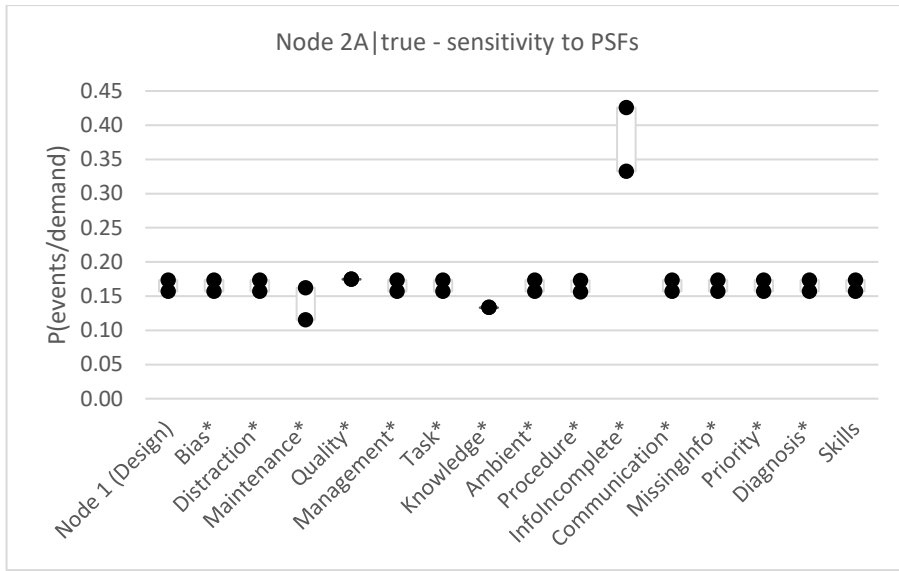
Appendix G

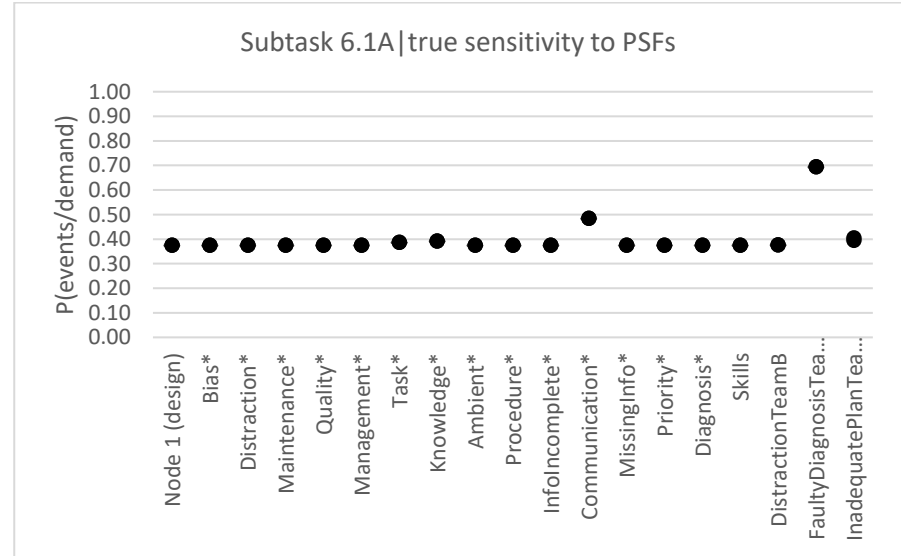
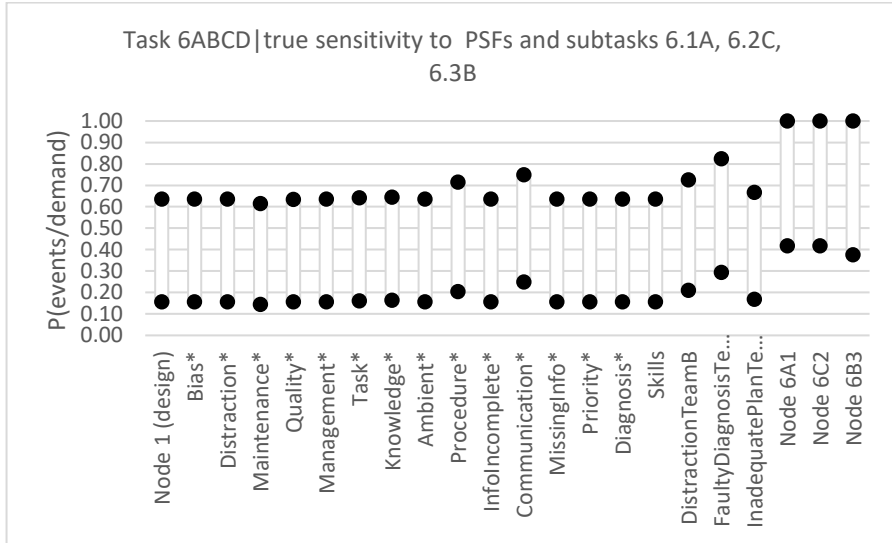
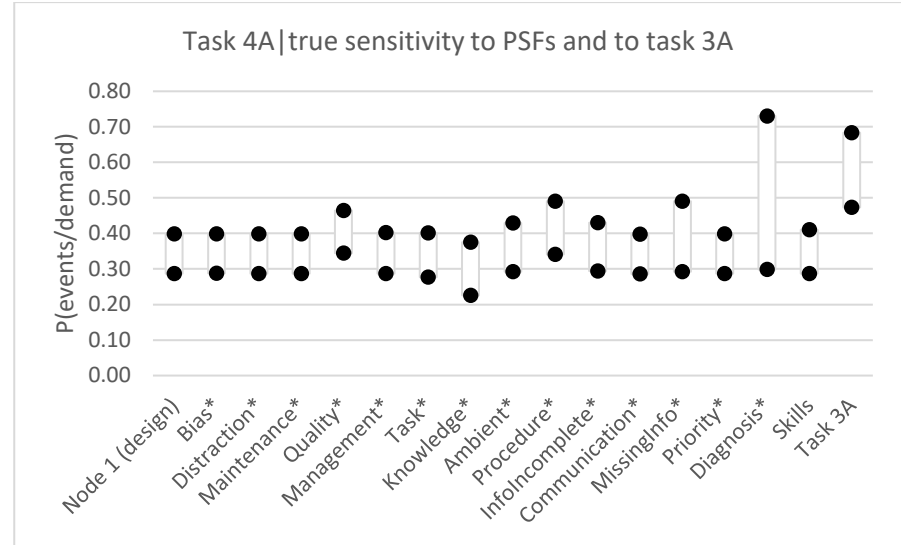
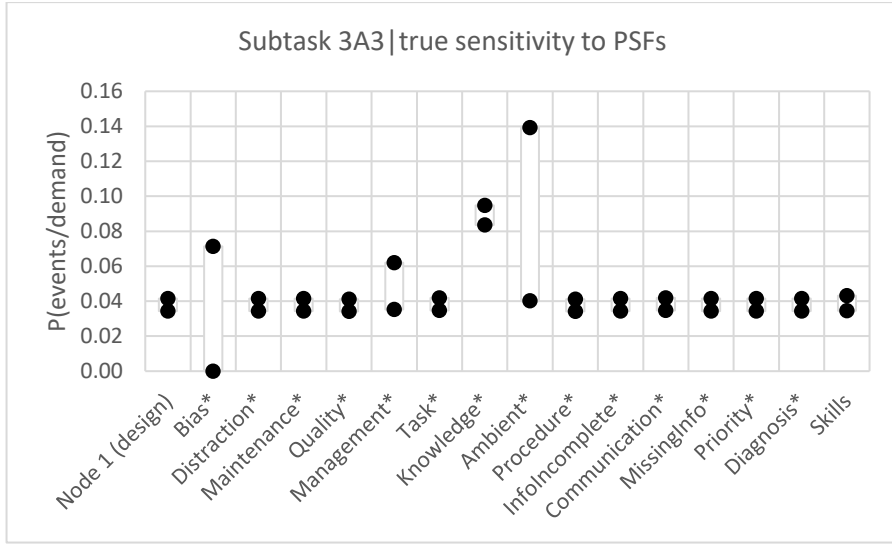
All graphs for diagnostic analysis simulated for Model #1.

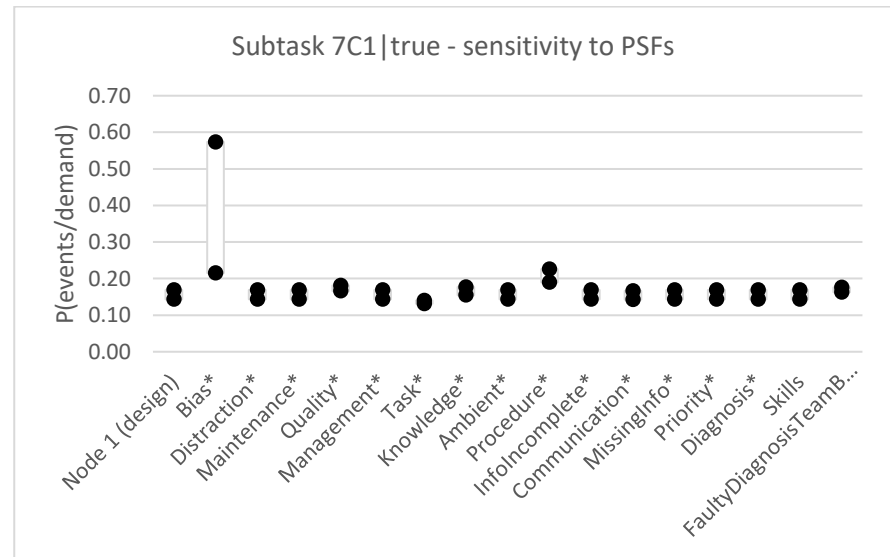
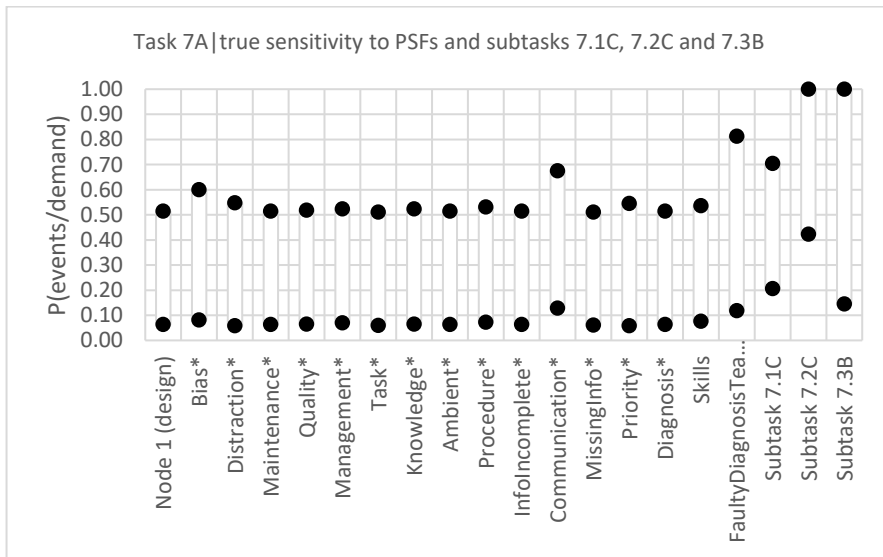
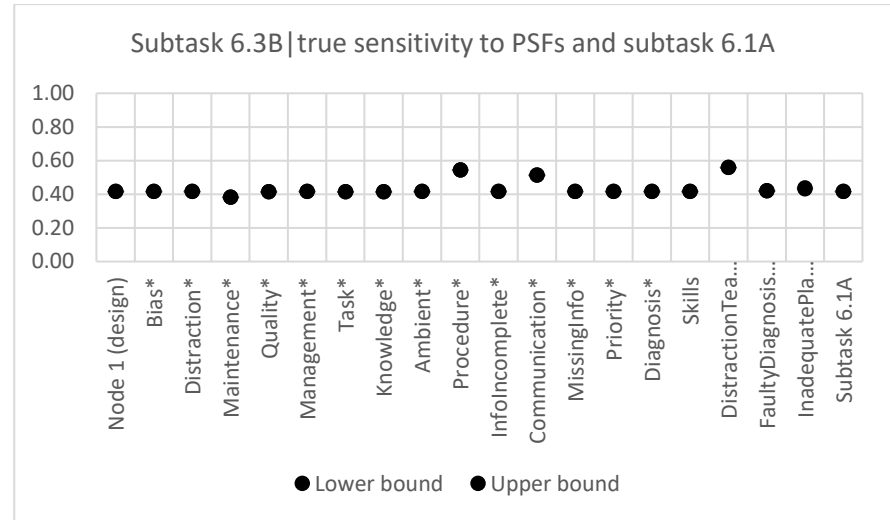
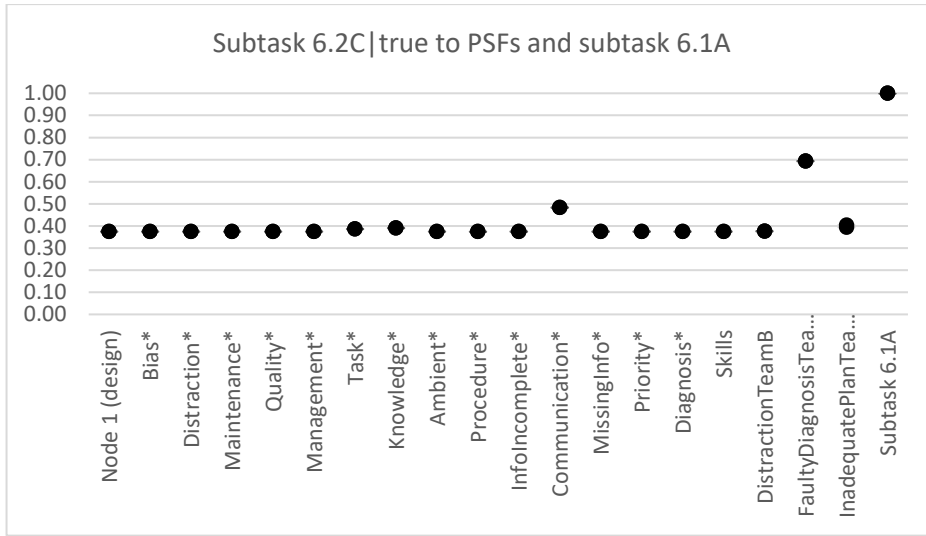


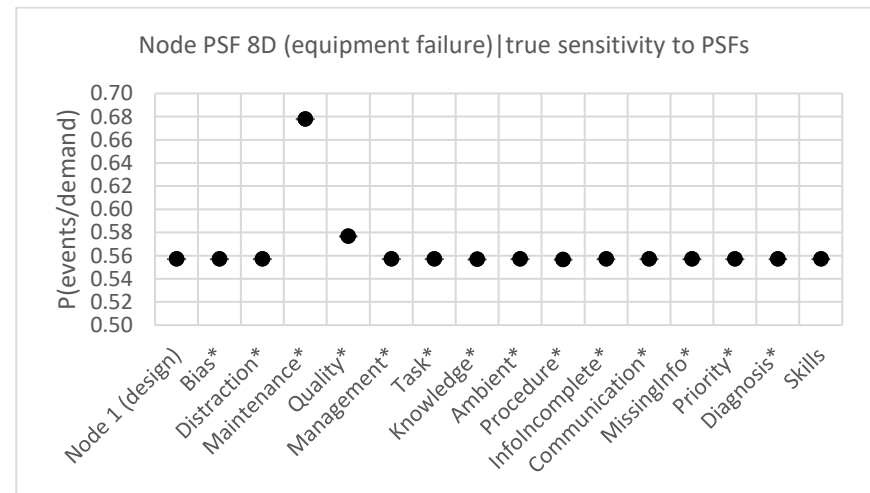
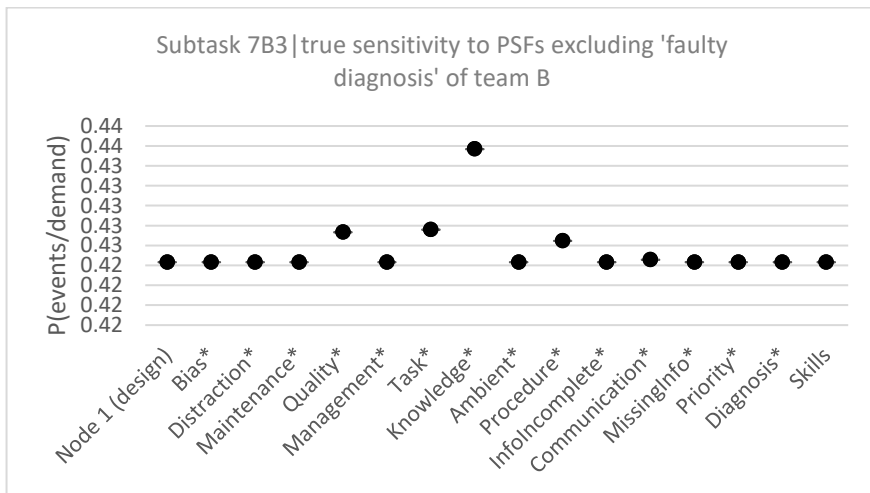
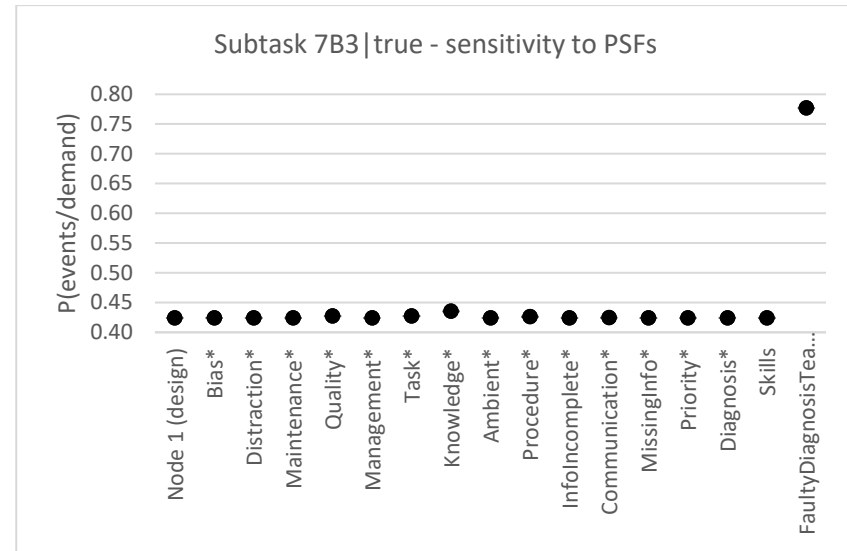
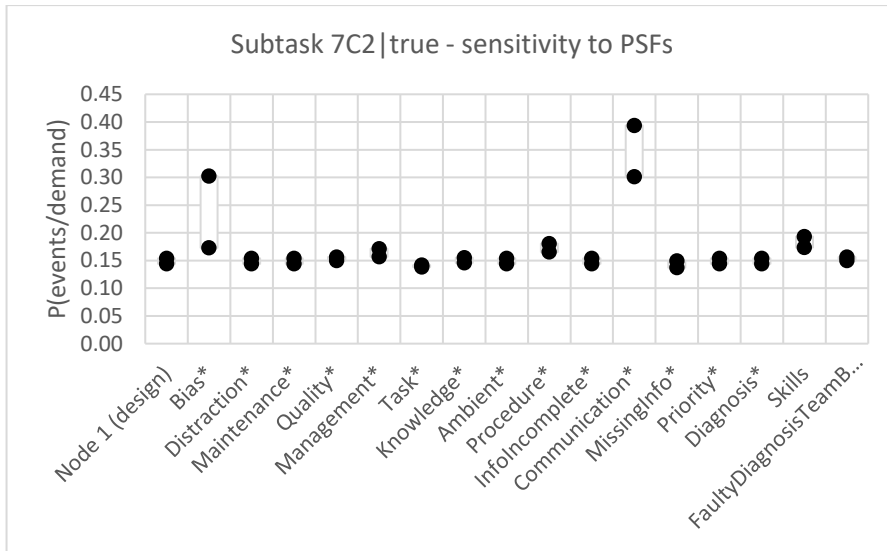


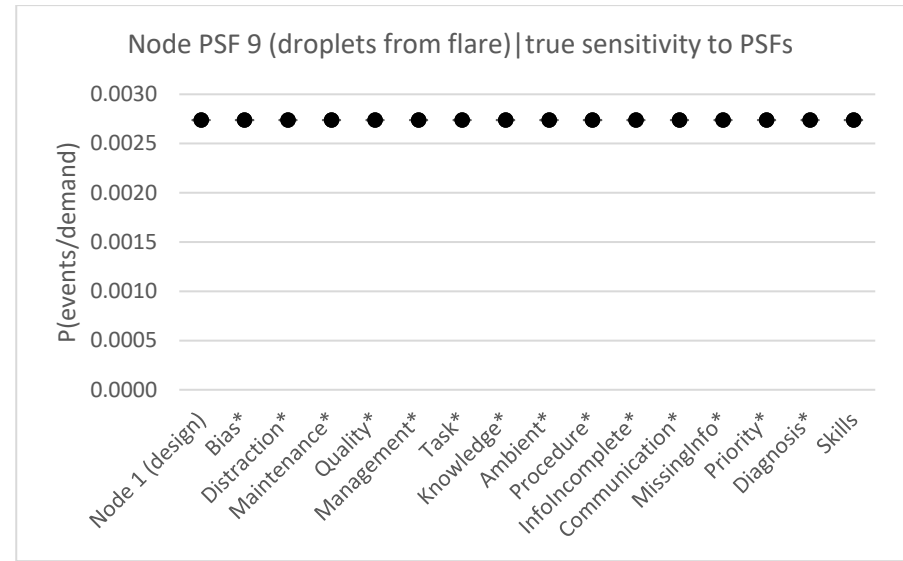
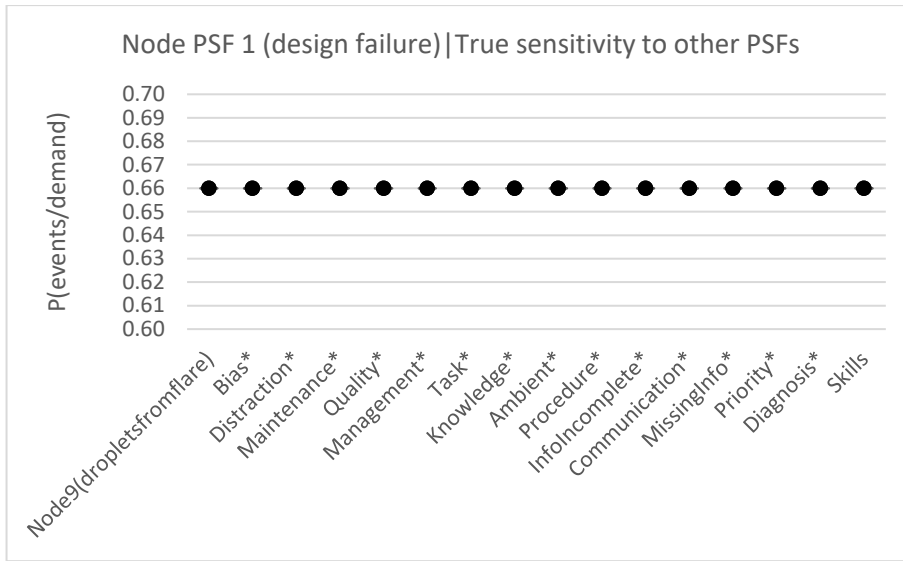






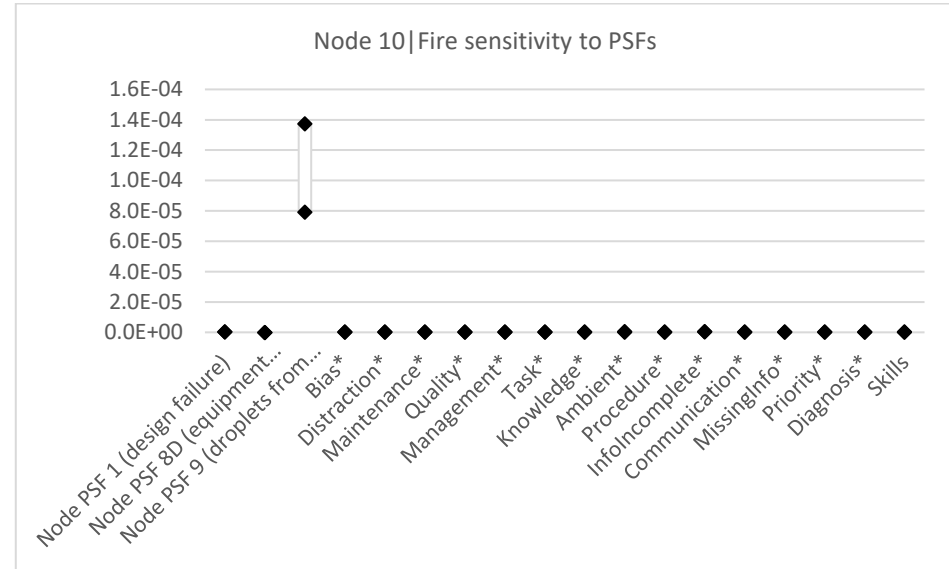
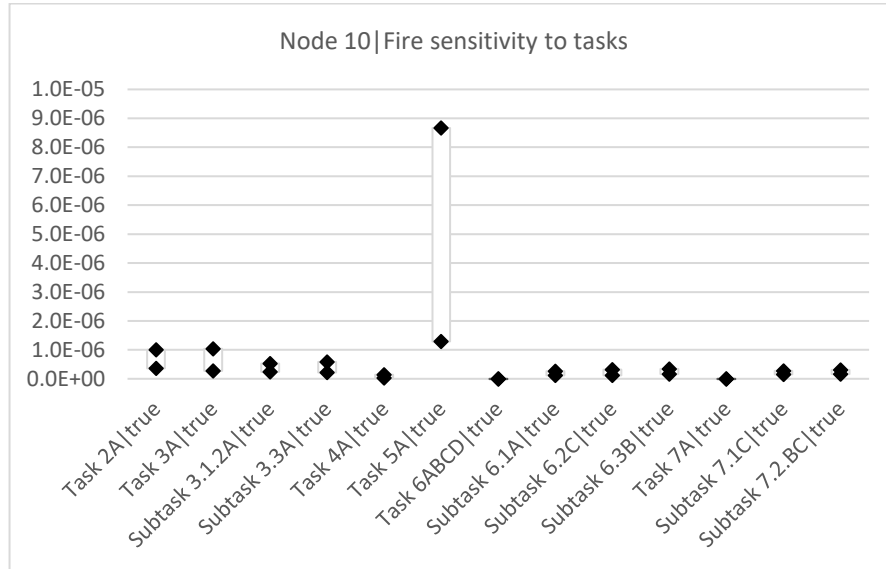


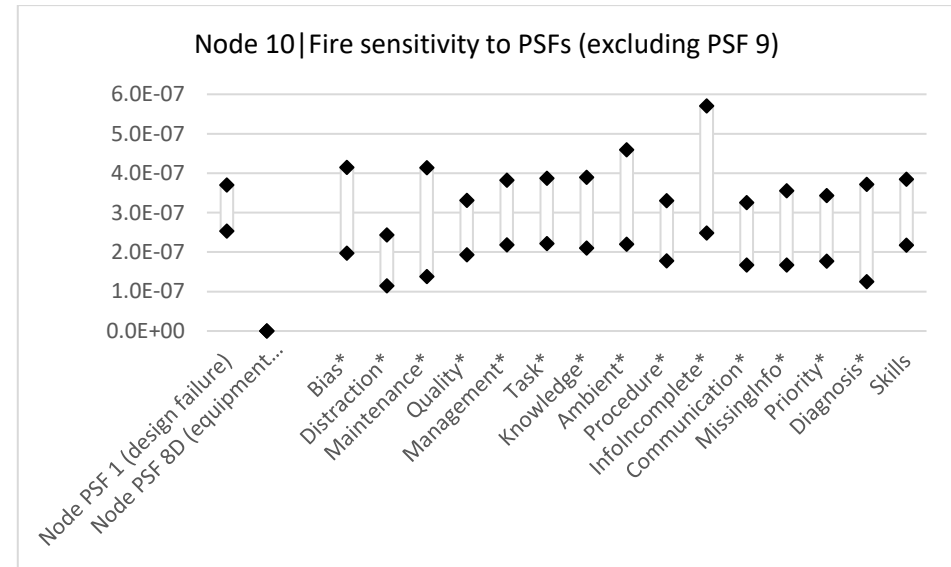
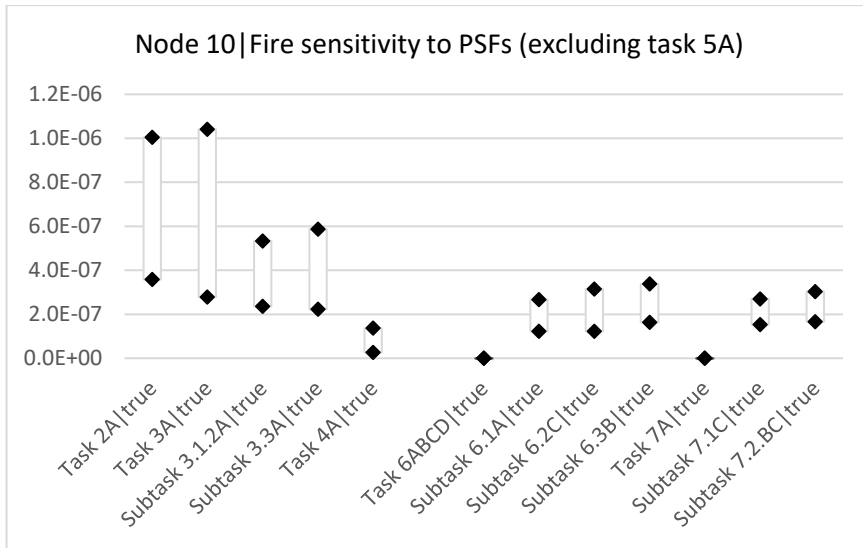


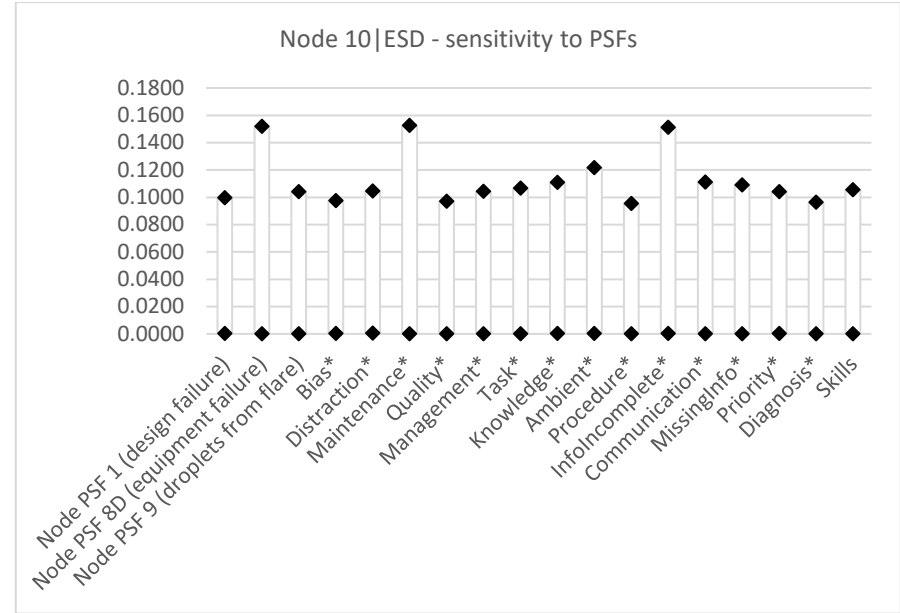
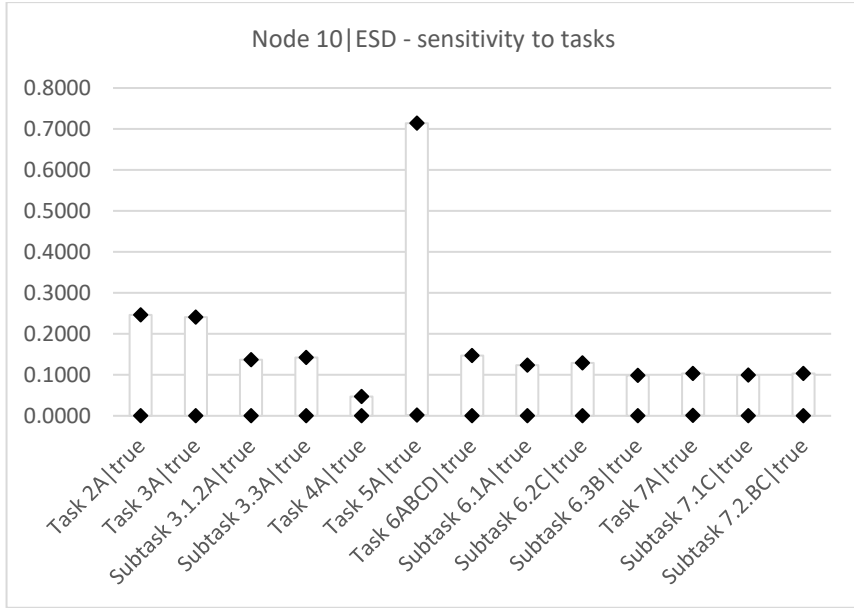


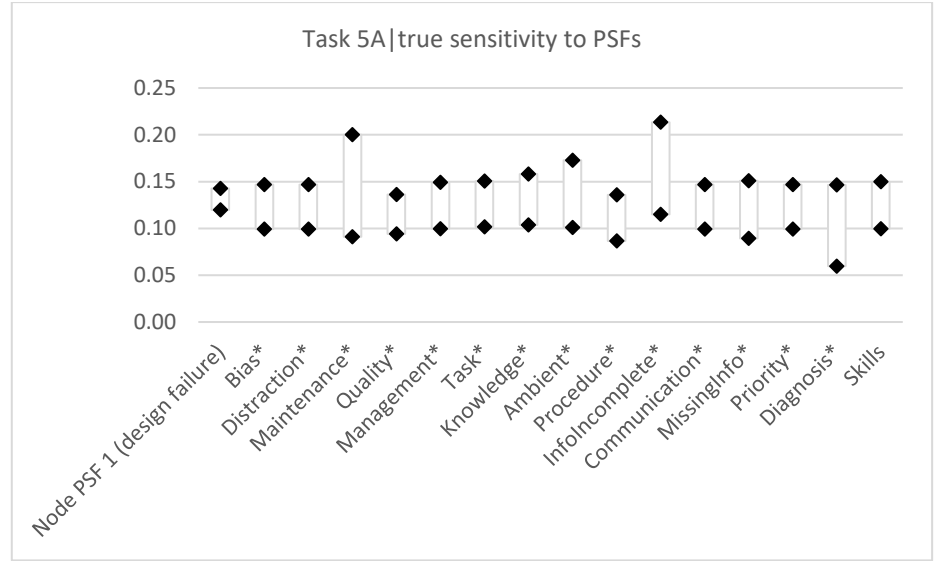
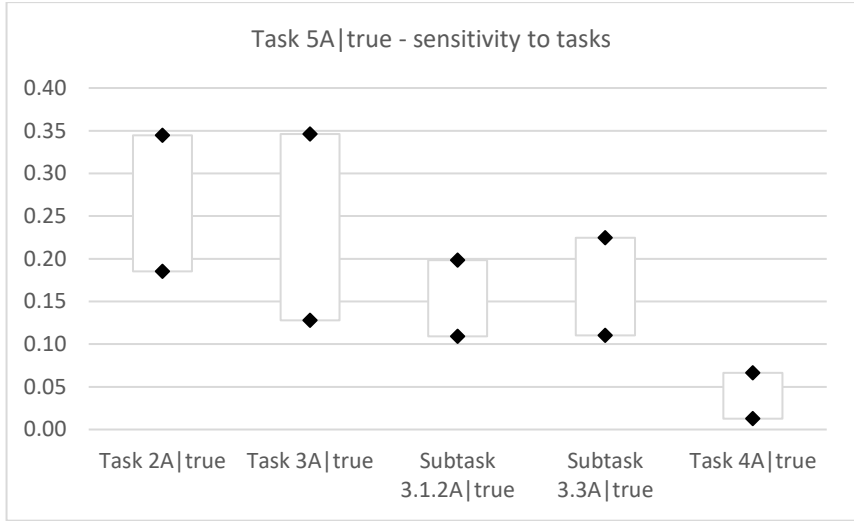
Appendix H

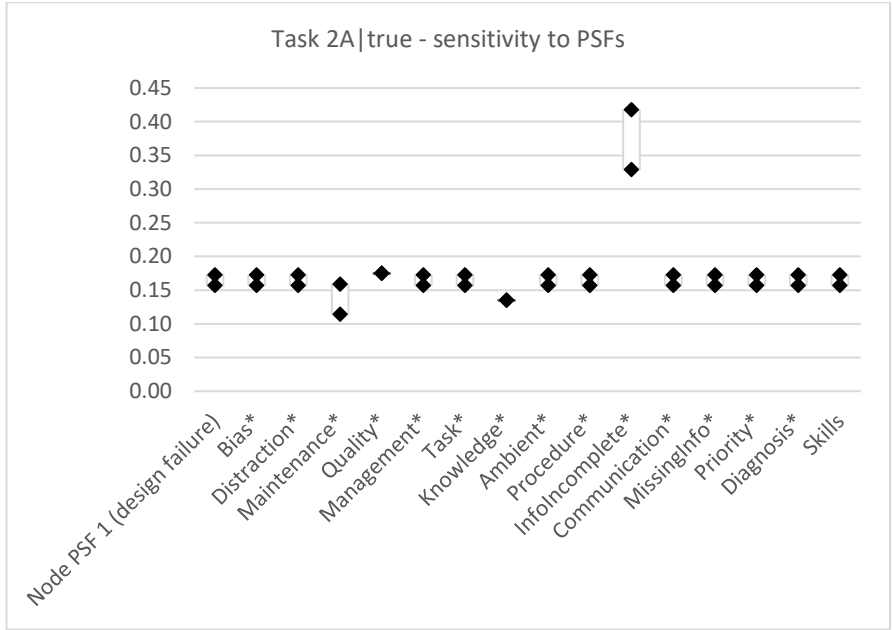
All graphs for diagnostic analysis simulated for Model #2.



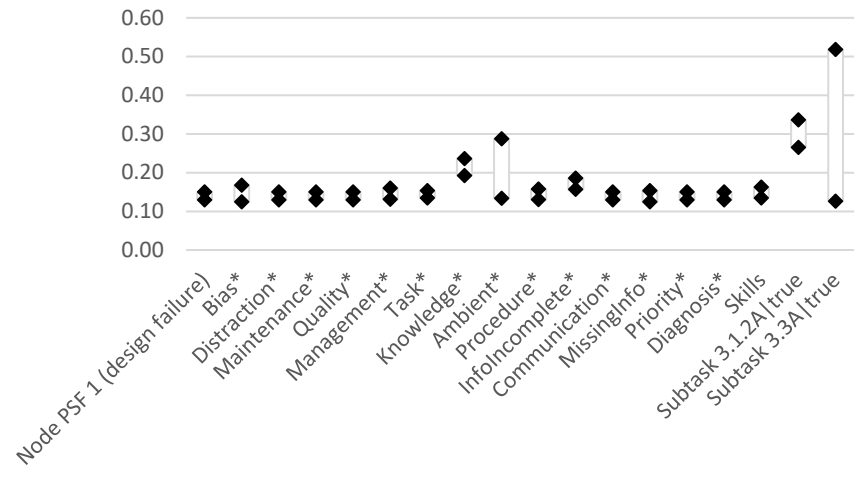


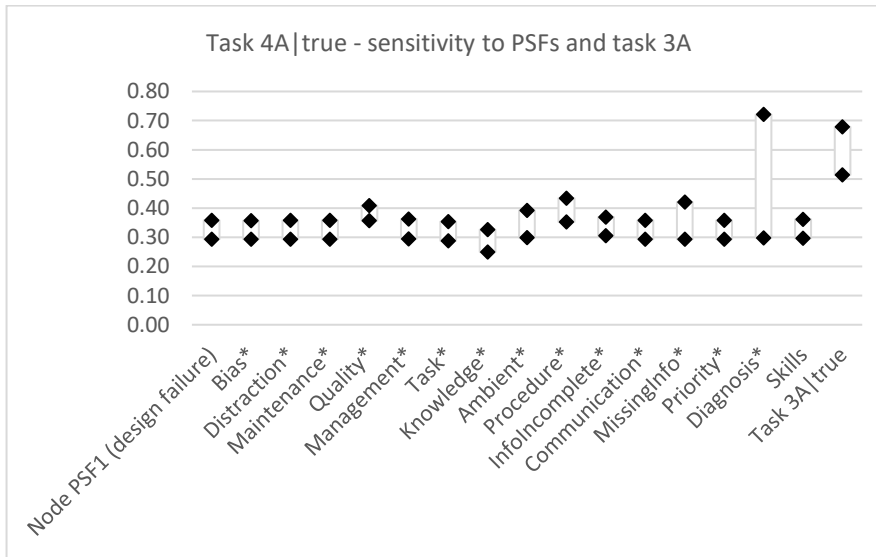
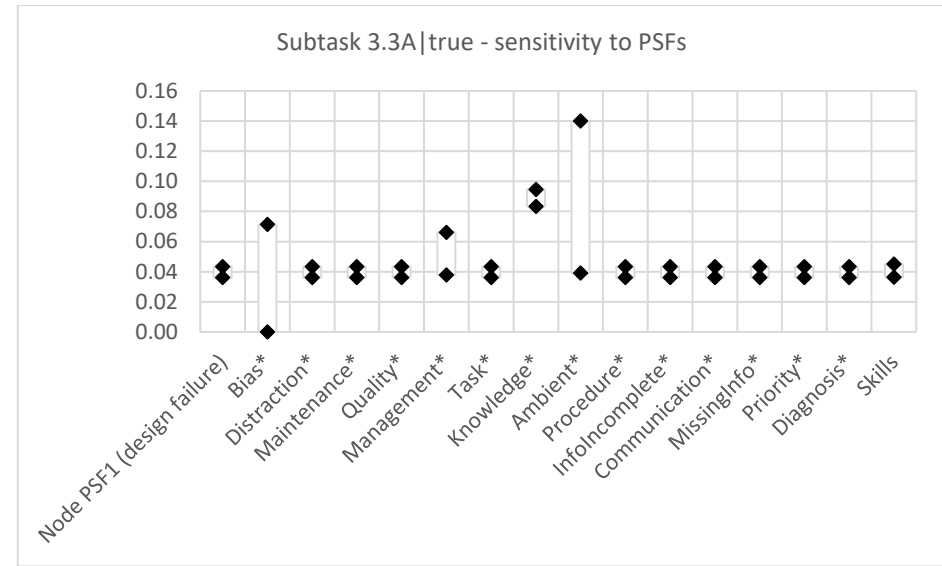
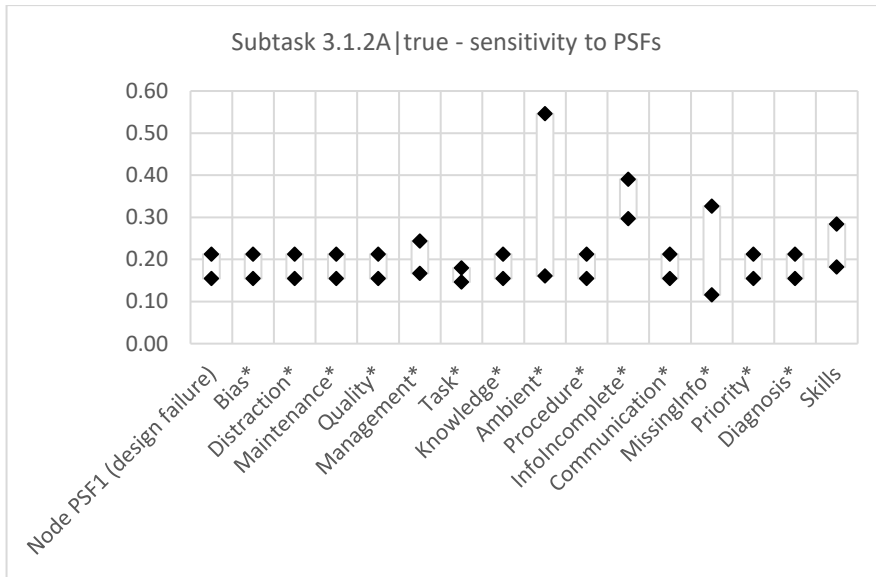


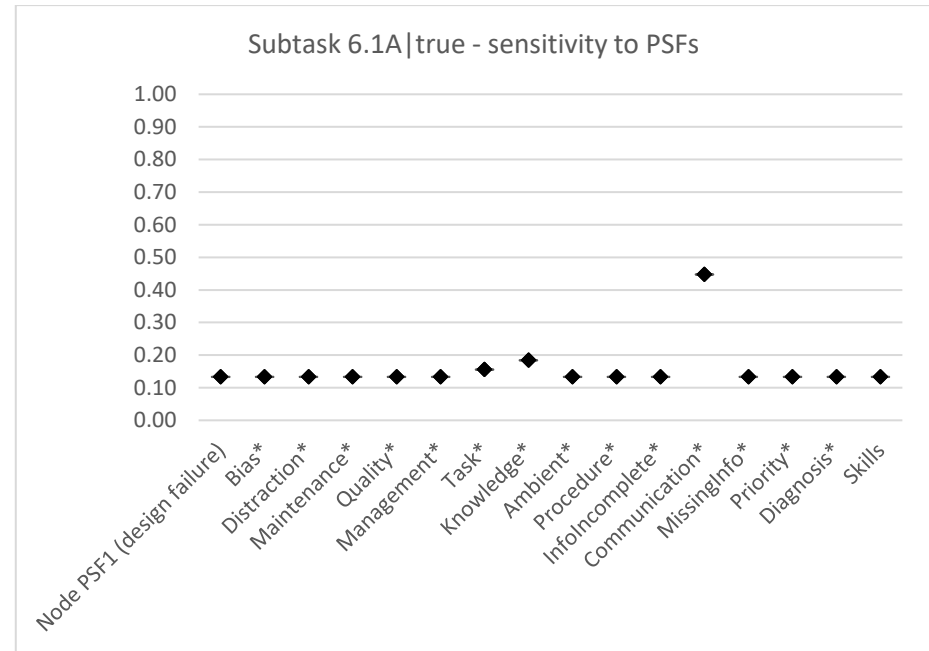
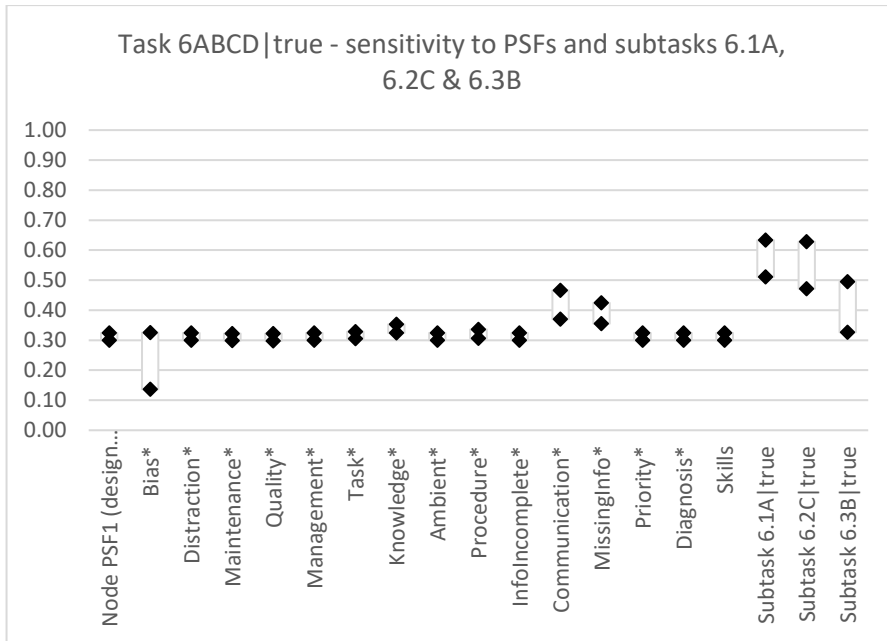


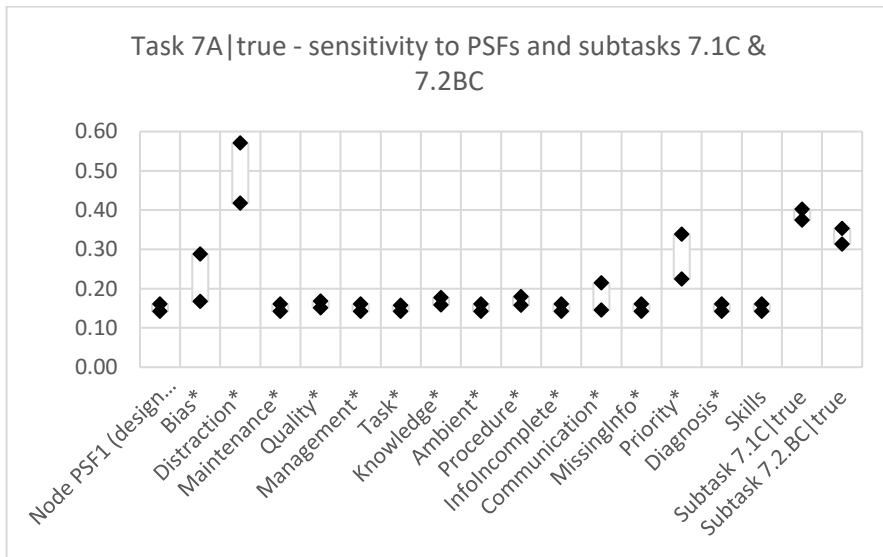
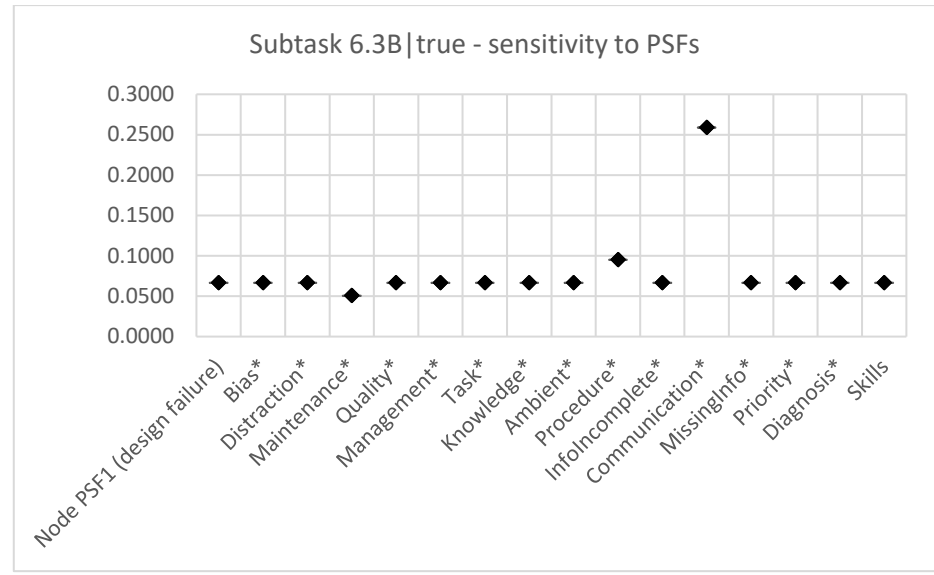
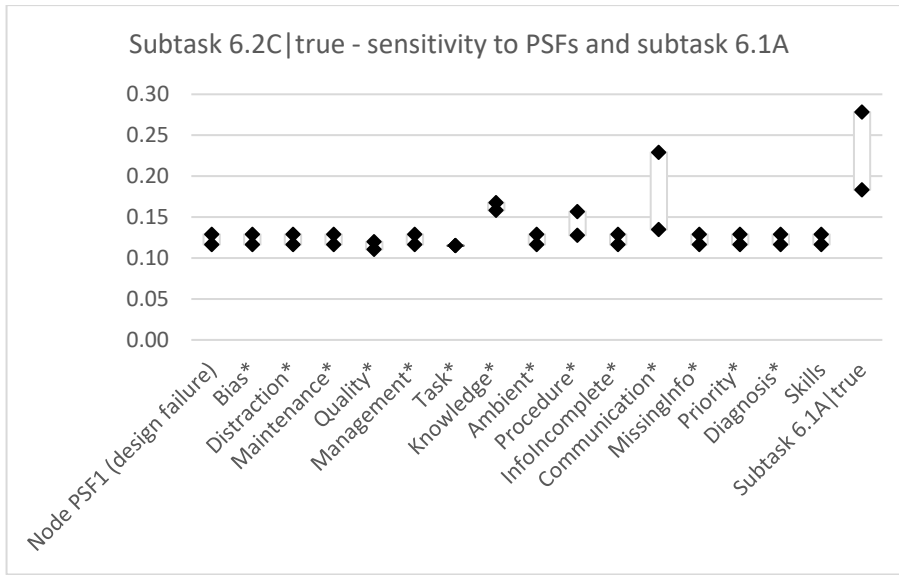


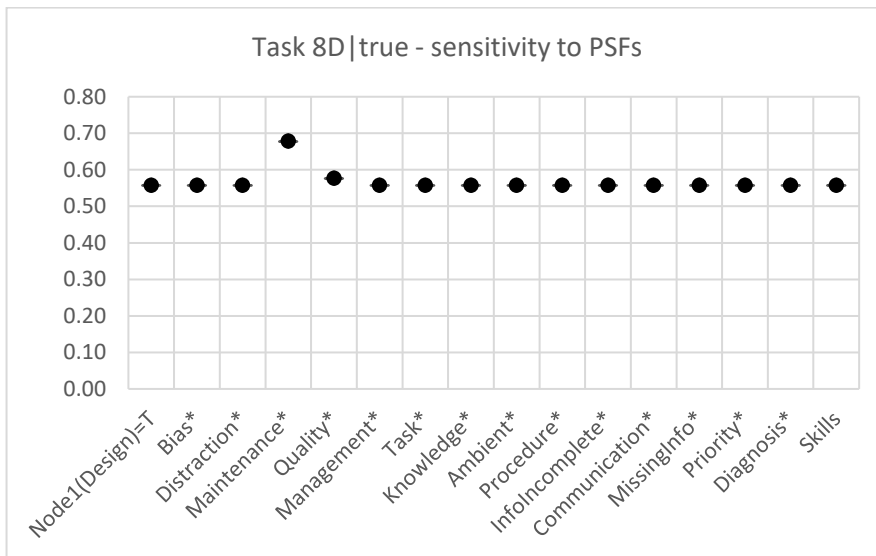
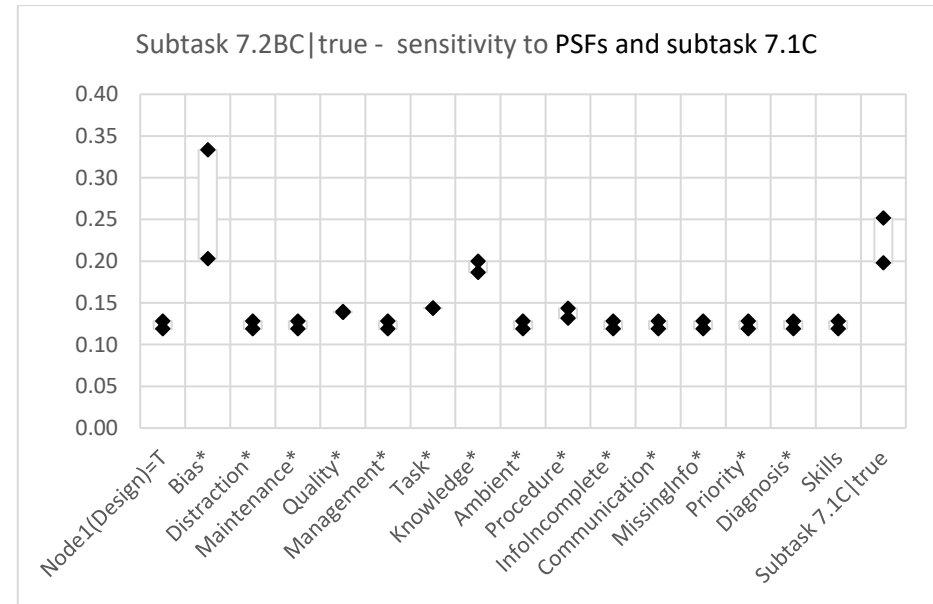
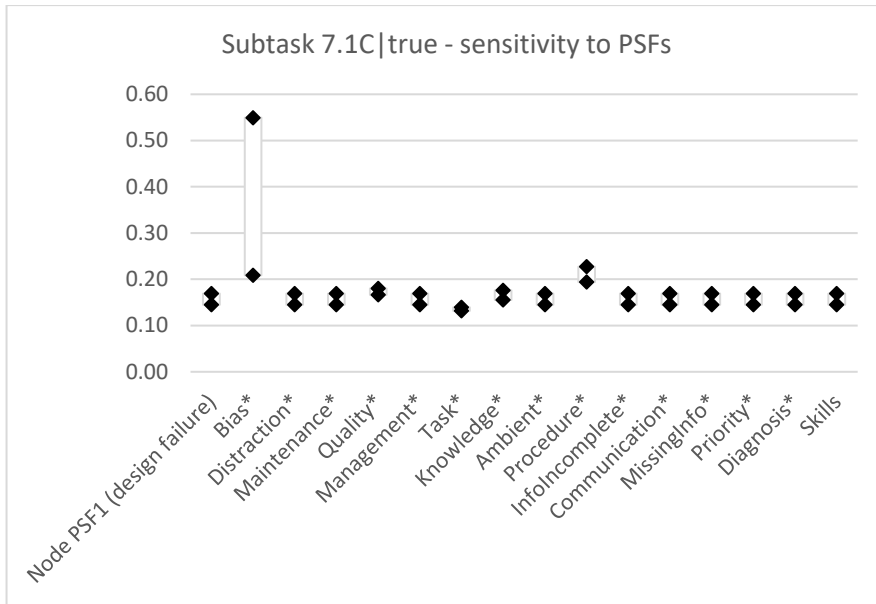
Task 3A | true sensitivity to PSFs and subtasks 3.1.2A and 3.3A

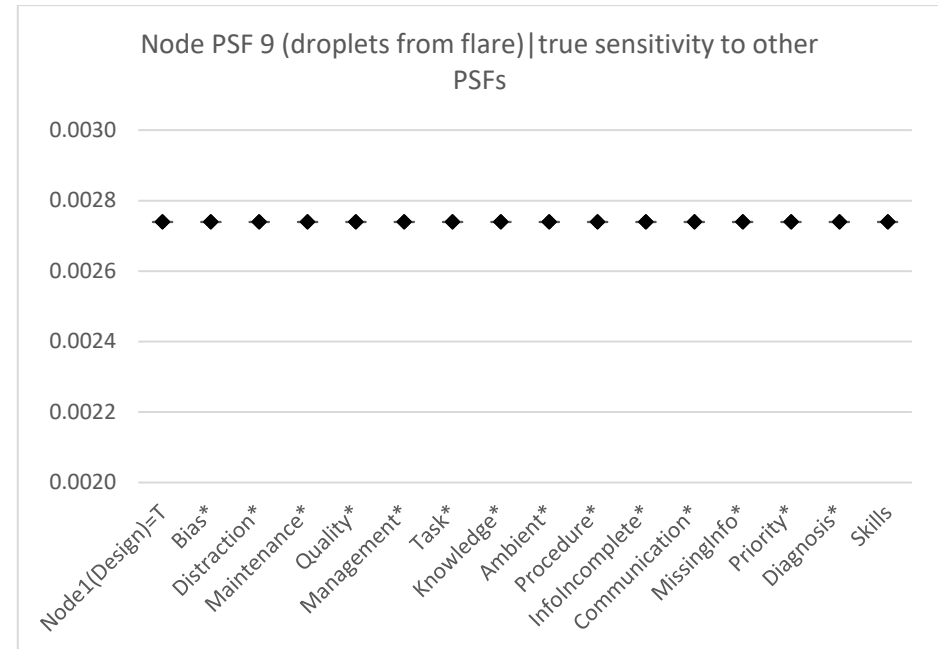
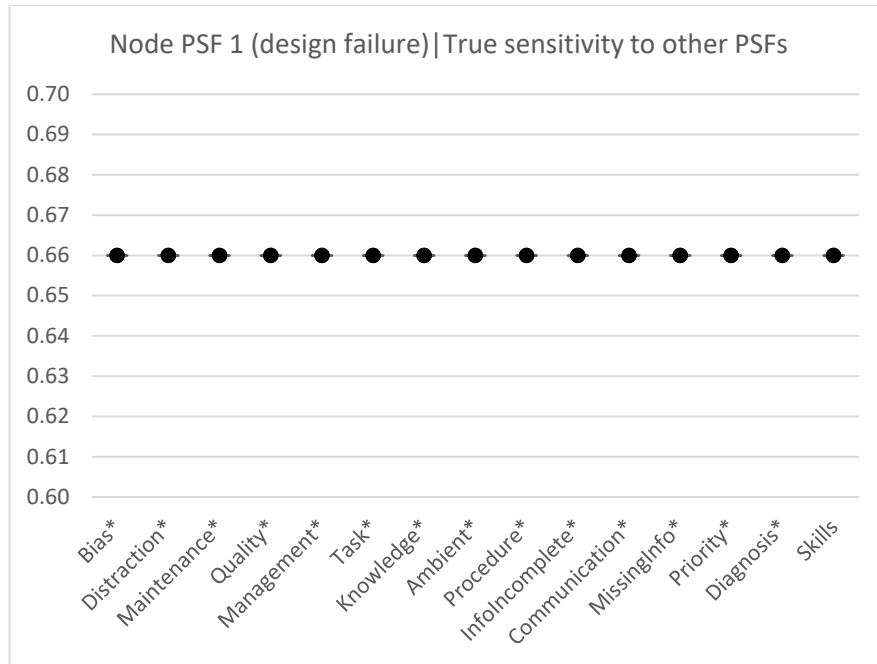












ABDI, H. 2007. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA, 508-510.

ALAN KEITH, P., AUBREY MAURICE, T., HANS STEFAN LEDIN, H., SAFETY, E., OFFSHORE, D., REDGRAVE, C., BOOTLE, MERSEYSIDE, L. H. S. H. S. L., H, H. & D, S. J. Ignition Hazards and Area Classification of Hydrocarbon Cold Vents by the Offshore Oil and Gas Industry 2012 2012.

ANP 2015. Investigation report of the explosion incident of the explosion incident occurred on 11/02/2015 in the FPSO Cidade de São Mateus Brazil: Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

ANP, A. N. D. P., GÁS NATURAL E BIOCOMBUSTÍVEIS. 2020a. *RE: Incident Data from Oil and Gas Exploration and Production*

ANP, A. N. D. P., GÁS NATURAL E BIOCOMBUSTÍVEIS. 2020b. *RE: Monthly bulletin with data on oil and gas production in Brazil, information on producing states, basins, fields and wells produced.*

ANTONUCCI, A., BRÜHLMANN, R., PIATTI, A. & ZAFFALON, M. 2009. Credal networks for military identification problems. *International Journal of Approximate Reasoning*, 50, 666-679.

- ANTONUCCI, A., DE CAMPOS, C. P., HUBER, D. & ZAFFALON, M. Approximating credal network inferences by linear programming. *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 2013 2013. Springer, 13-24.
- ANTONUCCI, A., HUBER, D., ZAFFALON, M., LUGINBÜHL, P., CHAPMAN, I. & LADOUCEUR, R. CREDO: a military decision-support system based on credal networks. *Proceedings of the 16th International Conference on Information Fusion*, 2013. IEEE, 1942-1949.
- API, A. 2010. API Recommended Practice 754. *Process Safety Performance Indicators for the Refining and Petrochemical Industries, 1st Ed.*, American Petroleum Institute, Washington, DC.
- ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCÍA, S., GIL-LÓPEZ, S., MOLINA, D. & BENJAMINS, R. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
- BARLOW, R. E. & WU, A. S. 1978. Coherent systems with multi-state components. *Mathematics of operations research*, 3, 275-281.
- BAYBUTT, P. 2016. Designing risk matrices to avoid risk ranking reversal errors. *Process Safety Progress*, 35, 41-46.
- BBC. 2019. Boeing: Which airlines use the 737 Max 8? *BBC*, p. Newspaper Article.
- BELL, J. & HOLROYD, J. 2009. Review of human reliability assessment methods. *Health & Safety Laboratory*, 78.
- BENCOMO, N. G. P. F. C. H. D. & BLAIR, G. *RE: GeNIe Modeler*.
- BENNETT, P. N., CHICKERING, D. M., MEEK, C. & ZHU, X. Algorithms for active classifier selection: Maximizing recall with precision constraints. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017. 711-719.
- BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993-1022.
- BOBBIO, A., PORTINALE, L., MINICHINO, M. & CIANCAMERLA, E. 2001. Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability Engineering & System Safety*, 71, 249-260.
- BORING, R. L. & BYE, A. Bridging human factors and human reliability analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2008. SAGE Publications Sage CA: Los Angeles, CA, 733-737.
- BROUGHTON, E. 2005. The Bhopal disaster and its aftermath: a review. *Environmental Health*, 4, 1-6.
- BROWNLEE, J. 2017. *A Gentle Introduction to the Bag-of-Words Model* [Online]. <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>. Available: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/> [Accessed 2019 2019].
- BROWNLEE, J. 2018. *The Model Performance Mismatch Problem (and what to do about it)* [Online]. <https://machinelearningmastery.com/the-model-performance-mismatch-problem/>. Available: <https://machinelearningmastery.com/the-model-performance-mismatch-problem/> [Accessed 27 June 2021 2021].
- BROWNLEE, J. 2021. *Cost-Sensitive Learning for Imbalanced Classification* [Online]. <https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/>. Available: <https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/> [Accessed 2021 2021].
- BUCKLAND, M. & GEY, F. 1994. The relationship between recall and precision. *Journal of the American society for information science*, 45, 12-19.
- BURGHERR, P. & HIRSCHBERG, S. 2008. A comparative analysis of accident risks in fossil, hydro, and nuclear energy chains. *Human and Ecological Risk Assessment*, 14, 947-973.

- BYE, A. 2018. Informing HRA by Empirical Data, Halden Reactor Project Lessons Learned and Future Direction. *Proceedings of PSAM 14-Probabilistic Safety Assessment and Management*, 16-21.
- CA AUTHORITY, C. A. 2016. CAP 737. Flight-crew human factors handbook. *London: Civil Aviation Authority*.
- CAIN, J. 2001. *Planning improvements in natural resource management. guidelines for using bayesian networks to support the planning and management of development programmes in the water sector and beyond*, Centre for Ecology and Hydrology.
- CANO, A., GÓMEZ, M., MORAL, S. & ABELLÁN, J. 2007. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *International Journal of Approximate Reasoning*, 44, 261-280.
- CCPS, C. F. C. P. S. 2010. *Guidelines for Risk Based Process Safety*, John Wiley & Sons.
- CGE, R. M. S. 2017. *The history of bowtie* [Online]. Available: https://www.cgerisk.com/knowledgebase/The_history_of_bowtie#Introduction [Accessed Web Page 2020].
- CHANG, Y. J., BLEY, D., CRISCIONE, L., KIRWAN, B., MOSLEH, A., MADARY, T., NOWELL, R., RICHARDS, R., ROTH, E. M. & SIEBEN, S. 2014. The SACADA database for human reliability and human performance. *Reliability Engineering & System Safety*, 125, 117-133.
- CHEN, S. H. & POLLINO, C. A. 2012. Good practice in Bayesian network modelling. *Environmental Modelling & Software*, 37, 134-145.
- CHRONOPOULOS, C. & GUZMAN, N. H. C. Is Smartness Risky? A Framework to Evaluate Smartness in Cyber-Physical Systems. 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference, 2020 2020.
- COOPER, S., RAMEY-SMITH, A., WREATHALL, J. & PARRY, G. 1996. A technique for human error analysis (ATHEANA). Nuclear Regulatory Commission.
- COX, D. R. 1958. Some problems connected with statistical inference. *Ann. Math. Statist*, 29, 357-372.
- COZMAN, F. G. 2000. Credal networks. *Artificial Intelligence*, 120, 199-233.
- CSB, U. S. C. S. A. H. I. B. 2011. Investigation Report of the Bayer CropScience Pesticide Waste Tank Explosion.
- CSB, U. S. C. S. A. H. I. B. 2014. Explosion and fire at the Macondo Well.
- CULLEN, L. W. D. 1993. The public inquiry into the Piper Alpha disaster. 49.
- DE VOS, D., DUDDY, M. & BRONNEBURG, J. The problem of inert-gas venting on FPSOs and a straightforward solution. Offshore Technology Conference, 2006 2006. Offshore Technology Conference.
- DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41, 391-407.
- DI FLUMERI, G., DE CRESCENZIO, F., BERBERIAN, B., OHNEISER, O., KRAMER, J., ARICÒ, P., BORGHINI, G., BABILONI, F., BAGASSI, S. & PIASTRA, S. 2019. Brain-Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing With Highly Automated Systems. *Frontiers in human neuroscience*, 13.
- DRUPSTEEN, L., GROENEWEG, J. & ZWETSLOOT, G. I. J. M. 2013. Critical steps in learning from incidents: using learning potential in the process from reporting an incident to accident prevention. *International Journal of Occupational Safety and Ergonomics*, 19, 63-77.
- EASA, E. *International Maintenance Review Board Policy Board* [Online]. <https://www.easa.europa.eu/domains/aircraft-products/international-maintenance-review-board-policy-board-IMRBPB#group-easa-downloads>. Available: <https://www.easa.europa.eu/domains/aircraft-products/international-maintenance-review-board-policy-board-IMRBPB#group-easa-downloads> [Accessed 20th December 2020].

- EI, E. I. & IOGP 2020. Report 454: Human factors engineering in projects. Energy Institute.
- ESTRADA-LUGO, H. D., DE ANGELIS, M. & PATELLI, E. 2019a. Probabilistic risk assessment of fire occurrence in residential buildings: Application to the Grenfell Tower.
- ESTRADA-LUGO, H. D., SANTHOSH, T. V., DE ANGELIS, M. & PATELLI, E. 2020. Resilience assessment of safety-critical systems with credal networks.
- ESTRADA-LUGO, H. D., TOLO, S., DE ANGELIS, M. & PATELLI, E. 2019b. Pseudo credal networks for inference with probability intervals. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 5.
- EVANS, J. S. B. T., HANDLEY, S. J. & OVER, D. E. 2003. Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 321.
- FENTON, N. & NEIL, M. 2012. *Risk assessment and decision analysis with Bayesian networks*, Crc Press.
- FERSON, S., KREINOVICH, V., GINZBURG, L. & SENTZ, F. 2003. Constructing Probability Boxes and Dempster-Shafer Structures. Sandia National Labs., Albuquerque, NM (US); Sandia National Labs
- FERSON, S., O'RAWE, J. & BALCH, M. 2014. Computing with confidence: imprecise posteriors and predictive distributions. *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*.
- FRENCH, S., BEDFORD, T., POLLARD, S. J. T. & SOANE, E. 2011. Human reliability analysis: A critique and review for managers. *Safety Science*, 49, 753-763.
- GERTMAN, D., BLACKMAN, H., MARBLE, J., BYERS, J. & SMITH, C. 2005. The SPAR-H human reliability analysis method. *US Nuclear Regulatory Commission*, 230, 35.
- GIBSON, W. H. & MEGAW, T. D. 1999. *The implementation of CORE-DATA, a computerised human error probability database*, HSE Books Norwich, UK.
- GOH, Y. M. & UBEYNARAYANA, C. U. 2017. Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, 108, 122-130.
- GOLDBERG, Y. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10, 1-309.
- GONÇALVES, F. C. C. & TRABASSO, L. G. 2018. Aircraft Preventive Maintenance Data Evaluation Applied in Integrated Product Development Process. *Journal of Aerospace Technology and Management*, 10.
- GOOGLE. 2018. *Machine Learning Crash Course* [Online]. <https://developers.googleblog.com/2018/03/machine-learning-crash-course.html>. Available: <https://developers.google.com/machine-learning/crash-course/classification/> [Accessed Web Page].
- GRECH, M. R., HORBERRY, T. & SMITH, A. Human error in maritime operations: Analyses of accident reports using the Leximancer tool. Proceedings of the human factors and ergonomics society annual meeting, 2002 2002. Sage Publications Sage CA: Los Angeles, CA, 1718-1721.
- GRECO, S. F., PODOFILLINI, L. & DANG, V. N. 2021. A Bayesian model to treat within-category and crew-to-crew variability in simulator data for Human Reliability Analysis. *Reliability Engineering & System Safety*, 206, 107309.
- GRIFFITH, C. D. & MAHADEVAN, S. 2015. Human reliability under sleep deprivation: Derivation of performance shaping factor multipliers from empirical data. *Reliability Engineering & System Safety*, 144, 23-34.
- GROTH, K. M. & MOSLEH, A. 2012. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226, 361-379.

- GROTH, K. M., SMITH, C. L. & SWILER, L. P. 2014. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliability Engineering & System Safety*, 128, 32-40.
- GROTH, K. M., SMITH, R. & MORADI, R. 2019. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. *Reliability Engineering & System Safety*, 191, 106507.
- HALLBERT, B., BORING, R., GERTMAN, D., DUDENHOEFFER, D., WHALEY, A., MARBLE, J., JOE, J. & LOIS, E. 2006. Human event repository and analysis (HERA) system, overview. *US Nuclear Regulatory Commission, Washington DC, Tech.Rep.NUREG/CR-6903*.
- HE, H. & MA, Y. 2013. Imbalanced learning: foundations, algorithms, and applications.
- HEIDARYSAFA, M., KOWSARI, K., BARNES, L. & BROWN, D. Analysis of Railway Accidents' Narratives Using Deep Learning. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018 2018. IEEE, 1446-1453.
- HENDERSON, J. & EMBREY, D. 2012. Guidance on quantified human reliability analysis. *Energy Institute, London*.
- HENRION, M. Some Practical Issues in Constructing Belief Networks. UAI, 1987. 161-173.
- HOLLNAGEL, E. 1998. *Cognitive reliability and error analysis method (CREAM)*, Elsevier.
- HSE, H. A. S. E. U. 2020a. *Optimising Offshore Working Patterns – Shared Research Project* [Online]. Available: <https://www.hse.gov.uk/aboutus/assets/docs/shared-research-offshore-working-patterns.pdf> [Accessed].
- HSE, U. 2010. Assessment of the adequacy of venting arrangements for cargo oil tanks on FPSO and FSU installations.
- HSE, U. 2020b. HSE Offshore Statistics, Offshore Hydrocarbon Releases 1992 – 2016
.
- HUGHES, P., FIGUERES-ESTEBAN, M. & VAN GULIJK, C. From negative statements to positive safety. 26th European Safety and Reliability Conference, 2017 2017. CRC Press/Balkema, 307.
- ILIEV, R., DEHGhani, M. & SAGI, E. 2015. Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, 7, 265-290.
- JENTSCH, F. G. 1993. Problems of systematic safety assessments: lessons learned from aircraft accidents. *Verification and Validation of Complex Systems: Human Factors Issues*. Springer.
- JUNG, W., PARK, J., KIM, Y., CHOI, S. Y. & KIM, S. 2020. HuREX—A framework of HRA data collection from simulators in nuclear power plants. *Reliability Engineering & System Safety*, 194, 106235.
- KIM, S. H., LEE, N. & KING, P. E. 2020. Dimensions of religion and spirituality: A longitudinal topic modeling approach. *Journal for the Scientific Study of Religion*, 59, 62-83.
- KIM, Y. 2020. Considerations for generating meaningful HRA data: Lessons learned from HuREX data collection. *Nuclear Engineering and Technology*.
- KIM, Y., PARK, J. & JUNG, W. 2017. A classification scheme of erroneous behaviors for human error probability estimations based on simulator data. *Reliability Engineering & System Safety*, 163, 1-13.
- KIM, Y., PARK, J., JUNG, W., CHOI, S. Y. & KIM, S. 2018. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. *Reliability Engineering & System Safety*, 173, 12-22.
- KIRWAN, B. 1994. *A guide to practical human reliability assessment*, CRC press.
- KIRWAN, B. 1997a. Validation of human reliability assessment techniques: part 1—validation issues. *Safety Science*, 27, 25-41.

- KIRWAN, B. 1997b. Validation of human reliability assessment techniques: Part 2—Validation results. *Safety Science*, 27, 43-75.
- KIRWAN, B. & AINSWORTH, L. K. 1992. *A guide to task analysis: the task analysis working group*, CRC press.
- KIRWAN, B., KENNEDY, R., TAYLOR-ADAMS, S. & LAMBERT, B. 1997. The validation of three Human Reliability Quantification techniques—THERP, HEART and JHEDI: Part II—Results of validation exercise. *Applied ergonomics*, 28, 17-25.
- KLETZ, T. Some Common Errors in Accident Investigations. *Safety and Reliability*, 2011 2011. Taylor & Francis, 4-13.
- KLETZ, T. A. 2001. *Learning from accidents*, Routledge.
- KNKT 2019. Aircraft Accident Investigation Final Report Boeing 737-8 (MAX) Lion Mentari Airlines KNKT.18.10.35.04. internet: KNKT.
- KUTER, U., NAU, D., GOSSINK, D. & LEMMER, J. F. Interactive course-of-action planning using causal models. Third International Conference on Knowledge Systems for Coalition Operations (KSCO-2004), 2004 2004. 37-52.
- KYRIAKIDIS, M., MAJUMDAR, A. & OCHIENG, W. Y. 2015. Data based framework to identify the most significant performance shaping factors in railway operations. *Safety Science*, 78, 60-76.
- LAUMANN, K., BLACKMAN, H. & RASMUSSEN, M. 2018. Challenges with data for human reliability analysis. *Proceedings of ESREL 2018*, 315-321.
- LÉGER, A., WEBER, P., LEVRAT, E., DUVAL, C., FARRET, R. & IUNG, B. 2009. Methodological developments for probabilistic risk analyses of socio-technical systems. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 223, 313-332.
- LEMMER, J. F. & GOSSINK, D. E. 2004. Recursive noisy OR-a rule for estimating complex probabilistic interactions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34, 2252-2261.
- LEVESON, N. 2020. Safety III: A Systems Approach to Safety and Resilience.
- LIMA, E. N., BENITES, R. D., MOSLEH, A. & MARTINS, M. R. 2019. A Methodology to Use Multi-Objective Optimization Criteria for an Offshore Topside Production System Since the Early Design Stages, and for The Unit Life Cycle.
- LIN, S.-W. & BIER, V. M. 2008. A study of expert overconfidence. *Reliability Engineering & System Safety*, 93, 711-721.
- LIU, P. & LIU, J. 2020. Combined Effect of Multiple Performance Shaping Factors on Human Reliability: Multiplicative or Additive? *International Journal of Human-Computer Interaction*, 36, 828-838.
- LOIS, E. 2009. *International HRA Empirical Study--phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Similar Performance Data*, Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.
- MALATO, G. 2015. *Why training set should always be smaller than test set* [Online]. <https://towardsdatascience.com/why-training-set-should-always-be-smaller-than-test-set-61f087ed203c>
- Towards Data Science. Available: <https://towardsdatascience.com/why-training-set-should-always-be-smaller-than-test-set-61f087ed203c> [Accessed Magazine Article].
- MARKS, S. & DAHIR, A. L. 2020. Ethiopian Report on 737 Max Crash Blames Boeing. p.Newspaper Article.
- MARTINS, M. R. & MATURANA, M. C. 2013. Application of Bayesian Belief networks to the human reliability analysis of an oil tanker operation focusing on collision accidents. *Reliability Engineering & System Safety*, 110, 89-109.

- MATLAB. 2019. *Support Vector Machines for Binary Classification* [Online]. <https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html>. Available: <https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html> [Accessed Web Page 2019].
- MATLAB & MATHWORKS. 2018. *Matlab documentation for confusion chart function* [Online]. <https://uk.mathworks.com/help/stats/confusionchart.html>. Available: <https://uk.mathworks.com/help/stats/confusionchart.html> [Accessed Web Page 2020].
- MCCALLUM, A. & NIGAM, K. A comparison of event models for naive bayes text classification. AAI-98 workshop on learning for text categorization, 1998 1998. Citeseer, 41-48.
- MKRTCHYAN, L., PODOFILLINI, L. & DANG, V. N. 2015. Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliability Engineering & System Safety*, 139, 1-16.
- MKRTCHYAN, L., PODOFILLINI, L. & DANG, V. N. 2016. Methods for building conditional probability tables of bayesian belief networks from limited judgment: an evaluation for human reliability application. *Reliability Engineering & System Safety*, 151, 93-112.
- MORAIS, C., ESTRADA-LUGO, D., JACQUES, T., TOLO, S., MOURA, R., BEER, M. & PATELLI, E. 2021 (in press). Robust data-driven human reliability analysis using Credal Networks (in press). *Reliability Engineering & System Safety Journal*.
- MORAIS, C., GARCIA, A., SILVA, B., FERREIRA, N. & PIRES, T. Explaining the explosion onboard FPSO Cidade de São Mateus from Regulatory Point of View. ESREL - Risk, Reliability and Safety: Innovating Theory and Practice 25-29 September 2016 2016 Glasgow, Scotland.
- MORAIS, C., MOURA, R., BEER, M. & PATELLI, E. 2018. Attempt to predict human error probability in different industry sectors using data from major accidents and Bayesian networks. *14th Probabilistic Safety Assessment and Management, PSAM 2018*.
- MORAIS, C., MOURA, R., BEER, M. & PATELLI, E. 2020. Analysis and estimation of human errors from major accident investigation reports. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 6.
- MORAIS, C., TOLO, S., MOURA, R., BEER, M. & PATELLI, E. Tackling the lack of data for human error probability with Credal network. Proceedings of the ESREL, 2019 2019a.
- MORAIS, C., YUNG, K. & PATELLI, E. Machine-learning tool for human factors evaluation-application to lion air Boeing 737-8 max accident. UNCECOMP 2019 and 3rd ECCOMAS Thematic Conference, 2019 2019b. National Technical University of Athens.
- MOSLEH, A., BIER, V. M. & APOSTOLAKIS, G. 1988. A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety*, 20, 63-85.
- MOURA, R., BEER, M., PATELLI, E. & LEWIS, J. 2017a. Learning from major accidents: Graphical representation and analysis of multi-attribute events to enhance risk communication. *Safety Science*, 99, 58-70.
- MOURA, R., BEER, M., PATELLI, E., LEWIS, J. & KNOLL, F. 2016. Learning from major accidents to improve system design. *Safety Science*, 84, 37-45.
- MOURA, R., BEER, M., PATELLI, E., LEWIS, J. & KNOLL, F. 2017b. Learning from accidents: interactions between human factors, technology and organisations as a central element to validate risk studies. *Safety Science*, 99, 196-214.
- MOURA, R., M., B., E., P., J., L. & KNOLL, F. 2020. Multi-Attribute Technological Accidents Dataset (MATA-D).

- MOURA, R., PATELLI, E., LEWIS, J., MORAIS, C. & BEER, M. 2017c. Human factors influencing decision-making: tendencies from first-line management decisions and implications to reduce major accidents. *Safety and Reliability—Theory and Applications*.
- MURPHY, K. 2007. Software for graphical models: A review. *International Society for Bayesian Analysis Bulletin*, 14, 13-15.
- MYUNG, I. J. 2003. Tutorial on maximum likelihood estimation. *Journal of mathematical psychology*, 47, 90-100.
- NEYMAN, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333-380.
- NIELSEN, T. D. & JENSEN, F. V. 2009. *Bayesian networks and decision graphs*, Springer Science & Business Media.
- NRC, U. N. R. C. 2014. The international HRA empirical study: lessons learned from comparing HRA methods predictions to HAMMLAB simulator data, NUREG-2127. *US Nuclear Regulatory Commission, Washington, DC*.
- OREDA. *Offshore and Onshore Reliability Data* [Online]. <https://www.oreda.com/>. Available: <https://www.oreda.com/> [Accessed 20th December 2020 2020].
- PARK, J. & JUNG, W. 2007. OPERA—a human performance database under simulated emergencies of nuclear power plants. *Reliability Engineering & System Safety*, 92, 503-519.
- PARK, J., KIM, Y. & JUNG, W. Use of a Big Data Mining Technique to Extract Relative Importance of Performance Shaping Factors from Event Investigation Reports. *International Conference on Applied Human Factors and Ergonomics, 2017* 2017. Springer, 230-238.
- PATELLI, E., ALVAREZ, D. A., BROGGI, M. & DE ANGELIS, M. An integrated and efficient numerical framework for uncertainty quantification: application to the nasa langley multidisciplinary uncertainty quantification challenge. *16th AIAA Non-Deterministic Approaches Conference, 2014* 2014. 1501.
- PATELLI, E., GHANEM, R., HIGDON, D. & OWHADI, H. 2016. COSSAN: a multidisciplinary software suite for uncertainty quantification and risk management. *Handbook of uncertainty quantification*, 1-69.
- PATELLI, E., TOLO, S., GEORGE-WILLIAMS, H., SADEGHI, J., ROCCHETTA, R., DE ANGELIS, M. & BROGGI, M. 2018. OpenCossan 2.0: an efficient computational toolbox for risk, reliability and resilience analysis.
- PING SHUN, K. 2018. *Accuracy, Precision, Recall or F1?* [Online]. <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>: Towards Data Science. Available: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9> [Accessed Magazine Article].
- PIRIE, W. 2004. Spearman rank correlation coefficient. *Encyclopedia of statistical sciences*, 12.
- PODOFILLINI, L. & DANG, V. N. 2013. A Bayesian approach to treat expert-elicited probabilities in human reliability analysis model construction. *Reliability Engineering & System Safety*, 117, 52-64.
- PODOFILLINI, L., MKRTCHYAN, L. & DANG, V. N. 2014. Aggregating expert-elicited error probabilities to build HRA models. *Safety and Reliability: Methodology and Applications*. CRC Press.
- PREISCHL, W. & HELLMICH, M. 2013. Human error probabilities from operational experience of German nuclear power plants. *Reliability Engineering & System Safety*, 109, 150-159.
- PREISCHL, W. & HELLMICH, M. 2016. Human error probabilities from operational experience of German nuclear power plants, Part II. *Reliability Engineering & System Safety*, 148, 44-56.

- PURSEL, M., GANT, S., NEWTON, A., BENNETT, D., O'SULLIVAN, L. & HOOK, P. Investigation of Cargo Tank Vent Fires on the GP3 FPSO, Part 1: Identification of Ignition Mechanisms and Analysis of Material Ejected from the Flare. *Hazards* 26, 2016 2016a.
- PURSEL, M., GANT, S., NEWTON, A., BENNETT, D., O'SULLIVAN, L. & HOOK, P. Investigation of Cargo Tank Vent Fires on the GP3 FPSO, Part 2: Analysis of Vapour Dispersion. *Hazards* 26, 2016 2016b.
- RAMOS, M., UTNE, I. B., VINNEM, J. E. & MOSLEH, A. 2018. Accounting for human failure in autonomous ships operations. *Safety and Reliability-Safe Societies in a Changing World ESREL 2018*, 355-63.
- RAMOS, M. A., DROGUETT, E. L., MOSLEH, A. & MOURA, M. D. C. 2020. A human reliability analysis methodology for oil refineries and petrochemical plants operation: Phoenix-PRO qualitative framework. *Reliability Engineering & System Safety*, 193, 106672.
- RANGEL, E. & SANGUEDO, C. A. 2018. Considerations on the New Requirements for Electrical Installations in Hazardous Locations. IEEE.
- RATSABY, J. & VENKATESH, S. S. Learning from a mixture of labeled and unlabeled examples with parametric side information. Proceedings of the eighth annual conference on Computational learning theory, 1995. 412-417.
- REASON, J. 2016. *Managing the risks of organizational accidents*, Routledge.
- RIBEIRO, L. C. F., AFONSO, L. C. S., COLOMBO, D., GUILHERME, I. R. & PAPA, J. P. 2020. Evolving Neural Conditional Random Fields for drilling report classification. *Journal of Petroleum Science and Engineering*, 187, 106846.
- RITCHIE, H. 2020. *What are the safest and cleanest sources of energy?* [Online]. <https://ourworldindata.org/safest-sources-of-energy#licence>: University of Oxford. Available: <https://ourworldindata.org/safest-sources-of-energy#licence> [Accessed 25/09/2021 2021].
- ROBERTS, M. E., STEWART, B. M. & AIROLDI, E. M. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111, 988-1003.
- ROBINSON, S. D., IRWIN, W. J., KELLY, T. K. & WU, X. O. 2015. Application of machine learning to mapping primary causal factors in self reported safety narratives. *Safety Science*, 75, 118-129.
- SALVI, O. & DEBRAY, B. 2006. A global view on ARAMIS, a risk assessment methodology for industries in the framework of the SEVESO II directive. Elsevier.
- SAMANIEGO, F. J. 1985. On closure of the IFR class under formation of coherent systems. *IEEE Transactions on Reliability*, 34, 69-72.
- SARKAR, S. & MAITI, J. 2020. Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis. *Safety Science*, 131, 104900.
- SHI, H. & LIU, Y. Naïve Bayes vs. support vector machine: resilience to missing data. International Conference on Artificial Intelligence and Computational Intelligence, 2011 2011. Springer, 680-687.
- SHIMAMURA, Y. 2002. FPSO/FSO: State of the art. Springer.
- SHIRAZI, C. H. 2009. *Data-informed calibration and aggregation of expert judgment in a Bayesian framework*.
- SIEGRIST, J. 2011. Mixing good data with bad: how to do it and when you should not. *Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management*.
- SKLET, S. 2006. Safety barriers: Definition, classification, and performance. *Journal of Loss Prevention in the Process Industries*, 19, 494-506.
- SMITH, E., ANNE KOOP, D. N. V. & KING, U. K. S. Guidance on Human Factors Critical Task Analysis. In: ICHEME, ed. *Hazards XXII Process Safety and Environmental Protection*, 2011.

- SOVACOO, B. K., KRYMAN, M. & LAINE, E. 2015. Profiling technological failure and disaster in the energy sector: A comparative analysis of historical energy accidents. *Energy*, 90, 2016-2027.
- STEMPFEL, Y. & DANG, V. N. 2012. Developing and evaluating the Bayesian Belief Network as a human reliability model using artificial data. *Advances in Safety, Reliability and Risk Manag.*
- STRÄTER, O. 2000. Evaluation of human reliability on the basis of operational experience. *Gesellschaft für Anlagen und Reaktorsicherheit (GRS) mbH.*
- SUN, Y., WONG, A. K. & KAMEL, M. S. 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23, 687-719.
- SUNDARAMURTHI, R. & SMIDTS, C. 2013. Human reliability modeling for the next generation system code. *Annals of Nuclear Energy*, 52, 137-156.
- SWAIN, A. D. & GUTTMANN, H. E. 1983. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. Sandia National Labs.
- TARGOUTZIDIS, A. 2010. Incorporating human factors into a simplified “bow-tie” approach for workplace risk assessment. *Safety Science*, 48, 145-156.
- TOLO, S., PATELLI, E. & BEER, M. Enhanced Bayesian network approach to sea wave overtopping hazard quantification. Proceedings of the 25th European safety and reliability conference, ESREL, Zurich, Switzerland, Sept, 2015. 7-10.
- TOLO, S., PATELLI, E. & BEER, M. 2017. Risk assessment of spent nuclear fuel facilities considering climate change. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3, G4016003.
- TOLO, S., PATELLI, E. & BEER, M. 2018. An open toolbox for the reduction, inference computation and sensitivity analysis of Credal Networks. *Advances in Engineering Software*, 115, 126-148.
- TRBOJEVIC, V. M. 2008. Optimising Hazard Management by Workforce Engagement and Supervision. RR637. *Optimising Hazard Management by Workforce Engagement and Supervision. Prepared by Risk Support Limited for the Health and Safety Executive, RR637.*
- TRIST, E. L. & BAMFORTH, K. W. 1951. Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human relations*, 4, 3-38.
- TROFFAES, M. C. M. 2007. Decision making under uncertainty using imprecise probabilities. *International journal of approximate reasoning*, 45, 17-29.
- TRUCCO, P., CAGNO, E., RUGGERI, F. & GRANDE, O. 2008. A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliability Engineering & System Safety*, 93, 845-856.
- VINNEM, J. E. 2001. *Operational safety of FPSOs: initial summary report*, Great Britain, Health and Safety Executive.
- WALLEY, P. 1991. Statistical reasoning with imprecise probabilities.
- WANG, S. I. & MANNING, C. D. Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012 2012. 90-94.
- WAYKOLE, R. N. & THAKARE, A. D. 2018. A Review of feature extraction methods for text classification. *IJAERD*, 4, 351-354.
- WICKENS, C. D., HOLLANDS, J. G., BANBURY, S. & PARASURAMAN, R. 2015. *Engineering psychology and human performance*, Psychology Press.
- WILLIAMS, J. A data-based method for assessing and reducing human error to improve operational performance. Conference Record for 1988 IEEE Fourth Conference on Human Factors and Power Plants, 1988. IEEE, 436-450.

- WISSE, B. W., VAN GOSLIGA, S. P., VAN ELST, N. P. & BARROS, A. I. Relieving the elicitation burden of Bayesian Belief Networks. BMA, 2008.
- XIANG, Y. & JIA, N. 2007. Modeling causal reinforcement and undermining for efficient CPT elicitation. *IEEE Transactions on Knowledge and Data Engineering*, 19, 1708-1718.
- XING, J., PARRY, G., PRESLEY, M., FORESTER, J., HENDRICKSON, S. & DANG, V. 2016. An Integrated Human Event Analysis Systems (IDHEAS) for Nuclear Power Plant Internal Events At-Power Application, NUREG-2199, Vol. 1. *Washington, DC: US Nuclear Regulatory Commission*.
- YANG, Z. L., BONSALE, S., WALL, A., WANG, J. & USMAN, M. 2013. A modified CREAM to human reliability quantification in marine engineering. *Ocean Engineering*, 58, 293-303.
- ZHANG, F., FLEYEH, H., WANG, X. & LU, M. 2019. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99, 238-248.
- ZHANG, W., YOSHIDA, T. & TANG, X. 2008. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21, 879-886.
- ZIO, E. 2009. Reliability engineering: Old problems and new challenges. *Reliability Engineering & System Safety*, 94, 125-141.
- ZIO, E. 2018. The future of risk assessment. *Reliability Engineering & System Safety*, 177, 176-190.
- ŽUBRINIĆ, K., MILIČEVIĆ, M. & ZAKARIJA, I. 2013. Comparison of Naive Bayes and SVM classifiers in categorization of concept maps. *International journal of computers*, 7, 109-116.