Robust data-driven human reliability analysis using credal networks

Caroline Morais[a,b], Hector Diego Estrada-Lugo[a], Silvia Tolo[c], Tiago Jacques[b], Raphael Moura[a,b], Michael Beer[a,d,e], Edoardo Patelli[a,f,*]

[a] Institute for Risk and Uncertainty, University of Liverpool, United Kingdom
[b] Agency for Petroleum, Natural Gas and Biofuels (ANP), Brazil
[c] University of Nottingham, United Kingdom
[d] Leibniz Universität Hannover, Germany
[e] Tongji University, China
[f] Centre for Intelligent Infrastructure, Department Civil and Environmental Engineering, Strathclyde University, United Kingdom

ABSTRACT

Despite increasing collection efforts of empirical human reliability data, the available databases are still insufficient for understanding the relationships between human errors and their influencing factors. Currently, probabilistic tools such as Bayesian network are used to model data uncertainty requiring the estimation of conditional probability tables from data that is often not available. The most common solution relies on the adoption of assumptions and expert elicitation to fill the gaps. This gives an unjustified sense of confidence on the analysis.

This paper proposes a novel methodology for dealing with missing data using intervals comprising the lowest and highest possible probability values. Its implementation requires a shift from Bayesian to credal networks. This allows to keep track of the associated uncertainty on the available data. The methodology has been applied to the quantification of the risks associated to a storage tank depressurisation of offshore oil & gas installations known as FPSOs and FSOs. The critical task analysis is converted to a cause-consequence structure and used to build a credal network, which extracts human factors combinations from major accidents database defined with CREAM classification scheme. Prediction analysis shows results with interval probabilities rather than point values measuring the effect of missing-data variables. Novel decision-making strategies for diagnostic analysis are suggested to unveil the most relevant variables for risk reduction in presence of imprecision. Realistic uncertainty depiction implies less conservative human reliability analysis and improve risk communication between assessors and decision-makers.

Keywords: Credal network, missing data, human reliability analysis (HRA), CREAM, FPSO/FSO, quantified bow-tie

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). 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) [1, 2]. Unfortunately, many of those databases had been discredited due to their large variability, especially if compared against the components reliability estimates [1]. 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 [3]. 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. This highlights the urgent need for novel tools and methodology able to tackle such limitations [4].

The starting point of this work is the research question if imprecise probability theory might help to capture and adequately model human reliability's variability, 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), acknowledgeable more variable than *hard barriers* which rely on hardware (i.e. physical or technical components) [5, 6]. 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. 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 [7]. 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 [8]. 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. While 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 [9]. 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 [8].

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 [10]. The approach proposed in this study expands on strategies developed by some of the authors in a former study [11].

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 [1, 2]. 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 [12] and overconfidence [13]. It might also be unfeasible if experts need to elicit a variable under many simultaneous conditions [14]. 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 [15, 16], simulators (e.g., HuREX, SACADA, HAMMLab, and ongoing efforts to develop a data framework to quantify the IDHEAS method) [17-20], derived from near-misses (i.e., incident events that could have resulted in severe consequences [5]) [21, 22], and finally analysis derived from major accidents [23, 24]. 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 [25]. Previous studies have offered suggestions on how to generate meaningful HRA empirical data, regarding preparation, collection, analysis, and application [26].

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 [27], but also can be recorded by specifically designed simulators [28]. In incidents databases, the data might be collected through extensive reading of investigation reports [29] or by using a machine-learning strategy of text recognition and classification [30]. 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 [31].

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 [29] 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 [23]. 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 [9]. 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 [1]. 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) [11]. 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 [27]. 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 [8].

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 [3, 19]. 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) [32], 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 [32]. A risk or hazard identification analysis is an important aid to identify which tasks are critical [2, 32]. 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 [2]. 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 [9], resembles the *bow-tie approach*, a popular qualitative risk analysis in Oil & Gas industry. This can be seen in Figure 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) [33, 34]. Bow-tie diagrams have been already used to model and quantify human factors by using a combination of fault and event trees [34, 35] and Bayesian networks [36].

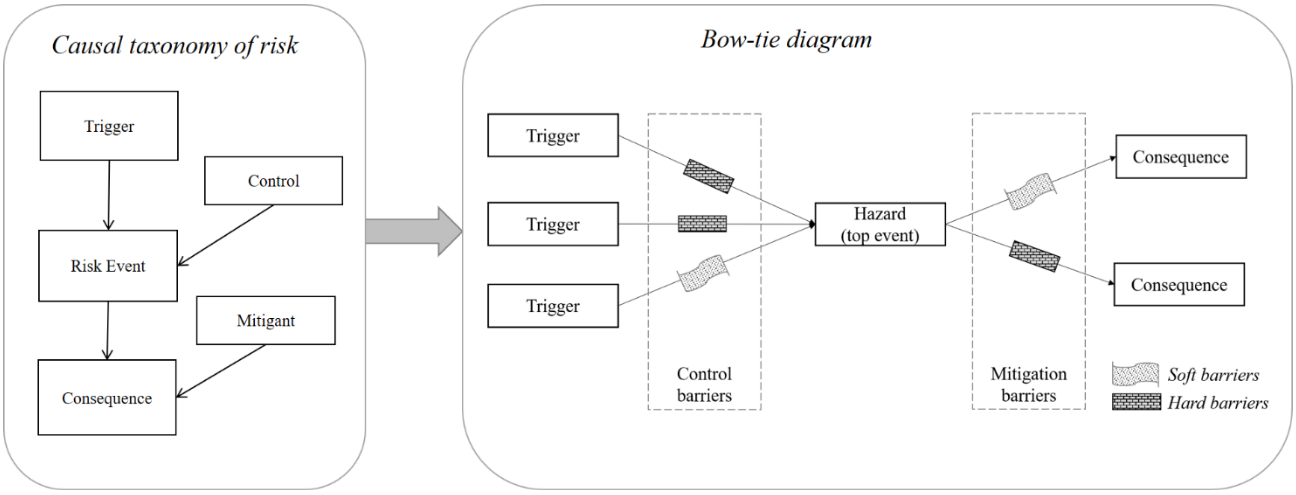


Figure 1. Similarity of the 'causal taxonomy of risk' between a Bayesian network and a 'bow-tie diagram'.

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 [1]. 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 [7]. 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* 2, nodes PSF1, PSF2 and PSF3 represent different performance shape 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.

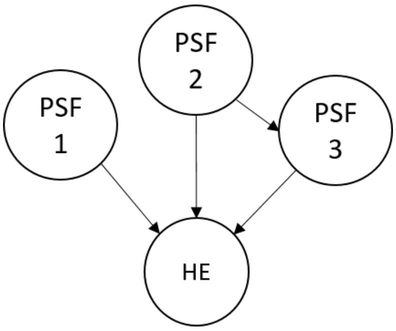


Figure 2. *Example of a simple Bayesian network used for modelling human error.*

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 [37].

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 $\boldsymbol{\pi_i}$:

Equation 1

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

where, x_i represents the status of random variable X_i, $\boldsymbol{\pi_i}$ 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 Eq. 1 becomes:

Equation 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

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

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 exact value of the interval probability (e.g. computation tree [37]), while approximation algorithms provide probabilities near the true value [38]. 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 [38, 39].

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 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 [9]. It is also possible to learn Bayesian network structure from data [27, 40], although this feature is considered more useful for data-rich applications. Usually this is not the case for human reliability data [8]. Instead of choosing between Bayesian networks or fault trees to model human reliability data, one can opt to transform Fault Trees into Bayesian networks [41] or even to combine both, as demonstrated by previous studies that have integrated human reliability Bayesian networks into systems' Fault Tree analysis [42-44]. Besides supporting the evaluation of reduction measures at the organisational level [43], 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* 1 representing the model in *Figure* 2, all states of a combination must sum to one, as defined by a probability axiom [9, 37].

Table 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, *Table* 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* [9], 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 [9]. 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 [26] ; for near-misses reports, events might be underreported to regulators [22]; for accident reports [23], even after scrutinised investigations [29] , some factors might not be observed or reported due to investigators' time, knowledge and bias constraints [45]. 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* [9]. This is also the standard approach used by some Bayesian networks software [46]. However, such strategy implicitly relies on 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 [42]. The functional interpolation [47] and the Cain calculator [48] 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 [8, 47]. *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 [8, 48]. 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 [8, 48]. 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 [42], or by socio-technical systems not necessarily behaving as coherent systems with multistate components [25]. 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 [19].

Expert elicitation is the most common strategy for filling gaps on data. 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 [12], disagreement [7] and overconfidence [13]. 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 [44]. 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 [49]. *Expert elicitation* are limited to the estimation of small CPTs due to humans' inability to estimate the influence of more than three factors simultaneously [14] or the impracticable large number of combinations leading to excessive elicitation burden [50].

Noisy-OR method is the most used model to populate CPTs from partial information, supporting both *expert elicitation* and empirical *data mining* [8, 51]. 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 [8]. To tackle these impediments, extensions have been proposed. The *noisy-MAX model* enabling multi-states nodes [52]; the *recursive noisy-OR* (RNOR) model allows multiple causes as input [53] and inhibition when multiple causes are present to allow the impact of each factor [54]. The *non-impeding noisy-AND tree* allow both reinforcement and undermining effects [51]. However, these Noisy-OR extensions generally address either dependent influences or multi-state nodes rather than both issues simultaneously [8].

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 [9]. 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 [25].

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 [55]. MLE was used in human reliability research to test a modelling approach where performance shaping factors have a joint effect on human error probability [56]. 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[56]. 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 [55].

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 [27, 40]. The approach is also known as *synthesis idiom* (determining synthetic nodes from parents by using a combination rule) [9]. 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* [2]. In a marine engineering application, CREAM [2] has been synthesised by incorporating fuzzy evidential reasoning and Bayesian inference logic to model dependency among common performance conditions [57]. In [27], 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 [58]. *Deriving data from underlying method relationships* reaffirms the importance of the qualitative assessment as changing the structure also changes the amount of information needed [19].

Although data generated in simulators has been traditionally used to validate probabilities obtained by experts [3, 19], recent research investigates its use to fill missing data. In [27], 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 [59], 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 [28]. A summary of lessons learned from challenges in data collection from simulators has been suggested by [26] , 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 [60], risk of fire in residential buildings [61] and railway [39]. 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 [11].

Credal networks are a generalisation of Bayesian networks sharing an identical graphical structure but being characterised by different probability values (Figure 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 [62]. This results in higher flexibility, allowing probabilities to be expressed also in the form of inequalities [10]. Figure 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.

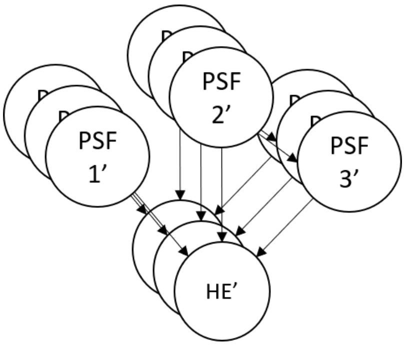


Figure 3. Credal network - a set of Bayesian networks characterised by different probability values.

A credal set, $K(X_i)$, consists of a group with a finite number of probability distributions $P(X_i)$ for a generic random variable X_i. More rigorously, according to the theory of imprecise probability, the credal set is a closed and convex set of probability mass functions [63]. 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: $\underline{P}(X_i = x_i) = \min_{K(X_i=x_i)}\left(P(X_i = x_i)\right)$ and a upper bound $\overline{P}(X_i = x_i) = \max_{K(X_i=x_i)}(P(X_i = x_i))$.

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 [10] . 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 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 6

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

and

Equation 7

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

Where the summation in Eq. 6 and 7 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 [39]:

Equation 8

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

Equation 9

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

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 10

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

Equation 11

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

In the above equations, the summation operator in the nominator acts over all variables, including the queried variable in state x_q *(x_1, . . . , 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, . . . , $x_n\backslash x_q$)*.

In credal networks the computation of the posterior probabilities of the queried nodes requires dedicated inference methods and often approximate approaches are inevitable if using continuous variables [38, 39]. The approximation algorithms used in credal networks can be divided in inner approximation (e.g., linear programming, Hill-climbing [64]) and outer approximation (e.g., branch and bound [64], pseudo-network [39]). The inner and the outer approximations provide probability bounds which enclose the exact probability interval (see *Figure* 4).

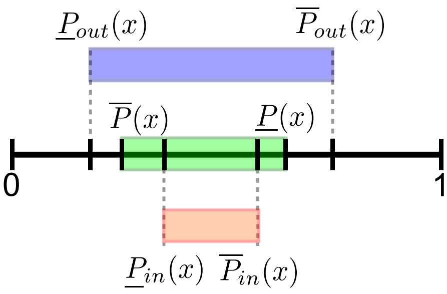


Figure 4. *Inference methods for credal networks*

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 [10, 39]. Freely available packages that implement algorithms to compute credal networks can be found in [10, 38, 65].

3.3. Defining the intervals to replace missing data combinations

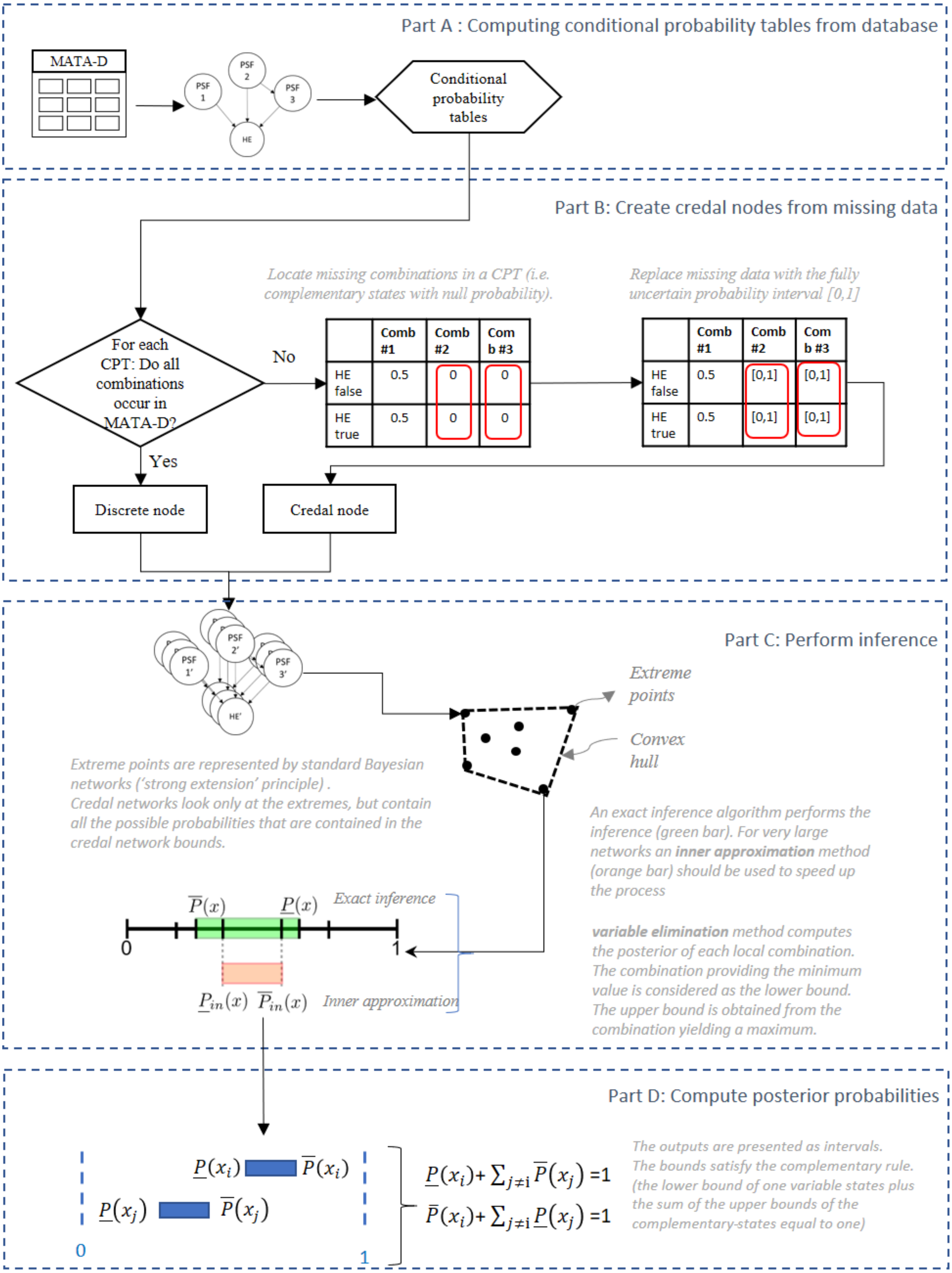
Credal networks are used for handling imprecise and incomplete beliefs of standard Bayesian models where the missing CPT combinations are replaced by intervals comprising the lowest and highest possible probabilities, i.e., zero and one [0,1]. Therefore following the example in in *Table* 1 the replace missing CPT combinations become: P(HE=T|PSF1=T,PSF2=F,PSF3=T)=[0,1] and P(HE=F|PSF1=T,PSF2=F,PSF3=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 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* 8 and *Equation* 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.

3.4. Overview of how the proposed methodology works

The methodology is composed by four main modules and summarised in Figure 5. *Part A* converts MATA-D to prior probabilities in conditional probability tables (detailed procedure is described in a previous study [25], 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 [66]. *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.


Part A : Computing conditional probability tables from database
MATA-D
PSF 1
PSF 2
PSF 3
HE
Conditional probability tables
Part B: Create credal nodes from missing data
Locate missing combinations in a CPT (i.e. complementary states with null probability).
Replace missing data with the fully uncertain probability interval [0,1]
For each CPT: Do all combinations occur in MATA-D?
No
Yes
Discrete node
Credal node
Part C: Perform inference
PSF 1'
PSF 2'
PSF 3'
HE'
Extreme points
Convex hull
Extreme points are represented by standard Bayesian networks ('strong extension' principle) .
Credal networks look only at the extremes, but contain all the possible probabilities that are contained in the credal network bounds.
An exact inference algorithm performs the inference (green bar). For very large networks an **inner approximation** method (orange bar) should be used to speed up the process
variable elimination method computes the posterior of each local combination. The combination providing the minimum value is considered as the lower bound. The upper bound is obtained from the combination yielding a maximum.
$\overline{P}(x)$
$\underline{P}(x)$
Exact inference
0
1
$\underline{P}_{in}(x)$
$\overline{P}_{in}(x)$
Inner approximation
Part D: Compute posterior probabilities
$\underline{P}(x_i)$
$\overline{P}(x_i)$
$\underline{P}(x_j)$
$\overline{P}(x_j)$
0
1
$\underline{P}(x_i)+\sum_{j\neq i}\overline{P}(x_j)=1$
$\overline{P}(x_i)+\sum_{j\neq i}\underline{P}(x_j)=1$
The outputs are presented as intervals. The bounds satisfy the complementary rule. (the lower bound of one variable states plus the sum of the upper bounds of the complementary-states equal to one)

	Comb #1	Comb #2	Com b #3
HE false	0.5	0	0
HE true	0.5	0	0

	Comb #1	Comb #2	Com b #3
HE false	0.5	[0,1]	[0,1]
HE true	0.5	[0,1]	[0,1]

Figure 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 associate accuracy [67]. 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. [68, 69]), challenges remain and the comparison of two of more variables affected by imprecision is not straightforward.

Let consider the simple example shown in *Figure* 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 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 [70]
- Different factors might have overlapping intervals and the most impacting factor might also be the most uncertain one. The *interval dominance* criteria [69] 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 [69].

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

To explain the identified criteria, the pairwise comparison of hypothetical factors shown in Figure 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.

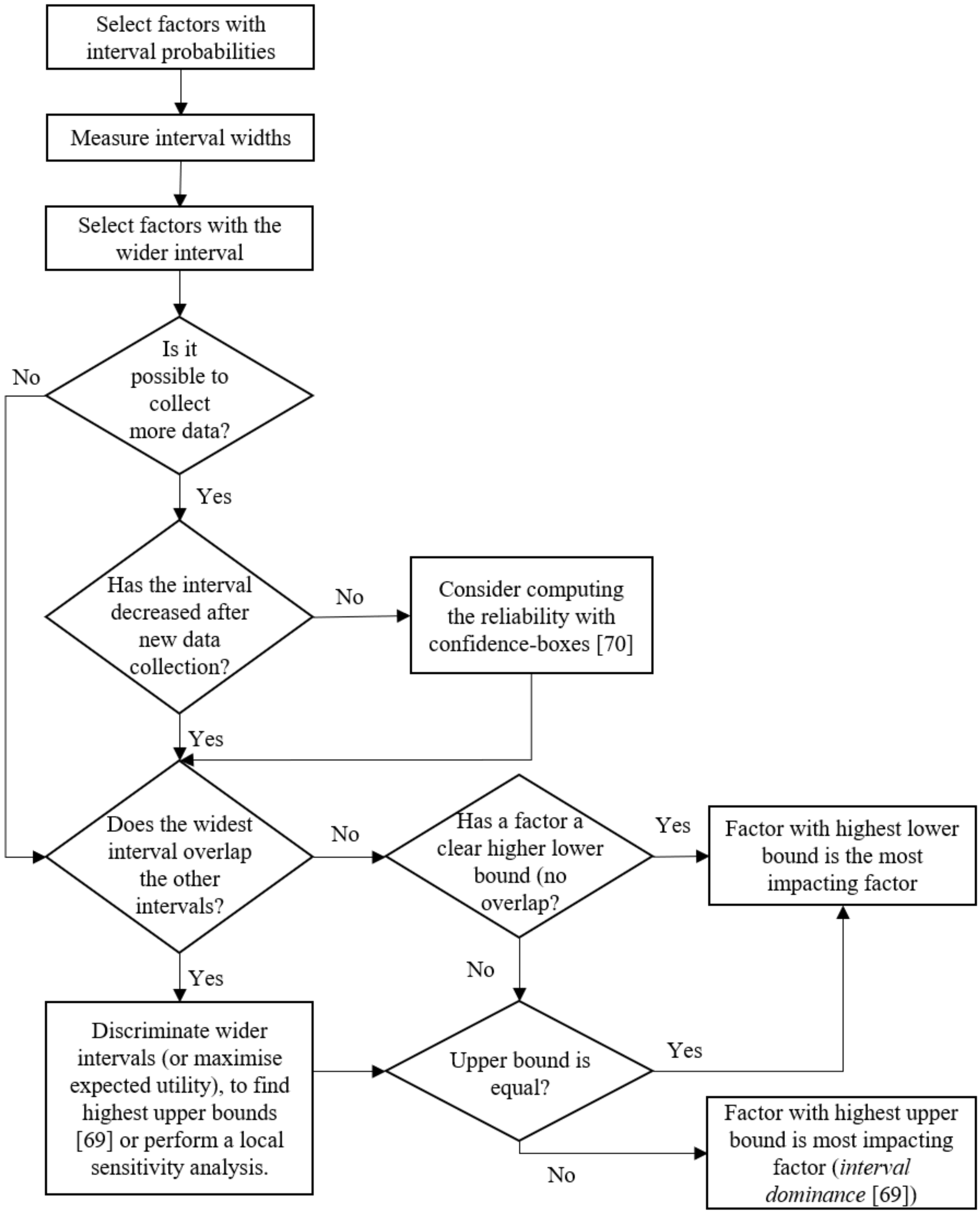
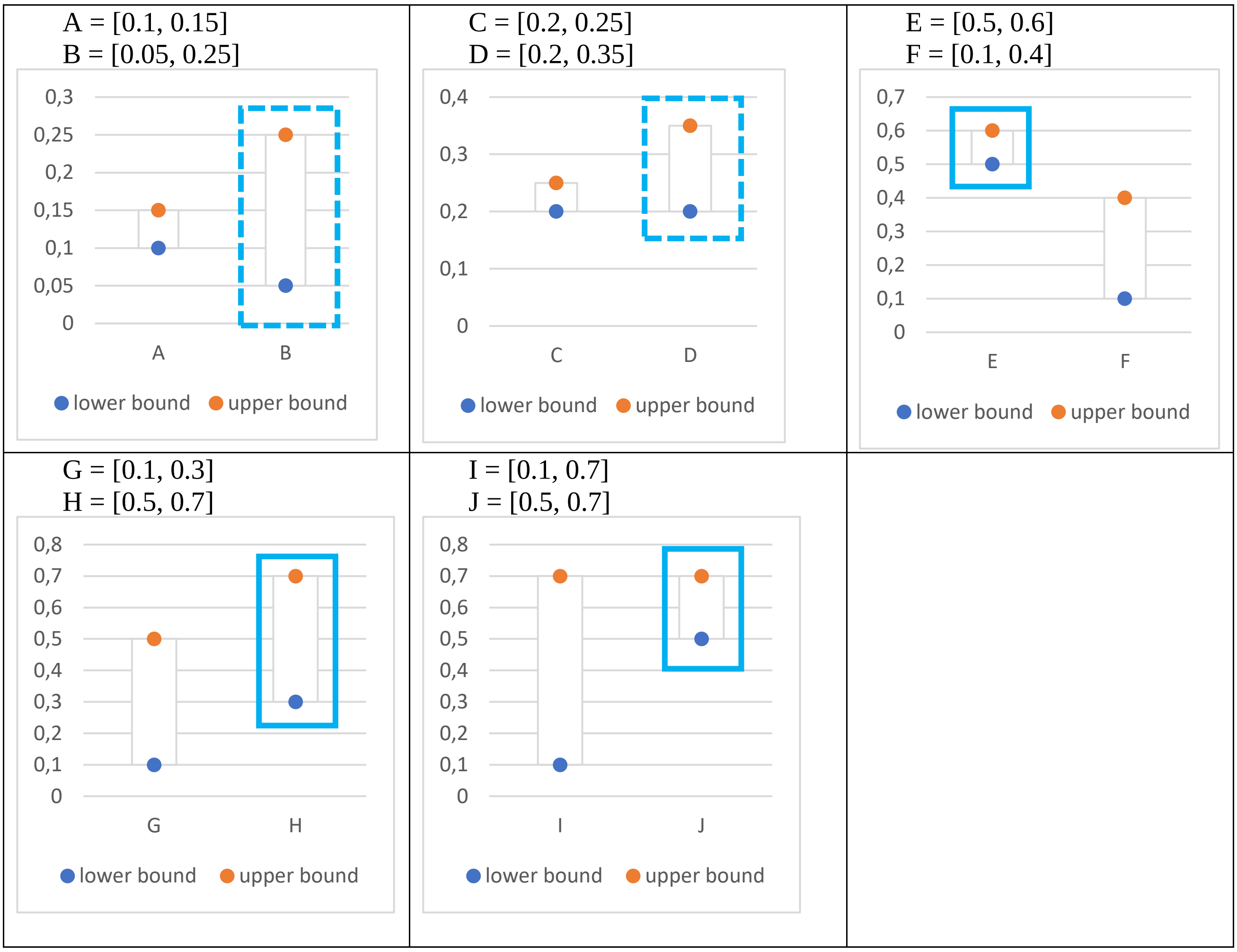

Select factors with interval probabilities
Measure interval widths
Select factors with the wider interval
Is it possible to collect more data?
No
Yes
Has the interval decreased after new data collection?
No
Consider computing the reliability with confidence-boxes [70]
Yes
Does the widest interval overlap the other intervals?
No
Has a factor a clear higher lower bound (no overlap?
Yes
Factor with highest lower bound is the most impacting factor
Yes
No
Discriminate wider intervals (or maximise expected utility), to find highest upper bounds [69] or perform a local sensitivity analysis.
Upper bound is equal?
Yes
Factor with highest upper bound is most impacting factor (*interval dominance* [69])
No

Figure 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* 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.



A = [0.1, 0.15]
B = [0.05, 0.25]

C = [0.2, 0.25]
D = [0.2, 0.35]

E = [0.5, 0.6]
F = [0.1, 0.4]

G = [0.1, 0.3]
H = [0.5, 0.7]

I = [0.1, 0.7]
J = [0.5, 0.7]

Figure 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 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 [60], part of the OpenCossan software [66, 71]. 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 [38, 39]. 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 [72]. 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 [73, 74]. 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 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 10) using a floating hose, or to an FSO via fixed pipelines [73]. Thus, FSOs do not have the production and process facilities (Figure 9).

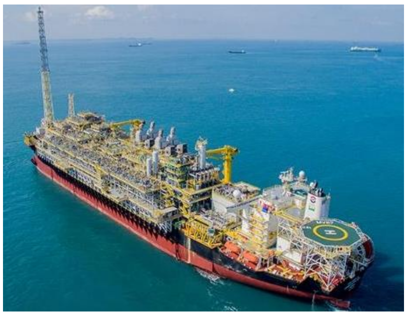


Figure 8. FPSO [1]

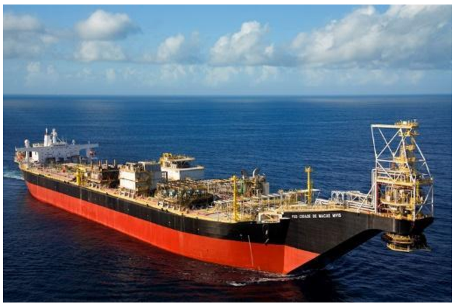


Figure 9. FSO[1]

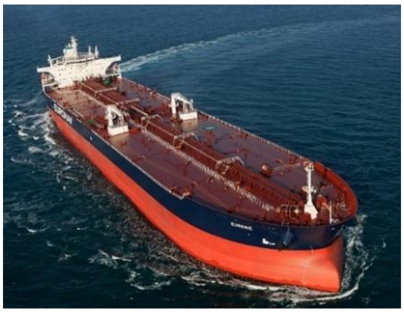


Figure 10. Shuttle tanker [2]

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 11) and avoid overpressure [72]. This depressurisation of oil cargo tanks is known as *cargo venting operation* [75]. 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 [72,74].

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 [75]. In addition, FPSOs/FSOs have their deck space more packed with equipment than tankers (as can be noted by comparing Figure 8 to Figure 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 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 [75].

[1] FPSO and FSO figure source: https://www.modec.com/fps/fpso_fso/lineup/index.html

[2] Shuttle tanker figure source: https://www.hellenicshippingnews.com/oil-tanker-demand-solid-but-trade-tensions-could-change-that/

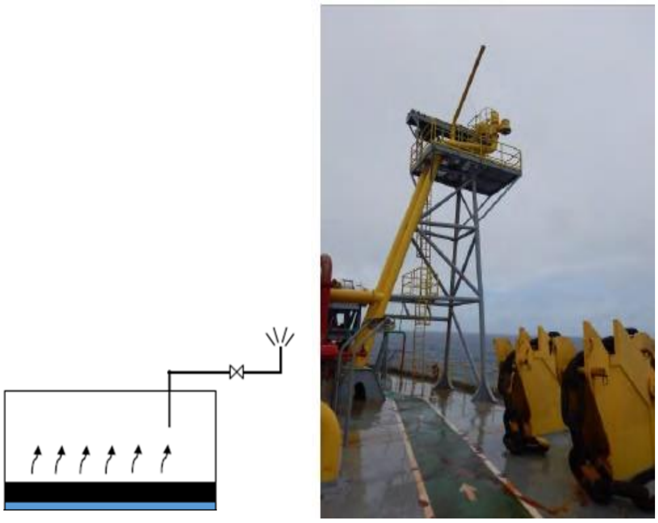


Figure 11. Scheme of a tank with its vent outlet and a photo of a vent outlet on a FPSO[3]

Accidents related to venting operation have the potentiality to create significant financial losses due to the loss or delay of production [73]. For instance, in Brazil, whilst duty holders are increasing their production of lighter crude oil [76], 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 [77], which cause financial loss. Past related incidents have been investigated on relation to the vapour content [74] and possible sources of ignition [78, 79], triggering the UK safety regulator to require duty holders to take appropriate measures to prevent fire and explosion [75].

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 [73, 75]. 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*

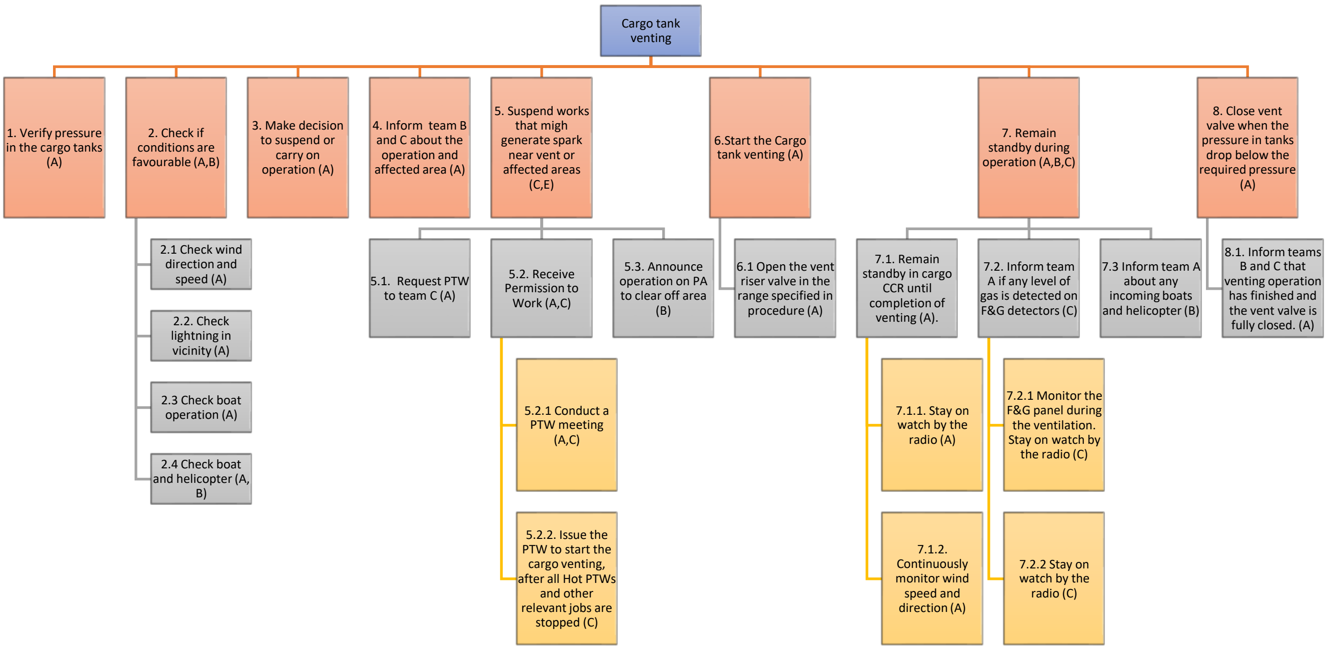
The qualitative part of the study defines the model structure. It was based on the operation's hierarchical task analysis: a structured way of condensing large amount of written information into a sequence of critical actions, screening potential human errors modes, performance shaping factors, and flagging 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 [75], related incidents [74, 78, 79], different design and operational measures [73] 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 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

[3] 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-fpso-cdsm-accident.pdf

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. The Supplementary material provides a detailed description of tasks, their potential human errors and PSFs and the full correlation table adapted from [2]. 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.



Cargo tank venting
1. Verify pressure in the cargo tanks (A)
2. Check if conditions are favourable (A,B)
3. Make decision to suspend or carry on operation (A)
4. Inform team B and C about the operation and affected area (A)
5. Suspend works that migh generate spark near vent or affected areas (C,E)
6.Start the Cargo tank venting (A)
7. Remain standby during operation (A,B,C)
8. Close vent valve when the pressure in tanks drop below the required pressure (A)
2.1 Check wind direction and speed (A)
2.2. Check lightning in vicinity (A)
2.3 Check boat operation (A)
2.4 Check boat and helicopter (A, B)
5.1. Request PTW to team C (A)
5.2. Receive Permission to Work (A,C)
5.3. Announce operation on PA to clear off area (B)
6.1 Open the vent riser valve in the range specified in procedure (A)
7.1. Remain standby in cargo CCR until completion of venting (A).
7.2. Inform team A if any level of gas is detected on F&G detectors (C)
7.3 Inform team A about any incoming boats and helicopter (B)
8.1. Inform teams B and C that venting operation has finished and the vent valve is fully closed. (A)
5.2.1 Conduct a PTW meeting (A,C)
5.2.2. Issue the PTW to start the cargo venting, after all Hot PTWs and other relevant jobs are stopped (C)
7.1.1. Stay on watch by the radio (A)
7.1.2. Continuously monitor wind speed and direction (A)
7.2.1 Monitor the F&G panel during the ventilation. Stay on watch by the radio (C)
7.2.2 Stay on watch by the radio (C)

Figure 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* [9] 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 13*, where discrete nodes are represented by rectangles (child nodes in green, root nodes in blue), and credal nodes by grey ellipses.

An alternative model #2 has been created and shown in *Figure* 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. 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 the Supplementary material, while a brief comparison of both models are presented in results session.

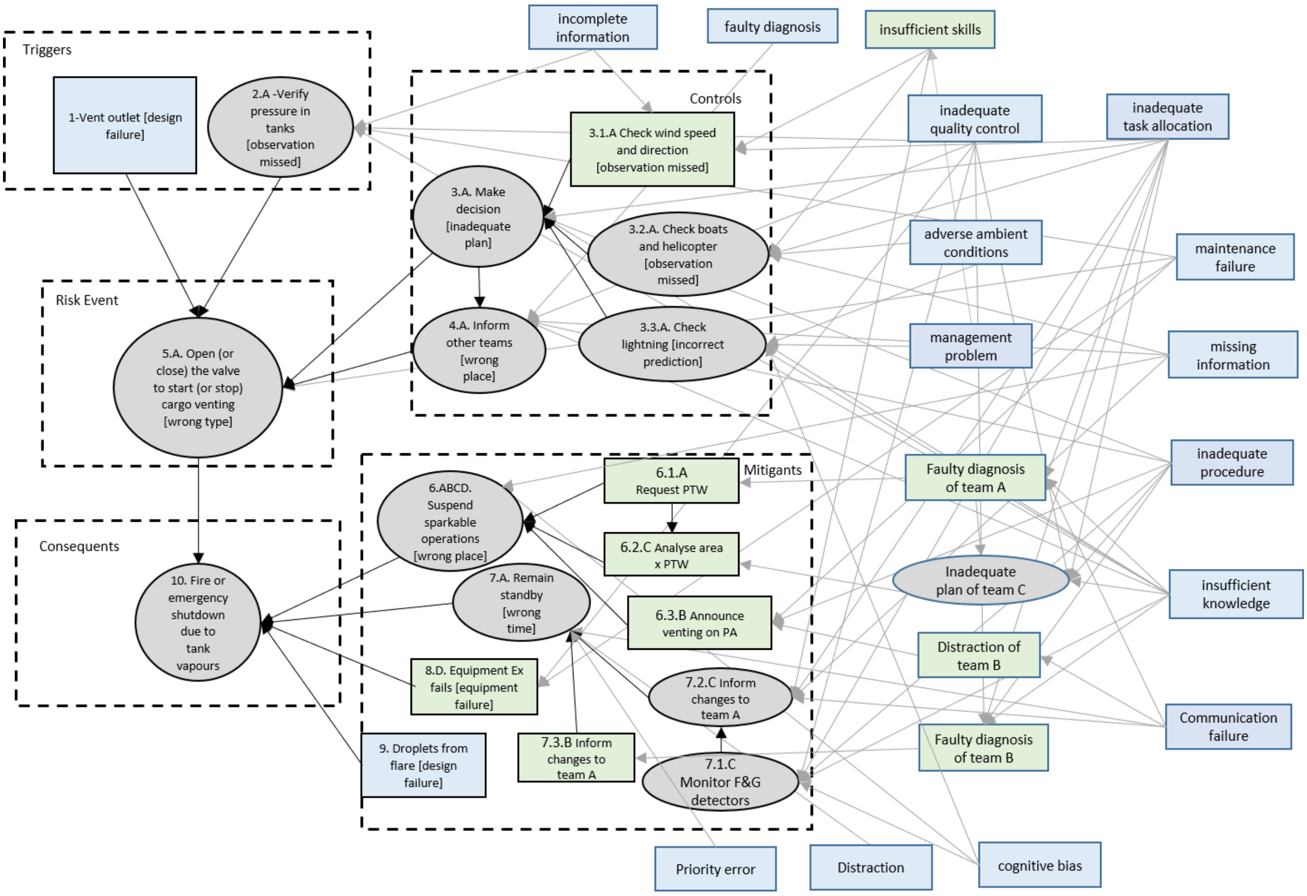

Triggers
1-Vent outlet [design failure]
2.A -Verify pressure in tanks [observation missed]
Risk Event
5.A. Open (or close) the valve to start (or stop) cargo venting [wrong type]
Consequents
10. Fire or emergency shutdown due to tank vapours
Controls
3.1.A Check wind speed and direction [observation missed]
3.A. Make decision [inadequate plan]
3.2.A. Check boats and helicopter [observation missed]
4.A. Inform other teams [wrong place]
3.3.A. Check lightning [incorrect prediction]
Mitigants
6.ABCD. Suspend sparkable operations [wrong place]
6.1.A Request PTW
6.2.C Analyse area x PTW
7.A. Remain standby [wrong time]
6.3.B Announce venting on PA
8.D. Equipment Ex fails [equipment failure]
7.2.C Inform changes to team A
9. Droplets from flare [design failure]
7.3.B Inform changes to team A
7.1.C Monitor F&G detectors
incomplete information
faulty diagnosis
insufficient skills
inadequate quality control
inadequate task allocation
adverse ambient conditions
maintenance failure
management problem
missing information
Faulty diagnosis of team A
inadequate procedure
Inadequate plan of team C
insufficient knowledge
Distraction of team B
Communication failure
Faulty diagnosis of team B
Priority error
Distraction
cognitive bias

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

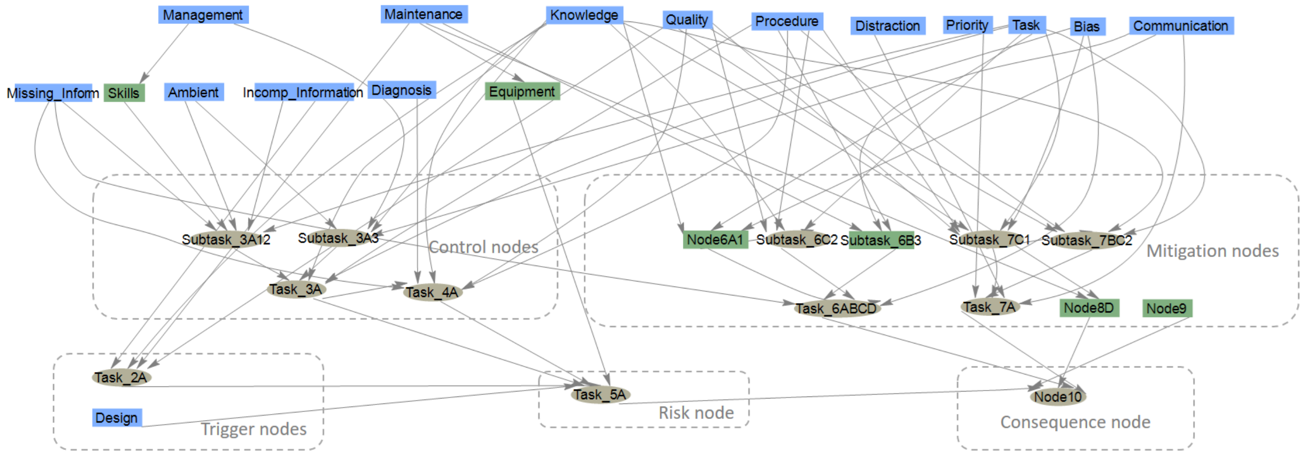


Figure 14. Model #2, some nodes were merged by using underlying CREAM method relationships

Table 2 presents a summarised description of nodes and links of model #1, while model #2 description is presented at Table 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 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 [23, 29]
Task 2A (Observation missed, a 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) *Note (1)*	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) *Note (1)*	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 [2])	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) *Note (2)*	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) *Note (2)*	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) *Note (2)*	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 Parent node of distraction of team B: communication failure	two (true/ false)	Discrete	MATA-D
Task 7A (Action performed at wrong time (an 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) *Note (3)*	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) *Note (3)*	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 hydrocarbon releases database [80]
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 [69]; UK FPSOs [70,71]; UK offshore hydrocarbon releases database [80]

Note (1): In 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. 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) ***(different from Model #1, due to subtasks)***	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) ***(different from Model #1)***	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) ***(different from Model #1)***	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
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) ***(different from Model #1)***	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
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) ***(different from Model #1)***	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
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) ***(different from model #1, due to some different PSFs)***	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
Node subtasks 7.2.BC (faulty diagnosis, cognitive function failure) ***(different from model #1)***	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
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 [2], 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) [23, 29]. 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 [25].

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 [80]. 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 [77]. The information about frequency of droplets from flare in FPSOs was obtained from [80], 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 [74, 78, 79].

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 Supplementary material). 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* 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 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 [23]	FALSE	0.44
	TRUE	0.56
Node PSF 9 *(Droplets from flare) from* ***[80]***	FALSE	$9.97x\ 10^{-1}$
	TRUE	$3.0\ x\ 10^{-3}$

Table 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 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 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* 6 shows the frequency this same context occurred in accidents recorded in MATA-D. Differently from CPTs shown in *Table* 4 and *Table* 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 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, 1]**	1.0	**[0, 1]**	1.0	**[0, 1]**	1.0	**[0, 1]**	1.0	**[0, 1]**
Incorrect prediction *TRUE*	0.01	0.07	0.09	0.0	0.0	0.0	0.12	**[0, 1]**	0.0	**[0, 1]**	0.0	**[0, 1]**	0.0	**[0, 1]**	0.0	**[0, 1]**

The complete CPTs for all nodes can be found on the Supplementary material. More details on how to convert the relative frequencies from MATA-D to the CPTs can be accessed on [25].

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

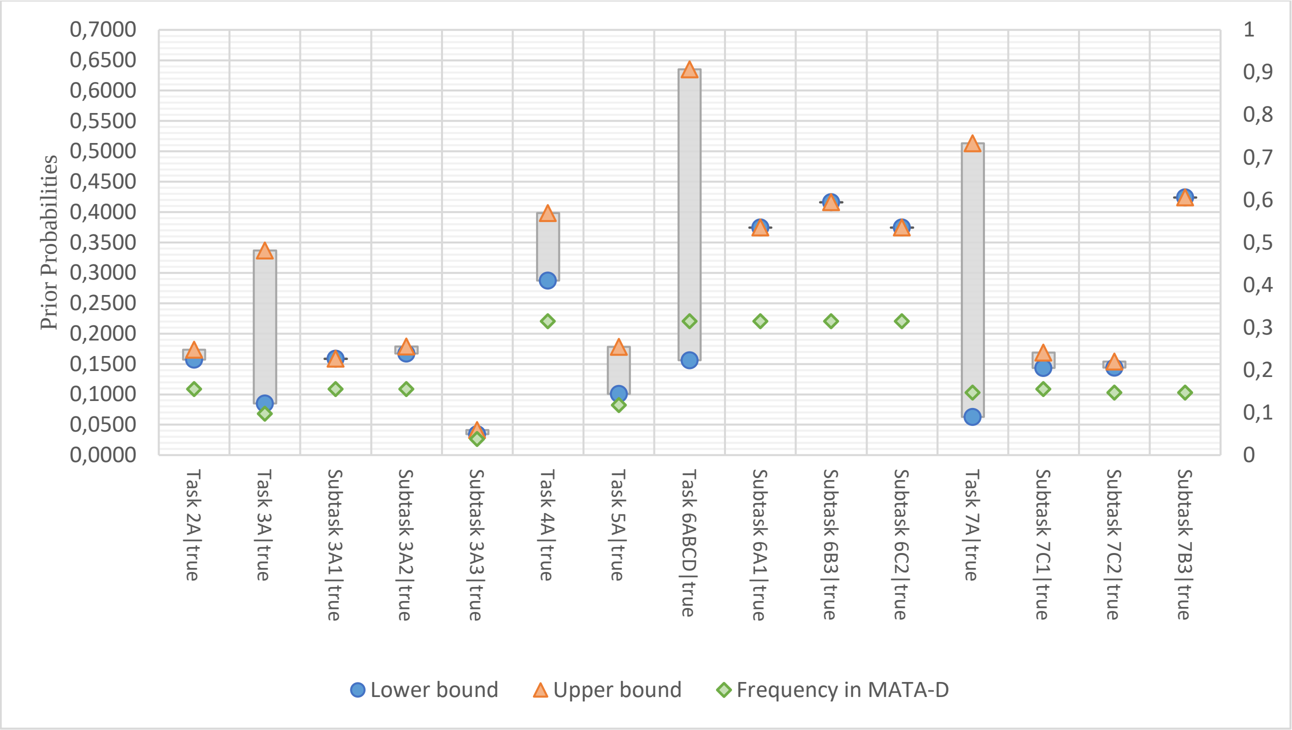
The results of the predictive analysis are presented in *Table* 7 for model #1, Figures 15 and 16 for the model #1 and Figures 17 and 18 for model #2, while some possible diagnostic analysis are presented from *Table* 8 and from Figure 19. In *Table* 7, the posterior probabilities are presented for all 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 15, which depicts the true states of trigger, control, mitigation and risk event nodes, and Figure 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 [23]. 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.

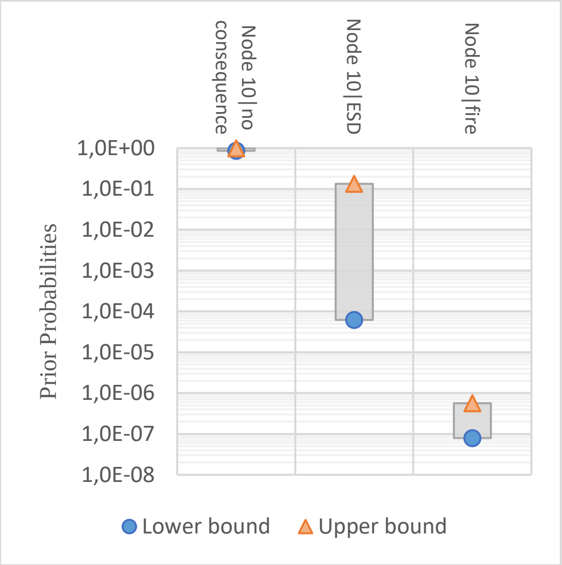
Table 7. Prediction of posterior probabilities in all variable states (model #1)

Event	**State**	**Lower bound**	**Upper bound**
TRIGGERS			
Task 2A *(observation missed)*	FALSE	0.83	0.84
	TRUE	0.16	0.17
CONTROL BARRIERS			
Task 3A *(inadequate plan)*	FALSE	0.66	0.92
	TRUE	0.08	0.34
Subtask 3.1A *(observation missed)*	FALSE	0.84	0.84
	TRUE	**0.16**	**0.16**
Subtask 3.2A *(observation missed)*	FALSE	0.82	0.83
	TRUE	0.17	0.18
Subtask 3.3A *(incorrect prediction)*	FALSE	0.96	0.97
	TRUE	0.034	0.04
Task 4A *(action in wrong place)*	FALSE	0.60	0.71
	TRUE	0.29	0.40
RISK EVENT			
Task 5A *(execution of wrong type)*	FALSE	0.82	0.90
	TRUE	0.10	0.18
MITIGATION BARRIERS			
Task 6 ABCD *(action in wrong place)*	FALSE	0.37	0.84
	TRUE	0.16	0.63
Subtask 6.1A *(action in wrong place)*	FALSE	0.62	0.62
	TRUE	0.38	0.38
Subtask 6.2C *(action in wrong place)*	FALSE	0.62	0.62
	TRUE	0.38	0.38
Subtask 6.3B *(action in wrong place)*	FALSE	0.58	0.58
	TRUE	0.42	0.42
Task 7A *(action performed at wrong time)*	FALSE	0.49	0.94
	TRUE	0.06	0.51
Task 7.1C *(observation missed)*	FALSE	0.83	0.86
	TRUE	0.14	0.17
Task 7.2C *(action performed at wrong time)*	FALSE	0.85	0.86
	TRUE	0.14	0.15
Task 7.3B *(action performed at wrong time)*	FALSE	0.58	0.58
	TRUE	0.42	0.42
CONSEQUENCE			
Node 10 *(consequence of hazard event)*	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}



806

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



808

809 *Figure 16. Posterior probabilities for the three states of the consequence node of model #1*

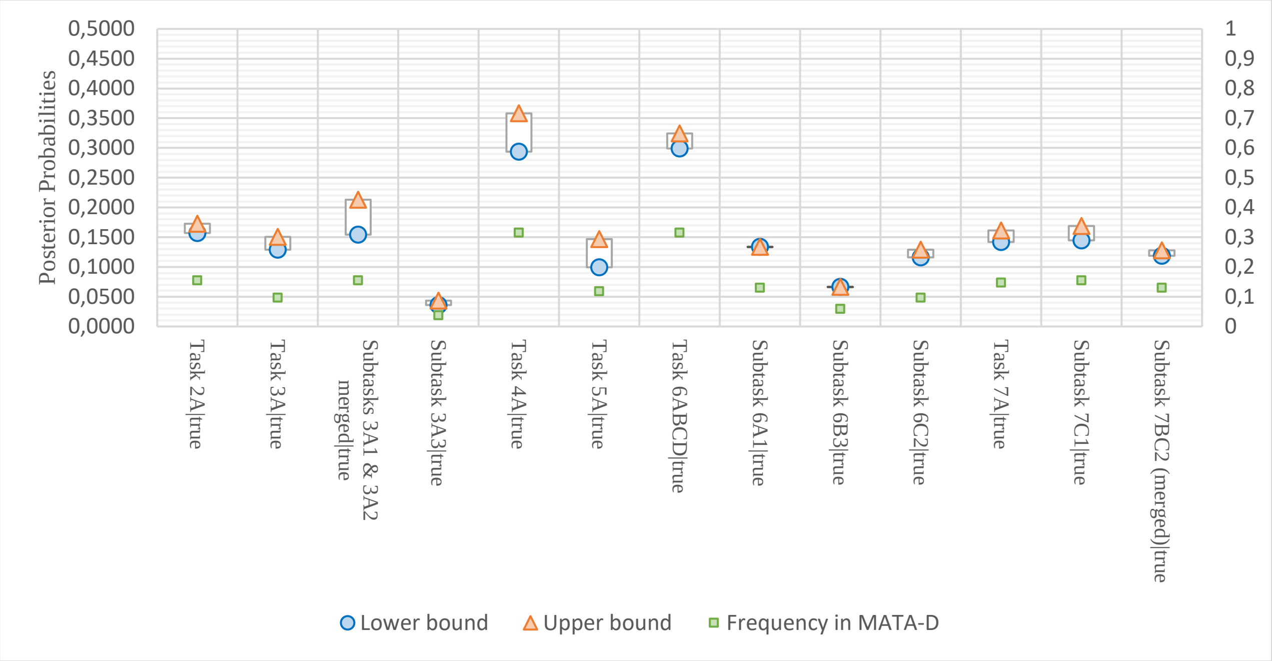


Figure 17. Point and interval posterior probabilities for the cargo venting human reliability model #2

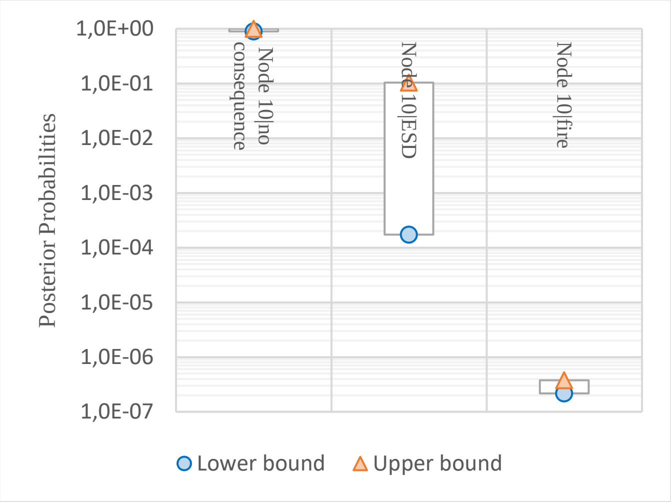


Figure 18. Posterior probabilities of three states of consequence node in model #2

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 the Supplementary material.

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 19 shows the sensitivity analysis for *task 5A* of model #1 to preceding tasks while Figure 20 presents the sensitivity analysis with respect to the PSFs. The probability values of the sensitivity analysis of task 5A are reported in *Table* 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).

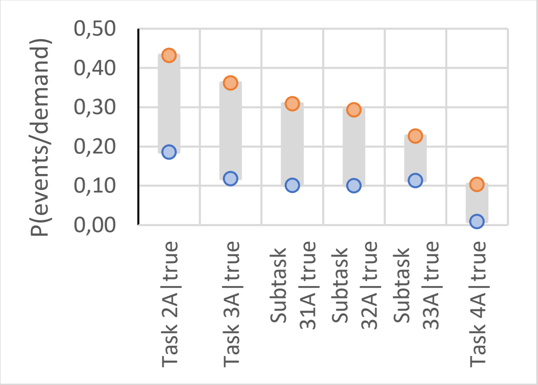


Figure 19.Task 5A|true - sensitivity to tasks (model #1)

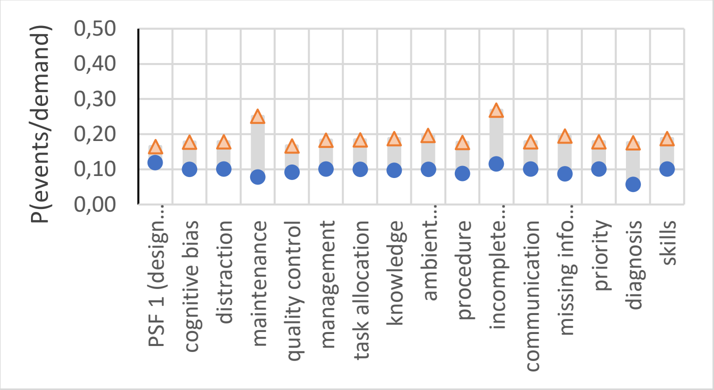


Figure 20. Task 5A|true - sensitivity to PSFs (model #1)

Table 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* 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.

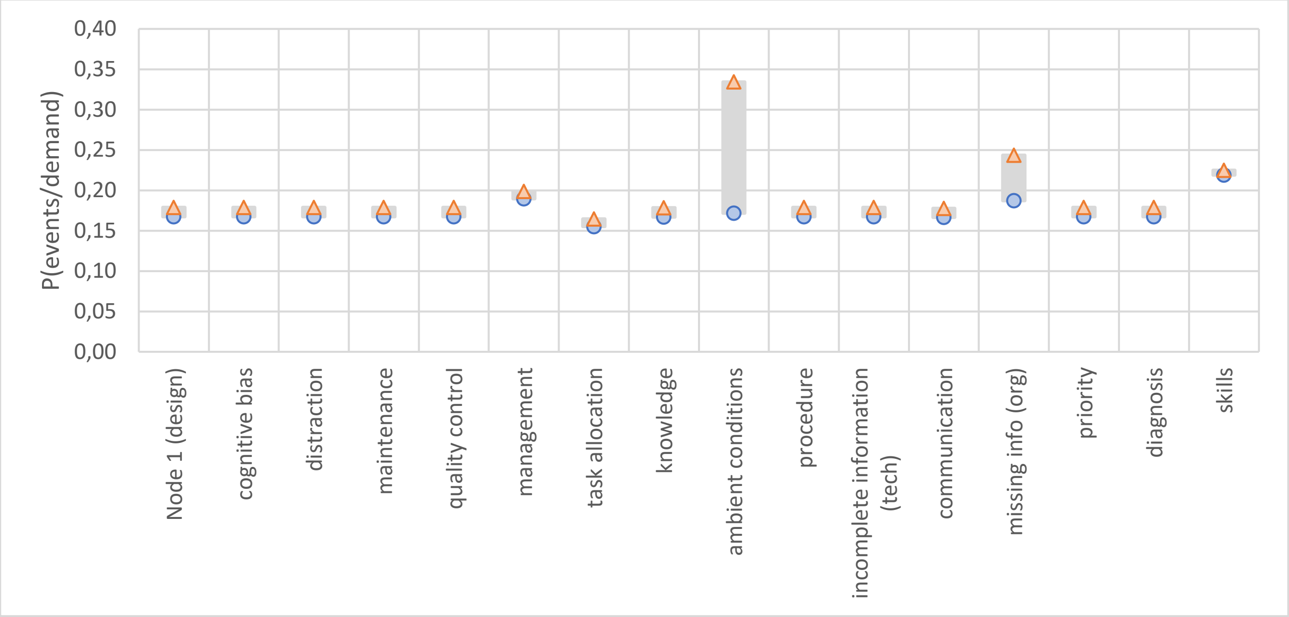


Figure 21. *Node 3.2A|true - sensitivity to PSFs*

Figure 22 to Figure 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.

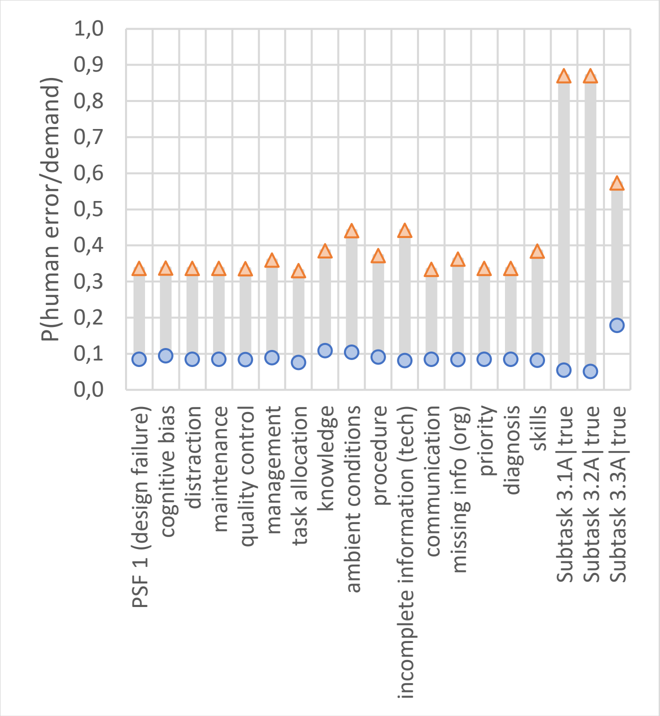


Figure 22. Node 3A|true - sensitivity to PSFs and subtasks 3.1A, 3.2A & 3A3 (model #1)

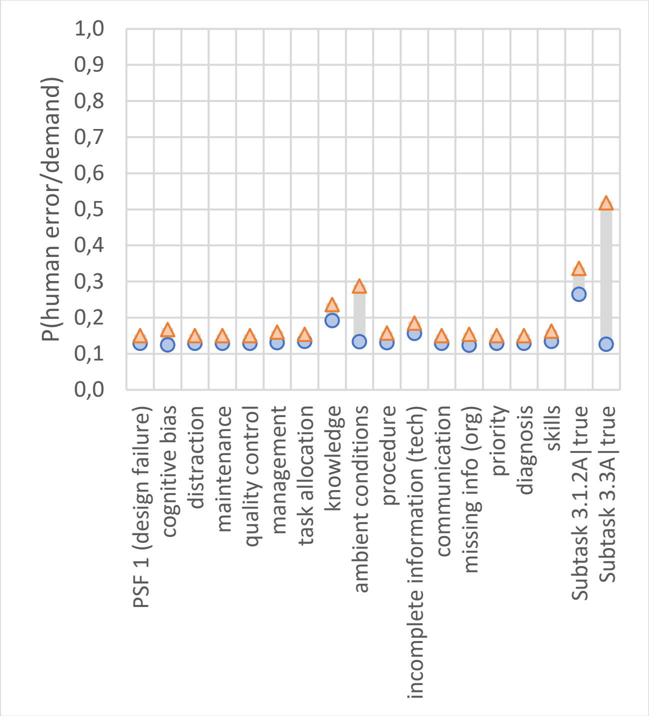


Figure 23. Task 3A|true sensitivity to PSFs and subtasks 3.1.2A and 3.3A (model #2)

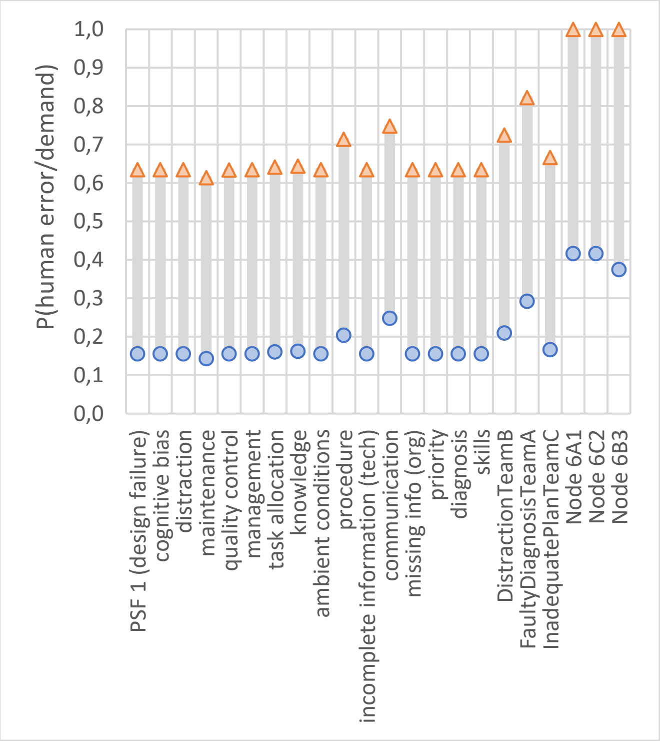


Figure 24. Task 6ABCD|true sensitivity to PSFs and subtasks 6.1A, 6.2C, 6.3B (model #1)

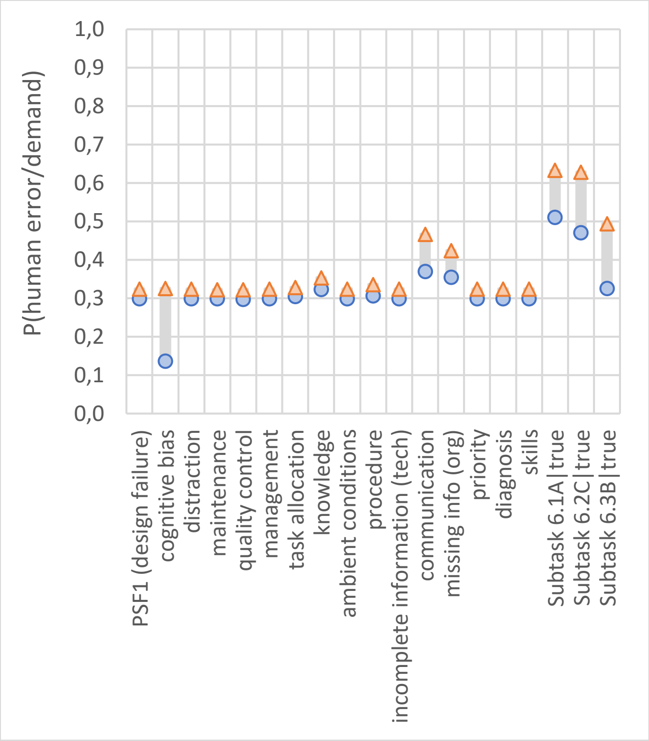


Figure 25. Task 6ABCD|true - sensitivity to PSFs and subtasks 6.1A, 6.2C & 6.3B (model #2)

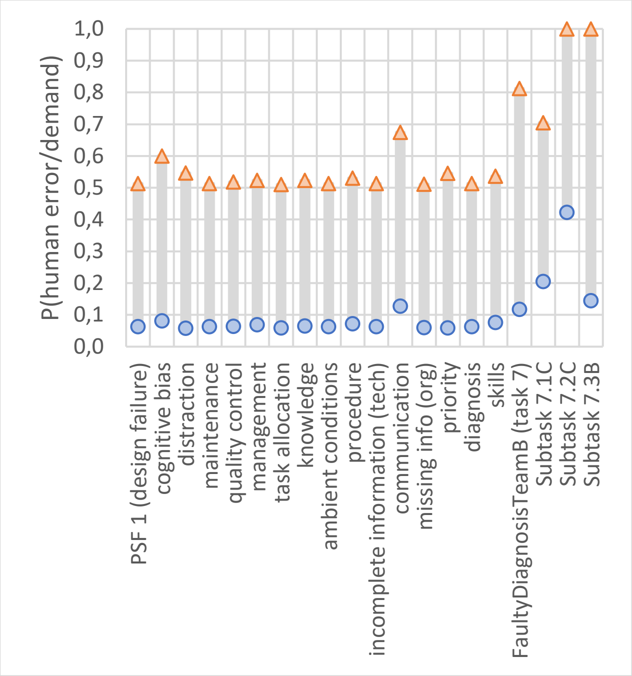


Figure 26. Task 7A|true sensitivity to PSFs and subtasks 7.1C, 7.2C and 7.3B (model #1)

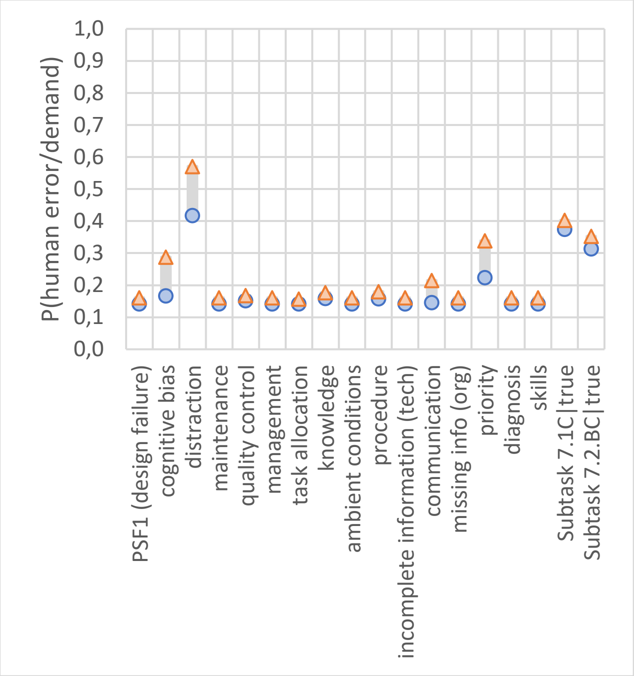


Figure 27. Task 7A|true - sensitivity to PSFs and subtasks 7.1C & 7.2BC (model #2)

Table 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* 28 is the graphical representation of intervals for ESD sensitivity, represented in logarithmic scale to facilitate the analysis of lower bounds. Figure 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 9. *Sensitivity analysis to tasks and PSFs of ESD and fire occurring as a consequence (model #1)*

Evidence on node	Node 10|ESD queried *P(event/days)*		Node 10 | fire queried *P(event/days)*	
	Lower bound	Upper bound	Lower bound	Upper bound
Performance Shaping Factors				
Node PSF 1 (design failure)	6.97 x10^{-5}	0.13	8.67 x 10^{-8}	5.50 x 10^{-7}
Node PSF 9 (droplets from flare)	9.47 x10^{-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	1.37 x10^{-4}	0.14	6.20 x 10^{-8}	5.71 x 10^{-7}
Distraction	9.56 x10^{-5}	0.13	7.00 x 10^{-8}	5.75 x 10^{-7}
Maintenance failure	5.63 x10^{-5}	0.19	4.37 x 10^{-8}	6.20 x 10^{-7}
Inadequate quality control	5.75 x10^{-5}	0.13	6.74 x 10^{-8}	4.93 x 10^{-7}
Management problem	6.77 x10^{-5}	0.14	7.77 x 10^{-8}	5.78 x 10^{-7}
Inadequate task allocation	5.64 x10^{-5}	0.15	7.41 x 10^{-8}	5.70 x 10^{-7}
Insufficient knowledge	6.02 x10^{-5}	0.14	7.67 x 10^{-8}	5.95 x 10^{-7}
Adverse ambient conditions	6.60 x10^{-5}	0.14	7.95 x 10^{-8}	6.17 x 10-7
Inadequate procedure	6.08 x10^{-5}	0.13	5.48 x 10^{-8}	5.25 x 10-7
Incomplete information (technology)	8.68 x10^{-5}	0.20	8.58 x 10^{-8}	9.43 x 10^{-7}
Communication failure	1.40 x10^{-4}	0.14	4.15 x 10^{-8}	4.80 x 10^{-7}
Missing information (organisation)	8.96 x10^{-5}	0.14	6.27 x 10^{-8}	6.83 x 10^{-7}
Priority error	7.45 x10^{-5}	0.13	7.43 x 10^{-8}	5.78 x 10^{-7}
Faulty diagnosis	4.80 x10^{-5}	0.12	5.18 x 10^{-8}	5.47 x 10^{-7}
Insufficient skills	7.40 x10^{-5}	0.14	7.57 x 10^{-8}	5.92 x 10^{-7}
Distraction of team B	5.17 x10^{-5}	0.14	5.69 x 10^{-8}	5.24 x 10^{-7}
Faulty diagnosis of team A	3.69 x10^{-5}	0.15	3.65 x 10^{-8}	4.88 x 10^{-7}
Faulty diagnosis of team B	1.01 x10^{-4}	0.14	2.73 x 10^{-8}	4.94 x 10^{-7}
Inadequate plan of team C	6.29 x10^{-5}	0.14	7.18 x 10^{-8}	5.60 x 10^{-7}
Tasks and subtasks				
Task 2A|true	1.34 x10^{-4}	0.32	1.19 x 10^{-7}	1.60 x 10^{-6}
Task 3A|true	1.26 x10^{-4}	0.27	1.06 x 10^{-7}	1.38 x 10^{-6}
Subtask 31A|true	8.79 x10^{-5}	0.20	7.86 x 10^{-8}	1.12 x 10^{-6}
Subtask 32A|true	7.02 x10^{-5}	0.20	7.44 x 10^{-8}	9.95 x 10^{-7}
Subtask 33A|true	1.14 x10^{-4}	0.16	8.14 x 10^{-8}	7.55 x 10^{-7}
Task 4A|true	1.61 x10^{-5}	0.07	6.89 x 10^{-9}	2.65 x 10^{-7}
Task 5A|true	5.10 x10^{-4}	0.84	3.90 x 10^{-7}	1.99 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	6.72 x10^{-4}	0.14	0	0
Subtask 7.1C|true	1.92 x10^{-4}	0.14	4.66 x 10^{-8}	4.91 x 10^{-7}
Subtask 7.2C|true	3.79 x10^{-4}	0.14	0	3.71 x 10^{-7}
Subtask 7.3B|true	1.32 x10^{-4}	0.14	0	5.17 x 10^{-7}

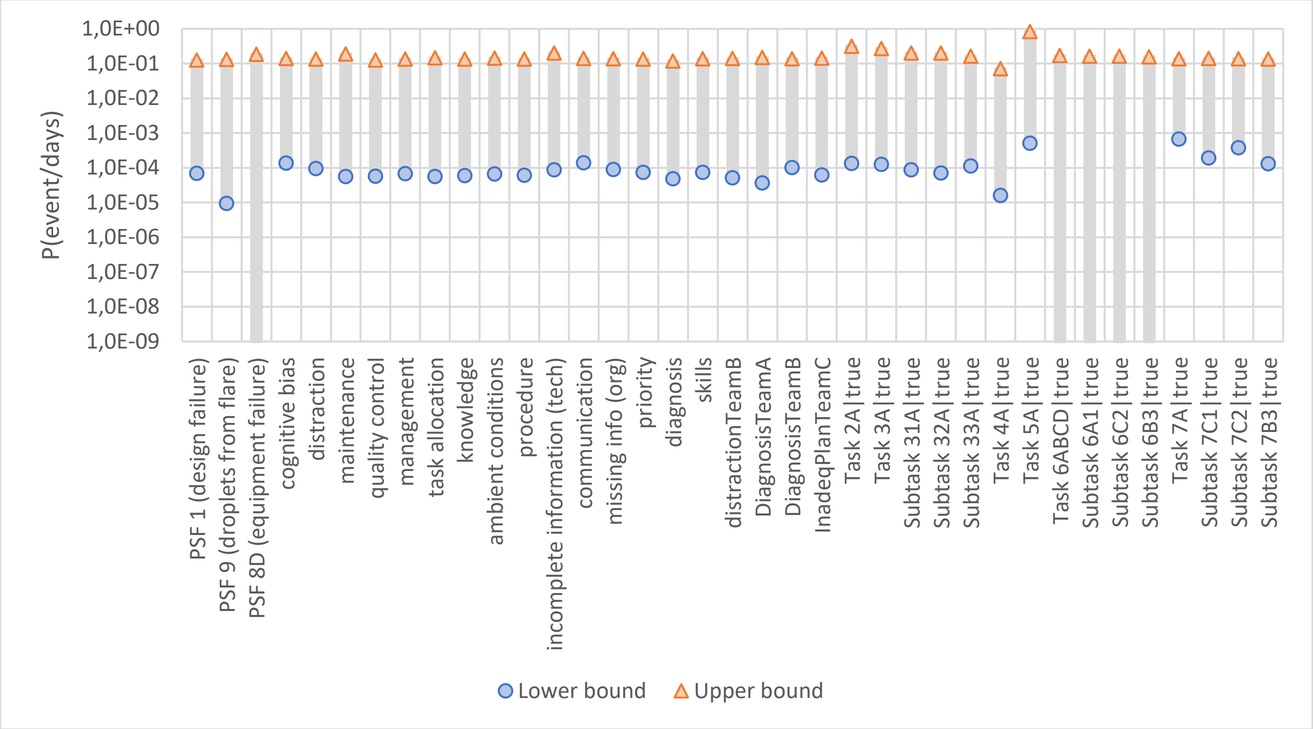


Figure 28. *Sensitivity Node 10|ESD (in log scale).*

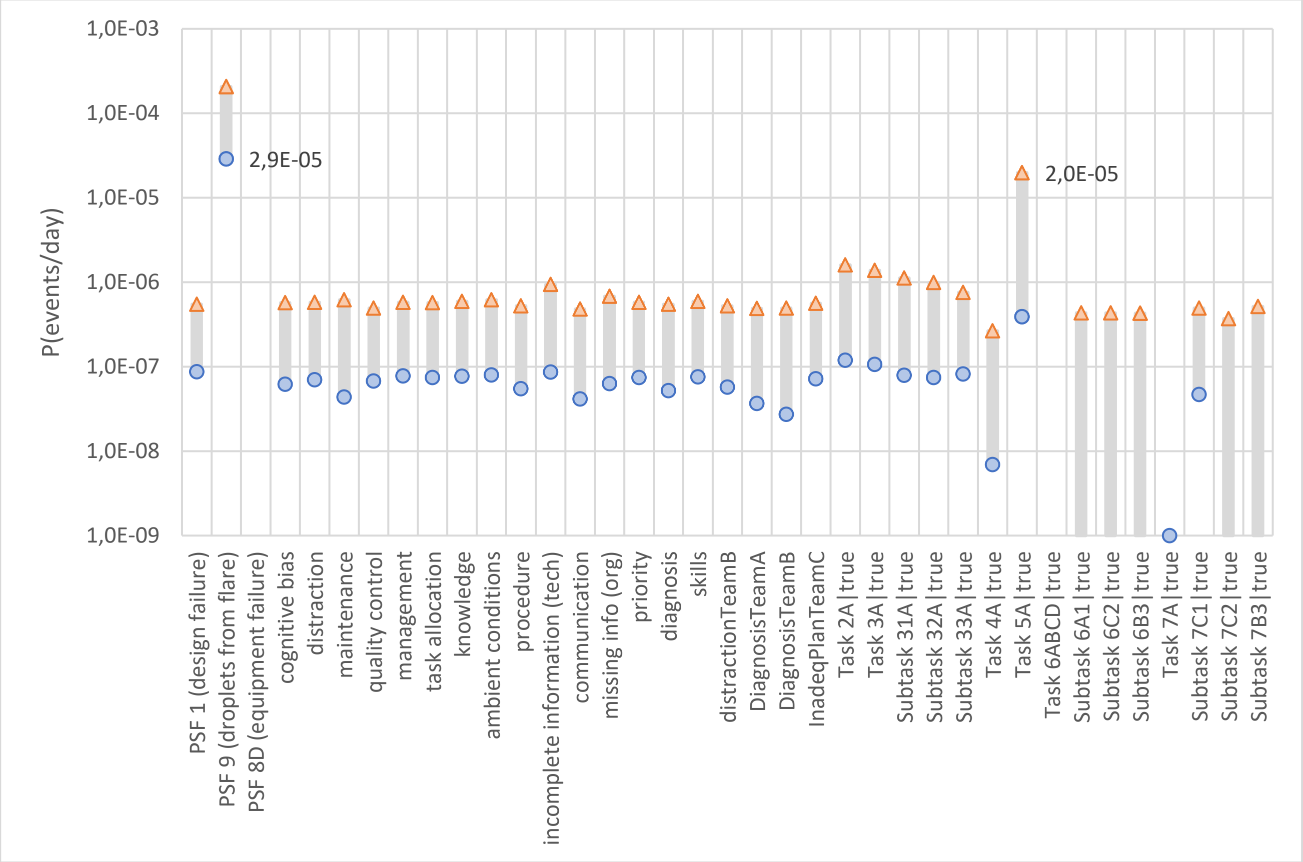


Figure 29. *Node 10|Fire - sensitivity to tasks and PSFs (log scale)*

Table 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 10. *Summary of most ~~influencing~~ important factors in ~~tasks of~~ model #1 ~~and #2~~ (in bold where ~~both~~ model~~s~~ #2 agrees)*

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 *(equals to 3.1.2A in model #2)*	**PSF incomplete information (tech factor)**	**PSF ambient conditions, followed by incomplete information (tech factor)**
Subtask 3.2A *(equals to 3.1.2A in model #2)*	**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	
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)**

4.5. 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 16).

It has been observed that, the majority of relative frequencies from MATA-D [23] 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

[81]. 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 [82]. 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 [25].

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 [82]. 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 [2]). The most important organisational factor is *maintenance failure* (i.e. missing or inappropriate management of maintenance leading to equipment not operational or indicators not working [2]), 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 26 and Figure 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* 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.

Acknowledgements

Caroline Morais gratefully acknowledges the Brazilian Oil & Gas regulator ANP (Agencia Nacional do Petroleo, Gas Natural e Biocombustiveis) for the support for her research. Hector Diego Estrada-Lugo gratefully acknowledges the Consejo Nacional de Ciencia y Tecnologıa (CONACyT) for the scholarship awarded by the Mexican government for graduate studies. Edoardo Patelli was partially supported by the EPSRC grant EP/R020558/2 Resilience Modelling Framework for Improved Nuclear Safety (NuRes).

Supplementary material *(see files on hyperlink provided)*

References

[1] Kirwan B. A guide to practical human reliability assessment: CRC press; 1994.
[2] Hollnagel E. Cognitive reliability and error analysis method (CREAM): Elsevier; 1998.
[3] Kirwan B. Validation of human reliability assessment techniques: part 1—validation issues. Safety Science. 1997;27:25-41.
[4] French S, Bedford T, Pollard SJT, Soane E. Human reliability analysis: A critique and review for managers. Safety Science. 2011;49:753-63.
[5] Reason J. Managing the risks of organizational accidents: Routledge; 2016.
[6] Sklet S. Safety barriers: Definition, classification, and performance. Journal of Loss Prevention in the Process Industries. 2006;19:494-506.
[7] Mkrtchyan L, Podofillini L, Dang VN. Bayesian belief networks for human reliability analysis: A review of applications and gaps. Reliability Engineering & System Safety. 2015;139:1-16.

[8] Mkrtchyan L, Podofillini L, Dang VN. Methods for building conditional probability tables of bayesian belief networks from limited judgment: an evaluation for human reliability application. Reliability Engineering & System Safety. 2016;151:93-112.

[9] Fenton N, Neil M. Risk assessment and decision analysis with Bayesian networks: Crc Press; 2012.

[10] Cozman FG. Credal networks. Artificial Intelligence. 2000;120:199-233.

[11] 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 ESREL2019.

[12] Mosleh A, Bier VM, Apostolakis G. A critique of current practice for the use of expert opinions in probabilistic risk assessment. Reliability Engineering & System Safety. 1988;20:63-85.

[13] Lin S-W, Bier VM. A study of expert overconfidence. Reliability Engineering & System Safety. 2008;93:711-21.

[14] Evans JSBT, Handley SJ, Over DE. Conditionals and conditional probability. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2003;29:321.

[15] Griffith CD, Mahadevan S. Human reliability under sleep deprivation: Derivation of performance shaping factor multipliers from empirical data. Reliability Engineering & System Safety. 2015;144:23-34.

[16] Di Flumeri G, De Crescenzio F, Berberian B, Ohneiser O, Kramer J, Aricò P, et al. 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. 2019;13.

[17] Jung W, Park J, Kim Y, Choi SY, Kim S. HuREX–A framework of HRA data collection from simulators in nuclear power plants. Reliability Engineering & System Safety. 2020;194:106235.

[18] Chang YJ, Bley D, Criscione L, Kirwan B, Mosleh A, Madary T, et al. The SACADA database for human reliability and human performance. Reliability Engineering & System Safety. 2014;125:117-33.

[19] NRC UNRC. The international HRA empirical study: lessons learned from comparing HRA methods predictions to HAMMLAB simulator data, NUREG-2127. US Nuclear Regulatory Commission, Washington, DC. 2014.

[20] Xing J, Parry G, Presley M, Forester J, Hendrickson S, Dang V. 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. 2016.

[21] 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: Springer; 2017. p. 230-8.

[22] Preischl W, Hellmich M. Human error probabilities from operational experience of German nuclear power plants. Reliability Engineering & System Safety. 2013;109:150-9.

[23] Moura R, M. B, E. P, J. L, Knoll F. Multi-Attribute Technological Accidents Dataset (MATA-D). 2020.

[24] Kyriakidis M, Majumdar A, Ochieng WY. Data based framework to identify the most significant performance shaping factors in railway operations. Safety Science. 2015;78:60-76.

[25] Morais C, Moura R, Beer M, Patelli E. Analysis and estimation of human errors from major accident investigation reports. ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg. 2020;6.

[26] Kim Y. Considerations for generating meaningful HRA data: Lessons learned from HuREX data collection. Nuclear Engineering and Technology. 2020.

[27] Groth KM, Mosleh A. 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. 2012;226:361-79.

[28] Sundaramurthi R, Smidts C. Human reliability modeling for the next generation system code. Annals of Nuclear Energy. 2013;52:137-56.

[29] Moura R, Beer M, Patelli E, Lewis J, Knoll F. Learning from major accidents to improve system design. Safety Science. 2016;84:37-45.

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

[31] Siegrist J. Mixing good data with bad: how to do it and when you should not. Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management2011. p. 368-73.

[32] Smith E, Anne Koop DNV, King UKS. Guidance on Human Factors Critical Task Analysis. In: IChemE, editor. Hazards XXII Process Safety and Environmental Protection2011.

[33] CGE RMS. The history of bowtie. 2017.

[34] Salvi O, Debray B. A global view on ARAMIS, a risk assessment methodology for industries in the framework of the SEVESO II directive. Elsevier; 2006. p. 187-99.

[35] Targoutzidis A. Incorporating human factors into a simplified "bow-tie" approach for workplace risk assessment. Safety Science. 2010;48:145-56.

[36] Léger A, Weber P, Levrat E, Duval C, Farret R, Iung B. Methodological developments for probabilistic risk analyses of socio-technical systems. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability. 2009;223:313-32.

[37] Nielsen TD, Jensen FV. Bayesian networks and decision graphs: Springer Science & Business Media; 2009.

[38] Tolo S, Patelli E, Beer M. An open toolbox for the reduction, inference computation and sensitivity analysis of Credal Networks. Advances in Engineering Software. 2018;115:126-48.

[39] Estrada-Lugo HD, Tolo S, de Angelis M, Patelli E. Pseudo credal networks for inference with probability intervals. ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg. 2019;5.

[40] Groth KM, Smith R, Moradi R. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. Reliability Engineering & System Safety. 2019;191:106507.

[41] Bobbio A, Portinale L, Minichino M, Ciancamerla E. Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. Reliability Engineering & System Safety. 2001;71:249-60.

[42] Martins MR, Maturana MC. Application of Bayesian Belief networks to the human reliability analysis of an oil tanker operation focusing on collision accidents. Reliability Engineering & System Safety. 2013;110:89-109.

[43] Trucco P, Cagno E, Ruggeri F, Grande O. A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. Reliability Engineering & System Safety. 2008;93:845-56.

[44] Ramos MA, Droguett EL, Mosleh A, Moura MDC. A human reliability analysis methodology for oil refineries and petrochemical plants operation: Phoenix-PRO qualitative framework. Reliability Engineering & System Safety. 2020;193:106672.

[45] Kletz T. Some Common Errors in Accident Investigations. Safety and Reliability. 1 ed: Taylor & Francis; 2011. p. 4-13.

[46] Bencomo NGPFCHD, Blair G. GeNIe Modeler.

[47] Podofillini L, Mkrtchyan L, Dang VN. Aggregating expert-elicited error probabilities to build HRA models. Safety and Reliability: Methodology and Applications: CRC Press; 2014. p. 1119-28.

[48] Cain J. 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; 2001.

[49] Podofillini L, Dang VN. A Bayesian approach to treat expert-elicited probabilities in human reliability analysis model construction. Reliability Engineering & System Safety. 2013;117:52-64.

[50] Wisse BW, van Gosliga SP, van Elst NP, Barros AI. Relieving the elicitation burden of Bayesian Belief Networks. BMA.

[51] Xiang Y, Jia N. Modeling causal reinforcement and undermining for efficient CPT elicitation. IEEE Transactions on Knowledge and Data Engineering. 2007;19:1708-18.

[52] Henrion M. Some Practical Issues in Constructing Belief Networks. UAI. p. 161-73.

[53] Lemmer JF, Gossink DE. Recursive noisy OR-a rule for estimating complex probabilistic interactions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2004;34:2252-61.

[54] Kuter U, Nau D, Gossink D, Lemmer JF. Interactive course-of-action planning using causal models. Third International Conference on Knowledge Systems for Coalition Operations (KSCO-2004)2004. p. 37-52.

[55] Myung IJ. Tutorial on maximum likelihood estimation. Journal of mathematical psychology. 2003;47:90-100.

[56] Stempfel Y, Dang VN. Developing and evaluating the Bayesian Belief Network as a human reliability model using artificial data. Advances in Safety, Reliability and Risk Manag. 2012.

[57] Yang ZL, Bonsall S, Wall A, Wang J, Usman M. A modified CREAM to human reliability quantification in marine engineering. Ocean Engineering. 2013;58:293-303.

[58] Moura R, Beer M, Patelli E, Lewis J. Learning from major accidents: Graphical representation and analysis of multi-attribute events to enhance risk communication. Safety Science. 2017;99:58-70.

[59] Groth KM, Smith CL, Swiler LP. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. Reliability Engineering & System Safety. 2014;128:32-40.

[60] Antonucci A, Brühlmann R, Piatti A, Zaffalon M. Credal networks for military identification problems. International Journal of Approximate Reasoning. 2009;50:666-79.

[61] Estrada-Lugo HD, de Angelis M, Patelli E. Probabilistic risk assessment of fire occurrence in residential buildings: Application to the Grenfell Tower. 2019.

[62] Estrada-Lugo HD, Santhosh TV, de Angelis M, Patelli E. Resilience assessment of safety-critical systems with credal networks. 2020.

[63] Walley P. Statistical reasoning with imprecise probabilities. 1991.

[64] Cano A, Gómez M, Moral S, Abellán J. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. International Journal of Approximate Reasoning. 2007;44:261-80.

[65] Antonucci A, De Campos CP, Huber D, Zaffalon M. Approximating credal network inferences by linear programming. European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty: Springer; 2013. p. 13-24.

[66] Patelli E, Tolo S, George-Williams H, Sadeghi J, Rocchetta R, de Angelis M, et al. OpenCossan 2.0: an efficient computational toolbox for risk, reliability and resilience analysis. 2018.

[67] Patelli E, Alvarez DA, 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 Conference2014. p. 1501.

[68] 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: IEEE. p. 1942-9.

[69] Troffaes MCM. Decision making under uncertainty using imprecise probabilities. International journal of approximate reasoning. 2007;45:17-29.

[70] Ferson S, O'Rawe J, Balch M. Computing with confidence: imprecise posteriors and predictive distributions. Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management2014. p. 895-904.

[71] Patelli E, Ghanem R, Higdon D, Owhadi H. COSSAN: a multidisciplinary software suite for uncertainty quantification and risk management. Handbook of uncertainty quantification. 2016:1-69.

[72] Vinnem JE. Operational safety of FPSOs: initial summary report: Great Britain, Health and Safety Executive; 2001.

[73] de Vos D, Duddy M, Bronneburg J. The problem of inert-gas venting on FPSOs and a straightforward solution. Offshore Technology Conference: Offshore Technology Conference; 2006.

[74] Alan Keith P, Aubrey Maurice T, Hans Stefan Ledin H, Safety E, Offshore D, Redgrave C, et al. Ignition Hazards and Area Classification of Hydrocarbon Cold Vents by the Offshore Oil and Gas Industry 2012.

[75] HSE U. Assessment of the adequacy of venting arrangements for cargo oil tanks on FPSO and FSU installations. 2010.

[76] ANP ANdP, Gás Natural e Biocombustíveis. Monthly bulletin with data on oil and gas production in Brazil, information on producing states, basins, fields and wells produced. 2020.

[77] ANP ANdP, Gás Natural e Biocombustíveis. Incident Data from Oil and Gas Exploration and Production 2020.

[78] 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 262016.

[79] 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 262016.

[80] HSE U. HSE Offshore Statistics, Offshore Hydrocarbon Releases 1992 – 2016 <br>. 2020.

[81] Kim Y, Park J, Jung W, Choi SY, Kim S. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. Reliability Engineering & System Safety. 2018;173:12-22.

[82] Swain AD, Guttmann HE. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. Sandia National Labs.; 1983.