Analysis and estimation of human errors from major accident investigation reports

Caroline Morais[a,b], Raphael Moura[a,b], Michael Beer[c,a,d] , Edoardo Patelli[a,*]

[a] Institute for Risk and Uncertainty, University of Liverpool, Chadwick Building, Peach Street, Liverpool L69 7ZF, United Kingdom
[b] National Agency for Petroleum, Natural Gas and Biofuels (ANP), Av. Rio Branco, 65, CEP: 20090-004, Centro, Rio de Janeiro, RJ, Brazil
[c] Institute for Risk and Reliability, Leibniz Universität Hannover, Callinstr. 34, 30167 Hannover, Germany
[d] Tongji University, Shanghai, China

*Contacting author: edoardo.patelli@liverpool.ac.uk

ABSTRACT

Risk analyses require proper consideration and quantification of the interaction between humans, organisation and technology in high-hazard industries. Quantitative Human Reliability Analysis approaches require the estimation of human error probabilities, often obtained from human performance data on different tasks in specific contexts (also known as performance shaping factors). Data on human errors are often collected from simulated scenarios, near-misses report systems, and experts with operational knowledge. However, these techniques usually miss the realistic context where human errors occur.

The present research proposes a realistic and innovative approach for estimating human error probabilities using data from major accident investigation reports. The approach is based on Bayesian Networks used to model the relationship between performance shaping factors and human errors.

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

1. INTRODUCTION

Despite the increasing level of automation and the advent of artificial intelligence [1], realistic risk assessments of high-hazard industries should ideally be performed through the analysis of the complex interaction between human, machine and organisational systems [2].

Human reliability analysis defines a collection of qualitative and quantitative methods used to account for human factors in social-complex industries in a systematic way [3]. Their main aims are to identify the possible human errors in a task (i.e. task analysis [4]), to quantify them (when needed) and to propose solutions to prevent or mitigate human errors [5]. The analysis uses the assumption that human errors are triggered by the interaction among individual, technological and organisational factors, the so-called performance-shaping factors.

Qualitative methods for human reliability provide only the identification of human errors and possible preventive or mitigation solutions. Quantitative human reliability methods provide the same functions as the qualitative methods, plus an estimation (or an adjustment) of the human error probabilities according to the defined performance shaping factors in a specific scenario. Different quantitative human reliability methods exist, including THERP [6], SPAR-H [7], HEART [8], CREAM [9] and ATHEANA [10]. These quantitative methods allow to find or adjust human error probabilities according to the performance shaping factors in the specific industrial context being assessed (organisational, technological and individual factors). However, human error probabilities obtained with quantitative methods are often affected by imprecision, sparse and/or incomplete human error data [11,12] leading to under-estimated or over-estimated probabilities [5]. This uncertainty may be one of the causes that are preventing industries from adopting risk assessments that account for human errors [13]. Although some safety regulators do accept qualitative analysis on human errors (e.g. [14]), human error probabilities are required by probabilistic safety (risk) assessments.

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

$$Human\ Error\ Probalitity = \frac{Number\ of\ observed\ errors}{Number\ of\ opportunities\ for\ error}$$

Equation 1

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

2. METHODOLOGY BACKGROUND

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

2.1. Data collection

Data collected from real operations are considered the most credible human error data, followed by data derived from real operations (i.e. incidents, near-misses and accidents), simulators and expert judgement (Figure 1) [5].

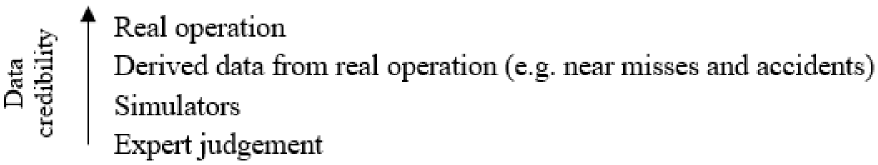

Data credibility
Real operation
Derived data from real operation (e.g. near misses and accidents)
Simulators
Expert judgement

Figure 1. Data credibility for Human Error Probability assessment (adapted from [5])

A summarised description of the strengths and pitfalls of each type of data are described below.

Expert judgement: Experts are individuals with recognised knowledge or skill in a specific domain. Sometimes expert elicitation is the only available data source [15] thus, their opinions are aggregated by adopting methods to reduce expert elicitation variability [16,17]. However, expert elicitation is considered the least credible source of data. This is because experts can be oriented by different sources of bias [15], be systematically overconfident about the accuracy of their judgements [18] and be experienced in the taxonomy used [5]. Ultimately, it is improbable to have a human reliability analysis that does not rely on expert judgement to some extent, as all methods start with a qualitative analysis of possible scenarios [19].

Simulators: Data from simulators are collected at mimicked control rooms or other workspaces where real operators perform specific tasks under normal or emergency scenarios. Data collected from simulators is often restricted to human-machine interfaces in control rooms. Often collected data needs to be calibrated by expert judgement adopting well kwown approaches, e.g. SACADA [20], HAMMLab [7, 21], HuREX [22], OPERA [23]. This approach is strong on detecting human errors, but weak on detecting all the performance shaping factors. This is due to the decontextualisation of the studied tasks [7], for instance operators know that their actions will not have any consequence and often know that their actions are being observed [5].

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

The direct task monitoring is the method where a real operational task is observed at the moment it is performed by an assessor or recorded and analysed after the event. It is considered one of the best data sources but it lacks data for tasks rarely performed. For instance, the database CORE-DATA has been partially generated with data derived from real operations [24].

Data from near-misses events are those that collect human errors and performance shaping factors from events that had the potential to cause considerable damage to assets and people but they had no relevant consequence [25-27]. This kind of data has the benefit of describing more errors related to hardware (such as manually operated valves) and relating human errors to performance shaping factors. However, near-miss reports are generally restricted to what needs to be communicated to the regulator [26, 27], thus relevant factors may not always be reported [28].

Data from major accident reports have the potential to deliver stronger relation between performance shaping factors and human errors [29,30]. This is because detailed analyses of the causes that led to the accidents are required and performed [31]. Despite the potential benefits, the strategy of using major accident data to estimate performance shaping factors and human error probabilities is not yet fully explored.

2.2. Bayesian networks

Bayesian network is a powerful graphical tool that has received an increasing interest due to their capability of providing efficient factorization of joint probability distributions, exploiting information about the conditional dependencies among variables [32]. Bayesian networks have also been used for the estimation of Human Error Probability on different industrial sectors, as described by the thorough review of Mkrtchyan et al. [33].

Let consider a simplified Bayesian network for modelling human error as shown in Figure 2. Each ellipse called ‘node’ represents variables such as ‘organisational factors’, ‘technological factors’, ‘person-related factors’, ‘cognitive errors’ and ‘execution errors’. The arrows represent the direction of the causal relationship between variables. In the model presented, the ‘organisational factors’ is defined as the parent node of ‘cognitive errors’ and, likewise, ‘cognitive errors’ as the child node of ‘organisational factors’. The ‘organisational factors’ is denoted a root node of the network, as it does not have parents. The causal relationships between variables is defined by Conditional Probability Distributions (CPDs). These distributions are usually represented by crisp values numerically coded in Conditional Probability Tables (CPTs) [34].

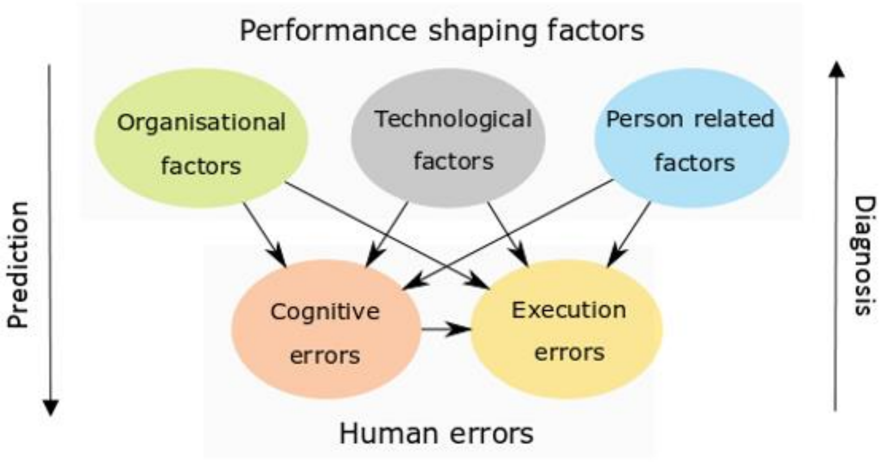


Figure 2. Simplified Bayesian network for human error probability

The main advantages of using Bayesian networks for Human Reliability Analysis are [33]:

- Deal with lack or incomplete data of human errors in complex industries by integrating expert judgement and other different sources of information in the model;
- Allow to consider dependencies among factors by using joint probabilities, to combat the frequent (and possibly mistaken) assumption of independencies between performance shaping factors and human errors.
- The acyclic graphs are easy to understand and potentially facilitate the communication between engineers, psychologists and social scientists in multi-disciplinary risk analysis.
- The possibility to update the marginal probabilities of the variables, when new information becomes available.

- Provide reasons for the results by allowing to identify which performance shaping factors are affecting individual human errors [35];
- The capability of performing "what if" scenarios analysis by fixing the state of variables, as well as to propagate the information in the direction of interest [36].

2.3. Identifying conditional dependencies from sparse data

Data for human error are usually sparse or missing. Although data can be collected from an increasing number and variability of accident reports (e.g. collecting reports from different safety regulators or from different industry sectors) some conditional dependencies might continuously fail to appear in the available data. Therefore, inferences of the human error probabilities are generally performed based on expert elicitation. Experts can contribute by providing direct probability values ('direct elicitation') or give their opinion through qualitative scales, questionnaires, relative judgements ('indirect elicitation') [33]. Alternative approaches are based on data derived from underling method relationships [37,38], or from specifically designed simulators [37, 39]. The discussion of the mathematical theory behind these approaches is beyond the scope of the present paper, however the interested reader can refer to [36, 40, 41]. Some basic background about conditional probability distributions (CPD) are provided in Appendix A.

2.4. Verification and Validation

Once the human error probabilities are obtained, they should be verified to test if the model works as it is supposed to work [33]. If the correct inputs are given, the appropriate outputs are seen [5]. In Jentsch words, we should ask ourselves: "Did we build the system right?" [42]. Verification can also be referred as 'internal validation', when used as a test to measure the variation between assessors, so the result can be repeated no matter the team or the day when the analysis is conducted [5].

Few published researches based on Bayesian network to infer human error probabilities have presented a verification process [33]. [43] have presented their verification results, after creating a set of hypothetical profiles at the extreme points, varying from the highest to the lowest level of each factor. [44] have conducted a sensitivity analysis focused on the 'context control modes' of the method CREAM, using expert judgement. They have suggested that in a successful model a slight change towards the negative effects of a 'context control mode' would result in the increment of the human error probability.

The literature suggests that higher levels of performance shaping factors would result in higher levels of human error probability, and that combinations of performance shaping factors would result on greater adverse impact on human error probability [3]. That means that Human Reliability should reflect the features of a Coherent System with multi-states components, where the performance of a system improves whenever any component or subset of component improves, and vice-versa [45,46].

To validate a model, one should test if the system does what is supposed to do in the real world: if the outputs have a good correlation to 'real world data' [5]. In Jentsch words, we should ask ourselves: "Did we built the right system?" [42].

A common method to validate a model is to conduct cross-validation, splitting available data sets into training and test sets. However, this approach is adopted in data-rich applications which is not the case presented in rare events such as human errors in major accidents [33]. For these events, Kirwan suggests the comparison of the new results with existing human error data of better or similar credibility level [5]. The measurable criteria used are correlation, accuracy, the degree of optimism/pessimism, and precision [5].

(i) Correlation: The degree of the predictive relationship is usually presented via a scatterplot of predicted versus actual human error probability. Although validations usually try to express parametric correlation (with the square of the correlation coefficient), the majority of validation

research conducted by the human reliability community have been expressed via non-parametric correlation [5,8,47], assuming that human behaviour does not rely on any assumption of the distribution function or the joint distribution of performance shaping factors.

The non-parametric correlation tests are Spearman's rank correlation coefficient [48] and Kendal's coefficient of concordance (Kendal's τ) [49]. Although both tests are different, the interpretation of both coefficients are similar: the correlation between the two variables will be high when observations have a similar (or equal) correlation of one. Likewise, if the coefficient value is next to zero, the correlation between the results from the model and the reference is small.

(ii) Accuracy: In risk assessment, an ideal accuracy level is when estimates lie within a factor of three of the 'true' values, but it is acceptable if falls within a factor of ten [5]. Model accuracies are often represented graphically in a scatterplot of the results against reference data using logarithms scale.

(iii) Precision: An aspect of precision is the degree to which the technique, when not accurate, is pessimistic rather than optimistic [5]. Pessimistic estimate is a prediction that goes into a more conservative direction. Conservative estimates lead to safer but at the same time more expensive design. Therefore, it is important to find strategies that provide more realistic HEPs to the industry. Histograms are also plotted to present how human error estimates were distributed into accuracy bands within pessimistic and optimistic factors of 3, 10 and 100.

3. PROPOSED APPROACH: USING DATASETS OF MAJOR ACCIDENTS REPORTS

3.1. Bayesian Network definition

All the steps required to build a Bayesian network from major accident reports are described below.

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

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

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

Definition of the structure: The Bayesian network structure (Figure 3) has the objective of capturing the dependencies between performance shaping factors and human errors, but also among performance shaping factors and human errors, and explicitly modelling their multi-level, hierarchical influences on each other.

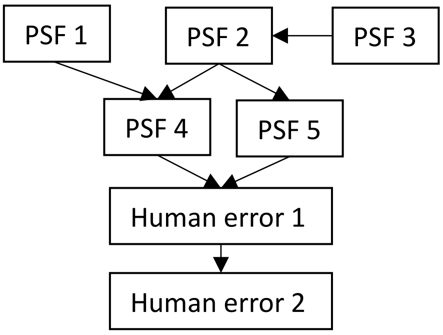


Figure 3. Example of a structure reflecting the causal relationships within variables

Experts with psychology and sociology knowledge might be elicited to obtain this type of structure (e.g. to identify the causal relationships of cognitive errors and organisational factors). Although one of the aims of this research was to avoid expert biases, it is acknowledgeable that at some level of the assessment the experts are essential – if not for eliciting the prior probabilities, they will be for the model structure or for the taxonomy used.

3.2. Assessment of the Conditional Probability Tables

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

Table 1. Example of a dataset with human errors and performance shaping factors (PSFs) identified for each accident.

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

Table 2 (a,b). Example of prior probabilities of root nodes PSF 1 and PSF 2.

PSF 1	
State 0	1
State 1	0

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

The conditional probabilities for child nodes depend on the structure defined on the network and are obtained by counting the frequency of all the possible combinations of the parent node's states in the dataset. The frequency is the number of accidents that present a specific combination divided by the number of accidents in the dataset. The frequencies obtained (Table 3) are then normalised, as the prior probabilities of the set of states of the child node must sum to one (Table 4). The same process is repeated for each combination of the conditional probability table. When this process is complete it is possible to compute the posterior probabilities for each node. The posterior probabilities of the state '1' of the child nodes designated to human errors will be the human error probabilities.

Table 3. Example of the conditional probability table for node 'Human error 1'

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

Table 4. Normalised conditional probability table

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

When the dataset used does not provide information for defining conditional distributions within certain variables states, the variable state "no data" is set to '1'. If this strategy were not used, the prior probabilities of states '0' and state '1' of the child node for that given combination would have both probabilities set equal to zero, making it impossible to compute the conditional probability table. In Ref [41] it is suggested to assigning equal probability to all the unknown combination of events. However, using the latter strategy, a researcher loses the information of what combinations do not lead to human errors according to the dataset, which can be potentially used in the future.

3.3. Validation and verification

The verification of the models is performed through what-if analysis, to test how the model behaved when analysing well-known scenarios [36]. To achieve that, some hypothetical scenarios have been created by fixing each state of each performance shaping factor node of the Bayesian network, and observing how the changes affected the human error probabilities.

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

4. CASE STUDY

4.1. MATA-D dataset

For the present research, the MATA-D dataset is adopted [29]. The dataset contains 238 major accident reports classified under the CREAM taxonomy [9]. A single taxonomy is used to describe both human errors and performance shaping factors for a variety of industrial sectors. Only trusted investigation boards have been used to build the dataset. Logical values, i.e. binary code of 1s or 0s, are used to designate whether or not a human error or factor was observed. This resulted in a matrix of zeros and ones with 238 rows (the number of accidents) by 53 columns formed by 39 performance shaping factors (Table 5) and 14 different type of human errors (Table 6).

Table 5. Performance shaping factors used in MATA-D dataset

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

Table 6. Human errors used in the MATA-D dataset

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

4.2. Bayesian Network

The methodology presented in Section 3 has been used to construct a Bayesian Network model from the MATA-D dataset and summarised in Table 7. The resulting structure of the Bayesian network is shown in Figure 4.

Table 7. Summary of the methodology to build the Bayesian Network model

Nodes and states	Structure	Conditional probability table	Verification and Validation
The nodes are variables defined in CREAM taxonomy [9]. From 39 possible performance shaping factors and 14 possible human errors, only six were not used, due to their absence on the accident reports. The root nodes have two states: '0' and '1' (following the logical entries of the MATA-D dataset) and child nodes have '0, '1' and 'no data'. The root nodes have two states: '0' and '1', to designate whether or not an evidence was encountered on an accident report. The child nodes have the states '0, '1' and 'no data'. The latter state is used when the dataset does not provide a specific combination between the parent nodes.	The connections between the nodes were defined according to relations based on expert judgement, from the same author of the taxonomy used to define the nodes [9]. He has named it the 'antecedent-consequence relation'. A different structure less reliant on expert judgement was proposed at [50], by using common patterns of PSFs and human errors identified on [51]. The structure depicts the influence between performance shaping factors, which means that some performance shaping factors are also child nodes. The structure represents the influence that performance shaping factors have upon each other. Eventually, this means that some performance shaping factors are also child nodes. All human errors are child nodes of the performance shaping factors.	The conditional probability tables for the root nodes were obtained directly from the frequencies of each performance shaping factor according to [29], e.g. design failure is equal to 66%, so at the conditional probability table the state '1' of the root node 'design failure' is 0.66 and the state '0' is the complement to one: 0.34. The frequency for combinations between factors and errors for the child nodes have been extracted from the dataset entries. Due to the high number of combinations between the states of the parent nodes that a child node has reached, obtaining the frequencies per combination from the dataset was not a trivial task. A code was used to read the table and extract the probability for each combination. For more information on the code and on how to use it, please contact the authors. The conditional probability tables for the root nodes are obtained directly from the frequencies of each performance shaping factor according to the dataset. The frequency for combinations between factors and errors are obtained also from the dataset inputs for each accident.	To verify any incoherence in the model, a what-if analysis was conducted by fixing the states of the variables. To validate the model, the estimates were tested against reference data published on [9] according to correlation, accuracy and precision. To verify any incoherence in the model, a what-if analysis is conducted by fixing the states of the variables. To validate the model, the estimates are tested against reference data according to correlation, accuracy and precision. If possible, the reference data should be obtained from operational experience.

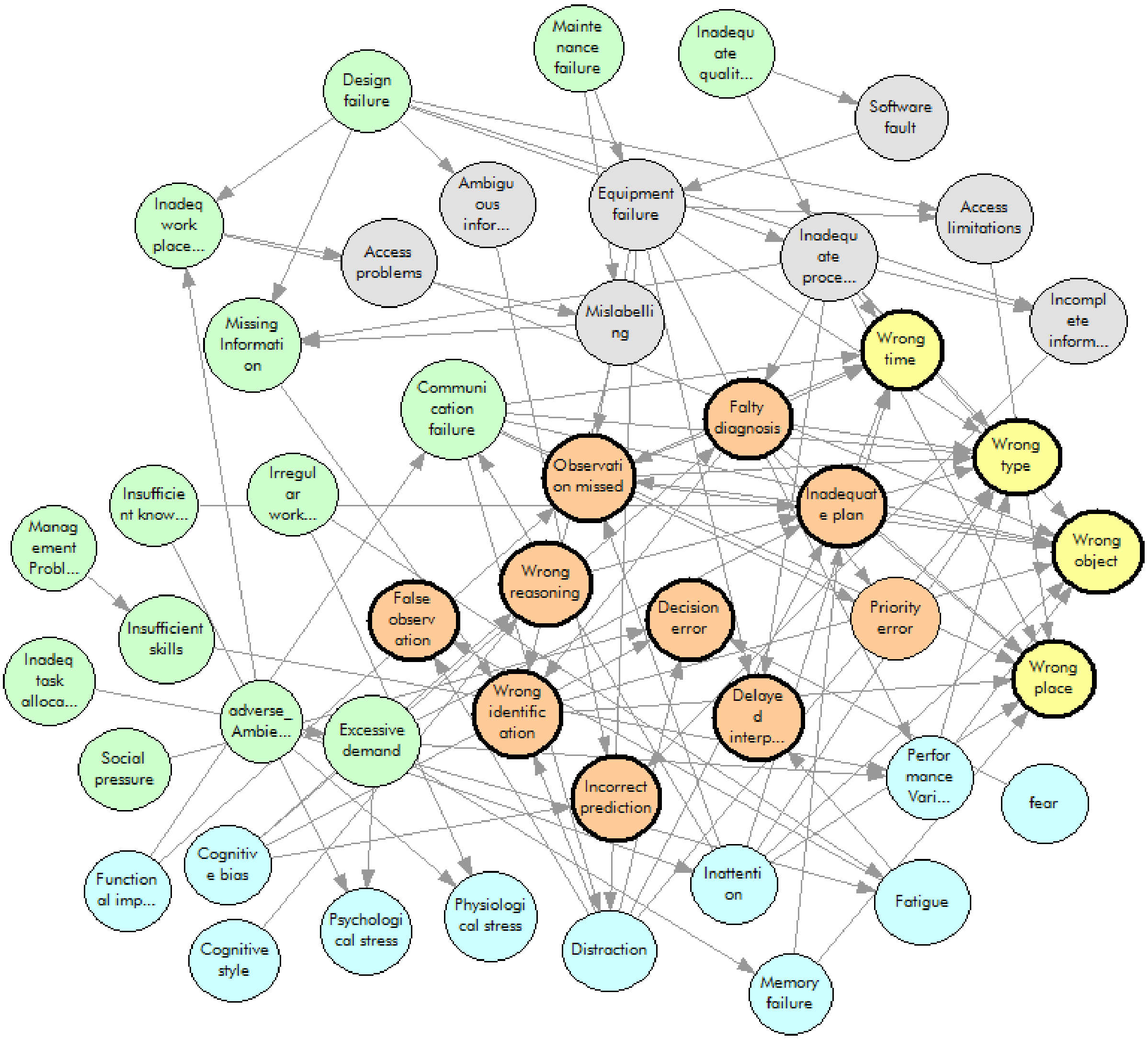


Figure 4. Model for predicting human error probabilities

4.3. Human Error Probabilities

The Human Error Probabilities (HEP) obtained analysing the MATA-D dataset are presented in Table 8 and graphically represented in a scatter plot in Figure 5. The state '0' indicates the probability of a specific human error not being triggered by a specific combination of performance shaping factors. The state 'no data' indicates the number of times a combination of those factors has not occurred in the dataset.

For the purpose of verification, the obtained probabilities have been compared against data from Ref. [9]. The interval of the reference is described by the lower and upper bounds for each human error.

Table 8. Human error probabilities from model compared with data reference [9]

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

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

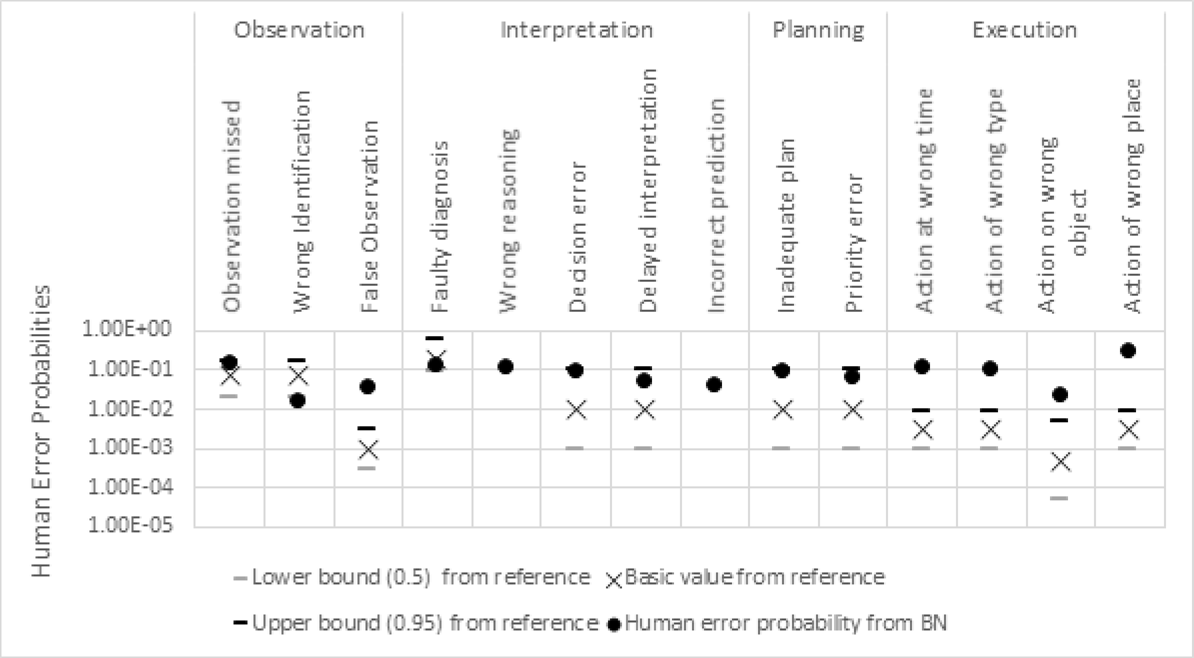


Figure 5. Human error probabilities from the proposed approach and from reference [9] plotted in a logarithmic scale

Figure 5 shows higher human error probabilities than the reference data. A possible interpretation of this trend might be attributed to the methods used to collect reference data [9], where all human errors were accounted for, including those that have not produced an accident. Thus more opportunities of errors were accounted on the denominator of Equation 1, making the resulting probabilities lower than those obtained with the present approach.

The human error estimates are the values obtained for the probabilities of the state ‘1’ of each child node. The results of state ‘0’ and the state ‘no data’ are presented in Table 9. A comparison of the results obtained for each state is also presented in Figure 6.

Table 9. Results of all states of human error probability nodes on the proposed model

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

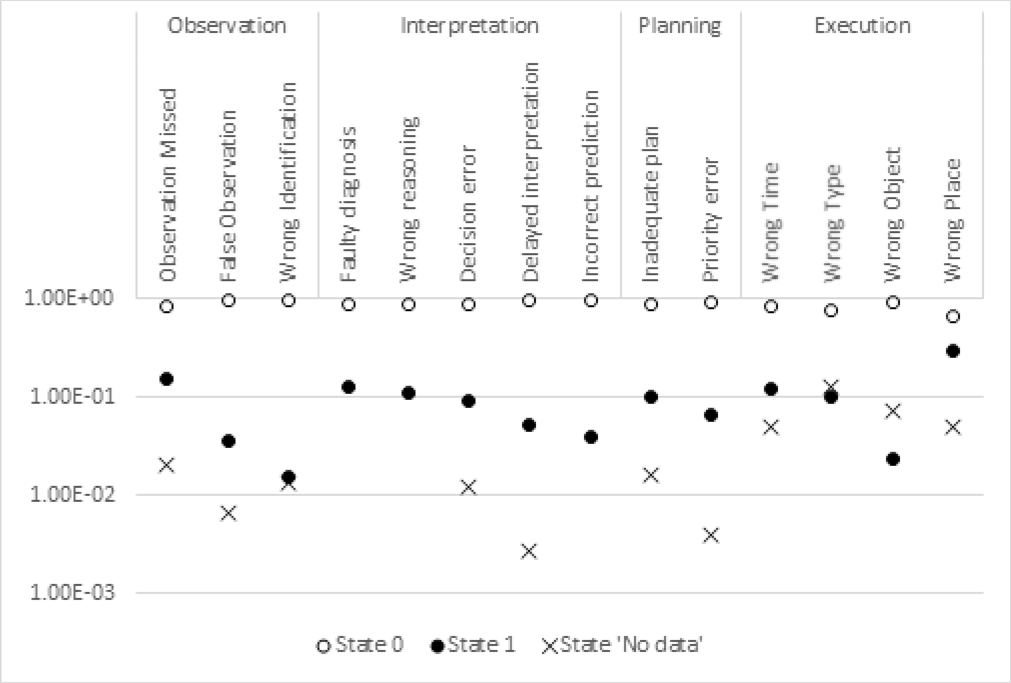
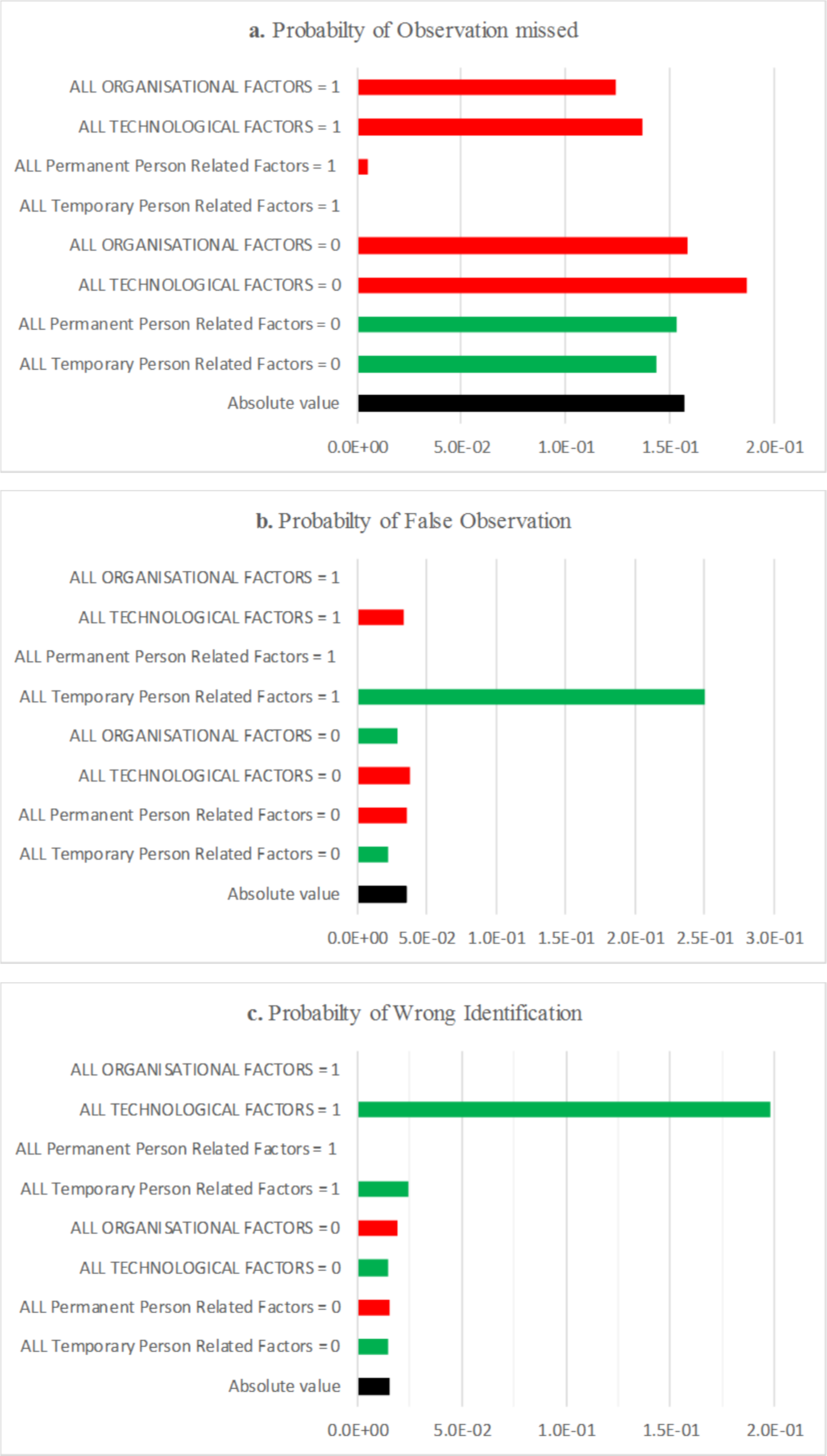


Figure 6. States estimates for the proposed model

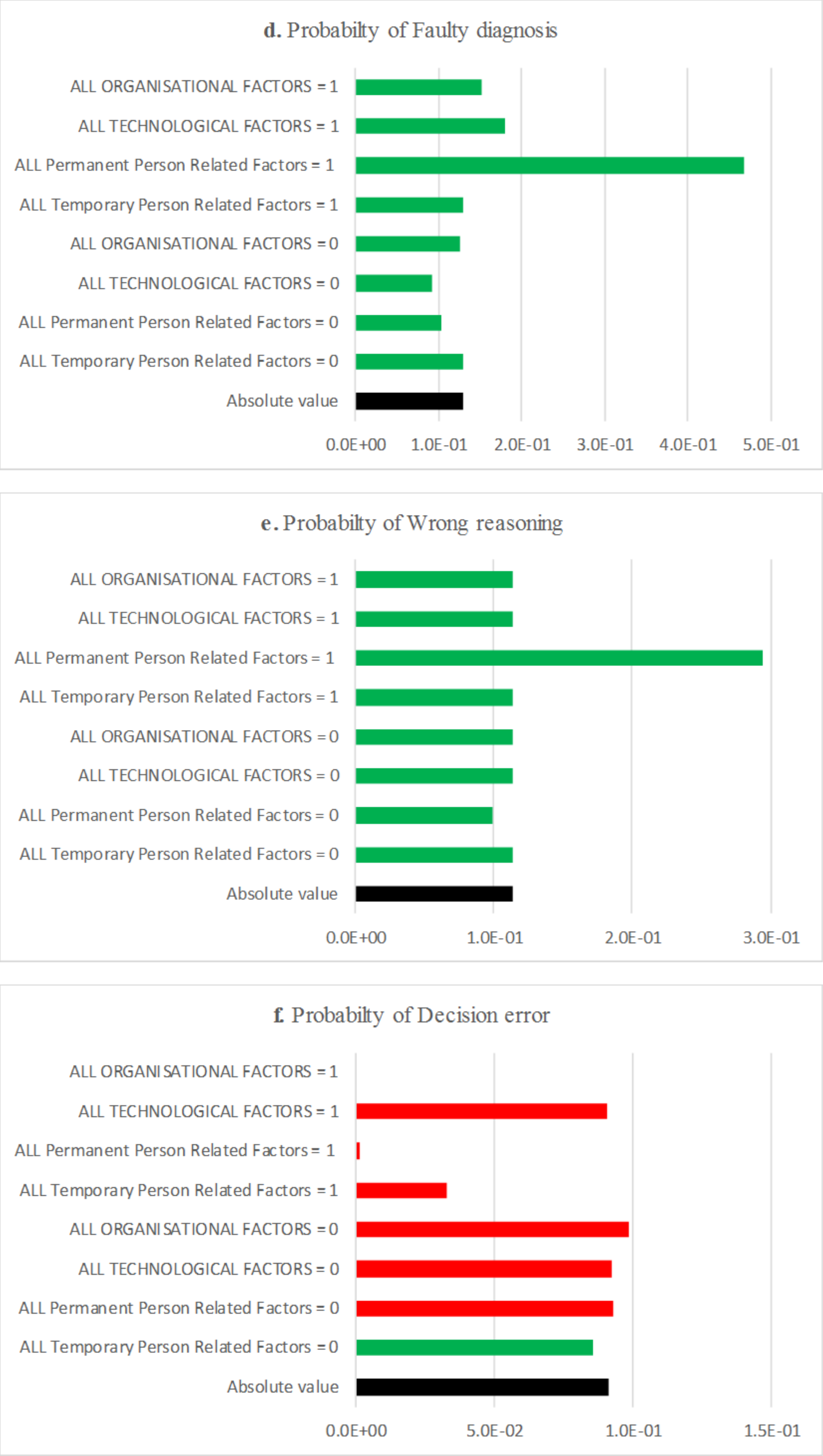
To test if the model outputs work as they were supposed to work, a what-if analysis was conducted, by fixing the states of sets of performance shaping factors one-at-the-time and summarised graphically in Figure 7. In Figure 7, the black bars in the charts represent the values estimated for the model; the green and red bars can be interpreted as a spectrum of human error probabilities after varying all performance shaping factors to their best and worst-case scenarios. The green bars represent the expected results for a specific variation, whereas the red bars represent the unexpected results. The expected results represent those values that are expected from a coherent system according to the formal definition used for reliability technological systems. For instance, in a coherent system, the probability of having a human error decreases if a set of performance shaping factors are set to zero (best-case scenario), and increases in case of performance shaping factors increased to 100% (worst-case scenario). The obtained figures show that in the scenario of having all the organisational factors failing to work, the cognitive error of missing an observation (i.e. 'Observation missed') would in fact decrease. This is possibly be explained by an increase in performance that humans might be using to compensate organisational errors. This reinforces the theory that humans are not only probable initiators of an event, but also the last chance to recover a problem initiated by organisational and technological factors [9].

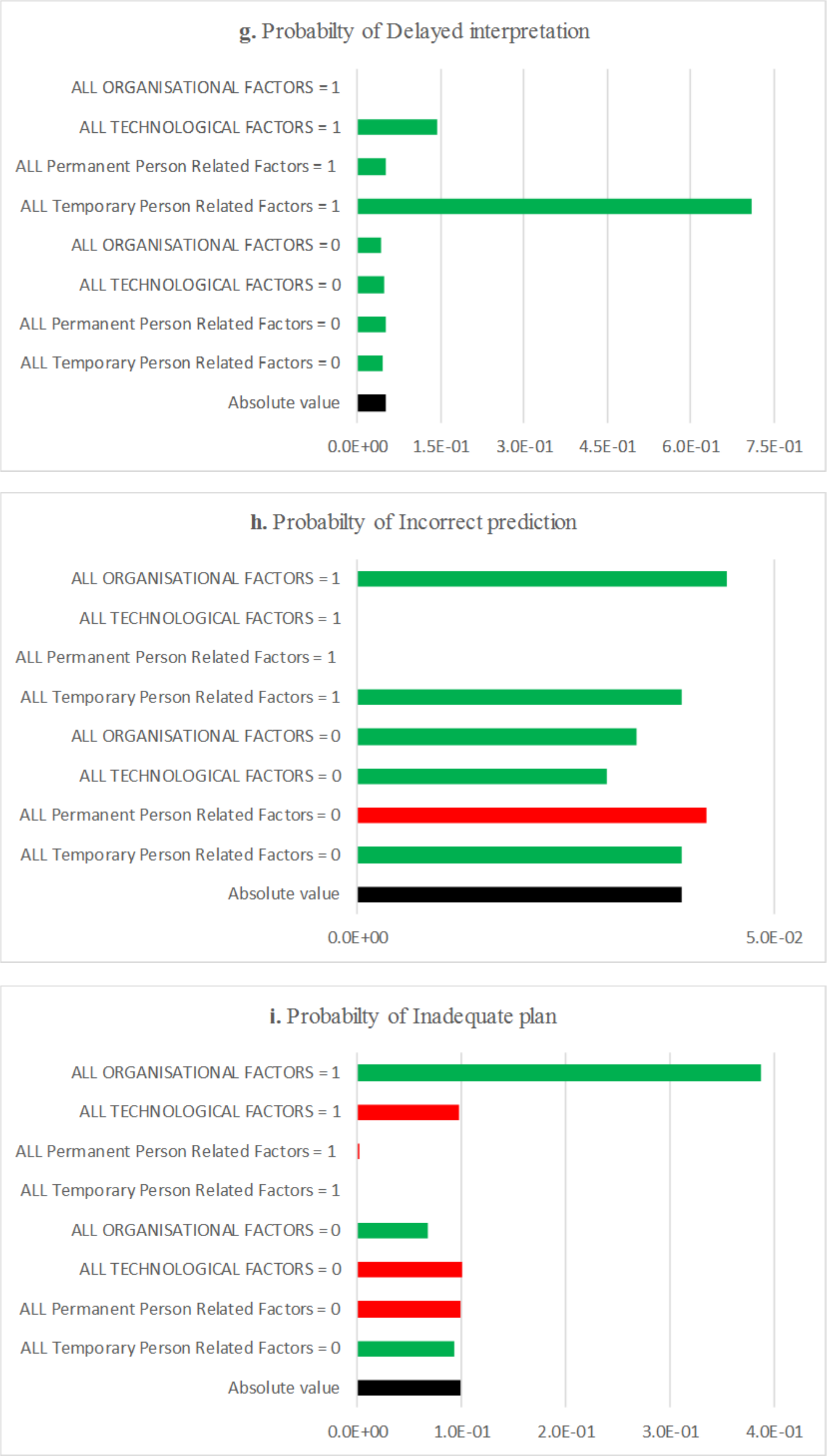


a. Probabilty of Observation missed

b. Probabilty of False Observation

c. Probabilty of Wrong Identification

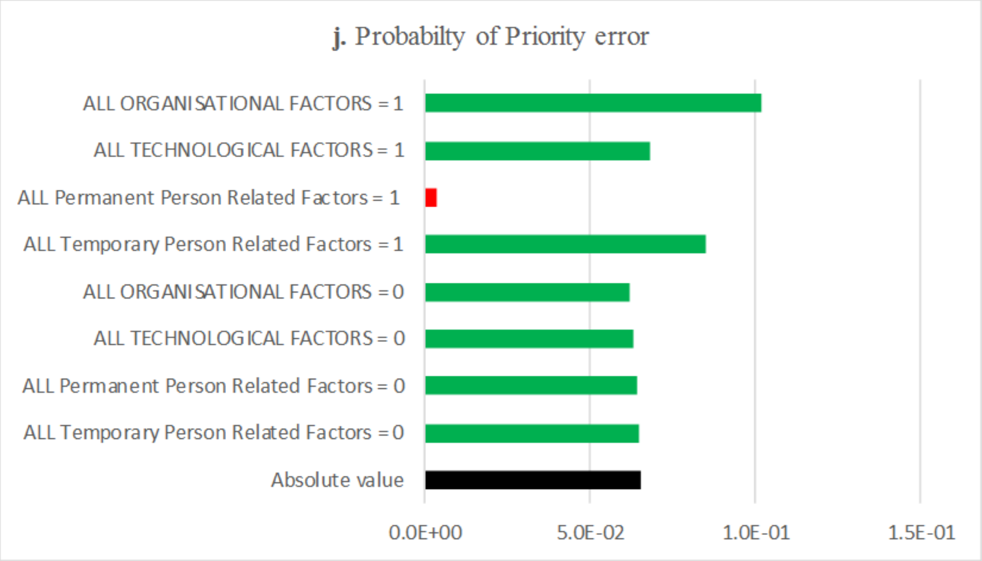

d. Probabilty of Faulty diagnosis
ALL ORGANISATIONAL FACTORS = 1
ALL TECHNOLOGICAL FACTORS = 1
ALL Permanent Person Related Factors = 1
ALL Temporary Person Related Factors = 1
ALL ORGANISATIONAL FACTORS = 0
ALL TECHNOLOGICAL FACTORS = 0
ALL Permanent Person Related Factors = 0
ALL Temporary Person Related Factors = 0
Absolute value
0.0E+00
1.0E-01
2.0E-01
3.0E-01
4.0E-01
5.0E-01
e. Probabilty of Wrong reasoning
ALL ORGANISATIONAL FACTORS = 1
ALL TECHNOLOGICAL FACTORS = 1
ALL Permanent Person Related Factors = 1
ALL Temporary Person Related Factors = 1
ALL ORGANISATIONAL FACTORS = 0
ALL TECHNOLOGICAL FACTORS = 0
ALL Permanent Person Related Factors = 0
ALL Temporary Person Related Factors = 0
Absolute value
0.0E+00
1.0E-01
2.0E-01
3.0E-01
f. Probabilty of Decision error
ALL ORGANISATIONAL FACTORS = 1
ALL TECHNOLOGICAL FACTORS = 1
ALL Permanent Person Related Factors = 1
ALL Temporary Person Related Factors = 1
ALL ORGANISATIONAL FACTORS = 0
ALL TECHNOLOGICAL FACTORS = 0
ALL Permanent Person Related Factors = 0
ALL Temporary Person Related Factors = 0
Absolute value
0.0E+00
5.0E-02
1.0E-01
1.5E-01

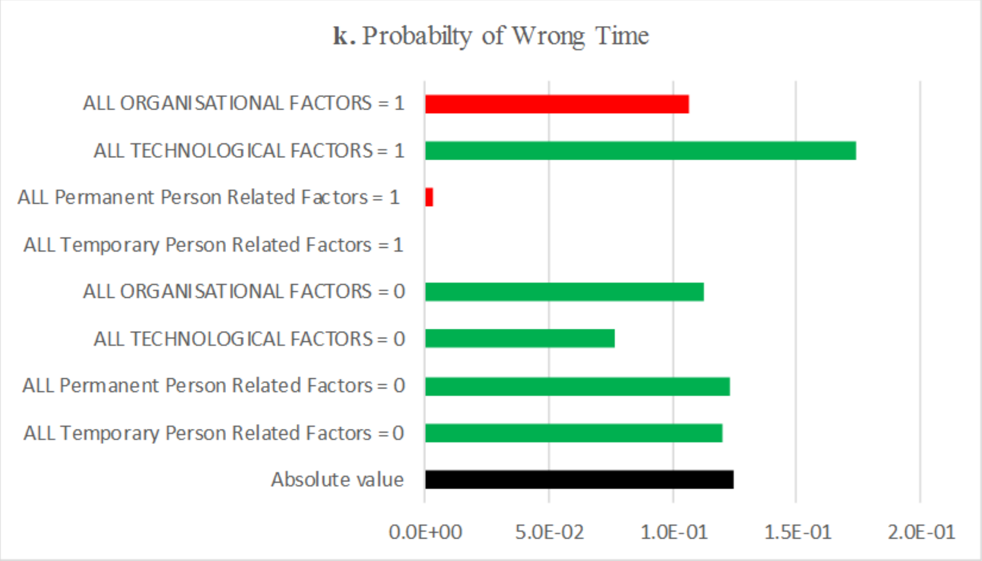


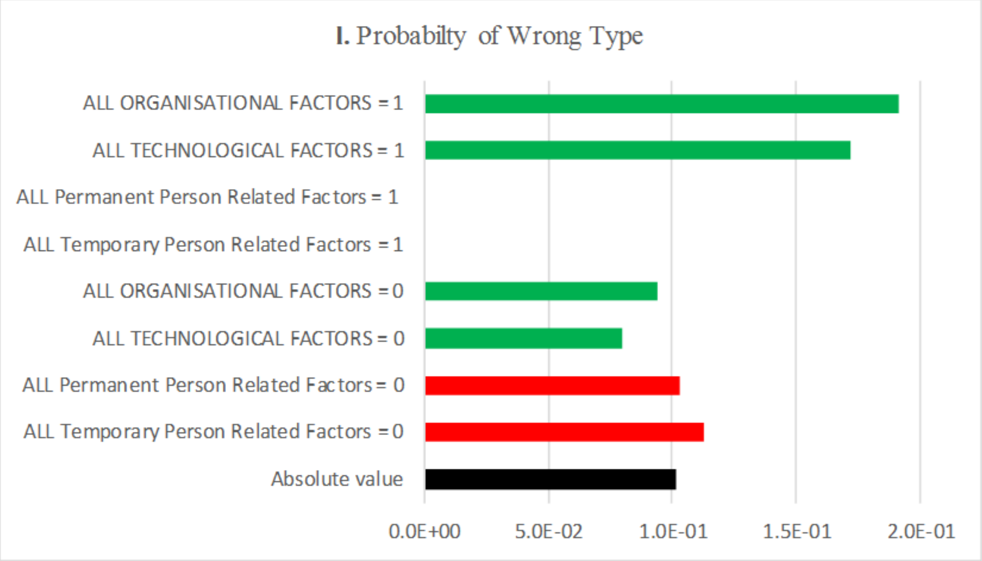
g. Probabilty of Delayed interpretation

h. Probabilty of Incorrect prediction

i. Probabilty of Inadequate plan


j. Probabilty of Priority error
ALL ORGANISATIONAL FACTORS = 1
ALL TECHNOLOGICAL FACTORS = 1
ALL Permanent Person Related Factors = 1
ALL Temporary Person Related Factors = 1
ALL ORGANISATIONAL FACTORS = 0
ALL TECHNOLOGICAL FACTORS = 0
ALL Permanent Person Related Factors = 0
ALL Temporary Person Related Factors = 0
Absolute value
0.0E+00
5.0E-02
1.0E-01
1.5E-01


k. Probabilty of Wrong Time
ALL ORGANISATIONAL FACTORS = 1
ALL TECHNOLOGICAL FACTORS = 1
ALL Permanent Person Related Factors = 1
ALL Temporary Person Related Factors = 1
ALL ORGANISATIONAL FACTORS = 0
ALL TECHNOLOGICAL FACTORS = 0
ALL Permanent Person Related Factors = 0
ALL Temporary Person Related Factors = 0
Absolute value
0.0E+00
5.0E-02
1.0E-01
1.5E-01
2.0E-01


l. Probabilty of Wrong Type
ALL ORGANISATIONAL FACTORS = 1
ALL TECHNOLOGICAL FACTORS = 1
ALL Permanent Person Related Factors = 1
ALL Temporary Person Related Factors = 1
ALL ORGANISATIONAL FACTORS = 0
ALL TECHNOLOGICAL FACTORS = 0
ALL Permanent Person Related Factors = 0
ALL Temporary Person Related Factors = 0
Absolute value
0.0E+00
5.0E-02
1.0E-01
1.5E-01
2.0E-01

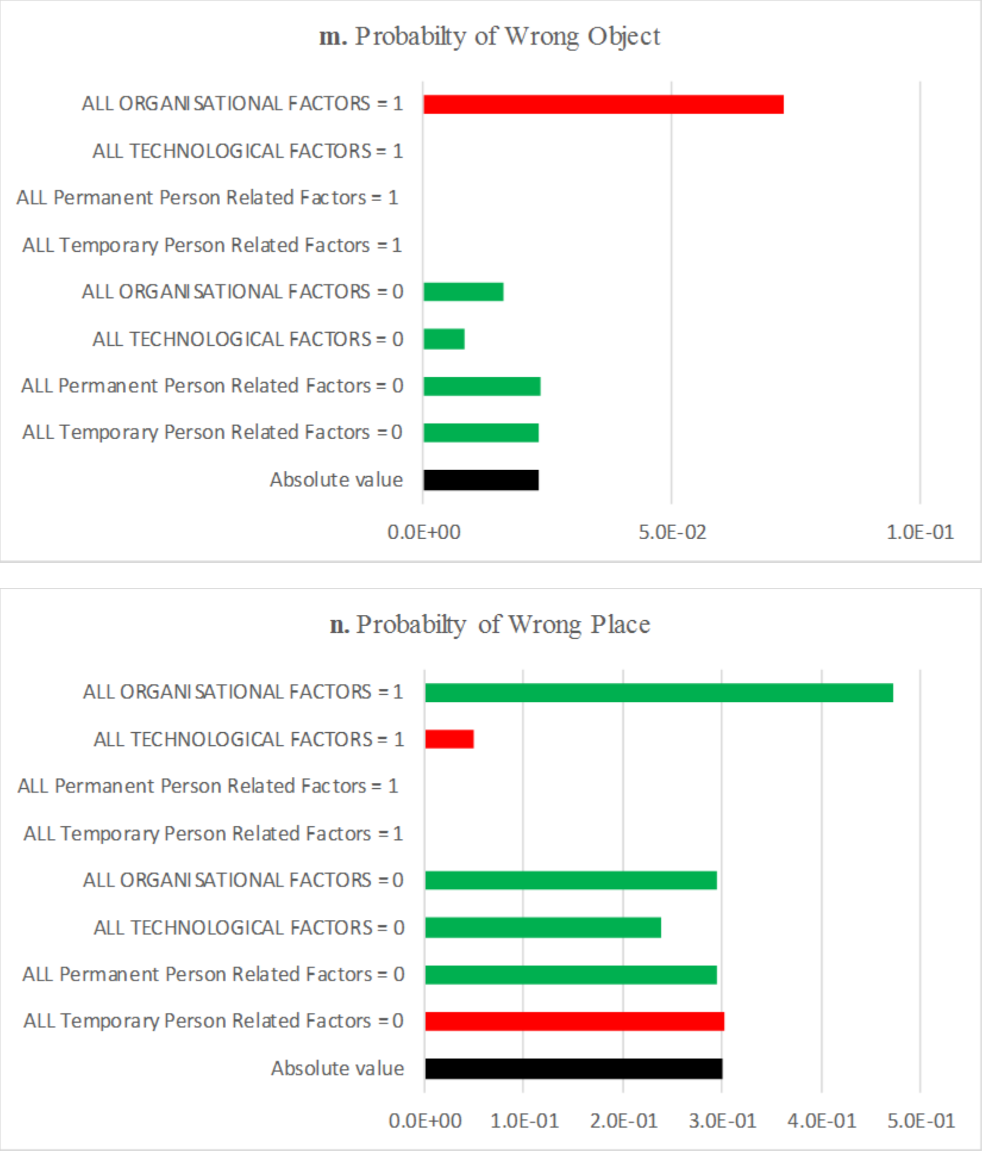


Figure 7 (a-n). Variation of the sets of PSFs for each HEP estimates

Variations of some sets of performance shaping factors had also resulted in zero probability human errors, as presented in Table 10. This shows that some human errors are impossible to occur under the specific conditions of performance shaping factors present in the MATA-D database.

Table 10. Sets of performance shaping factors variations producing zero human errors probability.

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

To validate the model, its outputs had been tested on the correlation, accuracy and precision to existing data obtained at [9]. The reference data were collected from simulators, expert judgement, laboratory controlled cognitive experiments and simulation studies of inspection tasks, (from simulated process plant and training schools). According to Hollnagel [9], data sources for human errors such as observation and execution were relatively well established at the time they were collected (approximately 1998). On the other hand, the author declared that interpretation and planning behaviours were mostly based on expert judgements. In addition, Ref [9] does not provide probabilities for 'wrong reasoning' and 'incorrect prediction'. To validate the model only the basic values provided in [9] are used.

Figure 8 shows a scatter plot of human error probability predicted from the model versus human error probability from the reference [9]. The present research has also tested non-parametric correlation, as human behaviour does not rely on any assumptions on the distribution function. The non-parametric correlation tests of Spearman's correlation coefficient and Kendal's coefficient of concordance are both presented in

Table 11. Both correlation coefficients are very small and not statistically significant. As shown on the scatterplot in Figure 8, seven of the human error probabilities estimated lied within a factor of 10 and five within a factor of 100 of the reference. To evaluate their accuracy within a factor of 3, the results were also plotted in a histogram (Figure 9). When not accurate, the histograms also illustrate if the estimates are pessimistic or optimistic if compared to the reference.

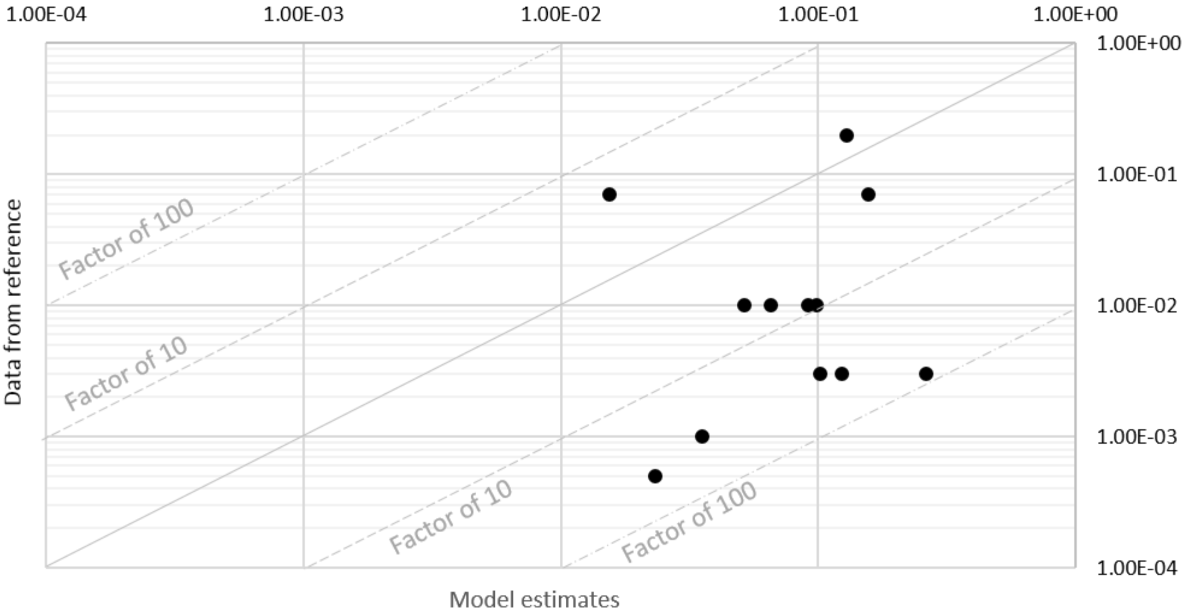


Figure 8. Human error probabilities (HEPs) from model versus HEPs from the reference in a logarithmic scale

Table 11. Non-parametric correlation results

Correlation between model outputs and values in the reference	Spearman’s correlation coefficient ($\square_s$)= 0.20115
	Kendal's coefficient of concordance (Kendal's □) = 0.3333

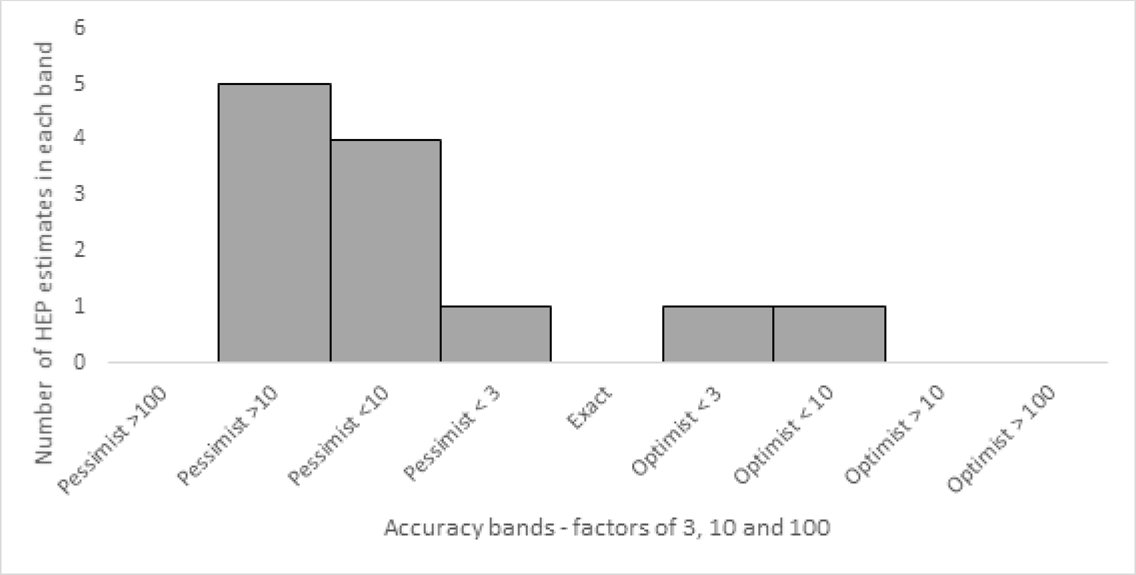


Figure 9. Histogram with accuracy bands

The model outputs had presented more pessimistic estimates rather than optimistic ones, meaning that the majority of HEPs estimated through both models tend to be higher than the reference. The histogram provided in Figure 9 shows how spread the results are. Table 12 presents the lower and upper bounds of human error probabilities after varying all performance shaping factors to their minimum and maximum values.

Table 12. Human error probability uncertainty after varying performance shaping factors

Cognitive and execution errors	Lower bound	Human error probability	Upper bound
Observation missed	5.30×10^{-3}	1.57×10^{-1}	7.75×10^{-1}
False Observation	1.00×10^{-3}	3.54×10^{-2}	3.27×10^{-1}
Wrong Identification	5.00×10^{-4}	1.54×10^{-2}	1.98×10^{-1}
Faulty diagnosis	9.15×10^{-2}	1.30×10^{-1}	4.69×10^{-1}
Wrong reasoning	9.95×10^{-2}	1.13×10^{-1}	2.94×10^{-1}
Decision error	1.40×10^{-3}	9.14×10^{-2}	2.72×10^{-1}
Delayed interpretation	2.10×10^{-2}	5.19×10^{-2}	7.10×10^{-1}
Incorrect prediction	2.30×10^{-3}	3.90×10^{-2}	8.49×10^{-2}

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

4.4. Discussion

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

The verification applied to the case study shows some human errors increasing if one or a set of performance shaping factor are decreased (and vice-versa). This may suggest and inadequacy of the used model or may also indicate that complex socio-technical systems do not necessarily behave as a coherent system. If right, the results of the case study suggest that some degraded performance shaping factors (or the combination of them) may cause also positive effects on human behaviour. Similar behaviour has been described by psychology research, which described that vigilance (the ability to maintain concentrated attention over prolonged periods of time) can actually decrease due to low levels of workload, an organisational shaping factor [52]. The verification step also has demonstrated that some human errors are unlikely to happen for specific states of performance shaping factors, as can be observed from some null human error probabilities.

The validation step has exposed a low correlation between the results obtained with the Bayesian network and the reference, as the model do not provide a predictive relationship with data from the reference used [9]. However, a new validation process must be conducted with data with similar source quality as the dataset (i.e. operational experience), as the data used as reference was partially obtained from simulators and expert elicitation.

The human error probabilities obtained from the model tend to be higher than the reference, meaning that if they are used to feed risk assessments they will lead to a safer design. The majority of results falls within a factor of 3 and 10 than within a factor of 100, which is normally acceptable to feed risk assessments. This validation aspect is important to develop because although higher than the real probabilities lead to safer design, they are not desirable as it can direct resources to the wrong risks.

The what-if analysis undergone in the verification and validation steps has also provided a spectrum of human error probability variations that can be seen as the uncertainty of estimates from different scenarios. In other words, varying the performance shaping factors in the Bayesian networks provides a distribution of human error probabilities, where uncertainty boundaries can be obtained.

To better capture the uncertainty associated with the dataset, two aspects of the data collection are suggested for future research. First, the data collection should be conducted by at least three experts, to improve the quality of the measure [17]. Second, the number of publicly available reports should be increased, allowing more experts to improve and test the dataset.

5. CONCLUSIONS

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

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

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

Verification and validation steps are provided to assess the accuracy of the estimated human error probabilities and the uncertainty related to the model or dataset used.

The approach presented in this paper have the potential to minimise the use of human reliability analysis methods to quantify and calibrate human error probabilities, thus minimising the need of expert elicitation – leaving for them the important mission of identifying critical tasks and the possible types of human errors associated, discussing possible controls and developing mitigation actions.

Acknowledgements

This research has been supported by ANP, the Brazilian Oil & Gas Regulator, and by EPSRC under the grant EP/R020558/1

6. Nomenclature

ATHEANA – Technique for Human Error Analysis
CORE-DATA - Computerised Operator Reliability and Error Database
CREAM – Cognitive Reliability and Error Analysis Method
HAMLAB - HAlden Man-Machine LABoratory
HEART – Human Error Assessment & Reduction Technique
HuREX - Human Reliability data Extraction
SACADA - Scenario Authoring, Characterization, and Debrief Application

Appendix

BNs can be represented by acyclic graphs, where nodes are connected to each other by arcs expressing dependencies among variables. The arcs directions must be coherent with the causal relationship of the connected variables. In the BN represented in Figure 10, the nodes A and B are called parent nodes of C, which is referred to as their child node. A and B are also called root nodes, as they do not have parents [36].

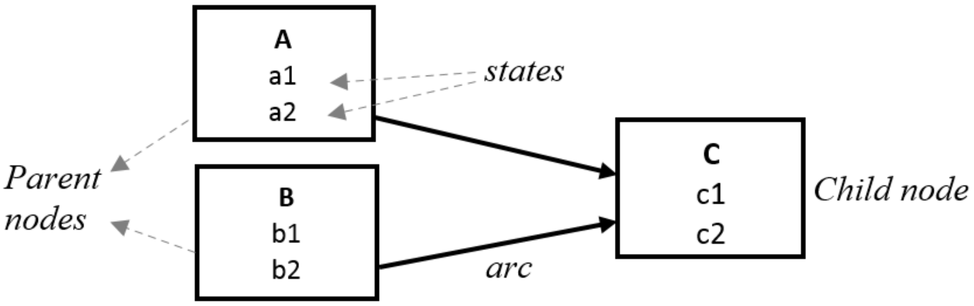


Figure 10. Directed acyclic graph of a Bayesian network

Figure 10 is the graphic representation of the conditional probability expressed in Equation 2 and Equation 3.

P(C=c1 | A=a1,B=b1) Equation 2

P(C=c2 | A=a1,B=b1) = 1-P(C=c1 | A=a1,B=b1) Equation 3

The Bayes' theorem expressed in Equation 3 provides the mathematical background for joint probabilities modelled by a generic BN with nodes X1, X2,…,Xn, where $p_i$refers to the outcomes assumed by the parents of the node Xi, which state is represented by x_i. The joint probability associated with this generic BN is represented by Equation 4.

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

Equation 4

If all nodes have a binary state, the number of combinations to consider in order to generate a child’s node conditional probability is two (a pair of combinations) to the power of the number of states of the parent nodes ($2^{\text{states of the parent nodes}}$). These possible combinations are usually organised in conditional probability tables, as the one represented in Table 13.

Table 13. Example of Conditional Probability Table for the BN of Figure 10

A	State 1		State 2	
B	State 1	State 2	State 1	State 2
State 1 of C	P(C=c1 | A=a1,B=b1)	P(C=c1 | A=a1,B=b2)	P(C=c1 | A=a2,B=b1)	P(C=c1 | A=a2,B=b2)

State 2 of C	P(C=c2 | A=a1,B=b1)
	Or
	1-P(C=c1 | A=a1,B=b1)
		P(C=c2 | A=a1,B=b2)
		Or
		1- P(C=c1 | A=a1,B=b2)
			P(C=c2 | A=a2,B=b1)
			Or
			1- P(C=c1 | A=a2,B=b1)
				P(C=c2 | A=a2,B=b2)
				Or
				1-P(C=c1 | A=a2,B=b2)

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

The inference computation in BNs can be obtained through some software packages, which allow the adoption of several algorithms, whether exact or approximate [53-55]. Those algorithms and modelling techniques are used as a starting basis and supporting tool for our development, which extrapolates towards an enhanced approach with novel features.

REFERENCES

[1] Ramos, Marilia Abilio, Ingrid Bouwer Utne, Jan Erik Vinnem, and Ali Mosleh. "Accounting for human failure in autonomous ship operations." *Safety and Reliability–Safe Societies in a Changing World. Proceedings of ESREL 2018, June 17-21, 2018, Trondheim, Norway* (2018).

[2] Zio, Enrico. "The future of risk assessment." *Reliability Engineering & System Safety* 177 (2018): 176-190.

[3] Henderson, J., and D. Embrey. "Guidance on quantified human reliability analysis." *Energy Institute, London* (2012).

[4] Kirwan, Barry, and Les K. Ainsworth. *A guide to task analysis: the task analysis working group*. CRC press, 1992.

[5] Kirwan, Barry. "Validation of human reliability assessment techniques: part 1—validation issues." *Safety Science* 27, no. 1 (1997): 25-41.

[6] Swain, Alan D., and Henry E. Guttmann. *Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report*. No. NUREG/CR--1278. Sandia National Labs., 1983.

[7] Gertman, D. I., H. S. Blackman, J. Byers, L. Haney, C. Smith, and J. Marble. "NUREG/CR-6883-The SPAR-H method." *Washington, DC: US Nuclear Regulatory Commission* (2005).

[8] Williams, J. C. "A proposed method for assessing and reducing human error." In *Proc. 9th Advances in Reliability Technology Symp.* Univ. of Bradford, 1986.

[9] Hollnagel, Erik. *Cognitive reliability and error analysis method (CREAM)*. Elsevier, 1998.

[10] Cooper, S. E., A. M. Ramey-Smith, J. Wreathall, and G. W. Parry. *A technique for human error analysis (ATHEANA)*. No. NUREG/CR-6350; BNL-NUREG-52467. Nuclear Regulatory Commission, Washington, DC (United States). Div. of Systems Technology; Brookhaven National Lab., Upton, NY (United States); Science Applications International Corp., Reston, VA (United States); NUS Corp., Gaithersburg, MD (United States), 1996.

[11] Bye, Andreas. " Informing HRA by Empirical Data, Halden Reactor Project Lessons Learned and Future Direction." In: Proceedings of the PSAM 14 2018 Conference: Probabilistic Safety Assessment and Management, UCLA, Los Angeles, 16-21 September, 2018.

[12] Kirwan, Barry. "Validation of human reliability assessment techniques: part 2—validation results." *Safety Science* 27, no. 1 (1997): 43-75.

[13] Zio, Enrico. "Reliability engineering: Old problems and new challenges." *Reliability Engineering & System Safety* 94, no. 2 (2009): 125-141.

[14] Bell, Julie, and Justin Holroyd. "Review of human reliability assessment methods." *Health & Safety Laboratory* 78 (2009).

[15] Mosleh, Adam, Van M. Bier, and George Apostolakis. "A critique of current practice for the use of expert opinions in probabilistic risk assessment." *Reliability Engineering & System Safety* 20, no. 1 (1988): 63-85.

[16] Mkrtchyan, Lusine, Luca Podofillini, and Vinh N. Dang. "Methods for building conditional probability tables of Bayesian belief networks from limited judgment: An evaluation for human reliability application." *Reliability Engineering & System Safety* 151 (2016): 93-112.

[17] Shirazi, Calvin Homayoon. "Data-informed calibration and aggregation of expert judgment in a Bayesian framework." PhD diss., 2009.

[18] Lin, Shi-Woei, and Vicki M. Bier. "A study of expert overconfidence." *Reliability Engineering & System Safety* 93, no. 5 (2008): 711-721.

[19] Laumann, Karin, Harold S. Blackman, and Martin Rasmussen. "Challenges with data for human reliability analysis." *Safety and Reliability–Safe Societies in a Changing World. Proceedings of ESREL 2018, June 17-21, 2018, Trondheim, Norway* (2018).

[20] Chang, Y. James, Dennis Bley, Lawrence Criscione, Barry Kirwan, Ali Mosleh, Todd Madary, Rodney Nowell et al. "The SACADA database for human reliability and human performance." *Reliability Engineering & System Safety* 125 (2014): 117-133.

[21] Lois, Erasmia. *International HRA Empirical Study--phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Similar Performance Data*. Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission, 2009.

[22] Kim, Yochan, Jinkyun Park, and Wondea Jung. "A classification scheme of erroneous behaviors for human error probability estimations based on simulator data." *Reliability Engineering & System Safety* 163 (2017): 1-13.

[23] Park, Jinkyun, and Wondea Jung. "OPERA—a human performance database under simulated emergencies of nuclear power plants." *Reliability Engineering & System Safety* 92, no. 4 (2007): 503-519.

[24] Gibson, W. Huw, and Ted D. Megaw. *The implementation of CORE-DATA, a computerised human error probability database*. HSE Books, 1999.

[25] Park, Jinkyun, Yochan Kim, and Wondea Jung. "Use of a Big Data Mining Technique to Extract Relative Importance of Performance Shaping Factors from Event Investigation Reports." In *International Conference on Applied Human Factors and Ergonomics*, pp. 230-238. Springer, Cham, 2017.

[26] Preischl, Wolfgang, and Mario Hellmich. "Human error probabilities from operational experience of German nuclear power plants, Part II." *Reliability Engineering & System Safety*148 (2016): 44-56.

[27] Preischl, Wolfgang, and Mario Hellmich. "Human error probabilities from operational experience of German nuclear power plants." *Reliability Engineering & System Safety* 109 (2013): 150-159.

[28] Kletz, Trevor. "Some Common Errors in Accident Investigations." In *Safety and Reliability*, vol. 31, no. 1, pp. 4-13. Taylor & Francis, 2011.

[29] Moura, Raphael, Michael Beer, Edoardo Patelli, John Lewis, and Franz Knoll. "Learning from major accidents to improve system design." *Safety science* 84 (2016): 37-45.

[30] Kyriakidis, Miltos, Arnab Majumdar, and Washington Y. Ochieng. "Data based framework to identify the most significant performance shaping factors in railway operations." *Safety science* 78 (2015): 60-76.

[31] API, ANSI. "API Recommended Practice 754." *Process Safety Performance Indicators for the Refining and Petrochemical Industries, 1st Ed., American Petroleum Institute, Washington, DC* (2010).

[32] Tolo, Silvia, Edoardo Patelli, and Michael Beer. "An open toolbox for the reduction, inference computation and sensitivity analysis of Credal Networks." *Advances in Engineering Software* 115 (2018): 126-148.

[33] Mkrtchyan, Lusine, Luca Podofillini, and Vinh N. Dang. "Bayesian belief networks for human reliability analysis: A review of applications and gaps." *Reliability engineering & system safety* 139 (2015): 1-16.

[34] Tolo, S., E. Patelli, and M. Beer. "Enhanced Bayesian network approach to sea wave overtopping hazard quantification." In *Proceedings of the 25th European safety and reliability conference, ESREL, Zurich, Switzerland, Sept*, pp. 7-10. 2015.

[35] Chen, Serena H., and Carmel A. Pollino. "Good practice in Bayesian network modelling." *Environmental Modelling & Software* 37 (2012): 134-145.

[36] Tolo, Silvia, Edoardo Patelli, and Michael Beer. "Risk assessment of spent nuclear fuel facilities considering climate change." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 3, no. 2 (2016): G4016003.

[37] Groth, Katrina M., and Ali Mosleh. "Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model." *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*226, no. 4 (2012): 361-379.

[38] Yang, Z. L., S. Bonsall, A. Wall, J. Wang, and M. Usman. "A modified CREAM to human reliability quantification in marine engineering." *Ocean Engineering* 58 (2013): 293-303.

[39] Sundaramurthi, Ranjitprakash, and C. Smidts. "Human reliability modeling for the next generation system code." *Annals of Nuclear Energy* 52 (2013): 137-156.

[40] Nielsen, T. D., and F. V. Jensen. "Bayesian Networks and Decision Graphs. Springer." (2009).

[41] Fenton, Norman, and Martin Neil. *Risk assessment and decision analysis with Bayesian networks*. Crc Press, 2012.

[42] Jentsch, Florian G. "Problems of systematic safety assessments: lessons learned from aircraft accidents." In *Verification and Validation of Complex Systems: Human Factors Issues*, pp. 251-259. Springer, Berlin, Heidelberg, 1993.

[43] Trucco, Paolo, Enrico Cagno, Fabrizio Ruggeri, and Ottavio Grande. "A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation." *Reliability Engineering & System Safety* 93, no. 6 (2008): 845-856.

[44] Yang, Z. L., S. Bonsall, A. Wall, J. Wang, and M. Usman. "A modified CREAM to human reliability quantification in marine engineering." *Ocean Engineering* 58 (2013): 293-303.

[45] Samaniego, Francisco J. "On closure of the IFR class under formation of coherent systems." *IEEE Transactions on Reliability* 34, no. 1 (1985): 69-72.

[46] Barlow, Richard E., and Alexander S. Wu. "Coherent systems with multi-state components." *Mathematics of operations research* 3, no. 4 (1978): 275-281.

[47] Kirwan, Barry, Richard Kennedy, Sally Taylor-Adams, and Barry Lambert. "The validation of three Human Reliability Quantification techniques—THERP, HEART and JHEDI: Part II—Results of validation exercise." *Applied ergonomics* 28, no. 1 (1997): 17-25.

[48] Pirie, W. "Spearman rank correlation coefficient." *Encyclopedia of statistical sciences* (1988).

[49] Abdi, H. "The Kendall rank correlation coefficient Encyclopedia of Measurement and Statistics ed N Salkind." (2007): 1-7.

[50] Morais C, Moura R, Beer M, Patelli E. Attempt to predict human error probability in different industry sectors using data from major accidents and Bayesian networks. In: *Proceedings of the Probabilistic Safety Assessment and Management (PSAM 14), September 16-21, Los Angeles; 2018.*

[51] Moura, Raphael, Michael Beer, Edoardo Patelli, John Lewis, and Franz Knoll. "Learning from accidents: interactions between human factors, technology and organisations as a central element to validate risk studies." *Safety Science* 99 (2017): 196-214.

[52] Authority, Civil Aviation. "CAP 737 Flight-crew human factors handbook." *London: Civil Aviation Authority* (2016).

[53] *BayesFusion, LLC. GeNIe Modeler. http://www.bayesfu-sion.com/, Accessed: 30 November 2017.*

[54] Patelli, Edoardo, Silvia Tolo, Hindolo George-Williams, Jonathan Sadeghi, Roberto Rocchetta, Marco de Angelis, and Matteo Broggi. "Opencossan 2.0: an efficient computational toolbox for risk, reliability and resilience analysis." In *Proceedings of the joint ICVRAM ISUMA UNCERTAINTIES conference*. 2018.

[55] Murphy, Kevin. "Software for graphical models: A review." *International Society for Bayesian Analysis Bulletin* 14, no. 4 (2007): 13-15.

Figure captions list

Table Captions list