Cite as:
Zwirglmaier K., Straub D., Groth K. (2017). Capturing cognitive causal paths in human reliability analysis with Bayesian network models. *Reliability Engineering & System Safety*, **158**: 117–129.

Capturing cognitive causal paths in human reliability analysis with Bayesian network models

Kilian Zwirglmaier, Daniel Straub
Engineering Risk Analysis Group, Technische Universität München (kilian.zwirglmaier@tum.de, http://www.era.bgu.tum.de)

Katrina M. Groth
Sandia National Laboratories, Albuquerque, NM

Abstract

In the last decade, Bayesian networks (BNs) have been identified as a powerful tool for human reliability analysis (HRA), with multiple advantages over traditional HRA methods. In this paper we illustrate how BNs can be used to include additional, qualitative causal paths to provide traceability. The proposed framework provides the foundation to resolve several needs frequently expressed by the HRA community. First, the developed extended BN structure reflects the causal paths found in cognitive psychology literature, thereby addressing the need for causal traceability and strong scientific basis in HRA. Secondly, the use of node reduction algorithms allows the BN to be condensed to a level of detail at which quantification is as straightforward as the techniques used in existing HRA. We illustrate the framework by developing a BN version of the *critical data misperceived* crew failure mode in the IDHEAS HRA method, which is currently under development at the US NRC (Xing et al., 2013). We illustrate how the model could be quantified with a combination of expert-probabilities and information from operator performance databases such as SACADA. This paper lays the foundations necessary to expand the cognitive and quantitative foundations of HRA.

Keywords: HRA; Bayesian networks; Bayesian updating; cognitive factors; causal paths

Acronyms

ACRS	Advisory Committee on Reactor Safeguards
ATHEANA	A Technique for Human Event Analysis
BN	Bayesian Network
CFM	Crew Failure Mode
CPT	Conditional Probability Table
DAG	Directed Acyclic Graph
DT	Decision Tree
HEP	Human Error Probability
HFE	Human Failure Event
HRA	Human Reliability Assessment
HSI	Human-System Interface
IDHEAS	Integrated Decision-Tree Human Event Analysis System
NRC	Nuclear Regulatory Commission
PDF	Probability Density Function
PIF	Performance Influencing Factor
PMF	Probability Mass Function
PRA	Probabilistic Risk Assessment
PSF	Performance Shaping Factor
SPAR-H	Standardized Plant Analysis Risk-Human Reliability Analysis method
THERP	Technique for Human Error Rate Prediction

1 Introduction

A comprehensive probabilistic risk assessment (PRA) is an essential element of safety and reliability assurance for many complex engineering systems. The aim of the PRA is to understand the possible failure scenarios, the corresponding adverse consequences, and the failure scenarios' probabilities. Most engineering systems can be characterized as human-machine systems, in which the human operator and the technical system are interacting. For that reason it is essential for a PRA to consider not only failures of technical components but also the effect of human actions and human inaction. Human reliability analysis (HRA) models human elements as part of PRAs; in general through identification and quantification of human failure events (HFEs) in PRA models. A variety of methods have been developed and applied in this field to determine human error probabilities (HEPs) corresponding to HFEs. Among the most important representatives are THERP (Swain and Guttmann, 1983), SPAR-H (Gertman et al., 2005) and ATHEANA (Cooper et al., 1996).

The limitations of existing HRA methods have been widely discussed previously (Woods, 1990, Hollnagel, 2000, Mosleh and Chang, 2004, Sträter, 2004, Boring et al., 2007, French et al., 2011, Groth and Swiler, 2013). Two interrelated shortcomings in existing HRA methods are the limited scientific basis used to develop those methods and the use of simplified modeling techniques, which lack causal structure and quantitative traceability.

Ongoing research into human performance is addressing the first shortcoming. The scientific foundations for human reliability have been explored and documented in the work by Whaley et al. (2012) on the psychological basis of HRA. In particular, they introduce a set of psychological failure mechanisms and proximate causes, which can lead to human failure events. Furthermore, they provide detailed insight into the factors that affect human performance (Performance Influencing Factors, PIFs), the dependency between those factors, and the causal pathways from those factors to human errors. International data collection activities offer insight into human performance in complex engineered systems (Park and Jung, 2007, CSNI, 2012, Chang et al., 2014), which provide new opportunities to improve the quantitative basis of HRA.

The second shortcoming, the lack of causal structure and quantitative traceability, is being addressed through advanced modeling efforts. Bayesian Network (BN) models (also called Bayesian Belief Networks), have becoming increasingly popular within HRA as a means for addressing these shortcomings because of their ability to explicitly model cause and effect combined with the ability to incorporate information from different sources (Baraldi et al., 2015, Mkrtchyan et al., 2015). Ongoing international research has demonstrated the ability of BNs both to capture the causal relationships among PIFs and to facilitate quantification of those relationships (Groth and Mosleh, 2012, Sundaramurthi and Smidts, 2013, Musharraf et al., 2014, Podofillini et al., 2014).

The psychological foundation has been leveraged in the development of two new HRA Methods, the IDHEAS (Integrated Decision-Tree Human Event Analysis System) method (Xing et al., 2013) and the PHOENIX method (Ekanem and Mosleh, 2014, Ekanem et al., 2016). Both IDHEAS and PHOENIX introduce the concept of crew failure modes (CFMs), a characterization of ways that a human failure event can occur during a crew interaction with the system. Both methods include a quantitative model relating PIFs to CFMs. However, the quantitative models in IDHEAS fall short of both causal and quantitative traceability; e.g. the motivation for the exclusion of cognitive mechanisms and PIFs from the method remains unclear (Stetkar, 2014). The PHOENIX method uses a BN model for quantification, but there are no directed arcs from one PIF to another, and thus the causal paths from the cognitive literature are not fully captured.

In this paper we propose a methodology to expand the scientific basis and traceability of HRA by capturing causal paths from cognitive literature in BN models. Furthermore we present a method for quantifying the BN model using Bayesian parameter updating to combine human performance data with expert elicitation results.

We introduce the methodology by developing a Bayesian network (BN) model for a single CFM from the IDHEAS method. We illustrate the procedure step by step, starting from the corresponding IDHEAS decision tree model, expanding the CFM model to a level where its cognitive foundation is modeled explicitly, and finally reducing the expanded model to a level where its quantification becomes straightforward. This process enhances the traceability between the HRA quantification models and the underlying cognitive literature basis. In addition we provide a method to quantify the new model based on expert elicitation and then show how a database can be used to update these expert elicited distributions, such that the final model is based on both expert knowledge and observed data.

2 Modeling and quantification tools

This section introduces Bayesian networks (BNs) and Bayesian updating, which provides the foundation for using a combination of experts' estimates and data for quantification.

2.1 Bayesian networks

Like decision, event- and fault trees, which are well known in the HRA community, BNs are a probabilistic modeling tool that is compatible with PRAs. In the last decade, BNs have been identified as a powerful tool for HRA, with multiple advantages over traditional HRA methods (Mkrtchyan et al., 2015). The graphical or qualitative part of a BN can be seen as a documentation of the causal dependencies between the random variables included in the model. An important difference to decision trees (DTs) is that BN models can explicitly represent the causality among the variables in the model and they do so in a computationally efficient way. The conditional independences underlying the graphical structure of the BN enable an efficient quantification of the model.

For brevity only a short introduction to the most important aspects of BNs is provided here. For a more in-depth treatment of BNs, the reader is referred to textbooks (Jensen and Nielsen, 2007, Kjaerulff and Madsen, 2013).

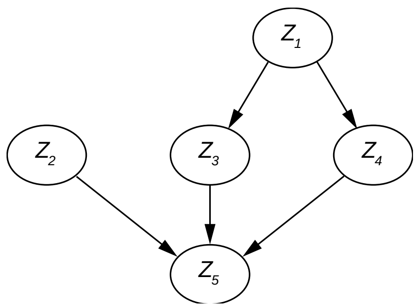

Figure 1. Example BN structure documenting the causal relationships between five variables (Z1 through Z5). In an HRA context, this model could be interpreted as the relationship between four PIFs (Z1, Z2, Z3, Z4) and HFE (Z5)

BNs are an efficient representation of a joint probability distribution $p(\mathbf{z})$ over a random vector $\mathbf{Z}$. Each node in the BN represents a random variable Z_i. The qualitative dependence structure between the random variables Z_i is represented by a directed acyclic graph (DAG). Family terms are used to describe relationships between random variables in a BN. In the BN of Fig. 1, Z_5 is a child of Z_2, Z_3 and Z_4, which in turn are its parents: $pa(Z_5) = \{Z_2, Z_3, Z_4\}$. Furthermore Z_1 to Z_4 are ancestors of Z_5, and Z_5 is a descendant of the former. Ideally, the graphical structure of a BN represents the causal relations among variables, but this is not a necessary condition. Interpreting the BN in Fig. 1 causally, one finds that Z_1 directly influences Z_3 and Z_4, but only indirectly influences Z_5.

In the BN, all random variables Z_i are specified by a conditional probability distribution given their parents, $p(z_i|pa(Z_i))$. For random variables without parents, this reduces to the marginal distribution $p(z_i)$. We restrict ourselves to BNs with discrete random variables, which are described by their conditional probability mass function (PMF). These are summarized in conditional probability tables (CPTs).

In the discrete BN, the joint probability distribution of all random variables is the product of all conditional PMFs:

$$p(\mathbf{z}) = \prod_{i=1}^{n} p(z_i|pa(Z_i)) \tag{1}$$

For the BN of Fig. 1, this reads as:

$$p(\mathbf{z}) = p(z_1) \cdot p(z_2) \cdot p(z_3|z_1) \cdot p(z_4|z_1) \cdot p(z_5|z_2, z_3, z_4) \tag{2}$$

Formulating the joint distribution as a product of conditional distributions facilitates the quantification of the model: It is significantly easier to elicit or quantify these conditional terms than more general joint distributions. In the example of Fig. 1, the random variable Z_4 is related to all other random variables (for an in-depth description of these relations see (Pearl, 1988)). However, the analyst need only specify $p(z_4|z_1)$, which is further simplified if Z_4 is causally dependent on Z_1. The BN structure then takes care of the dependence between Z_4 and the remaining random variables in the model. In this way, the BN supports an intuitive modeling process. In addition, the BN structure also reduces the number of parameters that need to be estimated.

The BN model supports practitioners in reasoning about the variables in the model. When observing some of the random variables in the BN, the conditional probability distributions of other random variables given the observations can be calculated with standard BN algorithms (e.g., the HEP given the states of some or all PIFs). In the process of reasoning, the parameters of the BN model remain untouched. In addition, real life situations or simulator experiments provide new information on the parameters of the BN model. The framework of learning and updating the parameters of the BN model with new data is called Bayesian parameter updating.

2.2 Bayesian parameter updating

Bayesian updating is applied to enhance the experts' estimates of the crew failure scenario probabilities with new data. We revisit the most important aspects of Bayesian updating; for a more in depth treatment we refer to (Kelly and Smith, 2009, Groth et al., 2014). The goal of Bayesian updating is to learn the distribution of one or more parameters $\boldsymbol{\theta}$. In the case of the HRA example considered in this paper, the parameters of interest are the crew failure scenario probabilities conditional on the PIFs, $p_{CFM|PIF_1,\dots,PIF_m}$. In other HRA applications, the parameters include PIF probabilities and HEPs. The prior PDF (Probability Density Function) $\pi_0(\boldsymbol{\theta})$ represents the belief in the state of $\boldsymbol{\theta}$ before considering the data, e.g., the probabilities based solely on expert elicitation. The data can be the result of one or more simulator experiments or operating events. Applying Bayes' rule (Eq. 3) allows one to combine the prior distribution with the data $\mathbf{x}$ to get the posterior distribution $\pi_1(\boldsymbol{\theta}|\mathbf{x})$, representing the belief in the state of $\boldsymbol{\theta}$ after observing $\mathbf{x}$:

$$\pi_1(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}) \tag{3}$$

where $f(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{x}$.

A typical database on crew performance in NPP control rooms or simulators contains the number of positive/negative outcomes in a number of trials. Since the outcome of each event is binary (e.g., success or failure), and assuming that the trials are independent of each other, this data can be modeled as a Bernoulli process. The parameter to estimate is the probability of failure θ and the observation is n_e, the number of times the crew failed to deal with the scenario in a correct way in a total of n observed/simulated scenarios. In this case, the likelihood function is $f(n_e|\theta)$ and is the binomial PMF (Probability Mass Function) with parameter θ:

$$f(n_e|\theta) = \binom{n}{n_e} \cdot \theta^{n_e} \cdot (1-\theta)^{n-n_e} \tag{4}$$

For binomial data, it is mathematically convenient to use a beta distribution to represent the prior beliefs on θ, because the beta distribution is a conjugate prior for the binomial likelihood function (Raiffa and Schlaifer, 1961). The beta PDF with parameters a_0 and b_0 is:

The use of conjugate priors greatly simplifies the mathematics of Bayesian updating. If the beta

$$\pi_0(\theta) = \frac{1}{B(a_0\,, b_0\,)}\theta^{a_0-1}(1-\theta)^{b_0-1} \tag{5}$$

distribution is used to model $\pi_0(\theta)$ and the likelihood function is the binomial PMF of Eq. 4, the posterior $\pi_1(\theta|n_e)$ is beta distributed as well; one example is shown in Fig. 2. In this case, the parameters of the posterior beta distribution can be calculated analytically as:

$$a_1 = a_0 + n_e \tag{6}$$

$$b_1 = b_0 + (n - n_e) \tag{7}$$

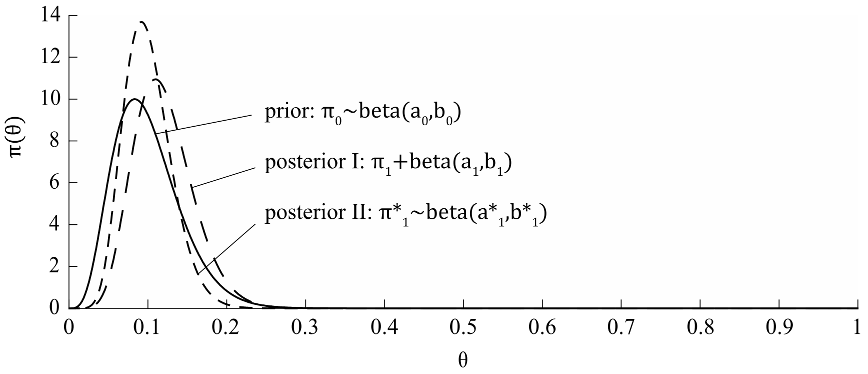

Figure 2. Updating the beta prior distribution with two different data sets. Posterior I is obtained by updating with a dataset of length $\mathrm{n} = 25$ in which $\mathrm{n_e} = 4$ occurrences of a HFEs are observed. Alternatively, posterior II is obtained by updating with a dataset of length $\mathrm{n} = 50$ with $\mathrm{n_e} = 5$ HFE occurrences.

Typically, the parameters of the beta prior, a_0 and b_0, are obtained from experts. Multiple techniques to elicit beta parameters from experts have been proposed (Bedford and Cooke, 2001, O'Hagan et al., 2006). A straightforward approach to estimate the parameters of $\pi_0(\theta)$ is to first elicit the expected value $E[\theta]$. A first constraint on the distribution parameters is then given through the definition of the expected value of a beta distribution:

$$E[\theta] = \frac{a_0}{a_0 + b_0} \tag{8}$$

In a next step, the standard deviation of the distribution needs to be determined, representing the experts' uncertainty on θ. This elicitation may be done graphically, as in Fig. 3, where beta PDFs are shown with an expected value $E[\theta] = 0.1$ and varying standard deviation.

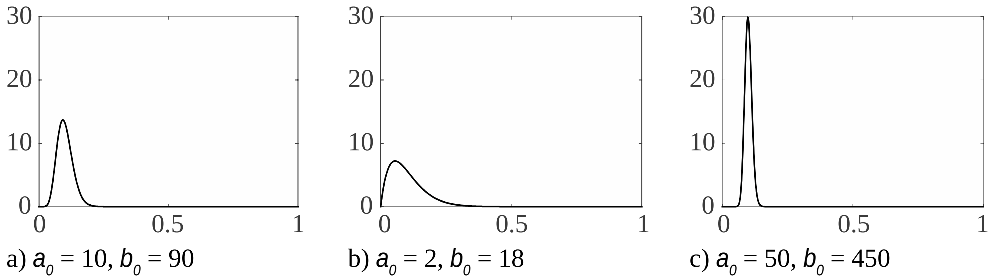

Figure 3. PDFs of beta distributions with mean 0.1 and varying standard deviations. The standard deviations of the distributions are a) 0.03, b) 0.07 and c) 0.01.

3 Crew failure modes in HRA

Two new HRA methods incorporate the concept of crew failure modes: the IDHEAS method developed by the U.S. NRC, and the PHOENIX method developed by the University of Maryland (Ekanem and Mosleh, 2014, Ekanem et al., 2016). PHOENIX and IDHEAS follow a similar modeling approach combining both qualitative and quantitative steps:

- Performing a qualitative task analysis and documenting crew failure paths in a crew response tree (CRT).
- Selecting applicable crew failure modes (CFMs) for each event in a CRT.
- Quantifying the individual CFMs (via Decision Trees for IDHEAS, and via BNs for PHOENIX) and combining probabilities of the relevant CFMs to calculate the human error probabilities (HEPs) for each event.
- HFE dependency analysis and possible recovery actions.

In this work we focus specifically on the modeling and quantification of the CFMs. In both methods, the CFMs are a crucial element, which translates the concept of human errors from a psychological perspective into crew errors that could lead to an HFE.

IDHEAS and PHOENIX each derive their CFMs from the psychological failure mechanisms (Whaley et all 2014). In both methods, PIFs are used to characterize the content of the task, and PIFs are used to quantify the probability of occurrence of a CFM. The two methods differ in the number of CFMs used, as well as the quantification approach. IDHEAS considers 14 Crew Failure modes (CFMs) representing failures that are typical for human performance in nuclear power plant control rooms. The CFMs in IDHEAS are summarized in Tab. 1. The PHOENIX method considers 19 different CFMs, which are summarized in Tab. 2.

In IDHEAS, each CFM is quantified using a DT[1], such as the one shown in Fig. 4. Each PIF is represented as a branch point in the DT. For simplicity, the IDHEAS developers chose to limit the number of PIFs in each DT to four.

Table 1. Crew Failure Modes used in the IDHEAS method (Xing et al., 2013)

Phase of response	Plant status assessment	Response planning	Execution
Crew Failure Mode (CFM)	Key alarm not attended to[†]	Delay implementation[†]	Fail to initiate execution
	Data misleading or not available	Misinterpret procedure[†]	Fail to execute response correctly
	Premature termination of critical data collection	Choose inappropriate strategy	
	Critical data misperceived		
	Wrong data source attended to[†]		
	Critical data not checked with appropriate frequency		
	Critical data dismissed/discounted[†]		
		Misread or skip step in procedure[*†]	
	Critical data miscommunicated[**†]		

[†] CFM for which data was collected.

[1] Note: The applied models are referred to as decision trees in the IDHEAS report (Xing et al. 2013). However, since there are no decisions involved, the tool should be more appropriately termed event tree in a PRA sense. Because this paper is mainly intended for the HRA community, we stick to the terms used in the IDHEAS report.

[*] May occur in either ‘Response Planning’ or ‘Execution’ phases.

[**] May occur in any of the three phases.

In PHOENIX, the CFMs are quantified using one BN that maps relationships between PIFs and all nineteen CFMs. PHOENIX considers nine “primary” PIFs, which all directly influence each of the CFMs. PHOENIX also includes an expanded qualitative BN model, which includes approximately 20 additional PIFs that have been collapsed into the nine primary PIFs. The BN model used in PHOENIX does not directly model interdependency between the PIFs.

Table 2. CFMs used in the PHOENIX method.

Information processing	Diagnosis/decision making	Action taking
Key alarm not responded to (intentional & unintentional)	Plant/system state misdiagnosed	Incorrect timing of action
Data not obtained (intentional)	Procedure misinterpreted	Incorrect operation of component/object
Data discounted	Failure to adapt procedure to the situation	Action on wrong component / object
Decision to stop gathering data	Procedure step omitted (intentional)	
Data incorrectly processed	Deviation from procedure	
Reading error	Decision to delay action	
Information miscommunicated	Inappropriate strategy chosen	
Wrong data source attended to		
Data not checked with appropriate frequency		

3.1 IDHEAS critical data misperceived

In the remainder of this paper, the CFM *critical data misperceived* is considered exemplarily to demonstrate the proposed framework. This CFM is presented to some detail in the following. *Critical data misperceived* captures situations such as the one in which a parameter has to be read from a control panel or the status of some piece of equipment is to be determined from an indication on the control panel and this piece of information is critical in the sense that its misinterpretation will lead to an incorrect response (Xing et al., 2013). Three PIFs are used to describe the context: *HSI/environment*, *workload*, and *training*[2], where HSI refers to Human-system interface. All the PIFs are binary with states labeled as {poor and good}, {high and low} or {no and yes}. In Fig. 4 the DT for the CFM *critical data misperceived* is shown. Each path through the DT represents one possible crew failure scenario. The analysts are provided with a

[2] This DT also contains a branch for recovery potential, which is used in most IDHEAS CFM. The meaning of “recovery potential” has been defined in a generic manner as “opportunities for correction given failure”. However these opportunities have not been clearly specified for the considered CFM, and thus we neglect this concept in the remaining sections of the paper.

set of two to five questions[3] for each PIF guiding them in determining the states of the PIFs (see Appendix A). Expert elicitation was used to assign probabilities to the different crew failure scenarios.

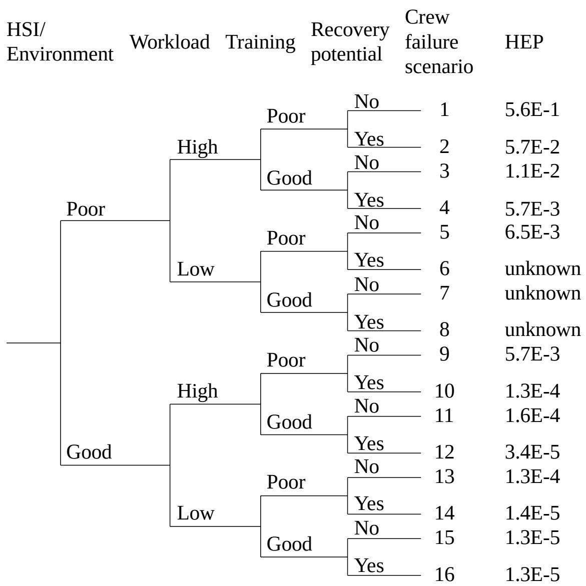

Figure 4. Decision tree for the crew failure mode critical data misperceived (Xing et al., 2013). The paths through the decision tree are numbered and for each path a probability was elicited from experts. E.g. the HEP for poor HSI/environment, high workload, poor training and no recovery potential is 0.56. (Note: The expert elicitation task has not been completed as of the writing of this paper; some probabilities are listed as “unknown” and some may change in the final IDHEAS report.)

4 Development of a BN structure for each CFM

As explained in section 2, the directed acyclic graph (DAG) part of a BN ideally represents the causal relationships between the random variables in the model. Furthermore, the structure also defines the information (i.e., the marginal and conditional probabilities) needed to quantify the BN. In this section, we illustrate the development of two BN structures for each CFM: a first BN that contains an expanded causal structure based on cognitive literature (Whaley et al., 2012) and PIF specification nodes corresponding to the questions in Appendix A; and a second BN obtained through reduction of the first structure.

Since the availability of data is the main bottleneck in HRA, we aim at developing a BN structure whose quantification requires roughly the same amount of information as the original DT. In section 4.3 we enhance the causal details in the original method by explicitly including the PIF specifications and by adding PIFs that are essential to the interpretation of the CFM. This model can be quantified or used un-quantified to help document the causal paths on which the model is based. In section 4.4, we demonstrate that node reduction algorithms can be used to reduce the BN with full causal details down to a structure, for which quantification is feasible

[3]These questions are not explicitly included in the DT. In section 4 we propose a way to directly include these questions in the model.

with respect to data availability. In the presented example, the final model is equivalent to the DT with explicit inclusion of the PIF specification nodes.

In the following subsections, we discuss the general idea behind the structure development approach and we develop the structure for the crew failure mode *critical data misperceived* step by step. Quantification of the models is discussed in section 5.

4.1 Summary of approach and models

We propose an approach for developing causal (BN) models for HRA starting from the psychological basis of the models. The following steps summarize the approach:

- Review of the cognitive foundation for each CFM to identify the main causal failure paths, the PIFs and possibly other relationships.
- Development of an exhaustive causal model including all identified causal failure paths, PIFs and relationships.
- Application of node reduction algorithms, to remove nodes from the model that are not quantifiable with feasible effort.
- Elicitation of experts and initial quantification of model.
- Updating the quantification with results from human performance databases.

4.2 BN model of original IDHEAS DT

Each DT used to quantify the IDHEAS CFMs includes a number of PIFs. It is straightforward to develop a BN structure out of these PIFs and the target node, which is the node representing the CFM event. Since the PIFs influence the state of the CFM node, generally the PIFs are modeled as parents of the latter. From the structure it is clear that quantifying these structures requires conditional probability distributions for the CFM node and marginal distributions for the PIF nodes.

The simple BN structure is shown in Fig. 5 for the CFM *critical data misperceived*. In this BN, the PIFs are assumed to be independent if the target node is not observed. The question if the PIFs are actually independent is not addressed within the original IDHEAS framework, since the IDEHAS decision trees provide HEPs only conditional on the states of all PIFs.

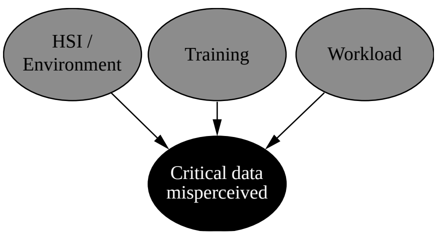

Figure 5. BN for the CFM *critical data misperceived* that corresponds to the original DT model.

4.3 BN model with full causal details

The BN model in Fig. 5, derived from the DT, reveals little about the cognitive paths leading to crew failure. This missing information is, however, essential to understanding the model with its features and limitations. The model is therefore expanded. We expand the model at two levels to the BN shown in Fig. 6. Firstly, we add an additional layer of nodes (white in Fig. 6). These nodes are intended to specify the causal paths leading to error based on cognitive psychology. These nodes are often too generic or abstract for analysts to directly determine the states, but critical for correctly modeling human performance. Secondly, an additional layer of PIF specification nodes is introduced (light grey in Fig. 6). These are based on questions and rules for the analysts that are provided in IDHEAS to support the determination of the states of the PIFs (Appendix A).

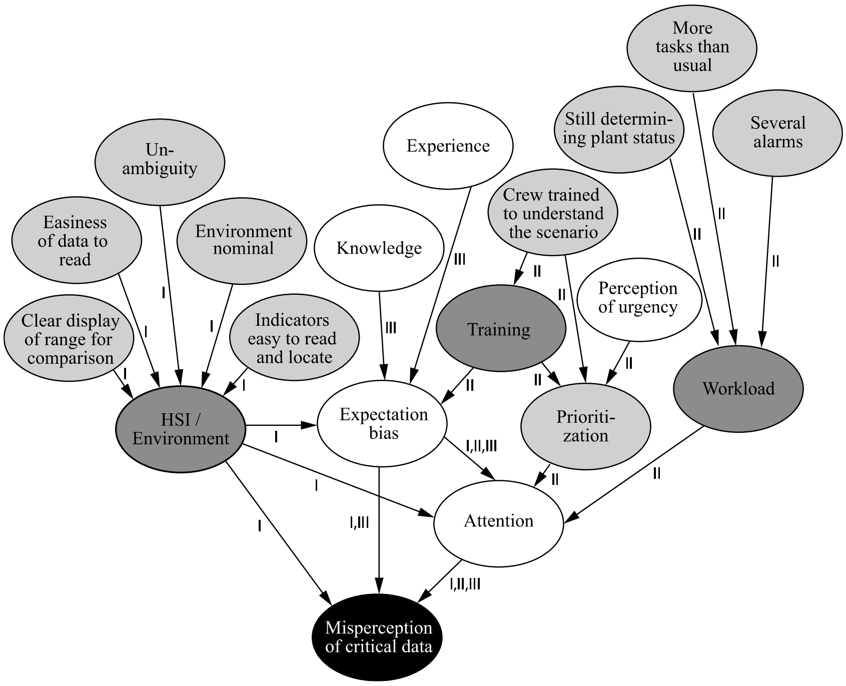

Figure 6. Fully expanded BN for the CFM *critical data misperceived*. The black node represents the target variable; dark grey nodes the PIF variables; light grey nodes the PIF specification variables and white nodes additional variables illustrating the causal paths. The causal paths I to III are indicated through roman numerals.

The literature serving as a foundation for IDHEAS (Whaley et al., 2012) summarizes the causal paths that can lead to a crew failure event, based on a comprehensive study of cognitive psychology. These paths can be implemented directly as nodes in the model to add additional causal details extracted from scientific literature. For the example CFM (Fig. 6) there are three main causal paths leading to data misperceived, following cognitive literature (Köhler, 1947, Broadbent, 1958, Tversky and Kahneman, 1974, Biederman, 1987, Endsley, 1995, Klein, 1998, Warner and Letsky, 2008):

- The first causal path (path I in Fig. 6) corresponds to the <u>misperception of data due to extreme *HSI/environment* conditions</u>. In this case, the quality of the HSI is so poor, or certain factors in the environment are so severe, that the information is degraded in such a way that it is misperceived. For example, the operators may be inundated with so many alarms that they experience sensory overload (Broadbent, 1958) and therefore misperceive the critical data. Technically this could be seen as an instrumentation failure rather than a HFE, but this instrumentation failure would manifest as a human failure event (Endsley, 1995, Klein, 1998).

- The second causal path (path II in Fig. 6) is attention degradation that leads to misperception. *Attention* can be degraded due to a combination of factors, including characteristics of the situation and the information (e.g., the *HSI and environment*), *high workload*, multiple *priorities*, and through the biases introduced by *training*, *knowledge*, and *experience*. *Training*, *workload* and *perception of urgency* cause the crew to *prioritize* certain tasks and direct *attention* to these. A misdirection of *attention* can lead to *misperception of critical data*. The *prioritization* and the crew members' *expectation biases* determines the amount of *attention* paid to the various pieces of information, which again may lead to *misperceiving the critical data* (Eriksen and St. James, 1986, Endsley, 1995).
- The third causal path (path III in Fig. 6) stems from *expectation biases* related to *experience* and *knowledge*, which can cause *misperception of critical data*. This can occur in a direct manner, e.g., situations where a person "sees what they want to see", or indirectly through changing the person's *attention* to focus on other data (Einhorn and Hogarth, 1981, Endsley, 1995).

As shown in the model (Fig. 6), the PIFs identified in the IDHEAS model influence the occurrence of the CFM through multiple causal paths. *HSI/environment* influences the target CFM through one direct causal path and additionally through two indirect causal paths. *Training* also influences the CFM (indirectly) through two different causal paths. The third causal path, *expectation bias*, is only indirectly captured in the original IDHEAS model.

The IDHEAS PIF specification nodes (light grey nodes) are intended to capture various aspects of the three PIFs, and are used in this model to demonstrate how observable questions can be explicitly included in the model.

The node *prioritization* has a dual role. Firstly, it represents a PIF question specifying training, which is "Is the significance of the decision that is based on obtaining this information correctly given a high priority compared to other concurrent tasks?". Secondly, *prioritization* is part of the second causal path. According to this path *training* influences *prioritization*. The link is thus directed from *training* to *prioritization*, and not like other PIF specifications the other way around (cf. Groth and Mosleh, 2012). To capture the influence of *prioritization* according to its role as a PIF specification node correctly, an additional dependence between *prioritization* and the node *crew trained to understand the scenario* needs to be introduced. Further discussions on the role of the node prioritization may be necessary, but are left for future research.

4.4 BN model reduction to facilitate IDHEAS-like quantification

The full model in Fig. 6 can be quantified using a variety of approaches. However, a secondary objective of this work is to develop a HRA model based on IDHEAS, and thus to limit the amount of additional information that must be elicited. To achieve this goal, the model in Fig. 6 is reduced to a form that more closely resembles the original IDHEAS DT, but is augmented with the PIF specifications from Appendix A.

To do so, the node removal algorithm by (Shachter, 1986) is applied to the BN of Fig. 6. This algorithm allows removing nodes, which have not received evidence, in a way that the (in)dependence assumptions incorporated in a BN are not altered (Straub and Der Kiureghian, 2010). The two principles of node removal are:

- Firstly, a node, which has not received evidence and which does not have children can be removed from the network. One refers to such nodes as barren.
- Secondly, the direction of a link between two nodes Z_i and Z_j can be reversed if Z_i inherits Z_j's parents and vice versa and if this does not cause the BN structure to become cyclic.

Nodes are eliminated by first reversing all links so that the nodes to be removed have no children, and then removing them. In this way, the joint probability of all remaining nodes in the BN is unaltered. Removing the white nodes from the BN in Fig. 6 results in the BN of Fig. 7.

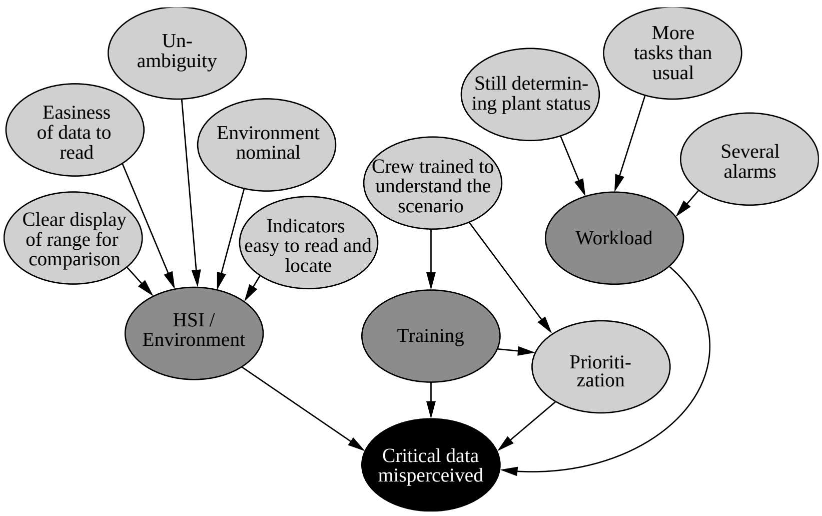

Figure 7. Reduced BN for the CFM *critical data misperceived*.

Exemplarily we describe the removal of the node *attention*. This node has one child, namely *critical data misperceived*. In order to make *attention* barren, its link pointing to *critical data misperceived* should be reversed. To not introduce any new independencies by doing so, both of the involved nodes need to inherit each other's parents. *Prioritization* and *workload* thus become parents of *critical data misperceived*. On the other hand *critical data misperceived* does not have any other parents, which are not at the same time parents of *attention*, with the exception of *attention* itself. Therefore the node *attention* does not inherent any additional parents. Since reversing the link between *attention* and *critical data misperceived* makes the former node barren, it can be removed.

4.5 Discussion/ implications of the models

Of special interest is the causal role of the node *prioritization* in this structure. According to causal path II discussed in section 4.2, high *workload* increases the probability of *misperceiving critical data* if the crew does not set correct priorities. In Fig. 6, *prioritization* is modeled as a child of *perception of urgency* and *training*. The causal interpretation is that both *training* and *perception of urgency* influence the probability of correct *prioritization*. Both *prioritization* and *workload* are parents of the node *attention* in Fig. 6. The combination of ineffective *prioritization* together with high *workload* will influence the *attention* paid to critical data.

Our derivation of the BN model from the cognitive paths proposed a direct dependency of *critical data misperceived* on the node *prioritization*, which is not considered directly as a PIF in IDHEAS. A detailed discussion on whether the inclusion of this is necessary or whether there are reasons to exclude this PIF is not within the scope of this paper. Since multiple cognitive literature sources indicate this dependency (Eriksen and St. James, 1986, Endsley, 1995), it is considered a critical PIF for accurately representing the cognitive factors.

The BN structure in Fig. 7 has advantages over the simple BN structure of Fig. 5. Firstly, in this model the analyst would directly answer the questions corresponding to the PIF specification nodes rather than assigning a PIF state based on implicit consideration of the questions, which is a much more abstract process[4]. The explicit inclusion of PIF specifications in the model expands the level of documentation provided by the model, enhances the traceability from analysis input to probability estimate, and reduces variability among analysts.

Secondly, if marginal probabilities are elicited for the PIF specification nodes, as was done in (Hallbert and Kolaczkowski, 2007), the BN in Fig. 7 can deal with missing information or uncertainty about one of the PIF specification nodes' states. For example, the HRA analyst may lack information about specific indicator designs, which may make it difficult to assess the state of *easiness of data to read*. In situations where the analyst does not have information about one or more PIFs, the analyst can use the prior probabilities in the BN rather than guessing or making unwarranted assumptions about the system.

Thirdly, the fully quantified BN can be used to reason about additional problems and gather additional insight. With identical analyst inputs, the BN structure in Fig. 7 will produce the same HEP assignments as the IDHEAS DT. However, the BN structure also offers the opportunity to reason about the PIFs, given knowledge of the CFM (and/or other PIFs). This provides an added benefit: the ability to identify, which PIFs (or PIF details) are likely to be present when we know there is an HFE. This gives insight into the probabilities of the causes

[4] This holds also if the original DT is considered together with the questions provided in the IDHEAS report.

or HFEs, which is a critical piece of information that can be used to prevent errors (Groth and Swiler, 2013).

5 CFM BN Quantification

This section describes the quantification of the BN structures developed in the previous section. We first present the straightforward quantification of the BN model in Fig. 5 based on the IDHEAS DT and how this simple model can be augmented with expert elicited data about the PIFs. Thereafter we show how the BN of Fig. 7 can be quantified using expert estimates, and finally how information from the SACADA (Scenario Authoring, Characterization, and Debriefing Application) (Chang et al., 2014) or similar databases can be used in this quantification. The BN of Fig. 6 can be quantified using a similar approach, but this is omitted for brevity.

5.1 Quantifying the BN model based on the original DT

As discussed in section 2, there is a conditional probability table (CPT) attached to each of the nodes in a BN. The CPT of the node *critical data misperceived* in the BN of Fig. 5 is identical to the conditional HEPs from the corresponding decision tree with one exception: the contribution of the recovery=yes branches is omitted because recovery is not clearly defined for this CFM. This CPT is shown in Tab. 3.

Table 3. CPT of the node *critical data misperceived*. The HEPs corresponding to the grey cells are marked as unknown in Fig. 4. For that reason, the estimates for the scenario [HSI = poor, workload = low and training = poor] were used. This corresponds to a conservative approximation, since changing the state of training from good to poor will certainly increase the HEP.

HSI/Env.	Poor				Good			
Workload	High		Low		High		Low	
Training	Poor	Good	Poor	Good	Poor	Good	Poor	Good
Recovery	No	No	No	No	No	No	No	No
Error	0.56	0.011	6.5E-3	6.5E-3	5.7E-3	1.6E-4	1.3E-4	1.3E-5
No error	0.44	0.989	0.9935	0.9935	0.9943	0.99984	0.99987	0.999987

Quantification of the BN also requires probability distributions for each of the PIF nodes. Unlike the conditional HEPs, these probability distributions are not provided by the original IDHEAS method.

The marginal distributions of the PIF nodes can be quantified using dummy distributions (e.g. assigning the same probability to each state of a PIF). In that case quantifying the BNs does not require any additional probability elicitations compared to quantifying the DT models. However, if dummy distributions are used for the PIF nodes, the BN, like the DT model, is only capable of giving ‘correct’ HEPs if the states of all PIFs are known (i.e., the BN model will predict HEPs identical to the DT, but additional benefits of the BN cannot be realized).

If the marginal distributions are actually elicited, the simple BN structure is capable of dealing with uncertainty about PIF states. In (Groth and Swiler, 2013) expert elicitations were used to quantify the CPTs of the PIFs, based on information elicited in (Hallbert and Kolaczkowski, 2007). Probability distributions for the PIFs of the CFM *critical data misperceived*, which are based on (Groth and Swiler, 2013) are given in Tab. 4.

Table 4. Illustrative probabilities quantifying the CPTs of the PIFs. The probabilities are based on (Groth and Swiler, 2013).

PIF	PIF state	$\Pr(PIF)$
HSI/environment	good	0.16
	poor	0.84
Training	good	0.67
	poor	0.33
Workload	low	0.84
	high	0.16

5.2 Quantification of BN model of DT with PIF specification nodes

We illustrate the quantification of the final BN model of Fig. 7. In this and similar BNs, there are three types of nodes to quantify: the CFM node (conditional on the PIFs), the PIF nodes (conditional on the PIF specification nodes), and the PIF specification nodes (marginal probabilities since these have no parents).

5.2.1 CFM node given PIFs

The parameters used in section 5.1 to quantify the CFM node were point estimates. (Xing et al., 2013) additionally provides HEPs with corresponding uncertainty estimates (i.e. quantile estimates). We fit beta distributions to these quantile estimates. For proof-of-concept, artificially generated data is used to update these fitted beta priors; the expected values of the posterior distributions are then used to quantify the final BN. Due to the updating as well as the fitting process, the parameters of the final BN do not exactly correspond to the parameters from Tab. 3.

For quantification of the BN in Fig. 5 following section 5.1, the target node needs to be defined conditional on three parents, i.e. the three PIFs. For the BN in Fig. 7, an additional direct dependency of the target node on the node *prioritization* was found. It is thus necessary to define the CPT of the target node on *HSI/environment*, *training*, *workload* and *prioritization*. Since IDHEAS does not provide probabilities for *critical data misperceived* conditional on the state of *prioritization*, simple assumptions are used here. For the purpose of illustration, we assume that *workload* and *prioritization* interact in a way that the conditional probabilities of *critical data misperceived* are equal to:

- the conditional probabilities derived for low *workload*, if the crew is able to *prioritize*;

- the conditional probabilities derived for low *workload*, if *workload* is low and the crew is not able to *prioritize*;
- the conditional probabilities derived for high workload, if *workload* is high and the crew is not able to *prioritize*;

5.2.2 PIF nodes

In section 5.1, marginal probability distributions need to be assigned to quantify the PIF nodes of the simple BN in Fig. 5. For the quantification of the extended BN in Fig. 7, the PIF nodes are defined conditional on PIF specification nodes. The IDHEAS report (Xing et al., 2013) provides rules on defining the states of the PIFs given the states of the PIF specification nodes. To be in line with the original method, we use these rules to quantify the CPTs of the PIF nodes. Presently, these rules are deterministic, which means they can be modeled as AND or OR relationships (deterministic nodes in the BN). These rules are provided for the CFM *critical data misperceived* as pseudo-code in Appendix B. Future research could focus on redefining these rules if it is found that these deterministic rules do not match reality, or if the relationship between the PIFs and PIF specifications is more nuanced than originally thought. It is, however, important that there is a common understanding on how the PIF specification nodes are linked to the PIF nodes, on how these nodes are included in the BN, and on how to quantitatively represent the dependence.

These rules can directly be transformed to CPTs quantifying the respective nodes in the BN. For example, the CPT for *training* given the two corresponding PIF specifications is provided in Tab. 5. The IDHEAS report defines *training* conditional on *prioritization* and *crew trained to understand the scenario*. However, to represent causality, the node *prioritization* is considered as a child of *training* rather than its parent in the BNs of Fig. 6 and 7. The CPT derived from the rules provided in the IDHEAS report needs thus to be reformulated using Bayes' rule[5].

Table 5. Deterministic CPT of training given the two corresponding PIF specification nodes

Crew trained to understand the scenario	Yes		No	
Prioritization	Yes	No	Yes	No
Good training	1	1	1	0
Poor training	0	0	0	1

5.2.3 PIF specification nodes

The PIF specification nodes require marginal probabilities. We elicit these probabilities from experts. Since a CFM like *critical data misperceived* typically does not apply in a NPP's normal state, it is important for the elicitation process that the experts understand that they are to give probabilities that are implicitly conditional on scenarios in which the CFM may apply. For example, if the CFM *critical data misperceived* and the PIF specification node *nominal*

[5] Note: No additional parameters need to be elicited in this case.

environment are considered, the experts need to give a probability of the event environment being nominal in situations where critical data is received. As proof-of-concept, the PIF specification nodes of the CFM *critical data misperceived* were quantified. The question, whether it is actually reasonable to elicit probabilities for such nodes conditional on being in an off-normal state, from experts should be further discussed. Such a discussion is outside the scope of this paper.

A small survey was carried out to illustrate the process to elicit the prior probabilities of the PIF specification nodes. The survey participants were two HRA experts with a background in cognitive psychology (experts I and II) and one former operator of a nuclear power plant on a submarine (expert III). Since the final probabilities should be elicited from actual nuclear power plant operators, the numbers given in this report are intended only for illustrative purposes of the framework. The survey is shown in Appendix C and the numbers given by the experts are summarized in Tab. 6.

Table 6. Results from the survey, carried out to elicit prior probabilities for the PIF specification nodes. Experts I and II are HRA specialists with a background in cognitive psychology and expert III is a former operator of a nuclear power plant on a submarine. The elicited numbers are probability estimates for the PIF specification nodes being in state “yes”.

PIF specification node	Expert			Mean
	I	II	III	
Indications clear and unambigous	0.9	0.8	0.9	0.87
Easiness of information to read	0.8	0.75	0.9	0.82
Clear display of range for comparison	0.8	0.8	0.95	0.85
Environment nominal	0.2	0.25	0.25	0.23
Indicators/source of data easy to read and locate	0.8	0. 8	0.5	0.7
Still determining plant status	0.7	0.65	0.98	0.78
Several alarms	0.8	0.9	0.98	0.89
More tasks than usual	0.15	0. 3	1.0	0.48
Crew trained to understand the scenario	0.8	0.8	0.98	0.86
Prioritization	0.95	0.9	0.9	0.92

The experts agree well on most questions. But there is a large spread in the experts estimates in questions 5 (Indicators/source of data easy to read and locate), 6 (Still determining plant status) and 8 (More tasks than usual). The mean values from Tab. 6 are directly used to quantify the marginal probabilities of the BN in Fig. 7, i.e. for the quantification of the PIF specification nodes we do not consider parameter uncertainties.

6 Updating with data

In this section, we illustrate how the SACADA database (Chang et al., 2014) could be used to update the probabilities of the CFM node in the IDHEAS-BN. These HEPs in IDHEAS are conditional on the relevant PIFs. Since SACADA and IDHEAS are not completely consistent, it is not always possible to deterministically decide in which states the IDHEAS PIFs are for a given SACADA case. Nevertheless, SACADA still provides information, which can and should be used to improve the quantitative side of IDHEAS. To this end, we define rules to probabilistically map SACADA onto IDHEAS. These rules allow estimating a probability of observing a specific context $\Pr(PIF_1 = s_1, \dots, PIF_m = s_m)$, which is defined through the states s_i of the PIFs, given a SACADA case. This concept is well known in HRA, e.g., from the quantification of SPAR-H (Gertman et al., 2005). Consider a crew failure scenario in a CFM, for which the corresponding probability $p_{CFM|PIF_1=s_1,\dots,PIF_m=s_m}$ is to be updated. The prior distribution of $p_{CFM|PIF_1=s_1,\dots,PIF_m=s_m}$ is beta with parameters a_0 and b_0. A database with n cases is used to update the distribution of $p_{CFM|PIF_1=s_1,\dots,PIF_m=s_m}$. In the case where the database is not completely consistent with the PIFs, one can rewrite Eqs. 6 and 7 to:

$$a_1 = a_0 + \sum_{i=1}^{n} \Pr(PIF_1 = s_1, \dots, PIF_m = s_m | \text{SACADA case } i) \cdot I_F(\text{SACADA case } i) \tag{9}$$

$$b_1 = b_0 + \sum_{i=1}^{n} \Pr(PIF_1 = s_1, \dots, PIF_m = s_m | SACADA\ case\ i)\big(1 - I_F(\text{SACADA case } i)\big) \tag{10}$$

where I_F is an indicator function, which is 1 if a failure was recorded in the SACADA database case i and 0 otherwise. Applying Eqs. 9 and 10 requires the conditional probabilities $\Pr(PIF_1 = s_1, \dots, PIF_m = s_m | \text{SACADA case } i)$. We use expert estimates to determine the distributions of the relevant IDHEAS specification nodes for a given SACADA case i, and BN inference algorithms to calculate $\Pr(PIF_1 = s_1, \dots, PIF_m = s_m | \text{SACADA case } i)$. Ideally the elicited experts should know both IDHEAS and SACADA well. Nevertheless if many SACADA indicators SI_i need to be mapped on PIF specification nodes PS_j from IDHEAS the expert elicitation becomes a tedious task. To simplify this process one can assign a factor a_{PS_j,SI_i} to each SACADA indicator, which represents its effect on the PIF specification node. Qualitatively the effect of a SACADA indicator SI_i on a PIF specification node PS_j can be summarized as:

- $a_{PS_j,SI_i} = 0$, if SI_i being in state *true* causes PS_j to be in state *false* with certainty;
- $0 < a_{PS_j,SI_i} < 1$, if SI_i being in state *true* decreases the probability of PS_j being in state *true*;
- $a_{PS_j,SI_i} = 1$, if SI_i does not have an influence on PS_j;
- $a_{PS_j,SI_i} > 1$, if SI_i being in state *true* increases the probability of PS_j being in state *true*;
- $a_{PS_j,SI_i} = \infty$, if SI_i being in state *true* causes PS_j to be in state *true* with certainty.

Assuming that the joint effect of m SACADA indicators $\{SI_1, \dots, SI_m\}$ on PS_j can be expressed as the product of the factors a_{PS_j,SI_i} corresponding to SI_i, one can write:

$$\Pr\left(PS_j = true \middle| SI_1, \dots, SI_m\right) = min\left(1, \Pr\left(PS_j = true\right) \cdot \prod_{i=1}^{m} a_{PS_j,SI_i}\right) \tag{11}$$

For proof-of-concept these factors are estimated in Tab. 7 for the CFM *critical data misperceived*. It is important to note that the numbers in this table only serve the purpose of illustration. No factors are assigned to the SACADA indicators marked in grey in this table, since these indicators are redundant. From $\Pr\left(PS_j = true \middle| SI_1, \dots, SI_m\right)$ the probability $\Pr(PIF_1 = s_1, \dots, PIF_m = s_m | \text{SACADA case } i)$ can be obtained through BN inference algorithms, which are implemented in any BN software.

Table 7. SACADA indicators, which can be related to PIF specification nodes in IDHEAS. Factors used to relate the two are given in the last column.

PIF specification node (IDHEAS) (PS_j)	SACADA indicator (SI_i)	Factor $\left(a_{PS_j,SI_i}\right)$
	HSI/environment	
Environment nominal	Noisy background (Table A1, Miscellaneous)	0
	Overloaded (Table A2, Status of alarm board)	0.7
	Multiple alarms (Table B3)	0.7
Indicators easy to read and locate	Slight change (Table A3, Degree of change)	0.8
	Distinct change (Table A3, Degree of change)	1.5
	No mimics (Table A3, Degree of change)	0
	Small indications (Table A3, Degree of change)	0.7
	Similar displays (Table A3, Degree of change)	0.8
	Slight changes (Table B4)	
	Labeling/mimic display issues (Table B4)	
	Training	
Crew trained to understand the scenario	Standard (Table A4, Familiarity)	∞
	Novel (Table A4, Familiarity)	0.2
	Anomaly (Table A4, Familiarity)	0.2
	Unfamiliar (Table B6)	0
	Procedure-scenario mismatch (Table B6) → Novel	0.2
	Prior Experience (Table B6) → Anomaly	0.2
Prioritization	Competing priorities (Table A5, Uncertainty)	0.5
	Conflicting guidance (Table A5, Uncertainty)	0.5
	Competing priorities (Table B6) redundant	
	Conflicting guidance (Table B6) redundant	
	Workload	
More tasks than usual	Normal (Table A1, Workload)	0
	Concurrent demand (Table A1, Workload)	
	Multiple concurrent demands (Table A1, Workload)	2
	Multiple demands (Table A1, Miscellaneous)	
	Coordination (Table A1, Miscellaneous)	1.1
Several alarms	Dark (Table A2, Status of alarm board)	0
	Busy (Table A2, Status of alarm board)	∞
	Overloaded (Table A2, Status of alarm board)	∞
	Multiple alarms (Table B3, Background)	
	Not applicable (Table B3, Background) redundant	

7 Example results with the "critical data misperceived" BN

With the established BN for *critical data misperceived*, HEPs conditional on different observations are investigated (Tab. 8). Case I gives the prior HEP before having knowledge about the states of the PIFs or the PIF specification nodes. The states of the PIF specification nodes occur in that case according to the probabilities elicited from the experts. The BN gives reasonable prior HEPs if the CPTs of the PIF specification nodes are elicited (either based on data, experts or similar sources) and not populated with dummy parameters. The capability of giving such probabilities sets the BN apart from the decision trees originally used to quantify IDHEAS.

Cases II and III represent the extreme cases of the CFM *critical data misperceived*. In Case II, the states of all PIF specification nodes are observed and all of them are in a favorable state. The HEP is therefore minimal for that case. In Case III, all PIF specification nodes are in an unfavorable state, hence the corresponding HEP is maximal. Both cases can also be derived from the original IDHEAS decision trees. Since evidence is here given to all PIF specification nodes, it is irrelevant if the CPTs of these nodes are elicited or populated with dummy parameters.

Cases IV and V represent cases with missing information. In Case IV some of the questions corresponding to the PIF specification nodes have not been answered. The same is true for Case V, which additionally demonstrates that evidence can also be given directly at the PIF level.

Table 8. HEPs for different observations.

	Case I	Case II	Case III	Case IV	Case V
Clear display of range for	-	Yes	No	-	-
Easiness of data to read	-	Yes	No	-	-
Unambiguity	-	Yes	No	Yes	-
Environment nominal	-	Yes	No	Yes	-
Indicators easy to read and locate	-	Yes	No	Yes	-
Crew trained to understand the	-	Yes	No	Yes	-
Prioritization	-	Yes	No	-	No
Still determining plant status	-	No	Yes	Yes	No
More tasks than usual	-	No	Yes	-	No
Several alarms	-	No	Yes	Yes	No
HSI/environment	-	-		-	Good
Training	-	-		-	-
Workload	-	-		-	-
HEP	0.01	0.00005	0.5	0.002	0.0003

Besides providing evidence at the level of PIFs or PIF specification nodes, it is possible to directly give evidence on the target node. It is for example possible to determine the distribution of the PIF nodes given a HFE as $Pr(HSI/Environment\ =\ poor\ |HFE = yes) = 0.998$,

$Pr(Workload = high \,|HFE = yes) = 0.996$, $Pr(Training = poor \,|HFE = yes) = 0.532$ and $Pr(Prioritization = no \,|HFE = yes) = 0.675$.

8 Discussion

We present a comprehensive framework for the application of BNs to address shortcomings of HRA with respect to scientific basis and traceability (both causal and quantitative). A main advantage of BNs is that they allow for models that are causally traceable. To this end, unobservable PIFs and concepts from psychology can be included in the BN structure and removed in a later step. Furthermore the quantification of BNs can rely on different information sources, such as data and expert elicitations.

Causal traceability is a major need in the field of HRA. We demonstrate how an expanded BN structure can qualitatively document the theoretical background of the method. Furthermore, we show how to reduce that structure to maintain causal traceability and to enable a more straightforward quantification than the full expanded structure. While both structures are quantifiable from a mathematical point of view, quantification of the expanded structure is difficult from an HRA-perspective, since data or experts that are capable of estimating the specific probabilities are not available.

If the BN is implemented in a software tool, the additional nodes of the expanded structure can be marked in separate color, to highlight that these nodes are necessary for the understanding of the causal relationships but are not quantified. While many recently developed HRA methods have a strong background in psychological research, this background becomes usually hidden for more applied users, who are presented only a reduced number of PIFs. By developing expanded BN structures and presenting them to users, the theoretical background becomes more traceable even if it may not be possible to provide it in full detail in this manner. It has been found by many researchers that the results of a HRA vary strongly with the analyst (Lois et al., 2009). This is currently a major point of criticism against HRA methods.

An example of how the proposed framework can help to increase causal traceability was presented in this paper by the application of the framework to the CFM *critical data misperceived*. An additional dependency between the node prioritization and the target node was revealed through the process of building an exhaustive BN structure and reducing it. For the purpose of traceability of the HRA method, it is important that the model developer is aware of additional causal details like these and communicates them to the analysts. It is then up to the model developer to decide whether quantification of these causal details is necessary or not.

Another major need in HRA, which is addressed here, is an exhaustive and rigorous quantification framework. It is generally known that HRA models are not capable or even intended to fully capture all aspects of human behavior. In spite of this, it is necessary to model human error, using all information and knowledge available. Many HRA researchers rely on quantifying their models either through experts or through data. Our proposed quantification framework combines these two, which is in line with the Bayesian understanding of probability used throughout PRA (Kelly and Smith, 2009) and is the only method to come up with sound probability estimates in an industry with scarce data. Using Bayesian updating allows using continuously more data to update the parameters of the BN, in order to improve the quality of the model.

A last point implicitly addressed in this paper is the applicability of BN-based HRA methods for every-day HRA practice. While HRA researchers may be tempted to embrace BNs simply for their powerful modeling features, HRA practitioners call for models that are applicable in their everyday practice. Not many of the BN HRA models developed up to this point satisfy this need. By developing a BN which is scalable to different sizes, we offer the potential to have the same HRA method meet the needs of both practitioners and researchers.

9 Conclusion

We propose a framework for developing BN models for HRA directly from causal dependencies found in cognitive literature. The framework is illustrated through the causal paths that were identified during the development of IDHEAS. In order to develop the BN structure, a two-level approach is proposed. In a first step, identified causal paths for a crew failure mode are modeled in a qualitative BN structure. Since quantification of such a BN structure is difficult, the model is reduced in a second step using node reduction algorithms. These algorithms allow for a well-founded simplification of the model, which does neither introduce new dependencies that are not justified through the original causal paths nor neglect existing dependencies. The proposed framework thus enhances the traceability and the scientific-basis of HRA methods. As proof-of-concept, the approach is applied to the IDHEAS crew failure mode *critical data misperceived*. In this process, an additional direct dependency of this event on *prioritization* was found. We additionally propose a quantification framework for the developed BN structure that combines both expert elicitations and observed data through Bayesian updating. In HRA reality, a full agreement between experts is almost never achieved and data is scarce. The combination of the two within a well-established framework therefore represents a promising strategy for estimating human error probabilities.

Acknowledgment

The authors would like to acknowledge the support from the survey respondents as well as April Whaley (INL), Stacey Hendrickson (SNL), Susan Stevens-Adams (SNL), and the NRC (especially Jing Xing).

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

BARALDI, P., PODOFILLINI, L., MKRTCHYAN, L., ZIO, E. & DANG, V. N. 2015. Comparing the treatment of uncertainty in Bayesian networks and fuzzy expert systems used for a human reliability analysis application. *Reliability Engineering & System Safety,* 138**,** 176-193.

BEDFORD, T. & COOKE, R. 2001. Probabilistic risk analysis: foundations and methods. 2001. Cambridge: Cambridge University Press.

BIEDERMAN, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review,* 94**,** 115.

BORING, R. L., GRIFFITH, C. D. & JOE, J. C. The Measure of human error: Direct and indirect performance shaping factors. Human Factors and Power Plants and HPRCT 13th Annual Meeting, 2007 IEEE 8th, 26-31 Aug. 2007 2007. 170-176.

BROADBENT, D. 1958. Perception and communication.

CHANG, J. Y., BLEY, D., CRISCIONE, L., KIRWAN, B., MOSLEH, A., MADARY, T., NOWELL, R., RICHARDS, R., ROTH, E. M., SIEBEN, S. & ZOULIS, A. 2014. The SACADA database for human reliability and human performance. *Reliability Engineering & System Safety,* 125**,** 117-133.

COOPER, S. E., RAMEY-SMITH, A. M., WREATHALL, J., PARRY, G. W., BLEY, D., LUCKAS, W. J., TAYLOR, J. H. & BARRIERE, M. T. 1996. A Technique for Human Error Analysis (ATHEANA).

CSNI 2012. Simulator studies for HRA purposes. Budapest, Hungary: Committee on the Safety of Nuclear Installations, OECD Nuclear Energy Agency.

EINHORN, H. J. & HOGARTH, R. M. 1981. Behavioral Decision Theory: Processes of Judgement and Choice. *Annual Review of Psychology,* 32**,** 53-88.

EKANEM, N. J. & MOSLEH, A. 2014. Phoenix – A Model-Based Human Reliability Analysis Methodology: Quantitative Analysis Procedure and Data Base. *PSAM 12.* Honolulu.

EKANEM, N. J., MOSLEH, A. & SHEN, S.-H. 2016. Phoenix – A model-based Human Reliability Analysis methodology: Qualitative Analysis Procedure. *Reliability Engineering & System Safety,* 145**,** 301-315.

ENDSLEY, M. R. 1995. Towards a theory of situation awareness in dynamic systems. *Human Factors,* 37**,** 32-64.

ERIKSEN, C. & ST. JAMES, J. 1986. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics,* 40**,** 225-240.

FRENCH, S., BEDFORD, T., POLLARD, S. J. T. & SOANE, E. 2011. Human reliability analysis: A critique and review for managers. *Safety Science,* 49**,** 753-763.

GERTMAN, D., BLACKMAN, H., MARBLE, J., BYERS, J. & SMITH, C. 2005. The SPAR-H Human Reliability Analysis Method. Idaho National Laboratory.

GROTH, K. & MOSLEH, A. 2012. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model. *Journal of Risk and Reliability* 226**,** 361-379.

GROTH, K. M., SMITH, C. L. & SWILER, L. P. 2014. A Bayesian method for using simulator data to enhance human error probabilities assigned by existing HRA methods. *Reliability Engineering & System Safety,* 128**,** 32-40.

GROTH, K. M. & SWILER, L. P. 2013. Bridging the gap between HRA research and HRA practice: A Bayesian network version of SPAR-H. *Reliability Engineering & System Safety,* 115**,** 33-42.

HALLBERT, B. & KOLACZKOWSKI, A. 2007. The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study.

HOLLNAGEL, E. 2000. Looking for errors of omission and commission or The Hunting of the Snark revisited. *Reliability Engineering & System Safety,* 68**,** 135-145.

JENSEN, F. V. & NIELSEN, T. D. 2007. *Bayesian networks and decision graphs,* New York [u.a.], Springer.

KELLY, D. L. & SMITH, C. L. 2009. Bayesian inference in probabilistic risk assessment—The current state of the art. *Reliability Engineering & System Safety,* 94**,** 628-643.

KJAERULFF, U. B. & MADSEN, A. L. 2013. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis,* Springer Publishing Company, Incorporated.

KLEIN, G. A. 1998. *Sources of power: How people make decisions*, MIT press.

KÖHLER, W. 1947. *Gestalt Psychology: An Introduction to New Concets in Modern Psychology,* New York: Liveright.

LOIS, E., DANG, V. N., FORESTER, J., BROBERG, H., MASSAIU, S., HILDEBRANDT, M., BRAARUD, P., PARRY, G., JULIUS, J., BORING, R. L., MÄNNISTÖ, I. & BYE, A. 2009. International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from

Comparing HRA Methods to Simulator Performance Data. Washington D.C.: US Nuclear Regulatory Commission.

MKRTCHYAN, L., PODOFILLINI, L. & DANG, V. N. 2015. Bayesian belief Networks for Human reliability analysis: a review of applications and gaps. *Reliability Engineering & System Safety*.

MOSLEH, A. & CHANG, Y. H. 2004. Model-based human reliability analysis: prospects and requirements. *Reliability Engineering & System Safety,* 83**,** 241-253.

MUSHARRAF, M., BRADBURY-SQUIRES, D., KHAN, F., VEITCH, B., MACKINNON, S. & IMTIAZ, S. 2014. A virtual experimental technique for data collection for a Bayesian network approach to human reliability analysis. *Reliability Engineering & System Safety,* 132**,** 1-8.

O'HAGAN, A., BUCK, C. E., DANESHKHAH, A., EISER, J. R., GARTHWAITE, P. H., JENKINSON, D. J., OAKLEY, J. E. & RAKOW, T. 2006. *Uncertain judgements: eliciting experts' probabilities,* John Wiley & Sons.

PARK, J. & JUNG, W. 2007. OPERA—a human performance database under simulated emergencies of nuclear power plants. *Reliability Engineering & System Safety,* 92**,** 503-519.

PEARL, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference,* San Francisco, CA, Kaufmann.

PODOFILLINI, L., MKRTCHYAN, L. & DANG, V. N. 2014. Quantification of Bayesian Belief Net Relationships for HRA from Operational Event Analyses. *International Conference on Probabilistic Safety Assessment and Management (PSAM 12).* Honolulu, HI (USA).

RAIFFA, H. & SCHLAIFER, R. 1961. *Applied statistical decision theory,* Boston, Division of Research, Graduate School of Business Adminitration, Harvard University.

SHACHTER, R. D. 1986. Evaluating influence diagrams. *Oper. Res.,* 34**,** 871-882.

STETKAR, J. W. 2014. Human Reliability Analysis Models. *Slides 15-28 from "ACRS Meeting with THE U.S. NUCLEAR REGULATORY COMMISSION October 2, 2014.".* United States Nuclear Regulatory Commission.

STRÄTER, O. 2004. Considerations on the elements of quantifying human reliability. *Reliability Engineering & System Safety,* 83**,** 255-264.

STRAUB, D. & DER KIUREGHIAN, A. 2010. Bayesian Network Enhanced with Structural Reliability Methods: Methodology. *Journal of Engineering Mechanics-ASCE,* 136**,** 1248-1258.

SUNDARAMURTHI, R. & SMIDTS, C. 2013. Human reliability modeling for the Next Generation System Code. *Annals of Nuclear Energy,* 52**,** 137-156.

SWAIN, A. D. & GUTTMANN, H. E. 1983. Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications. Washington DC: U.S. Nuclear Regulatory Commission.

TVERSKY, A. & KAHNEMAN, D. 1974. Judgment under uncertainty: Heuristics and biases. *science,* 185**,** 1124-1131.

WARNER, N. & LETSKY, M. 2008. Empirical model of team collaboration focus on macrocognition. *Macrocognition in teams***,** 15-33.

WHALEY, A., XING, J., BORING, R. L., HENDRICKSON, S., JOE, J. & LE BLANC, K. 2012. Building a psychological foundation for human reliability analysis (NUREG-2114). Washington DC: US Nuclear Regulatory Commission.

WOODS, D. D. 1990. On taking human performance seriously in risk analysis: Comments on dougherty. *Reliability Engineering & System Safety,* 29**,** 375-381.

XING, J., PRESLEY, M., PARRY, G., FORESTER, J., HENDRICKSON, S. & DANG, V. 2013. NRC/EPRI Draft Report for Peer Review: An Integrated Decision-Tree Human Event Analysis System (IDHEAS) Method for NPP internal at-power operation. NRC/EPRI

Appendix A:

Questions provided in IDHEAS (Xing et al., 2013) for the CFM *critical data misperceived*, to specify the meaning of the PIFs:

The PIF HSI/environment is considered to be in state poor only if any of the following questions are answered with “no” (Xing et al., 2013) p. 93:

- “Are the indications clear and unambiguous”?
- “Is the information easy to read“?
- “Is the range (or band) with which the information is to be compared clearly identified on the display“?
- “Is the environment in the location of the indicator/source of information nominal (i.e., not challenging due to noise, heat, humidity, etc.)“?
- “Are the indicators/sources of data easy to locate and read“?

The PIF workload is considered to be in state high only if any of the following questions are answered with yes (Xing et al., 2013) p. 93f:

- “Does the need to obtain information occur at a time when the operators are still in the process of determining the plant status?”
- “Does this occur at a time when there are several alarms or indications or tasks that need attention? “
- “Is the scenario one for which the number of tasks the crew has to perform in the time available higher than would be typically addressed in training?”

The PIF training is considered to be in state poor only if both of the following questions are answered with yes (Xing et al., 2013) p. 94f:

- “Has the crew been properly trained to understand and deal with scenarios in which the information source may provide difficulties?”
- “Is the significance of the decision that is based on obtaining this information correctly given a high priority compared to other concurrent tasks?”

Appendix B

Deterministic rule quantifying the node *HSI/environment*:

IF

- Indications are clear and unambiguous
- AND indications are easy to read

- AND the Range (or band) with which the information is to be compared is clearly identified on the display
- AND the environment in the location of the indicator/source of information is nominal
- AND the indicators/sources of data are easy to locate and read

THEN

HSI/environment is good

ELSE

HSI/environment is poor

Deterministic rule quantifying the node *training*:

IF

- The crew has been properly trained to understand and deal with scenarios in which the information source may provide difficulties
- OR the significance of the decision that is based on obtaining this information correctly is given a high priority compared to other concurrent tasks (referred to as prioritization in the model from Fig. 6 and 7)

THEN

Training is good

ELSE

Training is poor

Deterministic rule quantifying the node *workload*:

IF

- The need to obtain information occurs at a time when the operators are still in the process of determining the plant status
- OR this occurs at a time when there are several alarms or indications or tasks that need attention
- OR the scenario is one for which the number of tasks the crew has to perform in the time available is higher than would be typically addressed in training

THEN

Workload is high

ELSE

Workload is low

Appendix C

Survey to elicit probability distributions for the PIF specification nodes:

Purpose of this survey:

The purpose of this survey is to obtain probabilities, which can be used to illustrate a framework for quantifying IDHEAS. (Note that data will be used for proof-of-concept of the mathematics, to define how IDHEAS models could be modified to include information about the probability of PIFs. IDHEAS models will not be modified based on the results of this survey.

Background:

IDHEAS provides models that assign the probability of a human failure event, given the state of several performance influencing factors (PIFs). IDHEAS contains 14 models and approximately 20 PIFs.

Description of survey task:

Consider the crew failure mode *critical Data Misperceived*, which is defined as:

"A critical piece of information that is required to develop a plant status assessment is misperceived. A critical piece of data is one that, when misperceived in a certain way will lead to an incorrect response in that it leads to taking an incorrect or inappropriate path through the procedures or executing a response incorrectly." (*From SRM Vol. 3 – Draft IDHEAS method for internal procedural events*)

We consider a post-initiator event i.e. the NPP is already in an off-normal state, where the NPP-crew is confronted with critical data. For *critical data misperceived*, the IDHEAS model identifies three main factors (PIFs), which influence human performance: Human System Interface (HSI)/environment, training, and workload. Furthermore, IDHEAS provides several questions that analysts use to assign the state of those three PIFs. In this survey, you are asked to provide a probability for each of these questions.

Note:

- A (Bayesian) probability is a degree of belief rather than an actual physically measurable quantity
- You can give your answer in **one of the two** forms:
    a. The probability of event **X** occurring is ______.
    b. Event **X** is ______-times **more**/~~**less**~~ (~~**more**~~/**less**) likely than **not X**.

Name: ____

Company: ____

Position: ____

Brief description of your experience/background:

Basis for estimates in this document (e.g., „12 years of experience operating commercial NPP" „HRA database " „22 years of experience in HRA"):

Reminder: The purpose of the survey is to obtain the probability of these conditions, NOT to obtain the human error probability for these conditions.

Human System Interface (HSI)/environment:

1. Unambiguity and clearness of indications.
    a. The probability that the indications to this data are clear and unambiguous is ___.
    b. Indications are ______-times more/less likely to be clear and unambiguous than to be unclear and ambiguous.
2. Easiness of information to read.
    a. The probability that the information is easy to read is ______.
    b. Information is ______-times **more/less** likely to be easy to read than to be not easy to read.
3. Range (band) for comparison.
    a. The probability that the range (or band), with which the information is to be compared, is clearly identified on the display is ______.
    b. A display is ______-times **more/less** likely to have a clearly identified range (band) than to have a not clearly identified range (band).
4. Nominal environment.

   a. The probability of having a nominal environment (i.e. one that is not challenging due to noise, heat, humidity, etc.) is ______.
   b. During the event it is ______-times **more/less** likely that environment is nominal rather than non-nominal environment (challenging due to noise, heat, humidity, etc.).
5. Location and easiness to read of the indicators/sources.
   a. The probability that indicators/sources are easy to locate and read is ______.
   b. Indicators/sources are ______-times **more/less** likely to be easy to read and locate than to be not easy to read and locate.

Workload:

6. Crew still determining the plant status.
   a. The probability that the need to obtain information occurs at a time when the operators are still in the process of determining the plant status is______.
   b. Crew still determining the plant status is ______-times **more/less** likely than crew not in the process of determining the plant status.
7. Several alarms.
   a. The probability that the need to receive critical data occurs at a time when there are several alarms, indications or tasks that need attention is ______.
   b. A situation with Several alarms, indications or tasks that need attention at the same time is ______-times **more/less** likely than a situation without several alarms, indications or tasks.
8. More tasks in the available time than typically addressed in training.
   a. The probability that the number of tasks the crew has to perform in the available time is higher than it would be typically addressed in training is ____.
   b. A situation with a higher number of tasks than addressed in training is ______-times **more/less** likely than a situation with same or a lower number of tasks than typically addressed in training.

Training:

9. Crew trained to understand the scenario.
   a. The probability that the crew has been properly trained to understand and deal with the scenarios, in which the information source may provide difficulties is ______.
   b. Crew properly trained to understand and deal with the scenario is ______-times **more/less** likely than crew not properly trained to understand and deal with the scenario.
10. Significance of the decision that is based on this information.
   a. The probability that the decision based on obtaining this information correctly is given a high priority compared to other concurrent tasks is ______.

Giving the decision, which is based on the information, a high priority is ______-times **more/less** likely than not giving the decision a high priority.